








ChatGPT – a tool for assisted studying or a source of misleading medical information? AI performance on Polish Medical Final Examination

ChatGPT – pomoc naukowa przyszłości czy źródło fałszywych informacji?
Analiza odpowiedzi sztucznej inteligencji na przykładzie zadań
Lekarskiego Egzaminu Końcowego

Karol Żmudka¹ , Aleksandra Spychał¹ , Błażej Ochman² , Łukasz Popowicz³ , Patrycja Piłat³,
Jerzy Jaroszewicz¹ 

¹Department of Infectious Diseases and Hepatology, Faculty of Medical Sciences in Zabrze,
Medical University of Silesia, Katowice, Poland

²Department of Medical and Molecular Biology, Faculty of Medical Sciences in Zabrze,
Medical University of Silesia, Katowice, Poland

³Department of Psychiatry, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Katowice, Poland

ABSTRACT

INTRODUCTION: ChatGPT is a language model created by OpenAI that can engage in human-like conversations and generate text based on the input it receives. The aim of the study was to assess the overall performance of ChatGPT on the Polish Medical Final Examination (Lekarski Egzamin Końcowy – LEK) the factors influencing the percentage of correct answers. Secondly, investigate the capabilities of chatbot to provide explanations was examined.

MATERIAL AND METHODS: We entered 591 questions with distractors from the LEK database into ChatGPT (version 13th February – 14th March). We compared the results with the answer key and analyzed the provided explanation for logical justification. For the correct answers we analyzed the logical consistency of the explanation, while for the incorrect answers, the ability to provide a correction was observed. Selected factors were analyzed for an influence on the chatbot's performance.

RESULTS: ChatGPT achieved impressive scores of 58.16%, 60.91% and 67.86% allowing it pass the official threshold of 56% in all instances. For the properly answered questions, more than 70% were backed by a logically coherent explanation. In the case of the wrongly answered questions the chatbot provided a seemingly correct explanation for false information in 66% of the cases. Factors such as logical construction ($p < 0.05$) and difficulty ($p < 0.05$) had an influence on the overall score, meanwhile the length ($p = 0.46$) and language ($p = 0.14$) did not.

Received: 14.09.2023

Revised: 04.11.2023

Accepted: 06.12.2023

Published online: 16.04.2024

Address for correspondence: Karol Żmudka, Katedra i Klinika Chorób Zakaźnych i Hepatologii, Górnośląskie Centrum Medyczne im. prof. Leszka Gieca SUM, ul. Ziolowa 45/47, 40-635 Katowice, tel. +48 518 435 315, e-mail: s78586@365.sum.edu.pl



This is an open access article made available under the terms of the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license, which defines the rules for its use. It is allowed to copy, alter, distribute and present the work for any purpose, even commercially, provided that appropriate credit is given to the author and that the user indicates whether the publication has been modified, and when processing or creating based on the work, you must share your work under the same license as the original. The full terms of this license are available at <https://creativecommons.org/licenses/by-sa/4.0/legalcode>.

Publisher: Medical University of Silesia, Katowice, Poland



CONCLUSIONS: Although achieving a sufficient score to pass LEK, ChatGPT in many cases provides misleading information backed by a seemingly compelling explanation. The chatbot can be especially misleading for non-medical users as compared to a web search because it can provide instant compelling explanations. Thus, if used improperly, it could pose a danger to public health. This makes it a problematic recommendation for assisted studying.

KEYWORDS

artificial intelligence, public health, machine learning

STRESZCZENIE

WSTĘP: ChatGPT jest modelem językowym stworzonym przez OpenAI, który może udzielać odpowiedzi na zapytania użytkownika, generując tekst na podstawie otrzymanych danych. Celem pracy była ocena wyników działania ChatGPT na polskim Lekarskim Egzaminie Końcowym (LEK) oraz czynników wpływających na odsetek prawidłowych odpowiedzi. Ponadto zbadano zdolność chatbota do podawania poprawnego i wnikliwego wyjaśnienia.

MATERIAŁ I METODY: Wprowadzono 591 pytań z dystraktorami z bazy LEK do interfejsu ChatGPT (wersja 13 lutego – 14 marca). Porównano wyniki z kluczem odpowiedzi i przeanalizowano podane wyjaśnienia pod kątem logicznego uzasadnienia. Dla poprawnych odpowiedzi przeanalizowano spójność logiczną wyjaśnienia, natomiast w przypadku odpowiedzi błędnej obserwowano zdolność do poprawy. Wybrane czynniki zostały przeanalizowane pod kątem wpływu na zdolność chatbota do udzielenia poprawnej odpowiedzi.

WYNIKI: ChatGPT osiągnął imponujące wyniki poprawnych odpowiedzi na poziomie: 58,16%, 60,91% i 67,86%, przekraczając oficjalny próg 56% w trzech ostatnich egzaminach. W przypadku poprawnie udzielonych odpowiedzi ponad 70% pytań zostało popartych logicznie spójnym wyjaśnieniem. W przypadku błędnych odpowiedzi w 66% przypadków chatbot podał pozornie poprawne wyjaśnienie dla nieprawidłowych odpowiedzi. Czynniki takie jak konstrukcja logiczna ($p < 0,05$) i wskaźnik trudności zadania ($p < 0,05$) miały wpływ na ogólną ocenę, podczas gdy liczba znaków ($p = 0,46$) i język ($p = 0,14$) takiego wpływu nie miały.

WNIOSKI: Mimo iż ChatGPT osiągnął wystarczającą liczbę punktów, aby zaliczyć LEK, w wielu przypadkach podawał wprowadzające w błąd informacje poparte pozornie przekonującym wyjaśnieniem. Chatboty mogą być szczególnym zagrożeniem dla użytkownika niemającego wiedzy medycznej, ponieważ w porównaniu z wyszukiwarką internetową dają natychmiastowe, przekonujące wyjaśnienie, co może stanowić zagrożenie dla zdrowia publicznego. Z tych samych przyczyn ChatGPT powinien być ostrożnie stosowany jako pomoc naukowa.

SŁOWA KLUCZOWE

sztuczna inteligencja, zdrowie publiczne, nauczanie maszynowe

INTRODUCTION

ChatGPT is a natural language processing (NLP) system that allows users to engage in human-like conversations and generate text based on the input it receives. It can work with multiple languages including Polish and is capable of analyzing image-based data, then generate text upon the input it receives. The current free version of the tool uses GPT-3.5 (a generative pre-trained transformer), although there is the fee-based GPT-4. The subscription version is less popular, but offers substantially more accurate answers to both general and clinical tasks. Current research reveals that artificial intelligence can be used both to answer clinical questions but may also be utilized for educational purposes [1,2]. We believe that using the less insightful but substantially more popular free version would make a better model for a user who occasionally makes use of the chatbot. When it comes to the ability to answer clinical questions, there are numerous reports on the performance on different steps of USMLE (the United States Medical Licensing Examination). ChatGPT

based on GPT-3.5 made an important milestone in the development of AI with its ability to pass this exam on many occasions [1,3]. Moreover, reports confirm that the chatbot is able to answer clinical questions from non-English language-based or specialization exams [4,5]. It has also been reported that it can be employed to generate practice questions banks. Having said that, ChatGPT is a tempting tool to answer medical questions. However, ChatGPT on many instances can provide misleading information. In the following paper we will delve into the benefits of using ChatGPT but also analyze the potential threat to public health by assessing the performance of ChatGPT on the Polish Medical Final Examination (Lekarski Egzamin Końcowy – LEK). LEK consists of 200 medical multiple-choice questions on the subjects presented in Figure 1. Various aspects of medical knowledge are tested on the exam, both clinical and not. This makes the LEK database an excellent representation for a user seeking medical knowledge. Moreover, the exact same questions in the Polish and English version makes it a perfect model to test if language affects performance. Each task on LEK consists of a question and 5 distractors from which only one is correct.



Logical construction and the length of the question can differ significantly and potentially influence the performance. Although current reports prove that ChatGPT is indeed able to pass LEK, the knowledge about factors contributing to and the prevalence of misinformation is very limited. What seems to be the most concerning is the phenomenon of justifying wrong answers with a seemingly correct explanation, which poses substantial danger to public health.

We wanted to test the chatbot's general ability to answer medical knowledge associated questions, both clinical and not. Moreover, we wanted to establish if pre-selected factors contribute to the chatbot's performance, such as the language, length, logical construction and difficulty of the task. Lastly, we wanted to test the chatbot's ability to provide an explanation for certain decisions and the capability to correct itself after interaction with a user.

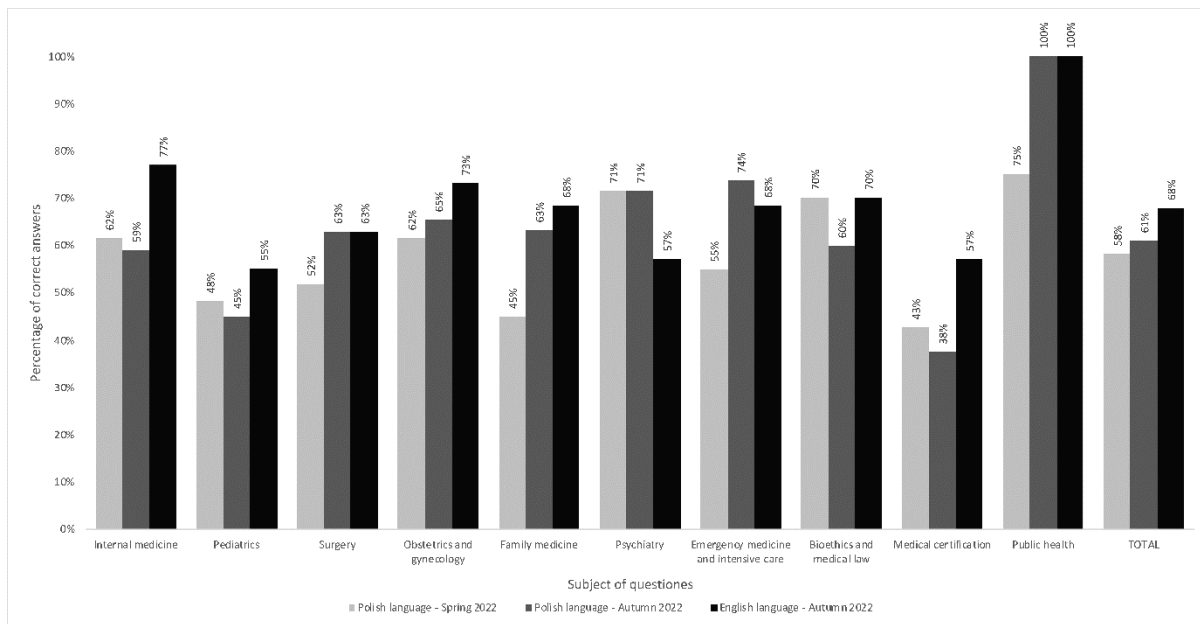


Fig. 1. Bar graph presenting ChatGPT percentage score in certain thematic categories on Spring 2022 Polish and Autumn 2022 Polish and English LEK exam.

Ryc. 1. Wykres słupkowy przedstawiający procentowy wynik ChatGPT w poszczególnych kategoriach tematycznych. Egzamin – wiosna 2022 w języku polskim oraz jesień 2022 w języku polskim i angielskim.

Then, with the gathered information, we analyzed the potential benefits and risks of the broad implementation of chatbots to public health. With the growing popularity of chatbots in the general population, it is likely that progressively more patients will be tempted to use such tools instead of looking for professional medical advice. Thus, using all the collected data we wanted to answer the question: can ChatGPT be a danger for public health or a good tool for assisted studying?

MATERIAL AND METHODS

We used the free ChatGPT versions: 13th February – 14th March which utilize GPT-3.5. Our material included 600 questions that appeared in three of the most recent, at the time of the study, LEK exams. Those exams were the Polish language versions of the LEK exams – Autumn 2022 and Spring 2022 and one English language version of the Autumn 2022 LEK exam. From those, 9 questions were withdrawn by the

LEK organizing committee. In the analysis, we included graphic tasks such as ECG charts. The questions were entered into the ChatGPT dialog text box and the provided answers were compared with the LEK answer key. The thematic structure compatible with the LEK organizing committee categorization is presented in Table I. Although the LEK exam includes only single choice questions, we subdivided them into the groups presented in Table II. This classification is based on the logical construction of the task and was determined by the person sampling the answers. As ChatGPT does not always provide an explanation, it was often necessary to ask for one. The provided explanations were analyzed for logical coherence. For the wrong answers with illogically coherent explanations, we confronted the chatbot and analyzed if the corrected answer was legitimate according to the answer key. All the answers were manually acquired by 4 medical students. The decision to classify a task to a certain logical construction category or if it was logically coherent was made by the person who entered the data into the chatbot and later it was done



independently by second reviewer. In the case of disagreement, it was discussed between the parties and a final decision was made.

Table I. Number of questions and percentage of total number of questions by thematic structure of Autumn 2022 Polish and English version and Polish version of Spring 2022 LEK exam

Tabela I. Liczba i procent ogólnej liczby pytań z podziałem według struktury tematycznej pytania z egzaminu w wersji polskiej i angielskiej – LEK jesień 2022, oraz w wersji polskiej – LEK wiosna 2022

Subject	Autumn 2022 PL/ENG	Spring 2022 PL
Internal medicine	39 (20%)	38 (19%)
Pediatrics	29 (15%)	29 (14%)
Surgery	27 (14%)	27 (14%)
Obstetrics and gynecology	26 (13%)	26 (13%)
Family medicine	19 (10%)	20 (10%)
Psychiatry	14 (7%)	14 (7%)
Emergency medicine and intensive care	19 (10%)	18 (9%)
Bioethics and medical law	10 (5%)	10 (5%)
Medical certification	8 (4%)	7 (4%)
Public health	6 (3%)	8 (4%)
Total	197	197

Table II. Number of questions and percentage of total number of questions by logical construction

Tabela II. Liczba i procent całkowitej liczby pytań z podziałem według konstrukcji logicznej pytania

Logical construction	Autumn 2022 PL/ENG	Spring 2022 PL
Select correct answer	145 (74%)	146 (74%)
Multiple choice	34 (17%)	29 (15%)
Select wrong answer	17 (9%)	22 (11%)
Multiple choice and select wrong answer	1 (< 1%)	0 (0%)
Total	197	197

To analyze the influence of task difficulty, we used the IDI – item difficulty index provided by the organizing committee for which 1 stands for “extremely easy” and 0 – “extremely difficult”. IDI is calculated by the following formula:

$$IDI = \frac{Ns + Ni}{2n}$$

where n is the number of examinees in each of the extreme groups (extreme groups up to 27% of the test takers with the best results and 27% of the test takers with the worst results in the entire group), Ns – the number of correct answers for the analyzed task in the group with the best results, Ni – the number of correct responses for the analyzed task in the group with the worst results. This is the range selected by Medical Examinations Center (Centrum Egzaminów Medycznych – CEM), and it has been defined for the difficulty index [6]. The 27% range for the extreme

group representation is a well-established and statistically justified interval for most types of tests [7]. Statistical analysis was conducted using R Studio statistical software (2022.07.2 Build 576). For quantitative non-normally distributed data we used the Mann–Whitney U test and the chi-square test for categorical variables. To assess the impact of selected variables on the odds of receiving a correct answer or a correct explanation, logistic regression was employed. The tables and graphs were prepared in MS Excel.

RESULTS

ChatGPT was able to pass all the three exams with the following scores: Spring 2022 PL – 56.63%, Autumn 2022 PL – 60.91 and Autumn 2022 ENG – 67.86%. Although we observed a higher score from the English version of the same exam, the proportion of correct answers did not differ by language significantly, χ^2 (df = 1, N = 394) = 2.17, p = 0.14. The results in the thematic subgroups in all the versions of the exam are presented in Figure 1. When comparing the proportion of correct answers in individual thematic categories between the three versions of the exam, no significant difference was observed. Among all the examined disciplines, the highest percentage of correct answers in the two Polish LEK exams was observed in public health (75% in the Spring 2022 exam and 100% in the Autumn 2022 exam) and the lowest in medical certification (43% in the Spring 2022 exam and 38% in the Autumn 2022 exam). However, while comparing the proportion of correct answers in each discipline with the proportion of correct answers in the rest of the questions, we did not observe a significant association between the discipline of the question and the proportion of correct answers (p > 0.05). It should be noted that such results could be affected by the low number of cases in certain thematic categories (Table I). A comparison of difficulty levels of those two thematic categories from the two Polish exams was done. The median IDI among the questions concerning public health equaled 0.9 (IQR = 0.79, 0.93), meanwhile among the questions concerning medical certification it amounted to 0.87 (IQR = 0.8, 0.89). The results indicated that there was no significant difference between the IDI of the medical certification and the public health questions U = 88, p = 0.47. Thus, it suggests that the difference in score is not associated with the different difficulty of questions in those categories but rather some other factors. The proposed explanation for this observation will be discussed later. When taking the total sample of the Polish exam questions, we observed that IDI influenced the proportion of correct answers. The median values of IDI for the correct and wrong answers were respecti-



vely 0.9 (IQR = 0.82, 0.95) and 0.83 (IQR = 0.66, 0.89); the distributions in the two groups differed significantly: Mann–Whitney $U = 10981$, $p < 0.01$ (Fig. 2). There were significantly higher odds of receiving correct answers for tasks with a higher IDI (meaning easier) – the odds ratio [OR] = 1.22, (95% CI [1.09, 1.38]) per 0.1 IDI increase. Nevertheless, when taking the total sample of Polish exam questions, we observed that IDI did not influence the proportion of logically coherent explanations – Mann–Whitney $U = 14164$, $p = 0.89$. Another factor which we believed could affect GPT-3.5 performance was the length of the task defined as the number of characters in the question and in the distractors. We found that there was no significant difference in the length of the task – Mann–Whitney $U = 19086$, $p = 0.46$ (Fig. 3). The last factor affecting the chatbot’s performance was the type of logical construction. The questions were separated into 3 types and later analyzed. One question type was assessed as a combination of 2 main ones, and thus this question was not included in the final analysis. The chi-square test of independence showed that there is a significant association between the type of logical construction and the proportion of correct answers χ^2 (df = 2, N = 387) = 9.217, $p < 0.01$. Figure 4 presents the percentage of correct answers in a particular category and the values of the χ^2 test results for comparison of individual subcategories. We found that

the *select the correct answer* type of questions are significantly more likely to be correct – [OR] = 2.01, (95% CI [1.27, 3.17]), $p < 0.01$. The odds ratio for receiving a correct answer were also calculated for the *select the wrong answer* type of questions [OR] = 0.64, (95% CI [0.33, 1.24]), $p = 0.19$ and the *multiple-choice* questions [OR] = 0.54, (95% CI [0.31, 0.93]), $p = 0.03$. Moreover, we analyzed if the type of question influenced the odds of receiving a logically coherent explanation. We found that in the case of the *multiple-choice* questions, the chances of receiving such an explanation was significantly lower – [OR] = 0.53, (95% CI [0.30, 0.94]), $p = 0.03$. Later, we analyzed the ability of GPT-3.5 to provide a logically coherent explanation and the ability to correct itself in all the questions. The results are presented in a chart (Fig. 5). From 235 correct answers, up to 80% (95% CI [75%, 85%]) were backed by a logically coherent explanation. What is disturbing is the fact that from 159 wrongly answered questions, 66% (95% CI [58%, 73%]) were backed by a seemingly correct explanation. When confronting chatbot about the other 37% illogically coherent explanations, we found that it was able to correct itself 78% (95% CI [64%, 88%]) of the time. Moreover, the answers that were backed by a logically coherent explanation are significantly more likely to be correct – the odds ratio [OR] = 2.11, (95% CI [1.33, 3.35]), $p < 0.01$.

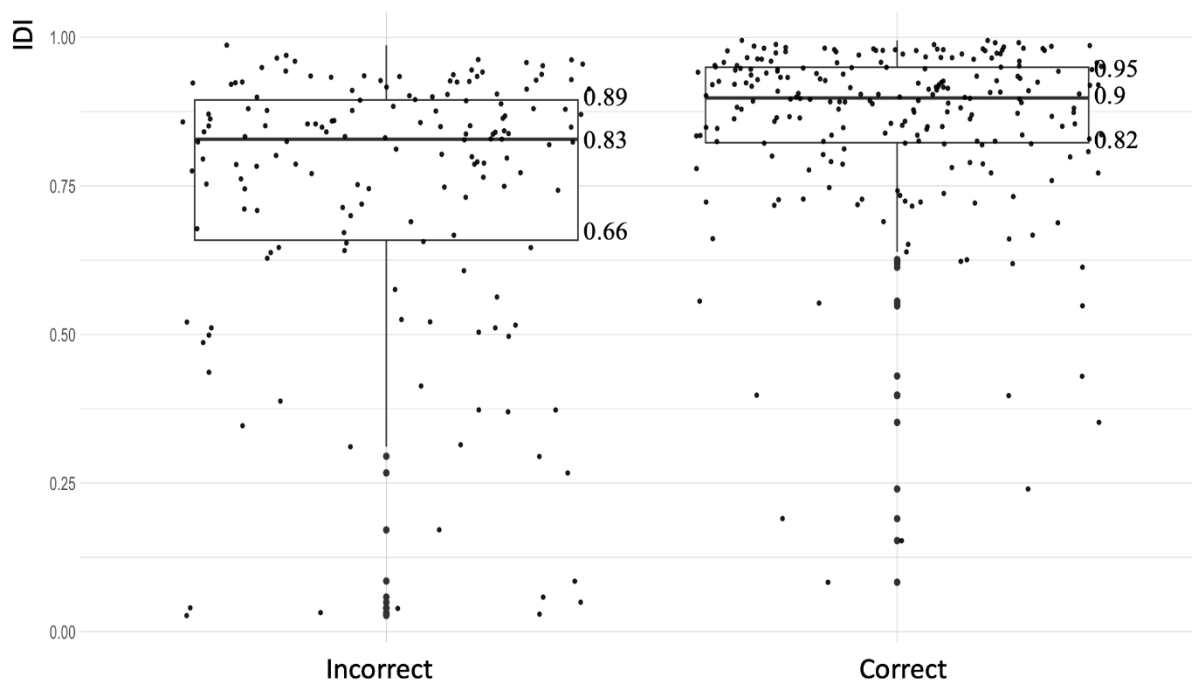


Fig. 2. Distribution of item difficulty index (IDI) in groups of incorrectly and correctly answered questions.
Ryc. 2. Rozkład wartości indeksu poziomu trudności pytania (IDI) w pytaniach, na które udzielono błędnych i prawidłowych odpowiedzi.

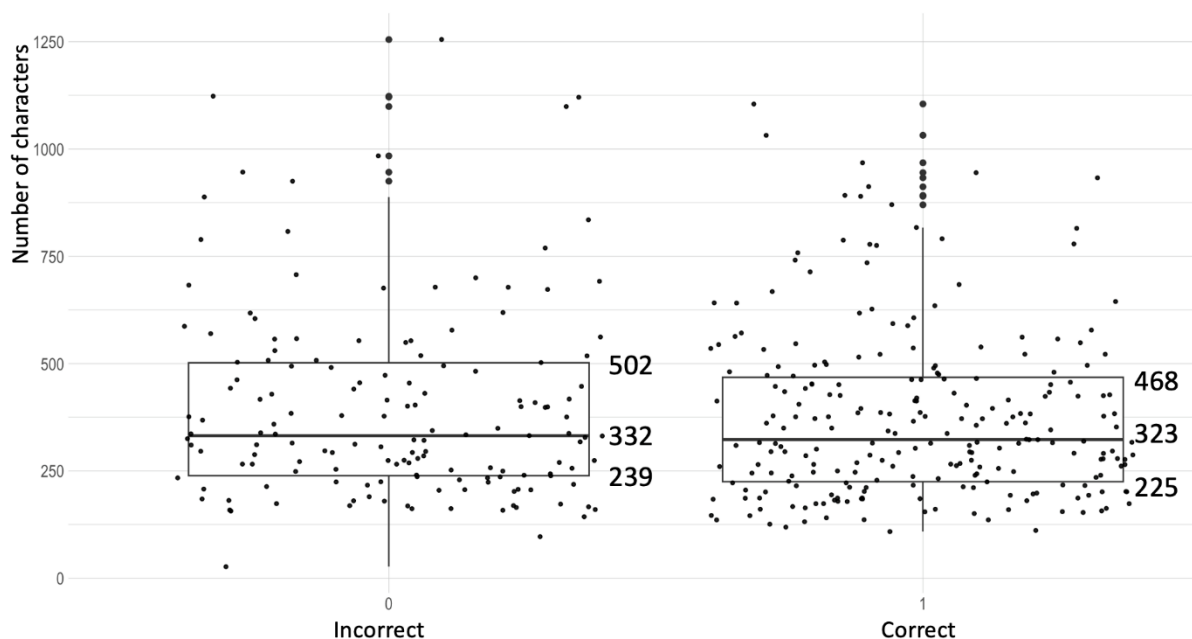


Fig. 3. Distribution of length of task measured as number of characters in question and distractors combined in groups of incorrectly and correctly answered questions.
Ryc. 3. Rozkład długości polecenia z dystraktorami wyrażony poprzez liczbę znaków w poleceniach, dla których udzielono nieprawidłowej i poprawnej odpowiedzi.

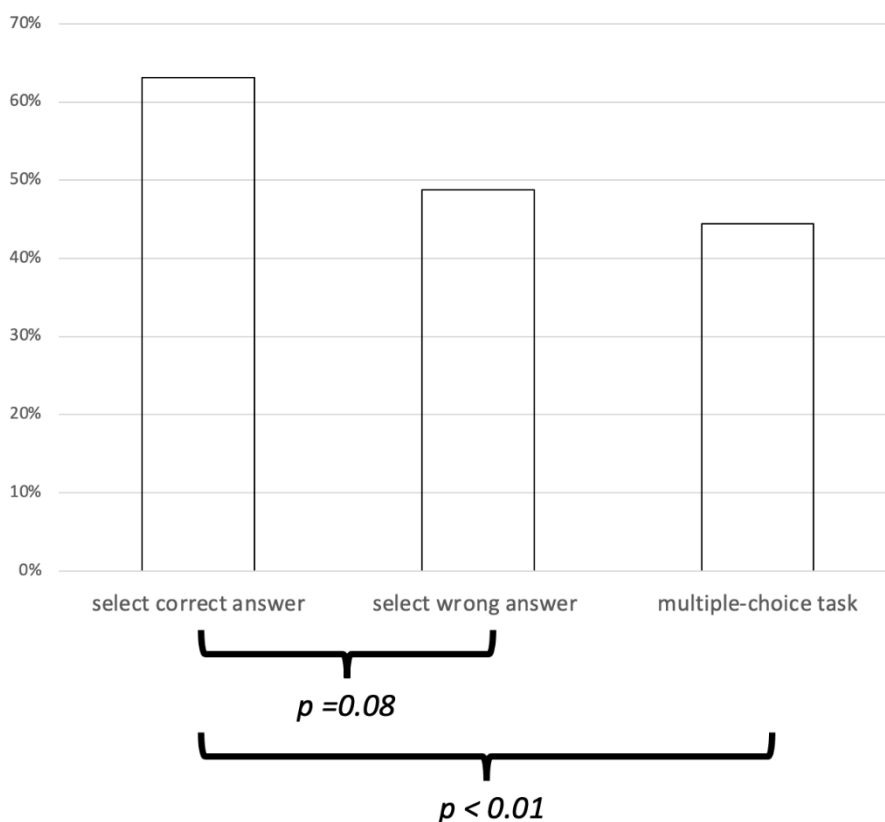


Fig. 4. Bar graph presenting total score of ChatGPT on Polish version of Spring and Autumn 2022 exams depending on logical construction of questions. P value for chi-square test between proportion of correct answers in different subgroups.
Ryc. 4. Wykres słupkowy przedstawiający łączny wynik ChatGPT dla polskiej wersji egzaminów wiosna i jesień 2022 z podziałem według konstrukcji logicznej polecenia. Przedstawiono wartość p dla testu chi-kwadrat, porównując proporcje poprawnych odpowiedzi w podgrupach.

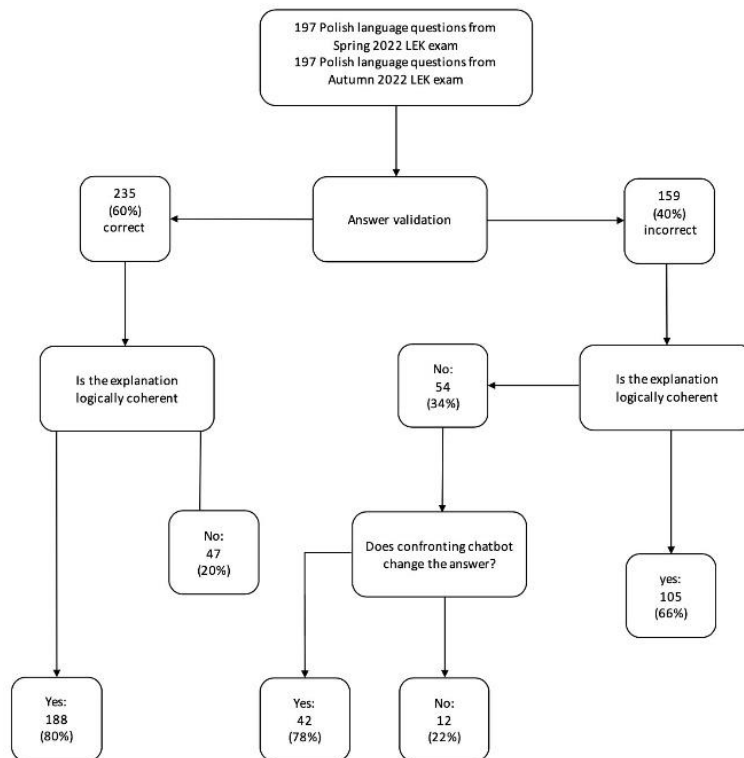


Fig. 5. Algorithm evaluating capability of ChatGPT to provide logically coherent explanation and ability to correct itself after user interaction.
Ryc. 5. Algorytm oceniający zdolność ChatGPT do przedstawienia logicznie spójnego wyjaśnienia wskazanej odpowiedzi oraz dokonania samokorekty po interakcji z użytkownikiem.

DISCUSSION

ChatGPT-3.5 demonstrated proficiency in passing the LEK examination during both the Spring 2022 and Autumn 2022 sessions. Our findings, in line with the conclusions of other researchers, demonstrate that Chat exhibits the capability to achieve a result ranging from 40% to 60% in licensing examinations required for medical practitioners [3,8,9,10]. In contrast to another study pertaining to the Polish LEK [9], our research confirmed ChatGPT’s successful performance in the Spring 2022 session of the examination. This divergence may be attributed to the methodology employed in data acquisition. While the cited study relied on automated data input via an API, our study involved the manual entry of questions, affording us the opportunity to make occasional refinements to the non-substantive, yet technically structured aspects of the inquiries. Moreover, the linguistic factor (Polish vs English) in the formulation of the questions for the Autumn 2022 LEK examination, despite the observed variations in the obtained outcomes, exhibited no statistical significance. While the language employed in the question formulation may be considered an inconsequential variable, it is worth noting that the linguistic aspect, along with the user’s country of origin and legal framework, can carry substantive implica-

tions. The training of the chatbot is predominantly reliant on English language data and English related facts and principles [11], which holds significance when considering its responses in the domain of medical law. Notably in our paper, ChatGPT presented the least proficiency in the medical certification part of the exam, which is based on local solutions and legal acts much different than their foreign counterparts. This observation aligns with the findings of Japanese researchers [12], who reported a similar issue. Specifically, when confronted with a question regarding palliative care for terminally ill patients, the chatbot erroneously identified euthanasia as the appropriate response, despite its illegality in Japan. This misclassification constitutes a cardinal error in its performance assessment. In contrast, ChatGPT exhibits exemplary performance when addressing inquiries pertaining to public health, notably, the difficulty index shows no substantial variance between the aforementioned question categories, thus accentuating the crucial role played by the dataset’s predominant focus on English language content. In summary, our findings indicate that language does not play a significant role compared to the sourcing of answers for the given questions, as established by other researchers as well. Furthermore, the length of the question proved inconsequential in relation to ChatGPT’s performance. The chatbot adeptly handles queries both extensive and concise in



nature. In this context, it is important to emphasize the remarkable ability of GPT-3.5 to effectively analyze tasks of a considerable length, such as half of the presented abstract (1000 characters). This observation is supported by researchers who have analyzed the chatbot's performance on ophthalmic exams. While longer tasks may pose potential challenges for GPT-3.5, we can reasonably expect optimal performance for standard-length clinical questions [13]. Nonetheless, the logical framework and format of the question (single-choice select correct vs. select wrong vs. multiple-choice) do bear significance. Although the chi-square test established that logical construction has an impact on the proportion of correct answers, it did not show a statistically significant difference in the mentioned proportion between the select the correct and select the wrong answer questions. Multiple-choice questions pose a comparatively a greater challenge for the chatbot. Moreover in the case of multiple-choice, ChatGPT had more difficulty in presenting a logically coherent explanation. Taking this into consideration, a user should be aware of the limitations of chatbots but also of factors that do not affect its performance for a superior quality of answers.

Performing the commands required during the LEK exam and similar exams, such as assigning patients to specific groups or selecting the most appropriate treatment approach based on the presented diagnostic test results and the patient's clinical characteristics, is a complex process that requires the integration of multiple stages of machine learning techniques and natural language processing on which the GPT-3.5 model is based. During information processing by GPT-3.5, the entered text undergoes tokenization as an initial step, which involves dividing the input text into smaller units known as tokens, which can be individual words, characters, or other units that enable better text processing and understanding. The response generated by GPT-3.5 relies heavily on the output of tokenized text from individual parameters and layers of the neural network, as well as the analysis of the context of the input text. Despite achieving sufficient answer percentages to pass the LEK exam and maintaining coherent logical explanations, GPT-3.5 does not always provide accurate answers or logically consistent explanations. Certainly, Chat occasionally provides a well-structured explanation, and such an explanation enhances the likelihood of a correct answer ($p < 0.01$). Does this imply that ChatGPT's understanding of the posed question is a guarantee of a good response? It is plausible, however, such deliberations tend to anthropomorphize Chat, and yet we are aware that the operation of large language models differs significantly from human cognition. The question remains whether the generation of incorrect answers and inconsistent explanations is due to the disruption of an appropriate integration of data from individual variables along with

the knowledge embedded in the neural network's materials, the outdated or inconsistent nature of GPT-3.5's own knowledge with current clinical knowledge, or the lack of real-time interaction ability, as suggested by the frequent achievement of correct answers when querying or emphasizing information that is the most relevant to selecting the correct answer. Based on our observations, the length of the text does not affect the correctness of the answer provided by GPT-3.5 ($p = 0.46$). Nonetheless, it should be noted that the final decision may also be influenced by the recognition of associations between analyzed variables, which may not be directly related but can have a mutual influence during the process of making appropriate clinical decisions, patient assignment to appropriate groups, and similar tasks that require the proper selection and integration of all the entered data. In addition to the mentioned difficulty of appropriately integrating data in the pipeline of input text-neural network-decision-making, the possible errors made by ChatGPT can also be influenced by the lack of the real-time interaction ability of GPT-3.5. This limitation in scenarios involving complex parameters or the absence of the necessary information for precise determination of the clinical situation, may lead to simplifications and wrong decisions. Owing to the possibility of introducing bias, it remains an extremely intricate task to determine whether the deciding factor for a negative response relies on declarative knowledge associated with straightforward facts readily available on the Internet or the procedural "skills" of the chatbot in the realm of diagnosis and clinical analysis. Nonetheless, based on our observation, the number of factors analyzed by Chat appears to exert a pivotal influence on the accuracy of both the response and the justification. Consequently, questions with a simpler structure and a proportional demand for more declarative knowledge seem to have a higher likelihood of eliciting a proficient response. Based on these limitations and the presented results in this manuscript, it can be concluded that GPT-3.5, despite being a useful tool for generating information, should not be recommended as a tool for making clinical decisions that require the analysis and integration of multiple clinical variables encompassing the characteristics of a specific patient.

In addition to assessing the efficacy of ChatGPT in relation to the LEK exam, we aimed to evaluate the applicability of the chatbot for educational and pedagogical purposes. To this end, we examined the logical coherence of the responses, effectively simulating a potential user's lack of knowledge. There are two reasons why we opted for this solution, although there are other models that include insight into the answers, accuracy or concordance [3]. We did not focus on those features as it is difficult to provide an objective classification for those factors. Instead, we



opted for logical coherence as the investigated factor to eliminate the subjective verdict of the researcher. Moreover, the opinion of the initial researcher was evaluated by another researcher to provide superior assessment. The second intention was to emulate a user with little to no medical knowledge. It is widely known that reaching for medical advice from unsupervised internet sources can lead to fatal results, especially when users rely on so-called “common sense” [14]. Thus, by eliminating the medical knowledge factor and only testing logical coherence, we created a model allowing us to assess the chatbot’s potential to give not only misleading but also seemingly credible information.

The chatbot exhibited the capacity to elucidate its reasoning in a logical and cohesive manner for the majority of the questions, providing responses that aligned with the designated answer key. However, a cause for concern arises from the observation that the same logical and coherent style of responses was consistently exhibited, even in cases where the provided answers were incorrect. This phenomenon has also been observed by other researchers [15,16]. This raises alarm, especially in light of the growing trend of online health information-seeking in Poland [17]. Furthermore, the ChatGPT-3.5 platform remains freely available and easily accessible, potentially encouraging users to engage with Chat due to its convenient accessibility. This aspect was also a contributing factor in selecting this particular version of Chat for our study. Nonetheless, the questions backed by a proper explanation were significantly more likely to be correct. It suggests that a user should aim to ask the chatbot for an explanation to ensure a higher chance of receiving a correct answer.

Given the potentially misleading responses to health-related inquiries and the occurrence of “hallucinations” within Chat (providing false information and non-existent sources while maintaining an appearance of logical and coherent discourse) [11,18], concerns regarding user safety are raised. Zuccon and Koopman [19] conducted a study in which ChatGPT-3.5 was exposed to a series of inquiries sourced from the TREC Health Misinformation Track 2021 and 2022. This collaborative initiative was specifically designed to furnish researchers with authoritative insights into the realm of health-related misinformation, thereby facilitating the acquisition of validated information in the domain of healthcare advice. The results of this study highlight the fact that while ChatGPT-3.5 demonstrates an 80% accuracy in determining the veracity of statements from TREC based on its training alone, the introduction of additional information suggesting false medical facts leads to a decrease in accuracy to 63%. Our team also observed a high level of suggestibility in ChatGPT-3.5 during the interactions, where the technique and manner of

questioning greatly influence the nature of the received response. This underscores the significance of acquiring the skill to formulate questions in an appropriate manner, thereby minimizing the risk of receiving false and potentially hazardous answers. Especially considering the fact that Chat does not always provide a disclaimer at the end of its responses, our observation may be attributed to the “test-like” nature of the posed medical-related queries. Nevertheless, in light of the potential risks, it is crucial that such disclaimers become a consistent feature.

We would like to highlight several significant limitations to our study. It is important to acknowledge that we did not establish statistically significant differences in the logical construction of the questions (identify the correct vs. select the false), likely attributable to the insufficient amount of data available. Furthermore, regarding the difficulty index mentioned earlier in the study, although it yields the expected result where Chat responds to questions according to the anticipated difficulty level, it is important to note that the current version of LEK consists of 70% questions from an explicit question bank, which significantly influenced the difficulty of the questions in light of the difficulty index, as it is a measure of student responses. To address this bias, it would be best to analyze more exams from previous years when the explicit question bank was not yet available. Additionally, there remains the issue of evaluating the logic and coherence of the explanations given by ChatGPT, although double checking by two researchers provides a higher objective value, there still could be a bias as both researchers are medical personnel.

CONCLUSIONS

1. The chatbot passed the LEK exam in all 3 instances, proving some capabilities to answer both clinical and medical-knowledge associated questions. However, a substantial number of wrong answers makes it a problematic recommendation to use as a reliable source of medical knowledge.
2. Users should be aware of the chatbot’s limitations and factors contributing to more accurate answers. By taking into consideration factors such as logical construction and the subject of a question, users can obtain more precise answers. Moreover, confronting the chatbot increases the chances of receiving correct information. On the other hand, the length of the task and the language surprisingly do not significantly impact the performance.
3. A high percentage of seemingly compelling explanations for false information can be potentially hazardous for non-medical users, posing a threat to public health.



Author's contribution

Study design – K. Żmudka

Data collection – K. Żmudka, A. Spychał, B. Ochman, Ł. Popowicz

Data interpretation – K. Żmudka, A. Spychał, B. Ochman, Ł. Popowicz

Statistical analysis – K. Żmudka, B. Ochman

Manuscript preparation – K. Żmudka, A. Spychał, B. Ochman, Ł. Popowicz, P. Piłat, J. Jaroszewicz

Literature research – A. Spychał, B. Ochman, Ł. Popowicz, P. Piłat

REFERENCES

1. Nori H., King N., McKinney S.M., Carignan D., Horvitz E. Capabilities of GPT-4 on medical challenge problems [Internet]. arXiv; 2023 [cited 2023 Jul 30]. Available from: <http://arxiv.org/abs/2303.13375>.
2. Newton P.M., Xiromeriti M. ChatGPT performance on MCQ exams in higher education: A pragmatic scoping review [Internet]. EdArXiv; 2023 Feb [cited 2023 Jul 6]. Available from: <https://osf.io/sytu3>.
3. Kung T.H., Cheatham M., Medenilla A., Sillos C., De Leon L., Elepaño C. et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* 2023; 2(2): e0000198, doi: 10.1371/journal.pdig.0000198.
4. Sabry Abdel-Messih M., Kamel Boulos M.N. ChatGPT in clinical toxicology. *JMIR Med. Educ.* 2023; 9: e46876, doi: 10.2196/46876.
5. Ali R., Tang O.Y., Connolly I.D., Zadnik Sullivan P.L., Shin J.H., Fridley J.S. et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery* 2023; 93(6): 1353–1365, doi: 10.1227/neu.0000000000002632.
6. opis_statystyk.pdf [Internet]. Centrum Egzaminów Medycznych; [cited 2023 Nov 17]. Available from: https://www.cem.edu.pl/aktualnosci/opis_statystyk.pdf.
7. Kelley T.L. The selection of upper and lower groups for the validation of test items. *J. Educ. Psych.* 1939; 30(1): 17–24.
8. Takagi S., Watari T., Erabi A., Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. *JMIR Med. Educ.* 2023; 9: e48002, doi: 10.2196/48002.
9. Rosol M., Gąsior J.S., Łaba J., Korzeniewski K., Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Medical Final Examination [Internet]. medRxiv; 2023 [cited 2023 Jul 1]. doi: 10.1101/2023.06.04.23290939. Available from: <https://www.medrxiv.org/content/10.1101/2023.06.04.23290939v1>.
10. Gilson A., Safranek C., Huang T., Socrates V., Chi L., Taylor R.A. et al. How does ChatGPT perform on the Medical Licensing Exams? The implications of large language models for medical education and knowledge assessment [Internet]. medRxiv; 2022 [cited 2023 Jul 3]. doi: 10.1101/2022.12.23.22283901. Available from: <https://www.medrxiv.org/content/10.1101/2022.12.23.22283901v1>.
11. OpenAI. GPT-4 technical report [Internet]. arXiv; 2023 [cited 2023 Jul 1]. Available from: <http://arxiv.org/abs/2303.08774>.
12. Kasai J., Kasai Y., Sakaguchi K., Yamada Y., Radev D. Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations [Internet]. arXiv; 2023 [cited 2023 Jul 1]. Available from: <http://arxiv.org/abs/2303.18027>.
13. Mihalache A., Popovic M.M., Muni R.H. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol.* 2023; 141(6): 589–597, doi: 10.1001/jamaophthalmol.2023.1144.
14. Powell J., Inglis N., Ronnie J., Large S. The characteristics and motivations of online health information seekers: cross-sectional survey and qualitative interview study. *J. Med. Internet Res.* 2011; 13(1): e20, doi: 10.2196/jmir.1600.
15. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* 2023; 11(6): 887, doi: 10.3390/healthcare11060887.
16. Hopkins A.M., Logan J.M., Kichenadasse G., Sorich M.J. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectr.* 2023; 7(2): pkad010, doi: 10.1093/jncics/pkad010.
17. Bujnowska-Fedak M.M. Trends in the use of the Internet for health purposes in Poland. *BMC Public Health* 2015; 15: 194, doi: 10.1186/s12889-015-1473-3.
18. Borji A. A categorical archive of ChatGPT failures [Internet]. arXiv; 2023 [cited 2023 Jul 3]. Available from: <http://arxiv.org/abs/2302.03494>.
19. Zuecon G., Koopman B. Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness [Internet]. arXiv; 2023 [cited 2023 Jul 2]. Available from: <http://arxiv.org/abs/2302.13793>.