Original Article
http://mjiri.iums.ac.ir
Medical Journal of the Islamic Republic of Iran (MJIRI)

Med J Islam Repub Iran. 2025 (11 Feb);39.24. https://doi.org/10.47176/mjiri.39.24

ChatGPT's Performance on Iran's Medical Licensing Exams

Alireza Keshtkar[1], Ali-Asghar Hayat[2], Farnaz Atighi[3], Nazanin Ayare[3], Mohammadreza Keshtkar[4], Parsa Yazdanpanahi[3], Erfan Sadeghi[5], Noushin Deilami[3], Hamid Reihani[3], Alireza Karimi[3], Hamidreza Mokhtari[3], Mohammad Hashem Hashempur[6]*

Received: 14 Jun 2024 Published: 11 Feb 2025

Abstract

Background: A 175 billion parameter transformer architecture is used by OpenAI's ChatGPT language model to perform tasks requiring natural language processing. This study aims to evaluate the knowledge and interpretive abilities of ChatGPT on three types of Iranian medical license exams: basic sciences, pre-internship, and pre-residency.

Methods: This comparative study involved administering three different levels of Iran's medical license exams, which included basic sciences, pre-internship, and pre-residency, to ChatGPT 3.5. Two versions of each exam were used, corresponding to the ChatGPT 3.5's internet access time: one during the access time and one after. These exams were inputted to ChatGPT in Persian and English. The accuracy and concordance of each question were extracted by two blinded adjudicators.

Results: A total of 2210 questions, including 667 basic sciences, 763 pre-internship, and 780 pre-residency questions, were presented to ChatGPT in both English and Persian languages. Across all tests, the overall accuracy was found to be 48.5%, with an overall concordance of 91%. Notably, English questions exhibited higher accuracy and concordance rates, with 61.4% accuracy and 94.5% concordance, compared to 35.7% accuracy and 88.7% concordance for Persian questions.

Conclusion: Our findings demonstrate that ChatGPT performs above the required passing scores on basic sciences and pre-internship exams. Moreover, ChatGPT could obtain the minimal score needed to apply for residency positions in Iran; however, it was lower than the applicants' mean scores. Significantly, the model showcases its ability to provide reasoning and contextual information in the majority of responses. These results provide compelling evidence for the potential use of ChatGPT in medical education.

Keywords: Medical education, Chat GPT, Artificial intelligence, Iran

Conflicts of Interest: None declared
Funding: This study was financially supported by Shiraz University of Medical Sciences.

Cite this article as: Keshtkar A, Hayat AA, Atighi F, Ayare N, Keshtkar M, Yazdanpanahi P, Sadeghi E, Deilami N, Reihani H, Karimi A, Mokhtari H, Hashempur MH. ChatGPT's Performance on Iran's Medical Licensing Exams. *Med J Islam Repub Iran.* 2025 (11 Feb);39:24. https://doi.org/10.47176/mjiri.39.24

Introduction

The rapid advancements in digital health have been greatly facilitated by the emergence of large language models (LLMs). These LLMs are large parameter space deep neural network models. These billion-parameter models are frequently trained using gigabytes or terabytes of text data. LLMs are a significant advancement in artificial intelligence (AI), opening up new possibilities for natural language processing and generation (1, 2).

An LLM Generative Pretrained Transformer (GPT) called ChatGPT (OpenAI; San Francisco, CA) was created to serve as a "chatbot" (1). With the appearance of this model, a sophisticated LLM was available to the general

↑*What is "already known" in this topic:*
Artificial intelligence has lots of applications in clinical practice and medical education. Nowadays, large language models, especially ChatGPT, are widely assessed in the standard medical examination to evaluate the knowledge of this model in medicine. However, there is limited knowledge of Iran's medical licensing exams.

→*What this article adds:*
We showed that ChatGPT can pass Iran's medical exams, including basic science and pre-internship, and obtain the minimal score needed to apply for residency positions in English. The accuracy and concordance were significantly different between the Persian and English languages.

Corresponding author: Dr Mohammad Hashem Hashempur, hashempur@gmail.com

[1] Research Center of Noncommunicable Diseases, Jahrom University of Medical Sciences, Jahrom, Iran
[2] Clinical Education Research Center, Department of Medical Education, School of Medicine. Shiraz University of Medical Sciences. Shiraz, Iran
[3] Student Research Committee, Shiraz University of Medical Sciences, Shiraz, Iran
[4] Student Research Committee, Yazd University of Medical Sciences, Yazd, Iran
[5] Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran
[6] Research Center for Traditional Medicine and History of Medicine, Department of Persian Medicine, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran

public in a user-friendly style for the first time. Recently, ChatGPT's performance has been evaluated through challenging tasks. For instance, standardized tests were used to compare the algorithm's performance against human participants intended for these tests. Despite lacking domain-specific training, ChatGPT has demonstrated remarkable performance, consistently achieving scores that are at or close to passing or threshold scores of postgraduate levels of specialization across various fields, including medicine, as evidenced by the successful completion of university multiple-choice questions (MCQ) exams of the Chinese National Medical Licensing Examination and the United States Medical Licensing Examinations (USMLE). In addition to evaluating LLM performance in medical licensing, these models have been explored in the medical field to provide personalized patient interaction and educate consumers about their health (3-6).

One of ChatGPT's main draws is its ability to understand the context and engage in meaningful, relevant conversations regarding the subject at hand (7). Numerous studies have been conducted on ChatGPT to showcase its potential application in the medical sector. In leveraging AI to produce precise and validated information for patients and the general public, it becomes crucial for medical students and healthcare professionals to assess the accuracy of AI-generated medical information (8-10). Therefore, it is of utmost importance to ascertain ChatGPT's proficiency in accurately answering medical examination questions (11).

Conventional medical education often relies on lectures and passive learning. Moreover, there is a reliance on standardized testing, which may not accurately reflect a student's abilities or readiness for practice. Additionally, the inability of medical students to engage in self-directed learning strategies can further complicate their learning process (12, 13). These challenges may leave medical graduates unprepared for healthcare demands. This has led to growing interest in using advanced language models like ChatGPT to transform traditional educational methods by providing interactive and personalized learning experiences (14). However, by research with ChatGPT, we recognized the potential impact of input language on the model's output accuracy (15). This prompted us to investigate the performance of ChatGPT in Persian medical exams and its understanding of Persian medical terminology, aiming to offer valuable insights for non-English speaking medical students looking to utilize this tool effectively in their studies.

Additionally, in our pursuit to enhance the utility of ChatGPT for medical students, we sought to assess the concordance of each question in a large sample size. By rigorously evaluating the model's accuracy across a diverse range of medical questions, our study aims to offer clear guidance on the optimal utilization of ChatGPT in the context of medical education.

Methods

Data inputting

This comparative study involved administering three different Iran medical license exams, which included basic sciences (BS), pre-internship, and pre-residency, to the web interface of ChatGPT 3.5 (OpenAI; San Francisco, CA) with the seven classical parameters including temperature, Top-p, Max Tokens, Frequency Penalty, Presence Penalty, Stop Sequences, and Number of Responses left at their default settings.

Two time-zone versions of each exam, corresponding to the ChatGPT 3.5's internet access time, which is limited to September 2021, were used: one from the exams organized before September 2021 and one after. All questions and their correct answers were sourced from the official website of the Iranian Ministry of Health and Medical Education. All questions were presented to ChatGPT in Persian language without any modifications and were in MCQ format. Additionally, the questions were translated into English using Google Translate® online and then submitted to ChatGPT without any correction. The study excluded questions that had pictures, as they could potentially distort ChatGPT's performance. To minimize memory retention bias, a new chat session was initiated for each question. At the beginning of every chat session, the questioner introduced himself as a medical student, and then the questions were input into the ChatGPT interface. No specific prompting strategies to elicit direct answers or explanations from the model were incorporated. Then, all the questions and answers were copied and delivered to two adjudicators to judge accuracy and concordance.

Adjudication and Bias Mitigation Strategies

We used a method of adjudication similar to Kung et al.(5). The criteria used for determining the accuracy were as follows:

- Accurate: The final answer matches the key of the national organization for educational testing.

- Inaccurate: Incorrect answer choice is selected, AI output is not an answer choice, AI returns a, or AI determines that not enough information is available.

For concordance, the criteria were:

- Concordant: Explanation affirms the answer.

- Discordant: Any part of the explanation contradicts itself.

To address the potential risk of bias and subjectivity, we implemented several mitigation strategies. We conducted two separate training sessions to ensure that the method of the study and the criteria of adjudication were clear to the adjudicators. Additionally, we used two students with the highest grades, who were blinded to each other, to determine the accuracy and concordance of the model's output. We employed blinding in the training sessions and ensured that the adjudicators were blinded to each other to reduce subjectivity. In cases where there were differences in accuracy and concordance judgment, accuracy was rechecked by using the answers provided by the Iran Ministry of Health and Medical Education, and the senior author arbitrated the discrepancies in concordance. Notably, the fact that only 20 questions in our large sample size needed to be arbitrated by the senior author demonstrated the low risk of subjectivity in our method.

Statistical analysis

Statistical analysis was performed by Statistical Package for Social Sciences (SPSS Inc. Released 2009. PASW Statistics for Windows, Version 18.0. Chicago: SPSS Inc.). Descriptive data were reported as numbers and percentages. Binary logistic regression and Chi-square tests were used to compare the variables of the study. A P-value less than 0.05 was considered statistically significant.

Results

A total of 2210 questions were given to ChatGPT, including 667 basic sciences (BS), 763 pre-internship (PI), and 780 pre-residency (PR) questions. Total accuracy was 48.5%, and total concordance was 91% among all examinations. The accuracy and concordance for English questions were 61.4% and 94.5%, respectively, and 35.7% and 88.7 % for Persian questions (Figure 1).

ChatGPT performed best in the BS exam with 72% accuracy in English. The lowest accuracy was in the PR exam, with 32.3% in Persian. The best concordance was in the BS exam with 98.2%, and the lowest was in the PR exam, with 84.1%. Detailed demographic data are presented in Table 1.

There was no significant difference in accuracy and concordance for the mentioned time zones (BS: 0.885-0.122, PI: 0.628-0.347, and PR: 0.825-0.065). We found significant differences in accuracy and concordance between Persian and English questions in all exams (BS accuracy: P-value < 0.001; OR (95%C.I.): 1.024 (0.741-1.415), BS concordance: P-value = 0.005; OR (95%C.I.): 0.612 (0.239-1.139), PI accuracy: $P < 0.001$; OR (95%C.I.): 2.531 (1.889-3.391), PI concordance: P-value = 0.002; OR (95%C.I.): 2.522 (1.403-4.534), PR accuracy: $P < 0.001$, OR (95%C.I.): 2.435 (1.825-3.249), PR concordance: P-value = 0.012; OR (95%C.I.): 1.826 (1.143-2.917) (Table 2 and Table 3).

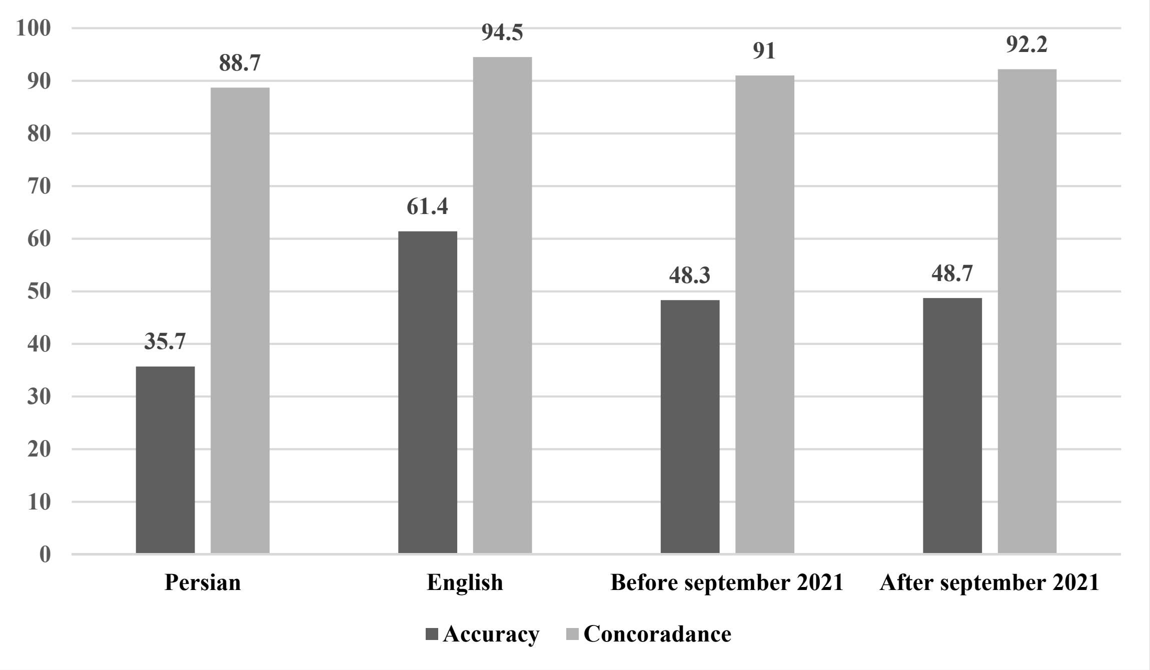

Figure 1. Accuracy and Concordance of all questions in different languages and time zones

Table 1. Descriptive characteristics of the study

Time zone	Exam	Language	Accuracy		Concordance	
			Accurate	Inaccurate	Concordant	Discordant
			Count (%)	Count (%)	Count (%)	Count (%)
Before September 2021	Basic science	Persian	56 (32.7)	115 (67.3)	156 (91.2)	15 (8.8)
		English	123 (72.4)	47 (27.6)	167 (98.2)	3 (1.8)
	Pre-internship	Persian	65 (34.0)	126 (66)	168 (88)	23 (12)
		English	110 (57.3)	82 (42.7)	183 (95.3)	9 (4.7)
	Pre-residency	Persian	77 (39.5)	118 (60.5)	164 (84.1)	31 (15.9)
		English	107 (54.9)	88 (45.1)	176 (90.3)	19 (9.7)
After September 2021	Basic science	Persian	64 (39.3)	99 (60.7)	146 (89.6)	17 (10.4)
		English	109 (66.9)	54 (33.1)	153 (93.9)	10 (6.1)
	Pre-internship	Persian	69 (36.3)	121 (63.7)	173 (91.1)	17 (8.9)
		English	111 (58.4)	79 (41.6)	182 (95.8)	8 (4.2)
	Pre-residency	Persian	63 (32.3)	132 (67.7)	173 (88.7)	22 (11.3)
		English	118 (60.5)	77 (39.5)	183 (93.8)	12 (6.2)

Table 2. Binary logistic regression for questions accuracy

Exams	Predictor	*P*-value	OR (95% C.I.)
Basic-science	Time (After Sep 2021 vs. Before)	0.885	1.024 (0.741-1.415)
	Language (English vs. Persian)	<0.001	4.096 (2.965-5.66)
Pre-internship	Time (After Sep 2021 vs. Before)	0.628	1.075 (0.803-1.44)
	Language (English vs. Persian)	<0.001	2.531 (1.889-3.391)
Pre-residency	Time (After Sep 2021 vs. Before)	0.825	0.968 (0.726-1.292)
	Language (English vs. Persian)	<0.001	2.435 (1.825-3.249)

Table 3. Binary logistic regression for the concordance of Chat-GPT answers

Exams	Predictor	*P*-value	OR (95% C.I.)
Basic science	Time (After Sep 2022 vs. Before)	0.122	0.612 (0.239-1.139)
	Language (English vs. Persian)	0.005	2.619 (1.347-5.09)
Pre-internship	Time (After Sep 2022 vs. Before)	0.347	1.301 (0.752-2.248)
	Language (English vs. Persian)	0.002	2.522 (1.403-4.534)
Pre-residency	Time (After Sep 2022 vs. Before)	0.065	1.545 (0.973-2.453)
	Language (English vs. Persian)	0.012	1.826 (1.143-2.917)

Table 4. Chi-square test for accuracy and concordance correlation

	Accuracy		*P*-value
	Inaccurate, N (%)	Accurate, N (%)	
Discordant	153 (82.3)	33 (17.7)	<0.001
Concordant	985 (48.7)	1039 (51.3)	

As shown in Table 4, 82.3% of the discordant questions were inaccurate, and 51% of the concordant questions were accurate. We found a statistically significant relationship between concordance and accuracy ($P < 0.001$).

Discussion

With the development of ChatGPT, a crucial turning point in the field of conversational AI has emerged, in which all aspects of society can be influenced, particularly the field of medicine, from research to teaching (16, 17).

The question sets used in this study included Iran’s three medical exams: BS, PI, and PR. The aforementioned exams include a series of standardized MCQs measuring both basic science knowledge and clinically-based problems requiring analytical thinking. ChatGPT performed above the required passing scores on the BS and PI exams. However, ChatGPT obtained lower scores in the PR exam, which is highly competitive with limited positions available, compared to applicants' mean scores, but it could achieve the minimal score required for application. Furthermore, ChatGPT was notably unaffected by the examination's time zone, despite it being presumed that ChatGPT would perform better on tests taken before September 2021, as the available data for ChatGPT 3.5 is restricted to September 2021.

The use of two languages in asking this chatbot's inquiries yielded an interesting finding. ChatGPT was unable to successfully pass any of the medical license exams in Persian, including both pre-September and post-September. A study by Khorshidi et al. on the 2023 Iranian residency entrance examination revealed ChatGPT 4’s excellent support for diverse language input. It was able to pass the PR exam by a score of 81.3% (18). The high accuracy of the Khorshidi et al. article could be a result of a potential memory retention bias that was not considered in their valuable article and the model's newer version. Nevertheless, the accessibility of ChatGPT 4 remains a concern, particularly for students in developing countries who often find it financially challenging. The significant difference between English and Persian in our study can be explained by the limited available Persian internet textual data compared to the English context. Moreover, it suggests that banned access to or limited acceptance of this technology in countries speaking Persian may have led to less data being generated from Persian interactions, which is essential for improving the model's language comprehension. Moreover, it is implied that Google Translate is successful in translating medical Persian texts while preserving core concepts. This suggests that utilizing Google Translate for English translation as the input language for ChatGPT can be an alternative option for individuals unable to afford or access newer versions of ChatGPT with better Persian language support.

The study points to a significant connection between the concordance and accuracy of the ChatGPT's responses. The ChatGPT frequently offers a logical and reasonable justification for the response choice when a question is correctly answered. Low concordance in incorrectly answered questions demonstrates that the chatbot has failed to understand the gist of the query. We did not investigate the questions with low concordance, which is a limitation of our study. However, since low concordance indicates that a model's capacity for analysis of the response is low, it is anticipated that queries with low concordance will focus more on analytical thinking than on scientific facts, which are easily accessible information for this model. A critical point worth mentioning is that using ChatGPT in practice when analytical thinking is needed, without the supervision of an expert medical consultant, may lead to life-threatening situations (19-22). As seen in our results, in the case of questions in which the presented patient had a case of multiple diseases, this model focused on less

[DOI: 10.47176/mjiri.39.24]

critical chief complaints and missed serious conditions.

ChatGPT can be applied to the field of medical education. According to the findings of previous studies and our research, ChatGPT's medical knowledge is becoming reliable, and its clinical reasoning ability is comparable to that of human medical learners. Additionally, the model's ability to converse in a human-like style makes it an interactive educational tool, even for medical students who do not speak English (5, 23). As an interactive virtual assistant, ChatGPT can facilitate personalized learning experiences by analyzing students' strengths and weaknesses, thereby tailoring educational content to meet individual needs. Furthermore, its ability to generate questions and simulate clinical scenarios allows for self-assessments. By providing immediate feedback on their assessments, ChatGPT can help students track their learning progress effectively. Additionally, students can practice their communication and diagnostic skills in a risk-free environment, preparing them for real-world patient interactions (24, 25). For instance, a medical student could submit a personal clinical case to ChatGPT and ask questions about its different aspects. The process of asking questions can spark further inquiry. By comparing the model's responses with his answers, the student can recognize his knowledge gaps and acquire additional information to supplement his existing understanding.

While the study demonstrates admirable accuracy (48.5%) for the model, it also shows that the glass is half empty, and it is essential to consider both the positive and negative consequences of relying on AI models for medical training (15). The new generation, which is mixed with this technology, cannot be prevented from using it (7). We suggest teaching students the correct way of using this model, not as a teacher for asking questions, but as an informed, quick study partner who is not safe from making mistakes to learn concepts more deeply.

In addition to these considerations, the application of ChatGPT in the context of medical education raises important ethical considerations warranting careful examination (26-28). It should be noted that ChatGPT, while proficient in generating human-like responses and providing educational support, is not a substitute for the expertise and judgment of licensed medical professionals. As such, a clear delineation of the roles and responsibilities of ChatGPT in medical education is essential to ensure that its use complements and enriches traditional teaching and assessment methodologies rather than supplanting them. Furthermore, the reliance on ChatGPT's medical advice without human verification may lead to inaccurate recommendations, potentially harming patients (29). In light of these ethical considerations, medical educators, practitioners, and policymakers must engage in establishing clear guidelines, ethical frameworks, and regulatory standards for the responsible use of AI in medical education, which are essential to mitigate potential risks and ensure the ethical and responsible integration of AI technologies in the medical field.

Limitations

There are certain limitations to this study. First, we utilized a single model, and the results are not generalizable for newer versions or other LLM chatbots. Even for the same LLM, the findings may differ by the time the research is published since LLMs can improve their function through ongoing user feedback. Moreover, by modifying model default parameters and incorporating specific prompting strategies, such as the chain of thoughts (CoT), the model performance can improve, and the results may differ. We did not customize the model, considering that most users may not be familiar with the parameters and prompting strategies, but it should be in mind as a potential limitation for the model's performance when interpreting the results. Another limitation that should be noted when interpreting the performance of the model in Persian is the scarcity of readily available Persian medical resources on the internet, which can result in insufficient representation of Persian medical terminology within Google Translate and models like ChatGPT. Furthermore, we did not consider the taxonomy of the questions, which helped determine ChatGPT's accuracy in handling more complex cases. We recommend future studies to incorporate this for a more precise evaluation.

Conclusion

Our findings demonstrate that ChatGPT performs above the required passing scores on basic sciences and pre-internship exams in the English language. Moreover, ChatGPT could obtain the minimal score needed to apply for residency positions in Iran in the English language; however, it was lower than the applicants' mean scores. Moreover, the model showcases its ability to provide reasoning and contextual information in the majority of responses, owing to its dialogic character when addressing inquiries.

Authors' Contributions

Design and conceptualization: A.K, A.H, M.H.H, Data collection: P.Y, A.Ka, H.M; Adjudication: N.D, H.R, Analysis, and interpretation: E.S; drafting the article and review: F.A, N.A, M.K., M.H.H.

Ethical Considerations

This study was approved by the Ethics Committee of Shiraz University of Medical Sciences (ethics code: IR.SUMS.MED.REC.1402.312).

Acknowledgment

Not applicable.

Conflict of Interests

The authors declare that they have no competing interests.

References

1. Hammer A. ChatGPT Can Pass the US Medical Licensing Exam and the Bar Exam [Internet]2023 January 23, 2023. Available from: https://www.dailymail.co.uk/news/article11666429/ChatGPT-pass-United-States-Medical-Licensing-Exam-Bar-Exam.html.
2. Lakshmi V. ChatGPT is on its way to becoming a virtual doctor, lawyer, and business analyst. Here'sa list of advanced exams the AI bot has passed so far. Businessinsider. 2023.

3. Varanasi L. ChatGPT is on its way to becoming a virtual doctor, lawyer, and business analyst Here's a list of advanced exams the AI bot has passed so far [Internet]2023. Available from: https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1.
4. <ChatGPT and exams Brief Report FINAL (1).pdf>.
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepano C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.
6. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health. 2023;2(2):e0000205.
7. Wu T, He S, Liu J, Sun S, Liu K, Han Q-L, et al. A brief overview of ChatGPT: The history, status quo and potential future development. IEEE/CAA Journal of Automatica Sinica. 2023;10(5):1122-36.
8. Atighi F, Yazdanpanahi P, Keshtkar A, Karimi A, Naseri A, Dabbaghmanesh MH. Illuminating the Path to Thyroid Disorder Management Using Artificial Intelligence: A Narrative Review. Shiraz E-Medical Journal. (In Press).
9. Keshtkar A, Ayareh N, Atighi F, Hamidi R, Yazdanpanahi P, Karimi A, et al. Artificial intelligence in diabetes management: Revolutionizing the diagnosis of diabetes mellitus; A literature review. Shiraz E‑Medical J. 2024;25:e146903.
10. Yazdanpanahi P, Atighi F, Keshtkar A, Hamidi R, Rezaeimanesh M, Karimi A, et al. The Current Progress of Artificial Intelligence in Approach to Thyroid Nodules: A Narrative Review. Shiraz E-Medical Journal.25(11).
11. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. Radiological Society of North America; 2023. p. e230163.
12. Majumder MAA, Haque M, Razzaque MS. Trends and challenges of medical education in the changing academic and public health environment of the 21st century. Frontiers in Communication. 2023;8:1153764.
13. Heydari M, Hashempur M, Shams M. Inappropriate Time Splitting Among Endocrine Topics in Undergraduate Medical Education. Education for Health. 2012;25(2):131-2.
14. Ahmed Y. Utilization of ChatGPT in Medical Education: Applications and Implications for Curriculum Enhancement. Acta informatica medica : AIM : journal of the Society for Medical Informatics of Bosnia & Herzegovina : casopis Drustva za medicinsku informatiku BiH. 2023;31(4):300-5.
15. Keshtkar A, Atighi F, Reihani H. Systematic review of ChatGPT accuracy and performance in Iran's medical licensing exams: A brief report. Journal of Education and Health Promotion. 2024;13(1):421.
16. Philipp Hacker AE, Marco Mauer. Regulating ChatGPT and other Large Generative AI Models. arxiv. May 12, 2023;v8.
17. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ. 2019;7:e7702.
18. Khorshidi H, Mohammadi A, Yousem DM, Abolghasemi J, Ansari G, Mirza-Aghazadeh-Attari M, et al. Application of ChatGPT in multilingual medical education: How does ChatGPT fare in 2023's Iranian residency entrance examination. Informatics in Medicine Unlocked. 2023;41.
19. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE, Miller V. High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. Cureus. 2023;15(5).
20. Sallam M, editor ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare; 2023: MDPI.
21. Chow JC, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. Frontiers in Artificial Intelligence. 2023;6:1166014.
22. Mijwil M, Aljanabi M, Ali AH. Chatgpt: Exploring the role of cybersecurity in the protection of medical information. Mesopotamian journal of cybersecurity. 2023;2023:18-21.
23. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312.
24. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. Pakistan journal of medical sciences. 2023;39(2):605-7.
25. Seetharaman R. Revolutionizing Medical Education: Can ChatGPT Boost Subjective Learning and Expression? J Med Syst. 2023;47(1):61.
26. Yu H. Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. Frontiers in Psychology. 2023;14:1181712.
27. Cotton DR, Cotton PA, Shipway JR. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International. 2023:1-12.
28. Nikolic S, Daniel S, Haque R, Belkina M, Hassan GM, Grundy S, et al. ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. European Journal of Engineering Education. 2023:1-56.
29. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Muller BP, Raptis DA, et al. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. J Med Internet Res. 2023;25:e47479.

[DOI: 10.47176/mjiri.39.24]