

Training anamnesis interviews with a chatbot virtual patient and integrated feedback: healthcare students' perspectives

Cornelia Schlick

cornelia.schlick@iu.org

IU International University of Applied Sciences

Katharina Rädel-Ablass

IU International University of Applied Sciences

Klaus Schliz

IU International University of Applied Sciences

Claudia Miersch

IU International University of Applied Sciences

Research Article

Keywords: virtual patient, large language model, anamnesis training, health professions education, automated feedback

Posted Date: August 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10473687/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Effective anamnesis is a core clinical competency, yet training opportunities with real or standardized patients are resource-intensive, difficult to scale, and particularly limited in online learning environments. Large language model (LLM)-powered virtual patients (VPs) offer a promising alternative, but most existing evidence focuses on medical students, and the integration of automated feedback and its perception across other health professions remains underexplored. This study examined healthcare students' perceptions of an enhanced GPT-4-powered VP with integrated automated feedback on professional and communication competence.

Methods

A cross-sectional study was conducted at IU International University of Applied Sciences (January–March 2025) among undergraduate students from eight health-related Bachelor's programs, including nursing, physiotherapy, occupational therapy, speech therapy, and nutritional science. Participants conducted an anamnesis interview with a chatbot-based VP simulating a patient with brain hemorrhage and multimorbidity, then received automated feedback before completing an online survey. Usability was assessed with the Chatbot Usability Questionnaire (CUQ); response quality, realism, behavioral engagement, feedback quality, and training benefit were assessed using purpose-developed, pilot-tested scales (5-point Likert), analyzed via medians, Cronbach's alpha, and Positive Response Proportions. Open-ended responses underwent qualitative content analysis following Mayring.

Results

Sixty students participated (83.3% female; mean age 37.0 years). Usability was rated high (CUQ mean 76.93, SD 12.51). Response quality and feedback quality received the highest ratings (Cronbach's $\alpha = 0.90$ and 0.86 , respectively), while realism scored lower ($\alpha = 0.87$). Despite limited perceived naturalness, most students reported behaving professionally, respectfully, and empathically toward the VP. Educational value was rated highly, especially regarding professional competence and communication skills. Qualitative findings highlighted interaction realism and the integrated feedback mechanism as the most valued aspects, alongside requests for more unpredictable patient behavior and multimodal interaction formats.

Conclusions

AI-based virtual patients with integrated automated feedback are perceived as a valuable, scalable tool for anamnesis training across diverse health professions, extending prior evidence beyond medical

education. While realism and emotional expressiveness remain areas for improvement, findings support the use of such systems as an effective formative learning tool. Future research should investigate the transfer of acquired skills to clinical practice and long-term learning outcomes.

INTRODUCTION

Effective anamnesis, commonly referred to as history taking, is a cornerstone of clinical practice across all health professions, forming the basis for accurate diagnosis, appropriate treatment planning, and the establishment of therapeutic patient–provider relationships.^{1,2} The systematic acquisition of anamnesis skills remains challenging in health professions education, particularly given the limited opportunities for repeated, standardized practice with real patients.³ This competency is equally essential for nurses, physiotherapists, occupational therapists, speech therapists, and nutritional scientists, all of whom rely on comprehensive patient interviews to deliver individualized care.⁴

Traditional training approaches, such as standardized patients (SPs) or peer role-plays, offer valuable opportunities for practicing patient communication.⁵ However, these methods face significant limitations, including high resource requirements, scheduling constraints, limited scalability, and substantial costs associated with training and compensating SPs.^{6,7} These challenges are particularly pronounced in online and distance learning settings, where access to face-to-face simulation experiences is inherently restricted.⁸ Furthermore, students may experience psychological stress during interactions with real or standardized patients, which can hinder their learning progress.⁹

Recent advances in artificial intelligence (AI), particularly large language models (LLMs) such as GPT-4, have opened new possibilities for health professions education.^{10,11} Several studies have demonstrated the potential of chatbot-powered virtual patients (VP) for practicing clinical reasoning and communication skills. Holderried et al. developed a GPT-powered chatbot that enabled medical students to practice anamnesis, reporting high medical plausibility of responses and positive usability ratings.⁹ Brügge et al. found that medical students who received LLM-generated structured feedback after simulated patient conversations significantly outperformed the control group in clinical decision-making.¹² Similarly, Wang et al. demonstrated that GPT-based simulations effectively support anamnesis training in medical students, with automated feedback on performance dimensions including history completeness and clinical reasoning.¹³

Although virtual patient simulations have been used in health professions education for some time, systems prior to the advent of LLMs have often been criticized for offering limited or no feedback – a critical shortcoming given that feedback is indispensable for the development of clinical and communication competencies.¹⁴ The linguistic capabilities of LLMs mark a substantial advance in this regard, enabling automated feedback that is more detailed, adaptive, and educationally meaningful than that of earlier VP systems, including unlimited availability, consistent responses, and immediate, personalized feedback at scale.¹⁴ Recent research further suggests that LLM-generated feedback on

anamnesis performance can approximate expert evaluation, supporting its potential as a scalable tool for health professions education.¹⁵

Despite these promising developments, important research gaps remain. Most existing studies have focused on medical students, leaving limited evidence on the applicability and effectiveness of AI-powered virtual patients for other health professions.^{9,12,13} Although Potter et al. recently piloted a conversational AI tool with nursing students and reported improved performance and high acceptability,¹⁶ broader evidence across health professions is still lacking. In addition, the systematic integration of AI-generated feedback targeting both professional competence and communication skills has received limited attention,¹⁵ and students' self-assessment of their behavior and engagement during virtual patient interactions remains underexplored.¹⁷

Building on our previous work, in which students from diverse healthcare programs rated a GPT-4-powered virtual patient positively in terms of realism, professional accuracy, and usability,¹⁸ the present study examined healthcare students' perceptions of an enhanced virtual patient system with integrated automated feedback. Specifically, we addressed the following research questions:

1. How do healthcare students evaluate the VP with regards to usability, response quality, and realism?
2. How do healthcare students assess their own attitude and behavior during interaction with the VP?
3. How do healthcare students perceive the quality of chatbot-generated feedback?
4. How do healthcare students appraise the educational value of chatbot-based anamnesis training?

METHODS

Study design and setting

This study employed a cross-sectional design to investigate healthcare students' perceptions of a VP system and its associated chatbot feedback. The study was conducted at the IU International University of Applied Sciences between January 15, and March 31, 2025.

Participants and recruitment

Eligible participants were undergraduate students enrolled in the Bachelor's programs in Nursing Education, Nursing Management, Health Education, Nursing, Physiotherapy, Occupational Therapy, Speech Therapy, and Nutritional Sciences. All students enrolled in these programs were invited to participate via email. Recruitment consisted of two calls for participation: one at the start of the study period and another midway through. The expected number of respondents was between 50 and 100.

Procedure

Participants received a study link via email and were introduced to the clinical case scenario, including a written description of the patient room and the option to view a photograph to enhance contextual realism. Only minimal patient information was provided (name and the presence of his wife); no clinical details were disclosed in advance.

Participants were instructed to conduct an anamnesis interview from the perspective of their respective profession, beginning with a greeting and ending with a farewell while structure and content were left to participants' professional judgment.

Upon completion, participants automatically received chatbot-generated feedback. They were then redirected to the online survey to complete the evaluation.

VP Implementation

The VP was powered by GPT-4. The model was deployed via IU International University of Applied Sciences' Guided Conversation Designer (GCD) platform, which allows structured prompt design and systematic recording of all interactions. The case scenario involved a patient ("Karl von Hausen") presenting with a brain hemorrhage following a bicycle accident, accompanied by language disorder, mild dysphagia, right hemiparesis, urinary diversion, hypertension, and obesity. The chatbot simulated both the patient and his wife. Prompt development and validation followed an interdisciplinary approach; a detailed description is provided in our previous publication.¹⁸

Feedback Implementation

The VP prompt was extended to provide automated feedback immediately following the anamnesis interview, addressing two key domains: professional competence (accuracy, completeness, and question depth and breadth, tailored to the student's professional background) and communication competence (clarity, empathy, and interaction skills). Feedback content and delivery were iteratively refined and validated with faculty from multiple healthcare disciplines to ensure relevance, clarity, and educational value. The feedback prompt is provided in the supplementary material.

Measures

The survey instrument comprised eight sections. Section 1 collected demographic and educational background information, including study program, study progression, prior vocational healthcare training, work experience, anamnesis experience, age, gender, and prior experience with AI-based learning. These variables were assessed to characterize participants' professional background and prior exposure to clinical practice and AI-supported educational tools.

Section 2 included the Chatbot Usability Questionnaire (CUQ),¹⁹ a validated 16-item instrument assessing perceived chatbot usability across dimensions such as user experience, personality, error

handling, and onboarding. Items were rated on a 5-point Likert scale. Following published scoring guidelines, responses were converted into a standardized score ranging from 0 to 100, with higher scores indicating better perceived usability.

Sections 3–7 assessed participants' perceptions of the VP system and its integrated feedback. These scales were specifically developed for the present study to capture five constructs: response quality (12 items), realism (7 items), behavioral engagement with the VP (5 items), feedback quality (8 items), and training benefit (5 items). Item development followed an interdisciplinary process involving healthcare educators from multiple disciplines to ensure content validity. The items were pilot-tested and revised prior to implementation. All items were formulated as statements and rated on a 5-point Likert scale.

Section 8 included open-ended questions to capture qualitative evaluations and suggestions for improvement.

The complete translated survey instrument is provided in the Supplementary Material.

Data analysis

Demographic and educational characteristics were analyzed descriptively using frequencies, percentages, means, and standard deviations as appropriate. For the CUQ, the overall mean score across the 16 items was calculated, and item-level descriptives were examined.

For the composite scales derived from Sections 3–7, internal consistency was assessed using Cronbach's alpha to justify aggregation into scale scores. Descriptive statistics (median, interquartile range) were calculated for each scale. Item-level analyses were conducted using the Positive Response Proportion (PRP), defined as the proportion of participants selecting the top two Likert-scale categories (ratings of 4 or 5). Negatively worded items were reverse coded to ensure that higher PRP values consistently reflected more favorable evaluations.

Open-ended responses were analyzed using qualitative content analysis following Mayring,²⁰ applying a content-structuring approach. An inductive–deductive procedure was used: main categories were derived deductively from the research questions, while subcategories emerged inductively from the data. Responses were paraphrased, grouped thematically, and iteratively assigned to categories. Coding was conducted by one researcher (KRA), supported by MAXQDA (Version 26). Approximately 20% of responses were independently coded by a second researcher; discrepancies were resolved through discussion and consensus. The final category system is provided in the Supplementary Material.

To further characterize VP interactions and feedback, descriptive analyses were conducted. VP conversations (n = 53; seven feedback texts could not be matched to their corresponding conversations and were therefore excluded) were examined for structural features, including the number of messages and total word count. Feedback texts were similarly analyzed for word count, the ratio of conversation-to-feedback words, and a qualitative feedback score (Table S1).

Ethical considerations and AI contributions

Ethical approval

for the study protocol was obtained from the IU International University Ethics Committee. Participation was voluntary, anonymous, and carried no academic consequences for students. Informed consent to participate was obtained from all participants prior to involvement in the study.

Language editing and stylistic refinement of the online survey and this manuscript were supported by ChatGPT (OpenAI, accessed 2025/6) and Claude (Anthropic, accessed 2026). All other aspects, including content, analysis, interpretation, and conclusions, remain the sole responsibility of the authors.

RESULTS

Demographics

A total of 60 students participated, of whom 50 (83.3%) identified as female and 10 (16.7%) as male, with a mean age of 37.03 years (SD = 9.14, range 20–61). They were enrolled in a variety of healthcare-related study programs and represented all stages of study progression (Table 1). The vast majority of participants had completed professional training prior to starting their academic studies. Only 7 (12%) participants indicated no prior healthcare training, while 34 (57%) had training in nursing, 10 (17%) in physical or occupational therapy, and 9 (15%) in other healthcare-related fields. Among those with prior training, 46 (87%) participants reported more than five years of work experience in the healthcare sector.

A large proportion of participants reported having prior experience using AI-based tools in their learning processes, and about three-quarters stated having practical experience with conducting anamnesis interviews (Table 1).

Table 1
Educational characteristics of the sample

Characteristic	Category	n (out of N = 60)	%
Study program	Bachelor nursing education	15	25.0
	Bachelor nursing management	10	16.7
	Bachelor nutritional science	10	16.7
	Bachelor physical therapy	9	15.0
	Bachelor nursing care	7	11.7
	Bachelor health education	7	11.7
	Bachelor occupational therapy	1	1.7
	Bachelor physician assistant	1	1.7
Study progression	Up to ¼ modules completed	13	21.7
	¼ to ½ of modules completed	16	26.7
	½ to ¾ of modules completed	16	26.7
	More than ¾ of modules completed	15	25.0
Anamnesis experience (self-rated)	No	12	20.0
	Yes	48	80.0
	- Expert level	12	25.0
	- Advanced level	33	69.0
	- Beginner level	3	6.0
AI experience in learning contexts	No	15	25.0
	Yes	44	73.3
	Not sure	1	1.7

Descriptives of the VP conversations and feedback

On average, a conversation consisted of 47.2 messages (SD = 30.5) and 1,192 words (SD = 571.3). The feedback contained, on average, 478.7 words (SD = 125.7), resulting in a mean conversation-to-feedback word ratio of 2.66 (SD = 1.5).

The qualitative feedback score was assessed using five predefined criteria addressing completeness, structure, and profession-specific relevance. Criteria were scored dichotomously (present = 1, absent =

0), resulting in a total score ranging from 0 to 5, with higher scores reflecting higher feedback quality (see Table S1). Each criterion was present in more than 83% of texts. Over 90% of feedback texts received a score of 4 or higher, and 35 of 53 texts (66%) met all quality criteria. Only one text received a score of 2 (see Figure S2).

Students' evaluation of the VP

The overall CUQ mean score was 76.93 [SD 12.51], indicating generally high usability. Item-level patterns are illustrated in Fig. 1. Participants showed the least agreement with error handling items, and around 20% indicated that the chatbot "seemed too robotic" and "was very complex".

For the Response Quality scale, students reported generally positive perceptions of the VP's response behavior. The median score was between 4 and 5 on the 5-point Likert scale (Fig. 2), with excellent internal consistency (Cronbach's $\alpha = 0.90$). PRPs for the 12 items ranged from 68% ($n = 41$, "follow-up questions answered") to 95% ($n = 57$, "credible responses").

For the Realism scale, the median score was 4 (Fig. 2), with good internal consistency (Cronbach's $\alpha = 0.87$). PRPs for the 7 items were generally lower than for response quality, ranging from 52% ($n = 31$, "reacted unpredictably like a real patient") to 86% ($n = 53$, "used terms like a real patient").

Students' attitudes and behavior during interaction with VP

For the Behavioral Engagement with the VP scale, the median score was 4 (Fig. 2), with good internal consistency (Cronbach's $\alpha = 0.76$). PRPs for the 5 items are displayed in Fig. 3. They span from 53% ($n = 32$, "felt natural speaking with the virtual patient") to 95% ($n = 57$, "acted respectfully toward the VP").

Students' perception of the feedback quality

For the Feedback Quality scale, the median score ranged between 4 and 5 (Fig. 2), with good internal consistency (Cronbach's $\alpha = 0.86$). PRPs for the 7 items are shown in Fig. 3. They were the highest across all scales, ranging from 82% ($n = 49$, "feedback necessary for learning success") to 97% ($n = 58$, "feedback was related to the content of the anamnesis").

Higher values indicate more favorable participant perceptions; Items marked with an asterisk (*) deviate from the original survey wording: they were initially negatively phrased and subsequently reverse-coded to ensure consistent interpretation.

Students' appraisal of the educational value

The educational value of VP-based training was assessed using both the Perceived Training Benefit scale and open-ended questions.

For the scale, the median score ranged between 4 and 5 (Fig. 2), with excellent internal consistency (Cronbach's $\alpha = 0.89$). PRPs for the 5 items ranged from 60% ($n = 36$, "prepare for challenging patients") to 85% ($n = 51$, "enhance professional competence"). The remaining other items also showed high PRPs:

“improve communication skills” (83%, n = 50), “transfer skills to real situations” (82%, n = 49), and “gain confidence with real patients” (PRP 80%, n = 50).

Qualitative responses are summarized in Fig. 4 (categories are detailed in Supplementary Table S2). Regarding the most beneficial aspects, interaction and realism of the conversation were the most frequently mentioned theme, followed by the integrated feedback mechanism. Further themes included the structure of the anamnesis, the inclusion of a family member, and the focus on communication flow. Suggested improvements centered on realism and complexity of the patient simulation, with participants requesting more unpredictable or challenging patient behavior, as well as more background information and clinical context prior to the interview, visual presentation of the patient, and voice-based interaction. Regarding perceived advantages, participants emphasized the value of practicing in a safe, consequence-free environment, highlighting error tolerance, and the opportunity to apply theoretical knowledge without the pressure of real clinical encounters. Disadvantages were primarily associated with the absence of nonverbal communication, which participants identified as limiting the authenticity of the interaction, alongside concerns regarding the emotional depth of the simulation.

DISCUSSION

Summary of findings

Overall, students evaluated the VP system and its integrated feedback positively across all assessed dimensions. The CUQ indicated generally high usability, although limitations in error handling and a somewhat robotic perception were noted by some participants. Response quality was rated highly, while realism, particularly emotional expressiveness and unpredictability, received more moderate ratings. Students' behavioral engagement revealed that most were able to act professionally, respectfully, and empathically toward the VP, even when the interaction was not perceived as fully natural or immersive. Feedback quality received the highest ratings across all scales, and educational value was consistently rated as high, particularly regarding professional competence and communication skills.

Interpretation of findings

Evaluation of the VP

The usability findings align with Holderried et al.,⁹ with both studies reporting a mean CUQ score of 77, indicating high usability. Handling errors and misunderstandings remained a challenge, with the lowest agreement for the item “coped well with errors or mistakes” (~ 40% in our study vs. 35% in Holderried et al.), likely due to limitations in natural language processing and the system's ability to flexibly interpret ambiguous or incomplete input. Differences in perceived complexity (20% in our study vs. 90% in Holderried et al.) may reflect variations in technical design, interface, or scenario characteristics rather than task difficulty per se.

The present findings on response quality and realism are broadly consistent with our previous study.¹⁸ Students generally perceived the VP's answers as coherent, contextually appropriate, and reliable, supporting its role in reinforcing structured anamnesis questioning and clinical knowledge recall. Perceptions of realism were more limited, particularly regarding emotional expressiveness, consistent with a recent review highlighting the difficulty of simulating authentic emotion in VP interactions.²¹ Students indicated a clear preference for more immersive interaction formats, such as audio-based or avatar systems, suggesting that incorporating multimodal components may further strengthen the experiential and relational aspects of communication training.

Attitudes and behavior during interaction with VP

Although participants did not consistently perceive the VP interaction as fully natural or immersive, they nevertheless reported high levels of professional, respectful, and empathic behavior. This indicates that learners can engage meaningfully and demonstrate key communication skills even when the scenario is not experienced as entirely realistic. In line with this, knowledge acquisition in emergency medicine training was found to be equivalent between low fidelity and high fidelity simulations, suggesting that higher levels of realism are not necessarily decisive for achieving learning outcomes.²²

Perceptions of feedback quality

Our findings suggest that AI-based systems, when carefully designed and appropriately prompted, can provide elaborate feedback immediately after task completion. Participants perceived this as helpful, pointing to practical advantages of AI, particularly through the provision of frequent and immediate feedback at scale. At the same time, these findings must be interpreted in light of the substantial heterogeneity of AI applications, educational contexts, and outcome measures. In medical education, especially, the quality, accuracy, and contextual appropriateness of AI-generated feedback therefore remain critical. Overall, our results align with research showing that feedback has a measurable moderate to strong impact on learning outcomes and that its effectiveness depends less on delivery mode or source than on the quality of information and the content of the feedback itself. This is particularly relevant because elaborated, immediate, and explanatory feedback tends to produce greater learning gains than simple corrective feedback or merely providing the correct answer.^{14,23–27}

Appraisal of educational value

The educational value of the VP-based training was rated very positively. Regarding realism, a more nuanced picture emerged: participants appreciated the authentic quality of the interaction but also expressed a desire for more complex and unpredictable scenarios, likely reflecting the clinically experienced composition of the sample. This aligns with recommendations to use simpler scenarios for novice learners to avoid cognitive overload, while more complex cases are appropriate for experienced personnel.²⁸ Students also emphasized the value of including a family member in the simulation, highlighting the importance of incorporating socially challenging interactions in training design. AI-based

VP systems offer a flexible framework for such adaptation, enabling the rapid generation of diverse cases with adjustable difficulty.

Educational implications

The present findings have several implications for the use of AI-based VP in health professions education. First, such systems appear particularly suitable for formative learning settings, supporting iterative practice and skill development rather than high-stakes assessment.²⁹ The consistently high ratings of feedback quality underscore the potential of AI-based systems to provide timely, structured guidance during supervised learning processes.

Second, simulation design should be aligned with learning objectives and learner experience rather than striving for maximal realism, with simpler cases for novice learners and richer, less predictable scenarios for more advanced students promoting higher-order clinical reasoning.²⁹

Third, the results point to the need to further develop AI-based VP beyond cognitive skill training. Although response quality was rated highly, lower ratings for emotional expressiveness suggest that current systems remain limited in supporting aspects such as empathy and relational communication. Furthermore, multimorbidity is underrepresented in many AI-based VP systems despite its clinical relevance.³⁰ Our study demonstrates that complex cases—such as a patient with traumatic brain injury and pre-existing conditions—can be satisfactorily represented, showing that AI can support more realistic and clinically challenging training scenarios. Future developments should therefore aim to better integrate these dimensions, for example through more immersive or multimodal interaction formats.

Importantly, the present study extends existing research, which has predominantly focused on medical students and residents,²⁹ by demonstrating that students from other health professions—including nursing, physiotherapy, and nutrition science—also perceive substantial educational value in chatbot-based anamnesis training. This suggests that AI-based VP is a versatile tool that can be applied across health professions to support the development of communication and clinical reasoning skills.

Finally, the implementation of such systems should be accompanied by appropriate pedagogical integration and supervision. As AI-generated outputs may vary and are not free from inaccuracies, their use requires structured guidance and opportunities for critical reflection. Embedding AI-based training within existing curricula and combining it with human or peer-supported feedback may help maximize its educational effectiveness while mitigating potential risks.²⁹

Limitations

Several limitations of this study should be considered. First, the sample consisted of students from a single university who already had experience with AI-based learning tools, and students with prior professional experience were overrepresented, limiting generalizability. Second, voluntary participation may have introduced self-selection bias, potentially inflating positive evaluations. Third, the study relied

on self-reported data, which are inherently subjective and may be influenced by social desirability. Fourth, the study captures immediate perceptions only and provides no evidence on whether experiences translate into long-term competence gains or improved patient outcomes.

In addition, AI-generated feedback was only analyzed structurally and evaluated by students, so it remains unclear how accurately it reflected actual anamnesis performance. Although students rated the feedback very positively, positive reception does not necessarily translate into improved learning outcomes. Research indicates that students often perceive human feedback as more credible and empathetic, whereas AI-generated feedback is seen as clearer and more detailed.^{31,32} Peer feedback and hybrid models combining AI and human feedback have shown high potential by combining credibility with detailed guidance.³³

Moreover, the dynamic development of AI models introduces variability in responses, as identical queries can yield different outputs over time. While this may challenge consistency in medical recommendations, it can also create realistic learning opportunities, reflecting situations where patients provide contradictory or incomplete information, requiring students to engage critically with AI-generated content.^{34,35}

Conclusion and directions for future research

This study demonstrates that AI-based VPs are a highly valued tool for practicing anamnesis across multiple health professions, including nursing, physiotherapy, and nutrition science. Students rated usability, response quality, and educational value positively, and reported being able to engage professionally, respectfully, and empathically during interactions even when scenarios were not fully immersive. Feedback was perceived as timely, structured, and supportive for learning. The findings further suggest that AI-based VPs can accommodate complex cases, including patients with multimorbidity, offering scalable and adaptable training opportunities tailored to learners' experience and competence level. While high realism and emotional expressiveness remain challenging, incorporating multimodal elements and carefully designed scenarios can enhance authenticity and support communication skills development.

Despite the limitations noted above, the study provides strong support for AI-based VP training as an effective formative learning tool across diverse health professions. Future studies should examine whether competencies acquired in virtual simulations transfer to real clinical settings, investigate long-term skill retention and patient outcomes, and explore the effects of different case complexities, multimodal formats, and adaptive AI-driven scenarios on learning outcomes.

Declarations

Ethics approval and consent to participate

All procedures adhered to established ethical standards. Participation was voluntary, anonymous, and carried no academic consequences for students. Ethical approval for the study protocol was obtained from the IU International University Ethics Committee.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by IU International University Incubator funding (Ref. ID: KI-Patient IPL).

Author Contribution

CS, KRA, KS and CM conceived and designed the study, and developed the prompt. CS and CM additionally developed the questionnaire. CS, KRA and CM collected and analysed the data. CS drafted the manuscript; KRA, KS and CM revised it. All authors read and approved the final manuscript.

Acknowledgement

The authors thank all students who participated in the study, and our research department for the financial and professional support.

Data Availability

The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

References

1. Keifenheim KE, Teufel M, Ip J, et al. Teaching history taking to medical students: a systematic review. *BMC Med Educ.* 2015;15(1):159. 10.1186/s12909-015-0443-x.
2. Kaplonyi J, Bowles K, Nestel D, et al. Understanding the impact of simulated patients on health care learners' communication skills: a systematic review. *Med Educ.* 2017;51(12):1209–19. 10.1111/medu.13387.

3. Lee J, Kim H, Kim KH, Jung D, Jowsey T, Webster CS. Effective virtual patient simulators for medical communication training: A systematic review. *Med Educ*. 2020;54(9):786–95. 10.1111/medu.14152.
4. Kelly S, Smyth E, Murphy P, Pawlikowska T. A scoping review: virtual patients for communication skills in medical undergraduates. *BMC Med Educ*. 2022;22(1):429. 10.1186/s12909-022-03474-9.
5. Eklics K, Csongor A, Hambuch A, Fekete J. Diverse Integration of Simulated Patients in Medical Education for Communication, Language, and Clinical Skills in Hungary. *Adv Med Educ Pract*. 2024;15:301–12. 10.2147/AMEP.S438102.
6. Räuchle J, Kühne F, Zacharias A, Große G, Weck F, Maaß U. Insights From Roleplays With Standardised Patients and Live Supervision for Psychotherapy and Counselling Training—A Qualitative Study. *Couns Psychother Res*. 2025;25(2):e70018. 10.1002/capr.70018.
7. Stamer T, Steinhäuser J, Flägel K. Artificial Intelligence Supporting the Training of Communication Skills in the Education of Health Care Professions: Scoping Review. *J Med Internet Res*. 2023;25:e43311. 10.2196/43311.
8. Liaw SY, Tan JZ, Lim S, et al. Artificial intelligence in virtual reality simulation for interprofessional communication training: Mixed method study. *Nurse Educ Today*. 2023;122:105718. 10.1016/j.nedt.2023.105718.
9. Holderried F, Stegemann–Philipps C, Herschbach L, et al. A Generative Pretrained Transformer (GPT)–Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study. *JMIR Med Educ*. 2024;10:e53961. 10.2196/53961.
10. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR Med Educ*. 2023;9:e48291. 10.2196/48291.
11. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and Generative Artificial Intelligence for Medical Education: Potential Impact and Opportunity. *Acad Med*. 2024;99(1):22–7. 10.1097/ACM.0000000000005439.
12. Brügge E, Ricchizzi S, Arenbeck M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Med Educ*. 2024;24(1):1391. 10.1186/s12909-024-06399-7.
13. Wang Z, Fan TT, Li ML, Zhu NJ, Wang XC. Feasibility study of using GPT for history-taking training in medical education: a randomized clinical trial. *BMC Med Educ*. 2025;25(1):1030. 10.1186/s12909-025-07614-9.
14. Wisniewski B, Zierer K, Hattie J. The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Front Psychol*. 2020;10:3087. 10.3389/fpsyg.2019.03087.
15. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, et al. A Language Model–Powered Simulated Patient With Automated Feedback for History Taking: Prospective Study. *JMIR Med Educ*. 2024;10:e59213. 10.2196/59213.
16. Potter J, Theobald K, Johnston S. Acceptability and satisfaction of a conversational AI tool for communication skills training in preregistration nursing: A pilot study. *Clin Simul Nurs*. 2026;110:101872. 10.1016/j.ecns.2025.101872.

17. Mukadam A, Suresh S, Jacobs C. Beyond Traditional Simulation: An Exploratory Study on the Effectiveness and Acceptability of ChatGPT–4o Advanced Voice Mode for Communication Skills Practice Among Medical Students. *Cureus* Published online May. 2025;19. 10.7759/cureus.84381.
18. Rädcl-Abläss K, Schliz K, Schlick C, et al. Teaching opportunities for anamnesis interviews through AI based teaching role plays: a survey with online learning students from health study programs. *BMC Med Educ*. 2025;25(1):259. 10.1186/s12909-025-06756-0.
19. Holmes S, Bond R, Moorhead A, Zheng J, Coates V, McTear M. Towards Validating a Chatbot Usability Scale. In: Marcus A, Rosenzweig E, Soares MM, editors. *Design, User Experience, and Usability*. Springer Nature Switzerland; 2023. pp. 321–39.
20. Mayring P. *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. 13., überarbeitete Auflage. Beltz; 2022.
21. Zeng J, Qi W, Shen S, et al. Embracing the Future of Medical Education With Large Language Model–Based Virtual Patients: Scoping Review. *J Med Internet Res*. 2025;27:e79091. 10.2196/79091.
22. Offenbacher J, Petti A, Xu H, et al. Learning Outcomes of High-fidelity versus Table-Top Simulation in Undergraduate Emergency Medicine Education: Prospective, Randomized, Crossover-Controlled Study. *West J Emerg Med*. 2021;23(1):20–5. 10.5811/westjem.2021.12.53926.
23. Morris R, Perry T, Wardle L. Formative assessment and feedback for learning in higher education: A systematic review. *Rev Educ*. 2021;9(3):e3292. 10.1002/rev3.3292.
24. Cai Z, Gui Y, Mao P, et al. The effect of feedback on academic achievement in technology-rich learning environments (TREs): A meta-analytic review. *Educ Res Rev*. 2023;39:100521. 10.1016/j.edurev.2023.100521.
25. Kaliisa R, Misiejuk K, López-Pernas S, Saqr M. How does artificial intelligence compare to human feedback? A meta-analysis of performance, feedback perception, and learning dispositions. *Educ Psychol*. 2026;46(1):80–111. 10.1080/01443410.2025.2553639.
26. Alghamdi LH, Alghizzi TM. Educators’ reflections on AI-automated feedback in higher education: a structured integrative review of potentials, pitfalls, and ethical dimensions. *Front Educ*. 2025;10:1704820. 10.3389/feduc.2025.1704820.
27. Qadir HM, Suleman MT, Khan RA, Sohaib M, Hasan MJ, Hussain SA. Optimizing learning outcomes: a deep dive into hybrid AI models for adaptive educational feedback. *J Big Data*. 2025;12(1):144. 10.1186/s40537-025-01187-6.
28. Bischof L, Steinacker AC, Von Trzebiatowski I, Klingenberg M. Simulationsbasierte Lehre in der Fort- und Weiterbildung von Ersteinschätzenden in Notaufnahmen – ein narratives Review. *Notf Rettungsmedizin*. Published online September 2, 2025. 10.1007/s10049-025-01620-6
29. Gandhi B, Morjaria L, Illamperuma I, Nadesan P, Arora A, Sibbald M. Large Language Model-Based Virtual Patients for Simulated Clinical Learning: A Scoping Review. *AI Med*. 2026;1(1):7. 10.3390/aimed1010007.
30. Li D, Lebai Lutfi S. Large Language Model–Based Virtual Patient Systems for History-Taking in Medical Education: Comprehensive Systematic Review. *JMIR Med Inf*. 2026;14:e79039.

10.2196/79039.

31. Kotlyar I, Krasman J. Student reactions to AI versus human feedback in teamwork skills assessment. *Int J Educ Technol High Educ*. 2025;22(1):57. 10.1186/s41239-025-00555-9.
32. Escalante J, Pack A, Barrett A. AI-generated feedback on writing: insights into efficacy and ENL student preference. *Int J Educ Technol High Educ*. 2023;20(1):57. 10.1186/s41239-023-00425-2.
33. Zhang A, Gao Y, Suraworachet W, Nazaretsky T, Cukurova M. Evaluating Trust in AI, Human, and Co-produced Feedback Among Undergraduate Students. *arXiv*. Preprint posted online 2025. 10.48550/ARXIV.2504.10961
34. Gaud D. Ethical Considerations for the Use of AI Language Model. *Int J Res Appl Sci Eng Technol*. 2023;11(7):6–14. 10.22214/ijraset.2023.54513.
35. Slimi Z, Villarejo Carballido B. Navigating the Ethical Challenges of Artificial Intelligence in Higher Education: An Analysis of Seven Global AI Ethics Policies. *TEM J* Published online May. 2023;29:590–602. 10.18421/TEM122-02.

Figures

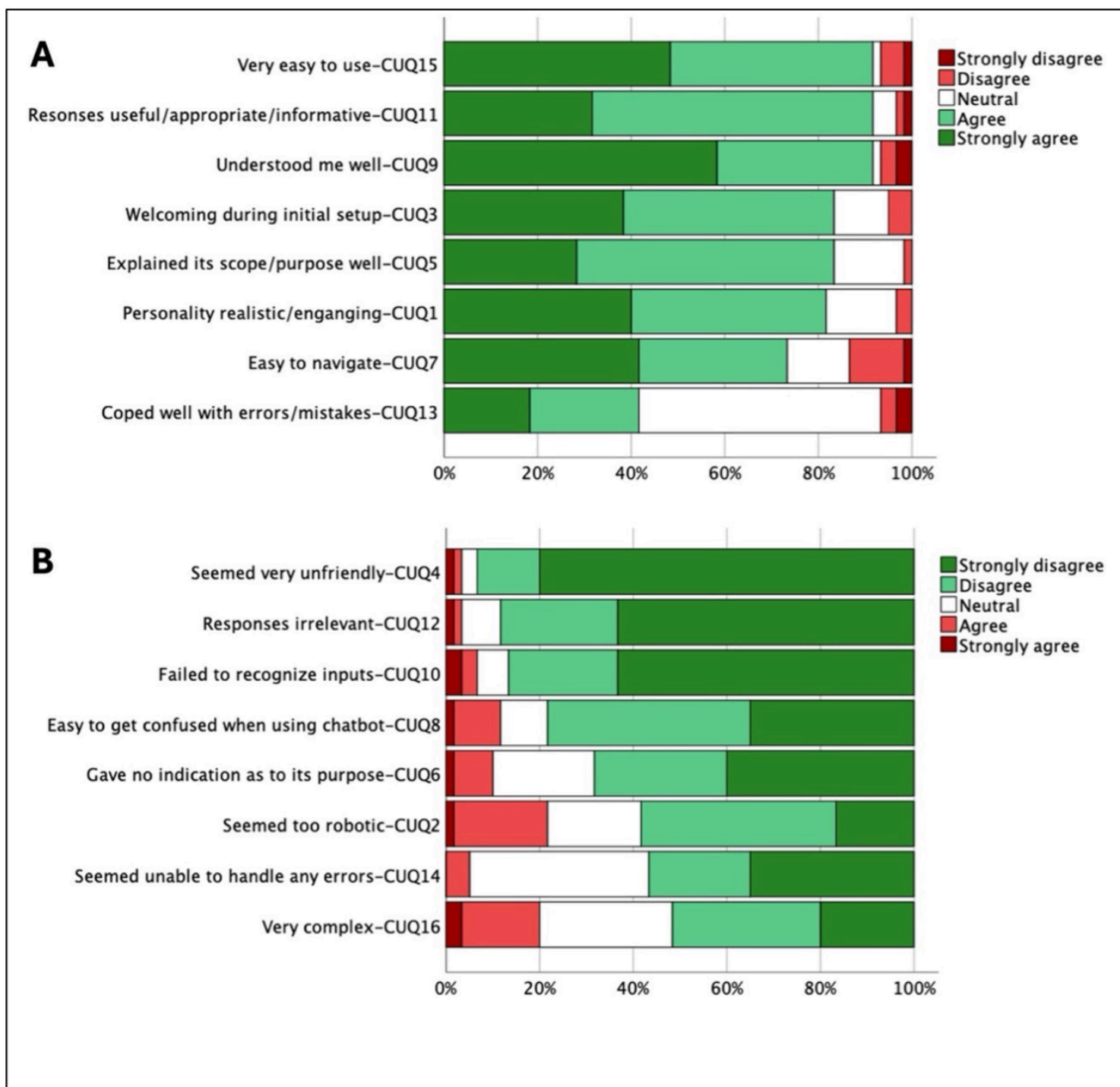


Figure 1

Item-level ratings of the Chatbot Usability Questionnaire (CUQ). Panel A displays the odd-numbered items assessing positive aspects of the chatbot, while Panel B shows the even-numbered items assessing negative aspects. Green bars indicate favorable responses: “agree” or “strongly agree” for positive items, and “disagree” or “strongly disagree” for negative items.

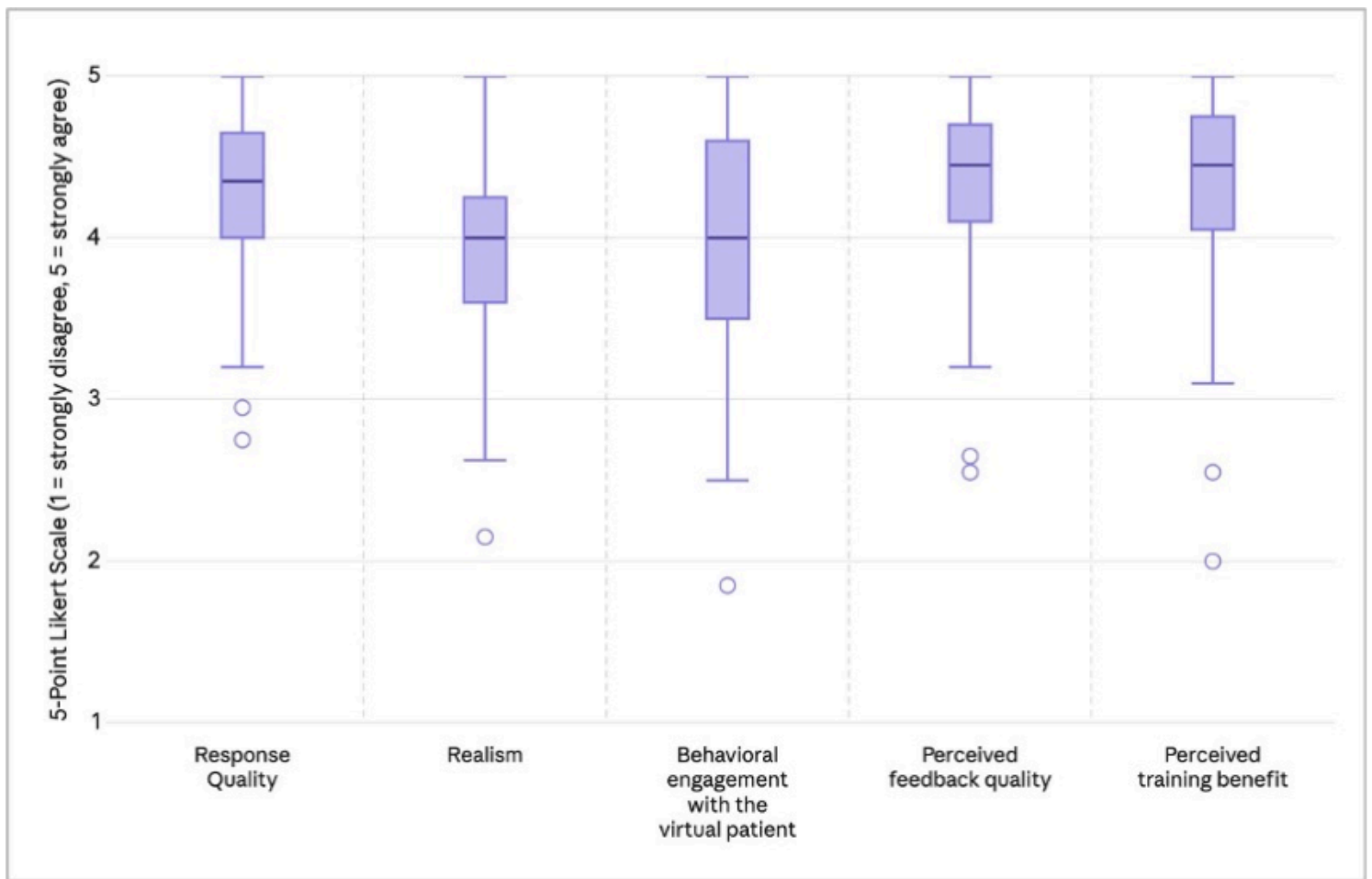


Figure 2

Median scores and interquartile ranges for the five composite scales: Response Quality, Realism, Behavioral Engagement with the Virtual Patient, Feedback Quality, and Perceived Training Benefit. Scores are based on a 5-point Likert scale (1 = strongly *disagree*, 5 = strongly *agree*), with higher values indicating more favorable evaluations.

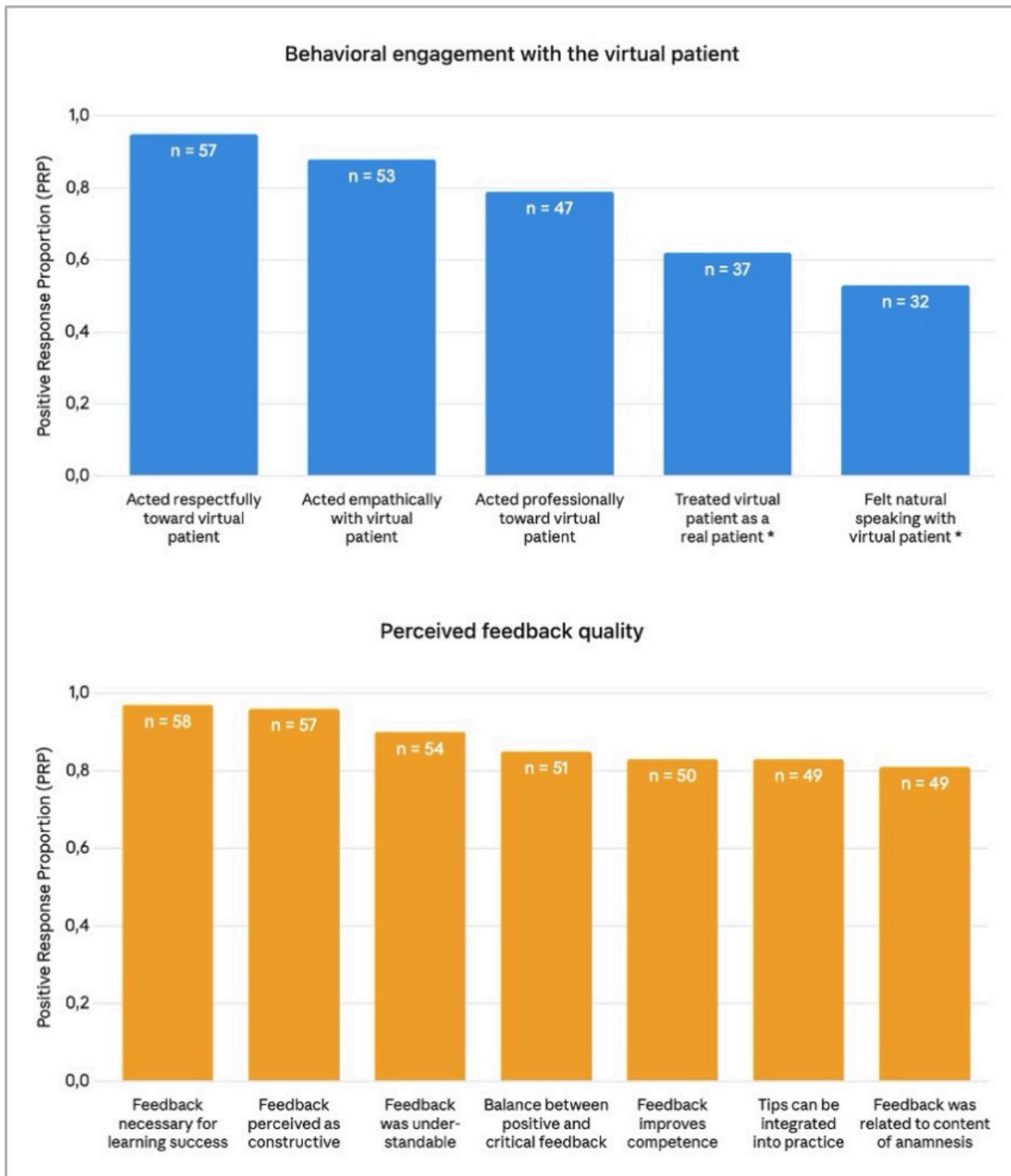


Figure 3

Positive Response Rates (PRP) for each item composite scales: *Behavioral Engagement with the Virtual Patient* and *Perceived Feedback Quality*.

Higher values indicate more favorable participant perceptions; Items marked with an asterisk (*) deviate from the original survey wording: they were initially negatively phrased and subsequently reverse-coded

to ensure consistent interpretation.

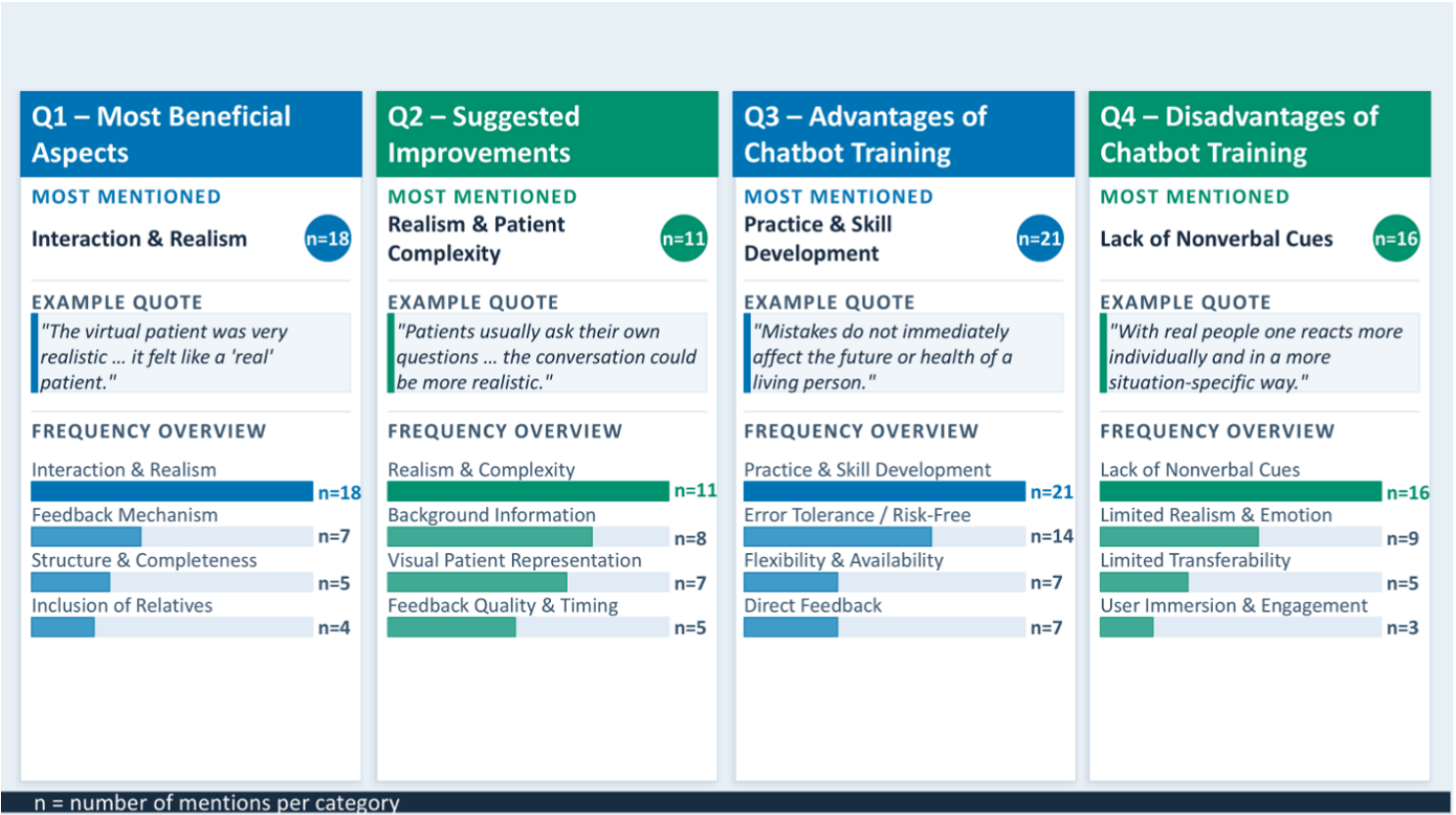


Figure 4

Summary of participants' responses to open-ended questions on the Virtual Patient Training.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SUPPLEMENT.docx](#)