© Pol J Radiol 2025; 90: e519-e525
DOI: https://doi.org/10.5114/pjr/208356

Received: 22.04.2025
Accepted: 17.07.2025
Published: 21.10.2025

Polish Journal of Radiology

http://www.polradiol.com

Original paper

Advancing AI in radiology: a comparative analysis of ChatGPT-o1-preview and ChatGPT-3.5 in the Polish National Specialization Exam

Adam Mitręga[1,2,#,A,B,D,E], Michał Bielówka[1,2,#,A,B,D,E], Dominika Kaczyńska[1,A,D,E,F], Natalia Denisiewicz[1,A,E,F], Mikołaj Magiera[1,E,F], Marcin Rojek[1,C,D], Maja Dreger[1,C,D], Jakub Kufel[3,A,E], Miłosz Zbroszczyk[3,A,E]

[1]Students' Scientific Association of Computer Analysis and Artificial Intelligence at the Department of Radiology and Nuclear Medicine of the Medical University of Silesia in Katowice, Poland
[2]Department of Biophysics, Faculty of Medical Sciences in Zabrze, Medical University of Silesia in Katowice, Zabrze, Poland
[3]Department of Radiodiagnostics, Interventional Radiology and Nuclear Medicine, Medical University of Silesia, Katowice, Poland

[#]These authors contributed equally to this work.

Abstract

Purpose: The aim of this study was to evaluate the performance of the ChatGPT-o1-preview language model in solving the Polish National Specialization Exam (PES) in radiology and imaging diagnostics and compare its results with previous versions of the model.

Material and methods: A set of 119 valid radiology exam questions from Spring 2023 was analyzed. Each question was classified by type, subtype, and clinical relevance. ChatGPT answered each question five times using a standardized prompt with a 5-point confidence scale. Performance was assessed using accuracy and declared and calculated difficulty indices. Statistical analysis was performed in Python with a significance level of $p < 0.05$, and results were compared with a previous model version.

Results: The model achieved a score of 93.33% correct answers, comparable to the average physician score of 94.86%. ChatGPT-o1-preview showed exceptional accuracy in "memory" questions, with over 96% correct answers. This result, significantly higher than that of the older ChatGPT-3.5 model (52%), demonstrates progress in artificial intelligence (AI) capabilities. The model also exhibited higher confidence in its responses, indicating better adaptation to medical exams.

Conclusions: Despite its high accuracy, the study was based on a relatively small set of questions, which limits the ability to fully assess the model's effectiveness. The results indicate the potential of AI as a tool to support clinical work, but further, more extensive research is necessary to evaluate its applicability and reliability in the medical environment.

Key words: artificial intelligence, ChatGPT, radiology, medical imaging, AI.

Introduction

Artificial intelligence (AI) is playing an increasingly important role in medicine, becoming the subject of extensive research [1, 2]. Its potential application in diagnostics, broadly defined data analysis, and clinical decision-making generates great interest.

At present, AI is widely used in daily life by many users – from voice assistants and automatic translators to content recommendation systems [3, 4]. The main reason for the growing popularity of AI is the reduction in time required

Correspondence address:
Natalia Denisiewicz, Students' Scientific Association of Computer Analysis and Artificial Intelligence at the Department of Radiology and Nuclear Medicine of the Medical University of Silesia in Katowice, Poland, e-mail: dndenisiewicz@gmail.com

Authors' contribution:
A Study design · **B** Data collection · **C** Statistical analysis · **D** Data interpretation · **E** Manuscript preparation · **F** Literature search · **G** Funds collection

for information retrieval, content creation, programming, analysis, and many other everyday tasks compared to performing them manually [5].

Since its initial versions, ChatGPT has undergone significant transformation. Early language models had limited ability to understand context and process complex information. With subsequent updates, their accuracy, ability to generate relevant responses, and capacity to analyze complex issues have significantly improved [6].

One area of AI model evaluation involves assessing their ability to pass medical exams. The Polish National Specialization Exam (PES) in radiology and medical imaging is a demanding test that evaluates not only theoretical knowledge but also the ability to analyze clinical cases [7].

The use of AI in clinical practice raises numerous controversies, mainly due to potential errors, biases, and the risk of providing unverified information, including false sources [8]. Additionally, legal issues and the lack of clear regulations limit the broad application of AI, such as ChatGPT, in medicine. Concerns also exist regarding liability for potential diagnostic or therapeutic errors based on information generated by language models [9].

The aim of this study is to evaluate the performance of the ChatGPT-o1-preview model in solving the PES exam in radiology and medical imaging and to compare its results with those of earlier model versions. The analysis will assess the model's effectiveness in solving exam tasks and examine how its capabilities have evolved over different versions.

Material and methods

Examination and questions

The questions were obtained using a Python-based scraper, courtesy of the Examination Center in Łódź. The analyzed set of questions was from the spring of 2023, corresponding to the set used in the study to which the model's performance is being compared. The collected data included the content of the questions and their status, classified as valid or rejected. Additionally, the analyzed information included statistics related to the given exam, such as the difficulty index, calculated based on the number of correct answers provided by physicians participating in the session. The specialization exam in radiology consists of 120 single-choice questions. The inclusion criteria for this study required that questions remain classified as non-excluded and contain only textual content. A total of 119 questions were included in the analysis, excluding one that had been removed from the database. Some questions were included despite retrospective substantive concerns, as long as these concerns did not result in the exclusion of the question from the database. For the purpose of a more detailed statistical analysis, all questions were divided into types and subtypes and by clinical significance. Three classification schemes were formulated, taking conceptual guidance from Bloom's Taxonomy [10]:

- Types: "Comprehension and critical thinking questions" or "Memory questions."
- Subtypes: "Description of imaging results," "Clinical proceedings," "Related to diseases," "Calculation and classification."
- Clinical relevance: "Clinical questions" and "Other questions."

Data collection and analysis

The continuous analysis project of language model performance, of which this study is a part, undergoes ongoing methodological improvements. As a result, some parameters collected during the evaluation of the previous model have been expanded or removed, posing a challenge in the statistical analysis of performance differences. One notable example concerns the internal confidence scores of the models. In the current evaluation, a refined method was used to collect confidence estimates across multiple interaction sessions, allowing for higher-resolution insights into the model's decision-making processes. However, this approach was not implemented during the earlier assessment of the previous model version. Consequently, the corresponding metrics are unavailable for that dataset and could not be reliably reconstructed or interpolated. This discrepancy limits direct comparison and has been acknowledged as a methodological limitation of the study.

In this study, each question was presented five times in separate sessions, preceded by a more detailed prompt, unlike the evaluation of the previous language model, where each question was asked only once. The prompt precisely defined the test guidelines, explained its format, ensured a reliable simulation of a single-choice test, and introduced a confidence scale from 1 to 5 (described in detail in words). Based on this scale, ChatGPT determined its confidence in the correctness of its answer, which was then used to calculate the Average Difficulty Index Declared by the Language Model.

Another key difference in this analysis is the introduction of a new coefficient – the Calculated Difficulty Index for the Language Model. This metric represents the ratio of the most frequently chosen response for a given question to the total number of test sessions conducted.

Statistical analysis

The methodology applied in this study follows an approach analogous to the previously conducted analysis, ensuring comparability of results while maintaining methodological consistency. A statistical significance threshold of $p < 0.05$ was adopted for all statistical analyses. The analyses were performed in a Python environment (version 3.11.1) using libraries such as numpy, scipy, pandas, matplotlib, and plotly, ensuring reproducibility and high precision of calculations. Appropriate statistical tests were applied to the analysis of individual variables. None

of the continuous variables – such as the Average Difficulty Index Declared by the Language Model, the Human Difficulty Index, and the Calculated Difficulty Index for the Language Model – followed a normal distribution, which justified the use of the non-parametric Mann-Whitney *U* test in analyses involving categorical variables such as Type, Answer correctness, and Division into clinical and other questions. Spearman's rank correlation test was used for associations among continuous variables.

Performance analysis of the new language model

To assess the relationship between continuous variables and binary categorical variables such as "Type," "Clinical," and "Did ChatGPT Respond Correctly," Mann-Whitney *U* tests were used. Meanwhile, relationships between pairs of continuous variables, such as the Calculated Difficulty Index for the Language Model, the Human Difficulty Index, and the Average Difficulty Index Declared by the Language Model, were analyzed using Spearman's rank correlation test, taking into account their discrete nature.

Comparison with the old language model

Comparing the performance of both language model versions using statistical tests was challenging due to the high efficiency of the new model and enhanced methodology applied in its evaluation. The primary criterion for analysis was the question: "Did ChatGPT pass the specialization exam?" Additionally, the distribution of the variable "Certainty of assessment according to a 5-degree scale" from the older model's performance analysis was compared with the analogous variable describing the new model, the Average Difficulty Index Declared by the Language Model, using the Mann-Whitney *U* test [7].

Results

Due to the exceptionally high accuracy of the language model's responses (7 incorrect vs. 112 correct out of 119 questions), conducting statistical analyses was challenging or even impossible. This was particularly the case when analyzing the distribution of question correctness in relation to question type. Given the insufficient sample size for statistical testing, the analysis was limited to presenting contingency tables (Tables 1 and 2).

Spearman's rank correlation was used to analyze a range of quantitative variables, including the Calculated Difficulty Index for the Language Model, the Average Difficulty Index Declared by the Language Model, and the Human Difficulty Index. The results are presented in Table 3.

Two statistically significant correlations were identified:

- between the Average Difficulty Index Declared by the Language Model and the Calculated Difficulty Index for the Language Model ($p = 0.01$, $r = 0.23$, weak correlation);
- between the Average Difficulty Index Declared by the Language Model and the Human Difficulty Index ($p < 0.001$, $r = 0.36$, weak correlation).

To compare the relationships of the quantitative variables listed in Table 3 with answer correctness (yes or no), a series of Mann-Whitney *U* tests were performed (Table 4).

A number of statistically significant correlations were identified. The division based on answer correctness was found to statistically significantly differentiate ($p < 0.05$) all three analyzed distributions: the Average Difficulty Index Declared by the Language Model ($U = 139.5$, $p < 0.001$, no $\bar{x} = 3.80$, yes $\bar{x} = 4.60$), Calculated Difficulty Index for the Language Model ($U = 144$, $p < 0.001$, no $\bar{x} = 0.60$, yes $\bar{x} = 1.00$), Human Difficulty Index ($U = 190$, $p = 0.02$, no $\bar{x} = 0.50$, yes $\bar{x} = 0.69$).

Table 1. Comparison of correct and incorrect answers by type

Did ChatGPT respond correctly?	Comprehension and critical thinking questions	Memory questions
Yes	84	28
No	6	1

Table 2. Comparison of correct and incorrect answers of clinical and non-clinical questions

Did ChatGPT respond correctly?	Clinical questions	Other questions
Yes	94	18
No	5	2

Table 3. Spearman's rank correlation results

Index		Spearman *R*	*p*-value
Average Difficulty Index Declared by the Language Model	Calculated Difficulty Index for the Language Model	0.230218	0.01
Average Difficulty Index Declared by the Language Model	Human Difficulty Index	0.364792	< 0.001
Calculated Difficulty Index for the Language Model	Human Difficulty Index	0.083245	0.37

Table 4. Mann-Whitney *U* tests results

Index	Category	*U* statistic	*p*-value
Calculated Difficulty Index for the Language Model	Type	1452	0.18
	Answer correctness	144.5	< 0.001
	Division into clinical and other questions	966	0.80
Average Difficulty Index Declared by the Language Model	Type	1096.5	0.19
	Answer correctness	139.5	< 0.001
	Division into clinical and other questions	1276	0.04
Human Difficulty Index	Type	1488	0.26
	Answer correctness	190	0.02
	Division into clinical and other questions	1156	0.24

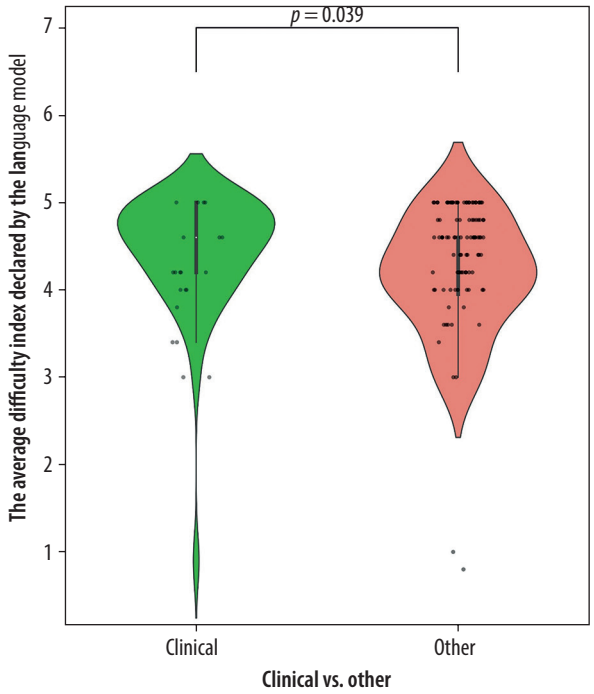

Figure 1. Comparison of Average Difficulty Index Declared by the Language Model between clinical and non-clinical questions

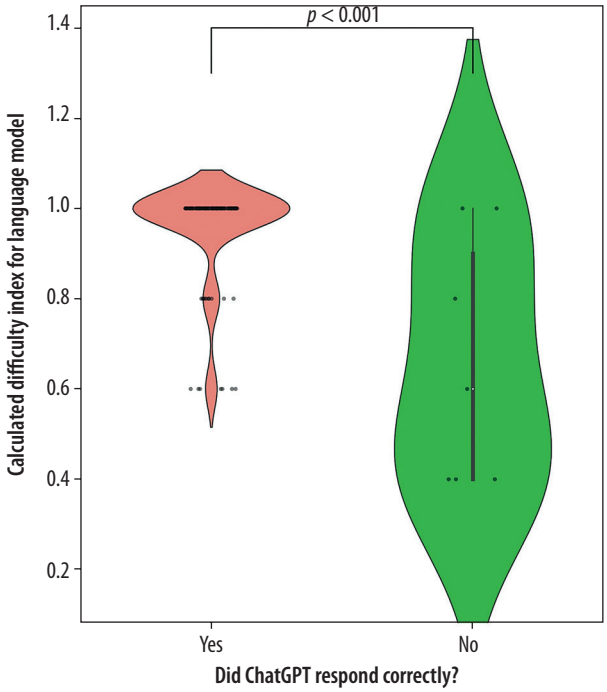

Figure 2. Comparison of Calculated Difficulty Index for the Language Model between correct and incorrect ChatGPT responses

Additionally, statistically significant differences were found between subpopulations based on the division of clinical and non-clinical questions ($p = 0.04$, $U = 1276$). The distributions of the mentioned subpopulations are presented in the figures. In Figure 1, which illustrates the differences in the distribution of the Difficulty Index Declared by the Language Model depending on whether a question belongs to the clinical or other category, several notable patterns can be observed. A key factor that must be emphasized is the considerably smaller number of questions classified as clinical. In addition to the observed statistical significance, the distributional differences between the two groups further suggest an underlying relationship: the language model tended to declare non-clinical questions as more difficult. Figure 2 presents the relationship between response correctness and the declared difficulty index, highlighting a key issue in evaluating the performance of new-generation language models. Despite the presence of statistical significance and the visual trend indicating that incorrectly answered questions tend to have higher declared difficulty, it is important to note the very limited representation of the incorrect answer group, which may limit the interpretability of this result. A similar limitation can be observed in Figures 3 and 4, which illustrate, respectively, the Difficulty Index Declared by the Language Model and its human-assigned counterpart. In both cases, the model tended to assign higher difficulty scores to questions it ultimately answered incorrectly.

Comparison with the old language model

A statistically significant difference was identified between the confidence coefficient of the new version of the lan-

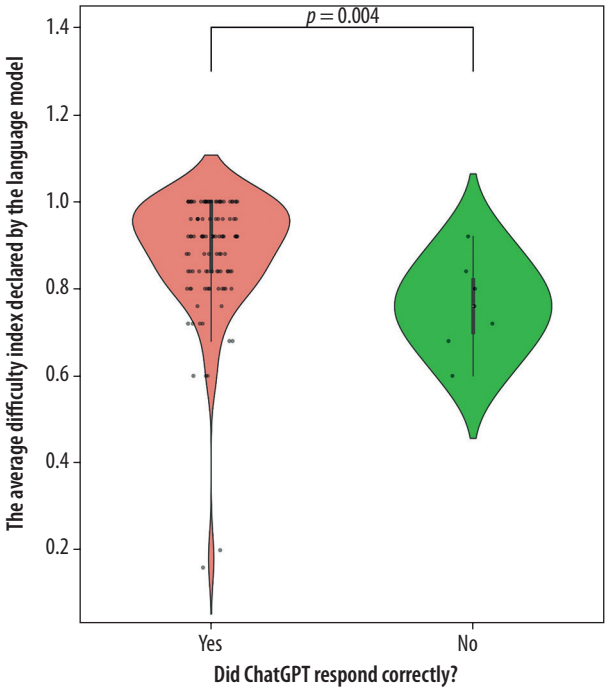

Figure 3. Comparison of Average Difficulty Index Assigned by the Language Model between correct and incorrect ChatGPT responses

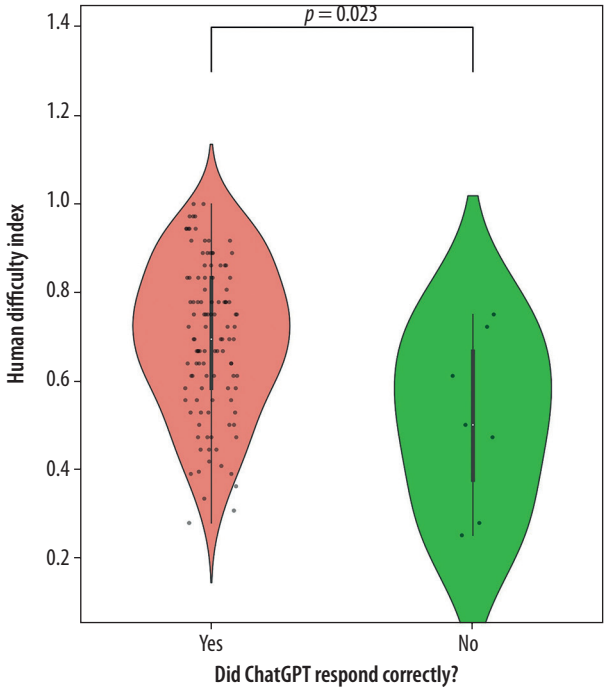

Figure 4. Comparison of Human Difficulty Index between correct and incorrect ChatGPT responses

guage model and the older version ($U = 4251$, $p < 0.001$, Old Model $\bar{x} = 4.00$, New Model $\bar{x} = 4.60$). The differences in the distributions of the confidence coefficient are illustrated in Figure 5, which illustrates differences in declared difficulty between the older and newer versions of the language model. The visualization demonstrates that the newer model generally rated the difficulty of the presented questions higher.

Discussion

The PES in radiology and medical imaging is the final exam for doctors specializing in this field. The exam consists of both an oral and a written part. To obtain the specialist title, candidates must pass the exam, which means achieving a score above 60% on the written part and receiving a positive evaluation on the oral part. Additionally, achieving a score above 75% on the written test exempts candidates from the oral part.

The detailed statistics on the pass rate for the PES exam in radiology and medical imaging for the years 2009-2018 show a pass rate of 94.86% [11].

The results obtained by the ChatGPT-o1-preview model demonstrate substantial progress in the capabilities of large language models (LLMs) in addressing highly specialized medical examinations. The model achieved an item-level accuracy of 93.3%, indicating strong performance across a broad set of exam questions.

While this accuracy figure highlights the model's potential, it is not directly comparable to the overall human pass rate, which reflects cumulative test performance and not question-level accuracy.

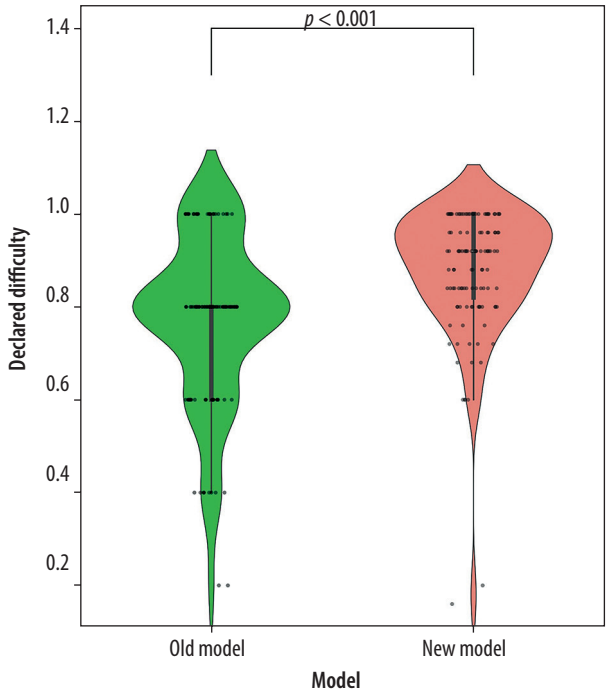

Figure 5. Comparison of declared difficulty between ChatGPT-3.5 (old) and ChatGPT-o1-preview (new) models

Notably, the examined AI outperformed its predecessor, ChatGPT-3.5, across all question types and subcategories, indicating a significant leap in performance and contextual understanding. In a study conducted by Kufel *et al.* [7], the effectiveness of the ChatGPT-3.5 language model was evaluated for solving the same PES exam in radiology and medical imaging. The older model scored 52%, not reaching the passing threshold of 60%. It per-

formed significantly worse compared to our result of 93.33%. In questions of the "clinical" subtype, it achieved 75%, compared to 95% achieved by the newer model. In the "comprehension and critical thinking" and "memory" categories, ChatGPT-3.5 achieved 55.56% and 44.83% correct answers, respectively. This means that in every case, it performed worse compared to ChatGPT-o1-preview, which achieved 93.33% and 96.55%, respectively.

Comparing the ChatGPT-o1-preview result in radiology (93.33%) to ChatGPT-3.5's results in PES exams in other fields of medicine such as dermatology and venereology [12] or allergology [13], the older model underperformed, achieving 49.58% and 52.54%, respectively.

In the new version of the model, ChatGPT-01-preview, the process of providing answers was designed to resemble human thought mechanisms. Through training, the model develops its thinking skills, tests various approaches, and learns to recognize and analyze its own mistakes [14]. This led to a significant improvement in passing the PES exam in radiology and medical imaging.

Despite these promising results, several limitations must be acknowledged. First, the evaluation was based solely on textual questions, omitting the visual component that is central to radiological practice. This limits the applicability of the findings to real-world clinical environments, where interpretation of imaging is crucial. Second, the study relied on a single exam session with 119 valid questions, which may not fully represent the variability and complexity of the broader medical curriculum. Moreover, although the model's declared confidence correlated moderately with both human and model difficulty indices, it remains uncertain whether this reflects genuine reasoning ability or statistical pattern recognition.

Future research should explore the model's performance using multimodal inputs that include medical imaging data, as well as a broader range of exam types and clinical specialties. Comparative studies involving other LLMs will also be valuable in assessing the current landscape of AI in medicine. Ultimately, while LLMs show potential as supportive tools in medical education and clinical decision-making, their role should be clearly defined and rigorously validated before any real-world implementation.

Conclusions

Based on the obtained results, ChatGPT-o1-preview demonstrated high accuracy in solving PES exam questions, correctly answering 93.3% of them. It achieved its highest accuracy in the "memory" subcategory (over 96% correct answers), although its performance was also consistently high across other subcategories. These findings contrast with those of previous studies involving ChatGPT-3.5, which demonstrated significantly lower performance.

This research highlights the potential of large language models as supportive tools in the medical field. To better assess their applicability and reliability, future investigations should involve broader and more diverse question sets across various medical subspecialties, as well as comparative benchmarking with other AI models. In particular, the integration of multimodal evaluation frameworks – including both text-based clinical scenarios and radiological images – would more accurately reflect real-world diagnostic tasks. Future studies could also explore the development of standardized head-to-head benchmarks to evaluate different model architectures (e.g., GPT-4, Med-PaLM, Gemini) under controlled conditions. Finally, incorporating response latency and user experience assessments may help evaluate the feasibility of real-time clinical deployment.

Disclosures

1. Institutional review board statement: Not applicable.
2. Assistance with the article: None.
3. Financial support and sponsorship: None.
4. Conflicts of interest: None.

References

1. Wu JH, Lin S, Moghimi S. Application of artificial intelligence in glaucoma care: an updated review. Taiwan J Ophthalmol 2024; 14: 340-351.
2. Zhao Z, Pi Y, Jiang L, Xiang Y, Wei J, Yang P, et al. Deep neural network based artificial intelligence assisted diagnosis of bone scintigraphy for cancer bone metastasis. Sci Rep 2020; 10: 17046. DOI: 10.1038/s41598-020-74135-4.
3. Chen R, Wang Q, Javanmardi A. A review of the application of machine learning for pipeline integrity predictive analysis in water distribution networks. Arch Computat Methods Eng 2025. DOI: 10.1007/s11831-025-10251-6.
4. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq [Preprint] 2023: rs.3.rs-2566942. DOI: 10.21203/rs.3.rs-2566942/v1.
5. Atkay S. Analysis of AI humanizer tools. In: Proceedings of the International Texas Congress on Advanced Scientific Research and Innovation. Houston, TX, 2024, pp. 40-48.
6. Nazir A, Wang Z. A comprehensive survey of ChatGPT: advancements, applications, prospects, and challenges. Meta-Radiology 2023; 1: 100022. DOI: https://doi.org/10.1016/j.metrad.2023.100022.
7. 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. DOI: 10.5114/pjr.2023.131215.

8. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: implications for clinical decision-making. PLOS Digit Health 2024; 3: e0000651. DOI: 10.1371/journal.pdig.0000651.
9. Naili YT, Mangkunegara IS, Purwono P, Baballe MA. Regulatory challenges in ai-based diagnostics: Legal implications of ai use in medical diagnostics. BIO Web Conf 2025; 152: 01034. DOI: https://doi.org/10.1051/bioconf/202515201034.
10. Anderson LW, Krathwohl DR (eds.). A Taxonomy for Learning, Teaching, and Assessing: a Revision of Bloom's Taxonomy of Educational Objectives. New York: Longman; 2001.
11. Centrum Egzaminów Medycznych [Internet]. [cited 2025 Apr 1]. Available from: https://www.cem.edu.pl/aktualnosci/spece/spece_stat.php.
12. Rojek M, Kufel J, Bielówka M, Mitręga A, Kaczyńska D, Czogalik Ł, et al. Exploring the performance of ChatGPT-3.5 in addressing dermatological queries: a research investigation into AI capabilities. Przegl Dermatol 2024; 111: 26-30.
13. Bielówka M, Kufel J, Rojek M, Mitręga A, Kaczyńska D, Czogalik Ł, et al. Evaluating ChatGPT-3.5 in allergology: performance in the Polish Specialist Examination. Alergologia Polska – Polish Journal of Allergology 2024; 11: 42-47.
14. Introducing OpenAI o1 [Internet]. [cited 2025 Apr 1]. Available from: https://openai.com/index/introducing-openai-o1-preview/.