Original Article

SAGE Open Medicine

SAGE Open Medicine
Volume 12: 1–7
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20503121241257777
journals.sagepub.com/home/smo
Sage

ChatGPT-3.5 passes Poland's medical final examination—Is it possible for ChatGPT to become a doctor in Poland?

Szymon Suwała[1], Paulina Szulc[2], Cezary Guzowski[2], Barbara Kamińska[2], Jakub Dorobiała[2], Karolina Wojciechowska[2], Maria Berska[2], Olga Kubicka[2], Oliwia Kosturkiewicz[2], Bernadetta Kosztulska[2], Alicja Rajewska[2] and Roman Junik[1]

Abstract

Objectives: ChatGPT is an advanced chatbot based on Large Language Model that has the ability to answer questions. Undoubtedly, ChatGPT is capable of transforming communication, education, and customer support; however, can it play the role of a doctor? In Poland, prior to obtaining a medical diploma, candidates must successfully pass the Medical Final Examination.

Methods: The purpose of this research was to determine how well ChatGPT performed on the Polish Medical Final Examination, which passing is required to become a doctor in Poland (an exam is considered passed if at least 56% of the tasks are answered correctly). A total of 2138 categorized Medical Final Examination questions (from 11 examination sessions held between 2013–2015 and 2021–2023) were presented to ChatGPT-3.5 from 19 to 26 May 2023. For further analysis, the questions were divided into quintiles based on difficulty and duration, as well as question types (simple A-type or complex K-type). The answers provided by ChatGPT were compared to the official answer key, reviewed for any changes resulting from the advancement of medical knowledge.

Results: ChatGPT correctly answered 53.4%–64.9% of questions. In 8 out of 11 exam sessions, ChatGPT achieved the scores required to successfully pass the examination (60%). The correlation between the efficacy of artificial intelligence and the level of complexity, difficulty, and length of a question was found to be negative. AI outperformed humans in one category: psychiatry (77.18% vs. 70.25%, $p = 0.081$).

Conclusions: The performance of artificial intelligence is deemed satisfactory; however, it is observed to be markedly inferior to that of human graduates in the majority of instances. Despite its potential utility in many medical areas, ChatGPT is constrained by its inherent limitations that prevent it from entirely supplanting human expertise and knowledge.

Keywords

ChatGPT, education, medicine, public health

Date received: 21 January 2024; accepted: 8 May 2024

Introduction

Artificial intelligence (AI) has made noteworthy progress in contemporary times and has gained increasing prevalence in diverse domains.[1] The recent emergence of ChatGPT, a chatbot based on Large Language Model created by OpenAI, has been a significant development in recent months. The ChatGPT platform was unveiled and made publicly available on 30 November 2022.[2] In just 2 months after its initial release, ChatGPT amassed 1 million users in 5 days and 100 million

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

Corresponding author:
Szymon Suwała, Department of Endocrinology and Diabetology, Nicolaus Copernicus University, Collegium Medicum, 9 Sklodowskiej-Curie Street, Bydgoszcz, 85-094, Poland.
Email: lekarz.szymon.suwala@gmail.com

Table 1. Thematic structure of medical final examination.

Category	Number of questions (%)	Additional requirements
Internal medicine	39 (19.5%)	Among tasks in these fields, at least 30 oncology-related questions (⩾21.3% of questions in these fields, ⩾15% of all questions in examination)
Pediatrics	29 (14.5%)	
Surgery	27 (13.5%)	
Obstetrics and gynecology	26 (13.0%)	
Family medicine	20 (10.0%)	
Psychiatry	14 (7.0%)	No additional requirements
Emergency medicine and intensive care	20 (10.0%)	
Bioethics and medical law	10 (5.0%)	
Medical jurisprudence	7 (3.5%)	
Public health	8 (4.0%)	

users, making it the application with the most rapid growth in history.[3] Primary objective of ChatGPT is to produce responses that are both contextually relevant and logically connected. The implementation of ChatGPT has garnered attention in domains that have conventionally relied on human ingenuity and efficiency, such as marketing, education, and customer service. Research has demonstrated the efficacy of ChatGPT in answering questions on authorized exams in various professions, including medicine. One of the first studies on this topic established that ChatGPT has the capacity to achieve a passing score (or near passing score) on the United States Medical Licensing Examination (USMLE)—three-step examination program for medical licensure in the United States.[4]

The Medical Final Examination is the Polish equivalent of the USMLE—successful completion of the exam enables candidates to apply for a license to practice medicine in Poland (as well as in the European Union, as per Directive 2005/36/EC of the European Parliament).[3] According to current law, final-year medical students or graduates of medical schools are eligible to undertake the examination.[5] The Medical Final Examination comprises a total of 200 test questions, covering various medical specialties such as internal medicine, pediatrics, surgery, obstetrics and gynecology, psychiatry, family medicine, emergency medicine and intensive care, bioethics and medical law, medical jurisprudence and public health (Table 1 contains data regarding the distribution of queries throughout different sections, including oncological topics)—attaining a minimum of 56% of the maximum achievable points is a prerequisite for successfully passing the examination.[5,6] The outcome of the Medical Final Examination holds significant importance for physicians, as it not only confers complete medical practice privileges but also serves as a pivotal factor in their selection for future specialized training programs.[7]

Since ChatGPT was capable of passing the USMLE and "becoming a doctor" in the United States, would it also be able to do so in Poland?

Material and methods

This study aimed to determine whether ChatGPT chatbot could pass the Medical Final Examination, which is required for practicing medicine in Poland—an exam is considered passed if at least 56% of the tasks are answered correctly.

To achieve this, ChatGPT, version 3.5 was presented with questions from 11 examination sessions held between 2013–2015 and 2021–2023 (the content and statistics of which were disclosed by the exam organizer, the Medical Examination Center; questions from the period 2016–2020 and their statistics were not publicly available at the time this article was written due to regulations).[8] The selection of ChatGPT-3.5 was primarily influenced by the limited timeframe between the deployment of ChatGPT-4 and the commencement of our research, constraints on research funding, and the decision made by the authors to make the chatbot accessible to all, without any exceptions, at no cost. ChatGPT was presented with a total of 2138 unique questions from 19 to 26 May 2023, in the form of 11 tests containing 192 to 198 questions (from each test, which normally contained 200 questions, authors excluded some questions from the original set, due to inconsistencies, errors, outdated content, or the need for figure analysis)—the questions in each test were provided in a sequential way, following the same order as in the actual exam, in the same chat window for each test session, without utilizing special, individual prompts or updates (as in a genuine examination). The answers provided by ChatGPT were compared to the official answer key, which had been reviewed for any changes resulting from the advancement of medical knowledge. The Supplemental File depicts an initial phase of the procedure involving the copying of questions from the exam organizer's website and their submission to ChatGPT in order to retrieve responses.

In order to facilitate later analysis, we classified all questions based on the different domains of the examination (internal medicine, pediatrics, surgery, obstetrics and gynecology, family medicine, emergency medicine and intensive care, psychiatry, bioethics and medical law, medical jurisprudence, and public health), into A-type and K-type assignments (as per the regulations of the Medical Examinations Center, A-type assignments require a single correct response, while K-type assignments require the correct set of statements[9] examples of both types of assignments are provided in Table 2), true or false

Table 2. Examples of A-type and K-type tasks used medical final examination in Poland.

A-type task	The most common cause of cancer-related deaths in women in Poland is: A. lung cancer; B. breast cancer; C. cervical cancer; D. ovarian cancer; E. colon cancer
K-type task	The following may lead to hypokalemia: (1) severe vomiting; (2) laxative-induced diarrhea; (3) loop diuretics; (4) alkalosis; (5) acidosis. The correct answer is: A. 1, 2, 3, 4; B. 1, 2, 3, 5; C. only 3; D. 1, 2, 3; E. all listed

Table 3. Detailed results of ChatGPT and humans in different sessions of medical final examination.

Session	% of correct answers (ChatGPT)	% of correct answers (humans)	*p*-Value
Spring 2013	53.40% (fail)	69.34%	<0.001
Fall 2013	64.59% (pass)	70.16%	0.142
Spring 2014	53.33% (fail)	69.52%	<0.001
Fall 2014	60.62% (pass)	70.73%	0.009
Spring 2015	57.58% (pass)	70.28%	0.001
Fall 2015	54.74% (fail)	68.53%	<0.001
Spring 2021	59.00% (pass)	81.95%	<0.001
Fall 2021	64.95% (pass)	82.87%	<0.001
Spring 2022	57.95% (pass)	80.87%	<0.001
Fall 2022	58.16% (pass)	85.03%	<0.001
Spring 2023	60.31% (pass)	83.13%	<0.001

statements expected from the question (e.g., "true statements are. . .," "the most likely is. . .," and "false statements are. . .," "the least likely is. . .") and theoretical or clinical nature of question. All answers of test-takers, the percentage of test-takers who selected the correct answer, and the question's difficulty index (ranging from 0 to 1, with a lower index indicating a more difficult question, according to the definition of Nitko adopted by Johari et al.[10]) were extracted from the Medical Examination Center's data.[8] The obtained data were statistically analyzed utilizing the mean and standard deviation, Student's *t*-tests, ANOVA, and Pearson's correlation coefficient due to the results of the Shapiro–Wilk tests, which indicated normal distribution of the variables. With a significance level of $\alpha = 0.05$, the entire analysis was conducted using Microsoft Excel and STATISTICA 13.0 software (TIBCO Software Inc., Palo Alto, CA, USA).

Results

A total of 2138 tasks were submitted to ChatGPT, with an average difficulty index of 0.744 ± 0.209. 84.85% of questions were classified as A-type—these questions were found to have a significant difference in difficulty when compared to K-type questions (0.752 ± 0.206 vs. 0.696 ± 0.219; $p = 0.031$). 1834 questions (85.78%) were found to contain true statements, and 431 (20.16%) included a patient case report—no significant differences were observed in the difficulty index for these two categories, as evidenced by *p*-values of 0.334 and 0.876, respectively. The mean length of questions was 380 ± 191 characters, and there was no significant correlation between question length and difficulty index ($p = 0.227$).

ChatGPT demonstrated a success rate of 58.61% of all questions, whereas human physicians achieved 75.60% ($p < 0.001$). We found a negative correlation between the length of questions (−0.069; $p = 0.001$) and task difficulty (0.265; $p < 0.001$) with the accuracy of AI-generated responses—ChatGPT's performance was poorer on longer and more challenging tasks.

An analysis was conducted on the percentage of correct answers provided by AI and doctors, divided into 11 examination sessions. In three sessions (Spring 2013, Spring 2014, and Fall 2015), ChatGPT was unable to attain the mandated passing threshold of 56%. Results attained by physicians were notably superior to those achieved by AI in almost every session (except Fall 2013). Detailed results are presented in Table 3.

The accuracy rates of both AI and human participants were compared in specific domains of tasks. The ChatGPT system exhibited the best performance in questions related to public health (82.56%) and psychiatry (77.18%). Although the absolute values in psychiatry-related tasks were more advantageous for AI, the difference was not statistically significant. In the remaining domains (except public health), physicians exhibited a significant and noteworthy advantage. Table 4 presents the complete results.

The study revealed that ChatGPT exhibited a noteworthy performance disparity between type-A and type-K questions, with a higher accuracy rate for type-A questions compared to type-K questions (61.69% vs. 40.99%; $p < 0.001$). A similar trend was observed for questions focused on identifying true statements rather than false ones (59.54% vs. 52.96%; $p = 0.031$). It is noteworthy that in the human population, there was an opposite trend observed in relation to type-A

Table 4. Detailed results of ChatGPT and humans in different domains of tasks in medical final examination.

Discipline	% of correct answers (ChatGPT)	% of correct answers (humans)	*p*-Value
Internal medicine	59.57%	77.88%	<0.001
Pediatrics	54.69%	71.22%	<0.001
Surgery	52.94%	76.88%	<0.001
Obstetrics and gynecology	57.44%	74.81%	<0.001
Psychiatry	77.18%	70.25%	0.081
Family medicine	55.81%	73.21%	<0.001
Emergency medicine and intensive care	58.54%	72.25%	<0.001
Bioethics and medical law	59.57%	76.93%	<0.001
Medical jurisprudence	40.79%	84.51%	<0.001
Public health	82.56%	83.01%	0.922

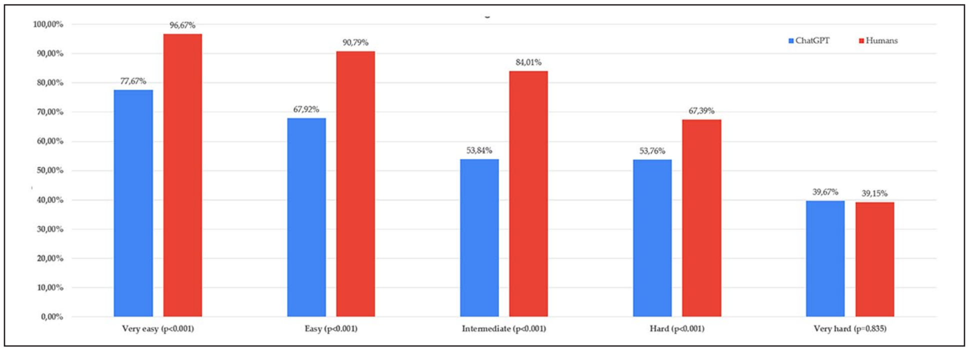

Figure 1. Results of ChatGPT and humans in questions grouped based on difficulty index quintiles.

and type-K questions (75.20% vs. 77.71%), but the difference was only marginally significant ($p = 0.062$). Physicians also prefer questions that aim to identify true statements over those that aim to identify false ones, but the difference was not statistically significant (75.03% vs. 74.69%; $p = 0.634$). There was no significant disparity observed in the accuracy of responses provided by ChatGPT and human participants, irrespective of whether the question pertained to a theory or a case report ($p = 0.615$ and 0.975).

All presented questions were classified into five equal quintiles based on their difficulty index. These quintiles were categorized as very easy (difficulty index ranging from 0.923 to 0.996), easy (0.855–0.922), intermediate (0.748–0.854), hard (0.567–0.747), and very hard (less or equal to 0.566). Subsequently, the accuracy rates of the responses were juxtaposed—consistent with prior findings regarding the correlation between difficulty index and performance, ChatGPT exhibited optimal proficiency on tasks categorized as very easy while demonstrating worse efficacy on challenging items. It is noteworthy that the responses furnished by ChatGPT did not exhibit any statistically significant (in *t*-test) between questions from the third and fourth quintiles, which correspond to intermediate and difficult questions, respectively ($p = 0.979$). Remarkably, despite the lack of statistical significance, AI exhibited a slightly superior performance compared to physicians in tackling the most challenging questions. Figure 1 presents the full results.

An interesting qualitative observation pertains to the selection of answer choices by ChatGPT. Despite the presence of distinct answer options labeled A, B, C, D, and E (with only one being correct), it was observed that in 64 tasks (2.99%), AI did not select any answer or indicate multiple answers as correct. The study findings indicate that a significantly increased risk of such situations was associated with A-type questions (RR 3.61; 95% CI: 1.14–11.43) and questions aimed at detecting false statements (RR 2.36; 95% CI: 1.39–4.02).

Discussion

AI has garnered worldwide attention (ChatGPT reached 100 million monthly active users just 2 months after launch, making it the fastest-growing consumer application in history[11]), but it has also raised concerns about algorithms and machines replacing human labor, a concern that has persisted since the industrial revolution.[12] Medicine requires a thorough approach to all issues and a vast knowledge base, especially for doctors who want to help their patients.[13,14] This study did not assess ChatGPT's therapeutic efficacy—however, given the rapid pace of AI, passing an examination that allows autonomous practice could be a worthwhile first step for future discussions in this area.

Kung demonstrated that ChatGPT had the ability to successfully complete the USMLE without any prior training.[4] The equivalent of this test in Poland is the Medical Final Examination, which, however, differs in that, unlike the USMLE, it is a single-component exam consisting solely of multiple-choice questions (with five answer choices per question).[5] Nevertheless, in our study ChatGPT proved able to handle it at least as well. However, ChatGPT did not pass MGRCGP:AKT (the Applied Knowledge Test of the Membership of the Royal College of General Practitioners)[15] and the Chinese National Medical Licensing Examination.[16] In Germany, ChatGPT in version 3.5 achieved a pass rate of medical license examination in one out of three cases, whereas in version 4.0 (which had a notable technological edge), it achieved a perfect success rate.[17] During a comparative evaluation conducted in Japan, ChatGPT-4 exhibited a superior performance ranging from 27.6% to 36.3%, depending on the specific question category in the test.[18] The efficacy of ChatGPT (in both versions) has also been demonstrated in other medical examinations, including medical biochemistry,[19] physiology,[20] microbiology,[21] parasitology,[22] as well as the European Exam in Core Cardiology[23] or Ophthalmic Knowledge Assessment Program (OKAP) exam.[24] Nonetheless, AI was unsuccessful in passing the American Heart Association Basic Life Support and Advanced Cardiovascular Life Support exam,[25] in Poland, also specialization exams in internal medicine[26] or radiology.[27] Determining the cause of substantial variations in the efficacy of ChatGPT, even within the same version (especially when it comes to GPT-4), is a challenging task. One potential option could be language differences. However, a study conducted by Panthier et al. examining the efficacy of ChatGPT in the French version of the European Board of Ophthalmology Examination indicated that the main factor influencing its effectiveness was not the language used;[24] nonetheless, it is crucial to consider the contrasting global significance and prevalence of French and Polish (309.8 million French speakers vs. 40.6 million Polish speakers according to Ethnologue data[28]). Additional observations and investigation are necessary.

Upon analysis of the examination results of human test-takers, a notable disparity in the proportion of accurate responses exists between the timeframes of 2013–2015 and post-2021, a trend that is not apparent in AI. The reason for this is probably that, as per the current regulations in Poland, a considerable segment of the Medical Final Examination (minimum 70%) comprises queries sourced from an open question bank, allowing for prior training. Article 14c of the Act on the Profession of Physician and Dentist, which was implemented at that time, resulted in a scenario where only 30% of new questions were created for each subsequent Medical Final Examination date, which was not previously accessible to the candidates.[29] Thus, this cannot be deemed a substantiation of ChatGPT's diminished efficacy subsequent to 2021; although it is known that ChatGPT's knowledge is limited to the year 2021, it is important to note that in our study, every question has been examined in relation to its adherence to current medical knowledge during this time.

ChatGPT demonstrated superior performance in the fields of public health and psychiatry, with accuracy rates of 82.56% and 77.18%, respectively. Interestingly, humans achieved their lowest performance in psychiatry tests (70.25%)—psychiatry is the only field in the Medical Final Examination in which ChatGPT-3.5 performed better than real doctors. The cause of this situation is still unknown to us, but a review of the available literature draws attention to the fact that aspects related to the psyche, psychology, and emotions appear to be the notable strengths of chatbots like ChatGPT. Franco D'Souza et al.[30] showed that ChatGPT-3.5 fred extremely well in clinical vignettes in psychiatry by receiving 61% grades "A," 31% "B," and only 8% "C." In the study by Elyoseph et al.[31] it has been proven that ChatGPT demonstrated significantly higher performance than the general population on all the Levels of Emotional Awareness Scale (LEAS) and can further improve its result. Of course, we must keep in mind that we are dealing only with a chatbot—Levkovich and Elyoseph[32] showed that ChatGPT-3.5 was able to underestimate the risk of suicide even in high-risk patients.

The ability to provide immediate solutions to inquiries is a chance to improve quality for medical practitioners, patients, and healthcare professionals. Notwithstanding, it appears highly unlikely that AI will have the capability to substitute medical practitioners in the immediate future. Even the most sophisticated algorithms and AI-enabled technologies cannot diagnose and cure illnesses, as DiGiorgio and Ehrenfeld[33] accurately noted. Our study demonstrated that ChatGPT has the potential to successfully clear the medical licensing examination in Poland—however, it is crucial to consider that medicine is not just a precise science but also an art that necessitates the application of critical thinking beyond algorithms. Additionally, it is important to emphasize the significance of utilizing an individualized approach to patient care that is based on interpersonal communication and knowledge. However, there exists a prospective application for ChatGTP or AI in the field of medicine, like the analysis of big or the creation of realistic descriptions of clinical cases, which serve as effective tools for students to learn and prepare for their profession.[34] It is interesting to mention the courteousness exhibited by AI and its prospective utilization in routine clinical practice—it has been demonstrated that in 79% of cases, patients perceived ChatGPT's responses to their urgent medical inquiries as being more empathetic and comprehensive when compared to those provided by human professionals.[35] On the other hand, ChatGPT's ability to empathize may influence our perception of chatbot mistakes, therefore warranting a sensible and careful approach to its actions.

Our research has a few limitations. Analysis was limited to the evaluation of ChatGPT's performance without conducting any comparative assessments with other AI or chatbot models. Additionally, it should be noted that ChatGPT undergoes regular updates—as previously stated in this article, employing ChatGPT-4 yields superior quality; however, it is a paid tool and not accessible to all individuals. It would be prudent to assess the efficacy of both ChatGPT versions 3.5 and 4.0 on a comparably extensive range of inquiries (similar work was done by Rosoł et al.,[3] showing the superiority of ChatGPT-4, but the study was based on a small number of questions) and potentially juxtapose them with other chatbots. In our work, we did not use any prompts that might have in any way influenced the effectiveness of the answers. Despite these limitations, our study provides significant perspectives on the advantages and limitations of ChatGPT in the setting of medical licensure examinations, such as the USMLE or the Polish Medical Final Examination.

Conclusions

The results of this study demonstrated the potential effectiveness of ChatGPT version 3.5 as an approach to passing the Medical Final Examination. There is evidence to suggest that ChatGPT and maybe other AI language models, despite their limitations, could be a valuable asset in patient care. The efficacy of the GPT3.5 model, while enough for passing the exam, was subpar and inferior to the performance of medical students and early-career doctors. To enhance proficiency in this domain, it is advisable to pursue more training for these models.

Acknowledgements

None.

Authors contribution

Szymon Suwała: conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing, visualization, project administration, funding acquisition. Paulina Szulc: conceptualization, investigation, formal analysis, data curation, writing. Cezary Guzowski: conceptualization, investigation, formal analysis, data curation, writing. Barbara Kamińska: conceptualization, investigation, formal analysis, data curation, writing. Jakub Dorobiała: investigation, formal analysis, data curation, writing. Karolina Wojciechowska: investigation, formal analysis, data curation, writing. Maria Berska: investigation, formal analysis, data curation, writing. Olga Kubicka: investigation, formal analysis, data curation, writing. Oliwia Kosturkiewicz: investigation, formal analysis, data curation, writing. Bernadetta Kosztulska: investigation, formal analysis, writing. Alicja Rajewska: investigation, formal analysis, writing. Roman Junik: conceptualization, methodology, resources, writing, project administration, funding acquisition, supervision.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Szymon Suwała https://orcid.org/0000-0002-5865-8484

Supplemental material

Supplemental material for this article is available online.

References

1. Farhi F, Jeljeli R, Aburezeq I, et al. Analyzing the students' views, concerns, and perceived ethics about chat GPT usage. *Comput Educ Artif Intell* 2023; 5: 100180.
2. Jeon J and Lee S. Large language models in education: a focus on the complementary relationship between human teachers and ChatGPT. *Educ Inf Technol (Dordr)* 2023; 28: 15873–15892.
3. Rosoł M, Gąsior JS, Łaba J, et al. Evaluation of the performance of GPT-3.5 and GPT-4 on the polish medical final examination. *Sci Rep* 2023; 13: 20512.
4. 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.
5. The Medical Examinations Center. Information about polish medical final examination. https://www.cem.edu.pl/english_info_lek.php (2023, accessed 19 May 2023).
6. The Medical Examinations Center. Thematic structure of the test—medical final examination. https://www.cem.edu.pl/english_lek_struktura.php (2023, accessed 19 May 2023).
7. Gąsiorowski J, Rudowicz E and Safranow K. Motivation towards medical career choice and future career plans of Polish medical students. *Adv Health Sci Educ* 2015; 20: 709–725.
8. The Medical Examinations Center. Question database—medical final examination. https://cem.edu.pl/pytcem/eula_lek_p.php (2023, accessed 19 May 2023).
9. The Medical Examinations Center. Instruction for composing questions in examinations. https://cem.edu.pl/spz/Instrukcja2016.pdf (2023, accessed 19 May 2023).
10. Johari J, Sahari J, Wahab DA, et al. Difficulty index of examinations and their relation to the achievement of program outcomes. *Procedia Soc Behav Sci* 2011; 18: 71–80.
11. 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.
12. Ahmadi A. ChatGPT: exploring the threats and opportunities of artificial intelligence in the age of chatbots. *Asian J Comput Sci Technol* 2023; 12: 25–30.
13. Sedaghat S. Success through simplicity: what other artificial intelligence applications in medicine should learn from history and ChatGPT. *Ann Biomed Eng* 2023; 51: 2657–2658.
14. Sedaghat S. Future potential challenges of using large language models like ChatGPT in daily medical practice. *J Am Coll Radiol* 2024; 21: 344–345.
15. Thirunavukarasu AJ, Hassan R, Mahmood S, et al. Trialling a large language model (ChatGPT) in general practice with

the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. *JMIR Med Educ* 2023; 9: e46599.

16. Wang X, Gong Z, Wang G, et al. ChatGPT performs on the Chinese National Medical Licensing Examination. *J Med Syst* 2023; 47: 86.
17. Meyer A, Riese J and Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written german medical licensing examination: observational study. *JMIR Med Educ* 2024; 10: e50965.
18. Takagi S, Watari T, Erabi A, et al. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. *JMIR Med Educ* 2023; 9: e48002.
19. Ghosh A and Bir A. Evaluating ChatGPT's ability to solve higher-order questions on the competency-based medical education curriculum in medical biochemistry. *Cureus* 2023; 15(4): e37023.
20. Subramani M, Jaleel I and Krishna Mohan S. Evaluating the performance of ChatGPT in medical physiology university examination of phase I MBBS. *Adv Physiol Educ* 2023; 47: 270–271.
21. 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.
22. 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.
23. 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 Digital Health* 2023; 4: 279–281.
24. Panthier C and 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(7): 706–711.
25. Long MT, Pedley A, Massaro JM, et al. The association between non-invasive hepatic fibrosis markers and cardiometabolic risk factors in the framingham heart study. *PLoS One* 2016; 11: e0157517.
26. Suwała S, Szulc P, Dudek A, et al. ChatGPT fails the internal medicine state specialization exam in Poland: artificial intelligence still has much to learn. *Pol Arch Intern Med* 2023; 133(11): 16608.
27. Bhayana R, Krishna S and Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology* 2023; 307: e230582.
28. Ethnologue. What are the top 200 most spoken languages? https://www.ethnologue.com/insights/ethnologue200/ (2024, accessed 01 February 2024).
29. Act on the profession of physician and dentist. https://lexlege.pl/ustawa-o-zawodach-lekarza-i-lekarza-dentysty/rozdzial-2a-lekarski-egzamin-koncowy-i-lekarsko-dentystyczny-egzamin-koncowy/5721/ (2023, accessed 01 February 2024).
30. Franco D'Souza R, Amanullah S, Mathew M, et al. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. *Asian J Psychiatr* 2023; 89: 103770.
31. Elyoseph Z, Hadar-Shoval D, Asraf K, et al. ChatGPT outperforms humans in emotional awareness evaluations. *Front Psychol* 2023; 14: 1199058.
32. Levkovich I and Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. *JMIR Ment Health* 2023; 10: e51232.
33. DiGiorgio AM and Ehrenfeld JM. Artificial intelligence in medicine and ChatGPT: De-Tether the physician. *J Med Syst* 2023; 47: 32.
34. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* 2023; 11: 887.
35. 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.