Iran J Nucl Med. 2024;32(1):60-65 [Serial No. 62]

Homepage: https://irjnm.tums.ac.ir/

ORIGINAL RESEARCH ARTICLE

Assessing ChatGPT's performance in national nuclear medicine specialty examination: An evaluative analysis

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

[1]*Department of Biophysics, Faculty of Medical Sciences, Medical University of Silesia, Zabrze, Poland*

[2]*Professor Zbigniew Religa Student Scientific Association, Department of Biophysic, Faculty of Medical Sciences, Medical University of Silesia, Zabrze, Poland*

[3]*Wielospecjalistyczny Szpital Powiatowy S.A. im. dr B. Hagera Pyskowicka 47-51,42-612, Tarnowskie Góry, Poland*

[4]*Department of Medical and Molecular Biology, Faculty of Medical Sciences, Medical University of Silesia, Zabrze, Poland*

ARTICLE INFO

Article History:
Received: 22 October 2023
Revised: 09 December 2023
Accepted: 11 December 2023
Published Online: 21 December 2023

Keyword:
Artificial intelligence
Computer science
Language model
Nuclear medicine exam

*Corresponding Author:
Dr. Michał Bielówka
Address: Professor Zbigniew Religa Student Scientific Association, Department of Biophysic, Faculty of Medical Sciences, Medical University of Silesia, Zabrze, Poland
Email: michalbielowka01@gmail.com

ABSTRACT

Introduction: The rapid development of artificial intelligence (AI) has sparked a desire to analyse its potential applications in medicine. The aim of this article is to present the effectiveness of the ChatGPT advanced language model in the context of the pass rate of the polish National Specialty Examination (PES) in nuclear medicine. It also aims to identify its strengths and limitations through an in-depth analysis of the issues raised in the exam questions.

Methods: The PES exam provided by the Centre for Medical Examinations in Łódź, consisting of 120 questions, was used for the study. The questions were asked using the openai.com platform, through which free access to the GPT-3.5 model is available. All questions were classified according to Bloom's taxonomy to determine their complexity and difficulty, and according to two authors' subcategories. To assess the model's confidence in the validity of the answers, each questions was asked five times in independent sessions.

Results: ChatGPT achieved 56%, which means it did not pass the exam. The pass rate is 60%. Of the 117 questions asked, 66 were answered correctly. In the percentage of each type and subtype of questions answered correctly, there were no statistically significant differences.

Conclusion: Further testing is needed using the questions provided by Centre for Medical Examinations from the nuclear medicine specialty exam to evaluate the utility of the ChatGPT model. This opens the door for further research on upcoming improved versions of the ChatGPT.

Use your device to scan and read the article online

How to cite this article: Kufel J, Bielówka M, Rojek M, Mitręga A, Czogalik Ł, Kaczyńska D, Kondoł D, Palkij K, Mielcarska S. Assessing ChatGPT's performance in national nuclear medicine specialty examination: An evaluative analysis. Iran J Nucl Med. 2024;32(1):60-65.

https://doi.org/10.22034/IRJNM.2023.129434.1580

INTRODUCTION

ChatGPT (Generative Pre-trained Transformer), a product of OpenAI, was launched on November 30, 2022. Growing popularity and competition motivate manufacturers to improve their products constantly. Consumers are surpassing their own expectations by proposing innovative approaches to utilizing ChatGPT to enhance their work. Medics also see potential in the product for improving the diagnostic and treatment process for their patients [1]. ChatGPT can generate descriptions of radiological examinations, answer questions related to treatment regimens, and find examples of scientific articles. With such measures, some of the work of nuclear medicine specialists will be automated and physicians will gain time for patient contact and more thorough clinical analysis of individual medical cases. Among skeptics of the application of AI technology in medical fields, there is an argument about the "tendency" of this language model to give erroneous results or cite false articles. These concerns seem justified, and to avoid them, the tool, in the form of ChatGPT, must be strictly controlled, and supervised, despite continuous improvements by manufacturers. Hence, it is unlikely that in the medical field AI technology will replace human labour in the coming years.

The question of ChatGPT's usefulness in solving medical exams has been raised previously by other authors. A review by Levin et al. reports 19 articles evaluating the effectiveness of the ChatGPT in multiple-choice questions in medical disciplines. Among them, two articles dealt with Plastic Surgery In-Service Examination (performance - 55.8% and 54.9%), two with United Stated Medical Licensing Examinations (USMLE) (performance - 60% and 52%), another two - anesthesia examinations (performance - 69.7% and 56.2%). The mean performance was 61.1%, however, the level of difficulty of the examinations included varied greatly [2].

Researchers at Charles Sturt University tested ChatGPT on exam questions from a nuclear medicine course. The Chatbot's answers performed significantly worse than the students [3]. The objective of this article is to showcase the efficacy of the ChatGPT advanced language model within the context of the success rate of the polish National Specialty Examination (PES) in nuclear medicine Furthermore, it aims to identify its strengths and limitations through an in-depth analysis of the issues raised in the exam questions.

The form of the exam is a test, containing single-choice questions. Analysis of this comparison will allow conclusions to be drawn as to the correctness of the answers given by AI algorithms. The analytical reasoning ability of the tested language model will be compared to human cognitive skills.

METHODS

Examination and questions

The prospective study was conducted based on a publicly available set of questions from the nuclear medicine specialty exam. The question set was downloaded from the website of the Centre for Medical Examinations in Łódź, Poland. The criteria for selecting the set from among those available was the date of the exam. The most recent available question set (spring 2016) was selected from those provided by the exam centre. The question set originally consisted of 120 questions. One question was excluded by the Examination Committee because it was not in line with current medical knowledge. Another two questions were excluded as containing graphic information that could not be presented in question form to the language model under examination. In the end, 117 questions were used for analysis.

For subsequent analysis, questions were evaluated against Bloom's classification, and two author's subcategories [4]. For the purposes of the study, two categories including comprehension and critical thinking questions and memory questions were subdivided in the context of Bloom's classification. In addition, each question was assigned to subcategories as pertaining to issues of clinical management, description of imaging results, related to pathologies, related to radiopharmaceuticals, or medical procedures. The last of the three classification systems used involved identifying each question as relating to physical or clinical issues.

Data collection and analysis

The language model GTP-3.5, version 25 July 2023, was used for the study. For a set of questions from the exam, 5 independent, separately initiated sessions were conducted to assess the probabilistic aspect of the model. The same prompt was used to initialize each of the five sessions (Figure 1). ChatGPT-3.5 does not define explicit answers to questions during the internal analysis, but generates answers in a probabilistic manner and provides the user with an answer selected from a pool of the most probable ones. This means asking the same question several times will generate different answers, with a probability derived directly from the model's internal 'confidence' in their veracity. To assess the internal confidence in the validity of the answer, the question had to be

asked several times in independent sessions. Asking the same question in a given session remains unreliable due to the contextual analysis capability of the algorithm. The model notices the situation of asking the same question again and interprets this as a suggestion from the user to change to a different answer.

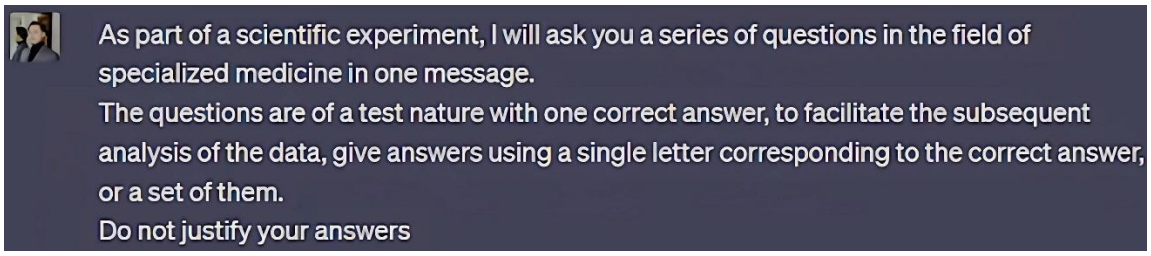

Figure 1. Prompt used to initialize each of the five sessions, English translation (the original prompt was called in Polish)

Statistical analysis

A series of statistical analysis were performed on the results obtained, assessing the correctness of the answers, taking into account the model's confidence coefficient and selected statistics provided by the Centre for Medical Examinations on the difficulty of the question. Despite the lack of access to the original data on which it was based, the Centre for Medical Examinations presents the methodology behind the calculation of the difficulty coefficient. The difficulty coefficient is calculated according to the following formula:

$$IDI = (Ns + Ni)/2n$$

where n represents the number of examinees in each of the extreme groups (extreme groups consisting of the top 27% of performers with the best results and the bottom 27% of performers with the worst results in the entire test), Ns-is the number of correct answers to the analysed task in the top-performing group, and Ni-is the number of correct answers to the analysed task in the bottom-performing group.

The certainty coefficient was expressed as the ratio of the number of dominant answers in successive sessions, to the number of sessions conducted (n=5). Determining which answer variant was dominant, a selection of PG model answers subject to evaluation as true or false in the context of the official answer sheet was made.

To assess the relationship between correctness of answers and question category membership, Pearson's chi-square test was used. To evaluate the quantitative variables (which consisted of the certainty coefficient and question difficulty) in the context of answer correctness, the Mann-Whitney U test with continuity correction was used. Spearman's rank order correlation test was used to assess the relationship between the question difficulty coefficient obtained from the Centre for Medical Examinations and the certainty coefficient calculated from model behaviour.

R Studio (Integrated Development Environment for R, R Studio, PBC, Boston, MA, USA) was used to perform the cited analyses. In all tests, p less than 0.05 was taken as significant.

RESULTS

Out of a pool of 120 questions in the specialty exam, 117 questions were used. Three questions were disregarded as they were marked as incompatible with current medical knowledge. Of the 117 questions asked, the ChatGPT had 66 correct answers and 51 incorrect answers, yielding a score of 56.41% (Table 1). The pass rate threshold for the exam is 60%.

Table 1. Correct and incorrect answers along with T-Student results

Correct answer	Number of questions	%	Mean Confidence	p value- Mean Confidence	Mean Difficulty	p value- Mean Difficulty
Yes	66	56.41026	84.8485	0.001765	78.0303	0.289357
No	51	43.58974	74.902		71.5686	

In terms of type, ChatGPT scored similarly in the range of 54.29-59.57% (Table 2). In questions divided by subtype, the score was between 52.38-83.33% (p=0.614) (Table 3). In terms of subject matter, ChatGPT also had a similar score in the range of 54.29-58.54% (Table 4).

Table 2. Division by type

Type	Correct answer			
	Yes	%	No	%
Memory questions	28	59.5745	19	40.4255
Comprehension and critical thinking questions	38	54.2857	32	45.7143

Table 3. Division by subtype

Subtype	Correct answer			
	Yes	%	No	%
Clinical management	11	52.38%	10	47.62%
Description of imaging findings	18	62.07%	11	37.93%
Related to diseases	5	83.33%	1	16.67%
Related to radiopharmaceuticals	21	52.50%	19	47.50%
Medical procedures	11	52.38%	10	47.62%

Table 4. Division by subject matter

Subject	Correct answer			
	Yes	%	No	%
Physical	48	58.54%	34	41.46%
Clinical	18	51.43%	17	48.57%

Statistical analysis using the Mann-Whitney test revealed that the questions that the ChatGPT answered correctly did not differ significantly in the difficulty index. The confidence index was higher in questions that the ChatGPT answered correctly (Figure 2). Furthermore, the difficulty index did not correlate with the certainty index. Moreover, both the certainty index and the difficulty index did not differ between question types (memory and thinking) and question topics (physical and clinical).

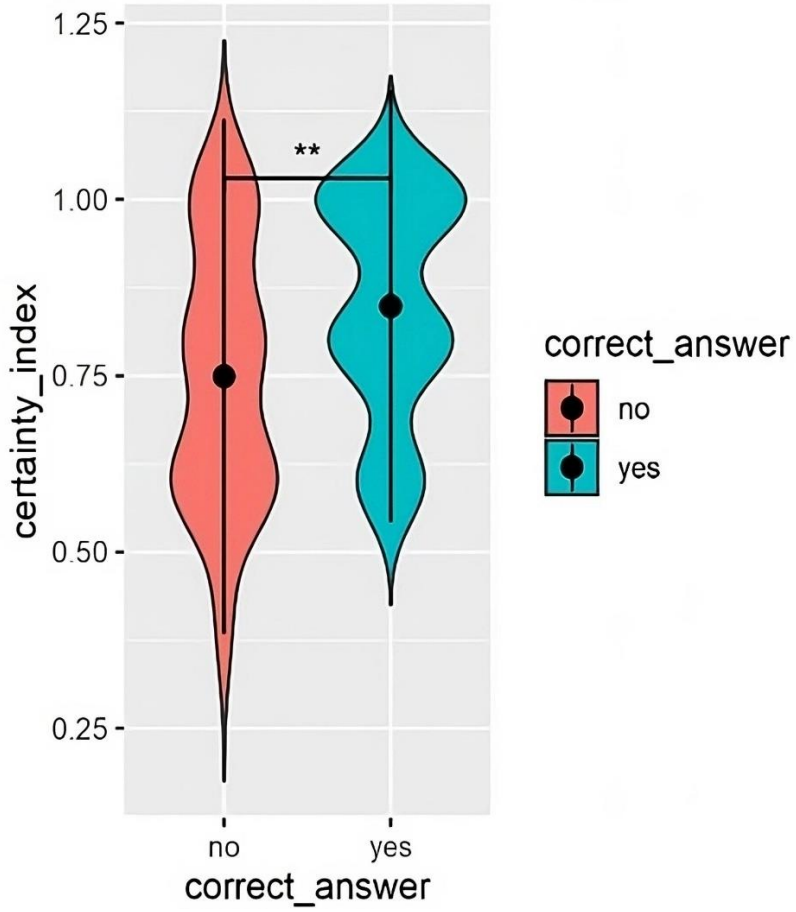

Figure 2. Confidence index depending on the correctness of the answer

DISCUSSION

The polish National Specialty Examination in nuclear medicine is the last part of the polish medical education system. It is the exam that doctors must pass to officially become a specialist in nuclear medicine. This exam assesses the doctor's knowledge, skills and competence in the field of nuclear medicine, including diagnosis and therapy using radioisotopes and molecular imaging techniques. In Poland, a score of 60% or more is required to pass the test. Similar qualifying examinations are used in many countries around the world.

In our study, ChatGPT-3.5 scored better (56%) than in the study by Currie et al. who tested the effectiveness of this language model for passing an exam and writing a nuclear medicine course paper covering material from the second and third years of a nuclear medicine science course [3]. His task included solving 2 computational exams, 6 writing tasks and 8 exams containing questions requiring longer answers to a problem. It performed unsatisfactorily in all computational tasks (31.7% compared to 67.2% for undergraduates) and in the written tasks (38.9% compared to 67.3% for undergraduates). In the other examinations, however, he scored better (51% compared to 57.4% for students), especially in questions requiring general and intermediate knowledge. On the other hand, the language model tested had the greatest difficulty in solving tasks requiring advanced and extremely detailed knowledge [3]. Similarly, in our experiment, ChatGPT had the most difficulty with specialist knowledge of radiopharmaceuticals and the least difficulty with more general knowledge related to human disease and interpretation of imaging findings.

A recent study by Oztermeli et al. presents optimistic results regarding the pass rate of the Medical Specialty Exam by ChatGPT. This is an exam prepared by the Student Selection and Placement Centre, which forms the basis of recruitment for resident doctors in Turkey. In this experiment, the language model tested answered 66% of the questions classified as questions with short content and 60.1% of the questions classified as those with longer content and more complex structure correctly [5]. These results are in line with the results of our experiment, in which the language model we studied performed better in memory-type questions than in comprehension and critical thinking questions-type questions requiring deeper understanding, context analysis and the ability to draw conclusions.

Unlike the aforementioned studies [3,5], in our study, we asked each question five times, because the language model under study generates responses probabilistically and provides the user with an answer selected from a pool of the most likely ones. We believe that this resulted in a deeper understanding of the model's breadth of knowledge and the identification of response generation patterns. This strategy allowed us to explore its text-generation process in more detail. Through repeated trials, the quality of the results obtained was increased, refining the consistency of the answers and eliminating random inaccuracies.

A result below the threshold for passing the exam obtained by ChatGPT could be the result of a number of factors. Hypothetically, the tool may not be provided with access to the necessary amount of data from the specialized literature or answers to questions in polish may pose more difficulty for the model (e.g., differences in terminology compared to English). Above all, however, ChatGPT is a linguistic model whose main purpose is not to provide scientifically supported answers from all specialized fields (although it is possible that this may be possible in the future). Regardless, the study demonstrates that OpenAI's language tool scored significantly lower on the nuclear medicine specialty exam than residents attempting to become specialists in this field. The study does not answer whether ChatGPT is capable of replacing nuclear medicine specialists; it was not designed to do so, moreover, it would be a methodological difficulty to find universal criteria requiring this AI tool to meet. The problems of working in the field of nuclear medicine go beyond providing more-or-less unreflective answers to the questions asked - they require criticality supported by the broad context of the situation, experience and awareness of the so-called "human factor." For this reason, it is impossible to assess the irreplaceability of specialist doctors.

Using the ChatGPT tool as a form of quality control of exam questions is also out of the question. Diagnostic tools of this type would require knowledge of the principles of their operation - the course of the decision-making process. The complex ChatGPT-type artificial intelligence model lacks the ability to quantifiably determine the level of difficulty of which it would be an indicator, as well as its regulation. The tool is still developing, so updates could make a difference in reliability.

However, it is worth leaning into the study conducted as an indicator of ChatGPT's usefulness

in medicine in general - an attempt to set the limits of its capabilities, as well as a milestone in the history of the development of publicly available artificial intelligence-based language tools. In the future, this will enable researchers to look at the state of knowledge prior to the use of more advanced and more effective tools.

One notable limitation of this study is the relatively small sample size, particularly when considering the complexity and diversity of the field of nuclear medicine. The study focused on a set of 117 questions from the polish National Specialty Examination in nuclear medicine. While this sample provided valuable insights into ChatGPT's performance on specific exam questions, the generalizability of the findings to a broader range of medical knowledge may be limited. Future studies with a larger and more diverse set of questions could offer a more comprehensive assessment of ChatGPT-3.5's capabilities in the field of nuclear medicine.

Another limitation is that the study utilized ChatGPT-3.5, which had its last knowledge update in January 2022. Given the rapid pace of advancements in AI, including language models, newer versions like ChatGPT-4 could potentially address some of the limitations identified in this study. Therefore, the findings might not fully reflect the current state of ChatGPT technology, and subsequent versions may exhibit improved performance. Advancements in newer versions of language models often include improvements in terms of understanding context, generating more coherent and contextually relevant responses, and addressing limitations identified in earlier versions. These improvements can be the result of larger and more diverse training datasets, fine-tuning processes, and enhancements in the underlying architecture. Upcoming studies should consider the latest versions of language models and explore how ongoing advancements impact their effectiveness in medical applications.

The study was conducted in polish, and the language specificity could impact ChatGPT's performance. The model might have been trained more extensively on English data, and certain medical terms or nuances in polish language medical questions may have posed challenges for the model. Examining ChatGPT-3.5's performance across multiple languages could provide a more comprehensive understanding of its language capabilities in the medical domain.

Furthermore, the questions used in the study were sourced from a specific exam format (single-choice questions) and covered a range of topics within nuclear medicine. The nature of these questions may not fully capture the complexity of real-world medical scenarios that clinicians encounter. Assessing ChatGPT's performance in a broader context, including clinical case studies or real patient scenarios, would provide a more realistic evaluation of its practical utility in the medical field.

CONCLUSION

The study outlined in this article demonstrates that ChatGPT-3.5 was unable to successfully pass the state specialty exam in nuclear medicine. It scored 56.41%, so it did not meet the minimum score threshold of 60%. The questions answered correctly by the AI algorithm did not differ significantly in the difficulty index and the confidence factor was higher in the questions answered correctly by this technology. Over a period of nine years (2009-2018), 112 people took the exam, with 111 passers (99.11%). Our study shows that humans are definitely better at solving the test than artificial intelligence. However, further testing using the official questions provided by the Medical Examination Centre is needed to truly assess the effectiveness of the ChatGPT in successfully passing the specialist exam; perhaps a possible change in the methodology of conducting the exam in the prompt will change the characteristics and accuracy of the answers provided by ChatGPT. It is also important to remember that the technology is improving all the time, the ChatGPT is still learning using LLM and its ability to correctly solve the test should evolve. Further testing of AI technology for nuclear medicine PES questions is needed to gain a more complete understanding of its application. Undoubtedly, the development of AI has the potential to positively impact the work of nuclear medicine doctors, but this requires further work on the technology.

REFERENCES

1. The Lancet Digital Health. ChatGPT: friend or foe? Lancet Digit Health. 2023 Mar;5(3):e102.
2. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: A systematic review and a meta-analysis. BJOG-INT J OBSTET GY.2023 Aug [cited 2023 Nov 25]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/1471-0528.17641
3. Currie G, Barry K. ChatGPT in nuclear medicine education. J Nucl Med Technol. 2023 Sep;51(3):247-54.
4. Foreland M. Bloom's Taxonomy. In: Orey M, editor. Emerging perspectives on learning, teaching and technology. North Charleston: CreateSpace; 2010.
5. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: An observational study. Medicine (Baltimore). 2023 Aug 11;102(32):e34673.