Original Paper

Exploring the Possible Use of AI Chatbots in Public Health Education: Feasibility Study

Francesco Baglivo[1*], MD; Luigi De Angelis[1*], MD; Virginia Casigliani[1], MD; Guglielmo Arzilli[1], MD; Gaetano Pierpaolo Privitera[1,2], MD; Caterina Rizzo[1], MD

[1]Department of Translational Research and New Technologies in Medicine and Surgery, University of Pisa, Pisa (PI), Italy

[2]Training Office, National Institute of Health, Rome, Italy

[*]these authors contributed equally

Corresponding Author:
Francesco Baglivo, MD
Department of Translational Research and New Technologies in Medicine and Surgery
University of Pisa
Via San Zeno 35
Pisa (PI), 56123
Italy
Phone: 39 3288348649
Email: f.baglivo@studenti.unipi.it

Abstract

Background: Artificial intelligence (AI) is a rapidly developing field with the potential to transform various aspects of health care and public health, including medical training. During the “Hygiene and Public Health” course for fifth-year medical students, a practical training session was conducted on vaccination using AI chatbots as an educational supportive tool. Before receiving specific training on vaccination, the students were given a web-based test extracted from the Italian National Medical Residency Test. After completing the test, a critical correction of each question was performed assisted by AI chatbots.

Objective: The main aim of this study was to identify whether AI chatbots can be considered educational support tools for training in public health. The secondary objective was to assess the performance of different AI chatbots on complex multiple-choice medical questions in the Italian language.

Methods: A test composed of 15 multiple-choice questions on vaccination was extracted from the Italian National Medical Residency Test using targeted keywords and administered to medical students via Google Forms and to different AI chatbot models (Bing Chat, ChatGPT, Chatsonic, Google Bard, and YouChat). The correction of the test was conducted in the classroom, focusing on the critical evaluation of the explanations provided by the chatbot. A Mann-Whitney U test was conducted to compare the performances of medical students and AI chatbots. Student feedback was collected anonymously at the end of the training experience.

Results: In total, 36 medical students and 5 AI chatbot models completed the test. The students achieved an average score of 8.22 (SD 2.65) out of 15, while the AI chatbots scored an average of 12.22 (SD 2.77). The results indicated a statistically significant difference in performance between the 2 groups (U=49.5, *P*<.001), with a large effect size (r=0.69). When divided by question type (direct, scenario-based, and negative), significant differences were observed in direct (*P*<.001) and scenario-based (*P*<.001) questions, but not in negative questions (*P*=.48). The students reported a high level of satisfaction (7.9/10) with the educational experience, expressing a strong desire to repeat the experience (7.6/10).

Conclusions: This study demonstrated the efficacy of AI chatbots in answering complex medical questions related to vaccination and providing valuable educational support. Their performance significantly surpassed that of medical students in direct and scenario-based questions. The responsible and critical use of AI chatbots can enhance medical education, making it an essential aspect to integrate into the educational system.

(JMIR Med Educ 2023;9:e51421) doi: 10.2196/51421

KEYWORDS

artificial intelligence; chatbots; medical education; vaccination; public health; medical students; large language model; generative AI; ChatGPT; Google Bard; AI chatbot; health education; public health; health care; medical training; educational support tool; chatbot model

Introduction

Artificial intelligence (AI) has been taking significant steps in various fields, including health care and education. The advent of AI chatbots, in particular those built on large language models (LLMs), has opened up new possibilities for enhancing medical education, transforming the way we train future health care professionals. LLMs are a type of generative AI that has been trained on a massively large corpus of textual data from the web, specifically architected to help generate text-based content. AI chatbots have been increasingly used in health care applications, for example, to provide education and support to patients with chronic diseases [1] and to increase COVID-19 vaccine confidence and acceptance [2].

The advent of LLMs has renewed interest toward the potential of AI in the education field, mainly to serve as an assistant for educators and as a virtual tutor for students [3]. For example, CS50, an introductory course in computer science held by Harvard University, plans to use AI to grade assignments, teach coding, and personalize learning tips [4]. Medical education is no exception, with papers exploring examples of AI chatbot applications, including the generation of accurate and versatile clinical vignettes, improving personalized learning experiences, and being an adjunct in group learning [5,6]. For example, the MedQA data set is a well-known data set containing multiple-choice questions collected from real-world professional examinations; it is used as an international benchmark to test the capabilities of AI models in the health care domain [7]. MedQA includes questions from the United States Medical Licensing Exam (USMLE), a set of 3 standardized tests of expert-level knowledge. A recent paper tested ChatGPT performances on the USMLE, showing results near the passing threshold of 60% accuracy. The authors suggested that, based on this result, ChatGPT may potentially support students in the medical education field [8].

The current state-of-the-art model on MedQA performance is Med-PaLM 2, reaching a score of 86.5% in this benchmark [9]. GPT-4 achieves strong performance in many languages, including Italian, on the massive multitask language understanding benchmark, which is a data set of multiple-choice questions not specific to the health care domain [10]. However, the testing of LLMs on medical multiple-choice questions in languages other than English is still limited and worth exploring for its implications on medical education around the world.

This paper aims to evaluate the feasibility of using AI chatbots as educational support tools in public health training in Italy, specifically in the context of vaccination. We compared the performance of different AI chatbot models in answering questions related to vaccination, providing insights into the potential and limitations of AI in medical education.

Methods

Study Design

In Italy, since 2015, there has been a national admission test to medical residency after medical school called Prova Nazionale per l'Ammissione dei Medici alle Scuole di Specializzazione di Area Sanitaria, hereafter referred to as the Italian National Medical Residency Test (SSM). This test consists of 140 expert-level multiple-choice questions regarding various medical subjects (eg, cardiology and orthopedics), and it is administered in Italian. Each candidate has a specific amount of time (usually 210 minutes) to answer the questionnaire. Based on the scoring on this test, a national ranking is drawn up and each candidate can choose the specific medical residency school they want to enroll in [11].

We chose to focus on vaccination-related questions from the SSM due to their relevance in public health training, their complexity, and their controversial nature in public discourse. The topic of vaccination is related to phenomena of extreme importance, such as infodemics and vaccine hesitancy, which the World Health Organization has identified as one of the top 10 threats to global health [12,13]. We conducted a comparative analysis of different AI chatbots based on LLMs, including Bing Chat, ChatGPT, Chatsonic, Google Bard, and YouChat, in answering a set of questions related to vaccination. These questions were selected from the SSM to ensure their relevance and applicability to the topic of vaccination [14-18]. Our study did not only assess the accuracy of the responses provided by these AI chatbots but also reported a use-case of AI chatbots as assistants for the correction of a test in a real-world scenario of medical education. The completeness of the information, reliability of the sources cited, and use of technical language was discussed between medical students and lecturers with research experience on the use of LLMs in public health. Since good performances by LLMs on medical question answering tasks are necessary but not sufficient to demonstrate their applicability in medical education, we also provided an example of the use of AI chatbots as education support tools for medical students.

LLM Chatbot Selection

In our study, we chose specific chatbots based on their availability and accessibility. We decided to select only chatbots based on LLMs that use a transformer architecture, as these models can be considered the current gold standard for natural language processing tasks. Our selection was driven by the chatbots' web-based user interface availability, which obviated the need for model application programming interface use. This methodology enabled us to assess the effectiveness of these chatbots when used by a nontechnical audience, like medical students. Although LLMs fine-tuned for the health care domain, such as Med-Palm2 [9], may demonstrate superior performance in certain contexts, it is pertinent to recognize that their access and use typically necessitate technical expertise via an application programming interface. Consequently, students without a technical background would encounter difficulties in using these resources routinely for academic endeavors.

Test Extraction and Item Classification

All questions related to vaccination were extracted from the SSM from 2015 (first year of the test) to 2022. The selection process involved a systematic search of questions using targeted keywords related to vaccination (a complete list of the keywords in Italian is provided in Multimedia Appendix 1). The inclusion criteria were as follows: (1) the question must contain any of the keywords and (2) the question must be related to the topic of vaccination. The selection was performed by a single reviewer, and a total of 15 questions were included (Multimedia Appendices 2 and 3).

Furthermore, the questions were classified into the following 3 categories based on their structure [19]:

1. Direct questions: These are straightforward questions that ask for specific information. For example, “What is the composition of the X vaccine?”
2. Scenario-based questions: These questions provide a scenario or case study and then ask a question related to that scenario. They usually require a more comprehensive understanding of a topic, as they often involve applying knowledge to a specific situation. An example from the list is, “A 52-year-old man, with a negative history for COVID-19 and vaccinated with three doses of anti-COVID mRNA vaccine, performs a serological test for anti-SARS-CoV-2 antibodies a month after the third dose. What serological profile do we expect to find?”
3. Negative questions: These questions ask which statements are false or true. They often require a more careful reading, as the use of negation can make them more complex. For instance, “Which of the following statements about vaccine composition is not true?”

Test Administration to Medical Students and Chatbots

The test was administered using Google Forms [20] to fifth-year medical students as part of their practical training session during the “Hygiene and Public Health” course in April 2023 at the University of Pisa, Italy, before completing all the planned lessons on the topic of vaccination. The form was accessible via a QR code and was anonymous. The test was given to fifth-year medical students because, at the University of Pisa, the public health course is held during the fifth year of medical school. The students were asked to complete the test in 30 minutes.

Subsequently, different AI chatbot models, namely Bing Chat, ChatGPT, Chatsonic, Google Bard, and YouChat, were asked the same set of questions. No prompt was given to the chatbots; the multiple-choice questions were directly copied and pasted into the chat. The responses of the AI chatbots were evaluated on the same scoring basis as the students’ responses, with correct answers scoring 1 point and incorrect or unanswered questions scoring 0 points.

The correction of the test was conducted in the classroom during a dedicated 120-minute session. This involved showing and discussing the solution to the questions provided by one of the chatbots, which was selected based on its performance on the task and its availability. In detail, the criteria for selecting the chatbot for the correction session were as follows: performance above 90% on the task, free web-based availability, and accessible without registration. The main focus of the correction was the critical evaluation of the explanations provided by the chatbot.

Medical students’ feedback was collected anonymously at the end of the training experience through a 3-item questionnaire with a Likert scale (1 to 10) regarding their general satisfaction, willingness to repeat the experience, and ease of use of the tool. In particular, the scale of the 3 items can be translated as follows:

- Item 1: 1=“dissatisfied with the experience,” 10=“very satisfied.”
- Item 2: 1=“I would not repeat the experience,” 10=“I would definitely repeat the experience.”
- Item 3: 1=“the tool is too difficult to be used,” 10=“the tool was very easy to be used.”

Mentimeter [21] was used to collect the feedback right after the correction of the test.

Statistical Analysis

A Shapiro-Wilk test was conducted to assess the data distribution. In order to investigate any differences in performance between the medical students and AI chatbots, a Mann-Whitney U test was conducted. The rank-biserial correlation was also calculated as a measure of effect size. The performances of the medical students and AI chatbots were compared within each question type, and the Mann-Whitney U test and rank-biserial correlation were calculated for each type of question. All analyses were conducted using Python (Python Software Foundation) with the pandas, matplotlib, seaborn, and scipy libraries. The source data are available in Multimedia Appendix 4 .

Ethical Considerations

The questionnaire administered in our study was an integral part of the educational activities of the course, serving as a self-assessment tool for the voluntarily participating students. It was designed to maintain the anonymity of the participants and did not collect any personal data. According to the University of Pisa teaching regulations, ethical approval was not necessary for this study as the data were completely anonymous from the beginning and collected by a link to a web platform where respondents could not be identified, and the results of university tests conducted during regular teaching activities are public and open.

Results

Test Completion

The test was completed by 36 medical students and 5 different AI chatbot models (Table 1). ChatGPT and Bing Chat were used in different versions. The total score for each participant was calculated out of a maximum of 15 points. The total number of students enrolled in the public health course was 96, of which 36 (37.5%) voluntarily completed the questionnaire.

Table 1. The performance of various artificial intelligence (AI) chatbot models in answering the 15 questions selected from the Italian National Medical Residency Test.

AI chatbot	Mode[a]	LLM[b] model	Score (N=15), n (%)	Not answered	Date of completion
ChatGPT	3.5	GPT3.5	12 (80)	1	April 14, 2023
ChatGPT	4.0	GPT4	15 (100)	—[c]	July 13, 2023
ChatGPT	4.0 plugin Scholar AI	GPT4	15 (100)	—	July 13, 2023
Bing Chat	Precise	—	15 (100)	—	April 13, 2023
Bing Chat	Creative	—	14 (93)	—	April 12, 2023
Bing Chat	Balanced	—	11 (73)	—	April 13, 2023
Google Bard	—	LaMDA[d]	7 (47)	2	July 13, 2023
YouChat	—	—	10	2	April 14, 2023
Chatsonic	—	GPT4	11	1	April 14, 2023

[a] The specific mode or version of the AI chatbot used.

[b] LLM: large language model.

[c] Not applicable.

[d] LaMDA: Language Model for Dialogue Applications

Shapiro-Wilk tests indicated normal distributions for the total scores of both chatbots and students but nonnormal distributions for all the subcategories of questions (direct, scenario-based, and negative) for both chatbots and students. For this reason, the Mann-Whitney U was chosen as the statistical test for all the comparisons.

On average, out of 15, medical students scored 8.22 (SD 2.65; median 8, IQR 4-12; range 3-15), while the AI chatbot models scored higher, with an average score of 12.22 (SD 2.77; median 12, IQR 8-15; range 7-15). The distribution of scores is displayed in Figure 1. Details regarding the accuracy of chatbots and medical students on each single question are provided in Figure 2.

Figure 1. This histogram represents the distribution of overall scores obtained by medical students (in blue) and AI chatbots (in green) on the vaccine-related Italian National Medical Residency Test questions. Each bar represents the stacked number of students or chatbots that achieved a particular score. The scores are represented on the x-axis, and the number of students or chatbots achieving each score is represented on the y-axis.

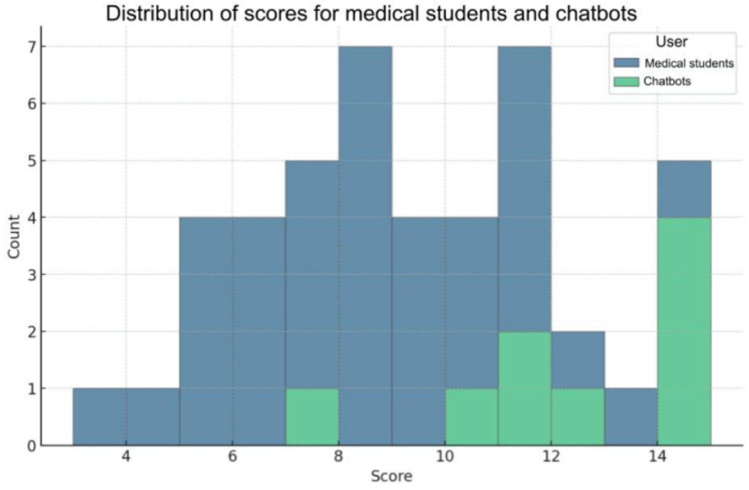
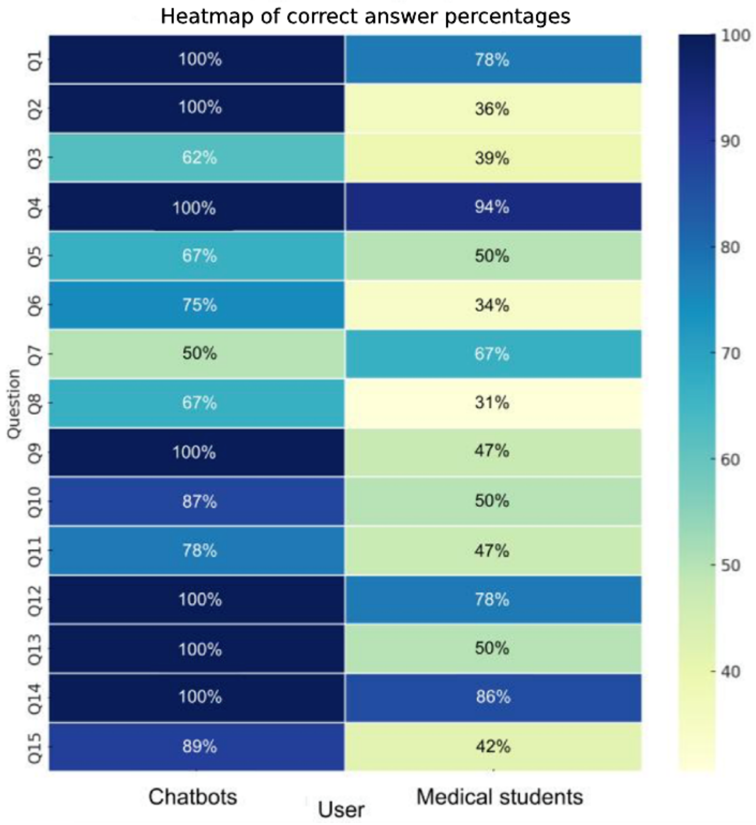


Figure 2. This heat map visualizes the percentage of correct answers provided by medical students and AI chatbot models for each question on the vaccine-related test. The questions are represented on the y-axis, and the user types (medical students or AI chatbot models) are represented on the x-axis. The color intensity in each cell corresponds to the percentage of correct answers, with darker shades representing higher percentages. The percentages are also annotated within the cells for easier reference.

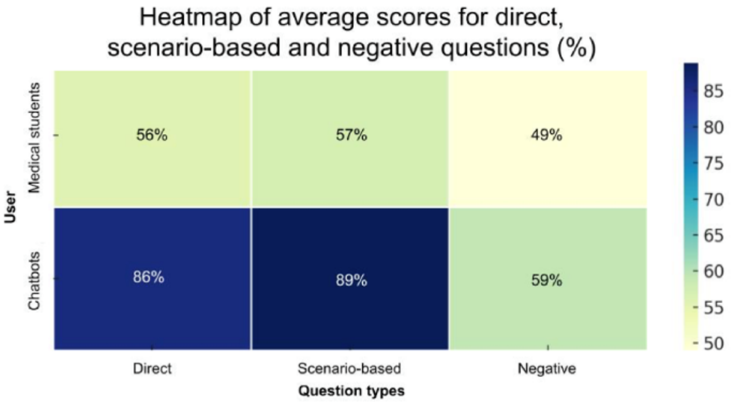


Performances on the Test: Comparison Between Medical Students and AI Chatbots

A Mann-Whitney U test was conducted to compare the total scores of the medical students and AI chatbots. The result indicated a statistically significant difference (U=49.5, *P*<.001). The rank-biserial correlation (r=0.69) suggested a large effect size, indicating a meaningful difference between the performances of the 2 groups.

The 15 items of the test were classified as following: 7 direct questions, 5 scenario-based questions, and 3 negative questions. Details of the classification can be found in Multimedia Appendix 2. Chatbots scored an average of 6.00 (SD 1.12; median 6, IQR 4-7) out of 7 on direct questions, 4.44 (SD 0.73; median 5, IQR 4-5) out of 5 on scenario-based questions, and 1.78 (SD 0.53; median 2, IQR 0-3) out of 3 on negative questions. Students scored an average of 3.89 (SD 1.14; median 4, IQR 2-6) out of 7 on direct questions, 2.86 (SD 1.31; median 3, IQR 1-5) out of 5 on scenario-based questions, and 1.47 (SD 1.00; median 1, IQR 0-2) out of 3 on negative questions. The percentage of correct answers to each type of question for both groups can be found in Figure 3.

Figure 3. This heat map displays the percentages of correct answers for the 2 groups of chatbots and medical students in each specific category of questions. The color intensity in each cell corresponds to the percentage of correct answers, with darker shades representing higher percentages. The percentages are also annotated within the cells for easier reference.



For direct questions, the Mann-Whitney U test showed a statistically significant difference in the scores of medical students and AI chatbots (U=33.5, *P*<.001). The rank-biserial correlation was 0.79, indicating a large effect size. For scenario-based questions, the Mann-Whitney U test also showed a statistically significant difference in scores (U=52.5, *P*=.002). The rank-biserial correlation was 0.68, suggesting a large effect size. However, for negative questions, there was no statistically significant difference in scores (U=137.5, *P*=.48). The rank-biserial correlation was 0.151, indicating a small effect size.

Report of the Educational Experience

Bing Chat (creative mode) was chosen to conduct the corrections in the classroom due to its good performance on the task (score: 14/15) and its availability at the time of our study.

Throughout the correction process, students actively participated in discussions, critically evaluating the explanations provided by the chatbot. Feedback collected postsession via Mentimeter revealed a high level of satisfaction and ease of use, with scores of 7.9 and 8.2, respectively, on a 10-point Likert scale. The students commended the novel and interactive format, stating that it added a fresh dimension to the traditional teaching approach, and showed a strong desire to repeat the experience (7.6/10).

Discussion

Principal Findings

Our paper explored the role of AI chatbots, particularly those constructed upon LLMs (ie, ChatGPT), in medical education and their potential to support learning and training in public health through a practical use-case experience. The results of our feasibility study showed that LLM-based chatbots can correctly answer complex health-related multiple-choice questions in Italian in the specific domain of vaccination, proving to possibly be a supporting educational tool in this specific setting. By using questions from the SSM, we not only evaluated the accuracy of the chatbots’ responses but also examined a real-world application of AI in providing an explained correction of a medical admission test. Bing Chat (creative mode) was chosen for the correction in class because, while it was not the best-performing chatbot, it provided longer and more in-depth answers to each question, thus providing a better ground for classroom discussion with students. ChatGPT was temporarily unavailable in Italy following an action of the President of the Italian Data Protection Authority for breaches of the European legislation on personal data processing and protection by OpenAI [22].

The chatbots analyzed exhibited high-level performance that was, on average, higher than the performance of the medical students. The chatbots showed a statistically significant superiority for direct and scenario-based questions, while they were less accurate on negative questions (not statistically significant). The performances of chatbots on this specific task relied on various factors and could be further improved by using prompt engineering and techniques such as chain-of-thought prompting [23].

Notably, good performance alone is not enough for the useful and safe adoption of these tools in real-world applications for medical education purposes. Especially in the medical domain, it is better to promote awareness of the benefits and limits of LLMs rather than prohibiting students from using them [24]. The critical evaluation of the answers provided by the chatbot not only enhanced students’ understanding of the correct responses but also stimulated conversations about the underlying concepts, resulting in a positive attitude of the participating students toward the tool [25]. In fact, the students reported a general satisfaction and a willingness to repeat the educational experience proposed in our study.

As suggested later by Cooper and Rodman [26], as medical educators, we took an activist approach trying to integrate AI into physician training, with the objective of preparing our students for safe and appropriate use of this technology in health care. In the current educational landscape, while the potential of LLMs as teaching tools is evident, their incorporation into traditional pedagogical methods demands planning. LLMs can be a useful support tool within different phases of teaching. In the introduction of new topics, they can act as supplementary informational sources, helping students to grasp foundational concepts quickly [5,27]. During in-depth discussions or tutorials, LLMs can serve as interactive tools to challenge students’ understanding, offering real-time feedback [4]. Moreover, in the revision phase, these models can be pivotal in addressing specific queries, clarifying doubts, and reinforcing knowledge through simulated question and answer sessions.

Our approach allows group discussions stimulating critical thinking about the potentiality and limits of AI chatbots in medical education. In fact, it is crucial to introduce students to the limitations of LLMs, such as their reliance on biased data, limited up-to-date knowledge, variable performances over time, and the potential for generating incorrect or false information [3,27]. The issue of “hallucinations" is particularly concerning in medical education and has to be properly discussed with students due to the possible fabrication of scientific references among other false or misleading information [28,29]. Even if the reliability of scientific references cited by ChatGPT and other LLMs is rapidly increasing thanks to their ability to browse the web and the use of plug-ins, such as ScholarAI, that seamlessly integrate peer-reviewed article searches into ChatGPT conversations, the need to demand a deep and critical check of the sources cited by LLMs and treat them as guilty until proven innocent remains [30].

A critical use of this tool should also be encouraged to counter the deskilling derived from an overreliance on it—students might eventually lose their abilities to produce original ideas and present proper arguments to prove their statements. Furthermore, chatbots cannot be used as substitutes in clinical reasoning, and specific training, through case studies and simulations, should be foreseen in medical school [31]. Since students, residents, and fellows are already using such tools, it is our duty to guide the academic community in raising awareness rather than prohibiting, or worse ignoring, the change.

Limitations

Our study has some limitations. The sample size of medical students was relatively small, which may limit the generalizability of the results. A larger sample size could offer a more comprehensive and reliable reflection of the performance of medical students. Moreover, all the questions used for the test focused solely on the topic of vaccination. While this focus provided valuable insights into the performance of the AI chatbots and students in this specific area, it may not fully represent their proficiency across a wider array of medical topics. Additionally, due to restrictions on time and availability, only one chatbot was thoroughly used and evaluated with the students in class. As highlighted in a recent paper [32], the performance of ChatGPT on different tasks seems to substantially change over time, at times worsening. Even if this behavior has not been demonstrated for medical questions yet, it could potentially reduce the long-term reliability of our results.

Future Perspectives

The present study offers insights into the potential role of AI chatbots as support tools in training. There are multiple stages in the individual training pathway where students can benefit from the support of this technology. The cited Harvard example [4] is just one of many potential applications. In the medical field, AI-powered chatbots can assist students in conducting targeted searches for scientific literature, helping them find relevant and reliable references for their studies. Essentially, the chatbots could serve as an interface that may guide students to the best available learning resources, discarding irrelevant or less useful materials. This approach could offer personalized training, catering to individual interests and personal learning needs. However, it should not only focus on knowledge components, potentially neglecting the development of competencies, as defined by the World Health Organization [33]. The implementation of mutable virtual simulation scenarios could address the implementation of specific skills and attitudes; in this use case, students could face a simulation that was not based on predetermined algorithmic scripts but rather on a virtual interlocutor with a variable and human-like approach powered by AI. In this way, it may be possible to develop an experiential approach similar to a specific real-world scenario (eg, an interview of parents on vaccine adverse effects), which would be useful for training students' communication and practical skills in public health.

Future Studies

In the future, we aim to investigate the performance of chatbots across all questions from the SSM to assess how well the AI models can navigate a broader and more diverse range of medical subjects. Such an analysis would allow us to deeply evaluate the ability of chatbots to comprehend and respond accurately in Italian, evaluating linguistic proficiency gaps that might need to be addressed in future model development for the tools to be actually used by Italian medical students. Another aspect of this future study would be a comparison between the AI chatbots' performance and the actual results obtained by Italian doctors, providing a significantly wider benchmark for the Italian language.

Further studies are needed to assess if the integration of AI tools in public health medical training may improve the acquisition of knowledge and performances in final exams. In this way, we think that starting with a practical example of the application of a chatbot based on LLMs can be a beginning for experimenting with AI in support of training for health professionals, with the prospect of expanding this range of application to orient us toward the innovations in training proposed by supranational and national organizations.

Our feasibility study provided a real-world example of the application of AI tools in support of training for health professionals in public health. It demonstrated a good reliability of the tools used and a high satisfaction of the students for this type of practical activity, supporting the possible use of AI for medical education in public health. Further studies should be encouraged to explore other possible applications of AI-based tools in health care training in order to assess if they improve the performance of the students and to guide their awareness and critical use.

Acknowledgments

We thank all the medical students at the University of Pisa who participated in the educational activity reported in this study.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

All raw data are available in Multimedia Appendix 4.

Authors' Contributions

FB and LDA conceived the study. FB, LDA, VC, and GA performed the literature search and drafted the manuscript. CR and GPP provided expert insights and contributed to the manuscript revision. All the authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Keywords used for the filtering of questions from the SSM to identify vaccine-related questions.

[DOCX File , 21 KB-Multimedia Appendix 1]

Multimedia Appendix 2

List of questions used in the study. This table presents the 15 questions selected through keywords from the Italian National Medical Residency Test (SSM) translated into English. Each question is categorized by type: direct, scenario-based, or negative. The questions cover various topics related to vaccination and are used to evaluate the performance of AI chatbots in providing accurate and comprehensive responses.
[DOCX File , 27 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Original questions in Italian extracted from Italian National Medical Residency Tests (SSMs) from 2015 to 2022.
[DOCX File , 26 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Raw data.
[XLSX File (Microsoft Excel File), 11 KB-Multimedia Appendix 4]

References

1. Schachner T, Keller R, V Wangenheim F. Artificial intelligence-based conversational agents for chronic conditions: systematic literature review. J Med Internet Res 2020 Sep 14;22(9):e20701 [FREE Full text] [doi: 10.2196/20701] [Medline: 32924957]
2. Lee KY, Dabak SV, Kong VH, Park M, Kwok SLL, Silzle M, et al. Effectiveness of chatbots on COVID vaccine confidence and acceptance in Thailand, Hong Kong, and Singapore. NPJ Digit Med 2023 May 25;6(1):96 [FREE Full text] [doi: 10.1038/s41746-023-00843-6] [Medline: 37231110]
3. Lo CK. What is the impact of ChatGPT on education? A rapid review of the literature. Educ Sci 2023 Apr 18;13(4):410 [doi: 10.3390/educsci13040410]
4. Rai S, Bloomberg. Harvard professor taps A.I. to help teach world's most popular online computer class. Fortune. 2023. URL: https://fortune.com/2023/06/03/ai-to-help-teach-harvard-university-online-computer-science-course/ [accessed 2023-07-31]
5. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-607 [FREE Full text] [doi: 10.12669/pjms.39.2.7653] [Medline: 36950398]
6. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
7. Jin D, Pan E, Oufattole N, Weng W, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci 2021 Jul 12;11(14):6421 [doi: 10.3390/app11146421]
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
9. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv. Preprint posted online on May 16, 2023 [FREE Full text] [doi: 10.1038/s41586-023-06455-0]
10. OpenAI. GPT-4 technical report. arXiv. Preprint posted online on March 15, 2023 [FREE Full text]
11. Tarchi L, Damiani S, Marini S, Cappelli C, Liuzzi G, Minerva M, et al. Do local curriculum scores correlate with national residency test results? A pluriannual, nationwide survey of Italian Medical Universities. Ital J Med 2021 Feb 12;15(2):99-106 [doi: 10.4081/itjm.2021.1470]
12. Mheidly N, Fares J. Leveraging media and health communication strategies to overcome the COVID-19 infodemic. J Public Health Policy 2020 Dec;41(4):410-420 [FREE Full text] [doi: 10.1057/s41271-020-00247-w] [Medline: 32826935]
13. Ten threats to global health in 2019. World Health Organization. 2019. URL: https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019 [accessed 2023-07-31]
14. Bing chat. Microsoft. 2023. URL: https://www.microsoft.com/en-us/edge/features/bing-chat?form=MT00D8 [accessed 2023-07-31]
15. ChatGPT. 2023. URL: https://chat.openai.com/ [accessed 2023-07-31]
16. Bard. Google. 2023. URL: https://bard.google.com/ [accessed 2023-07-31]
17. You.com. 2023. URL: https://you.com/search?q=who+are+you&tbm=youchat&cfr=chat [accessed 2023-07-31]
18. Chatsonic – best ChatGPT alternative for content creation. Writesonic. 2023. URL: https://writesonic.com/chat [accessed 2023-07-31]

19. Rauschert E, Yang S, Pigg R. Which of the following is true: we can write better multiple choice questions. Bulletin Ecologic Soc America 2018 Nov;100(1):e01468 [doi: 10.1002/bes2.1468]
20. Chaiyo Y, Nokham R. The effect of Kahoot, Quizizz and Google Forms on the student's perception in the classrooms response system. In: 2017 International Conference on Digital Arts, Media and Technology (ICDAMT). 2017 Presented at: ICDAMT; March 1-4, 2017; Chiang Mai [doi: 10.1109/icdamt.2017.7904957]
21. Mentimeter. URL: https://www.mentimeter.com/ [accessed 2023-07-31]
22. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023;11:1166120 [FREE Full text] [doi: 10.3389/fpubh.2023.1166120] [Medline: 37181697]
23. Zhang Z, Zhang A, Li M, Smola A, Shanghai ?, Tong J. Automatic chain of thought prompting in large language models. arXiv. Preprint posted online on October 7, 2022 [FREE Full text]
24. Yu H. Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. Front Psychol 2023;14:1181712 [FREE Full text] [doi: 10.3389/fpsyg.2023.1181712] [Medline: 37325766]
25. Moldt J, Festl-Wietek T, Madany Mamlouk A, Nieselt K, Fuhl W, Herrmann-Werner A. Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. Med Educ Online 2023 Dec;28(1):2182659 [FREE Full text] [doi: 10.1080/10872981.2023.2182659] [Medline: 36855245]
26. Cooper A, Rodman A. AI and medical education — a 21st-century Pandora's box. N Engl J Med 2023 Aug 03;389(5):385-387 [FREE Full text] [doi: 10.1056/nejmp2304993]
27. Safi Z, Abd-Alrazaq A, Khalifa M, Househ M. Technical aspects of developing chatbots for medical applications: scoping review. J Med Internet Res 2020 Dec 18;22(12):e19127 [FREE Full text] [doi: 10.2196/19127] [Medline: 33337337]
28. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023 Mar 06;9:e46885 [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
29. Goddard J. Hallucinations in ChatGPT: a cautionary tale for biomedical researchers. Am J Med 2023 Jun 25:00401-1 [doi: 10.1016/j.amjmed.2023.06.012] [Medline: 37369274]
30. ScholarAI. 2023. URL: https://scholar-ai.net/ [accessed 2023-07-31]
31. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 01;9:e48291 [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
32. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? arXiv. Posted online on July 18, 2023 2021 [FREE Full text]
33. Global competency framework for universal health coverage . World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240034686 [accessed 2023-07-31]

Abbreviations

AI: artificial intelligence
LLM: large language model
SSM: Italian National Medical Residency Test
USMLE: United States Medical Licensing Exam

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 31.07.23; peer-reviewed by T Kgosietsile, R AlSaad; comments to author 24.08.23; revised version received 29.08.23; accepted 27.09.23; published 01.11.23

Please cite as:
Baglivo F, De Angelis L, Casigliani V, Arzilli G, Privitera GP, Rizzo C
Exploring the Possible Use of AI Chatbots in Public Health Education: Feasibility Study
JMIR Med Educ 2023;9:e51421
URL: https://mededu.jmir.org/2023/1/e51421
doi: 10.2196/51421
PMID: 37910155