Med
Innovations
Journal

Received: 22.12.2025
Revised: 02.01.2026
Accepted: 09.01.2026
Published: 10.03.2026

Original Article

Analyzing ChatGPT 3.5's Performance on the Polish Specialty Exam in Palliative Medicine: Strengths and Limitations

Piotr Dudek[1], Dominika Kaczyńska[1*], Natalia Denisiewicz[1], Adam Mitręga[1,2], Michał Bielówka[1,2], Łukasz Czogalik[1] and Marcin Rojek[1]

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
2. Department of Biophysics, Faculty of Medical Sciences in Zabrze, Medical University of Silesia in Katowice, Jordana 18, 40-043 Zabrze, Poland

*Corresponding author: dominikakaczynska01@gmail.com

Abstract

Background: Artificial intelligence (AI) has growing applications in medicine. This study examines whether the ChatGPT-3.5 language model can pass the Polish Specialty Examination (PES) in Palliative Medicine and evaluates its accuracy in various medical question categories. Additionally, the study highlights limitations AI faces in this medical context.
Material and methods: A total of 120 PES questions from spring 2023 were used, presented in single-choice format with four options. Each question was asked five times, and an answer was considered correct if ChatGPT responded correctly in at least 3 of 5 attempts. A confidence score was calculated based on the frequency of correct answers. Questions were also categorized by Bloom's Taxonomy to assess complexity.
Results: The minimum passing score for the PES was 60%, while ChatGPT achieved 53.33%, falling short of the required threshold. Performance varied by question type, with higher accuracy in clinical management (63.16%) and memory-based questions (54.32%) compared to critical thinking (51.28%). The language model showed a notable confidence rate in correct responses.
Conclusion: ChatGPT-3.5's performance on the PES in Palliative Medicine is shaped by its access to knowledge bases, but it lacks the practical experience essential to medical practice. Human expertise, which incorporates social, emotional, and cultural insights, remains critical for personalized patient care, underscoring the importance of human advantage in palliative care settings.

Keywords: ChatGPT-3.5; artificial intelligence; large language models; medical specialty examination; palliative medicine;

1. Introduction

ChatGPT is an AI-based language model developed by OpenAI, with development beginning in 2018. It is considered one of the state-of-the-art generative models [1]. Compared to earlier models, ChatGPT demonstrates significantly enhanced analytical and contextual understanding. Built on a deep neural network, it has been developed to improve its ability to generate responses based on extensive text datasets [2, 3].

Artificial intelligence (AI) is widely used in medicine worldwide [4, 5], supporting clinicians in diagnosis, treatment planning, and decision-making by facilitating analysis of large volumes of medical data. In palliative medicine, AI applications include risk assessment, outcome prediction, and personalized palliative care planning [6, 7]. Numerous studies are currently examining the effectiveness of ChatGPT-3.5 in medical analysis [8, 9, 10].

While ChatGPT-3.5 is widely used by millions for everyday tasks [11, 12], its effectiveness in specialized fields such as palliative medicine is still debated. The Polish Specialty Examinations (PES), including the one in palliative medicine, rigorously test the specialist knowledge and practical skills required for patient care. These exams cover a wide range of topics related to caring for terminally ill patients, including clinical, ethical, and communication aspects, and require knowledge in pharmacotherapy, interdisciplinary care, and patient-family communication.

Though ChatGPT-3.5 is effective in generating responses, it has limitations. Its knowledge is based on data up to 2023, and it lacks the capacity for independent clinical evaluation [13, 14]. In palliative care, where a personalized approach is essential, ChatGPT-3.5 works by averaging information in its training data. Additionally, GPT-3.5 has access to only a limited amount of specialized medical literature necessary for fully educating a palliative medicine physician [15, 16].

The aim of this study is to evaluate the effectiveness of ChatGPT-3.5 in analyzing data and answering questions from the PES in palliative medicine, in order to assess its medical expertise in this area.

2. Materials and Methods

2.1 Study Design and Questions

This prospective study was conducted between July 20, 2024, and July 31, 2024, to evaluate the ability of ChatGPT to answer questions from the Polish Specialty Examination in Palliative Medicine, using the Spring 2023 examination session as a reference set. The analyzed material consisted of 120 single-choice questions (four answer options) obtained from the official archived database of the Centre for Medical Examinations (CEM) in Lodz.

The primary objective of the study was not to assess ChatGPT's performance across the entire conceptual scope of palliative care, but rather to test predefined hypotheses concerning model behavior and response characteristics within a fixed and internally consistent textual corpus. Consequently, representativeness with respect to the full domain of palliative medicine was not a methodological requirement for hypothesis verification.

The questions were analyzed as individual, cognitively isolated instances. Each query was processed independently, without shared context or memory between questions, in accordance with the "one call – one world" principle governing large language model inference. Under this framework, randomization of question order was not methodologically relevant, as no carry-over effects or contextual dependencies between questions could occur.

No thematic stratification of questions was applied. This decision was intentional and methodologically justified, as the original proportion of question types reflects the fixed blueprint used in the construction of Polish specialty examinations and remains stable across examination versions. Introducing artificial stratification would have altered these proportions and could have compromised the internal validity of the hypothesis-driven evaluation.

To enhance interpretability, questions were categorized according to Bloom's modified taxonomy [17,18]. The assessment of answer correctness was performed by a single researcher. Each question was classified as either a memory-based or comprehension/critical-thinking question, and as either clinical or non-clinical. Additional subcategories included clinical management, pharmacology, treatment protocols, anesthesia, psychology, medical procedures, clinical guidelines, and symptom recognition.

2.2 Data Collection and Analysis

Prior to data collection, ChatGPT-3.5 was oriented to the formal structure of the examination, including the number of questions, the single-correct-answer format, and the fixed number of answer options. All interactions were conducted in Polish to ensure full consistency with the original examination conditions.

Each question was posed to ChatGPT five times in separate sessions. This repeated querying was not intended to generate multiple dependent observations for statistical testing, but rather to estimate the model's internal response stability and confidence. For each question, the most frequently selected answer across the five independent runs was treated as the model's final response.

An answer was classified as correct if it was selected in at least 3 out of 5 attempts. This threshold follows established standards in the relevant literature and reflects the minimal semantically meaningful cutoff that allows differentiation between random answer selection and a consistent model preference. Given the five-trial design, lower thresholds (e.g., 2/5) would not sufficiently exceed chance-level performance, while higher thresholds (e.g., 4/5) would impose disproportionately strict criteria without methodological justification.

The confidence level for each question was calculated as the proportion of correct answers across the five repetitions (range: 0–1). All interactions were logged and archived for transparency and reproducibility (Supplementary Material 1).

2.3 Statistical Analysis

For statistical analysis, each question contributed a single binary outcome (correct vs. incorrect) and an associated confidence score. Thus, despite the repeated querying during data collection, the final analytical dataset consisted of 120 independent observations corresponding to distinct examination questions.

ChatGPT's performance was compared with the official answer key and with aggregated statistics published by the Centre for Medical Examinations in Lodz. Analyses focused on overall accuracy, as well as differences in performance across question types, cognitive levels, and thematic categories.

Pearson's chi-square test was used to assess associations between categorical variables, including correctness and question classification. ANOVA was applied to compare mean confidence levels across predefined groups [19], while the Mann–Whitney U test was used to evaluate differences in confidence between correct and incorrect responses.

Because the statistical analyses were confirmatory and hypothesis-driven, and because each question contributed only one final outcome to the dataset, the analysis does not constitute repeated-measures or exploratory testing. Consequently, no corrections for multiple comparisons were applied. Statistical significance was defined as $p < 0.05$. All analyses were conducted using RStudio (RStudio, PBC, Boston, MA, USA).

3. Results

The tested language model achieved a score of 53.33% correct responses (Table 1). For statistical analysis, results were categorized by type, subtype, and classified into "clinical" and "other" categories. Effectiveness within each category was calculated and results were compared (Tables 2, 3, and 4). A significance level of $p < 0.05$ was used for all tests.

Table 1. Correct and incorrect answers.

Correct answer	Number of questions	%
Yes	64	53,33%
No	56	46,67%

Table 2. Comparison of correct and incorrect answers by type.

Category	Correct answer			
	Yes	%	No	%
comprehension and critical thinking questions	20	51,28%	19	48,72%
memory questions	44	54,32%	37	45,67%

Table 3. Comparison of correct and incorrect answers divided into 'clinical' and 'other'

Category	Correct answer			
	Yes	%	No	%
clinical	49	53,26%	43	46,74%
other	15	53,57%	13	46,43%

Table 4. Comparison of correct and incorrect answers by subtype.

Category	Correct answer			
	Yes	%	No	%
clinical proceedings	12	63,16%	7	36,84%
medication	12	44,44%	15	55,56%
treatment	8	44,44%	10	55,56%
anesthesia	0	0,00%	1	100%
psychology	14	87,50%	2	12,50%
medical procedures	3	42,86%	4	57,14%
medical guidelines	5	41,67%	7	58,33%
signs and symptoms	10	50,00%	10	50,00%

The analysis showed a statistically significant correlation between answer accuracy and the confidence factor ($p < 0.000001$) (Figure 1), as well as a borderline significance for the correlation between question difficulty and answer accuracy ($p = 0.053$) (Figure 2). Additionally, there was a statistically significant correlation between difficulty levels in the "clinical" and "other" categories ($p = 0.031$) (Figure 3).

No significant correlation was found between answer accuracy and question type ($p = 0.91$), subtype ($p = 0.10$), or the "clinical" category classification ($p = 1.00$). Similarly, there was no significant relationship between difficulty and confidence rates within these subgroups.

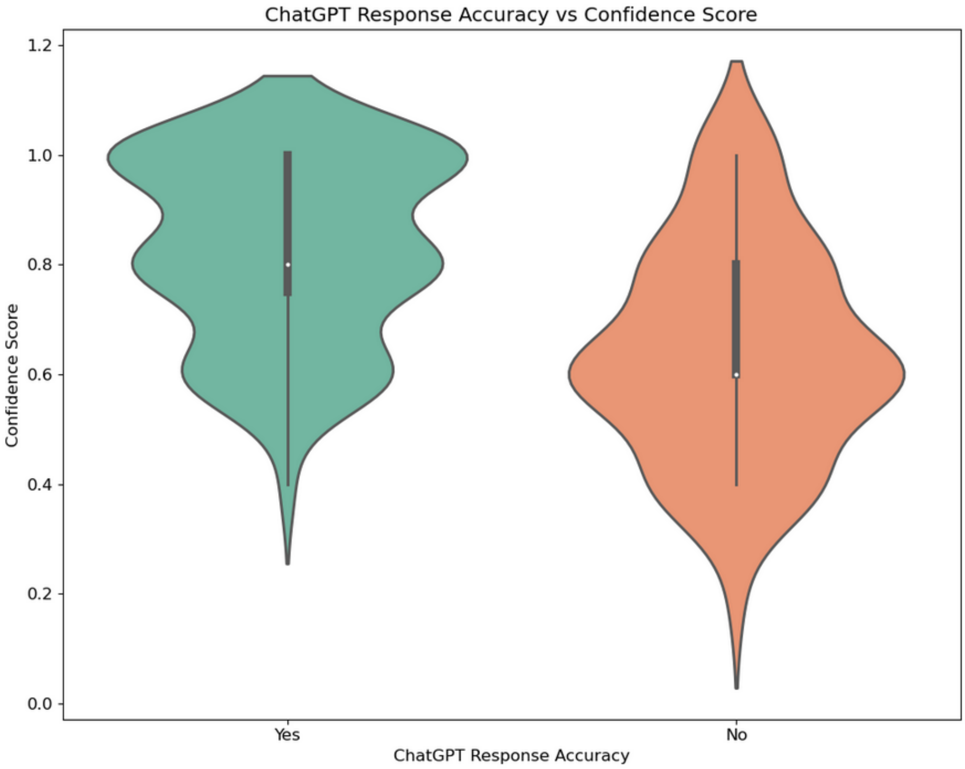

Figure 1. Comparison of the correctness of the answers of the tested language model versus the confidence factor.

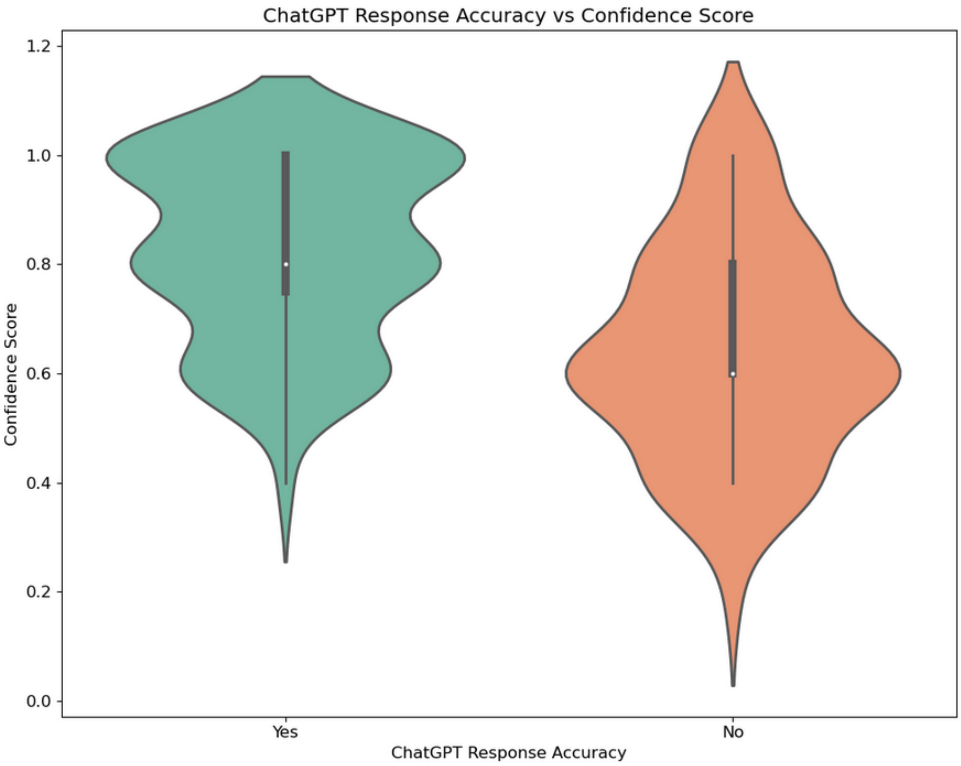

Figure 2. Comparison of the correctness of the answers of the tested language model versus the difficulty factor.

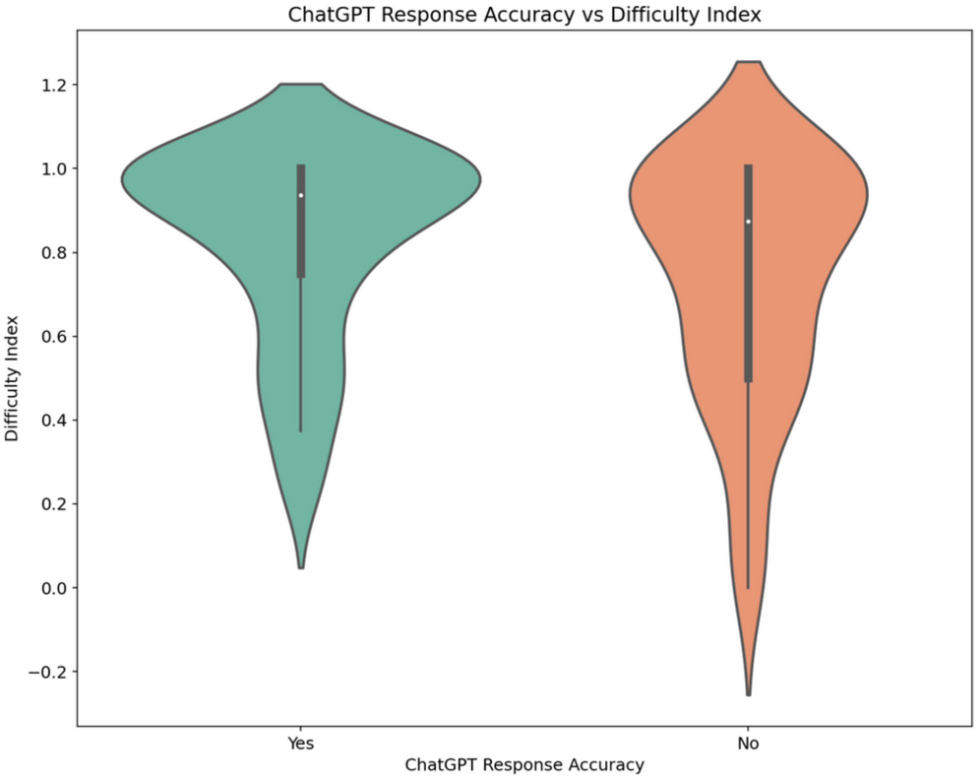

Figure 3. Comparison of the model's responses in the 'clinical' and 'other' categories versus the difficulty factor.

4. Discussion

In Poland, palliative medicine is recognized as a deficit specialty, with a shortage of specialists that impacts the availability and quality of care for patients with incurable diseases. Specialist training concludes with the State Specialist Examination, consisting of a test and an oral examination with problem-based questions. The exam includes 120 questions, and a score of at least 60% is required to pass, with a score of 75% in the test portion exempting candidates from the oral examination. From 2009–2018, the pass rate for the National Specialty Examination in palliative medicine was 97.58% [20].

In the present study, ChatGPT demonstrated an overall accuracy of 53.33%, correctly answering 64 of 120 questions, which remains below the defined passing threshold. The model exhibited comparable performance across the “clinical” (53.26%) and “other” (53.57%) question categories. Subtype analysis revealed the highest accuracy in psychology-related items (87.50%) and the lowest in anesthesiology (0.00%). The relatively strong performance in psychology may be explained by the predominantly theoretical character of psychological constructs and their extensive coverage in readily accessible academic sources.

A principal finding of this analysis was the strong and statistically significant association between response accuracy and the confidence factor ($p < 0.000001$), indicating that higher confidence ratings were closely aligned with correct responses and may therefore represent a meaningful proxy for answer reliability. In contrast, the relationship between question difficulty and accuracy did not achieve statistical significance, although a trend toward diminished performance with increasing difficulty was observed ($p = 0.053$).

To date, no studies have specifically evaluated the examination pass rates of artificial intelligence models in the field of palliative medicine, highlighting a notable gap in the existing literature. Given the growing interest in AI-assisted educational tools, further research is warranted to assess the potential role of such models in medical education and competency assessment. Moreover, the development of more specialized and domain-adapted algorithms may contribute to improving the quality of medical training and, consequently, standards of care in palliative medicine.

A similar study by Kufel et al. examined the PES pass rate in radiology and diagnostic imaging using ChatGPT, introducing a confidence factor from 1 to 5 to reflect the model's certainty in its answers. Their findings indicated higher confidence scores for correct answers, consistent with our results, which calculated confidence as the number of correct answers divided by 5. Despite only a single attempt per question in the radiology exam, ChatGPT achieved similar accuracy (52%) to our study with five attempts. Notably, ChatGPT performed better on clinical questions in radiology (75%) compared to palliative medicine (53.33%), possibly due to the smaller question pool in palliative care [21].

In a related study, Rojek et al. evaluated AI performance in the PES for dermatology and venereology, where ChatGPT scored 49.58%, below the 60% pass threshold. ChatGPT achieved a higher score (69.23%) on “medication” questions in dermatology compared to 44.44% in our study. Conversely, it performed better on “clinical proceedings” in palliative medicine (63.16%) than in dermatology (26.67%) [22].

According to an article by Bielówka et.al., the artificial intelligence model did not achieve a passing score (52.54%) on the PES exam in allergology. In this exam, ChatGPT scored higher in the ‘treatment’ questions obtaining 60% correct answers, compared to 44.44% in the PES in palliative medicine. It performed significantly better in the ‘symptoms and signs’ questions, obtaining 80% correct answers, compared to 50% in our examination. In contrast, he scored 53.33% in the ‘clinical procedures’ subtype, which means that he did worse than in the palliative medicine exam, where he obtained 63.16% correct answers [23].

Kufel et al. examined the performance of ChatGPT on the PES nuclear medicine exam, where it achieved a score of 56%, similar to the negative outcome in our study. In both studies, the model performed better on memory-based questions (59.57% in their study and 54.32% in ours) compared to comprehension and critical thinking questions (54.29% and 51.28%, respectively). This difference likely reflects the model’s design: AI is primarily equipped to retrieve and reproduce information based on its training data, lacking the capability to fully understand text or engage in abstract thinking [24].

In medicine, however, success depends not only on theoretical knowledge but also on practical experience. Large Language Models (LLMs) such as ChatGPT are inherently limited by their inability to accumulate clinical experience, which restricts their effectiveness in addressing practical, real-world medical problems.

4. Conclusion

The analysis of ChatGPT-3.5's performance on the Polish Specialty Examination (PES) in palliative medicine suggests that current AI models have limited effectiveness in passing highly complex medical exams. This limitation is largely due to the need for both practical and theoretical knowledge, which AI cannot fully emulate without clinical experience. Current AI constraints—such as restricted access to specialized medical literature and the inability to reliably distinguish credible from inaccurate information—significantly affect the exam outcomes for Large Language Models (LLMs). Additionally, the diverse types and varying difficulties of exam questions reduce the effectiveness of ChatGPT-3.5 in achieving correct responses.

In this study, ChatGPT-3.5 scored 53.33% on the PES, falling short of the 60% pass threshold. Continued testing with future versions of ChatGPT on structured question sets from the Centre for Medical Examinations could provide insights into its evolving accuracy and utility as a tool for assessing medical knowledge.

Despite advancements in AI and its expanding role in healthcare, the unique strengths of human clinicians in diagnosis and patient care remain essential. Physicians bring the ability to integrate complex social, cultural, and ethical considerations into patient care, elements that impact treatment effectiveness. Unlike AI, which relies on data patterns, experienced medical professionals draw on clinical practice and nuanced judgment, allowing for a more personalized and effective approach to each patient case.

Author Contributions: Conceptualization, P.D. and D.K.; Methodology, N.D.; Software, D.K.; Validation, P.D., D.K. and N.D.; Formal Analysis, M.B.; Investigation, A.M.; Resources, N.D.; Data Curation, D.K.; Writing – Original Draft Preparation, P.D.; Writing – Review & Editing, Ł.C.; Visualization, M.R.; Supervision, A.M.; Project Administration, P.D.;

Funding:
This study received no external funding.

Conflicts of Interest:
The authors declare no conflict of interest

References

1. Aydin, O., & Karaarslan, E. (2023). Is ChatGPT leading generative AI? What is beyond expectations? Academic Platform – Journal of Engineering and Science, 11(3). https://doi.org/10.21541/apjess.1293702
2. Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2(4), 100089. https://doi.org/10.1016/j.tbench.2023.100089
3. Bansal, G., Chamola, V., Hussain, A., et al. (2024). Transforming conversations with AI—A comprehensive study of ChatGPT. Cognitive Computation, 16, 2487–2510. https://doi.org/10.1007/s12559-023-10236-2
4. Ramesh, A., Kambhampati, C., Monson, J., & Drew, P. (2004). Artificial intelligence in medicine. Annals of The Royal College of Surgeons of England, 86(5), 334–338. https://doi.org/10.1308/147870804290
5. Busnatu, Ș., Niculescu, A. G., Bolocan, A., et al. (2022). Clinical applications of artificial intelligence—An updated overview. Journal of Clinical Medicine, 11(8), 2265. https://doi.org/10.3390/jcm11082265
6. Xie, W., & Butcher, R. (2023). Artificial intelligence decision support tools for end-of-life care planning conversations: CADTH Horizon Scan. Canadian Agency for Drugs and Technologies in Health.
7. Brender, T. D., Smith, A. K., & Block, B. L. (2024). Can artificial intelligence speak for incapacitated patients at the end of life? JAMA Internal Medicine, 184(9), 1005. https://doi.org/10.1001/jamainternmed.2024.2676
8. Jin, H. K., Lee, H. E., & Kim, E. (2024). Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: A systematic review and meta-analysis. BMC Medical Education, 24(1). https://doi.org/10.1186/s12909-024-05944-8
9. Farhat, F., Chaudry, B. M., Nadeem, M., Sohail, S. S., & Madsen, D. (2023). Evaluating AI models for the national pre-medical exam in India: A head-to-head analysis of ChatGPT-3.5, GPT-4, and Bard (Preprint). JMIR Medical Education. https://doi.org/10.2196/preprints.51523
10. Lim, Z. W., Pushpanathan, K., Yew, S. M. E., et al. (2023). Benchmarking large language models' performances for myopia care: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine, 95, 104770. https://doi.org/10.1016/j.ebiom.2023.104770
11. Ruby, D. (2023). ChatGPT statistics for 2023: Comprehensive facts and data. DemandSage.
12. Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., & Fahad, N. M. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. IEEE Access.
13. Kaneda, Y., Takahashi, R., Kaneda, U., et al. (2023). Assessing the performance of GPT-3.5 and GPT-4 on the 2023 Japanese nursing examination. Cureus, 15(8), e42924. https://doi.org/10.7759/cureus.42924

14. Liu, M., Okuhara, T., Chang, X., Shirabe, R., Nishiie, Y., Okada, H., & Kiuchi, T. (2024). Performance of ChatGPT across different versions in medical licensing examinations worldwide: Systematic review and meta-analysis. Journal of Medical Internet Research, 26, e60807. https://doi.org/10.2196/60807
15. Schmidl, B., Hütten, T., Pigorsch, S., et al. (2024). Assessing the use of the novel tool Claude 3 in comparison to ChatGPT 4.0 as an artificial intelligence tool in the diagnosis and therapy of primary head and neck cancer cases. European Archives of Oto-Rhino-Laryngology, 281(11), 6099–6109. https://doi.org/10.1007/s00405-024-08828-1
16. Schmidl, B., Hütten, T., Pigorsch, S., el al. (2024). Assessing the role of advanced artificial intelligence as a tool in multidisciplinary tumor board decision-making for recurrent/metastatic head and neck cancer cases. Frontiers in Oncology, 14, 1455413. https://doi.org/10.3389/fonc.2024.1455413
17. Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Addison Wesley Longman.
18. Bloom, B. S. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. Longmans, Green.
19. McHugh, M. L. (2011). Multiple comparison analysis testing in ANOVA. Biochemia Medica, 21(3), 203–209. https://doi.org/10.11613/bm.2011.029
20. Centrum Egzaminów Medycznych. (n.d.). Statystyki egzaminów specjalizacyjnych. https://www.cem.edu.pl/aktualnosci/spece/spece_stat.php
21. Kufel, J., Paszkiewicz, I., Bielówka, M., Bartnikowska, W., Janik, M., Stencel, M., et al. (n.d.). Will ChatGPT pass the Polish specialty exam in radiology and diagnostic imaging? Polish Journal of Radiology.
22. Rojek, M., Kufel, J., Bielówka, M., et al. (2024). Exploring the performance of ChatGPT-3.5 in addressing dermatological queries. Dermatology Review, 111(1), 26–30. https://doi.org/10.5114/dr.2024.140796
23. Bielówka, M., Kufel, J., Rojek, M., et al. (2024). Evaluating ChatGPT-3.5 in allergology: Performance in the Polish specialist examination. Alergologia Polska, 11, 42–47.
24. Kufel, J., Bielówka, M., Rojek, M., Mitręga, A., Czogalik, Ł., Kaczyńska, D., Kondoł, D., Palkij, K., & Mielcarska, S. (2024). Assessing ChatGPT's performance in national nuclear medicine specialty examination. Iranian Journal of Nuclear Medicine, 32(1), 60–65.