Society of Medical Education & Research

2025, Volume 5, Page No: 39-49
ISSN: 3108-4850

Annals of Pharmacy Education, Safety, and Public Health Advocacy

Enhancing Virtual Medical History Taking: Effects of Customized Guidelines in Two Serious Games for Medical Education

Robertas Damaševičius[1*], Rytis Maskeliūnas[2], Tomas Blažauskas[1]

[1]Department of Software Engineering, Kaunas University of Technology, 44249 Kaunas, Lithuania.
[2]Department of Multimedia Engineering, Kaunas University of Technology, 44249 Kaunas, Lithuania.

***E-mail** ✉ robertas.damasevicius@vdu.lt

Abstract

Serious games serve as safe educational environments where medical students can practice clinical skills without compromising patient safety. When combined with virtual patients through chatbots, these games provide a platform to practice medical history taking. This study explored the effect of self-directed learning, supported by a customized guideline, on history-taking performance in two different chatbot-based serious games. A total of 159 fourth-year medical students were randomly assigned to one of two serious games, both set in an emergency department but simulating different clinical cases. Each student completed the game at two time points, with a guideline provided between sessions. The two chatbots differed in interaction design: one required students to generate their own history-taking questions (free-entry). At the same time, the other offered an extensive predefined list of questions (long menu). Outcome measures included the history-taking information entered into the chatbots, scored quantitatively as a history score, as well as students' self-reported learning outcomes. At the first measurement point, students in the free-entry chatbot condition achieved higher mean scores (85.2 ± 27.7) compared to those using the long-menu chatbot (78.8 ± 35.7). Following the introduction of the guideline, students using the long-menu chatbot showed significant improvement in their second session ($t(315) = -2.918$, $p = .004$, $d = -0.229$). In contrast, no significant progress was observed in the free-entry group. Self-assessment results revealed no significant differences between the two game formats. Findings indicate that history-taking performance improves through self-directed learning in a long-menu chatbot setting, which relies on cued recall, but not in a free-entry chatbot that depends on free recall. Given that serious games represent partially artificial environments for training clinical history taking, future research should investigate how effectively students can transfer these skills to real-world clinical practice.

Keywords: Chatbot, Medical education, Serious game, History taking, Self-directed learning

Introduction

Medical history taking and clinical reasoning are closely connected [1], as shown by the fact that effective history taking directly contributes to a large proportion of accurate differential diagnoses [2]. The core of history taking lies in collecting information [3], which forms the foundation for developing diagnostic hypotheses—for example, obtaining details about a patient's alcohol consumption [4]. A thorough history provided by the physician, together with adequate patient information, is essential for reaching an accurate preliminary diagnosis and minimizing diagnostic errors [5]. This is particularly crucial in emergency medicine, where precise and complete information gathering is indispensable [6]. Given its importance, medical history taking must receive special emphasis during medical training.

Literature review

One established method for training history taking is the use of simulated patients (SPs) [7]. While SPs provide a safe and effective learning environment, they require significant resources [8, 9]. As a more cost-efficient alternative, virtual patients (VPs) are widely used, offering students consistent and safe opportunities to

Access this article online https://smerpub.com/

Received: 19 January 2025; **Accepted:** 15 April 2025

How to cite this article: Damaševičius R, Maskeliūnas R, Tomas Blažauskas T. Enhancing Virtual Medical History Taking: Effects of Customized Guidelines in Two Serious Games for Medical Education. Ann Pharm Educ Saf Public Health Advocacy. 2025;5:39-49. https://doi.org/10.51847/kNshKQSf5t

practice history taking [10]. When embedded in serious games, VPs can be placed within a broader storyline, allowing learners to perform additional tasks such as examinations or treatments. Serious games are already applied in teaching various aspects of history taking [11]. A realistic approach to training history taking with VPs is through chatbots, which are software systems enabling written or spoken interactions that mimic human conversation [12]. To date, only a limited number of studies have incorporated chatbots into medical education, either as part of serious games [13] or as stand-alone tools [14–16]. Evidence suggests that chatbot-based training can improve OSCE performance as well as learners' perceived competence and confidence [17]. Within serious games, chatbots function as integral design elements and should be grounded in educational theory to foster intrinsic motivation and enable systematic evaluation [18].

Theoretical frameworks frequently applied to serious games include Deci and Ryan's self-determination theory (SDT) [19] and Csikszentmihalyi's flow theory [20], both of which are commonly referenced in this context [21]. Serious games enable repeated practice in a safe setting without the constraints of time or resources [22], and are inherently designed to support self-directed learning [23]. Self-directed learning involves planning and managing one's own learning process [24, 25], providing students with autonomy, which is strongly linked to increased motivation [26–28].

Another strength of serious games is their external validity. Their realistic design enables knowledge and skills to be transferred more readily into real-world clinical practice [29]. For instance, a study on a trauma triage serious game found that physicians' in-game decision-making closely mirrored their real-life clinical behavior [30]. Still, research outside the serious games context highlights that transferring communication skills from theory to practice remains a significant challenge [31, 32]. Serious games, through their authenticity and interactive design, may therefore serve as effective tools to bridge this gap and support the transfer of history-taking skills to clinical practice.

Research aim and hypotheses

Given the importance of medical history taking and the potential of serious games to provide a safe environment for self-directed learning, this study investigated whether providing specific self-directed learning material—a customized history-taking guideline—improves student performance between two gameplay sessions, and whether this effect depends on the type of chatbot used. To explore this, two distinct serious games were compared: **EMERGE**, which uses a long-menu chatbot, and **DIVINA**, which requires students to formulate and enter their own questions.

In EMERGE, the long-menu format presents students with a list of possible questions that contain the keywords they type, thereby offering cues that support the recognition of relevant questions. In contrast, DIVINA requires students to formulate and input their questions independently. These two formats align with the constructs of **cued recall** and **free recall**, respectively, drawn from cognitive psychology. Cued recall involves memory retrieval supported by prompts, whereas free recall relies on unaided memory retrieval [33]. Because cued recall typically produces better outcomes than free recall [33], it was hypothesized that students would achieve higher initial history-taking scores in EMERGE than in DIVINA.

H1: Students will obtain significantly higher history scores in EMERGE session 1 compared to DIVINA session 1 and session 2.

The introduction of a history-taking guideline as self-directed learning material was expected to influence performance differently depending on the chatbot type. The guideline was designed to reactivate prior knowledge and help students develop an internal schema for history taking, which could then be applied during the second session. Developing such a schema reflects an element of clinical reasoning [34], and since history taking and clinical reasoning are closely interlinked [1], the guideline was assumed to enhance history-taking performance.

DIVINA's free-entry format provides greater flexibility, allowing students to apply their internal schema by formulating unrestricted questions, thereby fostering more thorough exploration and potentially leading to improved history-taking scores. In contrast, EMERGE constrains exploration by limiting students to menu-based options. Previous research has shown that learners tend to ask more and explore further in open-ended chatbot systems than in constrained ones [35]. Combining this exploratory potential with the structured knowledge gained from the guideline, it was hypothesized that history-taking performance would improve in DIVINA but not in EMERGE.

H2: After studying the history-taking guideline, students' history scores will improve in DIVINA but not in EMERGE across the two sessions.

Materials and Methods

Participants

This study was embedded within the mandatory cardiopulmonary module at Göttingen Medical School during the summer term of 2024. Ethical approval was granted by the local Institutional Review Board (application number: 21/3/24). While all students attended the module, participation in the research component—meaning consent to data analysis—was voluntary. A total of 159 fourth-year students agreed to take part (79 in EMERGE, 80 in DIVINA). All participants had previously completed a history-taking course, including a summative exam, two semesters before this module.

Study design

The module required students to attend four serious game sessions, each 90 minutes long. Only the first two sessions were relevant for the present study. Students were randomly divided into two groups: one group participated on-site using the serious game **EMERGE**, while the other group joined remotely and played **DIVINA.** This arrangement was determined by the technical setup of the two games. Both sessions were overseen by medical experts, who provided support upon request but did not intervene in the gameplay itself.

After the first session, all students were given a guideline for taking a history to support self-directed learning. They were instructed to use this material only in preparation for the second session, not during it. To monitor compliance, the post-session questionnaire included control items asking whether students had relied on the guideline during gameplay and whether they had reviewed it beforehand.

The questionnaire also measured students' perceived learning progress in history taking. These outcomes were analyzed using the comparative self-assessment (CSA) gain method [36], which required students at the end of session two to retrospectively rate their skills before session one and then assess their skills after the second session. The survey was conducted online using evasys (evasys GmbH, version 10.0).

Details of the study design and the medical cases used in both serious games are presented in **Table 1**.

Table 1. Depiction of the study procedure

Session	Group 1	Group 2	Diseases for both groups
1	DIVINA	EMERGE	STEMI, NSTEMI, musculoskeletal chest pain, hypertensive crisis
	Individual study of the history taking guideline		
2	DIVINA	EMERGE	STEMI, NSTEMI, musculoskeletal chest pain, hypertensive crisis, congestive heart failure, aortic stenosis
	Online questionnaire		

Diseases that were presented only in session two are written in italics.

Serious games

The study employed two different serious games, both simulating an emergency department where students assumed the role of a virtual physician on duty. Detailed descriptions are available in Aster *et al.* [37] for DIVINA and in Middeke *et al.* [38] for EMERGE. Each game allows learners to perform a range of emergency care tasks, including history taking, diagnostic examinations, treatment initiation, and patient discharge.

While the two games share common content and actions, they differ in visual design and, more importantly, in the way history taking is carried out via their chat systems. In **DIVINA**, students interact with a chatbot that requires them to compose and enter their own questions. In contrast, **EMERGE** utilizes a long menu chat format. Here, students are provided with a comprehensive list of potential questions. They can also type fragments of a question, which then generates a drop-down menu containing all matching options, from which they select the most appropriate phrasing.

Examples of these chat interfaces are presented in Aster *et al.* [37] and Middeke *et al.* [39], as well as in **Figure 1**. From a cognitive psychology perspective, DIVINA represents **free recall**, as it provides no cues for recognition, whereas EMERGE corresponds to **cued recall** through its structured menu system [33].

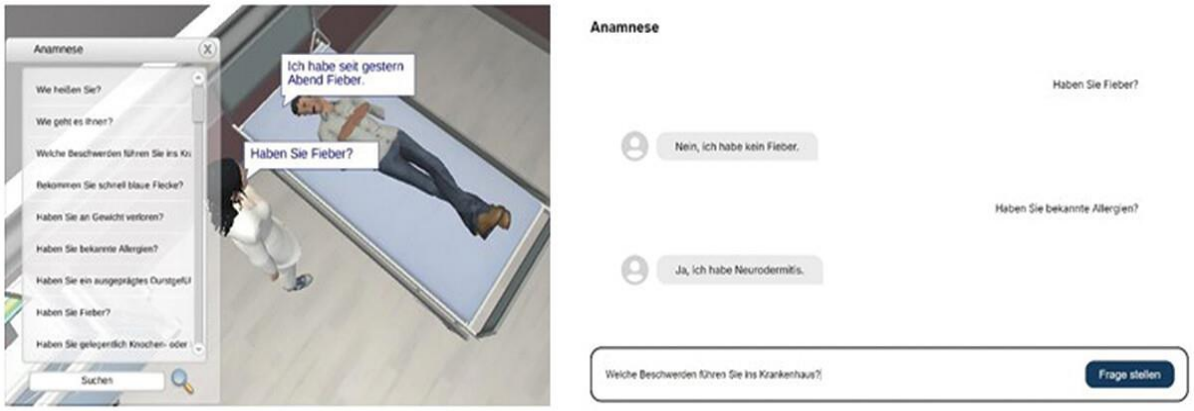

Figure 1. Screenshots of the chat interfaces in EMERGE and DIVINA. The left panel shows EMERGE (graphic by PatientZero Games GmbH), and the right panel displays DIVINA. All text appears in German, as the study was conducted entirely in that language

Development of the history-taking guideline

For this study, the authors created a history-taking guideline derived from the scoring checklist used for assessment. The guideline mirrored the checklist's structure, comprising five main sections: (1) current symptoms, including a focused pain anamnesis; (2) past medical history and pre-existing conditions, including cardiovascular risk factors; (3) lifestyle factors and additional risks; (4) medication history; and (5) social history.

Each section contained detailed sub-items that reflected the essential elements of an ideal history. For example, within pain anamnesis, aspects such as timing of symptoms and pain intensity were specified. To support learning, each section also included two model questions. For instance, under pre-existing conditions and history, examples included: *"Do you have any chronic illnesses?"* or *"Have you ever had an accident or a fall?"* Students were expected to use the guidelines as preparation material for self-directed learning, activating prior knowledge to be applied in the second session. Importantly, they were instructed to consult the guidelines only between sessions, not while playing the second session.

Data analysis

To evaluate the performance of history-taking, all qualitative entries submitted by students were converted into quantifiable data. Two independent raters, blinded to group assignment and session, assessed the data separately. Ratings were based on a history-taking checklist developed in an earlier study [35], which was further refined to address ambiguities, adapt scoring ranges, and align with emergency department contexts. Each of the 26 checklist items could receive up to 10 points, yielding a maximum possible score of 260. Students were not required to use the exact wording of the example questions; variations were accepted as long as the content was appropriate.

For self-assessment outcomes, CSA gain values were calculated as outlined by Raupach *et al.* [36]. First, incomplete data and unmatched response pairs were excluded. Mean retrospective ratings (skills before session 1) were then compared with mean post-session ratings (skills after session 2), applying the CSA formula. Data preparation involved multiple steps. Chats without any entries were excluded to ensure only valid datasets remained (i.e., at least one history-taking question was entered). Next, confounding cases were removed, including entries linked to diseases introduced only in session 2 (e.g., congestive heart failure, aortic stenosis) and data from students who had participated in only one session. The complete process of dataset derivation is illustrated in **Figure 2.**

All statistical analyses were performed using IBM SPSS Statistics (version 27) and Microsoft Excel. Normality assumptions were violated in most cases, except for the EMERGE session 2 and the proxy score for DIVINA session 1 used in ANOVA. However, given the sufficiently large sample size, both t-tests and ANOVAs were considered robust against such violations [40, 41]. Accordingly, an independent one-tailed t-test was used to test **H1**, while paired one-tailed t-tests and a mixed ANOVA were applied for **H2**.

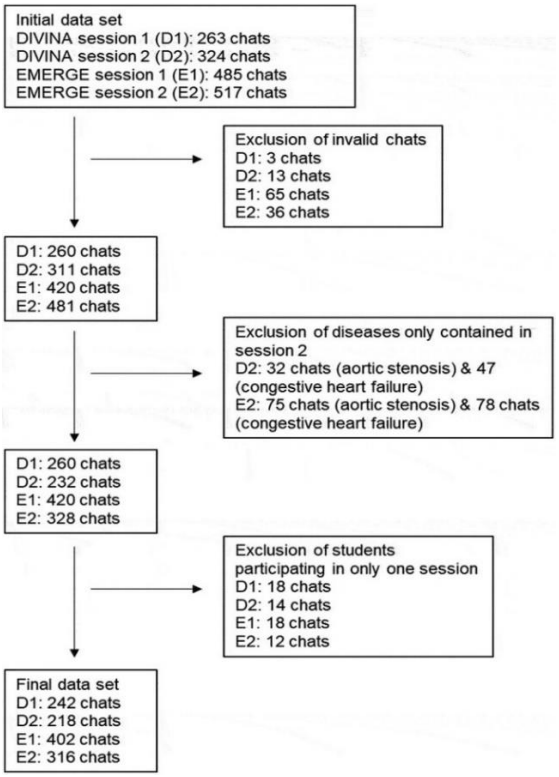

Figure 2. Process for deriving the final data set. Each session is abbreviated with a capital letter referring to the serious game and a number referring to the respective session. Resulting in D1 for DIVINA session 1, D2 for DIVINA session 2, E1 for EMERGE session 1, and E2 for EMERGE session 2

Results and Discussion

The final dataset included 242 and 218 chats from DIVINA sessions 1 (D1) and 2 (D2), respectively, and 402 and 316 chats from EMERGE sessions 1 (E1) and 2 (E2). On average, students submitted 3.51 valid chats in D1 and 3.16 in D2, compared to 5.66 chats in E1 and 4.45 in E2. Regarding guideline usage, the majority of participants reported not consulting the guideline during the second session (DIVINA: 89.9%, n = 69; EMERGE: 80.8%, n = 52), while only 60.9% of the DIVINA group and 47.2% of the EMERGE group indicated that they had used it for preparation.

History-taking performance was assessed using the mean score of two independent raters, who demonstrated excellent interrater reliability (intraclass correlation coefficients: D1 = .945, D2 = .960, E1 = .977, E2 = .965) [42]. Hypothesis 1, predicting higher history scores in E1 compared to D1 and D2, was not supported. Unpaired t-tests revealed no significant differences between E1 and D1 or between E1 and D2 ($P > .05$). Notably, the mean scores in E1 (78.8 ± 35.7) were slightly lower than those in D1 (85.2 ± 27.7) and D2 (88.3 ± 29.5).

For hypothesis 2, which anticipated an improvement in DIVINA (D2 > D1) but not in EMERGE (E2 = E1), paired t-tests were conducted. Although D2 showed a higher average score (88.3 ± 29.5) than D1 (86.5 ± 27.9), this increase was not statistically significant, $t(217) = -0.622$, $P = .267$, $d = -0.062$. Contrary to expectations, EMERGE participants achieved significantly higher scores in E2 (86.6 ± 35.0) compared to E1 (78.5 ± 34.6), $t(315) = -2.918$, $P = .004$, $d = -0.229$. This improvement appears attributable to an increased number of questions submitted in E2 (mean = 11) relative to E1 (mean = 10.1), whereas the number of questions in DIVINA remained stable across sessions (D1: 15.7, D2: 15.3).

A mixed ANOVA was conducted to examine potential interaction effects between session and game type, using a proxy variable representing the mean history score per student across all chats. The analysis confirmed sphericity, and Levene's test indicated homogeneity of error variances ($P > .05$), although Box's test revealed heterogeneity of covariance ($P < .001$). No significant interaction between session and group was observed, $F(1, 138) = 1.051$, $P = .307$, partial $\eta^2 = .008$, nor were there significant main effects for session, $F(1, 138) = 1.746$, $P = .189$, partial $\eta^2 = .012$, or group, $F(1, 138) = 0.172$, $P = .679$, partial $\eta^2 = .001$. The proxy variable was also used to generate **Figure 3**, illustrating the progression of history scores across sessions.

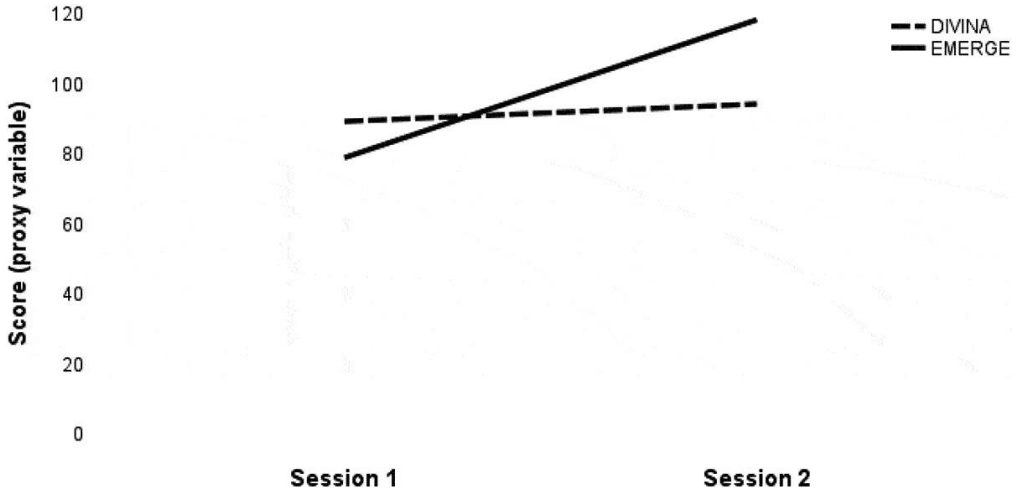

Figure 3. Course of the score improvements between the two sessions for each serious game. The proxy variable built for the analysis of the mixed ANOVA was also used for the creation of the figure

Comparative self-assessment (CSA) GAIN

In addition to objective performance measures, students' subjective learning gains were evaluated using a comparative self-assessment at the conclusion of session 2. Since all self-assessment data were collected anonymously, it was not possible to link subjective and objective data for combined analysis. Unlike the objective dataset, the subjective data were not pre-processed, but all valid data pairs from the retrospective and post-session assessments were included in the study.

Question-wise comparisons of CSA gain scores between the two serious games revealed no significant differences. As illustrated in **Figure 4**, participants in EMERGE exhibited, on average, slightly higher increases in self-assessed skills than those in DIVINA. However, improvements were modest for both games, with CSA gain scores remaining below 60%, indicating a moderate level of perceived learning success.

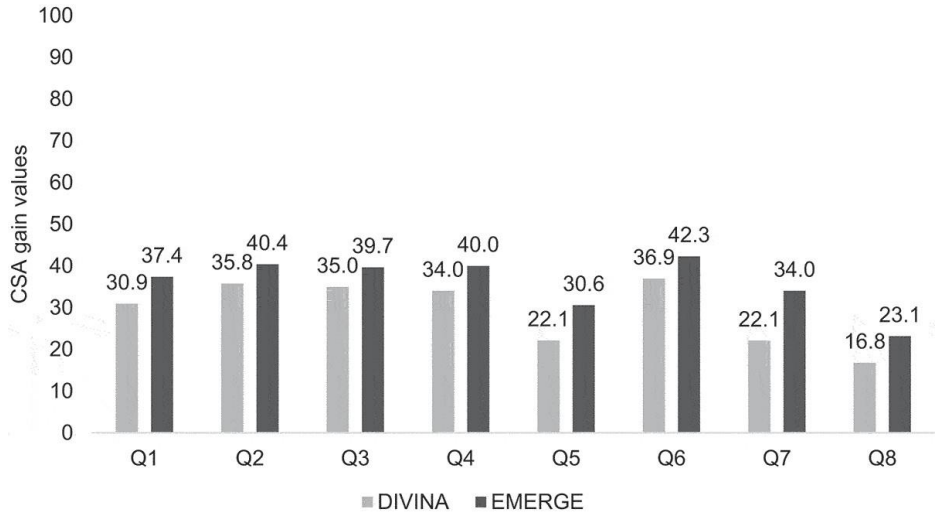

Figure 4. Mean CSA gain values for each question of both serious game groups

This study investigated whether providing a guideline to support self-directed learning enhances history-taking performance in chatbots used in serious games. Two types of chatbots were compared: a long-menu format relying on cued recall (EMERGE) and a free-entry format relying on free recall (DIVINA). The results indicate that the guideline led to improved history-taking in the long-menu chatbot but not in the free-entry chatbot.

Regarding the first hypothesis, the null hypothesis—that there would be no difference in performance between the first EMERGE session and the DIVINA sessions—could not be rejected. Descriptively, students in both DIVINA sessions achieved higher mean scores than in EMERGE session 1, suggesting that the assumption that a long-menu format would initially encourage students to ask more questions is not supported.

The second hypothesis examined whether supplementary self-directed learning material would improve performance in DIVINA but not in EMERGE. Although students in the second DIVINA session obtained slightly higher scores after using the guideline, the increase was not statistically significant. This suggests that the expected combination of a cognitive schema for history taking, facilitated by the guideline, and the freedom to explore in a free-entry chatbot did not yield the hypothesized benefit.

Unexpectedly, scores in EMERGE improved significantly between the two sessions. While no main effects of group or time, nor an interaction effect, were observed in the mixed ANOVA, the increase in EMERGE scores may reflect the advantage of cued recall: the long-menu format likely facilitated retrieval after knowledge reactivation through the guideline. This finding aligns with prior research indicating that cued recall often produces better performance than free recall [33, 43]. The improvement can be attributed to students asking more relevant questions, as scored via the predefined checklist, which is critical for clinical reasoning.

Although the guideline was intended to help students form an internal representation of history taking—a process analogous to clinical reasoning [34]—we did not collect qualitative data, such as focus groups, to directly confirm whether students developed such a schema. It remains unclear whether students were unable to apply reactivated knowledge in DIVINA because of limitations inherent to the free-entry format, such as the requirement to formulate full questions rather than relying on keywords.

Regarding guideline usage, more than half of DIVINA participants reported using it for preparation, compared to less than half in EMERGE. However, due to the anonymous nature of questionnaire responses, we were unable to link individual usage to performance outcomes, which prevented us from concluding whether students who used the guideline improved more than those who did not. Self-assessment data showed no significant differences between the games at the question level. Nonetheless, EMERGE participants demonstrated a slightly higher increase in perceived learning, consistent with objective performance results. This suggests that constrained chat systems with predefined questions may benefit more from interposed self-directed learning material.

Although CSA gains provide some insight into knowledge reactivation and recall, the absence of targeted questions regarding guideline content limits conclusions about individual differences in benefit. Overall, the findings highlight the potential for guided learning to enhance performance in cued-recall settings while indicating that free-entry chatbots may require additional strategies to facilitate effective application of self-directed learning.

Despite the observed differences between sessions, it is essential to note that students, on average, achieved only about one-third of the maximum possible points across both serious games and sessions. Although participants had previously completed a history-taking module, applying or reactivating this knowledge appears to be more challenging in a free-recall format, as in DIVINA, compared to the cued-recall format of EMERGE. Interestingly, a higher proportion of students using the free-entry chatbot reported consulting the guidelines for preparation than those using the long-menu system.

Serious games offer a safe learning environment and are widely employed in medical education [18], but their suitability for training history taking warrants careful consideration. First, serious games remain an artificial environment, lacking critical human factors such as non-verbal communication and empathy, which are essential in real-life history taking. This limitation may vary by chatbot type, as preliminary research suggests that AI systems like ChatGPT could potentially support training in empathic history taking [44]. Second, the transferability of skills acquired in serious games to real-life clinical settings is uncertain. Effective transfer requires not only the adaptation of pre-existing skills into the game environment but also the subsequent application of in-game learning to real patient interactions. Evidence on transfer performance is mixed; while some studies demonstrate the successful application of in-game learning [29], others report limited transfer [38, 45]. For instance, training with a stand-alone chatbot has been shown to improve outcomes in mock OSCEs [17]. Still, it remains unclear whether

similar benefits occur with chatbots embedded in serious games.

Several limitations of this study should be considered. First, the online study design could not entirely prevent students from consulting external resources, including the guidelines, during sessions. While most participants indicated they had not used the guideline during gameplay, responses may be influenced by social desirability bias. Nevertheless, if students had used the guideline during the session, it is likely that higher scores would have been achieved, supporting the assumption that guideline use during gameplay was minimal. Additionally, the DIVINA chatbot occasionally mismatched student questions with appropriate answers, requiring repeated queries. Such issues are common in natural language processing chatbots and highlight the need for more reliable virtual patient systems [46]. It was also not possible to determine whether a low number of questions reflected students' choices or external constraints such as time limits. Future studies could address this by analyzing only complete virtual patient encounters.

Second, the history-taking checklist was designed to balance comprehensiveness and rating efficiency, scoring only the first question within each rubric (e.g., "Do you smoke?"). Consequently, some relevant items, particularly regarding associated symptoms, were not captured. Future research could consider a more detailed checklist or complement quantitative scoring with qualitative analysis to better capture the richness of student history-taking behavior.

Third, the emergency department context may pose additional challenges for history-taking training. While accurate and comprehensive history taking is critical in urgent care, factors such as life-threatening conditions, time pressure, and the need to perform simultaneous investigations may impede performance. Although this study focused on the chatbot, differences in the games' visual designs could also have influenced outcomes. Testing chatbots within the whole game environment, rather than as stand-alone tools, introduces inherent variability that is difficult to control.

Implications

Serious games that incorporate chatbots are valuable tools in medical education, offering a safe environment for students to practice skills without risking patient safety [22]. The present study suggests that combining self-directed learning materials with a long-menu chatbot, which promotes cued recall, may be more effective for training history-taking skills than pairing such materials with a free-entry chatbot relying on free recall. Future research should explore the mechanisms underlying differences between various chatbot types and examine whether following prescribed expert questions or generating questions independently leads to more effective learning. It is conceivable that long-menu chatbots may be better suited for undergraduate training, providing structured guidance. In contrast, free-entry chatbots may be more suitable for postgraduate learners, offering greater autonomy and realism. Additionally, comparing students' performance in chatbots with objective exam results could help evaluate their educational effectiveness. Investigating whether learning transfers from serious games to real-life clinical situations or examination settings would further strengthen the evidence for their external validity.

Conclusion

This study examined the effect of supplemental self-directed learning material, in the form of a history-taking guideline, on students' performance using chatbots embedded in serious games. The results indicate that long-menu chatbots, which rely on cued recall, benefit more from interposed self-directed learning materials than free-entry chatbots. Although students initially performed better with the free-entry chatbot, providing a written guideline led to significant improvements only in the long-menu format. These findings highlight the importance of considering both chatbot type and the use of supporting materials when designing serious games for teaching medical history taking.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Fürstenberg S, Helm T, Prediger S, Kadmon M, Berberat PO, Oubaid V, et al. Assessing clinical reasoning in undergraduate medical students during history taking with an empirically derived scale for

clinical reasoning indicators. BMC Med Educ. 2020;20(1):368.
2. Peterson MC, Holbrook JH, Von Hales DE, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. West J Med. 1992;156(2):163–5.
3. Manalastas G, Noble LM, Viney R, Griffin AE. What does the structure of a medical consultation look like? A new method for visualising doctor-patient communication. Patient Educ Couns. 2021;104(6):1387–97.
4. Skotzko CE, Vrinceanu A, Krueger L, Freudenberger R. Alcohol use and congestive heart failure: incidence, importance, and approaches to improved history taking. Heart Fail Rev. 2009;14(1):51–5.
5. Schwitzguebel AJ, Jeckelmann C, Gavinio R, Levallois C, Benaïm C, Spechbach H. Differential diagnosis assessment in ambulatory care with an automated medical history-taking device: pilot randomized controlled trial. JMIR Med Inform. 2019;7(4):e14044.
6. Collins LC, Gablasova D, Pill J. 'Doing questioning' in the emergency department. Health Commun. 2023;38(12):2721–9.
7. Keifenheim KE, Teufel M, Ip J, Speiser N, Speiser S, Kopp V, et al. Teaching history taking to medical students: a systematic review. BMC Med Educ. 2015;15:159.
8. Cleland JA, Abe K, Rethans JJ. The use of simulated patients in medical education: AMEE guide no 42. Med Teach. 2009;31(6):477-86.
9. Kaplonyi J, Bowles KA, Nestel D, McKinley RK, O'Neill P, McLachlan J, 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. Maicher KR, Zimmerman L, Wilcox B, Danforth D, Price A, Danforth D, et al. Using virtual standardized patients to accurately assess information gathering skills in medical students. Med Teach. 2019;41(9):1053–9.
11. Alyami H, Alawami M, Lyndon M, Alghamdi M, Alzahrani A, Alzahrani A, et al. Impact of using a 3D visual metaphor serious game to teach history-taking content to medical students: longitudinal mixed methods pilot study. JMIR Serious Games. 2019;7(3):e13748.
12. Adamopoulou E, Moussiades L. An overview of chatbot technology. Artif Intel Appl Innovations. 2020;373–83.
13. Raupach T, de Temple I, Middeke A, Benaïm C, Spechbach H, Levallois C, et al. Effectiveness of a serious game addressing guideline adherence: cohort study with 1.5-year follow-up. BMC Med Educ. 2021;21(1):189.
14. Co M, John Yuen TH, Cheung HH. Using clinical history taking chatbot mobile app for clinical bedside teachings - a prospective case control study. Heliyon. 2022;8(6):e09751.
15. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Benaïm C, Spechbach H, Levallois C, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ. 2024;10:e59213.
16. Holderried F, Stegemann-Philipps C, Herschbach L, Benaïm C, Spechbach H, Levallois C, 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.
17. Raafat NH, Harbourne AD, Radia K, Benaïm C, Spechbach H, Levallois C, et al. Virtual patients improve history-taking competence and confidence in medical students. Med Teach. 2024;46(5):682–8.
18. Aster A, Laupichler MC, Zimmer S, Benaïm C, Spechbach H, Levallois C, et al. Game design elements of serious games in the education of medical and healthcare professions: a mixed-methods systematic review of underlying theories and teaching effectiveness. Adv Health Sci Educ Theory Pract. 2024;29(5):1825–48.
19. Deci EL, Ryan RM. The general causality orientations scale: self-determination in personality. J Res Personality. 1985;19(2):109–34.
20. Csikszentmihalyi M. Beyond boredom and anxiety. San Francisco (CA): Jossey-Bass; 1975.
21. Krath J, Schürmann L, von Korflesch HFO. Revealing the theoretical basis of gamification: a systematic review and analysis of theory in research on gamification, serious games and game-based learning. Comput Hum Behav. 2021;125:106950.
22. Sharifzadeh N, Kharrazi H, Nazari E, Benaïm C, Spechbach H, Levallois C, et al. Health education serious games targeting health care providers,

patients, and public health users: scoping review. JMIR Serious Games. 2020;8(1):e13459.
23. Tan AJQ, Lee CCS, Lin PY, Benaïm C, Spechbach H, Levallois C, et al. Designing and evaluating the effectiveness of a serious game for safe administration of blood transfusion: a randomized controlled trial. Nurse Educ Today. 2017;55:38–44.
24. Knowles MS. Self-directed learning: a guide for learners and teachers. New York (NY): Association Press; 1975.
25. Loyens SMM, Magda J, Rikers RMJP. Self-directed learning in problem-based learning and its relationships with self-regulated learning. Educ Psychol Rev. 2008;20(4):411–27.
26. Ryan RM, Deci EL. Self-regulation and the problem of human autonomy: does psychology need choice, self-determination, and will? J Pers. 2006;74(6):1557–85.
27. Dankbaar ME, Roozeboom MB, Oprins EA, Benaïm C, Spechbach H, Levallois C, et al. Preparing residents effectively in emergency skills training with a serious game. Simul Healthc. 2017;12(1):9–16.
28. Ryan RM, Deci EL. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp Educ Psychol. 2000;25(1):54–67.
29. Buijs-Spanjers KR, Harmsen A, Hegge HH, Spook JE, de Rooij SE, Jaarsma DADC, et al. The influence of a serious game's narrative on students' attitudes and learning experiences regarding delirium: an interview study. BMC Med Educ. 2020;20:289.
30. Mohan D, Angus DC, Ricketts D, Farris C, Fischhoff B, Rosengart MR, et al. Assessing the validity of using serious game technology to analyze physician decision making. PLOS ONE. 2014;9(8):e105445.
31. Berglund L, von Knorring J, McGrath A. When theory meets reality — a mismatch in communication: a qualitative study of clinical transition from communication skills training to the surgical ward. BMC Med Educ. 2023;23:728.
32. van den Eertwegh V, van Dulmen S, van Dalen J, Scherpbier AJJA, van der Vleuten CPM. Learning in context: identifying gaps in research on the transfer of medical communication skills to the clinical workplace. Patient Educ Couns. 2013;90(2):184–92.
33. Higham PA, Guzel MA. Cued recall. In: Seel NM, ed. Encyclopedia of the Sciences of Learning. Boston (MA): Springer; 2012. p. 868-71.
34. Young M, Thomas A, Lubarsky S, Ballard T, Gordon D, Gruppen LD, et al. Drawing boundaries: the difficulty in defining clinical reasoning. Acad Med. 2018;93(7):990-5.
35. Aster A, Lotz A, Raupach T. Theoretical background of the game design element "chatbot" in serious games for medical education. Adv Simul. 2025;10(1):Article 10.
36. Raupach T, Münscher C, Beissbarth T, Burckhardt G, Pukrop T. Towards outcome-based programme evaluation: using student comparative self-assessments to determine teaching effectiveness. Med Teach. 2011;33(8):e446-53.
37. Aster A, Hütt C, Morton C, Flitton M, Laupichler MC, Raupach T. Development and evaluation of an emergency department serious game for undergraduate medical students. BMC Med Educ. 2024;24:1061.
38. Middeke A, Anders S, Raupach T, Schülper M. Transfer of clinical reasoning trained with a serious game to comparable clinical problems: a prospective randomized study. Simul Healthc. 2020;15(2):75–81.
39. Middeke A, Anders S, Schuelper M, Raupach T, Schuelper N. Training of clinical reasoning with a serious game versus small-group problem-based learning: a prospective study. PLOS ONE. 2018;13(9):e0203851.
40. Glass GV, Peckham PD, Sanders JR. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Rev Educ Res. 1972;42(3):237-88.
41. Rasch D, Guiard V. The robustness of parametric statistical methods. Psychol Sci. 2004;46:175–208.
42. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994;6(4):284-90.
43. Tulving E, Pearlstone Z. Availability versus accessibility of information in memory for words. J Verbal Learn Verbal Behav. 1966;5(4):381-91.
44. Aster A, Ragaller SV, Raupach T, Marx A. ChatGPT as a virtual patient: written empathic expressions during medical history taking. Med Sci Educ. 2025:1-0. doi:10.1007/s40670-025-02342-7
45. Aster A, Scheithauer S, Middeke AC, Zegota S, Clauberg S, Artelt T, et al. Use of a serious game to teach infectious disease management in medical

school: effectiveness and transfer to a clinical examination. Front Med. 2022;9:863764.

46. Maicher K, Danforth D, Price A, Zimmerman L, Wilcox B, Liston B, et al. Developing a conversational virtual standardized patient to enable students to practice history-taking skills. Simul Healthc. 2017;12(2):124-31.