ChatGPT fails the Polish board certification examination in internal medicine: artificial intelligence still has much to learn

Szymon Suwała[1], Paulina Szulc[2], Aleksandra Dudek[2], Aleksandra Białczyk[2], Kinga Koperska[2], Roman Junik[1]

1 Department of Endocrinology and Diabetology, Nicolaus Copernicus University, Collegium Medicum, Bydgoszcz, Poland
2 Evidence-Based Medicine Student Scientific Club of Department of Endocrinology and Diabetology, Nicolaus Copernicus University, Collegium Medicum, Bydgoszcz, Poland

Correspondence to:
Szymon Suwała, MD, PhD,
Department of Endocrinology and
Diabetology, Nicolaus Copernicus
University, Collegium Medicum,
ul. Sklodowskiej-Curie 9,
85-094 Bydgoszcz, Poland,
phone: +48 52 585 42 40, email:
lekarz.szymon.suwala@gmail.com

Received: June 19, 2023.
Revision accepted:
November 15, 2023.
Published online:
November 20, 2023.
Pol Arch Intern Med. 2023;
133 (11): 16608
doi:10.20452/pamw.16608
Copyright by the Author(s), 2023

Introduction Internal medicine, as a field, is often referred to as the queen of medical science. Physicians specializing in internal medicine are required to possess extensive knowledge as well as a high degree of focus and self-discipline. Based on official data from the Polish Chamber of Physicians and Dentists,[1] as of September 30, 2023, there were a total of 31 184 internal medicine specialists practicing in Poland, which accounts for 16.51% of the total number of physicians practicing in the country. In accordance with the Polish law, a physician can become an internal medicine specialist after completing specialist training and passing the board certification examination. The assessment consists of 2 elements: a multiple-choice test that encompasses 120 questions with 5 possible answers of which only 1 is accurate, and an oral examination that can only be attempted upon successfully passing the written test (which requires scoring at least 60% of possible points). However, as of the beginning of 2023, a candidate may be excused from the oral examination if they achieve a minimum score of 75% in the written test.

Recently, artificial intelligence (AI) has made a noteworthy progress and gained widespread adoption across diverse domains. ChatGPT, a natural language processing model with a capacity of 175 billion parameters, developed by OpenAI, was launched on November 30, 2022. It is an AI model that has undergone extensive training on a vast corpus of textual data. Its primary function is to respond to user queries in a manner that is both coherent and contextually relevant. The deployment of AI has attracted worldwide attention, occasionally accompanied by apprehension, particularly in the fields that have conventionally relied on human creativity and productivity, such as marketing, science, and education. Concerns over being replaced or made redundant are growing among employees as a result of the accelerated development of AI and its expanding role in a variety of tasks. Such a fear has existed in our society at least since the industrial revolution. AI has shown promise in the medical field, with ChatGPT passing examinations such as the United States Medical Licensing Examination (USMLE),[2] the European Exam in Core Cardiology,[3] or the Ophthalmic Knowledge Assessment Program (OKAP) exam.[4] Although AI appears limited and incapable of substituting human thought and critical data analysis (which are an essential part of a physician's daily work), we attempted to investigate whether it is possible for an AI system to successfully pass an exam required to become an internal medicine specialist, which is a crucial medical specialty in Poland. To our best knowledge, this is the first study in the world to assess the capability of AI in the field of internal medicine.

Methods The aim of this study was to verify whether OpenAI ChatGPT is currently capable of passing the Polish board certification examination in the field of internal medicine, and to find correlations between the effectiveness of AI and the characteristics of the examination tasks. OpenAI ChatGPT, version of May 24, 2023, was presented with all questions from the Polish board certification examinations from the years 2013 to 2017 (10 sessions; questions from this period are publically available, published by the Medical Examination Center along with detailed statistics, such as the percentage of examinees selecting specific answers or the difficulty index [DI; ie, the ratio of the average total score of examinees to the range of full marks, according to the definition of Nitko adopted by Johari et al[5]; the lower the index, the more difficult the task]). The total number of questions (after removing questions that are impossible for ChatGPT to analyze, such as those containing images) was 1191. Each author of the study presented ChatGPT with original examination questions (without prompts or other specific commands) from chosen examination sessions once, simulating the course of a particular exam. An example, together with the obtained answer, is shown in Supplementary material. We decided not to include the questions added to the examination database after 2021 (despite their availability) as ChatGPT, based on the GPT-3.5 architecture, has a limited knowledge base that only extends up to 2021, and does not have real-time access to Internet data.[6]

According to the regulations of the Medical Examination Center, test queries may take the form of A-type tasks (with a single correct response) or K-type tasks (complex questions, requiring choosing a correct set of statements). In the analyzed sample, A-type tasks constituted 83.8% of all questions (998 out of 1191). We also verified the length of questions (median [interquartile range, IQR], 289 [231–370] characters, including spaces) and their DI (median [IQR], 0.692 [0.581–0.806]). The number of human examinees selecting the correct answer was obtained from the official database of the Medical Examination Center. Based on these data, the tasks were divided into equal quintiles, and categorized respectively as very short (≤218 characters), short (219–264 characters), medium-length (265–318 characters), long (319–396 characters), and very long (≥397 characters), and very easy (DI ≥0.828), easy (DI, 0.737–0.827), intermediate (DI, 0.652–0.735), hard (DI, 0.551–0.651), and very hard (DI ≤0.55).

The responses provided by ChatGPT were validated using the official answer key, which had been reviewed for any changes resulting from the advancement of medical knowledge.

Statistical analysis Statistical analysis was performed with STATISTICA 13.0 software (TIBCO Software Inc., Palo Alto, California, United States). Data are presented as median (IQR), and they were compared using nonparametric tests, such as the Mann–Whitney test, Kruskal–Wallis test, and Spearman correlation coefficient, due to the results of the Shapiro–Wilk tests, which indicated non-normal distribution of the variables. A *P* value below 0.05 was assumed as significant.

Results In the analysis of 10 examination sessions that took place between the spring of 2013 and the autumn of 2017, ChatGPT demonstrated a correct answer rate of 47.5% to 53.33% (median, 49.37%), which was insufficient to pass the examination. In all sessions, the performance of ChatGPT was significantly inferior to that of human examinees (whose results ranged from 65.21% to 71.95%; median, 69.92%). Comprehensive outcomes are presented in FIGURE 1. There was a significant difference in the mean DI of questions between individual examination sessions (*P* <0.001).

Upon analyzing the categorization of questions into A-type and K-type, it was observed that both human participants and ChatGPT exhibited superior performance on A-type questions (ChatGPT: median [IQR] result, 52.88% [50.12%–55.88%] vs 29.38% [21.63%–34.13%]; *P* = 0.01; human examinees: median [IQR] result, 70.16% [66.85%–71.18%] vs 66.39% [63.96%–67.4%]; *P* = 0.003). The difference between AI and human examinees was significant in both categories (*P* <0.001). There was a significant correlation between the type of question and the outcomes of ChatGPT (β = 0.16; SE = 0.03; *P* <0.001) and human responders (β = 0.09; SE = 0.03; *P* = 0.002), as determined in the linear regression analysis. There was also a substantial difference in the average DI between the A-type and K-type questions (median [IQR], 0.697 [0.589–0.81] vs 0.656 [0.538–0.769]; *P* <0.001); questions that required the selection of a correct set of statements were more challenging.

The outcomes of ChatGPT showed notable differences based on the quintile of question length. ChatGPT performed the best on questions of minimal length (median [IQR] result, 59.17% [53.13%–64.71%]), followed by long (50.86% [41.67%–54.55%]), very long (49.29% [45.45%–58.54%]), short (46.14% [43.48%–55.18%]), and medium-length questions (39.56% [38.46%–50%]) (*P* = 0.003). Regarding human examinees, the difference was not significant; very short questions were associated with the highest results (median [IQR] result, 73.9% [68.58%–74.63%]), followed by very long (70.1% [68.6%–72.68%]), long (69.32% [64.82%–71.33%]), short (69.05% [67.09%–72.92%]), and medium-length (68.9% [67.32%–70.65%]) questions (*P* = 0.13). There was no significant link between the length of questions and the results of ChatGPT and human participants, as shown in the linear regression analysis (ChatGPT: β = –0.05; SE = 0.03; *P* = 0.1; humans: β = –0.04; SE = 0.03; *P* = 0.12). On the other hand, we found a significant, but very weak, correlation between the length of the question and the DI (R = –0.08; *P* <0.001).

With respect to question difficulty, it was discovered that the performance of ChatGPT gradually declined as the task difficulty increased, which is consistent with human behavior. ChatGPT obtained a median (IQR) score of 63.48% (57.89%–68.97%) on very easy, 55.63% (54.17%–66.67%) on easy, 41.67% (35%–50%) on intermediate, 41.88% (39.13%–54.55%) on hard, and 37.12% (25%–43.47%) on very hard questions (*P* <0.001). Humans examinees scored 91.18% (90.28%–92.21%) on very easy, 81.63% (81.33%–82.63%) on easy,

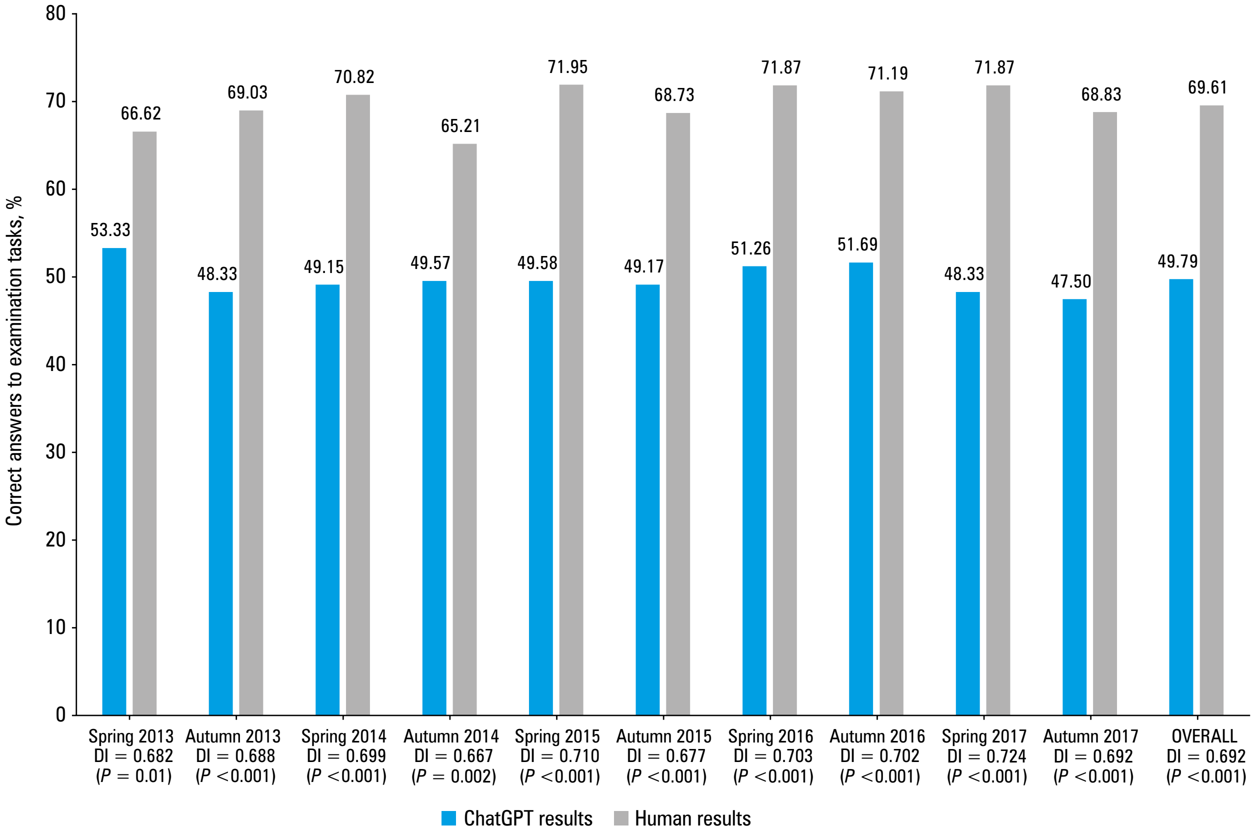

FIGURE 1 Mean results of ChatGPT and human examinees in individual examination sessions and overall, including the difficulty index (DI) and *P* value for the difference between AI and human performance

71.47% (70.46%–72.63%) on intermediate, 61.97% (61.36%–63%) on hard, and 41.92% (40.77%–43.33%) on very difficult tasks (*P* <0.001). A significant difference was observed between the results of human examinees and ChatGPT for each quintile (*P* <0.001), with the exception of the quintile that included the most challenging tasks (*P* = 0.17). In the linear regression analysis, we found a significant association between the DI and the results of ChatGPT (β = 0.18; SE = 0.03; *P* <0.001). In the case of human examinees, considering the methodology for determining the DI, the result was obvious (β = 0.96; SE = 0.01; *P* <0.001).

Additionally, the performance of ChatGPT in answering questions concerning specific fields of internal medicine was checked. ChatGPT most often responded correctly to questions concerning allergology (71.43%), followed by those on infectious diseases (55.26%), endocrinology (54.64%), nephrology (53.51%), rheumatology (52.83%), hematology (51.51%), gastroenterology (50.97%), pulmonology (46.71%), and diabetology (45.1%). It obtained the lowest result (43.72%) on questions regarding cardiology.

Discussion AI has made significant advancements in recent years, and gained considerable popularity in various fields. Previous applications of AI in health care involved tasks such as cataloging and interpreting big data or developing and implementing diagnostic–therapeutic algorithms (such as detection of skin lesions suspected to be tumors and preliminary assessment of electrocardiograms or radiographic images).[7,8] AI usage seems to be of great aid given the underfunding of health care systems, the problem of professional burnout among medical professionals, and personnel shortages.[9]

The implementation of natural language processing AI models has gained global recognition, as evidenced by ChatGPT's achievement of an estimated 100 million active users monthly within 2 months of its launch, thereby setting a record as the most rapidly growing consumer application in history.[10] ChatGPT has proven effective in passing professional medical examinations—one of the notable findings was reported by Kung et al,[2] who demonstrated that ChatGPT was capable of passing the USMLE without prior training (which encouraged researchers similar to us to conduct their own research and led others to conclude that it may be a good opportunity to harness AI for educational development and creation of new question databases).[11] According to the available literature, at the time of writing this article, ChatGPT has also performed well on other medical exams, including ones in physiology,[12] microbiology,[13] parasitology,[14] as well as postgraduate exams in radiology,[15] the European Exam in Core Cardiology,[3] or the OKAP exam.[3] The knowledge of AI, as evaluated through the American Board of Orthopaedic Surgery Examination, was found to be equivalent to that of a novice resident in the corresponding areas of specialty.[16] However, ChatGPT did not successfully pass the American College of Gastroenterology examination[17] or the American Heart Association Basic Life Support and Advanced Cardiovascular Life Support exam.[18] The aforementioned reports indicate that the abilities of AI are limited, and at present, it is difficult for AI to compete with the expertise of trained medical professionals, particularly in the field of internal medicine.

The performance of ChatGPT was unsatisfactory in the challenging examination that required extensive medical knowledge of internal medicine. With results ranging from 47.5% to 53.33%, it did not meet the minimum requirement of 60% of correct responses. Interestingly, the best result was achieved on questions from the exam conducted in the spring of 2013 (the earliest session included in the current analysis), while the worst result was obtained on questions from the autumn of 2017 (the latest session analyzed). This may be related to different retention of information from the training data of ChatGPT (longer exposure to older than to recent information). Of note, the knowledge base of ChatGPT is presently limited to the year 2021. Nevertheless, it is worth mentioning that the success rate of ChatGPT between the first and last session did not show a gradual downward trend, which may contradict the previous speculation. Currently, it is not feasible to make comparisons between our findings and those of other investigations due to the lack of literature that incorporates responses to questions from multiple examination sessions carried out over a few years.

According to our assessment, ChatGPT (similarly to humans) performed significantly better on questions that were simpler in construction and required focusing on single pieces of information rather than on finding a set of correct statements. Failure of ChatGPT to answer complex questions may potentially be attributed to information errors, whereby the AI system failed to recognize a pivotal element of information. Gilson et al[6] identified 2 additional factors contributing to the inadequacy of AI in answering questions, namely logical errors (where ChatGPT adequately identified the pertinent information but did not properly convert it to an answer) and statistical errors (centered around an arithmetic mistake). In their study, logical errors were the most common cause of AI failure. Information errors were ranked as the second most prevalent type of error in their study, closely following logical errors. Due to the same reasons, ChatGPT may exhibit superior performance when presented with short queries—it is evident that the shorter the question, the easier it is for the machine algorithm to identify the key information necessary for a correct answer. However, the reason behind the poorest performance of ChatGPT on questions of medium length and not the longest ones is yet to be determined. It cannot be excluded that this was merely a chance event.

It is difficult to establish whether the failure of ChatGPT on the Polish board certification examination in internal medicine can be attributed to linguistic differences between Polish and English. A study by Panthier et al,[19] investigating the effectiveness of ChatGPT in the French version of the European Board of Ophthalmology Examination, suggested that the primary determinant was not the language. Nevertheless, it is important to take into account the contrasting worldwide prominence of French and Polish.

Our study has a few limitations. The analysis focused solely on the assessment of ChatGPT's performance, without engaging in any comparative evaluations with alternative AI models. Additionally, it should be noted that ChatGPT undergoes regular updates, and the version employed in our research may not necessarily reflect the most current iteration at the time of publication. Regardless of these limitations, our investigation offers valuable insights into the advantages and disadvantages of ChatGPT in the context of its application in the field of medicine.

It is unlikely that AI will be able to replace health care professionals in the near future, particularly in the field of internal medicine—even the most sophisticated algorithms and technologies facilitated by AI are incapable of diagnosing and treating diseases without human input. However, medicine is a field in which the utilization of AI language processing models may be beneficial. For example, the courteous behavior displayed by ChatGPT and its potential application in regular clinical settings are noteworthy. A study comparing physician and AI chatbot responses to urgent medical inquiries posted on a public social media forum showed that 79% of patients perceived the responses provided by ChatGPT to be more empathetic and comprehensive than those offered by human professionals.[20] Undoubtedly, it is worthwile to follow the development of AI, especially ChatGPT, to be able to take advantage of its rapid progress.

SUPPLEMENTARY MATERIAL

Supplementary material is available at www.mp.pl/paim.

ARTICLE INFORMATION

ACKNOWLEDGMENTS We thank Cezary Guzowski and Bernadetta Kosztulska for their assistance in conducting the study.

FUNDING Collegium Medicum of Nicolaus Copernicus University partially financed the study using sources designated for basic research tasks.

CONFLICT OF INTEREST None declared.

OPEN ACCESS This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), allowing anyone to copy and redistribute the material in any medium or format and to remix, transform, and build upon the material, including commercial purposes, provided the original work is properly cited.

HOW TO CITE Suwała S, Szulc P, Dudek A, et al. ChatGPT fails the Polish board certification examination in internal medicine: artificial intelligence still has much to learn. Pol Arch Intern Med. 2023; 133: 16608. doi:10.20452/pamw.16608

REFERENCES

[1] Polish Chamber of Physicians and Dentists website. Statistical data. https://nil.org.pl/rejestry/centralny-rejestr-lekarzy/informacje-statystyczne. Accessed November 23, 2023.

2 Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digital Health. 2023; 2: e0000198.

3 Skalidis I, Cagnina A, Luangphiphat W, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Heart J Digit Health. 2023; 4: 279-281.

4 Antaki F, Touma S, Milad D, et al. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci. 2023; 3: 100324.

5 Johari J, Sahari J, Wahab DA, et al. Difficulty index of examinations and their relation to the achievement of programme outcomes. Procedia Soc Behav Sci. 2011; 18: 71-80.

6 Gilson A, Safranek CW, Huang T, 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.

7 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019; 25: 44-56.

8 Lo CK. What is the impact of ChatGPT on education? A rapid review of the literature. Educ Sci (Basel). 2023; 13: 410.

9 Meskó B, Hetényi G, Győrffy Z. Will artificial intelligence solve the human resource crisis in healthcare? BMC Health Serv Res. 2018; 18: 545.

10 Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ. 2023; 9: e46885.

11 Biswas S. Passing is great: can ChatGPT conduct USMLE exams? Ann Biomed Eng. 2023; 51: 1885-1886.

12 Subramani M, Jaleel I, Krishna Mohan S. Evaluating the performance of ChatGPT in medical physiology university examination of phase I MBBS. Adv Physiol Educ. 2023; 47: 270-271.

13 Das D, Kumar N, Longjam LA, et al. Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. Cureus. 2023; 15: e36034.

14 Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof. 2023; 20: 1.

15 Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology. 2023; 307: e230582.

16 Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery examination? Orthopaedic residents versus ChatGPT. Clin Orthop Relat Res. 2023; 481: 1623-1630.

17 Suchman K, Garg S, Trindade AJ. Chat Generative Pretrained Transformer fails the multiple-choice American College of Gastroenterology self-assessment test. Am J Gastroenterol. 2023; 10: 14309.

18 Fijačko N, Gosak L, Štiglic G, et al. Can ChatGPT pass the life support exams without entering the American Heart Association course? Resuscitation. 2023; 185: 109732.

19 Panthier C, Gatinel D. Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: a novel approach to medical knowledge assessment. J Fr Ophtalmol. 2023; 46: 706-711.

20 Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023; 183: 589.