Exploring the performance of ChatGPT-3.5 in addressing dermatological queries: a research investigation into AI capabilities

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

[1]Students' Scientific Association of Computer Analysis and Artificial Intelligence at the Department of Radiology and Nuclear Medicine of the 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]Multi-specialty District Hospital S.A. Dr. B. Hager Pyskowicka, Tarnowskie Góry, Poland
[5]Department of Medical and Molecular Biology, Faculty of Medical Sciences in Zabrze, Medical University of Silesia in Katowice, Poland
[6]Faculty of Medical Sciences in Katowice, Medical University of Silesia, Katowice, Poland

Dermatol Rev/Przegl Dermatol 2024, 111, 26–30
DOI: https://doi.org/10.5114/dr.2024.140796

CORRESPONDING AUTHOR:
Dr. 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

ABSTRACT

Introduction: In the 21st century's era of rapid technological advancement, ChatGPT-3.5, an artificial intelligence (AI) language model, is scrutinized for its application in dermatology. Using 119 questions from the National Specialist Examination (PES), we assess ChatGPT-3.5's performance by comparing it to human skills and addressing ethical implications.

Objective: Our primary aim is to evaluate ChatGPT-3.5's proficiency in responding to 119 dermatology questions from the PES. The study emphasizes ethical considerations and compares the model's knowledge and skills to those of human dermatologists.

Material and methods: Utilizing the 2023 PES question database, questions were categorized by Bloom's taxonomy and thematic content. ChatGPT-3.5, version of 3 August 2023, answered 119 questions in five sessions, allowing for a probabilistic evaluation. Statistical analyses, conducted using R Studio, assessed correctness, confidence, and difficulty.

Results: ChatGPT-3.5 achieved a 49.58% correct response rate, below the 60% passing threshold. No significant differences in difficulty or correlations between difficulty and certainty were observed. Varied performance across question types highlighted strengths and weaknesses. Despite suboptimal results, ChatGPT-3.5's differential performance offers insights, suggesting future improvements. The study advocates for ongoing research into AI integration in dermatology, envisioning a promising role for AI in assisting dermatologists.

Conclusions: Ethical considerations are crucial for effective AI introduction, minimizing errors, and enhancing dermatological healthcare quality, fostering optimism for AI's evolving role in dermatology.

Key words: medical education, artificial intelligence, dermatology, venereology, ChatGPT-3.5.

INTRODUCTION

The 21st century is a time of rapid technological development, computers, and automation in many aspects of life. ChatGPT, available since 30 November 2022, has been rapidly gaining more users worldwide. This product from OpenAI is a language model based on artificial intelligence (AI) designed to respond to user queries globally [1]. By utilizing advanced techniques such as deep learning, machine learning, artificial neural networks, and natural language processing, artificial intelligence attempts to replicate the cognitive processes of humans [2, 3]. OpenAI continues to refine its product by releasing successive versions of ChatGPT to expand its applications in both professional and daily life [4]. In medicine, researchers are also exploring the use of this technology to enhance diagnostics, develop new studies, and create medications [2, 5].

Among various medical fields, there is growing interest in employing AI in dermatology. The primary tools for dermatologists include visual assessment of macroscopic images and the use of a dermatoscope. The development of AI technology can not only assist dermatologists, but also benefit non-dermatology professionals and improve doctor-patient communication through teledermatology [6].

The authors of this publication aimed to compare the knowledge and skills currently demonstrated by AI (using ChatGPT) with human skills in dermatology. To achieve this goal, 119 questions from the National Specialist Examination (PES) test section were posed to ChatGPT. The PES questions are single-choice and assess the ability to conclude from the information provided in the question. To achieve a positive result, a doctor must answer at least 60% of the questions correctly, the same criterion being applied in the study [7]. Although the use of artificial intelligence in medicine poses ethical and legal dilemmas, the application of AI in various stages of dermatological diagnostics and scientific research under human supervision should be considered. AI technology can contribute in the future to faster diagnosis of skin changes, including tumours, inflammatory conditions, allergic reactions, patient education, and the advancement of dermatology and venereology as medical disciplines.

OBJECTIVE

Our main aim is to assess the effectiveness of ChatGPT-3.5 in addressing 119 dermatology-related queries sourced from the PES. By conducting this evaluation, we aim to comprehend both the strengths and limitations of ChatGPT-3.5 in the domain of dermatology, thereby delivering valuable insights to the ongoing discourse concerning AI applications in healthcare.

MATERIAL AND METHODS

Examination and questions

This study was conducted using a publicly available database of questions from the National Specialist Examination, accessible on the website of the Medical Examination Centre in Lodz, Poland [8]. Each examination consists of 120 single-choice questions. The selection criterion was the subject matter within the field of dermatology and venereology, as well as the timing of the examination. The latest available exam from the spring of 2023 was chosen. One question was excluded from the study because it was inconsistent with current medical knowledge. Ultimately, the analysis included 119 questions.

The categories were created by the authors to segregate question types in order to obtain better statistical results. The categorisation of questions according to Bloom's taxonomy [9] was done by 2 independent researchers, if there were conflicts, they were solved by a third person. The same applied to the categorisation of questions into thematic categories, such as medical procedures, clinical proceedings, diagnostics, medication, and those related to diseases. Furthermore, in a similar manner, a categorization has been established, delineating two primary question types: those pertaining to comprehension and critical thinking, and those centred around memory. This division seeks to distinguish between inquiries that assess understanding and analytical skills, and those focused on simple recall.

Data collection and analysis

The GPT-3.5 language model version as of 3 August 2023 was utilized to provide answers to the questions. For each examination question in the prepared set (n = 119), five separate and entirely independent question-answering sessions were conducted. This approach allowed for the exploration of the probabilistic nature of the examined language model, which, when responding to questions, provides the most probable answer according to its internal mechanisms. Therefore, it is possible that multiple questions will yield divergent responses, reflecting the inherent uncertainty of the model. Conducting questions in separate sessions prevented the language model from being influenced by its previous responses. Each question was preceded by an identical prompt, facilitating a fair simulation of a single-choice test, limiting responses to a single letter.

Statistical analysis

The number of correct answers provided by ChatGPT-3.5 was calculated based on the criterion that considered a response correct if it was obtained in at least

3 out of 5 sessions. A certainty coefficient of the language model was introduced as the ratio of the number of dominant responses to a given question to the total number of conducted sessions. The significance between the distributions of questions answered correctly and incorrectly by ChatGPT-3.5 and the types and subtypes of questions was evaluated using Pearson's χ^2 test. Shapiro-Wilk test was applied to evaluate distribution of continuous data. To compare differences in numeric variables including difficulty index, certainty coefficient between questions answered correctly and incorrectly, Mann-Whitney *U* and Kruskal-Wallis ANOVA tests were performed. To assess the relationship between numeric variables, Spearman's rank-order correlation was used. *P*-values < 0.05 were considered statistically significant. All statistical analyses were conducted using the R Studio environment (an open-source integrated development environment for the R language) version 1.1.46.

RESULTS

The number of correct answers provided by ChatGPT was 59 out of 119 points (49.58%) (table 1). The performance was assessed for different types of questions and their subtypes. Results were compared between types: "memory questions" and "comprehension and critical thinking questions", as well as subtypes such as "medical procedures", "clinical proceedings", "diagnostics", "medication" and "related to diseases" (tables 2, 3).

Table 1. Correct and incorrect answers

Correct answer	Number of questions	%
Yes	59	49.58
No	60	50.42

Table 2. The division into "memory questions" and "comprehension and critical thinking questions". χ^2 test, $p = 0.79$

Category	Correct answer			
	Yes	%	No	%
Comprehension and critical thinking questions	27	50.94	26	49.06
Memory questions	32	48.48	34	51.52

Table 3. Division into subtypes. χ^2 test, $p = 0.20$

Topic	Correct answer			
	Yes	%	No	%
Medical procedures	3	37.50	5	62.50
Clinical proceedings	4	26.67	11	73.33
Diagnostics	15	55.56	12	44.44
Medications	9	69.23	4	30.77
Related to diseases	28	50.00	28	50.00

Statistical analysis using the Mann-Whitney *U* test (fig. 1), Spearman's rank-order correlation, and Kruskal-Wallis ANOVA (tables 4, 5) revealed that questions for which ChatGPT provided a correct answer did not significantly differ in difficulty, and the difficulty index did not correlate with the confidence index.

The confidence index was higher for questions to which ChatGPT provided a correct answer. However, both the difficulty coefficient and confidence index did not differ between question types and subtypes.

DISCUSSION

The specialization in dermatology and venereology is one of the most sought-after in Poland (in 2023, 392 applications were submitted for 40 residency positions), requiring a high number of points to be obtained during the Medical Final Examination for a positive qualification [10]. The Specialist Examination in Dermatology and Venereology is a single-choice exam and constitutes the final stage of specialization, equivalent to obtaining the title of a specialist in this field. Like every speciality exam in Poland, it consists of both theoretical and practical

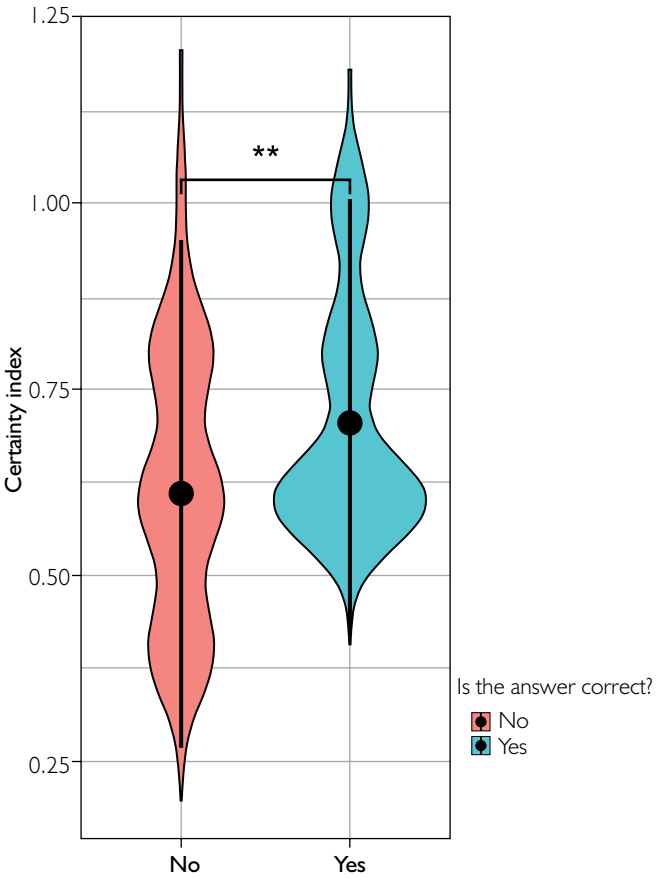

Figure 1. Comparison of certainty index between questions answered correctly and incorrectly by ChatGPT-3.5. Mann-Whitney *U* test, $p < 0.01$

parts. To pass the theoretical exam, one must correctly answer at least 60% of the questions, and achieve a score above 75% in order to be exempt from the practical exam. Detailed statistics on the pass rates of the National Specialist Examination for Dermatology and Venereology from 2009 to 2018 show a pass rate of over 91% [11].

In our study, ChatGPT achieved a score of 49.58%, providing 59 correct and 60 incorrect answers, equivalent to a negative result on the specialist exam. The artificial intelligence model performed poorly on questions related to clinical proceedings (26.67%) and medical procedures (37.50%), obtaining the lowest scores in these two subcategories. Surprisingly, it demonstrated a high accuracy rate of 69.23% in prescribing appropriate treatments, achieving the highest score in this subcategory. Another surprising finding is that ChatGPT performed worse on "memory" questions (48.48%) than on "comprehension and critical thinking" questions (50.94%). The better performance of the AI model in the "comprehension and critical thinking" category does not stem from ChatGPT's ability to process multiple data simultaneously, as the AI model processes text sequentially, step by step. There was no significant difference in frequencies of correct and incorrect answers between types and subtypes of questions.

In a study by Lewandowski *et al.*, ChatGPT-4 exceeded the pass threshold in all three dermatology specialty certificate tests, achieving a minimum of 80% and 70% of correct answers for English and Polish versions, respectively. Furthermore, ChatGPT-4 answered "clinical image" questions with an average accuracy of 92.98% and 84.21% for English and Polish questions, respectively, which appears to be a very good result [12].

In the study by Passby *et al.*, both ChatGPT-3.5 and ChatGPT-4 received positive results in the dermatology Specialty Certificate Examination, answering 84 randomly selected single-choice questions from a sample bank of dermatological questions. ChatGPT-3.5 scored an overall result of 63.1%, while ChatGPT-4 achieved 90.5% [13].

Joly-Chevrier *et al.* used 241 multiple-choice questions from the publicly available Term-in-Review question bank for their study. ChatGPT-3.5 answered correctly 59.3% of them, reaching 80% of correct answers in the "basic science and structure of the skin" and "paediatric dermatology" categories. AI performed the worst in the "benign and malignant neoplasm" category, achieving only 30% correctness [14].

In a study conducted by Kufel *et al.*, the same language model was examined in terms of the pass rate of the National Specialist Examination (PES) in radiology and imaging diagnostics within the Polish education system. The study also introduced a confidence index

Table 4. Comparison of the difficulty index between question subtypes. Kruskal-Wallis ANOVA test

Subtype	Median	q1	q3	*P*-value
Medical procedures	0.841	0.591	0.864	0.099
Clinical proceedings	0.636	0,500	0.727	0.099
Diagnostics	0.682	0,545	0.909	0.099
Medications	0.682	0,545	0.864	0.099
Related to diseases	0.591	0.455	0.750	0.099

Table 5. Comparison of the confidence index between question subtypes. Kruskal-Wallis ANOVA test

Subtype	Median	q1	q3	*P*-value
Medical procedures	0.60	0.60	0.80	0.96
Clinical proceedings	0.60	0.40	0.80	0.96
Diagnostics	0.60	0.60	0.80	0.96
Medications	0.60	0.60	0.80	0.96
Related to diseases	0.60	0.60	0.80	0.96

for the language model. However, unlike this study, it was determined based on a direct question posed to ChatGPT, which assessed the certainty of its answers on a scale from 1 to 5. In both studies, it was observed that the confidence index was higher for questions to which ChatGPT provided a correct answer. This observation may indicate the potential of both methods to assess the reliability of this language model's responses to a given problem, considering the probabilistic nature of AI responses. Similar accuracy results (52%) were achieved, despite only one attempt at answering an exam question, as opposed to 5 attempts in this study. Additionally, it was demonstrated that ChatGPT performed significantly better in clinical management (75% of correct responses) in the field of radiology compared to clinical proceedings in dermatology and venereology (26.67% of correct responses). This finding is surprising but may be attributed to a small question sample [15].

During data collection and AI model training, such as ChatGPT, all freely available online knowledge sources were utilized. Unfortunately, much of the Polish medical literature, including dermatology and venereology, is not available online or lacks free access, posing a limitation for ChatGPT during data collection. In cases where necessary information for providing answers is not available in the same language, ChatGPT resorts to foreign data, and translating them into Polish may present an additional challenge [16].

CONCLUSIONS

The article presents a study in which ChatGPT-3.5 was used to respond to questions from the National

Specialist Exam. The results indicate that the model achieved a correct answer rate of 49.58%, which is below the 60% threshold required to pass the exam. In questions correctly answered by ChatGPT, the difficulty index did not significantly differ and did not correlate with the confidence index. Moreover, the confidence coefficient was higher for questions to which ChatGPT provided a correct answer.

Between 2009 and 2018, 476 individuals took the exam, with 456 candidates obtaining a positive result, yielding a pass rate of 95.8%. This underscores a clear advantage of human performance over artificial intelligence in test solving. However, with the dynamic development of artificial intelligence and the creation of increasingly efficient language models, promising improvements in their competencies can be expected. This provides an optimistic outlook for the coming years, suggesting the emergence of artificial intelligence that could assist dermatologists in their daily work.

Further research into the utilization of artificial intelligence in dermatology is essential, considering the improvement of model results and technical solutions. Effectively integrating these technologies into medical practice is necessary to minimize errors and enhance the quality of healthcare in the field of dermatology.

FUNDING

No external funding.

ETHICAL APPROVAL

Not applicable.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

References

1. Introducing ChatGPT [Internet]. [cited 2023 Sep 1]. Available from: https://openai.com/blog/chatgpt
2. **Niewęgłowski K., Wilczek N., Madoń B., Palmi J., Wasyluk M.:** Zastosowania sztucznej inteligencji (AI) w medycynie. Med Og Nauk Zdr 2021, 27, 213-219.
3. **Patel S., Wang J.V., Motaparthi K., Lee J.B.:** Artificial intelligence in dermatology for the clinician. Clin Dermatol 2021, 39, 667-672.
4. Introducing GPTs [Internet]. [cited 2023 Dec 21]. Available from: https://openai.com/blog/introducing-gpts
5. **Fishman E.K., Weeks W.B., Lavista Ferres J.M., Chu L.C.:** Watching innovation in real time: the story of ChatGPT and radiology. Can Assoc Radiol J 2023, 74, 622-623.
6. **Young A.T., Xiong M., Pfau J., Keiser M.J., Wei M.L.:** Artificial intelligence in dermatology: a primer. J Invest Dermatol 2020, 140, 1504-1512.
7. Centrum Egzaminów Medycznych [Internet]. [cited 2023 Dec 21]. Available from: https://www.cem.edu.pl/spec.php
8. Centrum Egzaminów Medycznych [Internet]. [cited 2023 Sep 1]. Available from: https://cem.edu.pl/index.php
9. **Foreland M.:** Bloom's Taxonomy. In: Emerging Perspectives on Learning, Teaching, and Technology [Internet]. Global Text Project; 2010. Available from: https://textbookequity.org/Textbooks/Orey_Emergin_Perspectives_Learning.pdf
10. www.rynekzdrowia.pl [Internet]. 2023 [cited 2023 Dec 22]. Są wyniki jesiennego naboru na specjalizacje. Hitem m.in. radiologia, dermatologia i psychiatria. Available from: https://www.rynekzdrowia.pl/Nauka/Sa-wyniki-jesiennego-naboru-na-specjalizacje-Hitem-m-in-radiologia-dermatologia-i-psychiatria,251762,9.html
11. Centrum Egzaminów Medycznych [Internet]. [cited 2023 Dec 22]. Available from: https://www.cem.edu.pl/aktualnosci/spece/spece_stat2.php?nazwa=Dermatologia%20i%20wenerologia
12. **Lewandowski M., Łukowicz P., Świetlik D., Barańska-Rybak W.:** An original study of ChatGPT-3.5 and ChatGPT-4 Dermatological Knowledge Level based on the Dermatology Specialty Certificate Examinations. Clin Exp Dermatol 2023; llad255.
13. **Passby L., Jenko N., Wernham A.:** Performance of ChatGPT on dermatology Specialty Certificate Examination multiple choice questions. Clin Exp Dermatol 2023; llad197.
14. **Joly-Chevrier M., Nguyen A.X.L., Lesko-Krleza M., Lefrançois P.:** Performance of ChatGPT on a practice dermatology board certification examination. J Cutan Med Surg 2023, 27, 407-409.
15. **Kufel J., Paszkiewicz I., Bielówka M., Bartnikowska W., Janik M., Stencel 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.
16. **Porter E., Murphy M., O'Connor C.:** Chat GPT in dermatology: Progressive or problematic? J Eur Acad Dermatol Venereol 2023, 37, e943-e944.

Received: 4.01.2024
Accepted: 25.02.2024
Online publication: 27.06.2024

How to cite this article
Rojek M., Kufel J., Bielówka M., Mitręga A., Kaczyńska D., Czogalik Ł., Kondoł D., Palkij K., Mielcarska S., Bartnikowska W: Exploring the performance of ChatGPT-3.5 in addressing dermatological queries: a research investigation into AI capabilities. Dermatol Rev/Przegl Dermatol 2024, 111, 26–30. DOI: https://doi.org/10.5114/dr.2024.140796.