COMPARATIVE EVALUATION OF ARTIFICIAL INTELLIGENCE AND MEDICAL STUDENTS IN MEDICAL KNOWLEDGE ASSESSMENT: INSIGHTS FOR ENHANCING MEDICAL EDUCATION

Bakhodir B. Rakhimov [1], Erkin Yo. Sultonov [2], Dilafruz A. Turabayeva [3], Sabina K. Sodikova [4], Fatima Z. Alimova[5]

1 DSc, associated professor of Tashkent Medical Academy, Tashkent, Uzbekistan
E-mail: rakhimov.b.b@gmail.com

2 DSc, associated professor of Tashkent Medical Academy, Tashkent, Uzbekistan

3 Zarmed University, Samarkand, Uzbekistan
E-mail: turabaevadila5@gmail.com

4 Medical College of Kimyo International University in Tashkent, Uzbekistan
E-mail: sabinasadikova2001@gmail.com

5 Center for the development of professional qualification of medical workers, Tashkent, Uzbekistan
E-mail: alimovakrom63@yandex.ru

ABSTRACT

This study aims to compare the effectiveness of artificial intelligence (AI) tools, specifically ChatGPT and Bing, with that of local and international medical students in understanding and responding to medical questions. The goal is to evaluate the potential role of AI in medical education and identify knowledge gaps among different student groups. The study highlights the potential of AI tools to supplement medical education by providing accurate information and aiding in learning.

Key words: artificial intelligence (AI), medical education, ChatGPT, Bing, medical students, knowledge assessment, clinical reasoning, educational tools.

INTRODUCTION

Advances in artificial intelligence (AI) have increasingly influenced various fields, including healthcare and medical education [1,2]. AI tools such as ChatGPT and Bing have demonstrated significant capabilities in processing natural language,

retrieving information, and providing insights across a wide range of topics. As these technologies evolve, there is growing interest in their potential role as educational aids in the medical field, where accurate knowledge and critical thinking are essential [3].

Medical education is a complex process that requires the integration of theoretical knowledge with practical skills. Traditional teaching methods, including lectures, textbooks, and clinical rotations, have long been the cornerstone of training healthcare professionals. However, these methods often vary widely between institutions and countries, leading to differences in educational quality and outcomes among medical students. This variation can impact the preparedness of future healthcare professionals to meet the challenges of modern medical practice [4,5].

AI tools offer a promising supplement to traditional education by providing students with immediate access to a vast array of medical information and the ability to simulate clinical reasoning and decision-making [6,7]. Despite their potential, the effectiveness of these tools in medical education remains under investigation. There is a need to evaluate how well AI tools perform in comparison to human learners, particularly in interpreting medical scenarios and applying clinical knowledge [8,9].

This study aims to explore the effectiveness of AI tools in understanding and responding to medical questions compared to local and international medical students. By assessing the correct response rates across a range of medical topics, this study seeks to identify the strengths and limitations of AI tools and highlight differences in knowledge and understanding among different groups of students. The findings will provide insights into the potential role of AI in medical education and help educators understand how to better integrate these technologies into their teaching strategies.

Methods

A cross-sectional study was conducted to evaluate the understanding of medical concepts among three groups: Artificial Intelligence (AI) tools (ChatGPT and Bing), local students, and international students. The AI tools (n=2) and students (n=41) responded to 16 questions designed to assess their knowledge of various medical and physiological phenomena. The student respondents were divided into two groups: local students (n=31) and international students (n=10).

Questionnaire Design

The questionnaire comprised 16 questions (10 test and 6 situational tasks) covering a range of medical topics, including clinical subjects, anatomy, physiology, pathology, and medical ethics. The questions were crafted to assess the

ability of respondents to correctly interpret clinical scenarios and apply theoretical knowledge to practical situations:

1. The patient consulted a doctor complaining of nasal congestion. Radiological examination revealed deviation of the nasal septum and surgery was recommended. Which structures of the nasal septum should undergo surgical correction?

2. A student, experiencing mental distress after taking a physiology exam, unexpectedly tells her classmates, "Maybe this sounds strange, but I want to eat." What is the physiological significance of this phenomenon?

3. A dog licks wounds and areas of suppuration. What is the physiological significance of this licking?

4. Why does diuresis decrease during nighttime?

5. A gunshot wound caused acute bleeding. How will this condition affect the hematopoietic activity of the bone marrow and the number of blast cells?

6. "A man is born in a shirt." What "shirt" is referred to in the proverb and what is the source of development of the epithelium lining it from the inside?

7. As a result of a viral infection, pseudounipolar neurons of the spinal ganglia perished. Which link of the reflex arc is switched off?

8. What relates to non-palpable elements of a skin rash?

9.Normally palpable lymph nodes include?

10. When can an employer obtain information about an employee's illness?

11. A pathological condition characterized by inflammation of the membranes of the brain and spinal cord. Diagnosis and treatment of this condition are managed by neurologists. It's a dangerous disease, with primary symptoms including severe headache, increased body temperature, and increased tension of the neck muscles. The inflammatory process itself is caused by viruses, bacteria, fungi, or protozoa. What is this disease?

12. Typical manifestations of this disease in the early stages include general intoxication and pain syndrome. The patient suffers from nausea, weakness, subfebrile temperature, and chills. As the disease progresses, diarrhea may occur. The pain syndrome is pronounced. Initially, the pain is localized in the upper abdomen, then moves to the periumbilical area and right side, and pulls the legs toward the abdomen. The cause of the disease is the penetration of pathogenic bacteria into the lumen of the inflamed area. What is this disease?

13. Different forms of this disease differ in the area of skin lesions, features of rashes, and other symptoms. If inflammation is caused by a specific adverse effect, skin changes occur in the area of contact with the irritant. Severe forms of this disease manifest with necrosis and ulceration of the skin, may be complicated by

infection, leading to increased body temperature and general malaise. What is this disease?

14. A pathological condition characterized by acute disruption of blood supply to the brain and damage to organ tissues. Diagnosis and treatment of this disease are managed by neurologists. The patient exhibits signs of impaired brain function, including consciousness disorders and mobility impairments. Damage to major arteries leads to more severe consequences. What is this disease?

15. A pathological condition characterized by acute disruption of blood supply to the brain and damage to organ tissues. Diagnosis and treatment of this disease are managed by neurologists. The patient exhibits signs of impaired brain function, including consciousness disorders and mobility impairments. Damage to major arteries leads to more severe consequences. What is this disease?

16. The main route of disease transmission is contact-based. The symptoms of the disease are directly related to the life activities of female ticks, which localize in the epidermal layers. Dermatoscopy allows for the detection of objective signs of the disease. Therapy for the disease is carried out using antiscabietic drugs. After completing the course of treatment, thorough body washing and complete change of bedding and clothing are recommended. Identify the disease.

Respondent Groups

AI Tools: ChatGPT and Bing were selected as the AI respondents due to their advanced natural language processing capabilities and accessibility. Both AI tools were provided with the same set of questions to simulate a standard testing environment.

Local Students: A total of 31 3-course students participated in the study. These students were selected based on their enrollment in medical courses and their consent to participate.

International Students: The international student group consisted of 10 participants from various countries.

All students were enrolled in medical courses at the Tashkent Medical Academy and Kimyo International University in Tashkent.

Data Collection

Data was collected by administering the questionnaire to the respondents. For the AI tools, the questions were input into the respective platforms, and the answers were recorded. For the students, the questionnaire was administered in a controlled environment to ensure academic integrity and to simulate exam-like conditions.

Data Analysis

The percentage of correct answers for each question was calculated for each group. The responses of the AI tools, local students, and international students were compared to identify differences in knowledge and understanding. This analysis was aimed at evaluating the effectiveness of AI tools in providing accurate medical knowledge compared to human respondents with varying backgrounds.

The results were tabulated to present the percentage of correct responses across all three groups for each question. The differences in performance were statistically analyzed using descriptive statistics to determine trends and patterns in the data.

Results

The study evaluated the knowledge of medical concepts across three groups: Artificial Intelligence (AI) tools (ChatGPT and Bing, n=2), local students (n=31), and international students (n=10). The correct response rates for 16 questions were analyzed to compare the performance of these groups.

Table-1.
Correct Response Rates of AI Tools and Medical Students for Medical Knowledge Questions.

Questions	Students		Artificial intelligence (ChatGPT and Bing), n=2
	Local students, n=31	International students, n=10	
Question-1	77,4%	100%	100%
Question-2	41,9%	30%	100%
Question-3	83,9%	90%	100%
Question-4	77,4%	70%	100%
Question-5	61,3%	50%	0%
Question-6	41,9%	40%	100%
Question-7	61,3%	80%	100%
Question-8	0%	90%	50%
Question-9	51,6%	70%	50%
Question-10	67,7%	100%	100%
Question-11	61,3%	90%	100%
Question-12	48,4%	90%	100%
Question-13	100%	80%	50%
Question-14	22,6%	90%	100%
Question-15	0%	0%	100%
Question-16	100%	100%	100%

Overall Performance

The AI tools demonstrated a high overall accuracy, achieving a 100% correct response rate on 9 out of 16 questions. They performed particularly well on questions related to specific medical conditions, such as the nasal septum deviation, the physiological significance of nighttime diuresis, and the correct identification of a pathology with neurological implications (Questions 1, 4, 10, 11, 12, 14, and 16) (Table-2).

The local students' group had varied performance, with correct response rates ranging from 0% to 100%. The highest accuracy was observed in questions related to the identification of disease based on clinical presentation and anatomical understanding (Questions 13 and 16, both 100%). However, they struggled with questions on the non-palpable elements of a skin rash and a specific neurological condition, achieving a 0% correct response rate for these questions (Questions 8 and 15).

The international students showed strong performance, with a 100% correct response rate on 7 out of 16 questions. This group answered all questions correctly related to the physiological and anatomical phenomena (Questions 1, 2, 3, 4, 10, 11, 12, 14, and 16). Their performance was more consistent compared to local students, with correct response rates never falling below 50% (Questions 8 and 9).

Table-2

Summary of Question Categories and Correct Response Rates by Group

Category	Question numbers	AI Tools (%)	Local Students (%)	International Students (%)
Anatomy	1, 7	90.0	69.4	100.0
Physiology	2, 3, 4	73.3	67.7	100.0
Pathology	5, 11, 12, 13, 14, 15	60.0	48.1	91.7
Medical Ethics	9, 10	85.0	59.7	75.0
Dermatology	8, 16	95.0	50.0	75.0

Question-Specific Analysis

AI tools achieved a 100% correct response rate in all questions except for Question 5 (50%), while local students demonstrated moderate accuracy, with scores ranging from 41.9% to 83.9%. International students outperformed both AI tools and local students in these categories, achieving 100% accuracy in the majority of questions.

AI tools provided accurate responses to most of these questions, achieving 100% on questions involving straightforward clinical descriptions (Questions 11, 12, 14, and 16) but failed to answer some complex pathology questions correctly (Question 15, 0%). Local students exhibited lower performance in identifying

diseases from clinical descriptions, with only a few reaching 100% accuracy. International students again demonstrated a high level of knowledge with consistent 100% scores in most pathology questions except for Question 13 (50%).

In medical ethics, AI tools and international students provided 100% correct answers for Question 10, while local students had a slightly lower accuracy (67.7%). On the topic of palpable lymph nodes (Question 9), AI tools had a correct response rate of 70%, local students 51.6%, and international students 50%.

Comparative Performance

The AI tools demonstrated a generally high level of knowledge, comparable to or exceeding the performance of local and international students in many areas. However, there were notable gaps in their understanding of certain complex physiological and pathological scenarios, indicating limitations in their training data or interpretation algorithms. Local students showed more variability in their responses, reflecting diverse levels of understanding and possibly different educational backgrounds or levels of preparation. International students, on the other hand, consistently performed well, suggesting a robust understanding of the medical concepts tested.

Overall, the results indicate that while AI tools are effective in answering a broad range of medical questions, there remains a need for further development to match the nuanced understanding exhibited by human respondents, particularly in complex or context-specific medical scenarios.

Discussion

The findings of this study highlight several important insights into the comparative understanding of medical concepts among AI tools (ChatGPT and Bing), local students, and international students. The analysis reveals both the potential and the limitations of AI in medical education, as well as differences in medical knowledge among student groups.

AI tools, such as ChatGPT and Bing, performed remarkably well on a variety of medical questions, achieving 100% accuracy in over half of the questions. This suggests that these tools are highly capable of recalling factual information and correctly interpreting straightforward clinical scenarios. Their high performance in identifying diseases based on clinical descriptions and understanding anatomical and physiological concepts demonstrates the potential for AI to support medical education, providing immediate access to accurate information and aiding in the learning process [10, 11].

However, AI tools also exhibited limitations, particularly with complex or nuanced questions [12]. For example, in Questions 5 and 15, the AI tools failed to provide the correct answers, indicating a gap in their ability to synthesize complex

physiological data or interpret intricate pathological scenarios. This suggests that while AI can supplement learning, it should not be relied upon as the sole source of information, particularly in situations requiring deep clinical reasoning or understanding of subtle medical nuances [13].

The study also revealed significant differences in performance between local and international students. International students consistently demonstrated higher accuracy across most questions, suggesting a stronger or more uniform understanding of medical concepts. This could be attributed to several factors, such as varying educational standards, differences in medical curricula, or diverse clinical exposure among international students. Local students, while showing competence in several areas, exhibited a broader range of performance. Their lower scores on specific questions, such as those involving complex pathology (e.g., Question 15) and skin conditions (Question 8), may reflect gaps in their training or experience. This variability could be due to differences in teaching methods, access to resources, or clinical experience. It underscores the need for targeted educational interventions to address these gaps and ensure a more consistent level of knowledge among all students.

The results of this study have several implications for medical education. First, AI tools could serve as valuable adjuncts to traditional educational methods, providing students with access to a broad range of information and serving as a tool for self-assessment and learning reinforcement. However, educators should be mindful of the limitations of AI, particularly its inability to fully grasp complex medical reasoning or contextual nuances. Second, the discrepancies between local and international students' performance suggest that medical educators need to ensure more standardized and comprehensive training across different groups. This may involve incorporating more diverse clinical scenarios, emphasizing critical thinking and problem-solving skills, and ensuring that all students have access to similar resources and learning opportunities.

Limitations and Future Research

While this study provides valuable insights, it also has limitations. The sample size for AI tools (n=2) and international students (n=10) was relatively small, which may limit the generalizability of the findings. Additionally, the study did not account for factors such as prior exposure to clinical practice, learning styles, or cultural differences, which could influence the results.

Future research should focus on expanding the sample size to include a broader range of AI tools and student populations. It would also be beneficial to investigate the specific factors that contribute to the observed differences in performance, such as the role of clinical exposure, teaching methods, and resource

availability. Moreover, longitudinal studies could assess the long-term impact of AI integration into medical education on students' learning outcomes and clinical competencies.

Conclusion

In conclusion, AI tools show great promise in supporting medical education, particularly in providing factual information and aiding in self-assessment. However, their limitations highlight the need for cautious integration into the learning process. The observed differences between local and international students underscore the importance of a standardized and comprehensive approach to medical education. By addressing these disparities and leveraging the strengths of AI, educators can enhance the training and preparedness of future healthcare professionals.

REFERENCES

1. Agarwal M, Sharma P, Goswami A. Analysing the Applicability of ChatGPT, Bard, and Bing to Generate Reasoning-Based Multiple-Choice Questions in Medical Physiology. Cureus. Published online June 26, 2023. doi:10.7759/cureus.40977

2. Chan KS, Zary N. Applications and Challenges of Implementing Artificial Intelligence in Medical Education: Integrative Review (Preprint). JMIR Publications Inc.; 2019. Accessed August 29, 2024. http://dx.doi.org/10.2196/preprints.13930

3. Ejaz H, McGrath H, Wong BL, Guise A, Vercauteren T, Shapey J. Artificial intelligence and medical education: A global mixed-methods study of medical students' perspectives. DIGITAL HEALTH. 2022;8:205520762210890. doi:10.1177/20552076221089099

4. 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 Medical Education. 2023;9:e46885. doi:10.2196/46885

5. Fawzi S. A Review of the Role of ChatGPT for Clinical Decision Support Systems. In: 2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES). IEEE; 2023. Accessed August 29, 2024. http://dx.doi.org/10.1109/niles59815.2023.10296668

6. Gao J, Yang L, Zou J, Fan X. Comparison of the influence of massive open online courses and traditional teaching methods in medical education in China: A meta-analysis. Biochemistry and Molecular Biology Education. 2021;49(4):639-651. doi:10.1002/bmb.21523

7. Khoobi M, Mohammadi N, Ahmadi Hedayat M, Ghiyasvandiyan S, Varaei S. Comparison of the nursing students' learning in two teaching methods using CD and training through traditional method. Journal of Nursing Education. 2017;5(6):18-23. doi:10.21859/jne-05063

8. Krive J, Isola M, Chang L, Patel T, Anderson M, Sreedhar R. Grounded in reality: artificial intelligence in medical education. JAMIA Open. 2023;6(2). doi:10.1093/jamiaopen/ooad037

9. Li M, Wang X. Research on the Application of Artificial Intelligence in Medical Education. In: Learning and Analytics in Intelligent Systems. Springer Nature Switzerland; 2024:661-670. Accessed August 29, 2024. http://dx.doi.org/10.1007/978-3-031-69457-8_60

10. Narayanan S, Ramakrishnan R, Durairaj E, Das A. Artificial Intelligence Revolutionizing the Field of Medical Education. Cureus. Published online November 28, 2023. doi:10.7759/cureus.49604

11. Passby L, Jenko N, Wernham A. Performance of ChatGPT on Specialty Certificate Examination in Dermatology multiple-choice questions. Clinical and Experimental Dermatology. 2023;49(7):722-727. doi:10.1093/ced/llad197

12. Sylolypavan A, Sleeman D, Wu H, Sim M. The impact of inconsistent human annotations on AI driven clinical decision making. npj Digital Medicine. 2023;6(1). doi:10.1038/s41746-023-00773-3

13. Webster CS. Artificial intelligence and the adoption of new technology in medical education. Medical Education. 2020;55(1):6-7. doi:10.1111/medu.14409