International Journal of Multidisciplinary Sciences and Arts
E-ISSN :2962-1658
Volume 5, Number 2, April 2026
https://doi.org/10.47709/ijmdsa.v5i2.5343

Evaluating Advanced AI in Oncology Education and Clinical Knowledge Assessment

Yasar Ahmed[1*], Hatim Ibrahim[2],Simaa Hamid[3]
[1,2]Department of Medical Oncology, St. Vincent's University Hospital, Ireland
[3]Independent Researcher, Dublin, Ireland
[1*]drhammor@gmail.com

open access

***Corresponding Author**

Article History:
Submitted: 16-01-2025
Accepted: 20-12-2025
Published: 23-02-2026

Keywords:
Artificial Intelligence; Medical Oncology; Multimodal Large Language Model; ChatGPT.

ABSTRACT

The rapid advancement of artificial intelligence (AI) has introduced powerful tools like the Multimodal Large Language Model (MLLM) with the potential to revolutionize medical practices, including oncology. This study investigates the performance of two such MLLMs, GPT-4o and Gemini Advanced, in answering oncology examination questions from the American Society of Clinical Oncology Self-Evaluation Program (ASCO-SEP) Question Bank. We extracted 832 multiple-choice questions from this bank, covering various oncological tasks such as diagnosis, treatment recommendations, and basic science knowledge. Both models were presented with these questions, and their responses were evaluated against the official answer key. Gemini Advanced outperformed GPT-4o, achieving 74.84% accuracy compared to 60% for GPT-4o. Further analysis revealed that Gemini Advanced consistently outperformed GPT-4o across all task categories, particularly in making diagnoses, ordering and interpreting test results, and recommending treatment or patient care. Both models encountered the most difficulty with questions related to pathophysiology and basic science knowledge. These findings suggest that while both MLLMs demonstrate a significant understanding of oncological knowledge, there remains room for improvement, particularly in handling complex clinical scenarios and integrating basic science knowledge. This study contributes to the growing body of evidence assessing the capabilities and limitations of AI in medical oncology, highlighting its potential role in augmenting clinical practice and medical education.

INTRODUCTION

The rapid evolution of **Multimodal Large Language Model (MLLM)** and natural language processing has ushered in a new era of artificial intelligence (AI) with the potential to revolutionize various fields, including medicine(1). Among these MLLMs, ChatGPT, a generative pre-trained transformer developed by OpenAI, has garnered significant attention for its ability to generate comprehensive and human-like responses to a wide array of queries(2). Gemini Advanced is Google's next generation foundational large language model (LLM) designed to be multimodal, highly efficient at tool and API integrations, and built to enable future innovations, like memory and planning.(3)

These powerful AI tools utilize their neural networks to interact with users, providing remarkably accurate and contextually relevant answers. MLLMs find applications across diverse domains, ranging from social sciences and language translation to the complex field of medical sciences(1,3). Within the medical field, MLLMs have demonstrated potential in various areas, including diagnostics, treatment recommendations, clinical decision-making, and even scholarly writing. Although still in development, the potential of MLLMs to enhance medical practices and improve patient care is undeniable.

The influence of advanced AI language models reaches various medical specialties, notably oncology. While prior research has investigated their performance in other medical fields, their capabilities within oncology remained largely uncharted. Recent studies have examined ChatGPT's performance in medical licensing examinations across different countries, with encouraging outcomes. These studies underscore the model's capacity to accurately answer medical multiple-choice questions and excel in diverse question types, including basic science evaluation, diagnosis, and decision-making.

ChatGPT's sophisticated natural language processing capabilities empower it to generate precise responses and demonstrate exceptional performance in intricate decision-making scenarios

Beyond its clinical applications, this research also delves into advanced AI language models 's potential to enhance medical education for both patients and healthcare providers(4). By providing readily accessible and comprehensible information, ChatGPT and Gemini could empower patients to make more informed decisions about their care and facilitate effective communication between patients and their oncologists. Additionally, they both could

E-ISSN :2962-1658
Volume 5, Number 2, April 2026
https://doi.org/10.47709/ijmdsa.v5i2.5343

serve as a valuable educational resource for oncologists, offering just-in-time information and support for clinical decision-making.

Studies have investigated the ability of ChatGPT to correctly answer questions about medical education (5), dental medicine (5),family medicine (6),paediatric(7) cardiology (8), Gastroenterology (9)ophthalmology (10),respiratory medicine(11) and nephrology (12).Furthermore, a recent studies revealed that AI is capable of passing national licensing examinations **Worldwide**(13–15)

However, despite the growing body of evidence supporting the capabilities of LLMs in medicine, their specific application and efficacy within the field of oncology remain relatively unexplored. While prior research has investigated their performance in other medical fields, their capabilities within oncology, a complex and rapidly evolving discipline, have remained largely uncharted. This knowledge gap underscores the need for dedicated research to evaluate the performance of advanced AI language models specifically in the context of oncology.

This study aims to bridge this gap by specifically evaluating the performance of GPT-4o and Gemini Advanced on a comprehensive oncology examination, the ASCO-SEP Question Bank. By doing so, we seek to assess the current state of AI's ability to understand and apply complex oncological knowledge, identify areas where these models excel or falter, and shed light on their potential implications for clinical practice and medical education

The ASCO-SEP Question Bank is an online, self-assessment tool designed to help oncology professionals prepare for their board certification exams(16). Developed by the American Society of Clinical Oncology (ASCO) , it provides access to hundreds of practice questions and detailed explanations that cover the full spectrum of oncology topics The American Society of Clinical Oncology Self-Evaluation Program (ASCO-SEP) Question Bank currently boasts over 1300 questions, meticulously designed to mirror the style and complexity you'll encounter in the actual board certification exams. The question bank consists of single-best-answer multiple-choice questions addressing the many facets of cancer care, including diagnosis, treatment, and supportive care.

This study aims to evaluate the accuracy of advanced AI language models (specifically GPT-4 and Gemini) in answering oncology examination questions from the American Society of Clinical Oncology Self-Evaluation Program (ASCO-SEP), using it as a benchmark. Furthermore, the study seeks to compare the performance of GPT-4o and Gemini Advanced on the ASCO-SEP examination, identifying the strengths and weaknesses of each model. The ultimate goal is to ascertain whether the knowledge base of these AI language models aligns with the established standards expected of practicing oncologists

MATERIAL AND METHODS

Study Design and Data Sources

This comparative cross-sectional study was conducted between September 20, 2024, and September 30, 2024. No human participants were involved. The data were collected from two large language models (LLMs): GPT-4o and Gemini Advanced, developed by OpenAI and Google, respectively.

GPT-4o is recognized for its strengths in text generation, reasoning, and creative writing, while Gemini Advanced is designed for handling multiple multiple types of media or complex problem-solving across domains like math and coding

Data Collection and Preparation

The ASCO Question Bank, accessed via the ASCO-SEP 2024 Digital Edition, served as the benchmark for evaluating the performance of GPT-4o and Gemini Advanced in medical oncology knowledge.

A total of 832 single-best-answer multiple-choice questions (MCQs) were extracted from the ASCO-SEP Question Bank, encompassing the full spectrum of cancer care. The question bank covers seven primary categories in oncology: Tumor Types, Tumor Modalities, Supportive Care, Basic Science, Clinical Trials, Prevention & Screening, and Epidemiology & Statistics

The questions assess the following tasks performed by physicians:

- Making a diagnosis
- Ordering and interpreting test results
- Recommending treatment or other patient care
- Assessing risk, determining prognosis, and applying principles from epidemiologic studies
- Understanding the underlying pathophysiology of disease and basic science knowledge

Crucially, the information tested, including landmark publications referenced, was established before September 2023 (the ChatGPT knowledge cutoff date). The ASCO-SEP 2024 Digital Edition was used to ensure questions predate this cutoff, aligning the chatbot's training data with the questions' publication timeframe."

E-ISSN :2962-1658
Volume 5, Number 2, April 2026
https://doi.org/10.47709/ijmdsa.v5i2.5343

Model Interaction and Response Evaluation

Both GPT-4o and Gemini Advanced were presented with the 832 MCQs in their original format. Questions with visual data were excluded. Each model generated responses independently, without human intervention. Although no justification was explicitly requested, both models occasionally provided explanations

For each model, a new chatbot session was initiated. Each question was input individually, and the model's response was recorded. A single session per model was used, acknowledging potential influence from previous interactions within the session.

Responses were compared to the official ASCO-SEP answer key , and correct answers were those matching the key. Accuracy was calculated as the percentage of correct answers out of the total

Data Analysis

Accuracy scores were calculated for both models. Descriptive statistics summarized model performance. The chi-square test assessed differences in accuracy between GPT-4o and Gemini Advanced. The paired Wilcoxon signed-rank test evaluated differences in performance across task categories.

Data analysis was performed using SPSS Statistics for Windows, Version 21.0 (IBM Corp., Armonk, NY, USA). P-values < 0.05 were considered statistically significant."

Ethical Consideration

This study was exempt from requiring informed consent and formal ethical approval as the data used were considered sufficiently generic and did not contain any identifiable patient information

RESULTS

This study primarily aimed to evaluate and compare the performance of GPT-4 and Gemini Advanced in answering oncology-related examination questions, using the ASCO-SEP as a benchmark. The accuracy of each model was assessed, and their performance was further analyzed across different cancer types/disciplines and tasks commonly performed by oncologist.

Overall Performance

The performance of GPT-4o and Gemini Advanced was evaluated using 832 multiple-choice questions from the ASCO-SEP 2024 Digital Edition, spanning 15 cancer types or disciplines. Overall, Gemini Advanced outperformed GPT-4o, achieving 74.84% accuracy (623 correct answers) compared to GPT-4o's 60% accuracy (500 correct answers). This difference was statistically significant ($p = 0.025$).

Performance by Cancer Type/Discipline

Both models exhibited variable accuracy across different cancer types and disciplines. Gemini Advanced's accuracy ranged from 61.4% for 'Understanding underlying pathophysiology of disease & basic science knowledge' to 81.5% for 'Making a diagnosis'. GPT-4o's accuracy ranged from 33.0% for questions related to basic science and pathophysiology to 66.9% for 'Making a diagnosis'. 'Geriatric Oncology' and 'Diagnosis' saw the fewest correct responses for both models, suggesting these areas may be particularly challenging.

Performance by Task

Further analysis compared model performance on tasks commonly performed by oncologists. No statistically significant difference was found in overall performance across tasks ($p = 0.0625$). However, the most pronounced difference was observed in understanding underlying pathophysiology and basic science, where Gemini Advanced achieved 61.4% accuracy compared to GPT-4o's 33.0% ($p = 0.000$). Both models encountered the most difficulty with questions in this category.

Other Notable Findings

- GPT-4o performed no better than random guessing on questions about landmark studies ($p = 0.25$).
- Both models showed relatively lower accuracy in assessing risk, determining prognosis, and applying principles from epidemiologic studies.
- Gemini Advanced consistently outperformed GPT-4o across all task categories, particularly in making diagnoses, ordering and interpreting test results, and recommending treatment or patient care

International Journal of Multidisciplinary Sciences and Arts
E-ISSN :2962-1658
Volume 5, Number 2, April 2026
https://doi.org/10.47709/ijmdsa.v5i2.5343

Table 1. Comparison of GPT-4o and Gemini Advanced in terms of correct answers number by cancer type/discipline

CANCER TYPE OR DISCIPLINE	GPT-4o		Gemini Advanced		*p*-value
	(N)	(%)	(N)	(%)	
Palliative/Supportive Care/Survivorship (84)	47	56.0	60	71.4	0.054
Pharmacology & Anticancer Therapeutics (50)	36	72.0	38	76.0	0.819
Clinical Research Methodology & Ethics (25)	14	56.0	19	76.0	0.232
Genetics/Tumor Biology (17)	10	58.8	13	76.5	0.463
Hematologic Neoplasms (84)	49	58.3	64	76.2	0.021
Breast Cancer (101)	57	56.4	77	76.2	0.004
Gastrointestinal Cancer (109)	71	65.1	83	76.1	0.101
Thoracic Cancer (92)	52	56.5	70	76.1	0.008
Genitourinary Cancer (101)	61	60.4	77	76.2	0.023
Gynecologic Cancer (34)	22	64.7	26	76.5	0.424
Head, Neck, Thyroid, & Central Nervous System (34)	19	55.9	26	76.5	0.124
Skin Cancer, Sarcomas, and Unknown Primary Site (50)	28	56.0	37	74.0	0.093
Geriatric oncology (17)	12	70.6	10	58.8	0.719
Diagnosis (17)	9	52.9	12	0.6	0.480
Patient Management (17)	13	76.5	11	64.7	0.706
Total (832)	500	60%	623	74.84%	

Table 2. The performance of GPT-4o and Gemini advanced in Oncology Examination Questions

Task performed	GPT-4o Correct responses (%)	Gemini Advanced Correct responses (%)	*P* value
Making a diagnosis	66.9%	81.5%	0.018
Ordering and interpreting results of tests	56.5%	77.1%	0.002
Recommending treatment or other patient care	55.4%	69.6%	0.038
Assessing risk, determining prognosis, & applying principles from epidemiologic studies	48.9%	65.7%	0.016
Understanding underlying pathophysiology of disease & basic science knowledge applicable to patient care	53.0%	61.4%	0.000
Overall Accuracy	60%	74.84%	0.0625

DISCUSSION

The evolving landscape of artificial intelligence (AI) in oncology has sparked significant interest in its potential to support medical professionals. Large language models (LLMs) like ChatGPT and Gemini Advanced have shown varying degrees of success in medical fields, particularly oncology. Studies on LLMs in clinical medicine consistently report improved performance from ChatGPT-3.5 to ChatGPT-4o, evident in its higher accuracy rates in nephrology, oncology, and neurology examinations(12,17,18) This progression highlights the rapid advancements in AI and its growing capacity to handle complex medical queries.

In our study, we evaluated the performance of GPT-4 and Gemini Advanced on oncology exam questions and decision-making scenarios. Previous research has shown that ChatGPT-3.5 and 4.0 perform well on structured multiple-choice questions, as demonstrated in both the European Society for Medical Oncology (ESMO) examination and the ASCO-SEP(17,19,20) These findings align with our results, where both AI models answered a significant proportion of questions correctly, with GPT-4 achieving approximately 72% accuracy(17).

Our study, along with others(17–22) , collectively explores the capabilities and limitations of LLMs in medical oncology. There's a clear consensus that LLMs, especially advanced versions like GPT-4 and Gemini Advanced, demonstrate a remarkable ability to encode and apply oncology knowledge, achieving high accuracy on various oncology examinations. This aligns with our findings, where Gemini Advanced outperformed GPT-4, and both exceeded the average human candidate scores on the ASCO-SEP examination (21,22).

A consistent finding across studies is the relative weakness of LLMs in handling patient management questions, particularly in complex clinical scenarios requiring nuanced decision-making and integration of multiple factors. This underscores the continued importance of human clinical judgment and the need for further development in this area. Our observation that both GPT-4 and Gemini Advanced demonstrated the most difficulty with questions related to pathophysiology and basic science knowledge aligns with Longwell et al.(17), who noted that incorrect answers were often linked to information retrieval errors, especially with recent publications.

E-ISSN :2962-1658
Volume 5, Number 2, April 2026
https://doi.org/10.47709/ijmdsa.v5i2.5343

Consistent with our findings, ChatGPT struggles particularly with clinical decision-making aspects that require contextual understanding and integration of recent clinical guidelines and studies(22). For example, in questions demanding nuanced reasoning—such as assessing treatment options or tailoring care to individual patients—AI models often underperform compared to human oncologists (21). Additionally, our analysis revealed that ChatGPT's performance in handling dynamic information, like new treatment protocols or landmark trials, lags behind human expertise (17).

Moreover, while ChatGPT excels at answering basic factual questions, its lower performance in areas like treatment planning and patient management highlights the ongoing need for human oversight. Recent studies support this, suggesting that incorrect AI outputs can lead to clinical risks if not carefully scrutinized(21). Furthermore, AI's current inability to manage real-time patient data or incorporate emotional and ethical aspects of care underscores its technological limitations(7,20).

However, discrepancies arise when comparing the performance of different LLM versions across studies. While our study found Gemini Advanced outperformed GPT-4 on the ASCO-SEP questions, Number of studies (17,20,22) reported superior performance for a proprietary LLM 2 (likely GPT-4) compared to an earlier version (likely GPT-3.5) on a combination of ASCO, ESMO, and original questions. These differences could be attributed to variations in the question sets, the specific versions of the LLMs used, and the evaluation methodologies employed.

The limitations highlighted in these studies emphasize the dynamic nature of medical oncology. The field is constantly evolving with new research findings and treatment guidelines, necessitating continuous updates and training of LLMs to maintain their accuracy and relevance. This is further supported by the observation in Longwell et al. (17)that incorrect answers were more common when questions required knowledge of recent publications.

The study reveals a notable performance difference between GPT-4o and Gemini Advanced across various oncological tasks. Gemini Advanced consistently outperforms GPT-4o, particularly in making diagnoses, ordering and interpreting test results, and recommending treatment or patient care. This discrepancy likely stems from several factors including differences in model architecture and training data(3), and task-specific strengths(23), complex problem-solving (24) and Efficiency in Tool and API Integrations.

Despite these limitations, the potential applications of LLMs in oncology are vast. Beyond examination performance, they could assist in drafting patient communication, generating clinical reports, and supporting decision-making(17–22,25). However, the potential risks associated with incorrect or outdated information, as evidenced by the high likelihood of harm associated with incorrect answers in Longwell et al.(17), underscore the need for careful implementation and ongoing evaluation.

These findings emphasize the role of AI as an adjunct rather than a replacement in medical decision-making. As models continue to evolve, there is potential for significant improvements, particularly in addressing the challenges of patient management and integrating new clinical research in real-time. The use of AI in oncology will likely grow, but human judgment remains irreplaceable, especially in high-stakes decision-making where nuanced, patient-centered care is critical.

The Medical Oncology exam from the American Board of Internal Medicine (ABIM) does not have a set pass mark or percentage. Instead, it uses a standardized score scale ranging from 200 to 800, with the pass mark typically around 60%, although it can vary depending on the overall performance of the candidates(26). Based on these results, both GPT-4o and Gemini Advanced achieve scores at or near the historical pass mark.

The Medical Oncology Certification Exam is a challenging exam, demanding not only a profound understanding of medical oncology but also the ability to apply this knowledge to intricate clinical scenarios. Candidates typically undergo 10-13 years of rigorous training and dedicated preparation to attain the required level of expertise. Remarkably, our results demonstrate that GPT-4 and Gemini Advanced, even in their beta versions, were able to achieve scores within the range of the exam's historical pass mark. The exam's questions consist of text vignettes with nuanced scenarios that require deductive reasoning (27). Successfully answering these questions necessitates a rational approach and a significant amount of knowledge, which may explain the success of these LLMs with this particular task.

Our study has some limitations. The analysis was based solely on whether the models selected the correct answer, without considering factors like question complexity or length. Future studies could incorporate more nuanced evaluation metrics to gain deeper insights into the models' strengths and weaknesses.

CONCLUSION

This study suggest that while both LLMs demonstrate a significant understanding of oncological knowledge, there remains room for improvement, particularly in handling complex clinical scenarios and integrating basic science knowledge. The discrepancies in performance between GPT-4o and Gemini Advanced highlight the influence of model architecture, training data, and task-specific strengths on the accuracy and capabilities of LLMs in medical oncology.

Despite their limitations, LLMs like GPT-4o and Gemini Advanced hold considerable potential for augmenting clinical practice and medical education in oncology. Future applications may include drafting patient communication, generating clinical reports, and supporting decision-making.

E-ISSN :2962-1658
Volume 5, Number 2, April 2026
https://doi.org/10.47709/ijmdsa.v5i2.5343

However, it is crucial to highlight AI's strengths without overlooking its limitations: We have shown that it is able to effectively process medical information and provide appropriate answers to questions, however, it is currently not a substitute for critical thinking, innovation, and creativity, some of the key attributes that doctors are expected to showcase.

As AI continues to advance, it is essential to conduct further research to fully understand the capabilities and limitations of LLMs in oncology and to establish clear guidelines for their responsible implementation in clinical practice and medical education

Author Contributions:

- **YA:** Conceptualization, Writing - original draft, Methodology, Formal analysis
- **HI:** Validation, Writing - review & editing
- **SH:** Methodology, Validation, Writing - review & editing, Software

Declaration of competing interest

Authors declare no conflict of interest.

Data availability

Data will be made available on request

REFRENCES

1. Raiaan MAK, Mukta MdSH, Fatema K, Fahad NM, Sakib S, Mim MMJ, et al. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. IEEE Access. 2024;12:26839–74.
2. Butte AJ. Artificial Intelligence-From Starting Pilots to Scalable Privilege. JAMA Oncol. 2023 Oct 1;9(10):1341–2.
3. Rane N, Choudhary S, Rane J. Gemini Versus ChatGPT: Applications, Performance, Architecture, Capabilities, and Implementation [Internet]. Rochester, NY; 2024 [cited 2024 Sep 26]. Available from: https://papers.ssrn.com/abstract=4723687
4. Ahmed Y. Utilization of ChatGPT in Medical Education: Applications and Implications for Curriculum Enhancement. Acta Inform Medica. 2023;31(4):300–5.
5. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent Off Publ Am Acad Esthet Dent Al. 2023 Oct;35(7):1098–102.
6. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc JCMA. 2023 Aug 1;86(8):762–6.
7. Le M, Davis M. ChatGPT Yields a Passing Score on a Pediatric Board Preparatory Exam but Raises Red Flags. Glob Pediatr Health. 2024 Mar 24;11:2333794X241240327.
8. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Heart J Digit Health. 2023 May;4(3):279–81.
9. Suchman K, Garg S, Trindade AJ. Chat Generative Pretrained Transformer Fails the Multiple-Choice American College of Gastroenterology Self-Assessment Test. Am J Gastroenterol. 2023 Dec 1;118(12):2280–2.
10. Mihalache A, Popovic MM, Muni RH. Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. JAMA Ophthalmol. 2023 Jun 1;141(6):589–97.
11. Luo H, Yan J, Zhou X. Evaluating artificial intelligence responses to respiratory medicine questions. Respirology. 2024;29(7):640–3.
12. Nicikowski J, Szczepański M, Miedziaszczyk M, Kudliński B. The potential of ChatGPT in medicine: an example analysis of nephrology specialty exams in Poland. Clin Kidney J. 2024 Jul 1;17(8):sfae193.
13. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023 Feb 9;2(2):e0000198.
14. Jin HK, Lee HE, Kim E. 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 Med Educ. 2024 Sep 16;24(1):1013.
15. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. J Med Internet Res. 2024 Jul 25;26(1):e60807.
16. ASCO-SEP for Training Programs - Informational Page | ASCO Education [Internet]. [cited 2024 Sep 27]. Available from: https://education.asco.org/product-details/ascoSEPtrainingprograms

E-ISSN :2962-1658
Volume 5, Number 2, April 2026
https://doi.org/10.47709/ijmdsa.v5i2.5343

17. Longwell JB, Hirsch I, Binder F, Gonzalez Conchas GA, Mau D, Jang R, et al. Performance of Large Language Models on Medical Oncology Examination Questions. JAMA Netw Open. 2024 Jun 18;7(6):e2417641.
18. Chen S, Kann BH, Foote MB, Aerts HJWL, Savova GK, Mak RH, et al. Use of Artificial Intelligence Chatbots for Cancer Treatment Information. JAMA Oncol. 2023 Oct;9(10):1459–62.
19. Barbour AB, Barbour TA. A Radiation Oncology Board Exam of ChatGPT. Cureus. 15(9):e44541.
20. Chow R, Hasan S, Zheng A, Gao C, Valdes G, Yu F, et al. The Accuracy of Artificial Intelligence ChatGPT in Oncology Examination Questions. J Am Coll Radiol [Internet]. 2024 Aug 2 [cited 2024 Sep 26];0(0). Available from: https://www.jacr.org/article/S1546-1440(24)00675-6/fulltext
21. Odabashian R, Bastin D, Jones G, Manzoor M, Tangestaniapour S, Assad M, et al. Assessment of ChatGPT-3.5's Knowledge in Oncology: Comparative Study with ASCO-SEP Benchmarks. JMIR AI. 2024 Jan 12;3(1):e50442.
22. Filippov E, Lizogub O, Kovalenko I, Golubykh K, Khunkhun R. Performance of ChatGPT on the European Society for Medical Oncology (ESMO) Exam: Comparative Analysis (Preprint). JMIR Prepr. 2024 Jan 20;
23. Hochmair HH, Juhász L, Kemp T. Correctness Comparison of ChatGPT-4, Gemini, Claude-3, and Copilot for Spatial Tasks. Trans GIS [Internet]. [cited 2024 Oct 1];n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.13233
24. Rane N, Choudhary S, Rane J. Gemini Versus ChatGPT: Applications, Performance, Architecture, Capabilities, and Implementation [Internet]. Rochester, NY; 2024 [cited 2024 Oct 1]. Available from: https://papers.ssrn.com/abstract=4723687
25. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. JAMA Oncol. 2023 Oct 1;9(10):1437–40.
26. How Hard Is the ABIM Certification Exam? ABIM Exam Explained. [Internet]. 2022 [cited 2024 Sep 28]. Available from: https://challengercme.com/blog/articles/2022/06/how-hard-is-the-abim-internal-medicine-board-exam
27. AHMED Y, TAHA MH, KHAYAL S. Integrating Research and Teaching in Medical Education: Challenges, Strategies, and Implications for Healthcare. J Adv Med Educ Prof. 2024 Jan 1;12(1):1–7.