Informatics in Medicine Unlocked 41 (2023) 101314

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Application of ChatGPT in multilingual medical education: How does ChatGPT fare in 2023's Iranian residency entrance examination

Hamid Khorshidi [a], Afshin Mohammadi [b], David M. Yousem [c], Jamileh Abolghasemi [d], Golnoosh Ansari [c,e], Mohammad Mirza-Aghazadeh-Attari [c], U Rajendra Acharya [f], Ali Abbasian Ardakani [g,*]

[a] Department of Information Engineering, University of Padova, Padova, Italy
[b] Department of Radiology, Faculty of Medicine, Urmia University of Medical Science, Urmia, Iran
[c] Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins Medical Institutions, Baltimore, MD, USA
[d] Department of Biostatistics, School of Public Health, Iran University of Medical Sciences, Tehran, Iran
[e] Department of Radiology, Northwestern University Feinberg School of Medicine, 420 E Superior St, Chicago, IL, USA
[f] School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, Queensland, Australia
[g] Department of Radiology Technology, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

ARTICLE INFO

Keywords:
ChatGPT
USMLE
Medical education
Language model

ABSTRACT

Background: ChatGPT is a large language model (LLM) artificial intelligence instrument trained on massive amounts of text data extracted from the internet and/or user input. In the present article, we aim to apply the latest version of ChatGPT to the Iranian Medical Residency Examination.

Methods: The Iranian Medical Residency Examination is composed of 200 multichoice questions covering all domains of medicine. We used ChatGPT to translate questions into English, French, and Spanish. We fed the questions as multiple-choice questions and allowed ChatGPT to provide comprehensive answers and justifications for its choices.

Results: ChatGPT was able to answer 161 (81.3% = 161/198) questions correctly when the Persian language was used. When the questions were translated into English, French, and Spanish, ChatGPT answered six, one, and five additional questions correctly, respectively. When comparing the different languages, there was no significant difference in the functioning of ChatGPT in different languages using either the McNemar test or the Binomial test.

Conclusion: ChatGPT can deliver above-average performance in the Iranian Medical Residency Examination, demonstrating its potential for using language models in medicine.

1. Introduction

Artificial intelligence (AI) and its many potential applications have been gaining traction in various fields, including medicine. The fascination with the rapid advancement of AI has led scholars to coin the term "fourth industrial revolution" for the potentially transformative effects of AI [1].

One of the most well-known additions to the expansive set of AI tools is the chat generative pre-trained transformer (ChatGPT), developed by OpenAI (San Francisco, CA, USA). ChatGPT is a large language model (LLM), an advanced artificial intelligence system designed to understand and generate human-like text based on the patterns and information it has learned from training data. It can engage in conversation, answer questions, provide explanations, and generate creative content across various topics and contexts by use of a neural network called "transformers", specifically designed to process sequential data, such as text, by capturing contextual relationships between words or symbols. The GPT model is characterized by many weights and is trained on an unfathomably large amount of data usually provided on the world wide web. The latest version of ChatGPT (ChatGPT-4) contains 100 trillion parameters and is being trained on a massive corpus of text data extracted from the internet and text input from users of the previous

* Corresponding author. Department of Radiology Technology, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, P.O.Box: 1971653313, Tehran, Iran.
E-mail addresses: Ardakani@sbmu.ac.ir, A.ardekani@live.com (A. Abbasian Ardakani).

https://doi.org/10.1016/j.imu.2023.101314
Received 24 June 2023; Received in revised form 28 July 2023; Accepted 29 July 2023
Available online 31 July 2023

versions. ChatGPT-4 is able to process large sequences of data, learn patterns, dynamically weigh the significance of different sequences of the input data and generate human-like responses [2].

As in many professions, scholars have aimed to utilize chatbots, computer programs, or an AI-based system designed to simulate human conversation or interaction to increase the quality and safety of services in the medical field. For some time, LLMs have been employed in patient communication and consumer health education [3,4], with growing interest in using such tools in content creation for patients, streamlining the process of obtaining informed consent, and facilitating medical documentation [5]. Real-world examples have been the use of chatbots as tools in helping clinicians explain medical procedures, risks and complications and benefits of medical or surgical therapeutic options to patients, and translating technical medical jargon and elaborate clinical concepts into easy-to-understand scripts [6,7].

Most recently, the utility of LLMs has been investigated in the field of medical education and medical research. Chatbots have been used to perform numerous tasks, ranging from generating clinical notes and clinical vignettes to design educational materials and educational texts, such as presentations, slideshows, or even medical mnemonics [7–9]. Interesting investigations have shown that ChatGPT performs well in different levels of medical exams, ranging from undergraduate shelf exams to comprehensive medical boards and core specialty examinations [9,10]. Though multiple authorities exist for the implementation of medical accreditation in different countries and settings, these tests usually incorporate standardized multiple-choice questions (MCQs) [11]. These MCQs are usually in the form of True-False or One-Best-Answer Items and aim to measure learning objectives in different levels, including knowledge (includes recall of memorized data by the respondent), combined comprehension and application of knowledge (requires the respondent not just to recall information, but also use them in different clinical contexts), and problem-solving (requires the respondent to understand concepts, their relationships and the aptitude to analyze them) [12]. Most recently, it has been shown that ChatGPT can perform well in simulation tests for the United States Medical Licensing Examination (USMLE) and can perform equal to competency levels expected from a near-graduating medical doctorate student [13–15]. ChatGPT-4 easily exceeds the threshold for passing the USMLE examination.

In the present study, we aim to apply the latest version of ChatGPT to the Iranian Residency Entrance Exam. This heavy knowledge-based MCQ exam encompasses nearly all fields of medicine, including all four major branches of the medical sciences (internal medicine, pediatrics, obstetrics, gynecology, and surgery) and a rather comprehensive assortment of minor surgical specialties, diagnostic radiology, medical ethics, biostatistics, and epidemiology [16]. For the first time, we also aim to investigate if the model's functionality depends on the language of the questions, feeding the model questions in the Persian script as the original language of the text, English as the dominant language of academia, and Spanish and French as the other two dominant languages in medical education and research [17,18].

2. Methods

2.1. Utilization of ChatGPT

The latest version of ChatGPT (Version 4, May 24th) was accessed between May 25th, 2023, and July 15, 2023. This chatbot was fed with a complete set of questions in Persian, the native language of the questions. Each question was copied and passed directly from the Print version of the examination provided by the Iranian Ministry of Health and Medical Education (MOHME), the test organizer. ChatGPT was also utilized to translate the questions into three of the most common languages utilized in day-to-day medical correspondence around the globe: English, French and Spanish [19].

2.2. Source of multiple-choice questions: the Iranian residency entrance exam

Unlike the resident Matching process in the United States, and similar to many other countries, the Iranian Medical Board, with the MOHME, administers a Centralized Test composed of 200 multichoice questions to rank individuals wanting to enter residency programs in University-hospitals affiliated with any of the Medical Sciences universities governed by the MOHME. Annually, more than 12.000 applicants apply to over 54 different University health systems in over 26 medical specialties, with 4300 individual residency spots available for the applicants. Applicants are composed of final-year medical students or physicians who have completed medical school, thus, all applicants have a minimum of 6 years of dedicated pre-clinical and clinical medical education prior to taking the exam.

The exam covers all fields of clinical medicine, encompassing over 17 general topics, including Internal Medicine, General Surgery, Pediatrics, Obstetrics and Gynecology (OB-GYN), Radiology, etc. The suggested study material for this exam exceeds 20 textbooks with an estimated burden of more than 26.000 pages [20]. This exam penalizes incorrect answers with 0.33% of the mark of a correct question, and respondents have to obtain a minimal mark of 150/600 to be able to apply for any given position. The median score of the applicants in this examination oscillates around 350/600 and a score above 500/600 usually puts the applicants in the top 2% [15].

Of the 200 questions in the 2023's Iranian Residency Entrance Exam, 198 questions were included, and 2 questions were excluded because they had multiple media associated with them. Both of these questions pertained to surgical cases, one being a question related to general surgery and the other being related to orthopedic surgery.

Table 1 shows the reference material for the examination and the used textbooks or guidelines in more detail.

2.3. Statistical analysis

The number of correct and incorrect answers were reported for each of the 17 topics of the examination in each of the four languages. The correct percentage of answers was recorded for each topic and the entirety of the examination.

The agreement between the Iranian residency entrance exam keys and ChatGPT answers was determined for each language input using Fleiss' kappa coefficient. The strength of agreement is categorized into five main groups according to the value of Fleiss' kappa coefficient: 0–0.20 for slight, 0.21–0.40 for fair, 0.41–0.60 for moderate, 0.61–0.80 for substantial, and 0.81–1.00 for excellent. In addition, the McNemar test was used to evaluate the differences in correct answer patterns of ChatGPT across four language inputs. Finally, the binomial statistical test was employed to compare the accuracy of ChatGPT based on four different languages. All statistical analyses were performed using SPSS software (version 24, IBM, Armonk, New York). A p-value less than 0.05 was considered statistically significant. Fig. 1 highlights the five main phases involved in this study schematically.

3. Results

The 198 questions were fed into the model. ChatGPT answered 161 (81.3%) questions correctly when the Persian language was used. When the questions were translated into English, French, and Spanish, ChatGPT answered six, one, and five additional questions correctly, respectively. A detailed summary of the performance is presented in Table 2. ChatGPT performed above 80% in 10 of the 17 medical topics and scored above 75% in the four major topics of the exam (Internal Medicine, Surgery, Pediatrics, OB-GYN). Interestingly, in the topics of Psychiatry, Pharmacology, Urology, and Medical Ethics, ChatGPT achieved a perfect mark (100%).

There was a high degree of agreement between ChatGPT and the

Table 1
Study material proposed by the Iranian Health Ministry for the Iranian Residency Entrance examination.

Topic	Proposed study material, circulated by the Iranian Health Ministry
Internal Medicine	1 Edward J. Wing, Fred J. Schiffman. Cecil Essentials of Medicine//Elsevier/10th Ed./2021 2 Loscalzo J. Harrison's Principles of Internal Medicine/ 21st Ed./McGraw-Hill/2022 3 Ebrahim NematiPour, "Diseases of the CaardioVascular System"/second edition/2020 4 National Guidelines for treating and managing Tuberculosis
Pediatrics	1 Karen J. Marcdante et al. Nelson Essential of Pediatrics. 9th edition. W. B. Saunders/2022 2 Breastfeeding Updates for the Pediatricians/February 2013 3 National Vaccination Guidelines 4 National Guidelines for well-child visits
Neurology	1 Proceedings of the Iranian Society of Neurology "Pathologies of the Central and peripheral nervous system"/2019/2nd Edition
Psychiatry	1 Pocket handbook of psychiatry/Kaplan & Sadock's/ 6th edition/Lippincott Williams & Wilkins/2018 2 Ghalebandi F. Clinical psychology and behavioral sciences"/2017
Dermatology	1 Lookingbill and Marks' Principles of Dermatology James G. Marks JR, Jeffry J. Miller, 6th, Edition, 2019 2 Mortazavi H, "Skin conditions",/2020 edition
Pathology	1 Robbins Basic Pathology/10th Edition/Copyright. 2018/Elsevier Inc
Radiology	1Learning Radiology Recognizing the Basics/William Harring/4th edition/2020
Surgery (General)	1Essentials of General Surgery and Surgical Specialties, 6th Edition- Peter F. Lawrence, 2019
Obstetrics and Gynecology	1Charles R. B. Beckmann et al. Obstetrics & Gynecology. 8th edition. American College of Obstetricians and Gynecologists. 2019
Urology	1SimForoosh N, "General Urology"/Third Version, 2021
Orthopedics	1Alami-Harandi B, "Orthopedics and fractures", 6th edition, 2019
Ophthalmology	1Javadi MA, "General ophthalmology", Second edition, 2018
ENT	1Handbook of Otolaryngology/Head and Neck Surgery/ David Goldenberg, Bradley J. Goldstein/Thieme/Second Edition/2017
Medical Ethics	1Larijhani MB, "The Physician and Ethical Considerations", Second Version
Epidemiology and Biostatics	1Introduction to Biostatistics and Research Methods/fifth Edition/P·S.S. Sundar Rao,DR. PH/2012 b y PHI Learning Private Limited, New Delhi 2 Yari P, "Epidemiology of Common diseases in Iran", Version two, 2021
Pharmacology	1Katzung Bertam G. Pharmacology: Examination & Board Review/McGraw-Hill/12th edition/2019

correct answers which were designated by the exam officials. The highest amount of agreement was between the English Version of the Exam and the Corrected answer sheet (Kappa = 0.791). In addition, and the lowest agreement was seen for the Persian version (Kappa = 0.751), which was the original language of the questions. More information is presented in Table 3.

When comparing the different languages, there was no significant difference in the functioning of ChatGPT in different languages using either the McNemar test or the Binomial test (Tables 4 and 5). The McNemar test compares the pattern of ChatGPT response to the questions by one-on-one comparison with the original test language and shows that the input language had no significant effect on the pattern of response. This test shows that if a particular question was selected in any given language, the model selected the same choice in all other languages as well.

The Binominal test determines indicated that there is no significant difference in overall accuracy among the four included languages, and as seen there is no significant difference (Table 5).

Table 2
Performance of ChatGPT on 2023's Iranian residency multiple-choice question tests based on seventeen question topics and four languages.

Question Topic	Number of Question	Number of Correct Responses (Percentage)			
		Persian	English	French	Spanish
Pediatrics	26	20 (76.9)	22 (84.6)	21 (80.8)	22 (84.6)
Obstetrics and Gynecology	18	16 (88.9)	16 (88.9)	15 (83.3)	15 (83.3)
General Surgery	23	16 (69.6)	20 (87)	19 (82.6)	19 (82.6)
Internal Medicine	45	35 (77.8)	34 (75.6)	32 (71.1)	36 (80)
Psychiatry	**8**	**8 (100)**	**8 (100)**	**8 (100)**	**8 (100)**
Pathology	9	8 (88.9)	8 (88.9)	8 (88.9)	8 (88.9)
Radiology	6	4 (66.7)	4 (66.7)	4 (66.7)	3 (50)
Infectious Diseases	10	9 (90)	9 (90)	10 (100)	9 (90)
Neurological Disorders	8	7 (87.5)	7 (87.5)	7 (87.5)	7 (87.5)
Pharmacology	**6**	**6 (100)**	**6 (100)**	**6 (100)**	**6 (100)**
Epidemiology	6	4 (66.7)	5 (83.3)	5 (83.3)	5 (83.3)
Otolaryngology (ENT)/Head and Neck Surgery	6	4 (66.7)	4 (66.7)	4 (66.7)	4 (66.7)
Ophthalmology	6	5 (83.3)	6 (100)	6 (100)	6 (100)
Urology	**6**	**6 (100)**	**6 (100)**	**6 (100)**	**6 (100)**
Orthopedics	7	7 (100)	5 (71.4)	5 (71.4)	6 (85.7)
Dermatology	6	4 (66.7)	5 (83.3)	4 (66.7)	4 (66.7)
Medical Ethics	**2**	**2 (100)**	**2 (100)**	**2 (100)**	**2 (100)**
All Questions	**198**	**161 (81.3)**	**167 (84.3)**	**162 (81.8)**	**166 (83.8)**

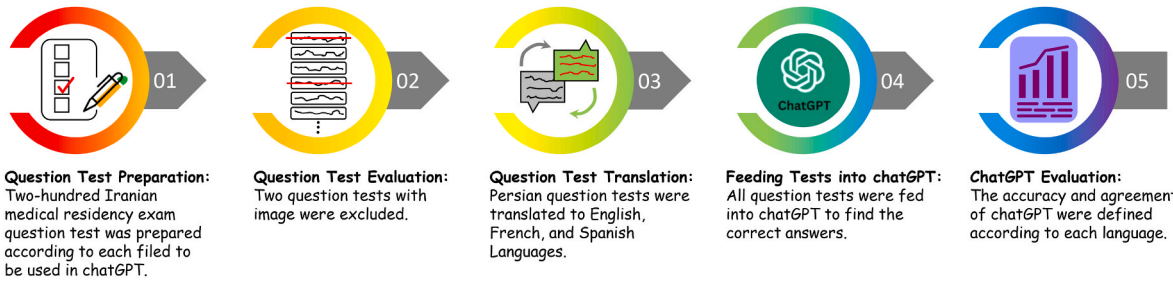

Fig. 1. Flow chart of the study highlighting the five main phases of the methodology.

Table 3
The agreement of ChatGPT with the Iranian residency entrance exam based on four different language inputs.

ChatGPT Responses →		Persian				Kappa	English				Kappa	French				Kappa	Spanish				Kappa
		A	B	C	D	0.751	A	B	C	D	0.791	A	B	C	D	0.757	A	B	C	D	0.784
Exam Answer Keys	A	40	5	1	3		39	3	1	6		38	2	3	6		39	3	2	5	
	B	3	41	2	3		4	42	1	2		5	40	1	3		3	42	2	2	
	C	3	5	45	1		5	3	46	0		5	3	46	0		5	3	46	0	
	D	3	4	4	35		3	3	0	40		3	4	1	38		4	2	1	39	

Table 4
Comparison of ChatGPT patterns in selecting from multiple choices in four different language inputs using McNemar statistical test. Values represent P values and show that no statistically significant difference was witnessed.

ChatGPT Responses	Persian	English	French	Spanish
Persian	–	0.307	1.000	0.383
English	0.307	–	0.180	1.000
French	1.000	0.180	–	0.344
Spanish	0.383	1.000	0.344	–

Table 5
Comparison of ChatGPT accuracies based on the four different language inputs using a Binomial statistical test. Values represent P values and show that no statistically significant difference was witnessed.

ChatGPT Responses	Persian	English	French	Spanish
Persian	–	0.424	0.897	0.508
English	0.424	–	0.503	
French	0.897	0.503	–	0.891
Spanish	0.508	0.891	0.594	–

4. Discussion

In this study, we present results obtained from applying the latest version of a large language model called ChatGPT-4 on the Iranian Medical Residency Entrance Exam. Our results demonstrate that this chatbot is able to perform well compared to exam takers, as it correctly answers greater than 80% of the questions, which would put it well above the average score of the human test takers.

The Iranian Medical Residency Examination is a standardized testing session that takes 3 h and the respondents answer 200 multiple-choice questions [21]. Due to the limited number of residency spots compared to the number of applicants, this exam is considered one of the most competitive centralized testing sessions in the domain of medicine [22]. The test is composed of questions covering all fields of medicine, with both knowledge-based questions and those related to the clinical management of patients, professional conduct, and medical ethics. The highest performance of ChatGPT was in topics such as Pharmacology, Psychiatry, Urology, and Infectious Diseases. These fields are usually composed of questions that directly assess the examinee's clinical "textbook" knowledge/memory of medical conditions and semantic information, which could be easily retrieved from online data sources, thus enabling models that are developed on such information to perform exceptionally well.

However, ChatGPT losses functionality as more abstract ideas and conditions are investigated in questions. Importantly, questions requiring the physician to contemplate the subjective symptoms of patients, those requiring careful splitting of conditions with very similar clinical pictures, questions focusing on particular steps of clinical guidelines, or those questions which require the examinee to provide specialized medical suggestions to patients are among those which were more frequently answered erroneously [23]. This particular finding was also previously mentioned by other scholars who had suggested that language models (including ChatGPT) would perform well on lower domains of knowledge, more specifically those concerned with the recall of facts and simple explanation of ideas or concepts (bloom taxonomy) [24,25]. In one interesting observation, ChatGPT-3.5 was used to go through questions that would be present in Radiology Board-Style examinations. From a total of 150 questions, ChatGPT was able to choose the correct answer in 104 questions (69%), while being able to provide a correct answer in only 53 questions out of 89, which were classified as high-order questions based on the Bloom taxonomy (60%). The lowest performance was seen in questions necessitating the application of clinical concepts to new scenarios (3 out of 10 questions), calculation of classification of patients (2 out of 8) and determination of disease associations (4 out of 7) [26].

Another important experiment was done by Kung et al. where ChatGPT version 3 was used to answer MCQs from the United States Medical Licensing Examination (USMLE) on the three forms necessary to pass the exams. STEP 1 of the examination focuses on the clinical aspects of basic medical sciences, STEP 2 and 3 focus on clinical cases and are mainly composed of clinical vignettes, requiring the respondent to analyze data, adjudicate on the importance of specific clinical findings and make a judgment on the most appropriate answers. ChatGPT achieved an accuracy of 45.4%, 54.1%, and 61.5%, for the first, second, and third steps, respectively. When indeterminate responses were classified as correct answers, the accuracy increased to 75%, 61.5%, and 68.8%, respectively. The authors also found a high degree of answer-explanation concordance, concluding that the language BOT had high degrees of internal consistency in its probabilistic models [14]. In all three forms, ChatGPT achieved a passing score, performing above the minimum threshold set for advancing through the exams. The results of this study also mimic ours, as we also show that the language model is able to perform solidly on questions assessing knowledge in different domains of medicine. However, our results are more promising, and we report higher accuracies compared to those that have been reported with STEP2/3 forms which resemble the Iranian Residency Entry Examination. This could be due to two reasons, one being that we utilized ChatGPT-4, the newer version of the language model, and the second being the differences in the questions between the two exams. The USMLE is mostly composed of clinical vignettes, which require a higher degree of medical reasoning and incorporate higher degrees of complex medical reasoning [27,28].

Another gap in the literature was evaluating the capability of ChatGPT to answer the same question in different languages. Our results indicated that the type of input language did not significantly affect both perception (Table 4) and performance (Table 5) of ChatGPT. Thereby, the use of ChatGPT when different languages are utilized will not be a limitation. The reduced accuracy in some subspecialties needs further review due to the nature of some questions (as above) and the reasoning required. In addition, our results highlight the high capability of ChatGPT in translation, where ChatGPT can translate high-level medical texts automatically without changing the concept and originality of questions.

Overall, the existing pieces of evidence hint at the potential role of ChatGPT in medical education and medical research, alongside more conventional uses such as acting as a medical information retrieval system for patients and clinicians, clinical decision support system, patient education, mental health support, and telemedicine and triage [23,29,30]. However, there are limitations in all of the studies published until now, which could complicate the routine implementation of

ChatGPT in clinical practice, including the small number of questions in many subspecialty fields, the accuracy of the answer key not being specifically validated, not assess time to complete the exam in the different languages, variability of correct answers depending upon geography i.e. the leading causes of death in Iran and France may be different in epidemiological type questions such as this, did not do reproducibility studies having the program test and retest for any changes over time, and the possibility that the BOT may become more or less accurate with more data fed into it with time.

Furthermore, one of the most critical issues is the non-standardized application of the tool in different fields, which limits generalizability. This is particularly important for educational purposes, as the tool could always function as a two-edged sword, providing misleading information or insufficient material [8]. Importantly, ChatGPT may not be trained on holistic data, especially with regard to professional texts, as demonstrated by the potential biases of chatbots in certain topics, such as political sciences. Thus, it is possible that the data fed to the model may not represent the full gamut of medical literature, or certain controversial medical topics may not have been disclosed to the chatbot. Furthermore, the chatbot can provide skewed answers to particular sensitive medical topics which may not represent accurate and pertinent medical information [31–33].

Future studies should focus on how the chatbot performs when faced with questions with varying taxonomies, or clinical vignettes. Scholars should also aim to devise methodologies that aim at increasing the reproducibility of answers provided by language models. Another important set of barriers that need to be overcome are those concerning ethical dilemmas when using language models. A significant portion of health information generated and shared could be sensitive and protected by federal and local rules and regulations, necessitating measures to protect such information (examples being information protected by the Health Insurance Portability and Accountability Act) [9]. Furthermore, routine utilization of ChatGPT in medical research and education faces challenges with plagiarism, as many previous studies have shown that ChatGPT and similar chatbots routinely fail to reference their resources properly (31).

5. Conclusion

ChatGPT scores are consistently good in four different languages, answering 161 (81.3%), 167 (84.3%), 162 (81.8%), and 166 (83.8%) questions correctly in Persian, English, French, and Spanish, respectively. Analysis showed that there were no significant differences in the languages in the perception and performance of ChatGPT, suggesting the role ChatGPT may have in medical education irrespective of language. ChatGPT worked particularly well in questions pertaining to topics with more focus on knowledge than patient-centered care, necessitating further research into how chatbots could further contribute to medical education and research, which requires more soft, interpersonal skills.

Funding

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

Declaration of competing interest

None.

Acknowledgement

None.

References

[1] Castro EMJ, Faria Araújo NM. Impact of the fourth industrial revolution on the health sector: a qualitative study. Healthcare informatics research 2020;26:328–34.
[2] Javan R, Kim T, Mostaghni N, Sarin S. ChatGPT's potential role in interventional radiology. Cardiovasc Intervent Radiol 2023;46:821–2.
[3] Das A, Selek S, Warner AR, Zuo X, Hu Y, Keloth VK, Li J, Zheng WJ, Xu H. Conversational bots for psychotherapy: a study of generative transformer models using domain-specific dialogues. Proceedings of the 21st Workshop on Biomedical Language Processing 2022:285–97.
[4] Tustumi F, Andreollo NA, Aguilar-Nascimento JE. Future of the language models in healthcare: the role of CHATGPT. Arquivos brasileiros de cirurgia digestiva : ABCD = Brazilian archives of digestive surgery 2023;36:e1727.
[5] Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. Croat Med J 2023;64:1–3.
[6] Javan R, Kim T, Mostaghni N, Sarin S. ChatGPT's potential role in interventional radiology. Cardiovasc Intervent Radiol 2023;46:821–2.
[7] Garg RK, Urs VL, Agrawal AA, Chaudhary SK, Paliwal V, Kar SKJm. Exploring the Role of Chat GPT in patient care (diagnosis and Treatment) and medical research. Syst Rev 2023;2023:23291311. 2006. 2013.
[8] Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - reshaping medical education and clinical management. Pakistan J Med Sci 2023;39:605–7.
[9] Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Frontiers in artificial intelligence 2023;6:1169595.
[10] Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, Fournier S. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? European heart journal. Digital health 2023;4:279–81.
[11] Gandomkar R, Changiz T, Omid A, Alizadeh M, Khazaei M, Heidarzadah A, Rouzrokh P, Amini M, Honarpisheh H, Laripour RJBME. Developing and validating a national set of standards for undergraduate medical education using the WFME framework. the experience of an accreditation system in Iran 2023;23:1–14.
[12] Collins J. Education techniques for lifelong learning: writing multiple-choice questions for continuing medical education activities and self-assessment modules, 26. Radiographics : a review publication of the Radiological Society of North America, Inc; 2006. p. 543–51.
[13] Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR medical education 2023;9:e45312.
[14] Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS digital health 2023;2:e0000198.
[15] Khoshpouri P, Mohseni A, Dabiri M, Ansari G, Zadeh FS, Ataeinia B, Saadat N, Sherbaf FG, Yousem DMJAr. International medical graduates in radiology residencies: demographics, performance, and visa issues. 2023.
[16] Shams F, Shams A. Emigration: an opportunity for Iranian physicians, a challenge for the government. Lancet (London, England) 2014;383:1039.
[17] Baethge C. The languages of medicine. Deutsches Arzteblatt international 2008;105:37–40.
[18] Pascual-Leone N, Liu JW, Beschloss A, Chenna SS, Saifi C. The language of all medical publications and spine publications from 1950 to 2020. North American Spine Society Journal (NASSJ) 2022;10:100118.
[19] Pascual-Leone N, Liu JW, Beschloss A, Chenna SS, Saifi C. The language of all medical publications and spine publications from 1950 to 2020. North American Spine Society journal 2022;10:100118.
[20] .
[21] Vice Chair of Education IMoH. Clinical resideny EExamination guide handbook. 2023.
[22] Gharebaghi R, Heidary F, Pourezzat AAJ. Serial deaths of young trainee physicians in Iran during COVID-19 pandemic; messages to policy makers. Frontiers in health services 2022;2:19.
[23] Li J, Dada A, Kleesiek J, Egger JJm. ChatGPT in healthcare: a taxonomy and systematic review. 2023. p. 2023. 2003. 2030.23287899.
[24] Lourenco AP, Slanetz PJ, Baird GLJR. Rise of ChatGPT: it may Be time to reassess how we teach and test radiology residents. Radiological Society of North America; 2023, 231053.
[25] Elsayed SJapa. Towards mitigating ChatGPT's negative impact on education: optimizing question design through bloom's taxonomy. 2023.
[26] Bhayana R, Krishna S, Bleakney RRJR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. 2023, 230582.
[27] Drake E, Phillips JP, Kovar-Gough I. Exploring preparation for the USMLE step 2 exams to inform best practices. PRiMER (Leawood, Kan.) 2021;5:26.
[28] Arzani A, Lotfi M, Abedi AJ. Experiences and clinical decision-making of operating room nurses based on benner's theory. Journal of Babol University Of Medical Sciences 2016;18:35–40.
[29] Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns, 11. Basel, Switzerland): Healthcare; 2023.
[30] Dave T, Athaluri SA, Singh SJ. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Frontiers in Artificial Intelligence 2023;6:1169595.

[31] Ferrara EJapa. Should chatgpt be biased? challenges and risks of bias in large language models. 2023.
[32] Ray PPJ, Systems C-P. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems 2023;3:121–54.
[33] Babaei N, Zamanzadeh V, Valizadeh L, Lotfi M, Samad-Soltani T, Kousha A, Avazeh MJH. A scoping review of virtual care in the health system: infrastructures, barriers, and facilitators. Home Health Care Serv Q 2023;42:69–97.