Alergologia Polska – Polish Journal of Allergology 2024; 11, 1: 42–47
DOI: https://doi.org/10.5114/pja.2024.135380
Received: 8.12.2023, accepted: 30.12.2023, online publication: 13.02.2024.

Alergologia POLSKA
Polish Journal of Allergology

ORIGINAL PAPER

Evaluating ChatGPT-3.5 in allergology: performance in the Polish Specialist Examination

Michał Bielówka[1], Jakub Kufel[2,3], Marcin Rojek[1,4], Adam Mitręga[1], Dominika Kaczyńska[1], Łukasz Czogalik[1], Michał Janik[1], Wiktoria Bartnikowska[5], Sylwia Mielcarska[6], Dominika Kondoł[7]

[1]Students' Scientific Association of Computer Analysis and Artificial Intelligence at the Department of Radiology and Nuclear Medicine, Medical University of Silesia, Katowice, Poland
[2]Department of Radiodiagnostics, Interventional Radiology and Nuclear Medicine, Medical University of Silesia, Katowice, Poland
[3]Department of Radiology and Nuclear Medicine, Medical University of Silesia, Katowice, Poland
[4]Students' Scientific Association at the Department of Microbiology and Immunology, Medical University of Silesia, Katowice, Poland
[5]Faculty of Medical Sciences in Katowice, Medical University of Silesia, Katowice, Poland
[6]Department of Medical and Molecular Biology, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Zabrze, Poland
[7]Dr B. Hager Memorial Multi-specialty District Hospital, Tarnowskie Góry, Poland

ABSTRACT

Introduction: The development of Artificial Intelligence (AI) and attempts to use it in medicine are increasingly becoming the subject of more scientific research.

Aim: The aim of this article is to present the effectiveness of the advanced language model, ChatGPT-3.5 in the context of the pass rate of the Polish National Specialist Examination (PES) in allergology. Additionally, it seeks to comprehend the potential applications of artificial intelligence in the field of medicine, particularly within allergology.

Material and methods: The study used the latest available PES exam prepared by the Medical Research Centre in Lodz. 118 questions were asked using the openai.com platform, which allows free access to the ChatGPT-3.5 model. All questions were classified according to Bloom's taxonomy to assess their complexity and difficulty, with additional three categorisations. Each question was asked five times.

Results: ChatGPT-3.5 did not pass the allergology PES, achieving a score of 52.54%. It was observed that the model performed better in answering memory questions (60%) compared to those requiring comprehension and critical thinking, where the results were slightly lower (45%). Moreover, within the categories of 'treatment', 'immune system' and 'symptoms', the model exceeded the passing threshold. Questions to which ChatGPT provided the correct answer significantly exhibited higher difficulty compared to those to which it provided an incorrect response.

Conclusions: The results indicate that ChatGPT's pass rate in the allergology PES is considerably lower than that of resident doctors specializing in this field. The potential applications of AI in medicine require further research to effectively support clinical practice among physicians.

KEY WORDS

artificial intelligence, allergology, language model, ChatGPT-3.

ADDRESS FOR CORRESPONDENCE

Michał Bielówka, Students' Scientific Association of Computer Analysis and Artificial Intelligence at the Department of Radiology and Nuclear Medicine, Medical University of Silesia, Katowice, Poland, e-mail: michalbielowka01@gmail.com

INTRODUCTION

An allergology specialist possesses extensive theoretical knowledge about the immunological and molecular bases of allergic diseases, as well as the morphology and physiology of the respiratory, digestive, and skin systems. They also have a range of diagnostic and therapeutic skills, including targeted examination of allergic symptoms in patients, conducting and interpreting skin tests, provocation tests, patient qualification for specific immunotherapy, and biological treatments [1]. In Poland, the specialization program in allergology lasts for 5 years, divided into a basic module comprising 2 years of training in internal medicine and a three-year specialized module. The culmination of this education process is the National Specialist Examination (PES), which consists of both a written and oral exam [2]. The written part of the PES comprises 120 single/multiple-choice questions, each question having 5 answer options. A passing score is achieved by candidates who obtain at least 60% of the test points [3].

The number of allergist specialists practicing in Poland is 1485 (data as of 30 September 2023) [4], which is relatively low, evidenced by a shortage of allergology clinics in approximately fifty percent of the counties in Poland, resulting in an exceptionally high patient-to-allergist ratio [5]. These statistics have concerned the authors of this publication, prompting them to investigate the capabilities of artificial intelligence (AI) in providing correct answers to the test questions included in the PES. During the analysis, the scope of questions was categorized thematically and according to specific competencies such as 'knowledge' and 'drawing conclusions'.

AIM

The aim of this analysis is to compare the results obtained by ChatGPT with human cognitive abilities and to contemplate the utilization of AI in the daily work of allergist doctors. Given the current low number of specialists in this field and the increasing incidence of conditions like asthma, the authors see potential in the application of AI by practicing doctors, foreseeing a reduction in waiting times for appointments. Furthermore, employing AI techniques in clinical studies could expedite medical advancements in the field of allergology.

MATERIAL AND METHODS

EXAMINATION AND QUESTIONS

The conducted study aimed to assess the ability of an artificial intelligence model to provide correct answers in the specialist examination in allergology. A set of 120 questions from the PES from the spring of 2023 was utilized, selecting the latest publicly available set. Two questions were excluded, one due to graphical content and the other deemed inconsistent with current medical knowledge, leaving a pool of 118 questions [6]. The qualified questions underwent classification according to Bloom's taxonomy and three parallel proprietary divisions.

The first proprietary division involved categorizing the scope of information referenced in the questions. This method led to the creation of categories such as 'clinical procedures', 'clinical guidelines', 'diagnostics', 'immunotherapy', 'genetics', 'immune system', 'treatment', 'symptoms', and 'disease-related'.

The subsequent division concerned the nature of the questions: memory-based and the ones requiring comprehension and critical thinking. The final division aimed to differentiate between 'clinical' questions and all others.

DATA COLLECTION AND ANALYSIS

The study was conducted using the GPT-3.5 language model as of 1 June 2023. Each question was posed multiple times in independent instances to determine the model's level of 'conviction regarding the correctness of the answer. The necessity to initiate a new chat multiple times stemmed from the risk of receiving the same question again, potentially suggesting the truthfulness of the previous response. Five sessions were carried out. In each session, a complete set (n = 118) of unique questions was presented, preceded by a prompt. The prompt aimed to streamline the collection of answers to questions by limiting them to a single letter and presenting the general concept of a single-test query.

STATISTICAL ANALYSIS

Analyses were conducted using the R Studio environment (an open-source integrated development environment for the R language) version 1.1.46. A response was considered correct if it was provided in at least three out of five

TABLE 1. Division by type

Type	Clinical/Other	Did ChatGPT respond correctly?	Number of questions	%
Comprehension and critical thinking questions	Clinical	No	27	54
		Yes	23	46
	Other	No	6	60
		Yes	4	40
Memory questions	Clinical	No	17	41.46
		Yes	24	58.54
	Other	No	6	35.3
		Yes	11	64.7

initiated instances of the GPT language model. Statistical significance was set at $p < 0.05$. The analysis of questions considered the model's confidence coefficient in its response (expressed as the ratio of the number of dominant responses in consecutive sessions to the total number of sessions ($n = 5$)), difficulty statistics from conducted exams (courtesy of CME – Medical Examinations Centre in Lodz), and whether the question belonged to any of the designated categories.

To assess quantitative variables in the context of response accuracy, the Mann-Whitney *U* test with a continuity correction was utilized. This method evaluated the relationship between response accuracy, difficulty obtained from CME, and the proprietary confidence coefficient. The Spearman's rank-order correlation test was employed to assess the relationship between the difficulty index of questions obtained from the Medical Examinations Centre and the confidence coefficient of the chatbot. The Pearson χ^2 test was used to evaluate the relationship between response accuracy and question category. To assess quantitative variables (comprising difficulty index and confidence coefficients of questions) concerning response accuracy, the Mann-Whitney *U* test with a continuity correction was utilized. The following formula represents the aforementioned relationship: .

$$p_{GPT} = \frac{\square_i max \Sigma_{j=1}^{n} \square \delta (x_i - x_j)}{n = 5}$$

TABLE 2. Division by topic

Topic	Correct answer			
	Yes	%	No	%
Clinical guidelines	8	50.00	8	50.00
Immunotherapy	6	54.55	5	45.45
Diagnostics	6	42.86	8	57.14
Clinical procedures	8	53.33	7	46.67
Treatment	6	60.00	4	40.00
Immune system	14	70.00	6	30.00
Genetics	1	33.33	2	66.67
Symptoms and signs	8	80.00	2	20.00
Disease-related	5	26.32	14	73.68

RESULTS

ChatGPT scored 52.54% (62/118 points) in the exam (Table 1).

During the statistical analysis, questions were divided into different categories.

When dividing questions into 'memory-based' and 'comprehension and critical thinking questions', ChatGPT scored 60% (35/58 points) and 45% (27/60 points), respectively (Table 1).

In the category divided into 'clinical' and 'other', ChatGPT scored 51.65% (47/91 points) and 55.56% (15/27 points), respectively (Table 1).

The questions were further categorized based on subjects, and the outcomes were found to be ranging from 26.32% to 80% (Table 2).

Using the Mann-Whitney *U* test, difficulty indices of questions and confidence coefficients of responses given by ChatGPT were compared. The results showed that questions ChatGPT answered correctly had significantly higher difficulty indices compared to those answered incorrectly. The confidence coefficient was higher for questions ChatGPT answered correctly (Figure 1). Furthermore, the difficulty index positively correlated with the confidence coefficient. However, the confidence coefficient did not differ between the question categories 'clinical' and 'other', nor between the categories 'memory questions' and 'comprehension and critical thinking questions'.

DISCUSSION

The Specialist Examination in Allergology represents a pivotal milestone for medical professionals aiming to attain specialization in this intricate field of medicine. This

comprehensive examination encompasses both practical and theoretical components, ensuring that candidates possess the requisite skills and knowledge to excel in allergology.

In Poland, achieving a minimum score of 60% in the Specialist Examination in Allergology is the benchmark for success and is crucial for obtaining specialization in this field. Additionally, successful completion of an oral examination, fulfilment of specified procedures, and the completion of a required period of practical training are also necessary components. These elements collectively contribute to acquiring the qualifications and skills essential for practicing as an allergology specialist in Poland.

The results of the examination of ChatGPT's performance provide valuable insights into its ability to answer questions across different categories and topics.

ChatGPT achieved an overall score of 52.54%, answering 62 out of 118 questions correctly. It obtained a comparable result to that obtained in the study conducted by Kufel *et al.* involving the same language model solving a Specialist Examination in Nuclear Medicine (56.41% of correct answers) [7]. While this score might be considered modest, it is important to remember that natural language understanding and generation is a complex task and any improvement in this area is noteworthy. Conversely, in Weng *et al.*'s study 'ChatGPT failed Taiwan's Family Medicine Board Exam', ChatGPT's accuracy in Family Medicine Board Exam rate was 41.6%, with 52 correct responses out of 125 questions [8]. Our study demonstrated a slightly better performance, possibly indicating ChatGPT's relative strength in that specific medical domain. It is significant to note that the Taiwan study encompassed diverse question types, including negative-phrase questions, multiple-choice questions, mutually exclusive options, case scenarios, and Taiwan's local policy-related questions. In contrast, the allergology study focused solely on medical and allergology-related questions. This variation in question types and subjects may account for the differences in accuracy rates between the two studies [8].

The aim of the study by Fuchs *et al.* was to assess how ChatGPT 3 and ChatGPT 4 perform when answering self-assessment questions related to dentistry in the Swiss Federal Licensing Examination in Dental Medicine (SFLEDM). In addition, the study examined allergy and clinical immunology in the European Examination in Allergy and Clinical Immunology (EEAACI) with priming and without priming [9]. ChatGPT 3 showed an average of 69 ±3.7% of correct responses in the EEAACI test without priming (as in our study), which is its best performance in an allergy exam in the literature. This is probably due to the small sample size of the questions asked, as

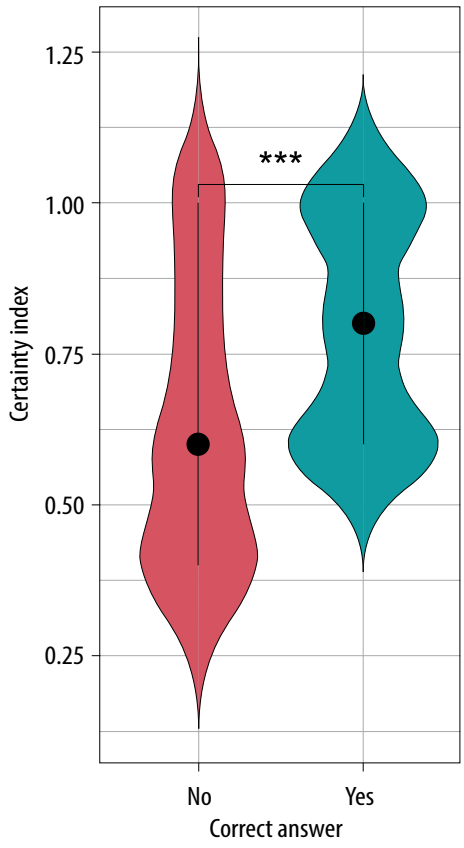

FIGURE 1. Comparison of correct answers with confidence coefficient

in the above-mentioned study, only 28 were asked during one EEAACI exam. Furthermore, our study showed that the ChatGPT performed differently on different categories of questions. For example, in the 'immune system' questions, it obtained an average of 70% of correct answers, while in the 'disease-related' category it obtained only 26.32%. However, the Fuchs *et al.* study did not include a breakdown of the questions into thematic categories or a specific list of questions asked of this language model. It did, however, show that a broader description of the context of the problem in question beforehand slightly helps the ChatGPT to answer the question correctly (3.9% improvement) [9].

As we proceed to examine the categorization of questions into 'memory' and 'comprehension and critical thinking', intriguing insights emerge. ChatGPT demonstrates a superior performance on 'memory' questions, boasting a commendable success rate of 60%, whereas on questions requiring critical thinking, it achieves a slightly lower score of 45%. It is worth noting that in our research about the Polish specialty exam in Radiology, ChatGPT demonstrated superior performance on questions that demanded critical thinking, scoring 55%, as opposed to questions that primarily relied on factual knowledge, where it scored 44% [10].

This observation implies that the model excels at factual recall and information retrieval, laying a solid foundation. However, it also underscores the room for improvement in tasks demanding more intricate abstract reasoning and critical thinking.

Upon further investigation into the differentiation of questions into 'clinical' and 'other' categories, we find a notable consistency in ChatGPT's performance. With a 51.65% accuracy rate in 'clinical' questions and a slightly higher 55.56% in 'other' questions, the model's ability to furnish precise responses demonstrates uniformity across these broad categories. In addition, in our study about the Polish specialty exam in Radiology, the ChatGPT achieved identical results of correct answers in both question categories: clinical (54.55%) and physical (54.55%). This uniformity holds significant promise for the model's versatility across diverse contexts, providing a valuable aspect of its adaptability.

A closer examination of performance based on specific topics reveals a spectrum of accuracy. Some areas, such as 'immune system' and 'symptoms', exhibit relatively high success rates, while others, including 'genetics' and 'related to diseases', present formidable challenges. These variations may be attributed to the rich diversity and intricacy of medical knowledge, serving as a formidable test for AI models like ChatGPT.

The results stemming from the Mann-Whitney *U* test unveil an engaging correlation between question difficulty and the model's confidence. It is particularly striking that ChatGPT exhibits increased confidence levels when addressing more challenging questions. This adaptability in confidence levels holds the potential to be a pivotal feature in enhancing the model's practical utility.

However, it is of paramount importance to maintain a balance between confidence and accuracy, thereby preventing the model from becoming overly confident, especially in demanding scenarios.

It is also worth mentioning that the use of artificial intelligence such as ChatGPT may have benefits as well as risks. These may relate to the accuracy of the language model's response itself, as well as ethical issues. This problem also applies to its possible other applications, e.g. in academic writing [11]. As yet, one has to be highly cautious in its potential usage.

CONCLUSIONS

Based on the results provided, ChatGPT could not pass the PES in allergology. The score of 52.54% did not meet the minimum score threshold of 60%. Nevertheless, the ChatGPT answered questions correctly with a significantly higher difficulty index than the questions that the ChatGPT answered incorrectly.

In addition, the ChatGPT scored more satisfactorily on the 'memory' question category (60.34%), relative to the 'comprehension and critical thinking' questions (45.00%). A relatively high score was also achieved in the category of questions on treatment (60.00%), signs and symptoms (80.00%), and queries about the immune system (70.00%).

For 9 years (2009–2018), 426 people took the exam, with 400 successful passers (93.9%) [12]. The study proves that humans perform better at solving the test than the proposed language model based on artificial intelligence. However, further research is needed using the official questions provided by the CEM, and testing the ChatGPT on the pass rates of state examinations in allergy. This will provide a more comprehensive understanding of the model and the characteristics of its performance in the topic above. It should also be borne in mind that the technology is improving all the time, ChatGPT is still learning with LLM and its ability to solve the test should improve. Undoubtedly, the development of artificial intelligence has the potential to positively impact the work of allergologists, but this requires further work on the technology.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Małolepszy J. History of Polish allergology over the last 58 years. Pol J Allergol 2022; 9: 217-25.
2. Centrum Medyczne Kształcenia Podyplomowego. Accessed: Nov. 21, 2023. [Online]. Available: https://www.cmkp.edu.pl/ksztalcenie/podyplomowe/lekarze-i-lekarze-dentysci/modulowe-programy-specjalizacji-lekarskich-2023.
3. Centrum Egzaminów Medycznych. Accessed: Nov. 21, 2023. [Online]. Available: https://www.cem.edu.pl/spec.php.
4. NIL - Informacje statystyczne. Accessed: Nov. 21, 2023. [Online]. Available: https://nil.org.pl/rejestry/centralny-rejestr-lekarzy/informacje-statystyczne.
5. Mapa potrzeb na lata 2022-2026 – Mapy potrzeb zdrowotnych – Ministerstwo Zdrowia. Accessed: Nov. 21, 2023. [Online]. Available: https://basiw.mz.gov.pl/mapy-informacje/mapa-2022-2026/.
6. Centrum Egzaminów Medycznych. Accessed: Nov. 21, 2023. [Online]. Available: https://cem.edu.pl/index.php.
7. Kufel J, Bielówka M, Rojek M, et al. Assessing ChatGPT's performance in national nuclear medicine specialty examination: an evaluative analysis. Iran J Nuclear Med 2023, doi: 10.22034/irjnm.2023.129434.1580.
8. Weng TL, Wang YM, Chang S, et al. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc 2023; 86: 762-6.
9. Fuchs A, Trachsel T, Weiger R, Eggmann F. ChatGPT's performance in dentistry and allergy-immunology assessments: a comparative study. Swiss Dent J 2023; 134 (5).
10. Kufel J, Paszkiewicz I, Bielówka M, et al. Will ChatGPT pass the Polish specialty exam in radiology and diagnostic imaging? Insights into strengths and limitations. Pol J Radiol 2023; 88: 430-4.

11. Dergaa I, Chamari K, Zmijewski P, Ben Saad H. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. Biol Sport 2023; 40: 615-22.
12. Centrum Egzaminów Medycznych. Accessed: Sep. 01, 2023. [Online]. Available: https://www.cem.edu.pl/aktualnosci/spece/spece_stat.php.