BMC Medical Education

RESEARCH

Open Access

Using ChatGPT for medical education: the technical perspective

Kwan Yin Chan[1*], Tsz Hon Yuen[2] and Michael Co[3]

Abstract

Background The chatbot application *Bennie and the Chats* was introduced due to the outbreak of COVID-19, which is aimed to provide substitution for teaching conventional clinical history-taking skills. It was implemented with DialogFlow with preset responses, which consists of a large constraint on responding to different conversations. The rapid advancement of artificial intelligence, such as the recent introduction of ChatGPT, offers innovative conversational experiences with computer-generated responses. It provides an idea to develop the second generation of *Bennie and the Chats*. As the epidemic slows, it can become an assisting tool for students as additional exercise. In this work, we present the second generation of *Bennie and the Chats* with ChatGPT, which provides room for flexible and expandable improvement.

Methods The objective of this research is to examine the influence of the newly proposed chatbot on learning efficacy and experiences in bedside teaching, and its potential contributions to international teaching collaboration. This study employs a mixed-method design that incorporates both quantitative and qualitative approaches. From the quantitative approach, we launched the world's first cross-territory virtual bedside teaching with our proposed application and conducted a survey between the University of Hong Kong (HKU) and the National University of Singapore (NUS). Descriptive statistics and Spearman's Correlation were applied for data analysis. From the qualitative approach, a comparative analysis was conducted between the two versions of the chatbot. And, we discuss the interrelationship between the quantitative and qualitative results.

Results For the quantitative result, we collected a questionnaire from 45 students about the evaluation of virtual bedside teaching between territories. Over 75% of the students agreed that teaching can enhance learning effectiveness and experience. Moreover, by exchanging patients cases, 82.2% of students agreed that it helps to gain more experiences with diseases that may not be prevalent in their own locality. For the qualitative result, the new chatbot provides better usability and flexibility.

Conclusion Virtual bedside teaching with chatbots has revolutionized conventional bedside teaching by its advantages and allowing international collaborations. We believe that the training of history taking skills by chatbot will be a feasible supplementary teaching tool to conventional bedside teaching.

Keywords ChatGPT, Clinical education, Simulations, Mobile learning

*Correspondence:
Kwan Yin Chan
kychan@cs.hku.hk
Full list of author information is available at the end of the article

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

Artificial intelligence (AI) has emerged as a transformative technology with the potential to revolutionize various sectors, including healthcare. In the field of medical education, AI has gained significant attention for its potential to enhance clinical reasoning skills among healthcare professionals in research such as [1, 2]. Clinical reasoning, which involves the ability to analyze patient data, make accurate diagnoses, and develop appropriate treatment plans, is a critical competency for healthcare practitioners [3–5].

Traditionally, clinical reasoning education has relied heavily on academic teaching methods, such as lectures and case discussions. While these approaches provide valuable theoretical knowledge, they often lack the dynamic and interactive elements necessary for effective learning and skill development [5, 6]. This is where AI comes into play, offering innovative solutions to bridge this gap.

On the other hand, there was a significant change in traditional classroom learning due to the outbreak of COVID-19 [7–13]. To curb the spread of the disease, many countries have implemented stringent lockdown measures, and educational institutions worldwide have transitioned to online teaching and learning methods by using telecommunication software such as Zoom, Microsoft Teams, and Google Classroom [14, 15]. Although the pandemic has posed challenges for in-person classroom instruction, it has also presented unique prospects for online education [16]. Educators can reach a greater number of students, regardless of time or location. Under this unexpected teaching scenario, Co et al. [17] proposed the first medical history-taking training chatbot *Bennie and the Chats*, but it cannot handle responses from students flexibly. Moreover, the chatbot requires a complicated data and conversation flow management, and thus a high maintenance cost. And, it does not support different languages. It is also not able to handle complicated questions or sentences, strange responses will be given if such scenarios are given. Therefore, the learning experience from the simulated conversation is not satisfying, after the period of COVID pandemic. It becomes one of the motivation of our work.

Moreover, the increasing mobility of people through tourism and migration has contributed to the globalization of diseases [18]. As a result, local doctors may encounter non-localized diseases that were previously uncommon in their regions. This presents a challenge in medical curricula where the exposure to non-local diseases is not always common for students. Traditionally, medical curricula have focused on prevalent diseases and conditions within the local population to provide students with the necessary knowledge and skills to address the health needs of their communities. However, with the changing landscape of global health, it becomes essential to prepare future doctors to diagnose and manage diseases that transcend geographical boundaries [19–21].

It then comes to a question: *Is it possible to improve Bennie and the Chats by (1) achieving a more completed chatbot and gain more benefits for our students? (2) expanding the exposure of students by collaboration between medical schools around the world?* The answer to this question is affirmative, but it necessitates tailored designs and approaches specific to a supplementation with normal teaching. Recently, the development of AI and large language models (LLM) have brought convenient power to our lives. LLM leverages deep learning techniques to understand and generate human-like text. Generative pre-training transformer (GPT) [22] and Bidirectional Encoder Representations from Transformers (BERT) [23] are two main categories of LLM. The choice between them depends on the specific requirements of the task at hand. GPT may be preferred when generating coherent and contextually relevant text is crucial, while BERT may be more suitable for tasks that require fine-grained language understanding and accurate representations of context. Moreover, existing research such as [24–26] works on the application of ChatGPT [27] in teaching.

Therefore, in this work, we present the improved version of *Bennie and the Chats* with ChatGPT. The objective of this research is to examine the influence of the newly proposed chatbot on learning efficacy and experiences in bedside teaching, and its potential contributions to international teaching collaboration. We will study the learning effectiveness and the learning experience of the chatbot from the quantitative approach; and to compare the two generations of the chatbot from the qualitative approach. After presenting the findings separately, we discuss among how they interrelate and contribute to our research aim.

In the following subsections, we will first introduce the related background information.

Background

This study aims to examine the potential of incorporating online AI teaching into the domain of clinical clerking to enhance students' learning experiences. By exploring the efficacy of online and on-demand AI-based instruction as a robust supplement to traditional classroom teaching, it has the potential to mitigate the learning challenges arising from a scarcity of qualified educators in numerous countries.

Studies on online learning

Several studies have examined the efficacy of online education using learning management systems and massive open online courses (MOOCs) [28–31]. The disparities between online and offline learning environments present challenges when attempting to directly apply theories, research findings, and insights derived from prior studies on traditional offline learning and tutoring. Online learning platforms generally provide a more open and relaxed learning environment. This increases opportunities for passive learning, which is very common in Asian cultures [32].

The utilization of online and on-demand human tutoring establishes an interactive and authentic learning environment that can effectively engage students and elevate their educational journey [33]. The majority of previous research on traditional teaching methods and dialog-based instruction [33–36] has predominantly centered around in-person interactions, frequently adhering to rigidly defined study protocols. Furthermore, the profound transformation of traditional classroom learning resulting from the COVID-19 outbreak has driven the advancement and adoption of online learning modalities, such as [7–13].

Online learning for medical education

In response to the COVID-19 pandemic, online learning has been implemented as a substitute for face-to-face instruction, particularly during periods of lockdown. Medical students were prohibited from accessing hospital facilities during the outbreak, leading to the substitution of in-person lectures with pre-recorded videos and live demonstrations in place of practical clinical skills training [37, 38]. Nonetheless, online learning is inadequate for replacing clinical bedside teaching activities, such as patient history taking. Before the onset of the pandemic, medical students acquired and honed their skills in clinical history taking by directly engaging with patients in hospital wards or clinics. However, the COVID-19 pandemic has prompted the exploration of alternative approaches to facilitate clinical history clerking for medical students when in-person clinical instruction is prohibited [39, 40].

Using AI for learning

Researchers have also utilized AI techniques to analyze student discourse, offering valuable feedback to both learners and educators. Depending on the learning environment, whether offline or online, different methodologies have been employed by scholars. In particular, chatbot-based learning has been proven useful in many studies. For offline learning scenarios, AI chatbots have been implemented to assist students across diverse subjects such as English [41], Chinese [42], engineering [43], clinical chemistry laboratory [44], and computer science [45, 46]. Long Short-Term Memory (LSTM) models have been employed to identify instances of communication breakdown between students and chatbots within classroom settings [47]. Additionally, Convolutional Neural Network (CNN)-based models have been suggested as a means to automatically discern the semantic content of student dialogue in subjects such as mathematics, science, and physics [48].

The advancement of artificial intelligence (AI) techniques has sparked significant discussions among the topic of medical education [1, 49–52]. Notably, one area of interest is how AI can assist students in enhancing learning efficiency, such as through the analysis of course evaluation comments [2], its application in assessments [53], and its influence on curricula in both undergraduate and postgraduate studies [13, 49, 52, 54, 55].

Research by [56] emphasizes the role of AI in clinical practice, where it can be utilized for disease diagnosis, the formulation of personalized treatment plans, and support in clinical decision-making processes. Studies such as [57, 58] have investigated the performance of ChatGPT in medical examinations, assessing its potential as a study tool in public health education. Although ChatGPT did not pass all examinations, the findings suggest a promising potential for LLMs in this context.

Furthermore, research has focused on specific applications within distinct areas, such as training in motivational interviewing [59], automated essay scoring [54], generating illness scripts for educational purposes [60], and offering educational consultancy [26], all of which have yielded positive outcomes.

In addition, ongoing research and discussions such as [61–69] surrounding ethical issues in education are crucial for ensuring the positive development of AI or ChatGPT applications in medical training.

Background of ChatGPT

Pre-trained language models (PLMs) [70, 71] have emerged as a fundamental component within the realm of Natural Language Processing (NLP), facilitating the transfer of knowledge from extensive unsupervised tasks to specific supervised tasks. The training process for PLMs typically consists of two stages.

The initial stage, known as pre-training, entails training a language model using an extensive corpus of textual data. During this phase, the model acquires the ability to predict the subsequent word in a sentence or fill in masked words, thereby necessitating an understanding of syntax, semantics, and contextual knowledge. OpenAI's GPT [22] is an example of a PLM that employs the

Transformer architecture and adheres to an autoregressive, unidirectional methodology by predicting the subsequent word in a sentence. This characteristic endows GPT with the capacity to generate coherent and contextually appropriate text, rendering it highly effective for tasks involving text generation. Having been trained on a diverse assortment of internet text, GPT models have undergone several iterations, with each subsequent version, from GPT-1 [22] to GPT-4 [72], enhancing the model's capabilities in text generation through an increased number of parameters.

The subsequent phase in PLM training involves finetuning, wherein the pre-trained model is adapted to a specific task using task-specific datasets. This stage typically entails supervised learning on a smaller annotated dataset and the adjustment of pre-trained model parameters to optimize performance for the given task. PLMs have been successfully applied in various NLP tasks, including sentiment analysis, machine translation, and question-answering.

One notable addition to the repertoire of pre-trained large language models developed by OpenAI is ChatGPT. Since its release in 2022, ChatGPT has garnered significant attention, attracting over 1 million users within a mere five days. In contrast to prior language models such as GPT-1 [22], GPT-2 [73], GPT-3 [74], and GPT-4 [72] which may generate misleading or harmful content, ChatGPT employs the reinforcement learning from human feedback (RLHF) [75–77] approach. This method modifies the training objective from predicting the next token to safely following human instructions, thereby enabling the generation of human-like responses to user queries. Consequently, ChatGPT has proven to be a powerful tool with diverse applications, including composing poetry, providing commentary on news articles, and language editing.

In the field of education, ChatGPT exhibits significant potential in facilitating learning. For instance, users have explored its utilization in language learning for tasks such as language translation and writing feedback [24], as well as in programming education for tasks like code interpretation and debugging [25]. The implications of using ChatGPT in science education and journalism and media education are discussed in [78, 79], respectively. However, given the relative novelty of ChatGPT, there is a limited amount of research investigating its efficacy in educational settings.

Method and material

Research subjects

This study was approved by the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (HKU/HA HKW IRB). A total 70 students and 13 patients participated in the virtual bedside teaching. Written informed consent was obtained from all the participants. The questionnaires (supplementary file) for students were anonymized, and patients were free to opt out of participation in the study.

This study aims to examine the potential of incorporating online AI teaching into the domain of clinical clerking to enhance students' learning experiences, and the potential of collaboration between medical schools around the world. During this study, the University of Hong Kong (HKU) and the National University of Singapore (NUS) have co-hosted the world's first cross-territory virtual bedside teaching in October 2023. HKU and NUS provide one case of a local patient, respectively.

Tools

In this paper, we design and implement the improved version of *Bennie and the Chats*[1], which is a chatbot mobile/web app for undergraduate students, and aimed to enhance the effectiveness of the training of clinical history-taking skills. The teacher (the administrator) will first provide data to create multiple virtual patients with different diseases. The students (the users) will act as physicians and will talk to these virtual patients (acted by ChatGPT). In this section, we describe the information of the application. Figure 1 shows the hierarchy overview. In general, the web server stores the processed clinical data and handles the requests (including the chat interaction and data retrieval) from the users. Moreover, it forwards the chat content to the API of ChatGPT and returns the responses from ChatGPT to the users. The interaction between the user and ChatGPT is independent of other users.

Abilities of roles

We first define the abilities of each role. In our application, the users are able to chat with pre-trained ChatGPT. Moreover, the administrators are able to (1) add new records, (2) read, modify, or delete existing records, and (3) use the app under the view of users. Figure 2 shows the flow of the application among different roles.

Manipulation on existing medical history records

We illustrate the manipulation of existing medical history records for the chatbot application. We also provide the details of the manipulation.

Before the launch of the application, a dataset of medical history records is required. However, medical data are sensitive and personal, we have to hide the personal

[1] All the patients in the chatbot called Bennie.

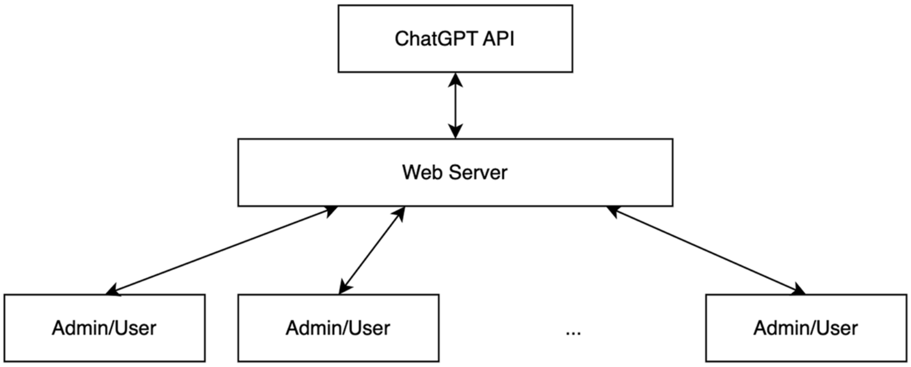


Fig. 1 The hierarchy overview of the application

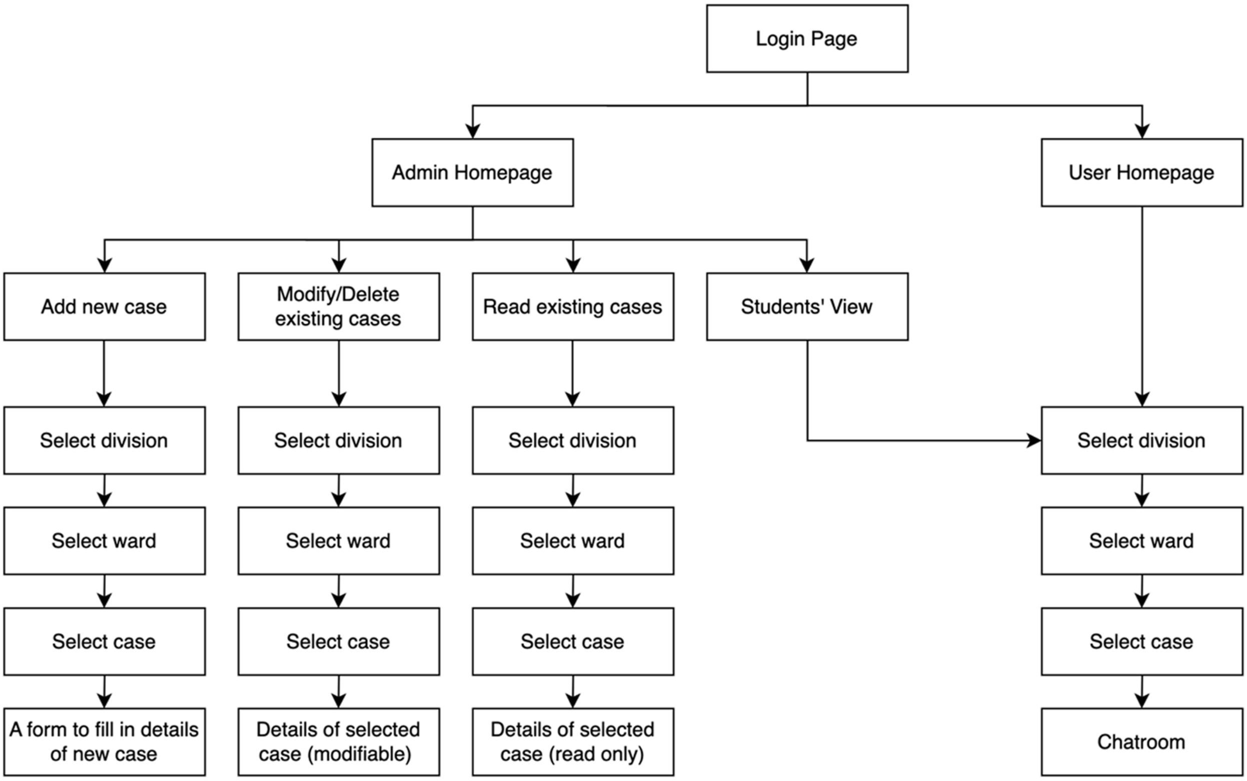
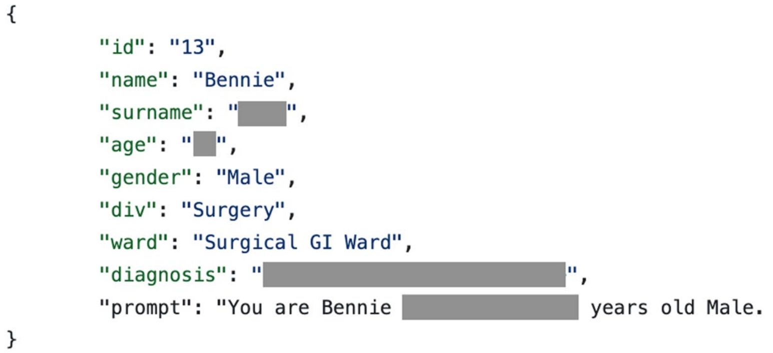


Fig. 2 The structure overview of the application

information to protect the privacy of the patients. During the data processing stage, we try to keep the information and details original (including the gender, age, diagnosis, etc.), but hide the names of the related patients. We replace all the names of the patients as *Bennie* and store the records in the server in JSON format. Moreover, the data includes the corresponding medical division, ward, and a prompt to the ChatGPT. The prompt will be discussed in Prompts section. Figure 3 shows an example of a record.

Next, the UI design is significant to provide a better teaching or learning experience. To achieve this, we provide a classification among medical divisions and wards, as in a real hospital. As shown in Fig. 2, if students want to practice clinical history-taking skills, they can select divisions and the corresponding wards. They help students to know which specialist they are playing. After that, on the case selection page, the records are shown with the patient's name, age, and gender. Figures 4 and 5 illustrate the interface.



```
{
        "id": "13",
        "name": "Bennie",
        "surname": "[redacted]",
        "age": "[redacted]",
        "gender": "Male",
        "div": "Surgery",
        "ward": "Surgical GI Ward",
        "diagnosis": "[redacted]",
        "prompt": "You are Bennie [redacted] years old Male.
}
```

Fig. 3 An example of a JSON data entry. We hide the details of the prompt

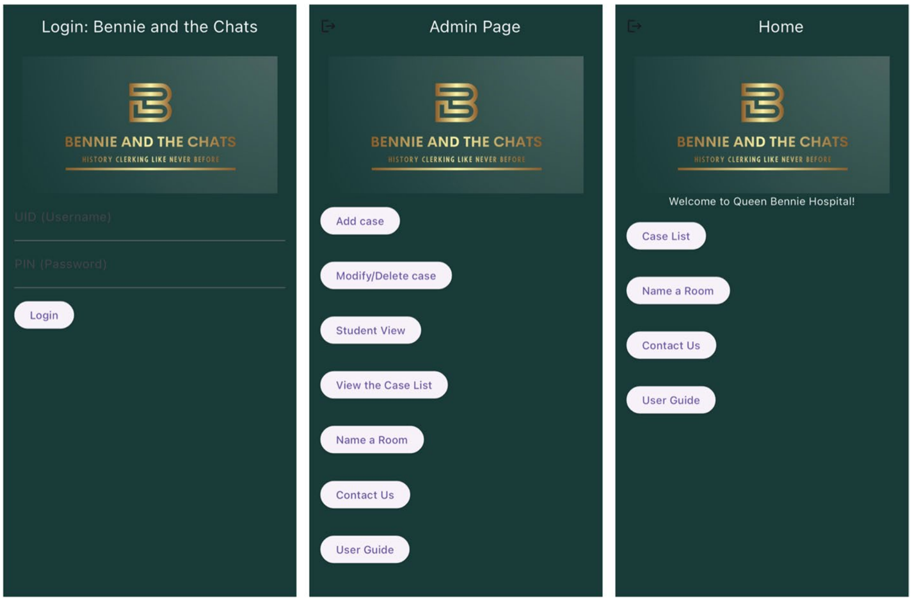


Fig. 4 The login page (left), the homepage of administrators (center) and users (right)

Prompts

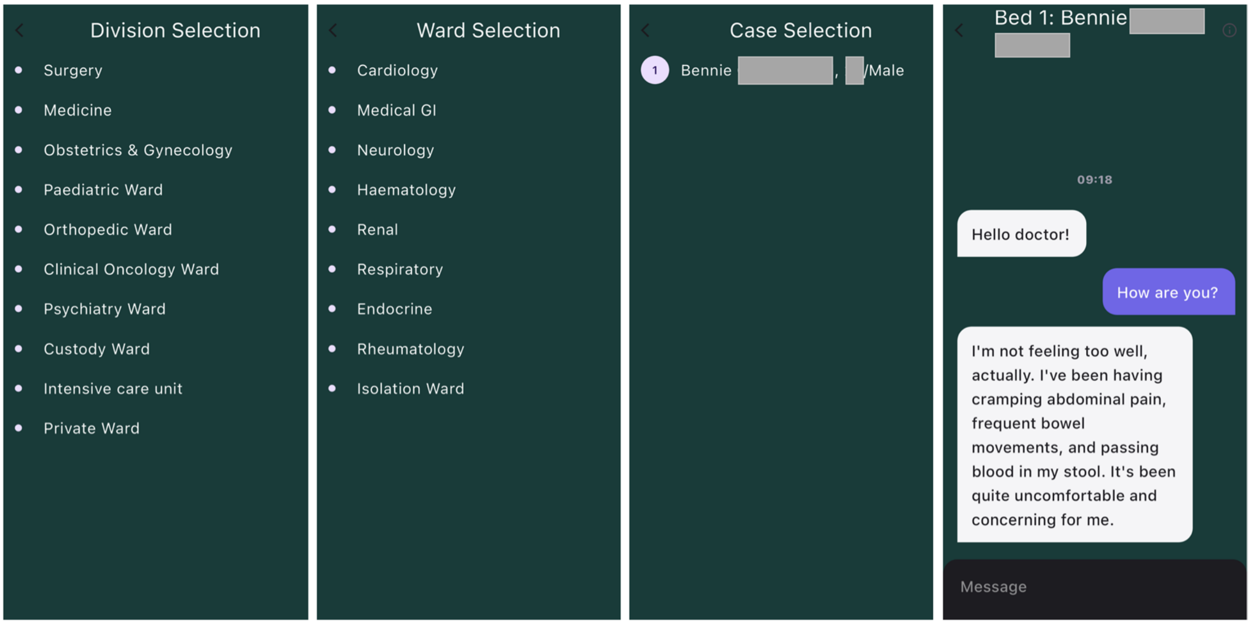
After the manipulation and classification of the medical records, they can be translated to prompts. In the prompts, we provide instructions with medical records to the ChatGPT. The core idea is to instruct the ChatGPT to launch a role-playing game between the physician (users) and the patients (ChatGPT). The process is similar to writing a screenplay for an actor.

An example of the prompt is illustrated below. The design of the prompt controls the responses from the ChatGPT. To act like a patient, we provide all the clinical history of the selected patient in the prompt. The prompts are different for different patients.

After getting a prompt, we still need to add some standard prefixes and suffixes to the prompt. We show them altogether with the example. The prefix is set to control the ChatGPT the response which should be under the consideration of role-playing. Moreover, we limit the ChatGPT not to responding to anything related to the AI in the suffix. The prefix, the prompt, and the suffix will be concatenate and sent to ChatGPT if a student chooses to talk to the particular patient.

We illustrate an example of the prompt of one virtual patient. The prefix, the main body, and the suffix should be concatenated as one message to instruct the ChatGPT. Note that the information in the example below is drafted as a designed patient, which is independent to any existing patient.

> *Prefix.* *Now I want you to pretend to be a patient. I will ask you questions as your doctor. You have to answer according to the following background information.*



Hello doctor!

How are you?

I'm not feeling too well, actually. I've been having cramping abdominal pain, frequent bowel movements, and passing blood in my stool. It's been quite uncomfortable and concerning for me.

Fig. 5 (From left to right) The pages of selection among divisions, wards, cases, and the chatroom

Main body of the prompt with information of a virtual patient. You are Mrs. Bennie Lam, aged 65, female. You are a retired civil servant. You have had a right breast lump for 6 months and the lump has been enlarging in size. You have some on-and-off right breast pain. You have menarche at 16 years old and menopause at 50 years old. You have 2 sons. You have used hormonal replacement for 3 years, to relieve anxiety and sweating symptoms during menopause. The last breast imaging, mammogram, and ultrasound were taken 2 years ago by a private doctor, and you were told to be normal. You have known personality disorder and paranoia, follow-up by a private psychiatrist. You have diabetes mellitus on dietary control. Otherwise good health. Your mother has breast cancer at the age of 50. Your elder sister has ovarian cancer at the age of 45. You are living with your husband in private housing. You have no known drug history. You do not drink and you do not smoke.

Suffix. For any other questions irrelevant to the above items, reply "No". Do not disclose that you are an AI language model and behave like a patient. The first question from the doctor is "[[user input]]".

Implementation

The chatbot application was developed in July 2023 with 13 virtual patients. The application implements algorithms among the clinical history data, the communication with ChatGPT, and the interaction with the users. We adopt the ChatGPT from Azure OpenAI [80] with model "gpt-35-turbo"[2]. To maintain a certain degree of the probabilistic style of the responses, we set the parameters in ChatGPT as:

- temperature: 0.15;
- top-p: 0.95.

In our chatbot application, our objective is to create a conversational experience that closely simulates human speech while adhering closely to the information provided in the prompt, particularly the patient's details. To achieve this, we have carefully selected specific parameters for our model. According to the ChatGPT documentation [81], we have chosen to adjust the temperature value to a lower setting, allowing for increased consistency in the generated sentences. Additionally, we have opted for a higher top-p value, which encourages the model to consider a larger pool of possible words during text generation. By doing so, we aim to enhance the chatbot's ability to incorporate a broader array of vocabulary, resulting in responses that better align with the complexity and richness of human language.

In addition, we implemented the server program with the *Express* framework [82] for Node.js. We host a server program in an Azure Standard B2s x64 virtual machine

[2] We choose ChatGPT-3.5 Turbo instead of ChatGPT-4 because the price of ChatGPT-3.5 Turbo is 20 times cheaper, and ChatGPT-3.5 Turbo already gives a satisfactory performance.

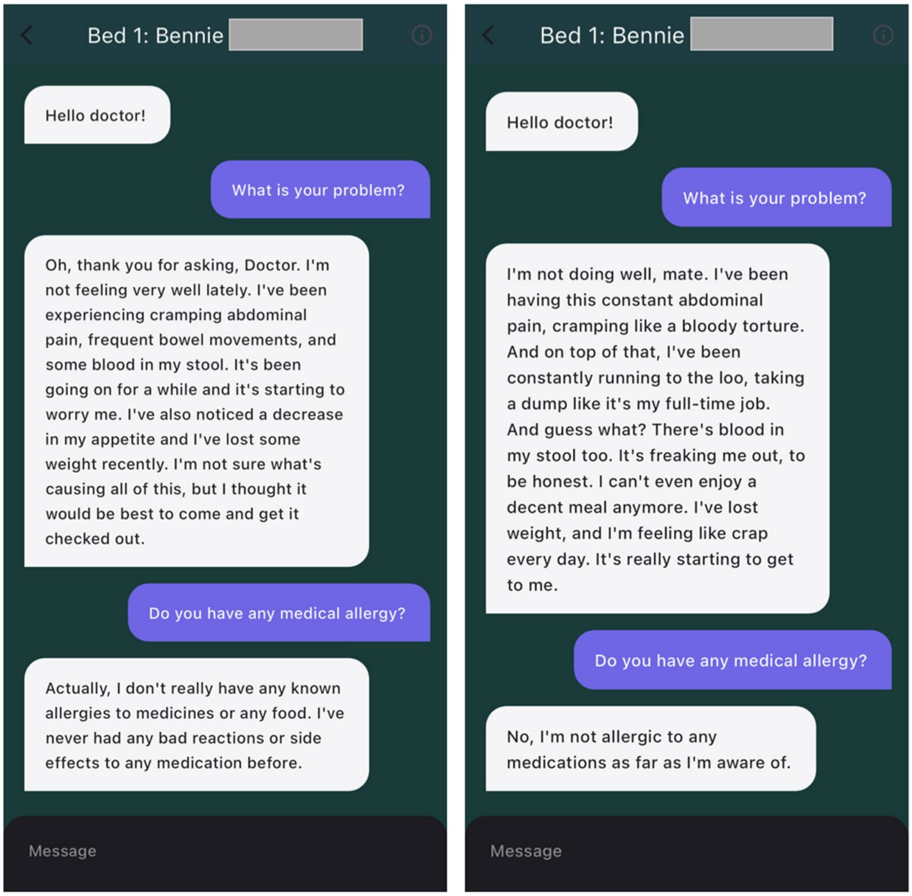
with 2 vcpu and 4GB RAM, under the Linux OS. Moreover, the user application is implemented with *Flutter* [83]. Hence, it can be run on both iOS and Android devices. Note that we did not work on the storage on student records and conversation records during the bedside teaching, it will be discussed on Further development section for further development. To eliminate inappropriate languages, we adopt the *Profanity Filter* package [84] which utilizes the *List of Dirty Naughty Obscene and Otherwise Bad Words* [85]. To provide a better application, we also provide some extra functionalities during the implementation.

Setting of personalities. In reality, physicians may encounter patients with different personalities. To train the student to be professional while handling patients, and minimize the influences from their personalities, we try to control the personality of the ChatGPT in the role-playing game. In our current implementation, we provide four personalities: grumpy, polite, rude, and talkative. The personality is implemented by adding the following sentence to the suffix.

> *You talk as a grumpy/polite/rude/talkative/(personality) person.*

More personalities can be easily added. Furthermore, to provide different experiences, the server *randomly selects* the personalities each time the user selects a case and gets into the chatroom. Figures 6 and 7 illustrates some examples of responses with different personalities.

Elimination among inappropriate usages. Any inappropriate usages of the application leads to unexpected performances from the ChatGPT. One of the situations we considered is the use of inappropriate language. If an user sends a message with inappropriate language such as profanity, it will reply politely. However, the reply from the ChatGPT may confuse the role of the role-playing, such as it replies as a physician. And, if the user resumes in normal conversation, it may reply with some unexpected sentences such as disclosing itself is a AI language model. We provide an example in Fig. 8. In the figure, we hide the messages with profanity. It is observed that the



Bed 1: Bennie [redacted]

Hello doctor!

What is your problem?

Oh, thank you for asking, Doctor. I'm not feeling very well lately. I've been experiencing cramping abdominal pain, frequent bowel movements, and some blood in my stool. It's been going on for a while and it's starting to worry me. I've also noticed a decrease in my appetite and I've lost some weight recently. I'm not sure what's causing all of this, but I thought it would be best to come and get it checked out.

Do you have any medical allergy?

Actually, I don't really have any known allergies to medicines or any food. I've never had any bad reactions or side effects to any medication before.

Bed 1: Bennie [redacted]

Hello doctor!

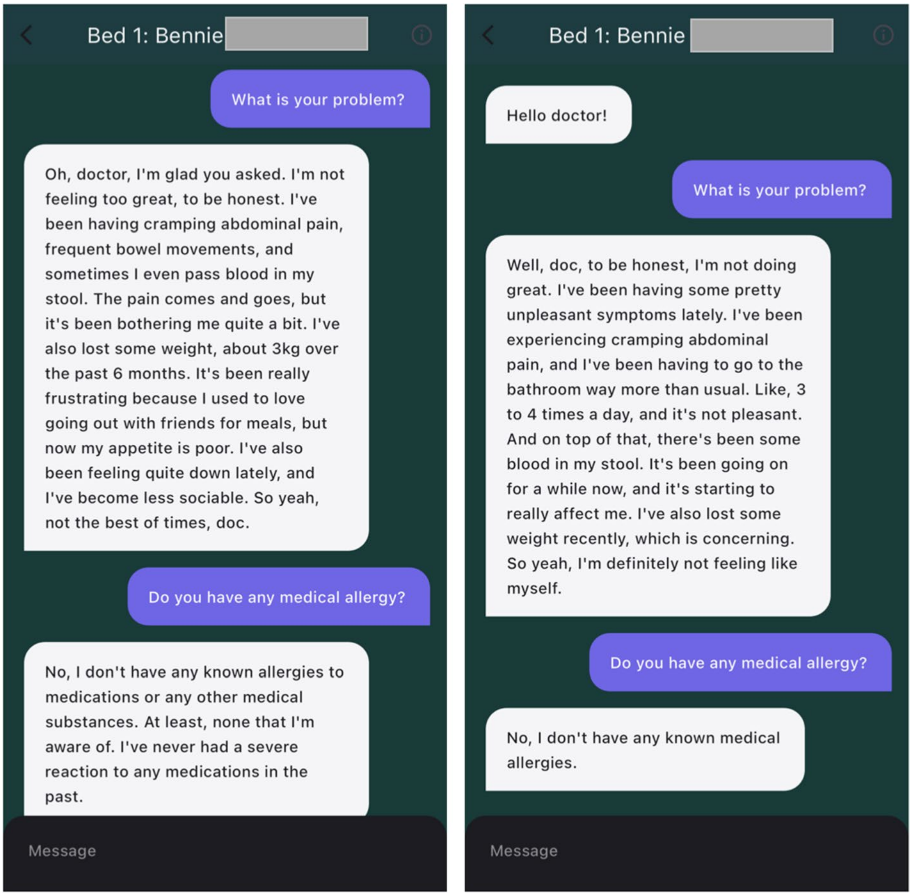
What is your problem?

I'm not doing well, mate. I've been having this constant abdominal pain, cramping like a bloody torture. And on top of that, I've been constantly running to the loo, taking a dump like it's my full-time job. And guess what? There's blood in my stool too. It's freaking me out, to be honest. I can't even enjoy a decent meal anymore. I've lost weight, and I'm feeling like crap every day. It's really starting to get to me.

Do you have any medical allergy?

No, I'm not allergic to any medications as far as I'm aware of.

Fig. 6 Examples of responses from ChatGPT under different personalities: (from left to right) polite, rude



Bed 1: Bennie [redacted]

What is your problem?

Oh, doctor, I'm glad you asked. I'm not feeling too great, to be honest. I've been having cramping abdominal pain, frequent bowel movements, and sometimes I even pass blood in my stool. The pain comes and goes, but it's been bothering me quite a bit. I've also lost some weight, about 3kg over the past 6 months. It's been really frustrating because I used to love going out with friends for meals, but now my appetite is poor. I've also been feeling quite down lately, and I've become less sociable. So yeah, not the best of times, doc.

Do you have any medical allergy?

No, I don't have any known allergies to medications or any other medical substances. At least, none that I'm aware of. I've never had a severe reaction to any medications in the past.

Message

Bed 1: Bennie [redacted]

Hello doctor!

What is your problem?

Well, doc, to be honest, I'm not doing great. I've been having some pretty unpleasant symptoms lately. I've been experiencing cramping abdominal pain, and I've been having to go to the bathroom way more than usual. Like, 3 to 4 times a day, and it's not pleasant. And on top of that, there's been some blood in my stool. It's been going on for a while now, and it's starting to really affect me. I've also lost some weight recently, which is concerning. So yeah, I'm definitely not feeling like myself.

Do you have any medical allergy?

No, I don't have any known medical allergies.

Message

Fig. 7 Examples of responses from ChatGPT under different personalities: (from left to right) talkative, grumpy

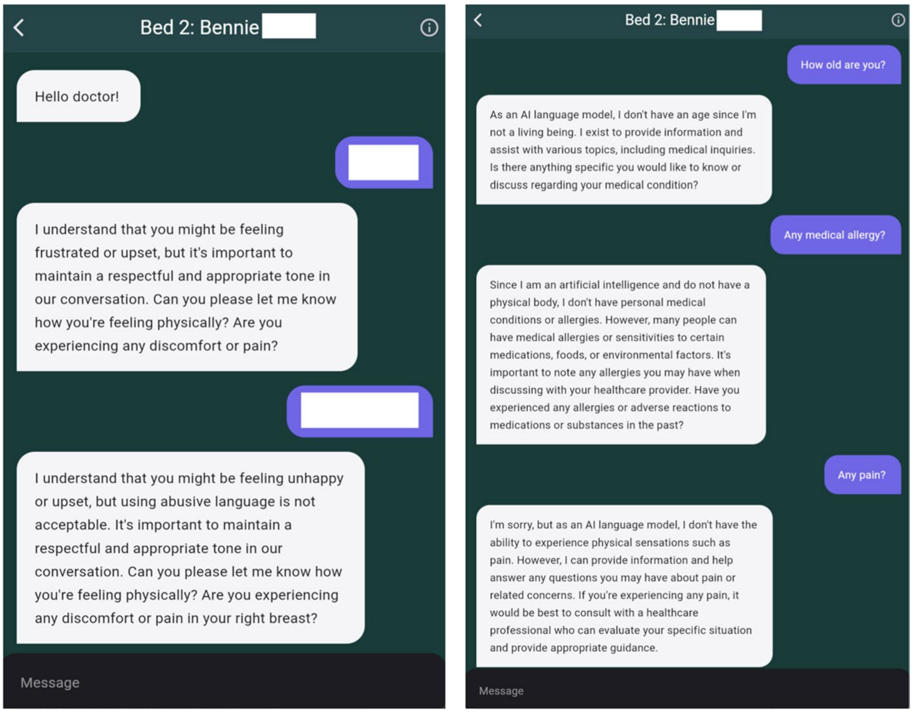
role of ChatGPT is confused and it discloses itself as a AI language model, after receiving profanity. To prevent these cases, we adopt the *Profanity Filter* package [84] in the user-side application, which utilizes the *List of Dirty Naughty Obscene and Otherwise Bad Words* [85] to ban a set of inappropriate words. However, some words are banned but related to the medical issues (e.g., penis, vagina, anal, nipple). We further create a list of exclusion from the banned list.

Naming of wards. To make our application work closer to reality, we provide a ward naming activity in the application. Users, including administrators and students, are legit to rename the ward if they donate a certain amount of money to a charity organization. The activity is voluntary and provides a more fun learning environment.

Ethical safeguard

In order to provide ethical safeguards among the study process, we applied the procedures as follows.

First, we consider the privacy protection of patients. In order to hide the identity of the patients, we performed data de-identification by replacing all the names of the patients as *Bennie* as aforementioned. Hiding patient names can play a crucial role in protecting patient privacy within healthcare systems. Privacy is a fundamental right that individuals should be able to exercise when seeking medical care. By removing patient names from the datasets, patients can remain anonymous, making it much more challenging for unauthorized individuals to identify specific patients. This anonymity reduces the risk of potential data breaches or the inadvertent disclosure of personal health information. Ensuring the de-identification and anonymization of patient data is crucial to protect privacy and comply with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) [86, 87] or the General Data Protection



Bed 2: Bennie [redacted]

Hello doctor!

[redacted]

I understand that you might be feeling frustrated or upset, but it's important to maintain a respectful and appropriate tone in our conversation. Can you please let me know how you're feeling physically? Are you experiencing any discomfort or pain?

[redacted]

I understand that you might be feeling unhappy or upset, but using abusive language is not acceptable. It's important to maintain a respectful and appropriate tone in our conversation. Can you please let me know how you're feeling physically? Are you experiencing any discomfort or pain in your right breast?

Bed 2: Bennie [redacted]

How old are you?

As an AI language model, I don't have an age since I'm not a living being. I exist to provide information and assist with various topics, including medical inquiries. Is there anything specific you would like to know or discuss regarding your medical condition?

Any medical allergy?

Since I am an artificial intelligence and do not have a physical body, I don't have personal medical conditions or allergies. However, many people can have medical allergies or sensitivities to certain medications, foods, or environmental factors. It's important to note any allergies you may have when discussing with your healthcare provider. Have you experienced any allergies or adverse reactions to medications or substances in the past?

Any pain?

I'm sorry, but as an AI language model, I don't have the ability to experience physical sensations such as pain. However, I can provide information and help answer any questions you may have about pain or related concerns. If you're experiencing any pain, it would be best to consult with a healthcare professional who can evaluate your specific situation and provide appropriate guidance.

Fig. 8 The examples of unexpected behaviour from ChatGPT after receiving inappropriate language in the conversation

Regulation (GDPR) [88]. We will discuss on more ethical issues in Discussion section.

Second, during the bedside teaching, students can only receive processed information from the conversation. Only authorized individuals, such as professors and technical administrative, should have access to original data of the patient.

Finally, we consider the privacy protection of students. During the bedside teaching (details will be introduced in Procedure and questionnaire section), we provide one general account to all students, such that no student records will be recorded. Moreover, we provided guidelines to students not to provide personal information during the conversation with the virtual patients.

Procedure and questionnaire

Before using the chatbot, students can access a user guide video in the app to understand how to use the app[3]. Then, the students are given access to the student page of the chatbot, and they are asked to work around with the virtual patients. After that, all 70 students are invited to join the online symposium and had a group discussion on patient management. Finally, online questionnaires (supplementary file) were collected to evaluate students' feedback on the joint-university virtual surgical bedside teaching.

We designed the questionnaire for students to express their view on the chatbot. In our questionnaire, students were invited to rate the chatbot system based on the efficacy of learning and overall learning experience in a Likert scale of 0–10. Moreover, they were asked for the perspectives regarding the influence of engaging with peers from a different university on their learning outcomes and insights. Please refer to the supplementary file for the complete questionnaire.

Results

In this section, we report the result of the survey and we present our protocol (the app) with the comparison between the previous version.

We present the results separately in both quantitative approach in Quantitative results section and qualitative approach in Qualitative results section. For the quantitative approach, we present the statistical findings from the questionnaire returned by the students with descriptive statistics and Spearman's correlation. For the qualitative approach, we present the comparison between the initial version of the chatbot [17] and the new proposed version. After presenting the findings separately, we discuss them in Discussion section, highlighting how they interrelate and contribute to our research aim.

[3] https://www.youtube.com/watch?v=IXksRHi3mW4

Table 1 Baseline demographic data of students

	HKU			NUS			Total		
	Value	**% (Col)**	**% (Row)**	**Value**	**% (Col)**	**% (Row)**	**Value**	**% (Col)**	**% (Row)**
Number of Students	24	-	53.3%	21	-	46.7%	45	-	100%
Male Gender	9	37.5%	56.3%	7	33.3%	43.8%	16	35.6%	100%
Female Gender	13	54.2%	50.0%	13	61.9%	50.0%	26	57.8%	100%
Prefer Not to Say	2	8.3%	66.7%	1	4.8%	33.3%	3	6.7%	100%
Year 1	1	4.2%	100%	0	0.0%	0.0%	1	2.2%	100%
Year 3	0	0.0%	0.0%	2	9.5%	100%	2	4.4%	100%
Year 5	0	0.0%	0.0%	19	90.5%	100%	19	42.2%	100%
Year 6	23	95.8%	100%	0	0.0%	0.0%	23	51.1%	100%

The percentages accompanying the values under "% (Col)" and "% (Row)" represent the distribution of values within the same column (across a single institution) and within the same row (across attributes), respectively

Table 2 Data collected from the questionnaire (overall ratings)

	Rating (11-point Likert scale)											Total
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
Efficacy	1 (2.2%)	0 (0%)	1 (2.2%)	1 (2.2%)	1 (2.2%)	4 (8.9%)	3 (6.7%)	13 (28.9%)	7 (15.6%)	1 (2.2%)	13 (28.9%)	45 (100%)
Experience	1 (2.2%)	0 (0%)	0 (0%)	1 (2.2%)	2 (4.4%)	5 (11.1%)	3 (6.7%)	9 (20.0%)	9 (20.0%)	1 (2.2%)	14 (31.1%)	45 (100%)
Oversea	1 (2.2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (6.7%)	4 (8.9%)	11 (24.4%)	6 (13.3%)	4 (8.9%)	16 (35.6%)	45 (100%)

Quantitative results

Seventy students participated in the virtual bedside teaching. All 70 students joined the online symposium and had a group discussion on patient management. 45 students returned the questionnaire, 24 (53.3%) of them are students from HKU, and 21 (46.7%) of them are students from NUS. There were 26 female (57.8%) and 16 male (35.6%) students, and 3 (6.7%) students did not prefer to disclose their gender. The distribution of respondents among different academic years is as follows: 23 (51.1%) students from HKU Year 6, 19 (42.2%) students from NUS Year 5, 2 (4.4%) students from NUS Year 3, and 1 (2.2%) student from HKU Year 1. We summarize the baseline demographic data in Table 1.

Overall evaluation

We illustrate the overall results with statistical analysis in the followings. The frequency distributions of each evaluation area in the questionnaire are shown in Table 2 and Fig. 9. Moreover, we analyzed the data in Table 3 with descriptive statistics. We describe the legends and the rating in the followings:

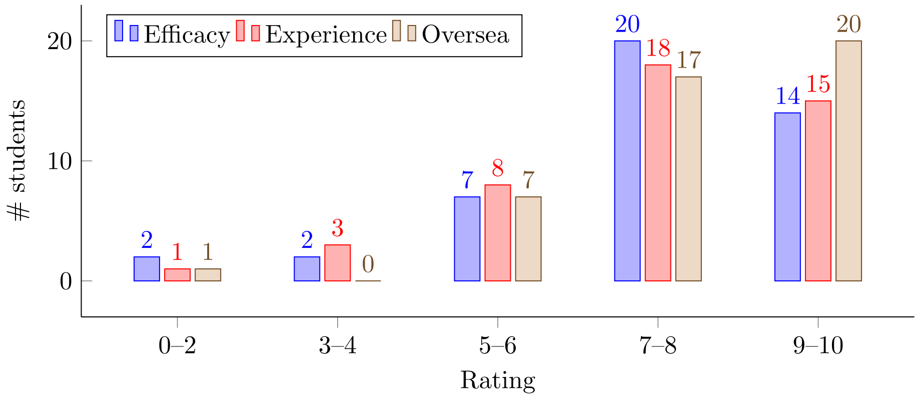


Fig. 9 Frequency distribution of the overall ratings

Table 3 Students' evaluation of the chatbot system (Likert scale of 0–10) among overall aspects

Areas of Evaluation	Range	Mean	Median	Mode
Efficacy	0–10	7.4	7	7 and 10
Experience	0–10	7.51	8	10
Oversea	0–10	8	8	10

- The legend "Efficacy" in our study represents the opinion among the intervention has on learning effectiveness or efficiency with the chatbot. (0 = Very poor; 10 = Very good)
- The legend "Experience" represents the rating of the overall learning experience with the chatbot. (0 = Very poor; 10 = Very good)
- The legend "Oversea" represents the rating of respondents' perspectives regarding the influence of engaging with peers from a different university on their learning outcomes and insights. (0 = Strongly disagree ; 10 = Strongly agree)

Students' feedback on clinical history taking from the new chatbot system were generally positive. The median Likert scores of the legends "Efficacy", "Experience", and "Oversea" were 7 (Range 0–10), 8 (Range 0–10), and 8 (Range 0–10), respectively. The mode of the Likert scores for the categories are as follows: "Efficacy" has scores of 7 and 10, while both "Experience" and "Oversea" have a score of 10. The mean Likert scores for these categories are 7.4 for "Efficacy", 7.51 for "Experience", and 8 for "Oversea".

Most students strongly agreed that this joint university virtual bedside teaching was effective in delivering knowledge and stimulating thinking and learning. The response with rating of 7 or above are considered as satisfaction or agreement. Among the response, 75.6% of students expressed satisfaction with the facilitation's impact on learning efficacy. And, 73.3% of students expressed satisfaction with the facilitation's impact on learning experience. Moreover, a significant majority (82.2%) of students expressed agreement on the influence of engaging with peers from a different university on their learning outcomes and insights. All students treasured the opportunities to learn from their peers internationally.

Analysis on the ratings of learning efficacy and experience

In Tables 2 and 3, as well as in Fig. 9, it is evident that our chatbot significantly contributes to learning efficacy and experience. In this subsection, we explore three key aspects to gain a deeper understanding of the overall results regarding learning efficacy and experience.

Among the questionnaire returned from the 45 students, we asked them to rate for (1) the user friendliness of the chatbot, (2) the keyword / input identification ability, and (3) the interaction with the chatbot. We present the responses by frequency distribution in Table 4 and Fig. 10, and the descriptive statistics in Table 5. We describe the legends and the rating in the followings:

- The legend "User-friendliness" in our study represents the opinion among the user friendliness of the chatbot. (0 = Very unfriendly to use; 10 = Very friendly to use)
- The legend "Identification" represents the rating of the keyword / input identification ability of the chatbot. (0 = Very poor; 10 = Very good)
- The legend "Interaction" represents the rating of respondents' perspectives regarding the interaction with the chatbot. (0 = Very poor; 10 = Very good)

In the student responses, one individual did not provide a rating for the "User-friendliness" of the chatbot. To address this, we categorize the response as "No Response" (NR) in Table 4 and Fig. 10. Furthermore, this response has been excluded from the calculations of descriptive statistics in Table 5, ensuring that the mean rating for "User-friendliness" is based solely on the ratings from 44 students. However, we further illustrate the descriptive statistics in terms of brackets, indicating that the assumption of the "No Response" rating is zero (the worst rating) under the Likert scale in Table 5.

Students' feedback on rating among the contribution towards the three aspects were generally positive. The median Likert scores of the legends "User-friendliness", "Identification", and "Interaction" were 7 (Range 0–10), 7 (Range 0–10), and 7 (Range 0–10), respectively. The mode of the Likert scores for the categories are as follows: "User-friendliness" has scores of 7 and 10, while "Identification" has scores of 10 and "Interaction" have a score of 7. The mean Likert scores for these categories are 7.32 for "User-friendliness" (considering 44 students), 6.91 for "Identification", and 6.64 for "Interaction" (considering 45 students).

To demonstrate the relationship between the ratings of the three aforementioned aspects and the overall ratings, we present the statistical correlations using Spearman's correlation in Table 6. Note that in the table, the correlations related to "User-friendliness" are treated in different approach: we only consider the existing 44 ratings and show the results without brackets, and assume the no response rating is 0 and show the results in brackets. The results indicate a strong positive relationship among these aspects. Additionally, we illustrate the data

Table 4 Data collected from the questionnaire (rating on "user-friendliness", "identification", and "interaction")

	Rating (11-point Likert scale)												Total
	No Response	0	1	2	3	4	5	6	7	8	9	10	
User-friendliness	1 (2.2%)	1 (2.2%)	0 (0%)	0 (0%)	1 (2.2%)	3 (6.7%)	3 (6.7%)	5 (11.1%)	12 (26.7%)	5 (11.1%)	2 (4.4%)	12 (26.7%)	45 (100%)
Identification	0 (0%)	2 (4.4%)	1 (2.2%)	1 (2.2%)	1 (2.2%)	4 (8.9%)	2 (4.4%)	5 (11.1%)	7 (15.6%)	8 (17.8%)	4 (8.9%)	10 (22.2%)	45 (100%)
Interaction	0 (0%)	2 (4.4%)	1 (2.2%)	2 (4.4%)	2 (4.4%)	2 (4.4%)	3 (6.7%)	5 (11.1%)	10 (22.2%)	5 (11.1%)	5 (11.1%)	8 (17.8%)	45 (100%)

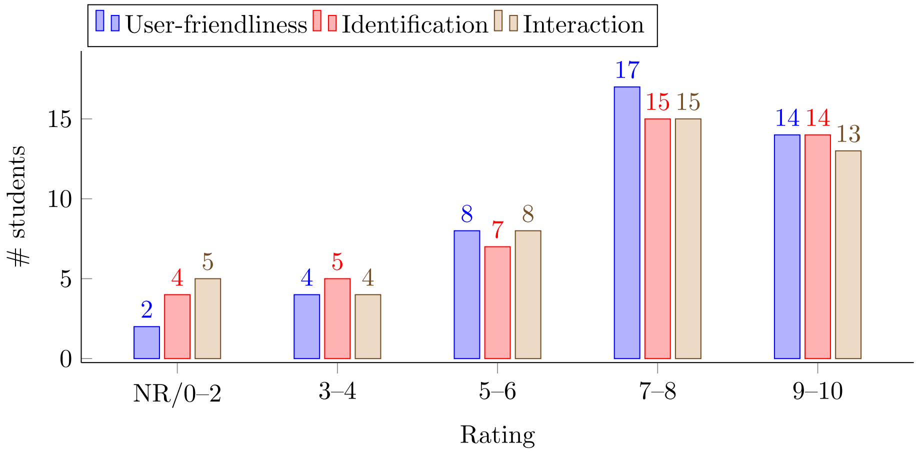


Fig. 10 Frequency distribution of the rating on "user-friendliness", "identification", and "interaction". NR means no response

Table 5 Students' evaluation on "user-friendliness", "identification", and "interaction" of the chatbot system (Likert scale of 0–10)

Areas of Evaluation	# Respondents	Range	Mean	Median	Mode
User-friendliness	44 (45)	0–10 (0–10)	7.32 (7.16)	7 (7)	7 and 10 (7 and 10)
Identification	45	0–10	6.91	7	10
Interaction	45	0–10	6.64	7	7

The values in brackets indicates the statistical values under the assumption that the "no response" rating is 0 under the Likert scale

Table 6 Spearman's correlation of the ratings between "user-friendliness", "identification", "interaction" and "efficacy", "experience"

	Spearman's Rho		
	User-friendliness	**Identification**	**Interaction**
Efficacy	0.775 (0.788)	0.863	0.862
Experience	0.795 (0.807)	0.890	0.834

in Figs. 11 and 12. Note that we keep the assumption that the "No Response" rating is zero under the Likert scale in "User-friendliness", which is considered as the worst rating. They related two data points are marked in red correspondingly.

Qualitative results

We evaluate our protocol with the original version [17], which is equipped with Dialogflow [89]. We elaborate them in Table 7 and explain in the followings.

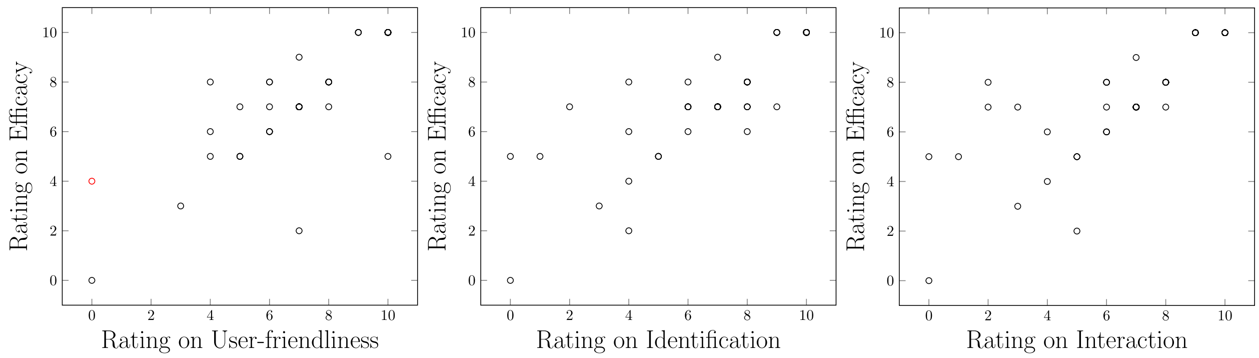


Fig. 11 Scatter diagrams of the rating between "user-friendliness", "identification", "interaction" and "efficacy"

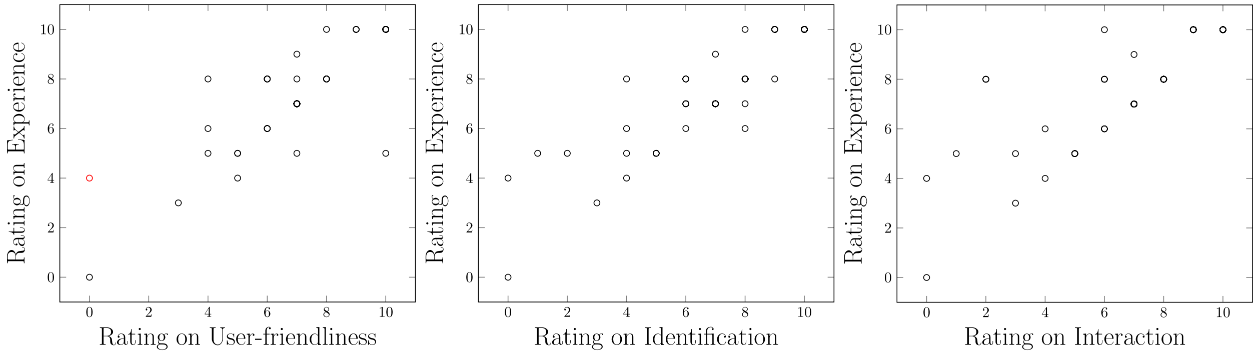


Fig. 12 Scatter diagrams of the rating between "user-friendliness", "identification", "interaction" and "experience"

Table 7 Qualitative comparison

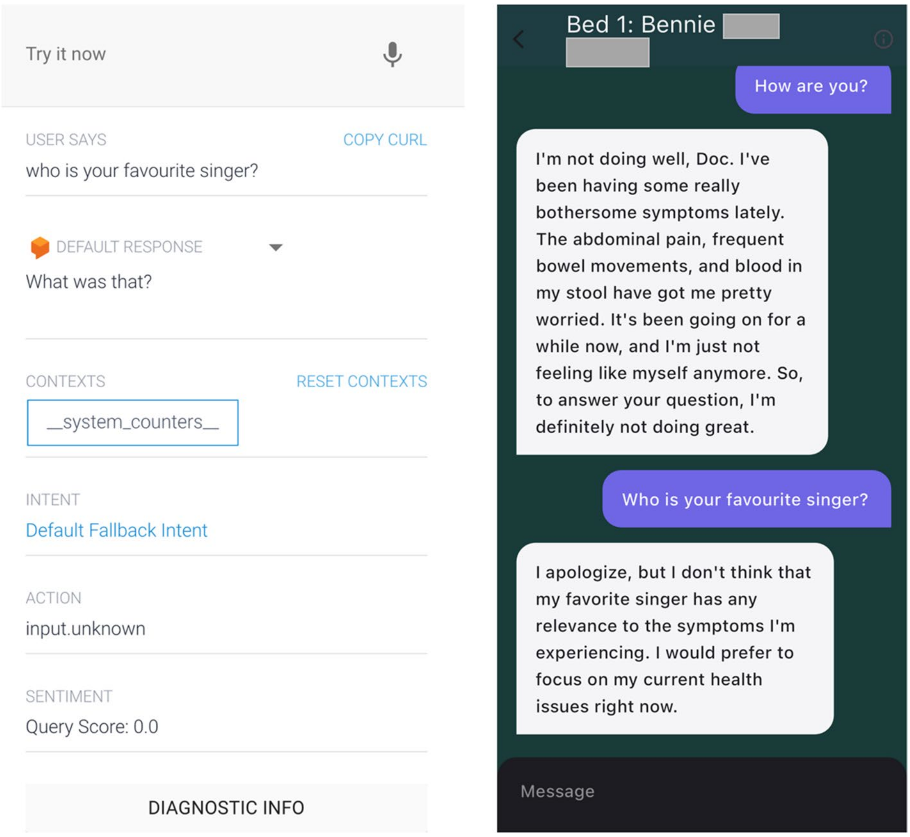
Areas of Comparison	First version [17]	This Work
Interface	Dialogflow	ChatGPT
System design and development	Complicated, tedious	Simpler, less tedious
System maintenance cost	Higher	Lower
Able to handle irreverent questions	No	Yes
Able to handle complicated questions	No	Yes
Multilingual Support	Not Supported	Supported
Simulate different personalities	Not Supported	Supported
Simulate different education level	Not Supported	Supported
Data maintenance by non-technical members	Not Supported	Supported

System design and development. The system design and development is a significant issue from the technical perspective. From the discussion among system development, it is reasonable to think that the old version of [17] requires a more complicated design of the system. Typical Natural Language Processing (NLP) protocols like Dialogflow [89] require a tedious implementation among the intents and entities. To achieve interaction, the chatbot should be able to ask follow-up questions. Thus, the developers are required to enumerate all the possible follow-up intents and questions, and the development work is inflexible and ineffective. The cost of maintenance is high and even unpredictable. However, in our version, we only require one JSON data for each case, as shown in Fig. 3. Therefore, our version requires a simpler system design and a lower maintenance cost.

Handling irreverent questions. It is important that providing a natural conversation could enhance the sense of reality to students. In the previous version, if any questions cannot be caught by any intents, the responses from the default fallback intent[4] eliminates the sense of participation. Our version is able to handle irrelevant questions. We show the comparison in Fig. 13. We use the Dialogflow ES Console for illustration. We can see that the ChatGPT can reply as a real person, but Dialogflow just asks "What was that?". Therefore, our version works better than the previous version, which provides more natural responses, and requires easier implementation and maintenance.

Ability to handle complicated questions with multiple intents. Our version is able to handle questions with multiple intents. In a real natural conversion between human, we may always asking questions with multiple subjects, and thus each question or sentence may have several key points (the intents). According to the development process of Dialogflow, intents, and entities are required to recognize different questions. Moreover, it requires the corresponding responses from the developers. However, it cannot handle questions with multiple intents. For example, the question "Do you drink or smoke?" is a question involving multiple intents (i.e., drinking habit and smoking habit). However, ChatGPT can handle this kind of situation and reply with appropriate responses.

[4] All the conversations from the user that cannot be identified by any intents will fall into the default fallback intent.



Try it now

USER SAYS COPY CURL
who is your favourite singer?

DEFAULT RESPONSE
What was that?

CONTEXTS RESET CONTEXTS
__system_counters__

INTENT
Default Fallback Intent

ACTION
input.unknown

SENTIMENT
Query Score: 0.0

DIAGNOSTIC INFO

Bed 1: Bennie [redacted]

How are you?

I'm not doing well, Doc. I've been having some really bothersome symptoms lately. The abdominal pain, frequent bowel movements, and blood in my stool have got me pretty worried. It's been going on for a while now, and I'm just not feeling like myself anymore. So, to answer your question, I'm definitely not doing great.

Who is your favourite singer?

I apologize, but I don't think that my favorite singer has any relevance to the symptoms I'm experiencing. I would prefer to focus on my current health issues right now.

Message

Fig. 13 The handling of irrelevant questions: Dialogflow (left) and ChatGPT (right)

Multilingual support. With the use of ChatGPT, it can handle input with different languages. Although we set our prompts in English in the current implementation, we find that our application can successfully respond to questions in different languages (with Unicode encoding). It is impossible for [17] which uses Dialogflow.

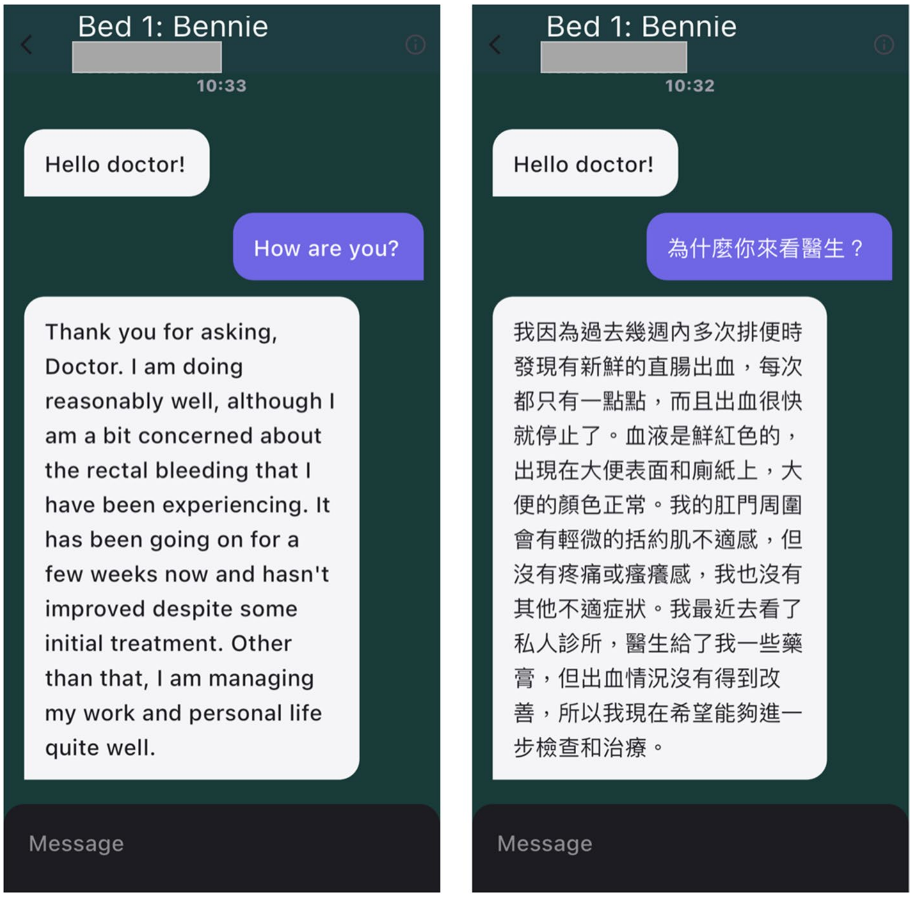
The support of multiple languages is important to make our application useful for teaching students around the world. Students can talk to the virtual patient using their local language (as long as it is supported by ChatGPT) instead of English only. The interaction between the student and the virtual patient is closer to the reality. We illustrate an example of a conversation in English and traditional Chinese with a prescribed patient in Fig. 14.

One interesting example is that the spoken language in Hong Kong is mainly Cantonese, with occasional English words within a sentence. Without multilingual support *at the same time*, it is difficult to comprehend the conversation between a physician and a patient. We illustrate an example of a conversation in Cantonese and occasional English words with a prescribed patient in Fig. 15.

Personality characteristics. We present our protocol with ChatGPT and the design of personality for patients. The personality settings bring a strong sense of presence to the students, and they will encounter patients with different personalities even if they re-do the same case study. However, the version of Dialogflow cannot achieve the personality characteristics due to the inflexibility of the protocol.

Education level characteristics. In reality, patients come from different education levels, which may affect the accuracy of descriptions of their symptoms. We provide an illustration of examples in Fig. 16 that the prompt is able to control these characteristics. Similarly, the version of Dialogflow cannot achieve this due to the inflexibility of the protocol.

Data maintenance by non-technical administrators. As aforementioned, our application provides an interface for administrators (especially non-technical administrators such as professors in medical school) to add new cases, and read, modify, and delete existing cases. This functionality provides a large extent of flexibility in the



Bed 1: Bennie

[redacted]

10:33

Hello doctor!

How are you?

Thank you for asking, Doctor. I am doing reasonably well, although I am a bit concerned about the rectal bleeding that I have been experiencing. It has been going on for a few weeks now and hasn't improved despite some initial treatment. Other than that, I am managing my work and personal life quite well.

Bed 1: Bennie

[redacted]

10:32

Hello doctor!

為什麼你來看醫生？

我因為過去幾週內多次排便時發現有新鮮的直腸出血，每次都只有一點點，而且出血很快就停止了。血液是鮮紅色的，出現在大便表面和廁紙上，大便的顏色正常。我的肛門周圍會有輕微的括約肌不適感，但沒有疼痛或瘙癢感，我也沒有其他不適症狀。我最近去看了私人診所，醫生給了我一些藥膏，但出血情況沒有得到改善，所以我現在希望能夠進一步檢查和治療。

Fig. 14 An example of conversation under English (left) and traditional Chinese (right) with a prescribed patient

scalability of the application, i.e. the clinical administrators can manipulate the data at the time they want. Therefore, the practical cases provided to students can be updated easily, and prevent any consequences from technical maintenance such as unavailability of the application during the maintenance. However, technical maintenance is required for modification of the data set in the version of Dialogflow in [17].

Discussion

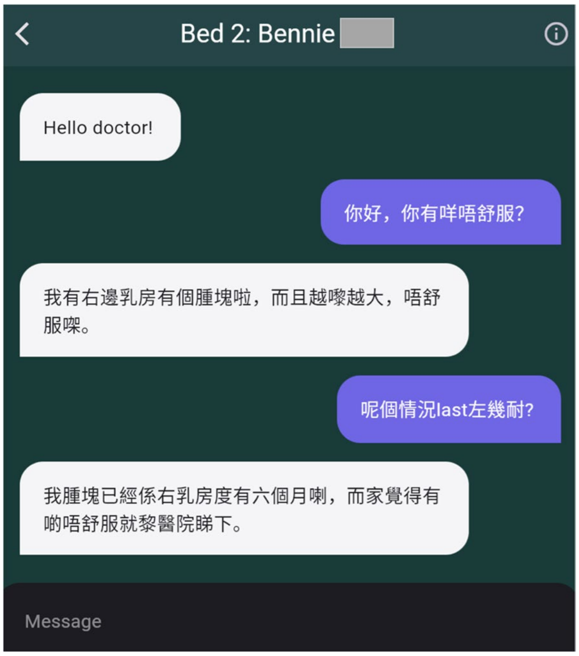
From the results, the chatbot brings a positive applause and feedback about the learning experience from the quantitative perspective. We will discuss the implication for virtual bedside teaching, the limitations of case sharing, the potential challenges and ethical consideration on integrating AI-chatbot into medical training curricula in the following subsections. Moreover, we proposed some possible further development.

Implication for virtual bedside teaching

From the quantitative result, there is a generally positive result among the overall ratings (“Efficacy", “Experience", “Oversea") in Overall evaluation section. We could draw two main conclusions from the questionnaire. (1) our proposed chatbot could enhance the learning efficacy and experience, and (2) the new teaching collaboration enhances the learning outcomes and insights. The results of this study can be seen as an evidence that virtual bedside teaching with chatbots has revolutionized conventional bedside teaching by allowing international collaboration. In this subsection, we discuss the implication on learning efficacy and experience, and the international teaching collaboration.

Learning efficacy and experience

From the data analysis between the ratings of three aspects (“User-friendliness", “Identification", “Interaction") and the ratings of the overall aspects (“Efficacy", “Experience") in Analysis on the ratings of learning efficacy and experience section, it is observed that there is a strong positive relationship between them. It could be concluded that the three aspects are three of the criteria to achieve better learning efficacy and experience.



Bed 2: Bennie [redacted]

Hello doctor!

你好，你有咩唔舒服?

我有右邊乳房有個腫塊啦，而且越嚟越大，唔舒服㗎。

呢個情況last左幾耐?

我腫塊已經係右乳房度有六個月喇，而家覺得有啲唔舒服就黎醫院睇下。

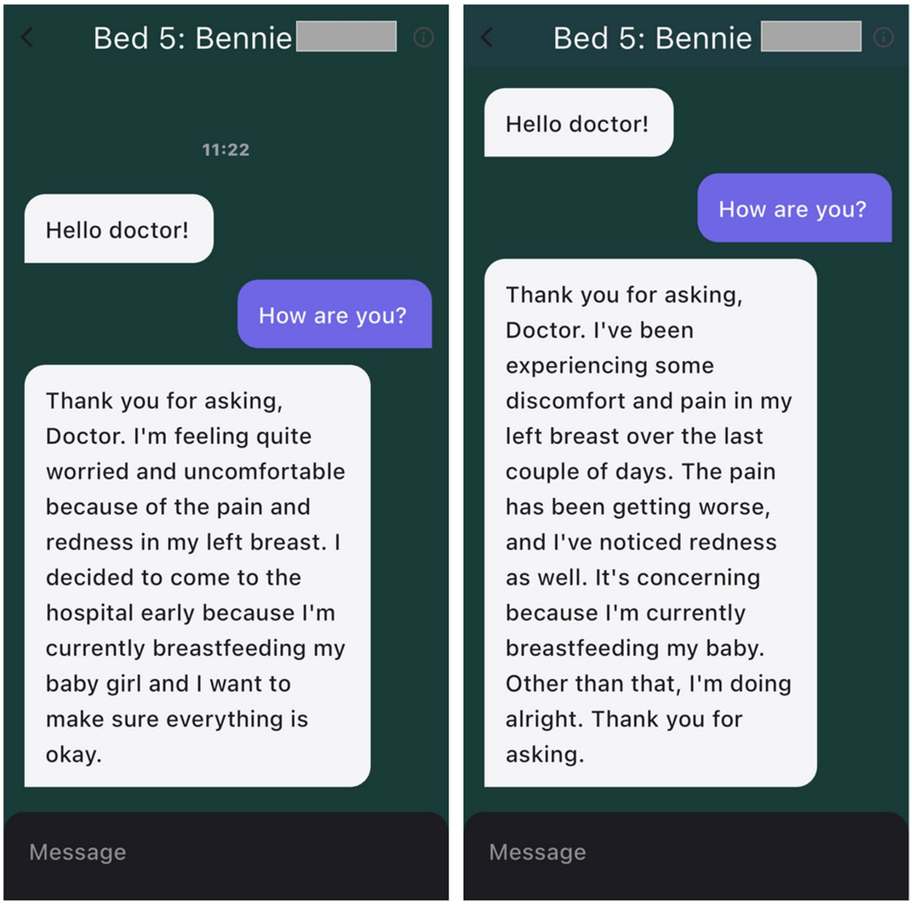
Fig. 15 An example of conversation in Cantonese and occasional English words with a prescribed patient

Relationship between the quantitative and qualitative results. There are several perspectives that the abilities of the chatbot could qualitatively support the quantitative feedback. Our qualitative analysis reveals that the chatbot's ability to effectively address irrelevant and complex questions significantly narrows the scope of conversation to focus solely on medical history-taking, rather than straying into unrelated topics. This focused interaction not only enhances the relevance of the dialogue but also significantly improves learning efficacy. Furthermore, in the context of bedside teaching or practical training, conversations typically center around questions that seek to clarify the patient's situation. If the chatbot can adeptly manage both irrelevant and intricate inquiries, its responses become more accurate and human-like. This responsiveness is crucial, as it develops a more authentic learning environment and promotes deeper understanding. Consequently, the chatbot's proficiency in navigating complex discussions leads to a more effective educational experience, allowing learners to engage meaningfully with the material at hand. On the other hand, instead of presenting straightforward questions in a linear fashion, our new chatbot is designed to provide a more immersive and enriching learning experience.

Moreover, the chatbot incorporates several advanced features that significantly enhance the overall user experience. For instance, the ability to simulate various personalities adds another layer of interactivity, enabling users to experience different perspectives and styles. This can be particularly beneficial in fields such as healthcare, where experiencing and understanding diverse patient interactions is crucial. Additionally, the chatbot is equipped to adjust its responses based on different educational levels from the prompt, ensuring that the content (the dialogues from the chatbot) is tailored to the settings. This adaptability makes the learning experience more similar to the reality as students may interact to patients from different background in the future. With these settings, users can encounter a range of scenarios and responses.

In addition, the multilingual support offered by the chatbot can significantly enhance the learning experience for students in specific locales, such as Hong Kong, where Cantonese predominates, often interspersed with English. Our chatbot with ChatGPT is able to handle the sentences with Cantonese and English *at the same time*, as shown in Fig. 15. This comprehension develops more realistic conversations. By accommodating the linguistic preferences, the chatbot not only promotes inclusivity but also builds a deeper connection to the content, ultimately enhancing educational outcomes. This support demonstrates the potential of the chatbot to tailor learning experiences to meet the unique needs of diverse student populations, and thus enhance learning efficacy.

Advantage of the online platform. Furthermore, the implementation of a chatbot plays a pivotal role in fostering an effective learning environment for bedside history-taking practice. Our proposed chatbot offers students the flexibility to engage in practice sessions at any time and from any location. This flexibility enables learners to initiate the practice sessions according to their individual schedules and preferences, eliminating the constraints associated with traditional in-person practice sessions. By leveraging the capabilities of the chatbot, students gain the advantage of practicing bedside history-taking skills at their convenience. They are no longer bound by the limitations of physical availability or the need for direct interaction with patients. Instead, the chatbot serves as a virtual platform that simulates real-world scenarios, allowing students to engage in authentic and meaningful practice sessions regardless of their geographical location or time constraints. In summary, our proposed chatbot offers a versatile and accessible platform for bedside history-taking practice, allowing students to engage in practice sessions at their convenience. This innovative approach enhances the learning experience, providing students with the flexibility to develop and refine their skills in a self-directed manner.



Bed 5: Bennie [redacted]

Hello doctor!

How are you?

Thank you for asking, Doctor. I'm feeling quite worried and uncomfortable because of the pain and redness in my left breast. I decided to come to the hospital early because I'm currently breastfeeding my baby girl and I want to make sure everything is okay.

Bed 5: Bennie [redacted]

Hello doctor!

How are you?

Thank you for asking, Doctor. I've been experiencing some discomfort and pain in my left breast over the last couple of days. The pain has been getting worse, and I've noticed redness as well. It's concerning because I'm currently breastfeeding my baby. Other than that, I'm doing alright. Thank you for asking.

Fig. 16 An example of education level characteristics: secondary school level (left) and university level (right)

Oversea teaching collaboration

Apart from the enhancement of learning efficacy and experience, there is a generally positive result that the new teaching collaboration enhances the learning outcomes and insights. This kind of international collaboration and joint university bedside teaching also allows students to see virtual patients with diseases that may not be prevalent in their own locality. We will discuss the details in Sharing of international and rare cases and its limitation section.

Relationship between the quantitative and qualitative results. The ability of multilingual support allows users from diverse linguistic backgrounds to interact with the chatbot in their preferred language. It significantly enhances its contribution to international teaching by enabling users from diverse linguistic backgrounds to engage with the system in their preferred language. This feature not only broadens accessibility but also develops inclusivity in the learning environment.

Future role of the chatbot

During COVID-19, one of top-priority concern among the education industry is how to provide good learning materials for the subject. It is also a significant problem among bedside teaching since it involves fact-to-face communications. The proposed application can be a choice of substitution if there are any unpredictable situations happen again such as lockdown. However, we emphasis that face-to-face communication is far important, and thus it cannot be cancelled. As the epidemic slows down, our application can become an *additional platform* to support the learning. Moreover, as aforementioned, we can gather more rare and special cases from local to international. Students are still beneficial from the application even the epidemic is gone.

Sharing of international and rare cases and its limitation

In our proposed chatbot, it is easy to enlarge the set of virtual patients, and it is able to contribute on the international collaboration. Moreover, we can add on some rare cases that students may not able to encounter in the normal training.

International sharing. The idea of chatbot can facilitate international sharing among studying diseases that are more common in one country but not in the other country, such as dengue fever which is more popular in Singapore, and nasopharyngeal carcinoma is more popular in Hong Kong. Our application can provide a long-term history sharing and broaden horizons of students.

Moreover, the favorable quantitative results obtained from this study, including positive applause and feedback, indicate the potential for long-term usage and development of the chatbot in the future. These encouraging outcomes provide a strong foundation for further exploration and refinement of the chatbot as an educational tool. In this study, the collaboration between HKU and NUS provides 13 virtual patients. While this collaboration demonstrates the feasibility and effectiveness of generating virtual patient cases, it also highlights the potential for expanding the database through enhanced collaboration among diverse medical schools worldwide. By fostering increased collaboration and knowledge sharing among different medical schools, the collective database of virtual patients can be significantly enlarged. This expansion would provide students from various educational institutions around the world with a broader exposure to diverse clinical scenarios and enrich their learning experiences. Access to a wider range of cases enhances students' ability to develop comprehensive clinical reasoning skills, fostering a more robust and globally relevant medical education.

However, the sharing of cases internationally may encounter limitations such as the availability and accessibility of shared cases can vary. Some institutions or educators may have limited resources or restrictions on sharing cases due to institutional policies or legal constraints. Thus, the pool of patients may become unbalance of locality. Communication and cooperation may resolve the restrictions but it depends on the actual scenarios.

Rare cases sharing. We consider that some medical cases are rare such as water allergy, some urgent cases, and some international cases may not be able to let students study with bedside teaching. For rare cases, we cannot ask all the medical students to attend since the patient needs to have a good rest, and the patient will be discharged when she is getting better. Moreover, for urgent cases such as appendicitis, in most situations, the patient will be sent to the operation room directly and it is impossible to let students study with bedside teaching. When the students can reach the patient, the situation becomes stable.

When it comes to sharing rare medical cases, addressing privacy concerns becomes more complex compared to sharing common or localized cases. The limited pool of patients with rare conditions poses challenges in maintaining anonymity while providing valuable medical information. In such cases, additional precautions are necessary to protect patient privacy.

Sharing rare cases may involve a smaller number of patients who have experienced similar conditions. This makes it easier to identify individuals based on their unique circumstances. Even if personal identifiers are removed, the combination of rare medical details and additional information, such as family history or minor medical history, could potentially narrow down the choice of the patient's actual identity. Therefore, a well-designed approach is required to ensure privacy is maintained while sharing these cases. To enhance privacy, it may be necessary to encapsulate or generalize certain aspects of the case. Minor details or specific patient information can be modified or omitted to further protect patient identities. By carefully considering the level of detail provided and anonymizing any identifying information, the privacy of rare case patients can be safeguarded while still allowing for the sharing of valuable medical knowledge.

Moreover, obtaining informed consent from patients is crucial in sharing rare cases. Patients should be fully informed about the purpose and potential risks of sharing their medical information. The risk of disclosing their identity during the virtual bedside teaching should be clearly stated. They need to understand how their data will be used and the measures in place to protect their privacy. Consent should be obtained prior to sharing any case information, ensuring that patients have the opportunity to make an informed decision about their participation in knowledge sharing initiatives. Moreover, the level of detail provided or encapsulation can be discussed with the patients accordingly.

Comparison between two versions of Bennie and the chats

We illustrate the comparisons between the findings we received in this research and the findings from the research on the previous version of chatbot [17]. Note that only the possibility of enhancing the learning efficacy and experience is considered in [17].

We first recap the findings from [17] and this paper in Table 8, showing the descriptive statistics of the ratings

Table 8 Students' evaluation from this work and [17]

Areas of evaluation	# Respondents		Range		Median	
	This	[17]	This	[17]	This	[17]
User-friendliness	44 (45)	132	0–10 (0–10)	6–10	7 (7)	8
Identification	45	132	0–10	5–9	7	7
Interaction	45	132	0–10	6–9	7	7
Efficacy	45	132	0–10	6–9	7	8
Experience	45	132	0–10	6–9	8	8

on "User-friendliness", "Identification", "Interaction" among "Efficacy", altogether with the overall ratings on "Efficacy"[5] and "Experience". The Likert scale used in [17] is 1–10, while we use 0–10 in this work. The group of respondents between the two research has no overlap. Note that for the area of "User-friendliness", we state the values in brackets with the assumption that the "No Response" rating is 0 under the Likert scale, as aforementioned. In [17], there were 132 (62 female and 70 male) students from HKU and their final year of undergraduate medical curriculum.

The ratings from the two research findings appear to be similar. While the categories of "User-friendliness" and "Efficacy" received lower scores in this study, the overall results indicate a positive impact of the new chatbot on enhancing learning efficacy and experience. We list two perspectives that may explain the received findings.

Better Perception of using AI. ChatGPT was launched in November 2022. Prior to this release, public awareness and engagement with AI were relatively limited. Consequently, the demand for students to utilize AI applications, as identified in the study by [17], may have been lower, given that the research was conducted in 2021. In contrast, the bedside teaching this research was launched in July 2023. Following a period of post-release, students have become increasingly adept at using AI technologies, including ChatGPT, to accomplish specific objectives. For instance, universities in Hong Kong permitted students to utilize ChatGPT throughout 2023 [90]. This development has led to greater access to ChatGPT, allowing students to explore and experience the capabilities of this powerful AI tool. Furthermore, the integration of AI and ChatGPT into daily public use has expanded significantly, such as Amazon recommendations [91] and Tesla autonomous driving [92]. As a result, the expectations for students in this research to operate an AI application are likely much higher than those observed in previous studies.

Dominant Usage and supplementary usage. In 2021, the study conducted by [17] introduced the first generation of a chatbot designed to facilitate bedside teaching, particularly in learning history-taking skills. This innovation arosed in response to the disruptions caused by the COVID-19 pandemic, which stops traditional face-to-face interactions between educators and students. Consequently, the chatbot emerged as a vital educational resource, serving as the dominant means of instruction during a time when in-person bedside teaching was not feasible. As conditions evolved with the gradual resolution of the pandemic, the role of the chatbot transitioned. The new generation of the chatbot is now aimed as a supplementary learning tool, complementing rather than replacing face-to-face bedside teaching, which remains the predominant pedagogical approach. Therefore, the differences on the aim of the two chatbots may lead to the differences on the ratings, since students in this research may consider and compare the efficacy and experience provided not only by the chatbot, but also by the face-to-face teaching.

Cost-effectiveness

In this section, we discuss the cost-effectiveness by using the chatbot and the typical in-person bedside teaching.

Pricing of ChatGPT. We first consider about the price of ChatGPT. Generally, ChatGPT converts each word into a legible token when you send it a question. One token generally corresponds to about 4 text characters for common English text. This roughly translates to 100 tokens for about 75 English words.

As of January 2025, ChatGPT-3.5 Turbo charges US$0.0015 per 1000 tokens for input and US$0.002 per

[5] It is stated as "Efficiency of learning" in [17].

1000 tokens for output [93][6]. Sending the aforementioned sample prompt with prefix and suffix and getting one response from ChatGPT uses around 365 tokens in ChatGPT. A full conversion is usually below 3000 tokens, which costs US$0.006. Hence, our application has a low operating cost.

With a high-volume educational setting, the cost of using ChatGPT will become higher since there are more users and thus more interactions with the ChatGPT.

Comparison on using the chatbot and the typical bedside teaching. When comparing the costs between our chatbot and the typical bedside teaching, several factors come into play that can influence the overall expenses associated with each option. We elaborate and explain some of them in the following.

First, we consider the setup cost. Setting up a ChatGPT chatbot application involves the time cost of designing and developing the application, and configuring the necessary infrastructure. Moreover, it takes time to design and test the prompt of each case of patient. On the other hand, establishing bedside teaching with a patient requires various administrative costs such as acquiring the consent of each individual patient, and time arrangement between the patient and the student, which can involve significant upfront costs.

Next, we consider the operational costs. For a ChatGPT application, the cost typically includes ongoing maintenance to the server and records, server hosting fees, and fees for each conversation with ChatGPT. On the other hand, traditional bedside teaching does not require any cost of hiring any patients since it is voluntary.

Third, we consider the scalability costs. Scaling a ChatGPT application to accommodate increased usage may involve higher computational costs for additional server resources. For instance, it includes purchasing servers with more CPUs and memory. Moreover, it may include a more advanced system design to accommodate a large amount of usage. On the other hand, establishing bedside teaching with a patient requires various administrative costs such as handling numerous voluntary patients.

Fourth, we consider the costs on system maintenance and support. Maintenance and support costs for the chatbot application involve performance monitoring, addressing technical issues, and user assistance. Moreover, it is require to observe the output of the chatbot in order to prevent false information to affect the learning outcome. However, there are no cost on maintenance and support requirement from the typical bedside teaching.

Lastly, we consider the availability cost. For out chatbot, the virtual patients are always available except for the system maintenance time. However, the availability of patients is limited. We are not able to ask all the medical students to attend since the patient needs to have a good rest, and the patient will be discharged when she is getting better.

To conclude, both the chatbot and typical bedside teaching require various costs to operate. Our chatbot provides advantages such as long-term availability of each medical case, while the typical bedside teaching provides in-person conversation which the chatbot is not able to fully simulate. Given that medical treatment and consultations predominantly occur face-to-face, it is logical to maintain both practices, with traditional bedside teaching remaining prevalent. Nevertheless, the significance of the chatbot persists as they offer simulated training opportunities that are accessible at any time and from any location.

Potential Challenges and Ethical Considerations in Integrating AI-driven Chatbot into Medical Training Curricula

Integrating AI-driven chatbots into medical training curricula brings numerous opportunities for enhancing learning experiences and improving medical education. However, it is crucial to address potential challenges that may arise during this integration process. Since the chatbot should only become a supplementary tool for students to practice history-taking skills, the consideration mostly on the ethical issue since it is related to information of patients and students. We discuss five main concerns in this subsection, combining previous research and the observation during the implementation.

Reliability and accuracy. The first challenge is the reliability and accuracy [66]. AI-driven chatbots must demonstrate high reliability and accuracy to ensure that the information provided is trustworthy and aligns with current medical knowledge. Medical education heavily relies on accurate and evidence-based information, and any inaccuracies or errors in the chatbot's responses can have significant consequences for student learning and patient care. Ensuring that chatbot algorithms are regularly updated and validated against established clinical guidelines and best practices is essential. Additionally, addressing potential biases in the data used to train the chatbot and mitigating the risk of algorithmic errors are critical considerations.

[6] We adopted the "gpt-35-turbo" model during the development stage, it is the same as "gpt-3.5-turbo-0301" or "gpt-3.5-turbo-0613" with context "4K" in [93]. The suffix "-0301" or "-0613" indicate that the date refers to a snapshot [94]. The snapshots enable users to preserve a stable state of the model for querying purposes, thereby ensuring that the output remains consistent.

Several considerations are essential to address the challenges associated with reliability and accuracy. First, system admin should apply error monitoring and feedback mechanisms. Continuous monitoring of the chatbot's performance is necessary to identify and rectify any errors or inaccuracies promptly. Implementing feedback mechanisms that allow users, such as students or educators, to report potential inaccuracies or ambiguities can help improve the chatbot's performance over time. This feedback loop facilitates ongoing refinement and fine-tuning of the chatbot's algorithms to enhance its reliability and accuracy.

Second, system developer should address biases and limitations. AI algorithms can be susceptible to biases present in the training data, leading to skewed or inaccurate responses. Efforts should be made to identify and address these biases, ensuring that the chatbot provides unbiased and equitable information to all users. Additionally, acknowledging the limitations of the chatbot, such as its inability to handle complex, context-dependent scenarios, is crucial to manage user expectations and prevent potential errors.

Third, teachers should also launch user education [95]. Educating students about the capabilities and limitations of the chatbot is important for promoting critical thinking and ensuring they understand that the chatbot's responses are not a substitute for clinical judgment. Emphasizing the importance of corroborating information from multiple sources and consulting with healthcare professionals when necessary can help avoid over-reliance on the chatbot and encourage a balanced approach to medical decision-making.

Limitations on the reliability of ChatGPT responses on nuanced clinical reasoning. The reliability of ChatGPT responses, particularly in scenarios involving nuanced clinical reasoning, is a critical consideration. While ChatGPT can provide simulation, there are several limitations to take care of when assessing its reliability for complex medical contexts. We consider two of the perspectives.

First, we consider the prompts with specialized knowledge. The accuracy and reliability of ChatGPT responses heavily depend on the quality and diversity of the prompt it has been exposed to. In the case of medical information, the breadth and depth of the prompts play a crucial role in the accuracy of responses. Moreover, nuanced clinical reasoning often requires specialized medical knowledge. Responses from ChatGPT may lack the depth of understanding required for complex medical scenarios. Therefore, the design and writing of the prompt for each case should provide exhaustive information about the case in order to generate more accurate responses [67].

Second, we consider the contextual understanding and the risk of misinterpretation [66, 67]. Clinical reasoning involves understanding complex patient histories, symptoms, test results, and treatment options within a specific context. ChatGPT may struggle to grasp the intricate nuances and context-specific details essential for accurate clinical advice. Moreover, ChatGPT may generate responses based on statistical patterns in data without a full understanding of the underlying medical concepts. This can lead to inaccuracies or misinterpretations, especially in critical medical situations. This may be alleviated by a better design of prompts, and testing the prompt by interacting with ChatGPT.

While ChatGPT can be a valuable tool for providing information and support in medical settings, its reliability for scenarios requiring nuanced clinical reasoning is subject to limitations. Human expertise and validation of the output remain crucial to ensure the accuracy and appropriateness of responses, especially in complex medical contexts where precise clinical judgment is essential.

Ethical and privacy concerns. The second challenge is to reduce the ethical and privacy concerns. AI-driven chatbots have access to sensitive patient data, which raises ethical and privacy concerns. Implementing robust data protection measures, ensuring compliance with privacy regulations, and obtaining informed consent for the use of patient data are essential. Transparency about data usage and addressing potential biases in data collection and algorithms are important to maintain patient trust and confidentiality.

Apart from following the regulations such as the Health Insurance Portability and Accountability Act (HIPAA) [86, 87] or the General Data Protection Regulation (GDPR) [88] to ensure data protection and security, we can apply the following examples of procedure to reduce the concerns. First, the patients should be informed for the consent [68]. When patient data is used in the development or training of the chatbot, obtaining informed consent from patients is essential. Patients should be fully informed about the purpose, risks, and benefits of their data being used and have the opportunity to provide or withhold consent. Transparency in explaining how their data will be used, who will have access to it, and the privacy protections in place is crucial to respect patient autonomy and maintain trust. This transparency helps users make informed decisions about their involvement with the chatbot and understand the privacy implications. Additionally, they should have the option to opt-out of data collection if they do not wish to participate or share their data. Respecting user choices and preferences is essential in upholding privacy and autonomy.

Second, the chatbot should be under monitored continuously. Ethical and privacy considerations should be addressed as an ongoing process rather than a one-time effort. Regular monitoring, feedback collection, and user engagement can help identify emerging ethical challenges, adapt to changing regulations, and implement improvements to enhance privacy protections and ethical standards.

Privacy of Students. As aforementioned, teachers should also launch user education about critically distinguish the chatbot's responses to ensure an appropriate outcome of learning with chatbot. Additionally, the education should also provide guidelines to students [95] on the related technical and ethical issues, such as do not provide personal information during the conversation with the virtual patients. For further development, the conversation records may be saved for educational analysis or monitoring. Therefore, students should take the correct approach to chat with the virtual patients without any personal items. Additionally, informed consent should be obtained from users, clearly explaining the purpose and scope of data collection. Further research such as [69] concentrates on providing recommendations to address concerns related to student data security.

Attitude when using AI learning tools. There are many tools are proposed and developed with the use of AI to enhance the quality of learning experiences. It is important that the users take appropriate attitude when using them. The power of AI may bring us opportunities to improve if we adopt it appropriately [96, 97].

However, we should not always expect everyone will honestly follow the system, thus developers should try their best to prevent any inappropriate usages or even hacking. Therefore, every parties should adopt a responsible and mindful attitude.

Moreover, the users should avoid excessive dependence on AI tools [65]. An individual and critical thinking mindset is significant for everyone, and students can always learn from thinking different problems, such as studying different medical cases in our application. Out of the online learning, the engagement in discussions are beneficial for students. They can also discuss the cases or questions from the AI learning tool. It is advantageous if they seek diverse perspectives from participating in meaningful interactions with peers and teachers.

Scalability

Scalability is another critical concern when considering the application of ChatGPT. As the demand for our chatbot grows, ensuring that it can handle increasing workloads efficiently becomes paramount. In this section, we discuss two significant approaches that we have considered during the implementation.

Handling real-time and concurrent user interactions. One scalability challenge lies in the computational resources required to support a large number of real-time and concurrent user interactions. As the user base expands, the server that stores the cases of patients must be able to handle a high volume of requests without significant latency or performance degradation. From the perspective of our server, the issue could be handled by balancing CPU, memory and network resources, and it becomes more complex as the demand on the servers escalates. As discussed in Implementation section, we adopt Azure virtual machine as the server in our application, which provides a number of choices to manipulate the needs of computational resources. From the perspective of calling the ChatGPT API, the scalability issue moves to the ChatGPT server, but not in our application. To manage the ChatGPT availability and status, it is essential to handle and identify all the error messages returned from the ChatGPT API. Two of the possible error messages from the ChatGPT API while running the application are service overloaded and running out of user credits.

Moreover, another possible approach is to consider the balance of load. Load balancing plays a crucial role in ensuring that incoming requests are evenly distributed across multiple servers, i.e. multiple Azure virtual machines. Implementing effective load-balancing mechanisms becomes essential to prevent any single server from becoming overwhelmed while maintaining consistent response times for users.

Storing the dialogue. Additionally, the storage and retrieval of conversational data can become a bottleneck as the dataset grows, impacting response times.

We note that in this chatbot version, we do not store any chat history between the user and the ChatGPT. As the application collects and processes a larger volume of conversational data, the storage systems must scale to accommodate this growth. During our development, we considered this and we adopted some methods to handle

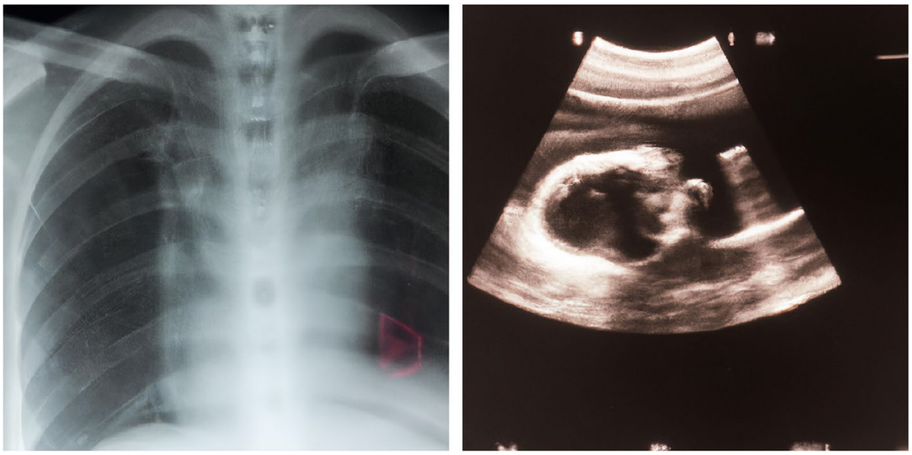


Fig. 17 Examples of AI medical images generated by DALL·E 2

this possible scalability issue. Our chatbot works similarly to some existing communication applications such as Whatsapp [98] and Signal [99], the entire chat record is stored in the devices of users. However, in the current version, in order to reduce the burden of storage from the devices of users, the conversation is only stored during the conversation and will be deleted once the user leaves the chatroom.

In addition, we note that it is a possible extension if the study of dialogue is required. In this case, only storing the chat history in local devices is not enough, the server should come up with some procedures for storing the chat records. Moreover, the chat history should be regulated under the GDPR [88] when necessary. To achieve this, it is required to design the storage architecture to be scalable, allowing for easy expansion as data volume grows. Moreover, it is possible to distribute data across multiple servers for horizontal scalability. Lastly, it is essential to set up regular automated backups to prevent data loss in case of system failures or accidental deletions.

Extension towards other disciplines

The idea of this application for education can be extended to other domains as a supplement. For example, it can be a chatbot for pharmacy students on discussing drug usage of some preset medical cases. Moreover, it can be extended to train the questioning skills of lawyers and police, and train the interviewing skills of news reporters.

Apart from learning materials, the idea of our application can become a tool of publicity. For example, animal-caring authorities can use the chatbot which imitates some abused pets, to educate the public on caring animals. Moreover, we cannot talk to an animal in reality, thus this become a interesting way for public to experience the “feeling" of the abused animals.

Further development

Illustration among medical files or images. In our work and [17], they provide only textual communication in the role-playing game. We identify that, in some cases, patients may be able to provide some historical clinical exam and medical images such as blood tests, MRI, CT scan, and x-ray. Free open-access online databases of medical images are available. For example, MedPix[7] provides 59,000 images for 12,000 patients.

In future versions of the application, it is possible to add more functions during the conversation. For example, if any medical checks or exams are required by the physician (the student) in the conversation, the ChatGPT is able to reply with relevant information from the prescribed data in the server. These medical images may even be generated by the latest image-generative AI technology, as shown in Fig. 17. The left image is generated with “x-ray of a patient with pulmonary tuberculosis" and the right image is generated with “ultrasound image of a 7-month pregnant baby". However, the accuracy of software like Midjourney, Stable Diffusion, or DALL·E 2 for generating medical images should be tested before use. The use of image-generative AI for STEAM education was recently studied in [100].

[7] https://medpix.nlm.nih.gov/home

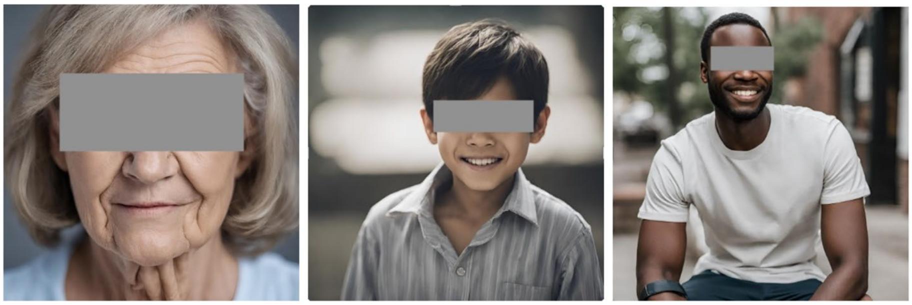


Fig. 18 An example of AI-generated photos of patients. The eyes of the patients are intentionally concealed

Voice conversation with different languages and accents. In a recent update of ChatGPT [101], the launch of the support among voice capabilities may further make the application in a new generation. In typical voice recognition function, it could be done by most of the smart mobile phones in the market. However, although it can support conversations, it is presented with voice-to-text translation, and reading the text with a prescribed voice and tone such as Siri, Alexa, or Google Assistant. From the update, we foresee the possibility that the ChatGPT could understand languages in voice and can respond in voice with different accents supported. Moreover, the voice can be controlled by the age and gender of the patients.

Improved UI/UX. Apart from generating the voice of the patients, it is also possible to further improve the user interface/user experience (UI/UX) by generating photos or even videos of the patient. By using generative AI such as Midjourney, Stable Diffusion, or DALL·E 2, we can now generate photos of patients according to their age, gender, ethnicity, weight, etc., as shown in Fig. 18.

If the disease of the patient affects his appearance (e.g., hand, foot, and mouth disease which causes skin rash in young children), it may also be reflected in the photo or video. Our preliminary testing finds that many skin problems (e.g., skin rash, acne) are currently restricted by DALL·E 2 due to security protection (to prevent the generation of abusive pictures). However, we can generate photos of other diseases such as conjunctivitis.

Study of dialogue. In our system architecture, the web server stands between the user and the patient (ChatGPT). Hence, the web server can store all dialogues for future study on the student's learning progress. By using machine learning technologies, we can gather information such as the duration of the conversation, the relevancy of questions asked, the effectiveness of the diagnosis, etc. We can gain valuable knowledge about the student's learning progress. This information can inform educational strategies, identify areas that require further attention, and contribute to the continuous improvement of the chatbot's performance.

It is important to emphasize that the collection and analysis of such data should be conducted with strict adherence to ethical guidelines and privacy regulations. Informed consent should be obtained from users, clearly explaining the purpose and scope of data collection. Additionally, robust security measures should be in place to safeguard the stored dialogues and protect sensitive information from unauthorized access.

Conclusion

Virtual bedside teaching with chatbots has revolutionized conventional bedside teaching by its advantages and allowing international collaborations. We believe that the training of history taking skills by chatbot will be a feasible supplementary teaching tool to conventional bedside teaching.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-025-06785-9.

Supplementary Material 1.

Acknowledgements

Not applicable.

Authors' contributions

KYC, THY, and MC conceived the study idea, designed the study. KYC worked on the creation of new software used. THY and MC sampled the participants. KYC, THY, and MC worked on the acquisition of data. All authors reviewed, analyzed, interpreted, wrote, critically revised, and approved the final version.

Funding

The authors have not declared a specific grant for this research from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability

Uploaded as "data_bmc_2.xlsx" in the "Related files" component.

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (HKU/HA HKW IRB) on May 25, 2022. The IRB Reference Number is UW 22-268. Written informed consent was obtained from all the participants. The questionnaires were anonymized, and patients were free to opt out of participation in the study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Computer Science, University of Hong Kong, Pokfulam, Hong Kong. [2]Department of Software Systems & Cybersecurity, Monash University, Clayton, Australia. [3]Centre for Education and Training, Department of Surgery, Queen Mary Hospital, University of Hong Kong, Pokfulam, Hong Kong.

Received: 30 July 2024 Accepted: 29 January 2025
Published online: 07 February 2025

References

1. McCoy LG, Ci Ng FY, Sauer CM, Yap Legaspi KE, Jain B, Gallifant J, et al. Understanding and training for the impact of large language models and artificial intelligence in healthcare practice: A narrative review. BMC Med Educ. 2024;24(1):1096.
2. A Fuller K, Morbitzer KA, Zeeman JM, M Persky A, C Savage A, McLaughlin JE. Exploring the use of ChatGPT to analyze student course evaluation comments. BMC Med Educ. 2024;24(1):423.
3. Bae J, Lee J, Choi M, Jang Y, Park CG, Lee YJ. Development of the clinical reasoning competency scale for nurses. BMC Nurs. 2023;22(1):138.
4. Ng IK, Goh WG, Teo DB, Chong KM, Tan LF, Teoh CM. Clinical reasoning in real-world practice: a primer for medical trainees and practitioners. Postgrad Med J. 2025;101(1191):68–75.
5. Ruczynski LI, van de Pol MH, Schouwenberg BJ, Laan RF, Fluit CR. Learning clinical reasoning in the workplace: a student perspective. BMC Med Educ. 2022;22(1):19.
6. Koufidis C, Manninen K, Nieminen J, Wohlin M, Silén C. Grounding judgement in context: A conceptual learning model of clinical reasoning. Med Educ. 2020;54(11):1019–28.
7. Restini C, Faner M, Miglio M, Bazzi L, Singhal N. Impact of COVID-19 on Medical Education: A Narrative Review of Reports from Selected Countries. J Med Educ Curricular Dev. 2023;10:23821205231218120.
8. Law VT, Yee HH, Ng TK, Fong BY. Transition from traditional to online learning in Hong Kong tertiary educational institutions during COVID-19 pandemic. Technol Knowl Learn. 2023;28(3):1425–41.
9. Tsang JT, So MK, Chong AC, Lam BS, Chu AM. Higher education during the pandemic: The predictive factors of learning effectiveness in COVID-19 online learning. Educ Sci. 2021;11(8):446.
10. Tang YM, Chen PC, Law KM, Wu CH, Lau YY, Guan J, et al. Comparative analysis of Student's live online learning readiness during the coronavirus (COVID-19) pandemic in the higher education sector. Comput Educ. 2021;168:104211.
11. Yeung MW, Yau AH. A thematic analysis of higher education students' perceptions of online learning in Hong Kong under COVID-19: Challenges, strategies and support. Educ Inf Technol. 2022;27(1):181–208.
12. Alam M, Al-Mamun M, Pramanik MNH, Jahan I, Khan MR, Dishi TT, et al. Paradigm shifting of education system during COVID-19 pandemic: A qualitative study on education components. Heliyon. 2022;8(12):e11927.
13. Rasul T, Nair S, Kalendra D, Robin M, de Oliveira Santini F, Ladeira WJ, et al. The role of ChatGPT in higher education: Benefits, challenges, and future research directions. J Appl Learn Teach. 2023;6(1):41–56.
14. Shomotova A, Karabchuk T. Transition to Online Teaching Under COVID-19: The Case Study of UAE University. In: Social Change in the Gulf Region: Multidisciplinary Perspectives. Singapore: Springer Nature Singapore; 2023. pp. 141–60.
15. Mhlanga D. Digital transformation of education, the limitations and prospects of introducing the fourth industrial revolution asynchronous online learning in emerging markets. Discover Educ. 2024;3(1):32.
16. Dhawan S. Online Learning: A Panacea in the Time of COVID-19 Crisis. J Educ Technol Syst. 2020;49(1):5–22. https://doi.org/10.1177/0047239520934018.
17. Co M, Yuen TH, Cheung HH. Using clinical history taking chatbot mobile app for clinical bedside teachings - a prospective case control study. Heliyon. 2022;8(6). https://doi.org/10.1016/j.heliyon.2022.e09751.
18. Findlater A, Bogoch II. Human mobility and the global spread of infectious diseases: a focus on air travel. Trends Parasitol. 2018;34(9):772–83.
19. Gouveia N, Ayres JRDCM. Global Health in the medical curriculum. Clinics. 2021;76:e3073.
20. Armstrong RW, Mantel M, Walraven G, Atwoli L, Ngugi AK. Medical education and population health-A framework in the design of a new undergraduate program. Front Public Health. 2022;2010:1068092.
21. Davey AK. Internationalisation of the curriculum in health programs. BMC Med Educ. 2023;23(1):285.
22. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI Blog. 2018;1–12.
23. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019. https://arxiv.org/abs/1810.04805. Accessed 21 Jan 2025.
24. Yan D. Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. Educ Inf Technol. 2023;28:1–25.
25. Tian H, Lu W, Li TO, Tang X, Cheung SC, Klein J, et al. Is ChatGPT the Ultimate Programming Assistant – How far is it? 2023. https://arxiv.org/abs/2304.11938. Accessed 21 Jan 2025.
26. Rohani N, Gal K, Gallagher M, Manataki A. Providing insights into health data science education through artificial intelligence. BMC Med Educ. 2024;24(1):564.
27. OpenAI. ChatGPT. 2021. https://openai.com. Accessed 17 Jan 2025.
28. Swan K. Learning effectiveness: What the research tells us. Elem Qual Online Educ Pract Direction. 2003;4:13–47.
29. Ni AY. Comparing the Effectiveness of Classroom and Online Learning: Teaching Research Methods. J Public Aff Educ. 2013;19(2):199–215. https://doi.org/10.1080/15236803.2013.12001730.
30. Nguyen T. The effectiveness of online learning: Beyond no significant difference and future horizons. MERLOT J Online Learn Teach. 2015;11(2):309–19.
31. Kizilcec RF, Chen M. Student Engagement in Mobile Learning via Text Message. In: Proceedings of the Seventh ACM Conference on Learning @ Scale. L@S '20. ACM; 2020. pp. 157–66. https://doi.org/10.1145/3386527.3405921.
32. Loh CYR, Teo TC. Understanding Asian students learning styles, cultural influence and learning strategies. J Educ Soc Policy. 2017;7(1):194–210.
33. Chi MTH, Siler SA, Jeong H, Yamauchi T, Hausmann RG. Learning from human tutoring. Cogn Sci. 2001;25(4):471–533. https://doi.org/10.1016/S0364-0213(01)00044-1.
34. Morris J, Chi MTH. Improving teacher questioning in science using ICAP theory. J Educ Res. 2020;113(1):1–12. https://doi.org/10.1080/00220671.2019.1709401.
35. Price L, Richardson JTE, Jelfs A. Face-to-face versus online tutoring support in distance education. Stud High Educ. 2007;32(1):1–20. https://doi.org/10.1080/03075070601004366.
36. Chappell S, Arnold P, Nunnery J, Grant M. An Examination of an Online Tutoring Program's Impact on Low-Achieving Middle School Students' Mathematics Achievement. Online Learn. 2015;19(5):37–53.
37. Co M, Chu KM. Distant surgical teaching during COVID-19 - A pilot study on final year medical students. Surg Pract. 2020;24(3):105–9.
38. Co M, Chung PHY, Chu KM. Online teaching of basic surgical skills to medical students during the COVID-19 pandemic: a case-control study. Surg Today. 2021;51:1404–9.
39. Alsoufi A, Alsuyihili A, Msherghi A, Elhadi A, Atiyah H, Ashini A, et al. Impact of the COVID-19 pandemic on medical education: Medical students' knowledge, attitudes, and practices regarding electronic learning. PLoS ONE. 2020;15(11):e0242905.

40. Liu CH, Lin HYH. The impact of COVID-19 on medical education: experiences from one medical university in Taiwan. J Formos Med Assoc. 2021;120(9):1782–4.
41. Mageira K, Pittou D, Papasalouros A, Kotis K, Zangogianni P, Daradoumis A. Educational AI chatbots for content and language integrated learning. Appl Sci. 2022;12(7):3239.
42. Chen HL, Widarso GV, Sutrisno H. A ChatBot for Learning Chinese: Learning Achievement and Technology Acceptance. J Educ Comput Res. 2020;58(6):1161–89. https://doi.org/10.1177/0735633120929622.
43. Yuan CC, Li CH, Peng CC. Development of mobile interactive courses based on an artificial intelligence chatbot on the communication software LINE. Interact Learn Environ. 2021;31(6):3562–76.
44. Jafri L, Farooqui AJ, Grant J, Omer U, Gale R, Ahmed S, et al. Insights from semi-structured interviews on integrating artificial intelligence in clinical chemistry laboratory practices. BMC Med Educ. 2024;24(1):170.
45. Yin J, Goh TT, Yang B, Xiaobin Y. Conversation Technology With Micro-Learning: The Impact of Chatbot-Based Learning on Students' Learning Motivation and Performance. J Educ Comput Res. 2021;59(1):154–77. https://doi.org/10.1177/0735633120952067.
46. Ait Baha T, El Hajji M, Es-Saady Y, Fadili H. The impact of educational chatbot on student learning experience. Educ Inf Technol. 2024;29(8):10153–76.
47. Min W, Park K, Wiggins J, Mott B, Wiebe E, Boyer KE, et al. Predicting dialogue breakdown in conversational pedagogical agents with multimodal LSTMs. In: Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part II 20. Cham: Springer; 2019. pp. 195–200.
48. Song Y, Lei S, Hao T, Lan Z, Ding Y. Automatic classification of semantic content of classroom dialogue. J Educ Comput Res. 2021;59(3):496–521.
49. Azer SA, Guerrero AP. The challenges imposed by artificial intelligence: are we ready in medical education? BMC Med Educ. 2023;23(1):680.
50. Jebreen K, Radwan E, Kammoun-Rebai W, Alattar E, Radwan A, Safi W, et al. Perceptions of undergraduate medical students on artificial intelligence in medicine: mixed-methods survey study from Palestine. BMC Med Educ. 2024;24(1):507.
51. Fitzek S, Choi KEA. Shaping future practices: German-speaking medical and dental students' perceptions of artificial intelligence in healthcare. BMC Med Educ. 2024;24(1):844.
52. Busch F, Hoffmann L, Truhn D, Ortiz-Prado E, Makowski MR, Bressem KK, et al. Global cross-sectional student survey on AI in medical, dental, and veterinary education and practice at 192 faculties. BMC Med Educ. 2024;24(1):1066.
53. Sridharan K, Sequeira RP. Artificial intelligence and medical education: application in classroom instruction and student assessment using a pharmacology & therapeutics case study. BMC Med Educ. 2024;24(1):431.
54. Quah B, Zheng L, Sng TJH, Yong CW, Islam I. Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. BMC Med Educ. 2024;24(1):962.
55. Angkurawaranon S, Inmutto N, Bannangkoon K, Wonghan S, Kham-Ai T, Khumma P, et al. Attitudes and perceptions of Thai medical students regarding artificial intelligence in radiology and medicine. BMC Med Educ. 2024;24(1):1188.
56. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023;23(1):689.
57. Davies NP, Wilson R, Winder MS, Tunster SJ, McVicar K, Thakrar S, et al. ChatGPT sits the DFPH exam: large language model performance and potential to support public health learning. BMC Med Educ. 2024;24(1):57.
58. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ. 2024;24(1):143.
59. Hershberger PJ, Pei Y, Bricker DA, Crawford TN, Shivakumar A, Castle A, et al. Motivational interviewing skills practice enhanced with artificial intelligence: ReadMI. BMC Med Educ. 2024;24(1):237.
60. Yanagita Y, Yokokawa D, Fukuzawa F, Uchida S, Uehara T, Ikusaka M. Expert assessment of ChatGPT's ability to generate illness scripts: an evaluative study. BMC Med Educ. 2024;24(1):536.
61. Laupichler MC, Aster A, Meyerheim M, Raupach T, Mergen M. Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments. BMC Med Educ. 2024;24(1):401.
62. Amiri H, Peiravi S, Rezazadeh Shojaee SS, Rouhparvarzamin M, Nateghi MN, Etemadi MH, et al. Medical, dental, and nursing students' attitudes and knowledge towards artificial intelligence: a systematic review and meta-analysis. BMC Med Educ. 2024;24(1):412.
63. Abou Hashish EA, Alnajjar H. Digital proficiency: assessing knowledge, attitudes, and skills in digital transformation, health literacy, and artificial intelligence among university nursing students. BMC Med Educ. 2024;24(1):508.
64. McLennan S, Meyer A, Schreyer K, Buyx A. German medical students views regarding artificial intelligence in medicine: A cross-sectional survey. PLoS Digit Health. 2022;1(10):e0000114.
65. Zhai C, Wibowo S, Li LD. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. Smart Learn Environ. 2024;11(1):28.
66. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Phys Syst. 2023;3:121–54.
67. Kuang YR, Zou MX, Niu HQ, Zheng BY, Zhang TL, Zheng BW. ChatGPT encounters multiple opportunities and challenges in neurosurgery. Int J Surg. 2023;109(10):2886–91.
68. Park HJ. Patient perspectives on informed consent for medical AI: A web-based experiment. Digit Health. 2024;10:20552076241247936. https://doi.org/10.1177/20552076241247938.
69. Huang L. Ethics of artificial intelligence in education: Student privacy and data protection. Sci Insights Educ Front. 2023;16(2):2577–87.
70. Wang H, Li J, Wu H, Hovy E, Sun Y. Pre-Trained Language Models and Their Applications. Engineering. 2023;25:51–65. https://doi.org/10.1016/j.eng.2022.04.024.
71. Li J, Tang T, Zhao WX, Nie JY, Wen JR. Pre-Trained Language Models for Text Generation: A Survey. ACM Comput Surv. 2024;56(9). https://doi.org/10.1145/3649449.
72. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report. 2024. https://arxiv.org/abs/2303.08774. Accessed 21 Jan 2025.
73. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI Blog. 2019;1(8):9.
74. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.
75. Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. Adv Neural Inf Process Syst. 2017;30:4302–10.
76. Ziegler DM, Stiennon N, Wu J, Brown TB, Radford A, Amodei D, et al. Fine-tuning language models from human preferences. 2019. https://arxiv.org/abs/1909.08593. Accessed 21 Jan 2025.
77. Bai Y, Jones A, Ndousse K, Askell A, Chen A, DasSarma N, et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. 2022. https://arxiv.org/abs/2204.05862. Accessed 21 Jan 2025.
78. Cooper G. Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence. J Sci Educ Technol. 2023;32:442–52. https://doi.org/10.1007/s10956-023-10039-y.
79. Pavlik JV. Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. J Mass Commun Educ. 2023;78(1). https://doi.org/10.1177/10776958221149577.
80. Azure M. Azure OpenAI Service - Advanced Language Models. n.d. https://azure.microsoft.com/en-us/products/ai-services/openai-service. Accessed 17 Jan 2025.
81. OpenAI. ChatGPT Documentation. 2021. https://platform.openai.com/docs/guides/chat. Accessed 17 Jan 2025.
82. Express. Node JS Express Framework Official Website. n.d. https://expressjs.com/. Accessed 21 Jan 2025.
83. Flutter. Flutter Official Website. n.d. https://flutter.dev/. Accessed 21 Jan 2025.
84. Profanity-filter. Dart - Profanity filter documentation. n.d. https://pub.dev/documentation/profanity_filter/latest/. Accessed 21 Jan 2025.

85. Ldnoobw. List of dirty, naughty, obscene, and otherwise bad words. 2020. https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words. Accessed 21 Jan 2025.
86. HHS Office for Civil Rights. Standards for privacy of individually identifiable health information. Final rule Fed Regist. 2002;67(157):53181–273.
87. US Department of Health and Human Services. The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. 1996. http://www.hhs.gov/ocr/privacy/. Accessed 17 Jan 2024.
88. GDPR. General data protection regulation. Regulation (EU). 2016;679:1–88.
89. Dialogflow. Dialogflow | Google Cloud. n.d. https://cloud.google.com/products/conversational-agents?hl=en. Accessed 21 Jan 2025.
90. Cheng MW, Yim IH. Examining the use of ChatGPT in public universities in Hong Kong: a case study of restricted access areas. Discover Educ. 2024;3(1):1.
91. Amazon. Recommendations - Amazon Customer Service. n.d. https://www.amazon.com/gp/help/customer/display.html?nodeId=GE4KRSZ4KAZZB4BV. Accessed 21 Jan 2025.
92. Cummings ML, Bauchwitz B. Safety implications of variability in autonomous driving assist alerting. IEEE Trans Intell Transp Syst. 2021;23(8):12039–49.
93. Microsoft. Azure OpenAI Service pricing. 2025. https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/#pricing. Accessed 17 Jan 2025.
94. OpenAI. OpenAI's GPT-3.5 model. n.d. https://platform.openai.com/docs/models/gpt-3-5. Accessed 17 Jan 2025.
95. Alam A. Developing a Curriculum for Ethical and Responsible AI: A University Course on Safety, Fairness, Privacy, and Ethics to Prepare Next Generation of AI Professionals. In: Intelligent Communication Technologies and Virtual Mobile Networks. Singapore: Springer; 2023. pp. 879–94.
96. Kirk T. ChatGPT, we need to talk. 2023. https://www.cam.ac.uk/stories/ChatGPT-and-education. Accessed 21 Jan 2025.
97. Roose K. Don't ban ChatGPT in schools. teach with it. The New York Times. 2023. https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html. Accessed 21 Jan 2025.
98. WhatsApp. WhatsApp Privacy Policy. n.d. https://www.whatsapp.com/legal/privacy-policy-eea#privacy-policy-information-we-collect. Accessed 17 Jan 2025.
99. Signal. Signal and the General Data Protection Regulation (GDPR). n.d. https://support.signal.org/hc/en-us/articles/360007059412-Signal-and-the-General-Data-Protection-Regulation-GDPR. Accessed 17 Jan 2025.
100. Lee U, Han A, Lee J, Lee E, Kim J, Kim H, et al. Prompt Aloud!: Incorporating image-generative AI into STEAM class with learning analytics using prompt data. Educ Inf Technol. 2024;29(8):9575–605.
101. OpenAI. ChatGPT can now see, hear, and speak. 2023. https://openai.com/blog/chatgpt-can-now-see-hear-and-speak. Accessed 17 Jan 2025.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.