Computer Methods and Programs in Biomedicine 244 (2024) 107964

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Development and evaluation of a software system for medical students to teach and practice anamnestic interviews with virtual patient avatars

Antonia Lippitsch [a], Jonas Steglich [a], Christiane Ludwig [a], Juliane Kellner [a], Linn Hempel [a], Dietrich Stoevesandt [a], Oliver Thews [b,*]

[a] Dorothea Erxleben Learning Centre Halle (DELH), University of Halle-Wittenberg, Germany
[b] Julius Bernstein Institute of Physiology, University of Halle-Wittenberg, Magdeburger Str. 6, Halle (Saale) 06112, Germany

ARTICLE INFO

Keywords:
Anamnestic interview
Medical students
Chatbot
Avatar
Practical exercise
Artificial intelligence

ABSTRACT

Background and Objectives: Taking a medical history is a core competence of the diagnostic process. At the beginning of their study medical students need to learn and practice the necessary techniques, initially focusing on good structuring and completeness. For this purpose, an interactive software system (ViPATalk) was developed in which the student can train to pose questions to virtual patient avatars in free conversation. At the end, the student receives feedback on the completeness of the questioning and an explanation of the essential items. The use of this software was compared to the traditional format of student role play in a randomized trial.

Methods: The central component of ViPATalk is a chatbot based on the AI language AIML, which generates an appropriate answer based on keywords in the student's question. To enable a realistic use, the student can enter the question via microphone (speech-to-text) and the answer generated by the chatbot is presented as a short video sequence, where the avatar is generated from a real image. Here, the transition between the sequences is seamless, resulting in a continuous movement of the avatar during the conversation.

Results: The learning success by practicing with ViPATalk was tested in an anamnestic interview with actors as simulated patients. The completeness of the conversation was evaluated with regard to numerous aspects and also certain behaviors during the conversation. These results were compared with those after practicing using peer role play.

Conclusions: It was found that practicing with ViPATalk was mostly equivalent to the students' role play. In the subsequent survey of the students, the wish was expressed that the ViPATalk software should also be used as an online tool for self-study and that there should be more cases for practicing.

1. Introduction

Taking a detailed medical history is an essential part of establishing contact with a patient. This involves recording the reason for the consultation and the current complaints on the one hand, but also all relevant information from the patient's history and environment on the other. Medical students must therefore learn and practice the necessary techniques [1]. Especially in the early section of the study, the aim is to learn a well-structured and as complete as possible medical history as a basic framework for the medical interview.

In the first section of medical school, the content of the anamnesis interview is first taught theoretically and then practiced in role plays [2]. For this purpose, the use of simulated patients (SP) is certainly a very good option. Simulated patients are usually medical non-professionals who are trained to perform a certain patient role in order to provide students a training opportunity. However, this form of teaching in small groups is very resource-intensive (actors, lecturers). For this reason, paired role plays between two students are often used. However, since the role of the interviewee is insufficiently defined, the practice effect is limited. If the anamnestic interview shall be practiced for different clinical pictures the student who plays the patient has to have good knowledge about the diseases and the symptoms. Another limitation of the role play is that it is time consuming especially for that student who is playing the "patient". Here a tool would be helpful in which the students can train the anamnestic dialog alone (without fellow student) for a large number of cases. However, since role play is a well established technique of teaching communication skills [3,4] new techniques have to be compared with it.

[*] Corresponding author.
E-mail address: oliver.thews@medizin.uni-halle.de (O. Thews).

https://doi.org/10.1016/j.cmpb.2023.107964
Received 16 August 2023; Received in revised form 27 November 2023; Accepted 30 November 2023
Available online 1 December 2023

Against this background, this paper deals with the development and evaluation of a training software for the anamnesis interview using a Virtual Standardized Patient (VSP) [5,6]. The aim of the study was firstly the development of a software teaching tool to practice dialogs for taking medical history of patients for medical students in the early phase of their study. First of all, it was tested whether simple rule-based chatbots are able to simulate the dialog and how the rules can be developed iteratively. The second question was to analyze whether a drawn image of a patient is sufficient to give the student the impression of a realistic dialog or whether a photorealistic representation is necessary. Thirdly, it was tested whether a computer-generated voice is sufficient to simulate a real patient or whether natural (recorded) speech is necessary. Finally, it was tested whether a speech-to-text tool can be used to give the student the possibility to interact with the software via spoken language. This developmental process (which took a period of more than 2 years) led to a system in which the conversation on the patient side is performed by a chatbot, which responds to the student's questions with an answer that matches the simulated clinical picture. From the graphical point of view optimal results were obtained with photorealistic visualization of the avatar and natural (not computer-generated) speech. The development also clearly indicated the need of a feedback function in which the students receive an assessment of the performance during the dialog. However, it became obvious that just providing a list of missing items which have not (but should be) included in the interview was not enough. For this reason, short additional videos for each clinical case were produced in which an experienced examiner explains the relevant, most important aspects which have to be covered in the anamnestic interview so that the students understand better the feedback list of missing items.

After the completion of the developmental process the software system was tested in terms of learning effect as defined by the completeness of the necessary questions and behavior during the conversation with a real (human) person in a randomized study. Since the projects aims at students at the beginning of their medical study these aspects are most important. However, for more advanced students many other aspects are also relevant (e.g., target-oriented dialog, empathy) but these issues cannot be addressed with the present software system. For assessing the software in routine practicing, the software tool was compared with the conventional role play between two students. This comparison was aimed to demonstrate that practicing with a virtual patient is at least non-inferior as compared to regular role play because in this case the software could be used time- and location independent on many different cases.

2. Materials and methods

2.1. Software system

The software system (called ViPATalk) consists of an input unit through which the student can ask his question, the actual chatbot that generates the appropriate answer, and an output unit in which an animated avatar presents the answer (Fig. 1). The system was implemented in the Delphi programming language. Screenshots of the software are shown in Suppl. Fig. S1.

2.1.1. Chatbot

For the present work, the approach of a simple rule-based chatbot was chosen. This chatbot structure is suitable for simple question-answer dialogs as they are typical in a medical history conversation (doctor asks a question to which the patient answers). Such a simple rule-based dialog system uses word patterns within the question sentence to guess the content of the query. If the recognition pattern matches the question asked, an appropriate response is issued. Rule-based chatbots have been known for quite a long time [7,8], but the structure is usually sufficient for medical history questioning, since a complex semantic analysis of the question is not necessary. For the definition of the rules, the syntax of the AIML (Artificial Intelligence Markup Language) [9] was used which is sufficient for chatbots with a limited language scope. AIML performs pattern recognition in the input text. The text is analyzed for certain keywords or word patterns and the answer matching this term is output. A semantic or syntactic analysis is not performed. A separate AIML rule must be defined for each keyword. Due to the multitude of linguistic possibilities to describe a context (e.g., use of different synonyms), extensive rule sets are created for each patient case to represent all facts of an anamnesis interview. The rules are processed within the program by means of a rule interpreter. For the present program for the interpreter library PASCALice was used [10], which takes over the execution of the rule processing, whereby the order of the rules within the database is insignificant. By the interpreter structure the system can be extended easily, around missing or incomplete question patterns arbitrarily to be added. Also, the database generated in this way can be easily transferred to other user environments (e.g., as an online teaching tool), as long as for this environment an AIML interpreter library exists (see Discussion).

If none of the keywords is recognized in the student's question, an output is generated asking the user to ask the question again with different wording. At the same time, the unrecognized question is saved by the ViPATalk software so that new keywords can be added to the list of rules if necessary. The extension of the database (definition of new AIML rules) is done manually in the present program version. All

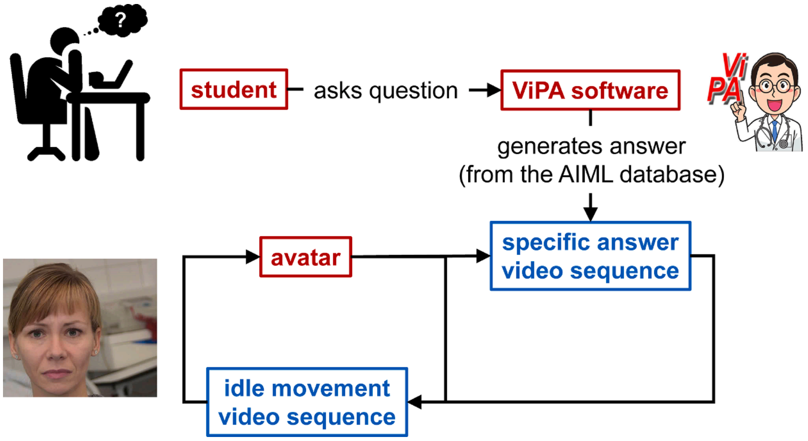

Fig. 1. Schematic program flow of the ViPATalk program. The student asks a question, the chatbot generates a suitable answer and the corresponding video sequence of the avatar is shown. Afterwards, the idle movement video sequence is shown again.

unrecognized questions by student users are logged by the system. At regular intervals, the lists of unrecognized word patterns are analyzed by the instructors and missing question patterns are added to the AIML database.

The AIML rules are divided into 9 topics to enable structuring of the medical history: (1) opening of the conversation, (2) current complaints, (3) own medical history, (4) medications and vaccinations, (5) risk factors and allergies, (6) vegetative medical history, (7) family medical history, (8) psychosocial medical history, (9) farewell.

2.1.2. Speech-to-text

For the anamnesis interview, the student can enter the question via the keyboard, although typing errors can lead to the chatbot not recognizing key words. To enable a more realistic interaction, the question can also be spoken in ViPATalk. For the conversion of the speech into an input text for the chatbot, the speech-to-text service of the company Google Cloud (Google Ireland Ltd., Dublin, Ireland) is used. For this, the audio file is transmitted over the Internet and ViPATalk receives the text back.

2.1.3. Virtual standardized patient (VSP) videos

The output of the answer generated by the chatbot, is initially shown on the screen as text. However, to generate a more realistic environment, a virtual representation of the patient (avatar) is also used. The avatar is initially based on a photograph of an actor (Suppl. Fig. S1B+C). This photo was then animated using CrazyTalk 8 software (Reallusion Inc. San Jose, CA, USA). In addition to a video sequence without text, where the avatar only performs an idle motion, individual short video scenes were generated for each of the chatbot's possible responses. The transitions of the video sequences were designed in such a way that no interruption is visible when the video changes. This creates the impression of a continuous movement of the patient and thus a real conversation situation. During the conversation, the avatar performs its idle movement, which is only interrupted for the answers generated by the chatbot (Fig. 1).

2.1.4. Feedback function

After completing an anamnesis interview, the student can retrieve feedback on his or her dialog. Here, the student receives a list of items that should have been asked in an anamnesis of the respective clinical case or that could have made an important contribution. For this purpose, when the AIML chatbot rules were generated, it was determined whether the answer was essential, important or secondary for the respective clinical picture. Based on this classification, the student receives an assessment of the completeness of his or her medical history, along with a reminder that the missing items should not be overlooked in future medical history discussions. To illustrate the importance of the essential and important items, a video was created for each virtual standardized patient case for the student to view after their exercise. In these videos, experienced clinical examiners present the cases again and explain in detail which aspects must be addressed most importantly in the case history for this clinical case.

2.2. Virtual standardized patient cases

Initially, the ViPATalk system was used to implement three clinical pictures: (1) case AS (female, 33 years, appendicitis, previous ectopic pregnancy), (2) case LD (female, 73 years, angina pectoris, suspected myocardial infarction) and (3) case MS (male, 68 years, esophageal variceal bleeding, diabetes mellitus type II, liver cirrhosis). For each case, the keywords for the chatbot and the resulting answers were defined according to the clinical picture and converted into AIML rules. For the AS case, this resulted in 1654 rules, for LD 2112 rules, and for MS 2247 rules. Photos of actors corresponding to the patient case were selected for the design of the videos. A single short video sequence was generated for each possible patient response, resulting in between 116 and 178 videos per case.

2.3. Evaluation trial

2.3.1. Study design

In order to demonstrate that a software tool for practicing medical history has *per se* a beneficial effect on learning, firstly the outcome of the teaching with ViPATalk should be compared to a group without any additional teaching. However, such an experimental design would not be ethically justifiable. Therefore, the teaching effect of ViPATalk was compared to the regularly method of using peer role play. In addition to several test runs during the development of the ViPATalk system, a controlled, randomized comparative study was subsequently conducted to test the application possibilities of the software. The focus of this comparison was on the question of whether the use of ViPATalk (intervention group; IG) in comparison to the usually used paired role play (control group; CG) would show a lower learning success with regard to the structuring and completeness of an anamnesis interview. The entire teaching session on anamnesis covered a period of 4 days with patient history training being a 1 h session per day. On the first day, all students received a theoretical introduction in the form of a lecture (Fig. 2). The second day was for hands-on practice, with students randomly assigned as matched pairs to the role play or ViPATalk group. On the third day, each study participant conducted a real-life medical history interview with a simulated patient. This interview was videotaped and subsequently used for quality assessment. On the fourth day, the tasks of the two test groups (roll play, ViPATalk) were swapped (Fig. 2) to ensure that no student would suffer a learning disadvantage in case of a difference between the two groups. The cross-over design on the fourth day was only due to the requirements by the ethic committee to give each student the possibility to learn with both techniques (in the case that one of the methods would be advantageous over the other). However, the results of the fourth day were not part of the further statistical analysis of the video recording on day 3. After completion of the fourth day, all study participants were able to submit their own experiences in an evaluation form.

2.3.2. Study population

The comparative study was conducted in 2020 and 2021 on a total of 168 3rd and 4th year medical students. The two groups (control group with peer role play and intervention group with ViPATalk) were balanced and thus included 84 students each. The study population consisted of 107 female and 61 male students. For the video recording on the third day, 84 were female and also 84 were male simulated patient cases.

2.3.3. Statistical analysis

The quality of the chatbot was quantified by the accuracy of the response generation. Here, it was judged successful if the question posed by the user led to the generation of an answer from the AIML rule base. An evaluation of the content, i.e., the evaluation of whether the individual answer generated by the system corresponded to the information desired by the student, could not be made, since in this dialog system it could not be inquired with which background a question was asked. Also, the further inquiry whether a generated answer corresponded to the expectations of the user was not practicable within longer anamnestic dialogues, which extended partly over more than 90 questions.

For data analysis, the video recordings were first analyzed according to a standardized list of topics, assessing whether or not a particular item was addressed in the case history (for a complete list of analyzed items, see Suppl. Tab S1). In addition, aspects of student interviewing and behavior (such as structured approach, friendliness, eye contact, or sitting position) were also assessed on a three-point scale. The results were first transferred to the spreadsheet Excel and analyzed descriptively in the form of frequency distributions. A possible group difference was calculated using the two-sided chi-square test or *t*-test as

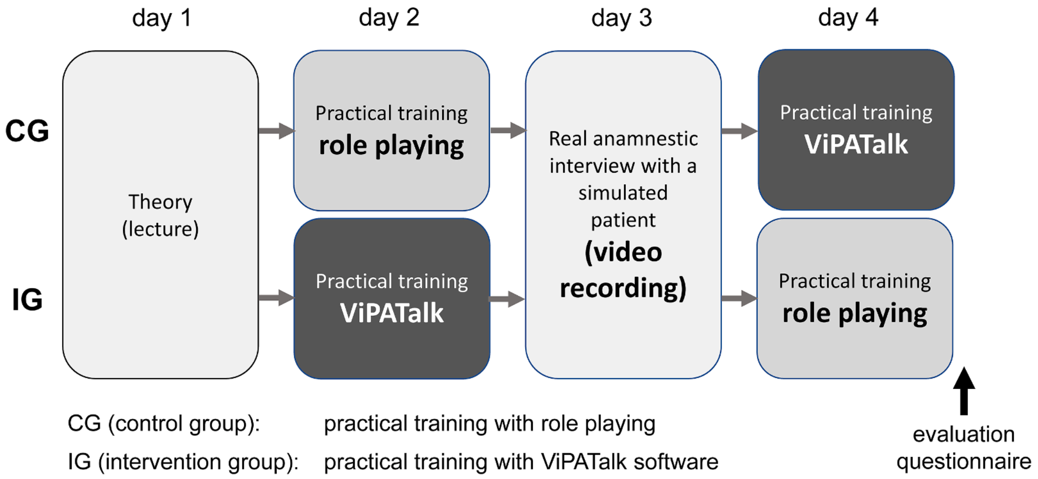
day 1
day 2
day 3
day 4
CG
IG
Theory (lecture)
Practical training role playing
Practical training ViPATalk
Real anamnestic interview with a simulated patient (video recording)
Practical training ViPATalk
Practical training role playing
CG (control group): practical training with role playing
IG (intervention group): practical training with ViPATalk software
evaluation questionnaire

Fig. 2. Sequence of the randomized trial to evaluate the ViPATalk software compared to students' peer role play (1-hour period).

appropriate, with α=5 % chosen as the significance level. An α-adjustment for multiple testing was not made. Therefore, the p-values are only descriptive.

In order to compare the impact of the virtual patient simulation with regular role play on the learning success not only the significant difference between the groups was analyzed but also whether both teaching forms lead to equivalent outcome or at least that teaching with ViPATalk is non-inferior compared to role play. For equivalence or non-inferiority testing different methods where used. For categorical data with binary outcome (yes/no) the methods described by Wellek were used [11,12]. For the equivalence or non-inferiority comparison of numerical data the TOST method was used [13,14]. The calculations were performed with R using the libraries "EQUIVNONINF" and "TOSTER". For these tests also α=5 % was used as significance level but without α-adjustment for multiple testing.

3. Results

3.1. Developmental steps of the software system

The step-by-step development process of the ViPATalk system involved chatbot functionality, animation of virtual patients (avatars), and speech recognition for entering questions.

3.1.1. Chatbot

The AI language AIML used for the chatbot uses simple pattern recognition to identify specific keywords or constellations of terms. Since it is not a linguistic analysis, any change in spelling is not recognized. In the present form the rules are defined for German language. Also, the answers generated by the system are in German. However, due to the flexible structure of the rules (in addition with an easy-to-handle rule editor) the dialogs can be easily transferred to other languages.

Thus, the tense used in the question (present, imperfect, etc.) matters, but so does the use of singular or plural form of a noun or the use of synonyms. Since students ask the question in a very wide variety of linguistic forms, there were a lot of unrecognized words at the beginning of the software development. Since ViPATalk stores all errors in the word recognition, the successful answering could be continuously improved by extending the AIML rules for the previously unrecognized questions. Therefore, each time the system was tested with students, the number of AIML rules increased greatly.

In the first development stage, the AIML rule base consisted of 887 rules for the AS case, 885 rules for the LD case, and 1025 rules for the MS case. The quality of the chatbot at this stage was tested on 474 dialogs (case AS: 177; LD: 163; MS: 134). On average, 30.7 questions (range: 10–93) (case AS: 28.4; LD: 28.3; MS: 36.6) were addressed to the chatbot by the users during each test run. Recognition of the question (with generation of the associated answer) was considered a successful response by the chatbot. In this initial development phase, the rate of unrecognized questions averaged 29.8 ± 13.0 % across all 474 test runs (case AS: 32.2 ± 15.1 %; LD: 29.0 ± 12.0 %; MS: 28.0 ± 10.6 %). Subsequently, the AIML database was significantly expanded based on the evaluation of the non-recognized question patterns. In the second stage of development, the rule base included 1654 rules for the AS case, 2112 rules for the LD case, and 2247 rules for the MS case. This rule base was tested on 294 test runs of the 3 simulated patients (case AS: 104; LD: 105; MS: 85). On average, 25.0 questions (range: 10–60) (case AS: 25.3; LD: 25.7; MS: 23.7) were directed to the chatbot during each test run. The recognition rate was significantly increased by expanding the rule database. On average, only 13.5 ± 9.0 % (case AS: 14.6 ± 8.8 %; LD: 14.1 ± 9.6 %; MS: 11.5 ± 8.3 %) of the questions were not recognized by the chatbot and thus could not lead to an answer. The analysis of the non-recognized questions showed three major reasons: (1) students use new unknown expressions in their questions, (2) typos within the text of the question and (3) incorrect speech-to-text recognition (leading to meaningless questions). The first aspect is the most important one because it shows that the rule database is still incomplete. For this reason, the database is continuously maintained and extended. Here an adaptive learning algorithm would be helpful so that the software adapts the rule base autonomously. Such algorithms will be part of the further development. For the correcting typos, a database with typical transposed letters and typing errors is included which replaces failures directly. The incorrect speech-to-text recognition is an aspect which cannot be influenced by the software system because it uses the speech-to-text functionality of Google. We recommend the students to speak loud and clear and use this feature in a quiet environment. In the future, the fraction of recognized question must be further increased. Even though the number of rules was markedly increased, the response time does not play a limiting role. On average, the response time of the chatbot was well below 1 ms for all three cases. Despite the significant improvement in the recognition rate of the chatbot, the expansion of the question pattern database is a key aspect for the further development of the system.

For the development of new cases for ViPATalk and for the maintenance or extension of the AIML rule base an easy-to-handle rule editor was designed. With this editor synonyms and variations of the recognition patter can be rapidly entered. It is also possible to use recognition patterns from previous cases and just to modify the respective answer. If the clinical case vignette is completed the definition of the dialog elements mostly does not take longer than 1 or 2 days.

3.1.2. *Avatars*

For the realistic representation of the simulated patients as interlocutors for the anamnesis interview, the avatars should move as naturally as possible, convert a written text (answer of the chatbot) into speech and simulate corresponding lip movements. In addition, the appearance of the avatar should correspond to the clinical case (e.g., age, posture, skin color). In a small pilot analysis, it was tested which kind of graphical presentation led to a more realistic appearance. In the first approach of designing patient avatars, commercial solutions sketched drawings were used. However, it became apparent that drawn images of individuals could not convey a natural appearance. Also, most commercial products almost exclusively depict young, healthy individuals. The second alternative were photographic images. Most of the respondents in this pilot phase stated that photographic images correspond better to the clinical appearance of patients. It was also asked which kind of output of the generated answer would be preferable. Here three alternatives were tested: (1) written text on the screen, (2) spoken text by a computer-generated speech or (3) natural spoken language by an actor. A clear majority voted for a speech output. However, the computer-generated speech (usually young, clearly spoken voices) was rated as not suitable to reflect the personality (age, education etc.) or the mood of acutely ill persons. For this reason, the animation of a photographic image was used for the present system (CrazyTalk 8 software). Here, lip and head movements are added to a given audio file to a photo and saved as a video. The use of a photograph allows the appearance to be adapted to the patient's case. Suppl. Fig. S1B+C show screenshots of two cases. All possible answers were spoken by actors and saved as audio files, so that an adapted speech image is achieved.

3.1.3. *Speech recognition*

In order to achieve a dialog between student and patient avatar that is as close to reality as possible, it should be possible to enter the question as spoken text. However, initial tests with the speech recognition function of personal computers showed that the recognition rate was very poor, so that a real dialog was not possible. To achieve sufficient recognition quality, it was necessary to switch to a more powerful platform, which was achieved with Google's online speech-to-text service. After the change, the recognition rate was sufficiently good, provided that clear and distinct speech was used and ViPATalk was used in a quiet environment.

3.2. *Randomized evaluation study*

Video analysis of the real patient interviews (SP) showed that after the theoretical introduction (lecture) and the first practice phase (with ViPATalk or the role play), a comprehensive medical history had not yet been achieved (Fig. 3, Suppl. Fig. S2). Taking all 42 analyzed items together the results showed that the ViPATalk and the role play groups ask on average almost the same number of questions (Fig. 4, "all categories"). The ViPATalk group asked 61.2 ± 10.5 % of all items whereas in the role play group it was 63.3 ± 10.1 %. The equivalence testing showed that both groups were statistically equivalent and the ViPATalk group was non-inferior (Suppl. Tab. S2). Analyzing the different aspects of the medical history separately showed that current complaints and previous illnesses of one's own or in family members were addressed by all students, but details of current complaints, such as duration, localization, or intensity, were asked by only 2/3 of the students. Childhood illnesses were inquired about by only half of the students. Some aspects, such as asking about current travel, were almost never asked. These results demonstrate that a one-day practice period is not sufficient for learning a complete, structured history. Comparing the exercise using student role play (control group CG) and the electronic ViPATalk platform (intervention group IG), the results differed only slightly. With regard to most items, there were no significant differences between the two forms of instruction (Suppl. Fig. S2). Equivalence testing showed that several items of the dialog were statistically equivalent or that the ViPATalk group was significantly non-inferior when compared with the role play group (Suppl. Tab S3). Only a small number of items showed deviations (Fig. 3). For example, the ViPATalk group (IG) asked slightly more frequently about the onset of current complaints, but somewhat less frequently about childhood illnesses. However, Fig. 3 also shows that there were no systematic differences between the two groups and may have arisen by chance. The differences found in the video analysis of the dialogs with human actors seems not to be the result of insufficient practicing in the ViPATalk group. For instance, questions about childhood diseases were asked during the training with ViPATalk quite regularly and these questions were mostly correctly recognized and answered by the chatbot. So it cannot explain the result that in the video analysis this item was significantly less often asked than in the role play group. On the other hand, it is remarkable that questions of the personal interpretations of the symptoms are more often asked in the ViPATalk group even though they practiced with a virtual "person" which will not have an interpretation of the symptoms. For that, it seems to most likely that the observed differences result by chance. It should kept in mind that in total 45 items were tested but without α-adjustment for multiple testing so that some comparisons may be "significant" by chance and the p-values are only descriptive.

If the items evaluated in the video analysis are combined according to topic groups (Fig. 4), minor differences seem to be indicated. The ViPATalk group (IG) seems to have asked about the current symptoms in

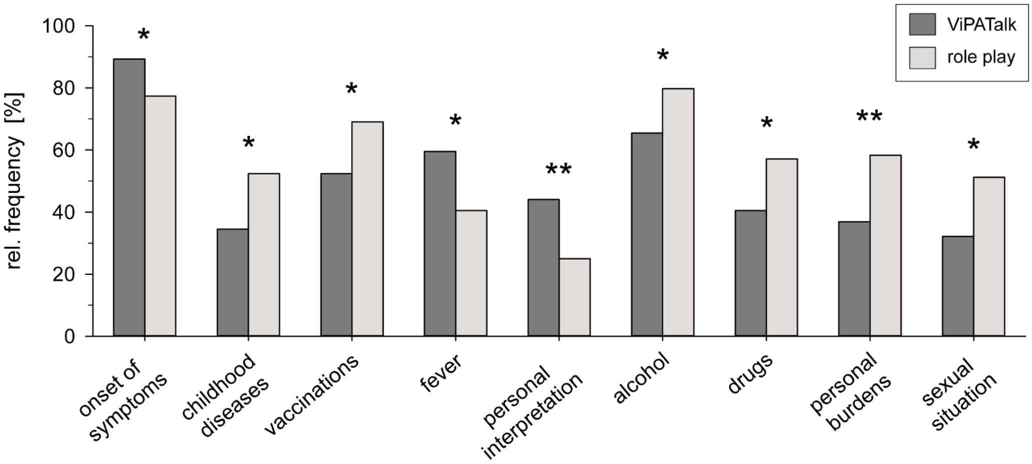

Fig. 3. Frequency distribution of whether the respective items were addressed in the real case history interviews with simulated patient (actors). $n = 168$; (*) $p < 0.05$, (**) $p < 0.01$ ViPATalk vs. role play.

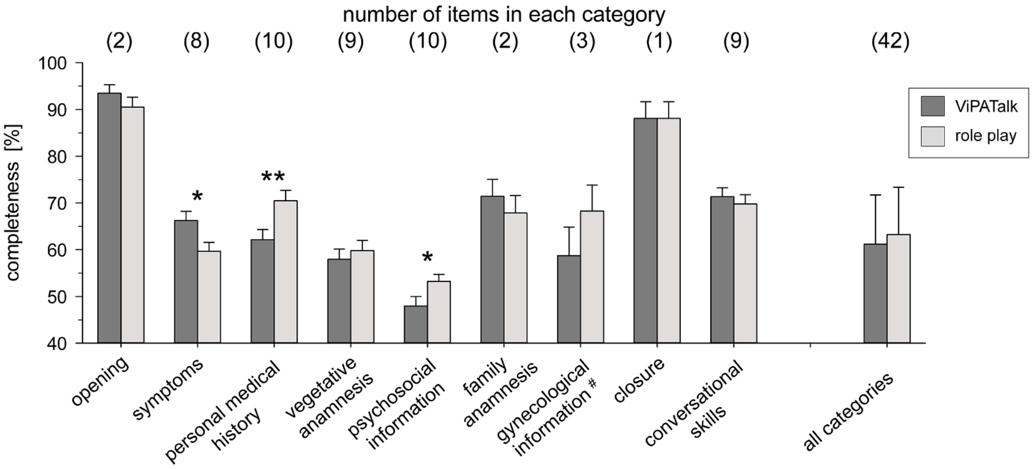

Fig. 4. Analysis of the completeness of the medical history regarding different categories in the video recording of the real conversation with a simulated patient (actors). The number of individual items in each topic group is given in parentheses. "all categories" includes all items without gynecological information and conversational skills. $n = 168$ ($^{\#}$ $n = 84$; cases with female actors); (*) $p<0.05$, (**) $p<0.01$ ViPATalk vs. role play.

more detail, whereas the role play group (CG) dealt somewhat more intensively with the patient's own previous illnesses and psychosocial situation. However, the observed differences are only slight. Suppl. Tab S2 shows the results of the statistical analysis whether both training methods are equivalent. Concerning the subject areas which were not statistically significant different (symptoms, personal medical history, psychosocial information) ViPATalk led mostly to equivalent results or showed non-inferior differences. At least these results indicated that ViPATalk and peer role play are equivalent forms of exercise. However, the partly incomplete case histories also prove that practice on a larger number of training cases is necessary for students to acquire a consolidated scheme of a structured case history.

In addition to completeness, students' behavior during the interview was also assessed (Suppl. Fig. S3). This included items such as whether the anamnesis interview was well structured, whether the patient was allowed to finish, or whether language understandable to laypersons was used. Again, there were no significant differences between the two test groups. Surprisingly, only the sitting position was significantly different in the two groups, with the ViPATalk group (IG) sitting more often facing the patient and at an appropriate distance.

3.3. Evaluation questionnaire

After experiencing both forms of teaching (role play and ViPATalk), the students were asked to complete a questionnaire to assess the use and significance of the software system. Overall, 41 % of the students rated the interaction with the virtual standardized patient as good, and 87 % of the respondents saw it as a useful addition to their previous teaching (Fig. 5A). The students rated the graphic design as mostly good (average rating 5.5 ± 1.1 out of 7 points). Whether the student role play or ViPATalk is the better form of teaching was judged differently (Fig. 5B). The students do not prefer one of the methods. The main points of criticism were seen on the one hand as the still unsatisfactory speech-to-text function for the voice input of the questions, and on the other hand the feedback function should still be improved in order to explain case-specifically why certain questions of the anamnesis are of particular importance. In general, the value of such a virtual training format was recognized and it was desired that such a format should also be offered outside of face-to-face teaching. Thus, almost 94 % of the students would like to be able to use the ViPATalk program from home (online) as well (Fig. 5C). There was also a desire for the system to contain significantly more different cases, with the majority of students favoring a number between 5 and 15 cases.

4. Discussion

This paper describes a computer system for self-study of medical history interviews for medical students in an early stage of their studies.

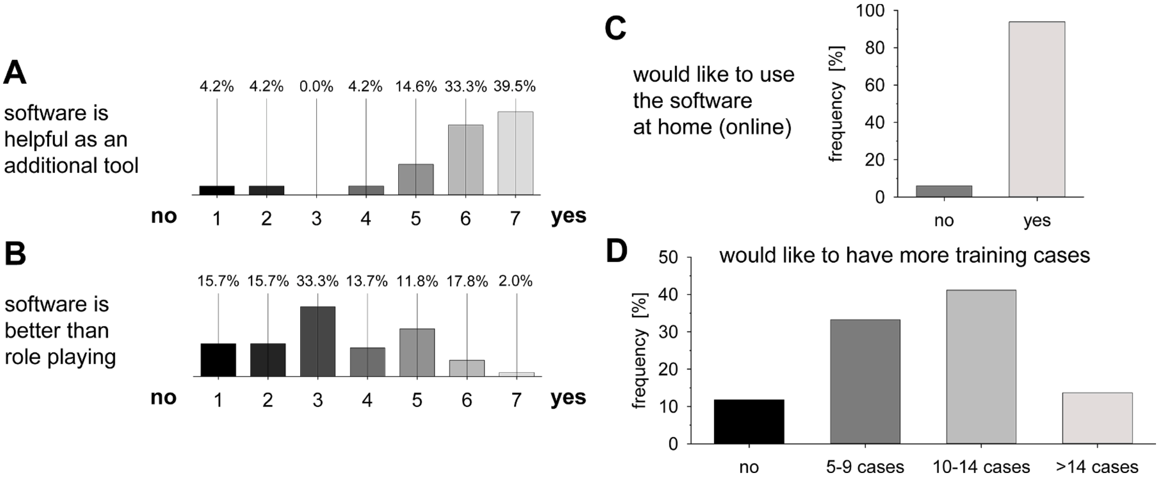

Fig. 5. Evaluation of the questionnaire after completion of the randomized study.

In developing the system, an attempt was made to create as realistic an environment as possible. On the one hand, this included voice input so that students could have a "free conversation" with the avatar whenever possible [6]. On the other hand, the representation of the patients should be as natural as possible and match the character of the case described (age, appearance, etc.). For the recognition of the question content, a pattern recognition approach of key terms was chosen. A content semantic analysis of the student question is not performed. Stored sentences are used as the answer. This approach is well suited for conversations with a limited vocabulary. In particular, during the anamnesis interview the physician asks clearly defined questions, which are answered relatively concisely by the patients. At least for cases in which the patients suffer from simple somatic disorders, this approach is sufficient for students in the early stages of their studies. In the case of complex contexts, such as psychiatric or psychosomatic disorders, in which linguistic correlations have to be detected, the simple pattern recognition system will not be sufficient. For this reason, such cases were excluded for the present system. Exclusive pattern recognition has two disadvantages. First, each student question must be formulated as a complete sentence. Also, follow-up questions in the form of chain questions (using linguistically incomplete sentences) cannot be understood by the system. Since this problem was already apparent during the initial testing, an introductory video was made created describing the appropriate way to handle the avatar.

Previous attempts also used AIML for rule-based chatbots in medical teaching [15]. However, in these attempts limitations of AIML with respect of the flexibility of pattern recognition were found. A general problem of simple rule-based pattern recognition is that different word forms (e.g., singular and plural forms or different tense forms) must each be programmed separately. For this reason, Stiff et al. changed to another AI-language ChatScript [16] which has the advantage to use sets of similar words concerning the meaning [17]. Since these sets can be imported from other existing databases the development of the rule base is faster and more flexible. It is also possible to extract these sets of keywords from other sources or combine it with neural networks to hybrid systems [16].

Since different word forms are a general problem of simple AIML rule-based pattern recognition, the present system was also expanded to sets of synonyms similar to those in ChatScript. To overcome the problem of different linguistic word forms which will lead to an immense increase of rules, the system is presently modified with simple natural language processing (NLP), which tokenizes and lemmatizes the entered question. This is done for German language with the Hanover Tagger [18]. This tagger converts the words of the question into their basic form, so that significantly fewer AIML rules are necessary. From the experiences with the ViPATalk software together with the described extensions pattern recognition seems to be a sufficient tool to generate adequate answers. For this reason, natural language understanding (NLU) [19] seems not to be necessary for a simple dialog chatbot for first-year medical students, in which simple chatbots are sufficient for more basic symptom-oriented diseases. This technique cannot be used in more complex dialogs for instance for psychosomatic or psychiatric cases [20].

Some other approaches to language analysis in medical history interviews have already been proposed in the literature. Furlan et al. used a Siamese long short-term memory (SLSTM) network [21,22] trained on a large number of anamnestic questions and linked it to the Systematized Nomenclature of Medicine (SNOMED CT) ontology. In their work, direct linkage to a medical entity was necessary because in their approach the history taking was only one part of the full diagnostic process. However, in testing, only moderate results were shown regarding history taking as part of the diagnostic process. Other research groups used longer text phrases for pattern recognition in the questions [23]. Campillos-Llanos et al. [24] used a semantic analysis of the question text, where entities, such as anatomical location, symptoms or timing, were also derived from the individual sentence parts. This information was then used to generate the answer. This approach is certainly promising but requires extensive and language-specific analysis.

The second important feature is a realistic graphical representation of the avatars and also a speech output that is as natural as possible. In preliminary investigations of the current project, it was investigated whether a text-to-speech feature would be suitable to output the answers generated by the chatbot. However, it was found that most computer-generated voices belong to young people and that the intonation and voice pitch do not match patients. For this reason, all answers were recorded by actors for the present system, whereby stresses and vocalizations were also adapted to the clinical picture. Regarding the external appearance of the avatars, several research groups have used graphically designed figures [23–25]. These drawn figures can be adapted to the particular clinical picture, however, no realistic appearance is created. The figures still appear artificial. For this reason, for the current project the way was chosen to animate real photographs, which corresponded in their appearance to the respective clinical picture, by means of software (with head, lip and eye movements). In this representation, the artificial animation is still recognizable, but due to the photographic template, a more natural impression is created than with drawn avatars.

A few approaches to a virtual learning environment for the diagnostic process have been described in the literature. In these systems, conversational guidance is integrated as part of the diagnostic process. The Shadow Health System (www.shadowhealth.com) is commercially available and is primarily aimed at nursing education and training. However, it is also used in pharmacy education [25–28]. In this regard, comparative studies have shown that practicing in a virtual environment brings confidence to subsequent real anamnesis interviews [26,27] and thus can increase educational success [25]. However, additional practice with standardized (simulated) patients has also been shown to be superior to training in the virtual environment alone [28]. In this respect, virtual case history training can only be a supplement but cannot substitute structured teaching.

Feedback after teaching is a very important aspect for all teaching methods (person-by-person or virtual teaching). In the ViPATalk system which primarily addresses the completeness of the necessary questions the students received as a first attempt a list of important aspects which were missing in their anamnestic interview. The analysis of the evaluation questionnaires clearly indicated that this feedback was not enough to understand why some aspects are relevant in the specific case. Therefore, feedback videos were created for each patient case where experienced clinical examiners present the case again and explain in detail which aspects are most essential. This important role of feedback in a teaching software, which can be equivalent to the direct feedback of clinical experts, has been demonstrated by others [29].

The comparative study between the training phase using ViPATalk and the peer role play did not reveal any clear differences (Figs. 3, 4). For this reason, it can be concluded that the software system is equivalent to the previously used training concept. The use of standardized patients will probably lead to better results [2], but often cannot be carried out due to the large number of students and is also not suitable as a sole training method, where each student realizes a large number of anamnesis interviews. Increasingly, online teaching is also desired and has also led to good results in comparative studies [30,31]. ViPATalk can be used very well in such an online environment. Evaluation of the system clearly indicated a desire by students to use the system with more cases as an online format from home. In the further development of the online implementation the pattern recognition of the questions will be improved by introducing a linguistic preprocessing of the text.

When comparing the use of ViPATalk with role play another aspect should be kept in mind. In the role play the student not only takes the role of the physician but in pairwise practicing another student also plays the patient. From playing the patient the student can also learn aspects of the symptoms or the impact of diseases on daily life. This aspect of learning from the patient's point of view is not possible when practicing with a simulation software.

In conclusion, the ViPATalk system seems to be a suitable platform for students to learn and practice the basic features of an anamnesis interview in the early part of their studies. The use of realistic avatars makes it easier for students to put themselves in the situation of a real conversation. Speech input also reinforces this impression. As a practice platform, the virtual system seems to be comparable to the learning success of a peer role play. Only for some specific items differences were found between both teaching methods. It has to be pointed out that in the present study the learning outcome was defined only by completeness of the necessary questions and the behavior during the conversation with a real (human) person. Other (also relevant) aspects of communication skills or acceptance of the software [32] were not considered. Even though the outcome of both teaching methods was not different, practicing with the software tool is more flexible (time- and location independent), it can be conducted alone (without fellow students) and on many different cases. The students judged the ViPATalk as a suitable additional tool for practicing anamnestic interviews. They also would like to use the system at home as an online tool with a larger number of patient cases. Because a primary focus of the system is the completeness of the history, its use is aimed preferentially at the early stages of training. Future studies can then investigate whether intensive training with ViPATalk can improve the outcome of real anamnesis interviews in the clinical setting.

CRediT authorship contribution statement

Antonia Lippitsch: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – review & editing. **Jonas Steglich:** Data curation, Formal analysis, Writing – review & editing. **Christiane Ludwig:** Investigation, Methodology, Writing – review & editing. **Juliane Kellner:** Investigation, Methodology, Writing – review & editing. **Linn Hempel:** Methodology, Writing – review & editing. **Dietrich Stoevesandt:** Methodology, Resources, Supervision, Validation, Writing – review & editing. **Oliver Thews:** Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The manuscript has not been submitted for publication to any other scientific journal. All authors on this paper are aware of and agree to the content of the manuscript and agree to be listed as authors. All authors declare no competing conflicts of interests.

Acknowledgment

The authors thank Ms. M. Rossov for creating the video sequences.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2023.107964.

References

[1] T. Seitz, B. Raschauer, A.S. Längle, H. Löffler-Stastka, Competency in medical history taking-the training physicians' view, Wien. Klin. Wochenschr. 131 (2019) 17–22.
[2] K.E. Keifenheim, M. Teufel, J. Ip, N. Speiser, E.J. Leehr, S. Zipfel, A. Herrmann-Werner, Teaching history taking to medical students: a systematic review, BMC Med. Educ. 15 (2015) 159.
[3] A. Gelis, S. Cervello, R. Rey, G. Llorca, P. Lambert, N. Franck, A. Dupeyron, M. Delpont, B. Rolland, Peer role-play for training communication skills in medical students: a systematic review, Simul. Healthc. 15 (2020) 106–111.
[4] D. Nestel, T. Tierney, Role-play for medical students learning about communication: guidelines for maximising benefits, BMC Med. Educ. 7 (2007) 3.
[5] A. Isaza-Restrepo, M.T. Gomez, G. Cifuentes, A. Arguello, The virtual patient as a learning tool: a mixed quantitative qualitative study, BMC Med. Educ. 18 (2018) 297.
[6] K.R. Maicher, A. Stiff, M. Scholl, M. White, E. Fosler-Lussier, W. Schuler, P. Serai, V. Sunder, H. Forrestal, L. Mendella, M. Adib, C. Bratton, K. Lee, D.R. Danforth, Artificial intelligence in virtual standardized patients: combining natural language understanding and rule based dialogue management to improve conversational fidelity, Med. Teach. (2022) 1–7.
[7] E. Adamopoulou, L. Moussiades, An overview of chatbot technology, Artif. Intell. Appl. Innov. 584 (2020) 373.
[8] G. Caldarini, S. Jaf, K. McGarry, A literature survey of recent advances in chatbots, Information 13 (2022).
[9] M.B. Marietto, R.V. Aguiar, G. de Oliveira Barbosa, W.W. Botelho, E. Pimentel, R. dos Santos França, V.L. da Silva, Artificial intelligence MArkup language: a brief tutorial, Int. J. Comput. Sci. Eng. Surv. 4 (2013) 1–20.
[10] K. Sullivan, PASCALice v1.5, GitHub repository, 2016.https://github.com/reshetnyakvkt/PascAlice (accessed 01.12.2023).
[11] S. Wellek, Testing Statistical Hypotheses of Equivalence and Noninferiority, 2nd ed., Chapman and Hall/CRC, New York, 2010.
[12] S. Wellek, Statistical methods for the analysis of two-arm non-inferiority trials with binary outcomes, Biom. J. 47 (2005) 48–61, discussion 99-107.
[13] D. Lakens, Equivalence tests: a practical primer for t tests, correlations, and meta-analyses, Soc. Psychol. Pers. Sci. 8 (2017) 355–362.
[14] J. Walker, Non-inferiority statistics and equivalence studies, BJA Educ. 19 (2019) 267–271.
[15] D.R. Danforth, M. Procter, R. Chen, M. Johnson, R. Heller, Development of virtual patient simulations for medical education, J. Virtual Worlds Res. 2 (2009) 4–11.
[16] A. Stiff, M. White, E. Fosler-Lussier, L. Jin, E. Jaffe, D.R. Danforth, A randomized prospective study of a hybrid rule- and data-driven virtual patient, Nat. Lang. Eng. 1 (2022) 1–42.
[17] E. Adamopoulou, L. Moussiades, Chatbots: history, technology, and applications, Mach. Lern. Appl. 2 (2020), 100006.
[18] C. Wartena, The hanover tagger (version 1.1.0) - lemmatization, morphological analysis and POS tagging in python, in: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), German Society for Computational Linguistics & Language Technology, 2019, pp. S1–24.
[19] I. F.T. Dolianiti, P. Antoniou, S. Konstantinidis, S. Anastasiades, P. Bamidis, C. B. Frasson, P. Vlamos, Chatbots in healthcare curricula: the case of a conversational virtual patient, P. Brain Function Assessment in Learning. BFAL 2020. Lecture Notes in Computer Science, Springer, Cham, 2020, pp. 137–147.
[20] A. Foster, N. Chaudhary, T. Kim, J.L. Waller, J. Wong, M. Borish, A. Cordar, B. Lok, P.F. Buckley, Using virtual patients to teach empathy: a randomized controlled study to enhance medical students' empathic communication, Simul. Healthc. 11 (2016) 181–189.
[21] R. Furlan, M. Gatti, R. Mene, D. Shiffer, C. Marchiori, A. Giaj Levra, V. Saturnino, E. Brunetta, F. Dipaola, A natural language processing-based virtual patient simulator and intelligent tutoring system for the clinical diagnostic process: simulator development and case study, JMIR Med. Inform. 9 (2021) e24073.
[22] R. Furlan, M. Gatti, R. Mene, D. Shiffer, C. Marchiori, A. Giaj Levra, V. Saturnino, E. Brunetta, F. Dipaola, Learning analytics applied to clinical diagnostic reasoning using a natural language processing-based virtual patient simulator: case study, JMIR Med. Educ. 8 (2022) e24372.
[23] A. Stevens, J. Hernandez, K. Johnsen, R. Dickerson, A. Raij, C. Harrison, M. DiPietro, B. Allen, R. Ferdig, S. Foti, J. Jackson, M. Shin, J. Cendan, R. Watson, M. Duerson, B. Lok, M. Cohen, P. Wagner, D.S. Lind, The use of virtual patients to teach medical students history taking and communication skills, Am. J. Surg. 191 (2006) 806–811.
[24] L.T. Campillos-Llanos, C. Bilinski, E. Zweigenbaum, P. Rosset, Designing a virtual patient dialogue system based on terminology-rich resources: challenges and evaluation, Nat. Lang. Eng. 26 (2020) 183–220.
[25] C.A. Taglieri, S.J. Crosby, K. Zimmerman, T. Schneider, D.K. Patel, Evaluation of the use of a virtual patient on student competence and confidence in performing simulated clinic visits, Am. J. Pharm. Educ. 81 (2017) 87.
[26] N.L. Borja-Hart, C.A. Spivey, C.M. George, Use of virtual patient software to assess student confidence and ability in communication skills and virtual patient impression: a mixed-methods approach, Curr. Pharm. Teach. Learn. 11 (2019) 710–718.
[27] B.D. Fidler, Use of a virtual patient simulation program to enhance the physical assessment and medical history taking skills of doctor of pharmacy students, Curr. Pharm. Teach. Learn. 12 (2020) 810–816.
[28] T. Zerilli, B.D. Fidler, C. Tendhar, Assessing the impact of standardized patient encounters on students' medical history-taking skills in practice, Am. J. Pharm. Educ. (2022) 8989.
[29] K.R. Maicher, L. Zimmerman, B. Wilcox, B. Liston, H. Cronau, A. Macerollo, L. Jin, E. Jaffe, M. White, E. Fosler-Lussier, W. Schuler, D.P. Way, D.R. Danforth, Using virtual standardized patients to accurately assess information gathering skills in medical students, Med. Teach. 41 (2019) 1053–1059.
[30] B. Duffy, R. Tully, A.V. Stanton, An online case-based teaching and assessment program on clinical history-taking skills and reasoning using simulated patients in response to the COVID-19 pandemic, BMC Med. Educ. 23 (2023) 4.
[31] S. Herbstreit, S. Benson, C. Raiser, C. Szalai, A. Fritz, F. Rademacher, G. Gradl-Dietsch, Experience with an OSCE anamnesis station via Zoom: feasibility, acceptance and challenges from the perspective of students, simulated patients and examiners during the COVID-19 pandemic, GMS J. Med. Educ. 39 (2022) Doc44.
[32] K. F.H. Frangoudes, M. Schiza, E.C. Matsangidou, M. Tsivitanidou, O. Neokleous, P. I. Zaphiris, An overview of the use of chatbots in medical and healthcare education,

A.. Learning and Collaboration Technologies: Games and Virtual Environments For Learning. HCII 2021. Lecture Notes in Computer Science, Springer, Cham, 2021, pp. 170–184.