Original paper

An investigative analysis – ChatGPT's capability to excel in the Polish speciality exam in pathology

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

[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 Radiology and Nuclear Medicine, Medical University of Silesia, Katowice, Poland
[3]Faculty of Medical Sciences in Katowice, Medical University of Silesia, Katowice, Poland
[4]Dr B. Hager Memorial Multi-Specialty District Hospital, Tarnowskie Góry, Poland
[5]Department of Medical and Molecular Biology, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Zabrze, Poland

This study evaluates the effectiveness of the ChatGPT-3.5 language model in providing correct answers to pathomorphology questions as required by the State Speciality Examination (PES). Artificial intelligence (AI) in medicine is generating increasing interest, but its potential needs thorough evaluation. A set of 119 exam questions by type and subtype were used, which was posed to the ChatGPT-3.5 model. Performance was analysed with regard to the success rate in different question categories and subtypes.

ChatGPT-3.5 achieved a performance of 45.38%, which is significantly below the minimum PES pass threshold. The results achieved varied by question type and subtype, with better results in questions requiring "comprehension and critical thinking" than "memory".

The analysis shows that, although ChatGPT-3.5 can be a useful teaching tool. However, its performance in providing correct answers to pathomorphology questions is significantly lower than that of human respondents. This conclusion highlights the need to further improve the AI model, taking into account the specificities of the medical field. Artificial intelligence can be helpful, but it cannot fully replace the experience and knowledge of specialists.

Key words: pathomorphology, artificial intelligence, language model, ChatGPT-3.5, specialty examination.

Introduction

ChatGPT is a product from Open AI that uses artificial intelligence (AI) resources to answer questions posed to it in the form of a chat [1]. When asked "What are you?" it responds "I am a language model (LLM) created by Open AI, based on the GPT-3.5 architecture, and my name is ChatGPT. My main task is to process and generate human-like text in response to user questions and commands. How can I help you today?". The tool is a LLM based on machine learning, deep learning, and artificial neural networks [2]. The developers of the application continue to enhance it, update it, and incorporate novel

features, resulting in its continuous growth [3]. The widespread availability of the tool has led many organisations, researchers, and politicians to fear its misuse for unethical and criminal purposes. In response to this concern, Open AI emphasises that it is constantly working to prevent abuse [4].

Among many scientific disciplines, medical practitioners have also taken an interest in the use of AI in their work, seeing in the tool an opportunity to reduce medical and therapeutic errors, and reduce diagnostic times which would undoubtedly be associated with an increase in the quality of life and survival of patients. The ethical issues of using AI in everyday medical practice remain controversial, and it should be remembered that it has certain limitations – it may have outdated information, give false positive or false negative results, or be exploited by hackers [5, 6].

The authors of this publication set out to test the ability of ChatGPT generation 3.5 to provide correct answers to questions on the State Speciality Examination (PES) in Pathomorphology, which is a Polish exam that verifies the readiness of pathomorphology trainees to become specialists in this field of medicine. The exam consists of a test and an oral part. This analysis uses single-choice test questions from the test part of PES to assess knowledge and the ability to draw logical conclusions. A passing score is considered to be a correct answer to at least 60% of the 120 questions [7, 8]. This study will provide an analysis of the AI skills used to answer PES questions.

Material and methods

Examination and questions

A publicly available set of exam questions, posted on the official website of the Centre for Medical Examinations in Łódź [9], used for the State Speciality Examination in Pathomorphology in spring 2023, was used for the prospective study – the selection criterion was the most recent exam available. It consists of a pool of 120 single-choice questions. They were divided using 2 proprietary divisions into types and subtypes. The first divided them into "memory" questions and "comprehension and critical thinking" questions. The second, on the other hand, referred to the subject of the question and included categories such as: "genetics", "pathophysiology", "microscopic findings", "signs and symptoms", "medical procedures", "aetiology of diseases", "epidemiology", and "medical guidelines". Finally, 119 questions were included in the study, after question number 68 was discarded due to its possible incompatibility with the latest medical knowledge. Qualifying questions were subjected to Bloom"s classification.

Data collection and analysis

The examination involved the ChatGPT-3.5 LLM as of 23 July 2023. Before each exam question was asked, it was presented with an identical prompt, describing the nature of the single-choice test and the nature of the desired answer, limiting it to a single letter corresponding to the solution. Each question was asked in separate, independent sessions, thus eliminating the risk of any disruptive influence on the answers through multiple questions, due to the ability to contextualise the algorithm. Questions were asked in Polish language. Five sessions were conducted for each of the 119 questions involved in the experiment. This allowed an analysis of the probabilistic nature of the language model's responses and its "belief" in the veracity of a given answer, as asking the same question several times can generate different results. ChatGPT did not use any additional resources beyond the basic architecture obtained during training process.

Static analysis

An answer was only considered correct if at least 3 out of 5 sessions had a correct result. A model confidence factor was introduced, representing the ratio of the number of its dominant responses to the number of sessions of a given question ($n = 5$).

A detailed statistical analysis was performed using R Studio (Integrated Development Environment for R, R Studio, PBC, Boston, MA, USA). Exam scores and statistics provided by the Medical Examination Centre, Łódź, Poland, were compared with those provided by ChatGPT-3.5. Performance across question types and their subtypes was calculated using Pearson's χ^2 test. The Mann-Whitney U test with continuity correction was used to assess the relationship of the coefficient of question difficulty and the confidence of the model, both in terms of correct answers and the difference between question types. A Kruskal-Wallis ANOVA test was also used to assess the model confidence coefficient and question difficulty between question subtypes. In addition, Spearman's rank order correlation was used to assess the correlation between the question difficulty rates and the confidence of the LLM under study. In all tests, p less than 0.05 was taken as significant.

Results

The ChatGPT score was 54/119 pts. (45.38%) (Table I).

Performance was counted and scores were compared across question types and subtypes.

By type: "comprehension and critical thinking" questions and "memory" questions, correct answers

accounted for 50.00% and 42.03%, respectively (Table II).

By subtype: "genetics", "pathophysiology", "microscopic findings", "symptoms and signs", "medical procedures", "aetiology of diseases", "epidemiology", "medical guidelines", the correct answers in each subtype ranged from 28.57% in the subtype "epidemiology" to 76.92% in the subtype "microscopic findings" (Table III).

In the statistical analysis, the significance level was set at $p < 0.05$.

Using the Mann-Whitney *U* test, it was found that questions answered correctly by the ChatGPT-3.5 differed significantly in the confidence index ($p = 0.000054$) but not in the difficulty index ($p = 0.12$) – the confidence index was higher in questions answered correctly by the ChatGPT (Fig. 1).

The coefficient of certainty and difficulty did not differ between question types ($p = 0.93$ and $p = 0.41$, respectively).

From the Kruskal-Wallis ANOVA test, the confidence and difficulty coefficient did not differ between question subtypes.

Using Spearman's rank order correlation, it appears that the difficulty index did not correlate with the certainty index ($p = 0.66$).

Table. I. Correct and incorrect answers

Correct answer	Number of questions	Percentage
Yes	54	45.38
No	65	54.62

Table. II. The division into "memory" and "comprehension and critical thinking" question types

Category	Correct answer			
	Yes	Percentage	No	Percentage
Comprehension and critical thinking questions	25	50.00	25	50.00
Memory questions	29	42.03	40	57.97

Discussion

In Poland, the State Specialist Examination in Pathomorphology is an examination designed to qualify individuals as specialists in this field of medicine. The examination consists of a practical and a theoretical part. In Poland, a score of at least 60% is considered to be a positive result of the specialist examination in pathomorphology. A score of at least 75% in this part exempts the candidate from the oral (practical) examination. Similar qualifying examinations are used in many countries around the world. In our study, ChatGPT did not manage to meet the pass rate threshold. Broken down by subtype, it scored highest in the "microscopic findings" subtype (76.92%) and lowest in the "epidemiology" subtype (28.57%). We can surmise that ChatGPT, using all publicly available online data, also made use of available atlases of pathomorphology specimens with their descriptions and explanations when learning the model.

Geetha *et al*. conducted a study in which they assessed the AI model's ability to solve the resident question bank derived from the American Society for Clinical Pathology. The resident question bank is used for continuing education and preparation for Pathology Committee examinations. The tests consist of 565 anatomical pathology (AP) questions and 151 clinical pathology (CP) questions. Each question is accompanied by 5 possible answers, of which only one is correct. Questions that contained images, diagrams, and other graphics were excluded from the study. Questions were categorised based on pathology residents' scores into easy (> 70%), intermediate (40–70%), and difficult (< 40%). ChatGPT scores were also compared with resident scores, and repetition was checked by asking the AI model the same questions

Table III. Subdivision into subtypes

Topic	Correct answer			
	Yes	Percentage	No	Percentage
Genetics	5	50.00	5	50.00
Pathophysiology	20	41.67	28	58.33
Microscopic findings	10	76.92	3	23.08
Signs and symptoms	5	45.45	6	54.55
Medical procedures	1	33.33	2	66.67
Aetiology of diseases	6	33.33	12	66.67
Epidemiology	2	28.57	5	71.43
Medical guidelines	5	55.56	4	44.44

again. A total of 258 questions meeting the inclusion criteria were tested. Of these, 162 were on the AP and 96 were on the CP. ChatGPT performed better on the CP section (60.42%) than on the AP section (54.94%). Furthermore, its performance was better on easy questions (68.47%) than on intermediate (52.88%) and difficult questions (37.21%). The study revealed that ChatGPT's knowledge of pathology is not comprehensive and falls significantly short compared to that of pathology residents in training. The results when asking the same questions again were comparable to the first round, with some differences in some subspecialty areas. ChatGPT showed better performance compared to residents in specialised areas such as dermatopathology, autopsy, gynaecological pathology, coagulation, haematology, and chemistry, demonstrating its potential as a teaching tool in the future. However, overall, the performance of the ChatGPT was 56.98% lower than the average performance of the peers (62.81%) [10].

In a subsequent study conducted by Bielówka *et al*., the identical LLM tested yielded an overall score of 52.54% when responding to the PES questions in the field of allergology. This is a slightly higher score than in this study, which may suggest that it has more expertise in allergology than in pathology. This is most likely due to the greater number of sources available by the ChatGPT [11].

In another study conducted by Kufel *et al*., ChatGPT-3.5 performance was tested by answering PES questions in radiology and diagnostic imaging specialty from Spring 2023. The authors presented a similar classification of questions based on comprehension and critical thinking as well as memory questions. The analysed LLM achieved a score of 44.8% correct answers in the memory questions category. Moreover, ChatGPT scored 55.5% while answering the comprehension and critical thinking questions, which is slightly better than in our study [12].

To our knowledge, the usefulness of ChatGPT in the passing rate of the National Speciality Examination in Pathology has not yet been tested in other countries in Europe and the world by other researchers. We are therefore unable to compare the results of our study.

Pathomorphology is an old and developing field, so the development of digital pathology, online histological atlases, and other teaching aids is of great interest. Often the key element of diagnosis, and at the same time the factor that most influences the treatment applied, is precisely the result of the histopathological examination. However, currently ChatGPT does not seem to be able to replace the experience and textbook knowledge of specialists in this field of medicine. Perhaps with the improvement of online sources of knowledge in pathology, the results of ChatGPT will be more promising in the future.

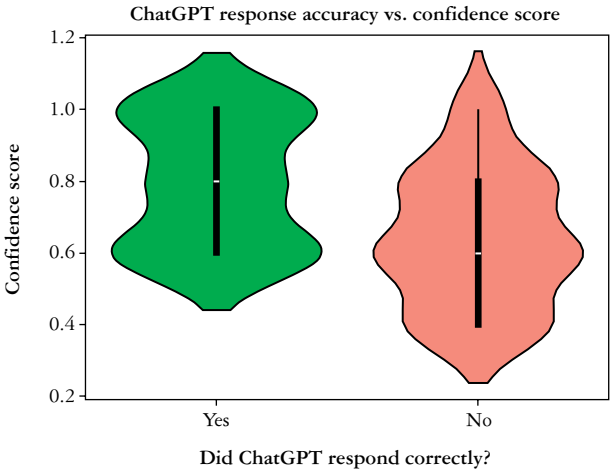

Fig. 1. Comparison of correctly and incorrectly answered questions with confidence index ($p = 0.000054$), Mann-Whitney U test

Conclusions

The study used ChatGPT-3.5 LLM to assess its ability to answer pathomorphology questions correctly, as required by the PES. The results show that the ChatGPT-3.5 achieved a 45.38% success rate, failing to meet the minimum threshold of 60% required for the PES.

Analysis of the breakdown by question type showed that the model performed better on "comprehension and critical thinking" questions (50.00%) than on "memory" questions (42.03%). Among the question subtypes, the highest performance was in the "microscopic findings" category (76.92%) and the lowest in the "epidemiology" category (28.57%).

Between 2009–2018, 166 doctors took the pathomorphology specialist exam, and 150 candidates achieved a pass rate of 90.36%, highlighting the significant advantage of humans over AI in solving tests [13].

For the ChatGPT model to become an effective tool to assist physicians in their daily work, it needs to be further refined and take into account the specificities of the field of pathomorphology. ChatGPT-3.5 may be a teaching tool in the future, but it cannot currently replace the knowledge and experience of specialists in the field.

Disclosures

1. Institutional review board statement: Not applicable.
2. We want to thank all the medical staff and medical assistants for their help in collecting biological samples and reviewing medical data.
3. Financial support and sponsorship: None.
4. Conflicts of interest: None.

References

1. Introducing ChatGPT. Available from: https://openai.com/blog/chatgpt (accessed: 01.09.2023).
2. Wu S, Roberts K, Datta S, et al. Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc 2019; 27: 457-470.
3. ChatGPT can now see, hear, and speak. Available from: https://openai.com/blog/chatgpt-can-now-see-hear-and-speak (accessed: 12.02.2024).
4. How OpenAI is approaching 2024 worldwide elections. Available from: https://openai.com/blog/how-openai-is-approaching-2024-worldwide-elections (accessed: 12.02.2024).
5. Jiang F, Jiang Y, Zhi H, et al, Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017; 2: 230-243.
6. Chen RJ, Wang JJ, Williamson DFK, et al. Algorithm fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng 2023; 7: 719-742.
7. PES – ważne zmiany – Polskie Towarzystwo Patologów. Available from: https://pol-pat.pl/index.php/2022/12/28/pes-wazne-zmiany/ (accessed: 12.02.2024).
8. Centrum Egzaminów Medycznych. Available from: https://www.cem.edu.pl/spec.php (accessed: 21.11.2023).
9. Centrum Egzaminów Medycznych. Available from: https://www.cem.edu.pl/pytcem/wyswietl_pytania_pes.php (accessed: 12.02.2024).
10. Geetha SD, Khan A, Khan A, Kannadath BS, Vitkovski T. Evaluation of ChatGPT"s pathology knowledge using board-style questions. medRxiv 2023.
11. Bielówka M, Kufel J, Rojek M, et al. Evaluating ChatGPT-3.5 in allergology: performance in the Polish Specialist Examination. Alerg Pol 2024; 11: 42-47.
12. 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: e430-e434.
13. Centrum Egzaminów Medycznych. Available from: https://www.cem.edu.pl/aktualnosci/spece/spece_stat2.php?nazwa=Patomorfologia (accessed: 12.02.2024).

Address for correspondence

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