УДК 004:378.22

DOI: 10.31655/2307-3373-2025-24-2-34-39

ARE GENERATIVE ARTIFICIAL INTELLIGENCE CHATBOTS CAPABLE OF SUCCESSFULLY PASSING THE PULMONOLOGY LICENSE EXAMINATION?

V. L. Poberezhets*,A,B,C,D,E,F, I. O. Radohoshchyn C,D,E,F

National Pirogov Memorial University, Vinnytsya, Ukraine

A – concept and design of the study; B – data collection; C – data analysis and interpretation; D – writing the article; E – editing the article; F – final approval of the article

Abstract. Since 2022, generative artificial intelligence (AI) chatbots have been rapidly integrated into various professional domains, including healthcare. Medicine, including specialties such as pulmonology, has also adopted these technologies, with generative AI demonstrating potential in interpreting imaging, explaining spirometry results, and supporting clinical decision-making and medical education. However, it is still debatable whether generative AI models can come close to the results of human physicians in official medical licensing testing.

Objective: To evaluate the performance of generative AI chatbots in answering pulmonology certification examination questions.

Materials and Methods: In December 2024, we presented examination tests from the database of questions for the certification of pulmonologists to the most widely used in Ukraine free chatbots with generative AI — ChatGPT version 3.5, Microsoft Copilot, and Gemini. These chatbots were instructed to answer 1095 test questions from the general database, after which the answers to questions about bronchial asthma (92 questions) and allergies (35 questions) were analysed.

Results: The accuracy of ChatGPT in solving pulmonary tests was 95 % (n = 1037 correct answers), Microsoft Copilot — 92 % (n = 1008 correct answers), and Gemini — 81 % (n = 890 correct answers). For questions about the diagnosis and treatment of allergies, Microsoft Copilot showed the best accuracy with 100 % correct answers (n = 35); ChatGPT scored 94.3 % correct answers (n = 33), and Gemini — 85.7 % correct answers (n = 30). ChatGPT correctly answered the question about bronchial asthma in 91.3 % of cases (n = 84), Gemini — 79.4 % (n = 73), and Copilot — 89.1 % (n = 82). All chatbots performed better on questions with a single correct answer compared to those with multiple correct answers: ChatGPT — 92.9 % vs. 75 %, Gemini — 83.3 % vs. 37.5 %, Copilot — 94 % vs. 37.5 % of correct answers.

Conclusions: Our research has shown that generative AI chatbots demonstrated high performance in solving the examination test for pulmonology certification, which can be considered a passing grade for a medical doctor in the respiratory field. In particular, this applies to the questions related to bronchial asthma and allergies. ChatGPT demonstrated the best accuracy, answering 95 % of all tests correctly. It was found that generative AI was significantly better at solving questions with a single correct answer compared to questions with multiple correct answers.

Key words: big language model, artificial intelligence, ChatGPT, bronchial asthma, allergy.

Introduction

The field of medicine receives new inventions and research data every day, and digital technologies have been rapidly introduced over the past few years. With the release of artificial intelligence (AI) to the public in 2022, active research into its use in all areas of work has begun. This introduction has not escaped the medical field, and pulmonology in particular. Thus, we and other scientists and clinicians in this field are faced with important questions: how to cooperate with neural networks and AI, whether they can help treat patients better, find non-standard solutions, improve the quality of service delivery, and whether it is advisable to use AI to analyze clinical cases in general?

AI was developed by scientists to facilitate our daily tasks and is now used in a variety of industries, with millions of requests for help coming in every day around the world. Preliminary research data shows that medical students actively use chatbots with generative AI in their studies to find information for classes, analyse clinical cases, clarify certain cases, etc. [1]. In addition, the research shows that generative AI is widely used not only by students but also by medical professionals [2–4].

Thus, large language models are already helping doctors in their routine work, and patients are actively using them for self-education, self-diagnosis, and disease control at home. The great popularity of patient use is based on the round-the-clock availability of such an "advisor", as well as the accessible, understandable language used by chatbots and a high level of empathy, as these generative AI models were originally developed to support communication effectively. Besides, artificial intelligence has great prospects in clinical work as a way to support effective clinical decisions and reduce the duration of diagnostics [5].

However, along with the significant opportunities, there are also great potential risks. For example, the misuse of AI by students in their studies or doctors in their workflow can lead to a general decline in the level of training. Another danger is the lack of training of such models in medical data, and as a result, insufficient accuracy. It is noteworthy that AI models can be both brilliant in solving complex issues and, at the same time, ridiculous in solving rather primitive clinical problems. This is because such models need to be trained from previous examples to provide any answers. When training them, developers often skip the stage of familiarizing the model with the simplest information, starting with more complex tasks, which leads to a lack of holistic knowledge that is so necessary for making correct clinical decisions. The next danger is the spread of misinformation in the community of both patients and healthcare professionals, as information provided by chatbots is often taken as truth, which makes it possible to spread false information among a large number of users. Another challenge in the use of AI is maintaining data confidentiality, as both training and generative AI require handling huge amounts of information [6]. In the case of healthcare, such information includes personal data of patients about their health status, which is sensitive and highly private information that can be disclosed to third parties if improperly stored or transmitted. The last danger in the use of AI in medicine is the increase in health inequality. This inequality is based on the different levels of digital literacy of citizens, when people without proper digital skills and access to a smartphone, laptop, or the Internet become completely cut off from the opportunities provided by AI. The elderly, rural residents, people with low socioeconomic status, and children are at risk.

But despite the significant threats, the benefits of AI are obvious, because at this stage of its development, AI can process a much larger amount of information more efficiently than any human. AI is being actively implemented in radiology and imaging diagnostics of respiratory diseases due to its ability to recognise pathological changes in chest X-rays, computed tomography, ultrasound, or other imaging methods. [3, 7]. When it comes to interpreting pulmonary function tests based on spirometry, body plethysmography, and DLCO (Diffusing Capacity of the Lung for Carbon Monoxide), AI has shown great promise as a method of assisting doctors in making clinical decisions, increasing their efficiency [8]. In addition to direct participation in clinical work, AI can be used to automate medical workflow and fill in medical records, which increases the time a doctor can spend with a patient [9].

Due to the high speed of data analysis and clinical decision-making, AI is increasingly being considered as a method of helping patients in emergency conditions and training in emergency medicine. Preliminary studies have shown a high rate of disease detection using imaging methods, correct classification of patients with low and high urgency, and a high level of accuracy in the assessment of emergency medicine examinations [2, 10].

But to make AI more advanced, we need to build a system for continuous evaluation and improvement of existing models and development of future ones. This process must involve all stakeholders: patients and their families, healthcare professionals and administrators, policymakers, non-governmental organizations, and developers of digital technologies and medical equipment. It is the fruitful cooperation between various healthcare system stakeholders that is crucial for the development of AI tools that will truly improve the quality of healthcare, be accessible to everyone, and not pose additional dangers or risks to users. Thus, it can be noted that AI chatbots are constantly improving their algorithms in cooperation with doctors and medical students: the test and clinical tasks that were previously solved incorrectly due to the limited AI algorithm have been constantly progressing over the past three years and are solved in a new style that meets clinical requirements.

In our study, we decided to investigate only one of the possible ways of using AI in pulmonology — to assess the ability of popular and widely available chatbots with generative AI to compete with pulmonologists in solving test tasks used for certification of medical specialists. Existing data on such AI capabilities are quite controversial and have not been studied on the example of pulmonology tasks, especially issues related to the treatment of bronchial asthma and allergies.

Objective

To evaluate the performance of generative AI chatbots in answering pulmonology certification examination questions.

Materials and Methods

We analysed the ability of chatbots with generative AI to solve exam tests for the successful certification of pulmonologists in Ukraine during December 2024. The three most popular free chatbots were selected for the analysis: ChatGPT version 3.5, Microsoft Copilot, and Gemini. These chatbots were given the task of solving all 1095 test questions from the general database and separately questions by keywords: “bronchial asthma”, “asthma”, and “allergies”. Thus, 35 tests on allergy and 92 tests on bronchial asthma were identified.

Results and Discussion

All three chatbots demonstrated strong overall performance, with over 80 % of test questions answered correctly Thus. ChatGPT proved to be the most accurate in solving pulmonary tests, with 95 % accuracy (n = 1037 correct answers), Microsoft Copilot scored 92 % (n = 1008 correct answers), and Gemini scored 81 % (n = 890 correct answers) (pic. 1).

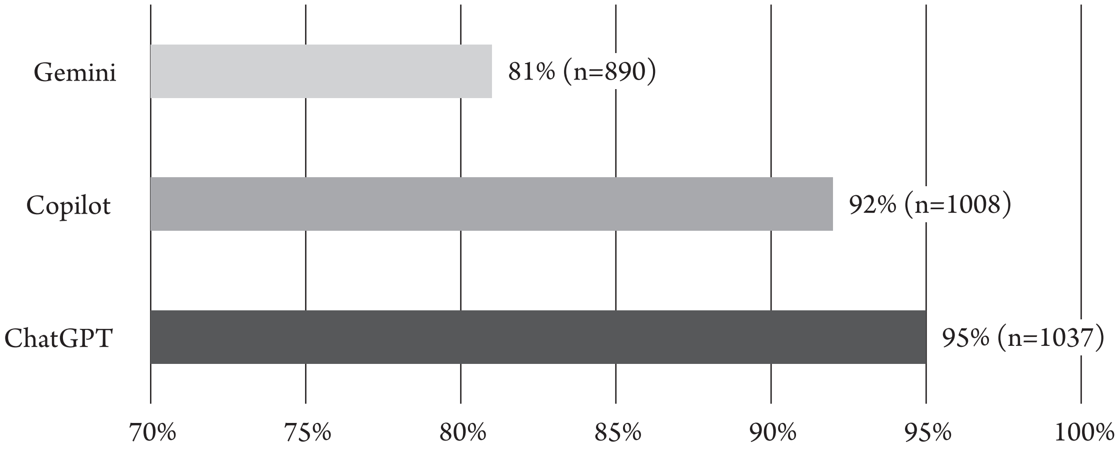
The analysis of questions related to the diagnosis and treatment of allergies (35 questions) revealed that Microsoft Copilot showed the best results with 100 % correct answers (n = 35); ChatGPT scored 94.3 % correct answers (n = 33), Gemini — 85.7 % correct answers (n = 30) (pic. 2).

Further analysis of the questions related to prevention, diagnosis, and treatment of asthma (92 questions, including 8 with multiple answers) showed the following results: ChatGPT — 91.3 % correct answers (n = 84), Gemini — 79.4 % correct answers (n = 73), Copilot — 89.1 % correct answers (n = 82) (pic. 3).

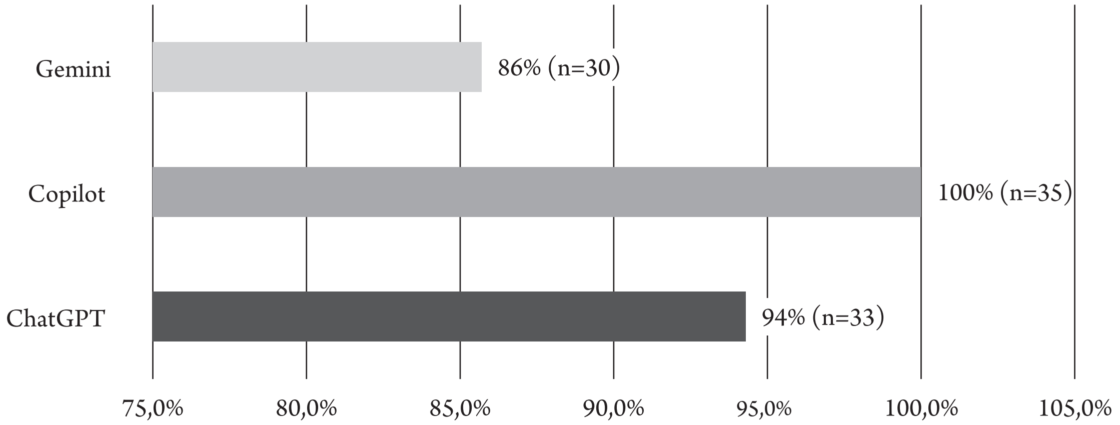
Analysis of the subgroup of single-answer tests (84 items) revealed the following accuracy rates: ChatGPT — 92.9 % correct answers (n = 78), Gemini — 83.3 % correct answers (n = 70), Copilot — 94 % correct answers (n = 79).

The analysis of tasks with multiple correct answers (8 tasks) showed a much worse ability of generative AI to solve such clinical problems: ChatGPT — 75 % correct answers (n = 6), Gemini — 37.5 % correct answers (n = 3), Copilot — 37.5 % correct answers (n = 3).

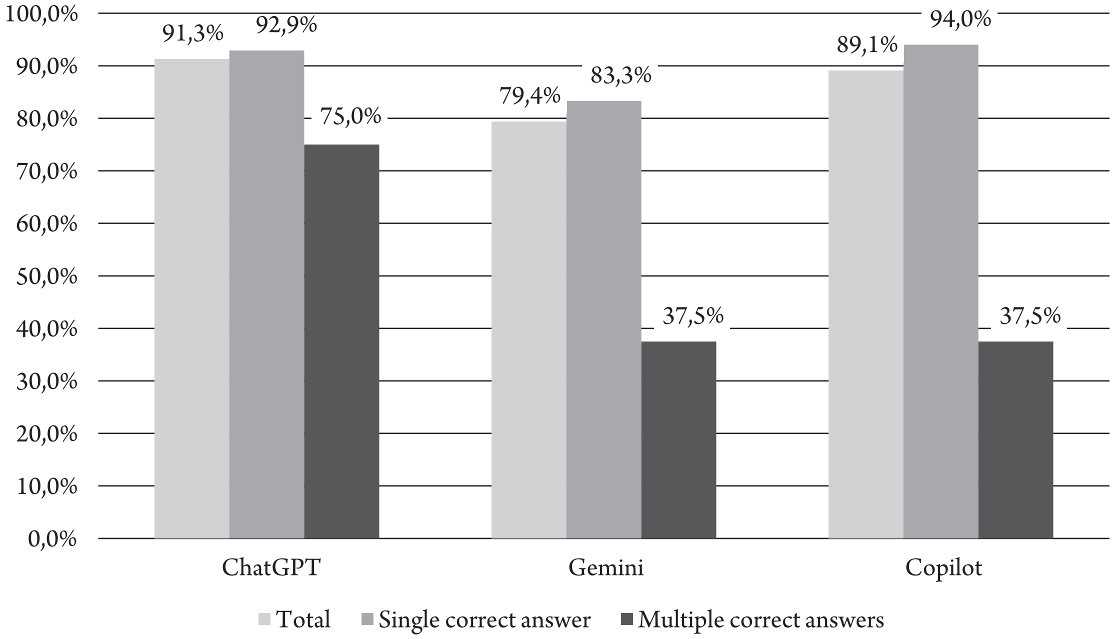
Our results confirm the high performance of chatbots with generative AI in solving clinical test questions, which confirms the previous data obtained by researchers from Saudi Arabia who demonstrated the high performance of ChatGPT, Claude, Copilot, and Gemini in solving multiple-choice tests of the USMLE (American Medical Licensing Examination) [11]. Similar results were shown by US researchers who developed an adapted assistant based on generative AI to prepare for the USMLE Step 1 exam [12]. In another study conducted in the UK, ChatGPT-4o scored a passing score (94 %) in both the UK Medical Practitioner’s Examination for Applied Knowledge (MPE) and USMLE Step 1 (89.9 %). An important discovery was that the chatbot’s performance did not decline, even when the wording of the questions was changed or when completely new questions were used that had never been published before. However, the study also highlighted a long-known weakness of generative AI — ChatGPT demonstrated a decrease in performance with questions containing images. This study showed that the performance and efficiency of chatbots with generative AI continue to improve [13]. In anoth-



Picture 1. Total accuracy of generative AI chatbots in pulmonary examination testing.



Picture 2. Accuracy of generative AI chatbots in the questions about allergy.



Picture 3. Accuracy of generative AI chatbots in the questions about bronchial asthma.

er study conducted by a collaboration of researchers from various British universities, another weakness of ChatGPT in solving tasks for obtaining a license to practice medicine in the UK was found — GPT-4 performed significantly better on questions about diagnosing diseases and conditions than on questions about treatment and management of the disease ($p = 0.015$) [14].

In another study, researchers from Qingdao University evaluated the ability of different ChatGPT models to pass the Chinese National Medical Licensing Examination. GPT-4o demonstrated significantly higher overall accuracy than GPT-4 and GPT-3.5, reaching accuracy rates of 84.2 % and 88.2 % in the 2020 and 2021 exams, respectively. Moreover, the highest accuracy (94.75 %) was observed in questions related to the digestive and respiratory systems. It is striking that such indicators were achieved by the chatbot in medical exams in languages other than English [15].

Given that the market for generative AI chatbots is developing extremely fast, it can take a much shorter period for a new player to emerge. This is exactly what happened when DeepSeek was introduced to the world in early 2025. This chatbot immediately attracted the attention of the global scientific community, and in April 2025, US researchers published the results of a study of the ability of this generative AI tool to solve the tasks of the USMLE medical licensing exam. In this study, the authors compared the new model with the already known ChatGPT and Llama by Meta AI. The performance of ChatGPT was higher than that of DeepSeek (95 % vs. 92 %, $p = 0.04$). The percentage of correct answers by Llama was only 83 %. [16].

These studies confirm our findings and even demonstrate that chatbots with generative AI continue to improve in solving medical licensing exams. However, there are still limitations and difficulties in the widespread use of such models in clinical practice, which require further research.

Conclusions

1. Our research has shown that chatbots with generative AI (ChatGPT 3.5, Microsoft Copilot and Gemini) have demonstrated high performance in solving the examination test for the certification of pulmonologists. In particular, with regard to questions related to bronchial asthma and allergies. These results were quite high and can be considered a passing grade for a medical doctor in the respiratory field.
2. The best results were demonstrated by ChatGPT, which correctly answered 95 % of all tests. The second result was demonstrated by Microsoft Copilot — 92 % correct answers, and Gemini - the third (81 % correct answers).
3. Microsoft Copilot showed the best results in solving questions related to the diagnosis and treatment of allergies, with 100 % correct answers.
4. When analysing the questions related to the prevention, diagnosis and treatment of asthma, ChatGPT scored 91.3 % of correct answers. Gemini had the worst accuracy with only 79.4% of correct answers.
5. A further analysis of the tasks revealed that generative AI was much better at solving questions with a single correct answer (in this group, all chatbots scored more than 83% of correct answers). Significantly worse results were found when evaluating questions with multiple correct answers. ChatGPT was the only chatbot that demonstrated the ability to solve 75 % of such questions correctly. Gemini and Copilot were not able to solve even half of such questions.
6. There is a need for further similar research to identify weaknesses in generative AI, which is crucial for improving existing AI models and will open opportunities for implementing such technologies in the healthcare system.

ЧИ ЗДАТНІ ЧАТ-БОТИ ІЗ ГЕНЕРАТИВНИМ ШТУЧНИМ ІНТЕЛЕКТОМ УСПІШНО СКЛАСТИ ЕКЗАМЕНАЦІЙНЕ ТЕСТУВАННЯ ДЛЯ АТЕСТАЦІЇ ЛІКАРІВ-ПУЛЬМОНОЛОГІВ?

В. Л. Побережець, І. О. Радогощин

Вінницький національний медичний університет ім. М. І. Пирогова, Вінниця, Україна

Резюме. Чат-боти із генеративним штучним інтелектом (ШІ) за досить короткий проміжок часу (із 2022 року) інтегрувались у всі сфери нашого життя, навіть якщо ми цього не помічаємо. Медицина та її окремі галузі, такі як пульмонологія, не стала виключенням і генеративний ШІ відмінно проявив свій потенціал у інтерпретації візуалізаційних методів досліджень, поясненні результатів спірометрії, допомозі у прийнятті клінічних рішень та навчанні. Однак досі залишається дискутабельним питання чи здатні моделі із генеративним ШІ наблизитись до результатів живих лікарів у офіційному медичному ліцензійному тестуванні.

Мета роботи: Оцінити здатність чат-ботів із генеративним штучним інтелектом у вирішенні екзаменаційного тестування для атестації лікарів-пульмонологів.

Матеріали та методи: У грудні 2024 року нами було запропоновано вирішити екзаменаційні тести із бази запитань для атестації лікарів-пульмонологів трьом найпоширенішим в Україні безкоштовним чат-ботам із генеративним ШІ — ChatGPT версія 3.5, Microsoft Copilot та Gemini. Даним чат-ботам було представлено завдання вирішити 1095 тестових завдань із загальної бази даних, після чого було здійснено аналіз відповідей на запитання про бронхіальну астму (92 запитання) та алергопатологію (35 запитань).

Результати: Точність ChatGPT у вирішенні пульмонологічних тестів склала 95 % (n = 1037 правильних відповідей), Microsoft Copilot — 92 % правильних відповідей (n = 1008), а Gemini — 81 % правильних відповідей (n = 890). У відповідях на запитання, що стосувались діагностики та лікування алергопатології найкращі результати показав Microsoft Copilot із 100 % правильних відповідей (n = 35); ChatGPT набрав 94,3 % правильних відповідей (n = 33), Gemini — 85,7 % правильних відповідей (n = 30). На запитання про бронхіальну астму ChatGPT відповів правильно у 91,3 % випадків (n = 84), Gemini — 79,4 % (n = 73), Copilot — 89,1 % (n = 82). Усі чат-боти показали кращі результати при відповіді на запитання, що мали єдину правильну відповідь ніж на запитання із множинними правильними відповідями: ChatGPT — 92,9 % проти 75 %, Gemini — 83,3 % проти 37,5 %, Copilot — 94 % проти 37,5 % правильних відповідей.

Висновки. Наше дослідження встановило, що чат-боти із генеративним ШІ продемонстрували високу результативність у вирішенні екзаменаційного тестування для атестації лікарів-пульмонологів, що можна вважати прохідним для лікаря-спеціаліста. Зокрема, це стосується і запитань щодо бронхіальної астми та алергопатології. Найкращий загальний результати продемонстрував ChatGPT, який правильно відповів на 95 % усіх тестів. Було виявлено, що генеративний ШІ значно краще справлявся із вирішенням запитань із єдиною правильною відповіддю порівняно із запитаннями із множинними правильними відповідями.

Ключові слова: велика мовна модель, штучний інтелект, ChatGPT, бронхіальна астма, алергопатологія.

REFERENCES

1. Poberezhets VL, Starychenko M. The Possibility of Using Large Language Models for Teaching Future Doctors (on the Example of ChatGPT). Asthma and Allergy. 2024;23(4):42-47. https://doi.org/10.31655/2307-3373-2024-4-42-47.
2. Paslı S, Şahin AS, İmamoğlu M, et al. Assessing the precision of artificial intelligence in emergency department triage decisions: insights from a study with ChatGPT. Am J Emerg Med. 2024;78:170–175. doi: 10.1016/j.ajem.2024.01.037.
3. Ostrovsky AM. Evaluating a large language model's accuracy in chest X-ray interpretation for acute thoracic conditions. The American journal of emergency medicine. 2024;93:99-102. doi:10.1016/j.ajem.2025.03.060.
4. Potapenko I, Boberg-Ans LC, Subhi Y, et al. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. Acta Ophthalmol. 2023;101:829–831. doi: 10.1111/aos.15661.
5. Drummond D, Adejumo I, Hansen K, et al. Artificial intelligence in respiratory care: perspectives on critical opportunities and challenges. Breathe (Sheff). 2024;20(3):230189. doi:10.1183/20734735.0189-2023.
6. Introduction to Large Language Models. Available from: https://developers.google.com/machine-learning/resources/intro-llms (last accessed 18.05.2025).
7. Lee HW, Jin KN, Kim Y J, et al. Artificial intelligence solution for chest radiographs in respiratory outpatient clinics: multicenter prospective randomized clinical trial. Ann Am Thorac Soc. 2023;20:660–667. doi:10.1513/AnnalsATS.202206-481OC.
8. Das N, Happaerts S, Janssens W, et al. Collaboration between explainable artificial intelligence and pulmonologists improves the accuracy of pulmonary function test interpretation. Eur Respir J. 2023;61:2201720. doi:10.1183/13993003.01720-2022.
9. Shah SJ, Crowell T, Jeong Y, et al. Physician Perspectives on Ambient AI Scribes. JAMA Netw Open. 2025;8(3):e251904. doi:10.1001/jamanetworkopen.2025.1904.
10. Berikol GB, Kanbakan A, Ilhan B, et al. Mapping artificial intelligence models in emergency medicine: A scoping review on artificial intelligence performance in emergency care and education. Turk J Emerg Med. 2025;25(2):67-91. doi:10.4103/tjem.tjem_45_25.
11. Bolgova O, Ganguly P, Mavrych V. Comparative analysis of LLMs performance in medical embryology: A cross-platform study of ChatGPT, Claude, Gemini, and Copilot. Anat Sci Educ. Published online May 11, 2025. doi:10.1002/ase.70044.
12. Cho Y, Park GL, Waite GN, et al. Development of a Universal Prompt as a Scalable Generative AI-Assisted Tool for USMLE Step 1 Style Multiple-Choice Question Refinement in Medical Education. Med Sci Educ. 2025;35(2):611-613. doi:10.1007/s40670-025-02334-7.
13. Newton PM, Summers CJ, Zaheer U, et al. Can ChatGPT-4o Really Pass Medical Science Exams? A Pragmatic Analysis Using Novel Questions. Med Sci Educ. 2025;35(2):721-729. doi:10.1007/s40670-025-02293-z.
14. Casals-Farre O, Baskaran R, Singh A, et al. Assessing ChatGPT 4.0's Capabilities in the United Kingdom Medical Licensing Examination (UKMLA): A Robust Categorical Analysis. Sci Rep. 2025;15(1):13031. doi:10.1038/s41598-025-97327-2.
15. Luo D, Liu M, Yu R, et al. Evaluating the performance of GPT-3.5, GPT-4, and GPT-4o in the Chinese National Medical Licensing Examination. Sci Rep. 2025;15(1):14119. doi:10.1038/s41598-025-98949-2.
16. Tordjman M, Liu Z, Yuce M, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. Nat Med. Published online April 23, 2025. doi:10.1038/s41591-025-03726-3.

Conflict of interests: None.
Sources of financing: Study did not receive any financial support.
Ethics: This study did not involve patients or the use of their data, and therefore did not require approval by an ethics committee.

Конфлікт інтересів: Відсутній.
Джерела фінансування: Це дослідження не отримало жодної фінансової підтримки.
Етика: Це дослідження не стосувалося пацієнтів та використання їхніх даних, тому й не потребувало схвалення етичним комітетом.

Цитування: Побережець ВЛ, Радогощин ІО. Чи здатні чат-боти із генеративним штучним інтелектом успішно скласти екзаменаційне тестування для атестації лікарів-пульмонологів? Астма та алергія. 2025;24(2):34–39. DOI: 10.31655/2307-3373-2025-24-2-34-39.

*Cited:*Poberezhets VL, Radohoshchyn IO. Are generative artificial intelligence chatbots capable of successfully passing the pulmonology license examination? Asthma and allergy (Ukraine). 2025;24(2):34–39. DOI: 10.31655/2307-3373-2025-24-2-34-39. Ukrainian.

Відомості про авторів

В. Л. Побережець*
Доктор філософії, асистент кафедри пропедевтики внутрішньої медицини, Вінницький національний медичний університет ім. М.І. Пирогова, кафедра пропедевтики внутрішньої медицини
Вул. Хмельницьке шосе, 96, 21029,
м. Вінниця, Україна;
Тел.: 38093-795-77-53, poberezhets_vitalii@vnmu.edu.ua
ORCID: https://orcid.org/0000-0003-2581-824X

І. О. Радогощин
Студент 4 курсу медичного факультету №2, Вінницький національний медичний університет ім. М.І. Пирогова.
Вул. Хмельницьке шосе, 96, 21029, м. Вінниця, Україна

Information about autors

V. L. Poberezhets
MD, PhD, Assistant Professor, Department of Propedeutics of Internal Medicine, National Pirogov Memorial University, Vinnytsya.
Khmelnytske highwat street 96, 20129, Vinnytsia, Ukraine

I. O. Radohoshchyn
Fourth-year student of the medical faculty №1, National Pirogov Memorial University, Vinnytsya.
Khmelnytske highwat street 96, 20129, Vinnytsia, Ukraine

Надійшла до редакції / Received: 19.05.2025 р.
Після доопрацювання / Revised: 26.05.2025 р.
Прийнято до друку / Accepted: 31.05.2025 р.