Acta Medica 2026; Early View: 139-145 DOI: 10.32552/actamedica.2026.1212

acta medica RESEARCH ARTICLE

The performance of ChatGPT and Google Bard in medical oncology board examination

Taha Koray Şahin[1], Murat Dinçer[2], Nuri Karadurmuş[3], Deniz Can Güven[1]

[1]Department of Medical Oncology, Hacettepe University, Ankara, Türkiye
[2]Department of Medical Oncology, Eskişehir Osmangazi University, Eskişehir, Türkiye
[3]Department of Medical Oncology, Gulhane School of Medicine, University of Health Sciences, Ankara, Türkiye

Abstract

Objective: Artificial intelligence (AI) is transforming healthcare, and large language models (LLMs) like ChatGPT and Google Bard have shown promise in providing medical information and decision support. The LLMs performed similarly or better than human participants in several board exams. However, their proficiency in complex clinical scenarios, like in oncology board exams, remains unclear. We aimed to assess the performance of three LLMs (ChatGPT 3.5, ChatGPT 4 and Google Bard) on the oncology board examination.

Materials and Methods: We utilized a question bank from the Turkish Society of Medical Oncology Board Exam encompassing 290 multiple-choice questions from 2021-2023. ChatGPT 3.5, ChatGPT 4, and Google Bard were asked to answer each question in both Turkish and English, providing explanations and confidence levels with their answers.

Results: The overall accuracy of LLMs was 59.3%, 42.8%, 36.2% for ChatGPT4, ChatGPT3.5, and Google Bard, respectively. The accuracy of ChatGPT 4 was significantly higher than that of ChatGPT 3.5 ($p<0.001$) and Google Bard ($p<0.001$), while the accuracy of ChatGPT3.5 was higher than that of Google Bard ($p<0.001$). Only the ChatGPT 4 was proficient in all three examination years (2021-2023). All LLMs performed better on translated questions than original Turkish ones. The LLMs were more accurate in general knowledge than case questions and were more confident in their answers for translated questions.

Conclusion: LLMs had moderate success in a medical oncology board exam, with only ChatGPT 4 demonstrating proficiency. The efficacy of LLMs in clinical decision-making requires further development, especially in native languages and complex case interpretations.

Keywords: large language models, ChatGPT, Google Bard, medical oncology, board, exam

Introduction

Artificial intelligence (AI) can transform every aspect of daily living, including healthcare [1]. The field of oncology was one of the most studied specialties in which AI emerged as a feasible way to improve patient care. In this regard, AI has already been used in cancer screening, molecular pathology, and big data analysis [2,3]. Furthermore, AI was promising as a decision-making tool after the earlier success of several models like IBM Watson [4]. However, the interest in this area was diminished after the limited clinical benefit of these models in complex clinical scenarios. Although not primarily developed for clinical decision and healthcare, the newly created large language models (LLMs) could potentially counter the limitations of previous

Corresponding author: Taha Koray Sahin • **Email:** takorsah@gmail.com

Received: November 12, 2025 **Accepted:** February 03, 2026 **Published online:** June 28, 2026

clinical decision support systems due to their training with larger datasets [5,6]. However, the potential of these LLMs in resolving complex clinical scenarios and generating true medical information has not been thoroughly investigated.

The most important prominent members of the recently developed LLMs are ChatGPT 3.5, ChatGPT 4, and Google Bard. In earlier studies, these LLMs were able to give accurate medical information in over the clinical scenarios [7,8]. Additionally, to evaluate their performance against human intelligence, these LLMs were tested in medical board exams. While the earlier results were very promising and demonstrated proficiency in board exams like United States Medical Licensing Examination (USMLE) [9] and radiology board exams [10], the LLMs' success was lower in specialties like ophthalmology [11] and neurology [12], in which more complex clinical scenarios and higher-order questions were more frequent. Additionally, the performance of individual LLMs, as well as the versions of the individual LLMs varied [11,13-15]. Despite this body of evidence in cardiology [16], radiology [17], and surgery boards [18], the performance of LLMs was not investigated in oncology board exams. Considering the huge burden of cancer, the generation of accurate information and the proficiency of LLMs in oncology is paramount. Therefore, we evaluated the performance of LLMs in an oncology board exam (Turkish Society of Medical Oncology Board Exam) and compared the performance of individual LLMs.

Methods

Sample questions

A question bank comprising the questions from the last three years (2021-2023) and including a total of 290 questions was used (Supplement). The questions for the board certification of Medical Oncologists from Turkey were created by a Turkish Society of Medical Oncology Proficiency Board Members, and the passing grades for each year were previously calculated according to the difficulty of this year's exam. The passing score was 53, 49, and 58 for 2021, 2022 and 2023, respectively. The full question set used in this study is openly available

The questions were multiple choice questions with five options and one correct answer with four distractors. The questions were retrieved from the website of the Turkish Society of Medical Oncology (members area). They were used with the approval of the Turkish Society of Medical Oncology Executive Board. The proficiency board members previously calculated the individual question difficulties, and these difficulty levels were used to compare the difficulty level annotated by the LLMs. Question topics and formats were categorized as case-based or general knowledge questions.

Data collection

The ChatGPT 3.5, ChatGPT 4, and Google Bard LLMs were used via the individual website interfaces. While the previous ChatGPT models were trained up to September 2021, this restriction was removed on September 27th, 2023, and both ChatGPT versions have access to live data via internet browsing. Similarly, the Google Bard could have live data from the internet. Therefore, all three LLMs were expected to provide answers in light of the most up-to-date data.

The researchers did not additionally pre-trained the LLMs before replying to the questions. The questions were asked in Turkish, and the LLMs were asked to answer in Turkish whenever possible. In another turn, the LLMs were also asked to translate the questions into English and give answers in English with explanations.

The following command was given to individual LLMs to gather data and all answer choices to this command and explanations were recorded.

"You are a medical oncologist and you are taking the oncology board exam. The board exam consists of multiple choice questions.

- Please write your answer so that there is only one correct answer among 5 options.
- Give an explanation
- Rate your confidence in your answer according to Likert performance with the following scales: 1 = do not trust [indicates that he/she does not know]; 2 = little confidence [i.e., maybe]; 3 = some interference; 4 = confidence [i.e., likely]; 5 = high confidence [stating the answer and explanation without doubt])
- Grading the question's difficulty level according to Likert performance with the following grades: 1=Very easy question, 2=Easy question, 3=Medium

question, 4=Difficult question, 5=Very difficult question.

Make sure you have these 4 in your output.

Answer each question with this format:

Answer:

Explanation:

Confidence level

Difficulty level:"

Statistical analyses and ethical considerations

The baseline question characteristics, and the accuracy of the LLMs were expressed with frequencies and percentages. The comparison of the accuracy of the individual LLMs and the comparison of the accuracy of the LLMs in original and translated versions of the questions were conducted with Chi-square tests. The comparison of median Likert scores for case questions vs general knowledge was conducted with Mann-Whitney U test. All statistical analyses were conducted with SPSS, version 25.0 (IBM Inc., Armonk, NY, USA), and a type 1 error level of 5% ($p<0.05$) was considered as the threshold limit for statistical significance.

Due to the use of a previously available question bank and no involvement of human subjects, the study is exempt from ethical approval. The study was conducted and reported according to STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) guidelines [19].

Results

A total of 290 questions were evaluated via three LLMs. Breast (n=31), lung (n=24), and colorectal cancers (n=29) were the most frequently assessed tumor types in the question bank. The distribution of case questions and general knowledge-based questions were even (49.7% vs. 50.3%). The treatment was the most frequently evaluated area (49.3%). The topic distribution for the questions is summarized in Table 1.

The overall accuracy of LLMs was 59.3%, 42.8%, 36.2% for ChatGPT4, ChatGPT3.5, and Google Bard, respectively. The accuracy of ChatGPT4 was significantly higher than that of ChatGPT3.5 ($p<0.001$) and Google Bard ($p<0.001$), while the accuracy of ChatGPT3.5 was higher than that of Google Bard ($p<0.001$). Only the ChatGPT4 was proficient in all three examination years (2021-2023). The three LLMs had moderate correlation in their accuracy for individual responses ($r=0.319$, $p<0.001$ for ChatGPT 3.5 vs. ChatGPT 4, $r=0.263$, $p<0.001$ for ChatGPT 3.5 vs. Google Bard and $r=0.288$, $p<0.001$ for ChatGPT 4 vs. Google Bard). The accuracy of the LLMs was higher for the translated questions compared to original language in all three LLMs (62.8 vs. 59.3%, $p<0.001$ for ChatGPT 4, 48.3 vs. 42.8%, $p<0.001$ for ChatGPT3.5and 43.1 vs. 36.2%, $p<0.001$ for Google Bard). While the accuracy was improved with the translated questions, the LLMS were not able to replicate the correct answers given in the original language for over 10% of the questions (Table 2). The LLMs were more accurate in general knowledge than case questions (Table 3). The ChatGPT 3.5 ($p=0.301$) and Google Bard ($p=0.378$) had similar accuracy across variable knowledge domains (treatment, diagnosis,

Table 1. General characteristics of the question bank

Feature	n (%)
Question Type	
General Information	146 (50.3)
Case Question	144 (49.7)
Question Category	
Diagnosis	39 (13.4)
Treatment	143 (49.3)
Prognosis	32 (11)
Toxicity	22 (7.6)
General Information	54 (18.6)
Question Topic	
Basic Science	55 (19)
Breast	31 (10.7)
Lung	24 (8.3)
GI	51 (17.6)
GU	28 (9.7)
GYN	22 (7.6)
HNC	15 (5.2)
Sarcoma	19 (6.6)
Hematology	16 (5.5)
Other	29 (10)

Table 2. Comparison of individual LLMs performance

		Translated Questions (English)			
		ChatGPT 3.5			
Original Questions (Turkish)		**Wrong**	**Correct**	**Total**	**p value**
ChatGPT 3.5	Wrong n, (%)	122 (42.10)	44 (15.20)	166 (57.20)	<0.001
	Correct n, (%)	28 (9.70)	96 (33.10)	124 (42.80)	
	Total	150 (51.70)	140 (48.30)	290 (100)	
		ChatGPT 4			
		Wrong	**Correct**	**Total**	**p value**
ChatGPT 4	Wrong, n (%)	78 (26.90)	40 (13.80)	118 (40.70)	<0.001
	Correct, n (%)	30 (10.30)	142 (49.00)	172 (59.30)	
	Total	108 (37.20)	182 (62.80)	290 (100)	
		Google Bard			
		Wrong	**Correct**	**Total**	**p value**
Google Bard	Wrong, n (%)	122 (42.10)	63 (21.70)	185 (63.8)	<0.001
	Correct, n (%)	43 (14.80)	62 (21.40)	105 (36.2)	
	Total	165 (56.90)	125 (43.10)	290 (100)	

prognosis, toxicity, general information), while the performance of ChatGPT 4 varied across knowledge domains (Table 3). The accuracy of the LLMs across tumor types demonstrated similar accuracy across most tumor types.

The median Likert score was 4 for ChatGPT4, ChatGPT3.5, and Google Bard (Figure 1). The LLMs' answers' certainty was higher in general knowledge than in case questions for Google Bard (p=0.001), while the certainty for general knowledge questions and case questions were similar in ChatGPT 3.5 (p=0.135) and ChatGPT 4 (p=0.111). The median difficulty was 3/5 for all three LLMs (Figure 2). The question difficulty was regarded as higher case questions than general knowledge questions for ChatGPT 3.5 (p<0.001), ChatGPT 4 (p<0.001) and Google Bard (p<0.001).

Discussion

In the present study, we observed that LLMs had moderate success in medical oncology proficiency. The performance of the individual LLMs significantly varied, with ChatGPT4 outperforming the two other LLMs. The LLMs underperformed in clinical cases, mirroring complex clinical scenarios in daily practice. Additionally, the performance of LLMs was lower in the native language compared to questions translated to English. To best our knowledge, our study is the first study evaluating the proficiency of LLMs in oncology proficiency.

The use of LLMs as an assistant to clinical practice garnered a lot interest in the last year, especially after the eye opening performance of the ChatGPT 3 in USMLE examination [20]. In the pivotal study which was broadcasted via even television news, the ChatGPT achieved around 60% accuracy and an acceptable reasoning for the responses (20). In a later work, Brin et al. compared the performance of ChatGPT 4 and ChatGPT 3.5. The ChatGPT4 outperformed ChatGPT 3.5 (correct response rate 90 vs. 62.5%) [21]. Additionally, the ChatGPT4 was consistent with repeated evaluations, while the ChatGPT 3.5 revised its' responses in 82.5% of the cases [21], further supporting the use of ChatGPT 4 in medical knowledge. However, later studies challenged these findings.

Regarding the accuracy of LLMs in cancer care, similar unequivocal results exist. The ChatGPT 4 was accurate and comprehensive, with over 85% of questions related to head and neck cancer knowledge [22]. The questions in the study were generated from the frequently asked questions of professional societies, support groups, and

Table 3. Accuracy of individual LLMs across question types

		ChatGPT 3.5		ChatGPT 4		Google Bard	
		Wrong	Correct	Wrong	Correct	Wrong	Correct
		Count (%)	Count (%)	Count (%)	Count (%)	Count (%)	Count (%)
Question Type	General Information	70 (47.9)	76 (52.1)	41 (28.1)	105 (71.9)	85 (58.2)	61 (41.8)
	Case Question	96 (66.7)	48 (33.3)	77 (53.5)	67 (46.5)	100 (69.4)	44 (30.6)
Knowledge Domain	Diagnosis	19 (48.7)	20 (51.3)	6 (15.4)	33 (84.6)	23 (59)	16 (41)
	Treatment	87 (60.8)	56 (39.2)	74 (51.7)	69 (48.3)	98 (68.5)	45 (31.5)
	Prognosis	19 (59.4)	13 (40.6)	13 (40.6)	19 (59.4)	21 (65.6)	11 (34.4)
	Toxicity	15 (68.2)	7 (31.8)	11 (50)	11 (50)	14 (63.6)	8 (36.4)
	General Information	26 (48.1)	28 (51.9)	14 (25.9)	40 (74.1)	29 (53.7)	25 (46.3)
Question Topic	Breast	22 (71)	9 (29)	12 (38.7)	19 (61.3)	17 (54.8)	14 (45.2)
	Lung	13 (54.2)	11 (45.8)	10 (41.7)	14 (58.3)	16 (66.7)	8 (33.3)
	GI	27 (52.9)	24 (47.1)	24 (47.1)	27 (52.9)	38 (74.5)	13 (25.5)
	GU	18 (64.3)	10 (35.7)	17 (60.7)	11 (39.3)	19 (67.9)	9 (32.1)
	HNC	10 (66.7)	5 (33.3)	8 (53.3)	7 (46.7)	11 (73.3)	4 (26.7)
	GYN	14 (60)	8 (36.4)	6 (27.3)	16 (72.7)	12 (54.5)	10 (45.5)
	Sarcoma	5 (26.3)	14 (73.7)	7 (36.8)	12 (63.2)	10 (52.6)	9 (47.4)
	Basic Science	30 (54.5)	25 (45.5)	18 (32.7)	37 (67.3)	35 (63.6)	20 (36.4)
	Hematology	10 (62.5)	6 (37.5)	6 (37.5)	10 (62.5)	10 (62.5)	6 (37.5)
	Other	17 (58.6)	12 (41.4)	10 (34.5)	19 (65.5)	17 (58.6)	12 (41.4)

social media, reflecting the potential of the LLMs in population-level healthcare education [22]. In contrast, ChatGPT provided accurate and comprehensive responses to only 53.1% of the questions regarding cervical cancer care in another study. [23]. It should be noted that this study used ChatGPT 3.5 [23], which underperformed in our analysis compared to ChatGPT 4. Whether the newer versions of the LLMs could improve healthcare information should be investigated.

There are several caveats regarding the use of LLMs in cancer care. First, the LLMs gather information from several sources with various publishing dates. In the earlier works with the ChatGPT in healthcare, the training time limitation up to September 2021 was an important problem. While this issue was resolved in September 2023, some of the recommendations by LLMs are still outdated. Although the LLMs gather information from several sources, the reasoning for selecting a particular reference is not fully delineated by artificial intelligence, as previously noted in IBM Watson studies [24]. This issue is particularly problematic in complex scenarios with more than one therapeutic option. Additionally, the LLMs could not understand and solve the complex clinical scenarios that are very common in clinical practice [25]. Moreover, LLMs may generate errors in treatment sequencing and timing in such cases and may assign high confidence to incorrect or potentially unsafe recommendations. Furthermore, in line with the general principles of the LLMs, the LLMs used very confident language even in the obviously wrong simple questions [13,26]. The reasoning for the problems requiring the best option. Lastly, less is known about the comparative efficacy of different LLMs. In several studies, the performance of newer versions of the LLMs was better compared to previous versions [11,13], and the ChatGPT outperformed Google Bard in a very recent study on radiology exam questions [17]. Further studies should separately evaluate the performance of different LLMs to delineate the best model for individual scenarios.

In conclusion, the LLMs had below acceptable performance in a national oncology board exam with only the ChatGPT 4.0 had proficiency in an oncology

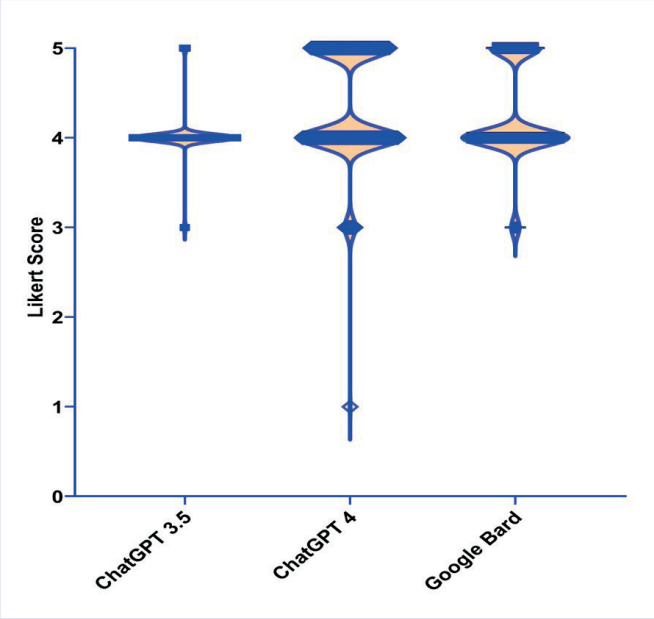

Figure 1. Confidence level rated on a 5-point Likert Scale (1 = very unconfident, 5 = very confident)

board examination. The LLMs' success in questions mirroring clinical practice was lower. Further research is needed to improve the proficiency of LLMs in cancer care-related information.

Author contributions

Conception and design: T.K.S., M.D., N.K., D.C.G.; Data acquisition: T.K.S., D.C.G.; Data analysis: T.K.S., D.C.G.; Data interpretation: T.K.S., D.C.G.; Drafting of the manuscript: T.K.S., M.D., N.K., D.C.G.; Critical revision of the manuscript: T.K.S., D.C.G.. All authors reviewed the results, approved the final version of the manuscript, and agreed to be accountable for all aspects of this study.

Ethical approval

Due to the use of a previously available question bank and no involvement of human subjects, the study is exempt from ethical approval.

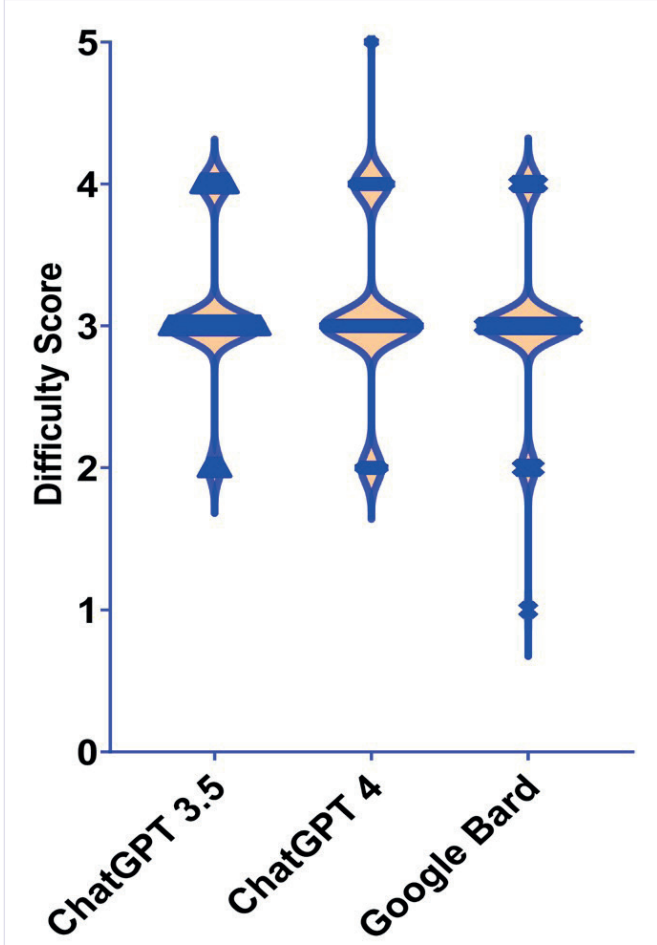

Figure 2. Difficulty level rated on a 5-point Likert Scale (1 = very easy, 5 = very difficult)

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflict of interest

The authors declare that this study was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

The authors declare that this study received no funding.

Generative AI statement

The authors declare that no generative AI or AI-assisted technologies were used in the writing or preparation of this study.

References

[1] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44-56. **[Crossref]**

[2] Bhinder B, Gilvary C, Madhukar NS, Elemento O. Artificial intelligence in cancer research and precision medicine. Cancer Discov 2021;11(4):900-15. **[Crossref]**

[3] Elemento O, Leslie C, Lundin J, Tourassi G. Artificial intelligence in cancer research, diagnosis and therapy. Nat Rev Cancer 2021;21(12):747-52. **[Crossref]**

[4] Malin JL. Envisioning Watson as a rapid-learning system for oncology. J Oncol Pract 2013;9(3):155-7. **[Crossref]**

[5] Benary M, Wang XD, Schmidt M, et al. Leveraging Large Language Models for decision support in Personalized Oncology. JAMA Netw Open 2023;6(11):e2343689. **[Crossref]**

[6] Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med (Lond) 2023;3(1):141. **[Crossref]**

[7] Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT - Reshaping medical education and clinical management. Pak J Med Sci 2023;39(2):605-7. **[Crossref]**

[8] Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Transactions on Benchmarks, Standards and Evaluations 2023;3(1):100105. **[Crossref]**

[9] Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of Large Language Models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312. **[Crossref]**

[10] Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: insights into current strengths and limitations. Radiology 2023;307(5):e230582. **[Crossref]**

[11] Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in Ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023;3(4):100324. **[Crossref]**

[12] Chen TC, Kaminski E, Koduri L, et al. Chat GPT as a Neuro-Score Calculator: analysis of a Large Language Model's performance on various neurological exam grading scales. World Neurosurg 2023;179:e342-e347. **[Crossref]**

[13] Schubert MC, Wick W, Venkataramani V. Performance of Large Language Models on a Neurology Board-Style Examination. JAMA Netw Open 2023;6(12):e2346721. **[Crossref]**

[14] Gunesli I, Aksun S, Fathelbab J, Yildiz BO. Comparative evaluation of ChatGPT-4, ChatGPT-3.5 and Google Gemini on PCOS assessment and management based on recommendations from the 2023 guideline. Endocrine 2025;88(1):315-22. **[Crossref]**

[15] Erul E, Aktekin Y, Danışman FB, et al. Perceptions, attitudes, and concerns on Artificial Intelligence applications in patients with cancer. Cancer Control 2025;32:10732748251343245. **[Crossref]**

[16] Skalidis I, Cagnina A, Luangphiphat W, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Heart J Digit Health 2023;4(3):279-81. **[Crossref]**

[17] Patil NS, Huang RS, van der Pol CB, Larocque N. Comparative performance of ChatGPT and Bard in a text-based radiology knowledge assessment. Can Assoc Radiol J 2024;75(2):344-50. **[Crossref]**

[18] Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023;93(6):1353-65. **[Crossref]**

[19] von Elm E, Altman DG, Egger M, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335(7624):806-8. **[Crossref]**

[20] Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198. **[Crossref]**

[21] Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023;13(1):16492. **[Crossref]**

[22] Kuşcu O, Pamuk AE, Sütay Süslü N, Hosal S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? Front Oncol 2023;13:1256459. **[Crossref]**

[23] Hermann CE, Patel JM, Boyd L, Growdon WB, Aviki E, Stasenko M. Let's chat about cervical cancer: assessing the accuracy of ChatGPT responses to cervical cancer questions. Gynecol Oncol 2023;179:164-8. **[Crossref]**

[24] Tupasela A, Di Nucci E. Concordance as evidence in the Watson for Oncology decision-support system. AI & SOCIETY 2020;35(4):811-8. **[Crossref]**

[25] Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. cureus. 2023;15(5):e39305. **[Crossref]**

[26] Rohrbach A, Hendricks LA, Burns K, Darrell T, Saenko K. Object hallucination in image captioning. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics; 2018, 4035-45. **[Crossref]**