Large language models (LLMs) in radiology exams for medical students: Performance and consequences

Die Leistungen von große Sprachmodelle (LLMs) in radiologischen Studentenprüfungen: Leistung und Auswirkungen

Authors

Jennifer Gotta[1], Quang Anh Le Hong[1], Vitali Koch[1], Leon D. Gruenewald[1], Tobias Geyer[2], Simon S. Martin[1], Jan-Erik Scholtz[1], Christian Booz[1], Daniel Pinto Dos Santos[1], Scherwin Mahmoudi[1], Katrin Eichler[1], Tatjana Gruber-Rouh[1], Renate Hammerstingl[1], Teodora Biciusca[1], Lisa Joy Juergens[1], Elena Höhne[1], Christoph Mader[1], Thomas J. Vogl[1], Philipp Reschke[1]

Affiliations

1 Department of Diagnostic and Interventional Radiology, Goethe University Frankfurt, Frankfurt am Main, Germany

2 Institute of Diagnostic and Interventional Radiology, Pediatric Radiology and Neuroradiology, Rostock University Medical Center, Rostock, Germany

Keywords

AI, medical, education

received 27.6.2024

accepted after revision 2.10.2024

published online 2024

Bibliography

Fortschr Röntgenstr

DOI 10.1055/a-2437-2067

ISSN 1438-9029

Georg Thieme Verlag KG, Rüdigerstraße 14,
70469 Stuttgart, Germany

Correspondence

Dr. Philipp Reschke
Department of Diagnostic and Interventional Radiology,
Goethe University Frankfurt, Theodor-Stern-Kai 7,
60696 Frankfurt am Main, Germany
Philipp.reschke@outlook.de

Supplementary Material is available at https://doi.org/10.1055/a-2437-2067.

ABSTRACT

Purpose The evolving field of medical education is being shaped by technological advancements, including the integration of Large Language Models (LLMs) like ChatGPT. These models could be invaluable resources for medical students, by simplifying complex concepts and enhancing interactive learning by providing personalized support. LLMs have shown impressive performance in professional examinations, even without specific domain training, making them particularly relevant in the medical field. This study aims to assess the performance of LLMs in radiology examinations for medical students, thereby shedding light on their current capabilities and implications.

Materials and Methods This study was conducted using 151 multiple-choice questions, which were used for radiology exams for medical students. The questions were categorized by type and topic and were then processed using OpenAI's GPT-3.5 and GPT- 4 via their API, or manually put into Perplexity AI with GPT-3.5 and Bing. LLM performance was evaluated overall, by question type and by topic.

Results GPT-3.5 achieved a 67.6 % overall accuracy on all 151 questions, while GPT-4 outperformed it significantly with an 88.1 % overall accuracy ($p < 0.001$). GPT-4 demonstrated superior performance in both lower-order and higher-order questions compared to GPT-3.5, Perplexity AI, and medical students, with GPT-4 particularly excelling in higher-order questions. All GPT models would have successfully passed the radiology exam for medical students at our university.

Conclusion In conclusion, our study highlights the potential of LLMs as accessible knowledge resources for medical students. GPT-4 performed well on lower-order as well as higher-order questions, making ChatGPT-4 a potentially very useful tool for reviewing radiology exam questions. Radiologists should be aware of ChatGPT's limitations, including its tendency to confidently provide incorrect responses.

Key Points

- ChatGPT demonstrated remarkable performance, achieving a passing grade on a radiology examination for medical students that did not include image questions.
- GPT-4 exhibits significantly improved performance compared to its predecessors GPT-3.5 and Perplexity AI with 88 % of questions answered correctly.
- Radiologists as well as medical students should be aware of ChatGPT's limitations, including its tendency to confidently provide incorrect responses.

Citation Format

- Gotta J, Le Hong QA, Koch V et al. Large language models (LLMs) in radiology exams for medical students: Performance and consequences. Fortschr Röntgenstr 2024; DOI 10.1055/a-2437-2067

ZUSAMMENFASSUNG

Ziel Das sich entwickelnde Feld der medizinischen Ausbildung wird durch technologische Fortschritte geprägt, einschließlich der Integration von Large Language Models (LLMs) wie ChatGPT. Diese Modelle könnten für Medizinstudenten unschätzbare Ressourcen sein, indem sie komplexe Konzepte vereinfachen und das interaktive Lernen durch persönliche Unterstützung verbessern. Diese Studie zielt darauf ab, die Leistung von LLMs in radiologischen Prüfungen für Medizinstudenten zu bewerten und Einblicke in ihre aktuellen Fähigkeiten und Auswirkungen zu geben.

Materialien und Methoden Diese Studie wurde mit 151 Multiple-Choice-Fragen durchgeführt, die für radiologische Prüfungen von Medizinstudenten verwendet wurden. Die Fragen wurden nach Typ und Thema kategorisiert und dann mithilfe von OpenAI's GPT-3.5 und GPT-4 über deren API verarbeitet oder manuell in Perplexity AI mit GPT-3.5 und Bing eingegeben. Die Leistung der LLMs wurde insgesamt nach Fragetyp und nach Thema bewertet.

Ergebnisse GPT-3.5 erreichte eine Gesamtgenauigkeit von 67,6% bei allen 151 Fragen, während GPT-4 mit einer Gesamtgenauigkeit von 88,1% signifikant besser abschnitt ($p < 0{,}001$). GPT-4 zeigte sowohl bei einfachen als auch bei komplexeren Fragen eine überlegene Leistung im Vergleich zu GPT-3.5, Perplexity AI und Medizinstudenten. Besonders hervorzuheben ist, dass GPT-4 bei den komplexeren Fragen deutlich besser abschnitt. Alle GPT-Modelle hätten die radiologische Prüfung für Medizinstudenten an unserer Universität erfolgreich bestanden.

Schlussfolgerung Zusammenfassend hebt unsere Studie das Potenzial von LLMs als zugängliche Wissensressourcen für Medizinstudenten hervor. GPT-4 schnitt gut bei Fragen niedriger und höherer Ordnung ab, was ChatGPT-4 zu einem potenziell sehr nützlichen Werkzeug für die Überprüfung von radiologischen Prüfungsfragen macht. Radiologen sollten sich der Grenzen von ChatGPT bewusst sein, einschließlich seiner Tendenz, selbstbewusst falsche Antworten zu geben.

Kernaussagen

- ChatGPT zeigte eine bemerkenswerte Leistung und alle Modelle bestanden die Radiologie-Prüfung für Medizinstudenten ohne Bildfragen.
- GPT-4 erzielte mit einer Gesamtgenauigkeit von 88% die höchste Punktzahl bei den Radiologie-Prüfungsfragen und übertraf damit GPT-3.5, Perplexity AI und Medizinstudenten deutlich.
- Radiologen sowie Medizinstudenten sollten sich der Einschränkungen von ChatGPT bewusst sein, einschließlich seiner Tendenz, selbstsicher falsche Antworten zu geben.

Introduction

The field of medical education is continually evolving with advancements in technology reshaping the way medical students are trained and assessed. One such technological innovation that has garnered significant attention in recent years is the integration of large language models (LLMs) [1]. A significant advantage of LLMs such as ChatGPT is their ability to provide explanations for solutions, thereby making it easier for students to understand exam architecture. Learning content can be tailored based on the user's knowledge level, and the chat function allows interactive learning.

LLMs, such as ChatGPT, are supported by deep neural networks and have been trained on vast datasets. These models have profound text analysis and generation capabilities, making them exceptionally promising tools for both medical practice and education [2, 3].

As an indispensable component of medical practice, radiology necessitates profound comprehension of intricate imaging studies and clinical implications. Medical students, on their journey toward becoming proficient healthcare professionals, undergo rigorous training and examinations to help them gain the requisite skills and knowledge. Although the field of artificial intelligence in diagnostic radiology has primarily centered on image analysis, there has been growing enthusiasm surrounding the potential applications of LLMs, including ChatGPT, within radiology [4, 5, 6]. These applications encompass a wide spectrum, including radiology education, assistance in differential diagnoses, computer-aided diagnosis, and disease classification [5, 6, 7]. If these LLMs can demonstrate accuracy and reliability, they have the potential to serve as invaluable resources for learners, enabling rapid responses to inquiries and simplification of intricate concepts. ChatGPT has already undergone investigation regarding its potential with respect to streamlining radiology reports and facilitating clinical decision-making [8, 9]. Furthermore, LLMs have already performed commendably in a diverse array of professional examinations, even without specialized domain pretraining [10]. In the realm of medicine, they showed convincing results with respect to medical examinations [11, 12, 13].

The aim of this study was to explore and evaluate the performance of LLMs in radiology examinations for medical students in order to provide insight into the present capabilities and implications of LLMs.

Methods

This exploratory prospective study was carried out from August to October 2023. We obtained informed consent from the head of the institute to utilize the institute's own radiology examination questions for medical students.

▶ **Table 1** Performance of LLMs and medical students stratified by question type and topic.

	Total	Students		GPT-3.5		GPT-3.5 + Bing		GPT-4	
		N	%	N	%	N	%	N	%
All questions	151	115.3	76.3	103	68.2	108	71.5	134	88.7
Bone	26	19.4	74.4	17	65.4	18	69.2	23	88.5
Breast	5	3.7	73.2	2	40	2	40	4	80
Cardiovascular	13	9.9	76.9	11	84.6	10	76.9	12	92.3
Chest	19	13.6	71.7	14	73.7	14	73.7	17	89.5
Gastrointestinal	16	12.8	80	11	68.8	12	75	15	93.8
Genitourinary	6	4.2	70	4	66.7	5	83.3	5	83.3
Head and neck	39	31.1	79.8	28	71.8	29	74.4	34	87.2
Physics	11	8.5	77.4	7	63.6	7	63.6	10	90.9
Systemic	16	12	75.1	9	56.3	11	68.8	14	87.5
Clinical management	37	29.1	78.7	26	70.3	27	72.9	33	89.2
Description of imaging findings	27	20.2	74.8	16	59.3	21	77.8	23	85.2
Diagnosis	23	17.2	74.7	16	69.6	14	60.8	22	95.7
Comprehension	28	21.3	76.3	21	75	23	82.1	25	89.3
Knowledge	36	27.5	76.3	24	66.7	23	63.9	31	86.1
Higher-order	87	66.5	76.4	58	66.7	62	71.3	78	89.7
Lower-order	64	48.8	76.3	45	70.3	46	71.9	56	87.5

Multiple-Choice Question Selection and Classification

200 multiple-choice questions, each featuring four incorrect answers and one correct answer, were identified using the database of our radiology institute. These questions were originally designed for use in the radiology examination for medical students at our hospital. The exclusion criteria comprised questions containing images (n = 40) and questions with multiple correct answers (n = 9). After this selection process, 151 questions remained. The questions were then either prompted through OpenAI's API for GPT-3.5 and GPT-4 or manually pasted into the user interface (UI) of Perplexity AI (GPT 3.5 +Bing). To avoid the influence of previous responses on the model's output, a new ChatGPT session was initiated for each query. All questions were asked in three separate ChatGPT sessions, and the average performance was calculated.

A simple prompt for the question was used in the following form:

> *Question:*
> *{question text}*
> *A: {answer A}*
> *B: {answer B}*
> *C: {answer C}*
> *D: {answer D}*
> *E: {answer E}*

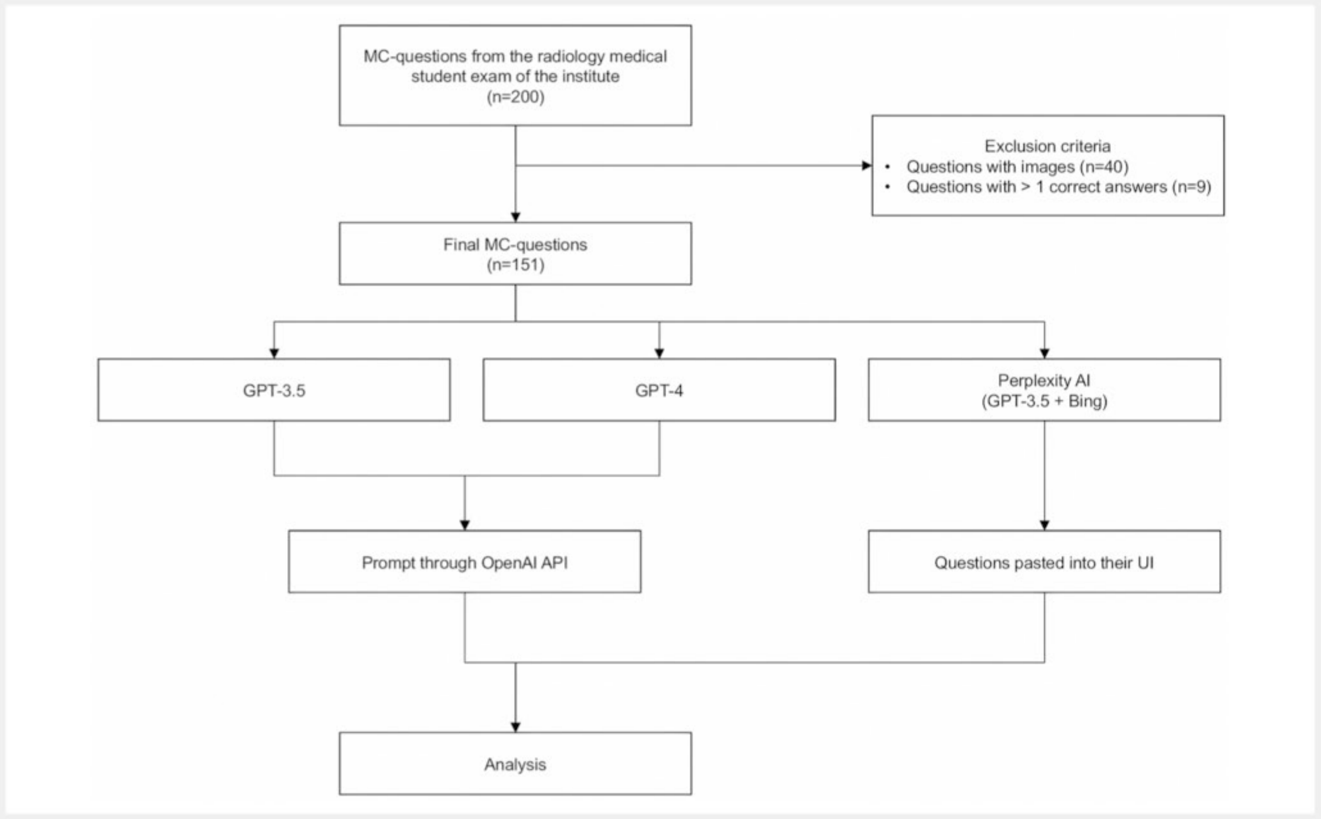
For the initial prompt we used:

> *You are an expert radiologist. Answer the following multiple-choice question in the form:*
> *<Single letter (answer)>*
> *<Text explaining the reason>*

The outputs were restructured and combined for statistical analysis. A passing score was considered to be 60 % or above. Additionally, the questions were categorized based on their type as either lower- or higher- order thinking questions, along with their subject matter, as detailed in ▶ **Table 1**. Lower-order thinking encompasses tasks related to remembering and basic understanding, while higher-order thinking involves the application, analysis, and evaluation of concepts. The higher-order thinking as well as the lower-order thinking questions were further subclassified by type (description of imaging findings, clinical management, comprehension, knowledge). Each question underwent independent classification by two radiologists. A flowchart of the study design is displayed in ▶ **Fig. 1** [14].

Large language models (LLMs)

ChatGPT (ChatGPT August 3, 2023 version, OpenAI) and Perplexity AI were used in this study. There are two versions of ChatGPT: ChatGPT, which is based on GPT-3.5, and ChatGPT Plus, which utilizes the more advanced GPT-4. In this study we used the two underlying LLMs directly via the OpenAI API. No specialized radiolo-


MC-questions from the radiology medical student exam of the institute (n=200)
Exclusion criteria
• Questions with images (n=40)
• Questions with > 1 correct answers (n=9)
Final MC-questions (n=151)
GPT-3.5
GPT-4
Perplexity AI (GPT-3.5 + Bing)
Prompt through OpenAI API
Questions pasted into their UI
Analysis

▶ **Fig. 1** Flowchart of the study design. From our initial 200 exam questions, 151 remained after excluding questions with images and questions with more than one correct answer. The questions were then prompted either by OpenAIs API for GPT-3.5 and GPT-4 or manually pasted into the UI of Perplexity AI (GPT 3.5 +Bing). The outputs were restructured and combined for statistical analysis. Abbreviations: MC: multiple choice; API: application programming interface; UI: user interface.

gy-specific pretraining was conducted for either of these models. It is important to highlight that GPT-3.5 and GPT-4, being server-contained LLMs, lack the capability to access the internet or external databases for information retrieval. In contrast, Perplexity AI (ChatGPT 3.5 +Bing) has the capacity to search the internet.

Medical students

The study included a cohort of 621 medical students who were in their first clinical semester, typically corresponding to their third year in medical school.

Prior to entering the clinical phase, the students completed two years of preclinical education, which included foundational courses in anatomy, physiology, biochemistry, pathology, and basic medical sciences. At the time of the study, the students had completed an introductory course in radiology. However, their exposure to advanced radiological topics was limited compared to more senior students and residents.

Statistical analysis

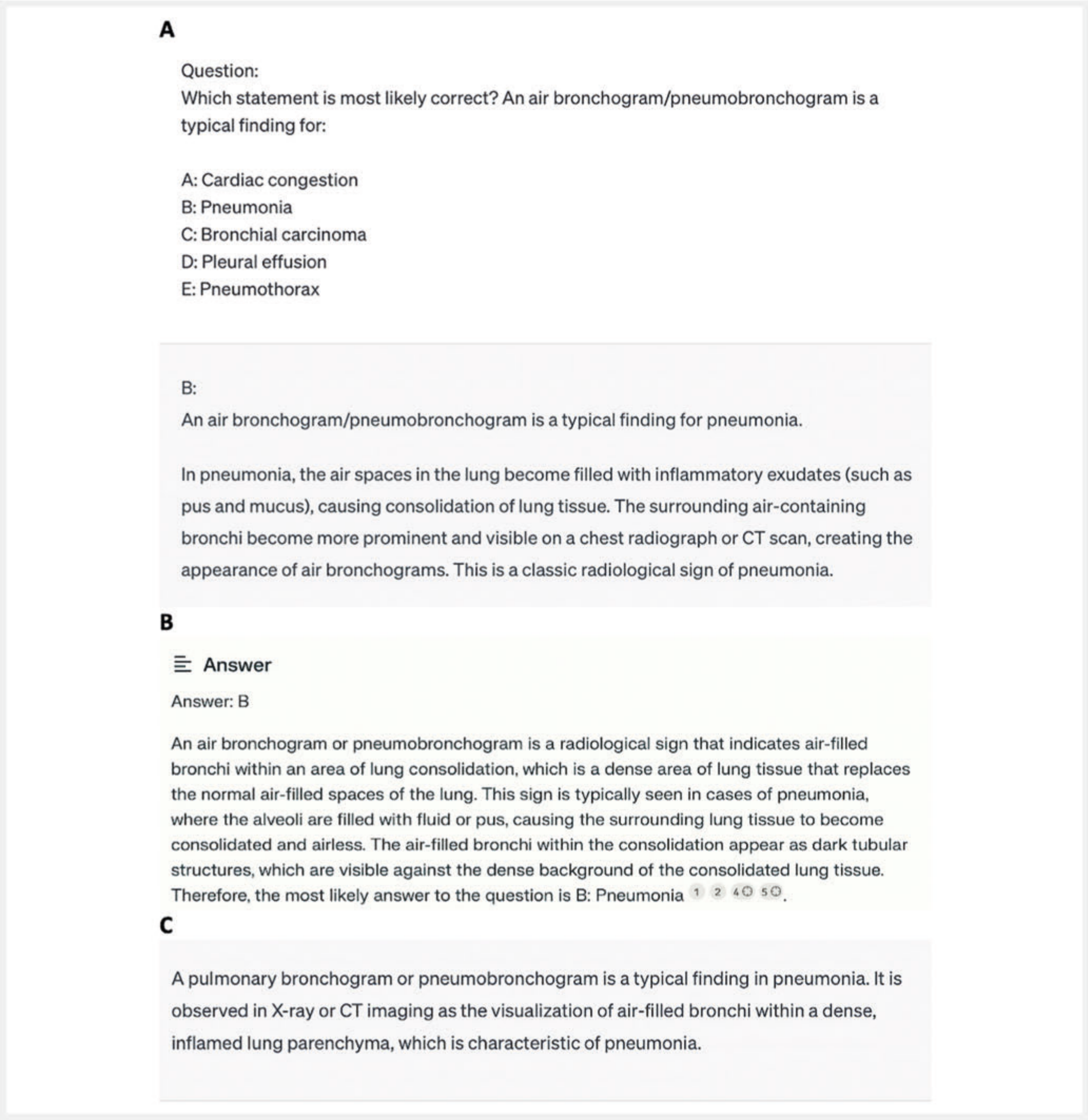
Statistical analysis was performed using Python (version 3.11). The McNemar test was used to determine the statistical significance of difference regarding the performance of the LLMs. This was also done for subgroups by question type and topic. For overall model performance, we utilized the widely used accuracy score.

To quantify the comparative performance of the LLMs and the medical students, we performed an odds ratio analysis. For each comparison, we set up 2 × 2 contingency tables that summarize the number of correct and incorrect answers for the two groups being compared. Thereafter, we calculated p-values using Fisher's Exact Test. A P-value of less than 0.05 was considered statistically significant. No correction-for-guessing was performed, since the passing score of our exam already accounts for guessing.

Results

Overall performance

The overall accuracy of GPT-3.5 for all 151 questions was 67.6 %. In contrast, GPT-4 achieved significantly higher accuracy compared to GPT-3.5 with an overall accuracy of 88.1 % ($p < 0.001$). No significant differences were observed between GPT-3.5 +Bing and GPT-3.5 ($p = 0.44$). In comparison, the overall accuracy of the medical students was 76 %. All LLMs would have passed the radiology exam for medical students at our university. ▶ **Table 1** shows the overall performance of the LLMs as well the performance stra-



A

Question:
Which statement is most likely correct? An air bronchogram/pneumobronchogram is a typical finding for:

A: Cardiac congestion
B: Pneumonia
C: Bronchial carcinoma
D: Pleural effusion
E: Pneumothorax

B:
An air bronchogram/pneumobronchogram is a typical finding for pneumonia.

In pneumonia, the air spaces in the lung become filled with inflammatory exudates (such as pus and mucus), causing consolidation of lung tissue. The surrounding air-containing bronchi become more prominent and visible on a chest radiograph or CT scan, creating the appearance of air bronchograms. This is a classic radiological sign of pneumonia.

B

Answer

Answer: B

An air bronchogram or pneumobronchogram is a radiological sign that indicates air-filled bronchi within an area of lung consolidation, which is a dense area of lung tissue that replaces the normal air-filled spaces of the lung. This sign is typically seen in cases of pneumonia, where the alveoli are filled with fluid or pus, causing the surrounding lung tissue to become consolidated and airless. The air-filled bronchi within the consolidation appear as dark tubular structures, which are visible against the dense background of the consolidated lung tissue. Therefore, the most likely answer to the question is B: Pneumonia 1 2 4 5.

C

A pulmonary bronchogram or pneumobronchogram is a typical finding in pneumonia. It is observed in X-ray or CT imaging as the visualization of air-filled bronchi within a dense, inflamed lung parenchyma, which is characteristic of pneumonia.

▶ **Fig. 2** GPT-3.5 /4.0 and Perplexity AI response to one of the questions. All picked the correct answer (option B). **A**: GPT-3.5 **B**: Perplexity AI; **C**: GPT-4.

tified by question type and topic and ▶ **Fig. 2** shows a question that was answered correctly by all LLMs.

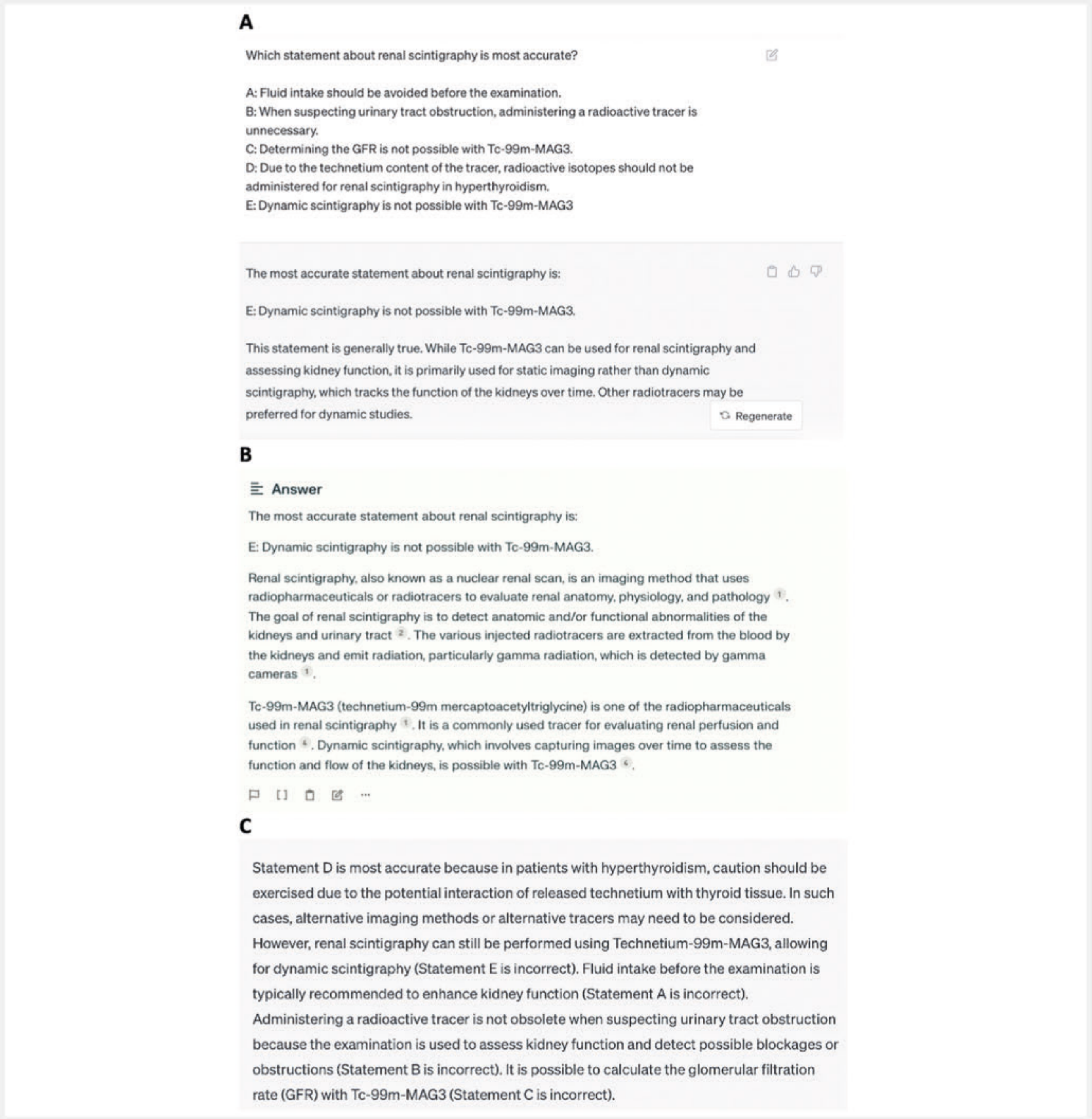
Performance by topic

Among the subgroups, GPT4 exhibited the highest performance in the gastrointestinal category, correctly answering 15 out of 16 questions, thus achieving an accuracy of 93.75 %. Compared with GPT3.5 and Perplexity AI, GPT-4 demonstrated significantly superior performance with regard to answering questions related to bone diseases (p = 0.03). However, subgroup analysis revealed no noteworthy variations in performance across the remaining subspecialty groups.

Questions answered incorrectly by all models

A total of seven questions were answered incorrectly by all models (**Table S1**). Among these, two questions pertained to the use of contrast agents in patients with renal insufficiency, while another related to MRI angiography in patients with a pacemaker.

The remaining questions that stumped all models demanded a nuanced understanding of specific details or specialized knowl-



Which statement about renal scintigraphy is most accurate?

A: Fluid intake should be avoided before the examination.
B: When suspecting urinary tract obstruction, administering a radioactive tracer is unnecessary.
C: Determining the GFR is not possible with Tc-99m-MAG3.
D: Due to the technetium content of the tracer, radioactive isotopes should not be administered for renal scintigraphy in hyperthyroidism.
E: Dynamic scintigraphy is not possible with Tc-99m-MAG3

The most accurate statement about renal scintigraphy is:

E: Dynamic scintigraphy is not possible with Tc-99m-MAG3.

This statement is generally true. While Tc-99m-MAG3 can be used for renal scintigraphy and assessing kidney function, it is primarily used for static imaging rather than dynamic scintigraphy, which tracks the function of the kidneys over time. Other radiotracers may be preferred for dynamic studies.

The most accurate statement about renal scintigraphy is:

E: Dynamic scintigraphy is not possible with Tc-99m-MAG3.

Renal scintigraphy, also known as a nuclear renal scan, is an imaging method that uses radiopharmaceuticals or radiotracers to evaluate renal anatomy, physiology, and pathology [1]. The goal of renal scintigraphy is to detect anatomic and/or functional abnormalities of the kidneys and urinary tract [2]. The various injected radiotracers are extracted from the blood by the kidneys and emit radiation, particularly gamma radiation, which is detected by gamma cameras [1].

Tc-99m-MAG3 (technetium-99m mercaptoacetyltriglycine) is one of the radiopharmaceuticals used in renal scintigraphy [1]. It is a commonly used tracer for evaluating renal perfusion and function [4]. Dynamic scintigraphy, which involves capturing images over time to assess the function and flow of the kidneys, is possible with Tc-99m-MAG3 [4].

Statement D is most accurate because in patients with hyperthyroidism, caution should be exercised due to the potential interaction of released technetium with thyroid tissue. In such cases, alternative imaging methods or alternative tracers may need to be considered. However, renal scintigraphy can still be performed using Technetium-99m-MAG3, allowing for dynamic scintigraphy (Statement E is incorrect). Fluid intake before the examination is typically recommended to enhance kidney function (Statement A is incorrect). Administering a radioactive tracer is not obsolete when suspecting urinary tract obstruction because the examination is used to assess kidney function and detect possible blockages or obstructions (Statement B is incorrect). It is possible to calculate the glomerular filtration rate (GFR) with Tc-99m-MAG3 (Statement C is incorrect).

▶ **Fig. 3** Response to a question answered incorrectly: Please be mindful that large language models (LLMs) frequently use assertive language in their responses, even when those responses are incorrect. Abbreviations: LLM: large language models.

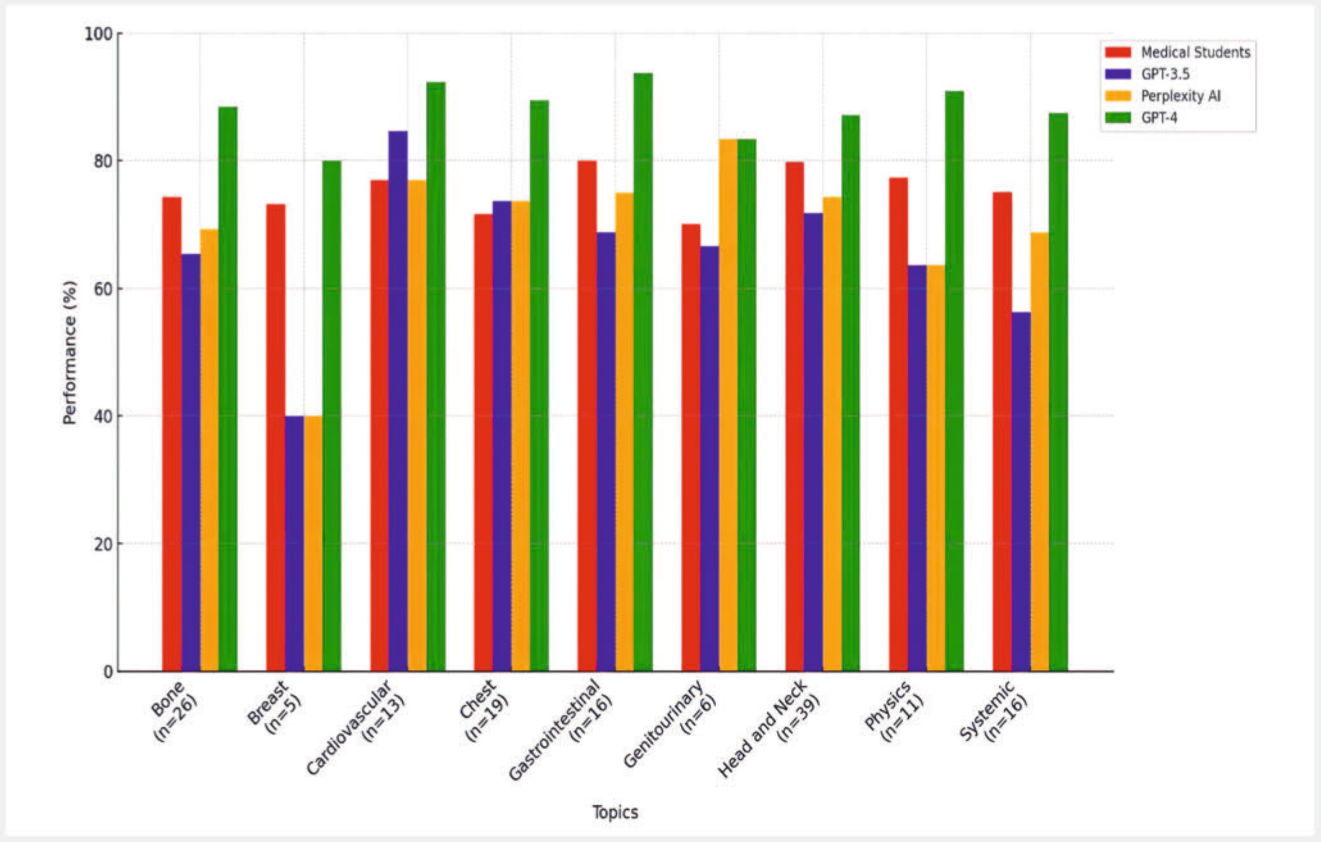
edge. For instance, one question pertained to renal scintigraphy, where the correct response hinged on the knowledge that Tc 99 m-MAG3 is primarily secreted by proximal renal tubules and, therefore, cannot be used to estimate glomerular filtration rate. ▶ **Fig. 3** illustrates a question that was answered incorrectly by all LLMs.

Performance by question type

GPT-4 demonstrated significantly superior performance in both lower-order and higher-order questions when compared to GPT-3.5 and Perplexity AI ($p = 0.01$ and $p < 0.001$, respectively).

GPT-4 achieved the best performance across all topics and categories compared to medical students, GPT-3.5, and Perplexity AI (▶ **Fig. 4**).

Within the subgroups, GPT-4 exhibited its highest performance when responding to higher-order questions related to diag-



▶ **Fig. 4** Performance comparison across medical topics: medical students vs. GPT models.

nosis. It provided correct answers for 22 out of 23 questions in this category, achieving an accuracy of 95.65 %.

In contrast, GPT-3.5 and Perplexity AI exhibited their highest performance with respect to the lower-order subgroup comprehension with accuracies of 75.00 % and 82.41 % (▶ **Table 1**). Perplexity AI demonstrated the weakest performance in the higher-order category diagnosis (60.9 %) and in the lower-order category knowledge (63.9 %), while GPT-3.5 had the weakest performance in the higher-order description of imaging findings (59.3 %) and the lower-order category comprehension (75 %). The average medical student achieved a similar performance for lower-order questions (76.27 %) compared to higher-order questions (76.39 %). The performance of the average student was relatively stable across all subgroups. The average student achieved the highest performance with regard to questions related to clinical management with an accuracy of 78.7 % and the lowest performance with regard to diagnosis with an accuracy of 74.7 % (▶ **Table 1**, ▶ **Fig. 5**, ▶ **Fig. 6**).

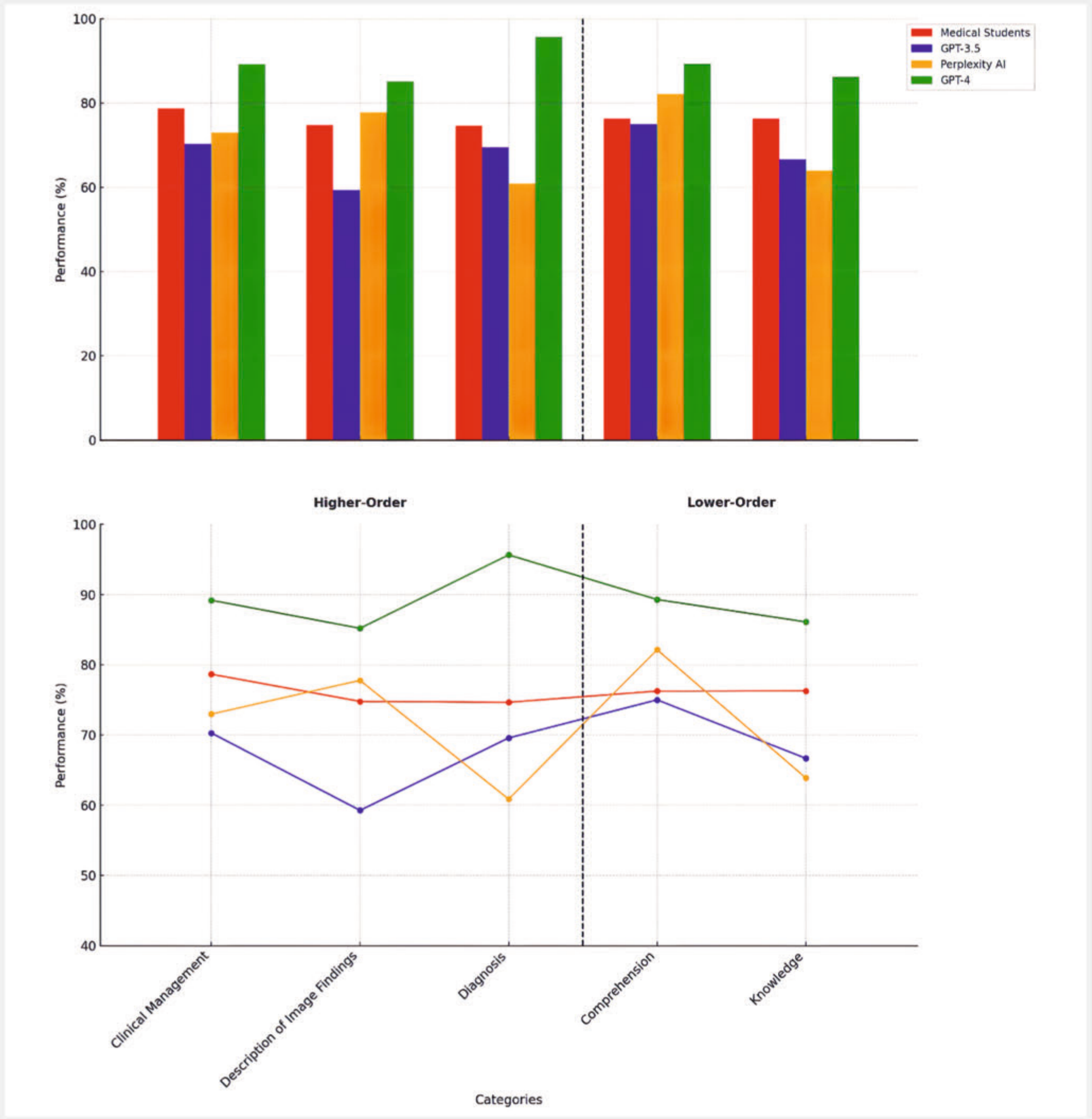
Odds ratio analysis

The odds ratio analysis confirmed that the overall performance of GPT-4 was significantly superior to that of GPT-3, Perplexity AI, and the medical students. The improved performance was particularly notable for higher-order questions, where GPT-4 showed the greatest improvement over the other GPT models and the students. For example, GPT-4 is 4.3 times more likely to correctly answer higher-order thinking questions than GPT-3.5 ($p < 0.001$). For lower-order thinking questions, while GPT-4 still performed better, the difference was not statistically significant compared to the medical students (▶ **Table 2**).

Discussion

The integration of LLMs into various domains has increased remarkably in recent years, with applications ranging from natural language processing to medical diagnostics. In the field of medical education, LLMs have shown immense potential to assist and enhance the learning experience for students, particularly in radiology – a discipline that demands profound understanding of complex medical concepts and terminology.

The present study provides several important key findings to understand how advancements in LLM technology can impact medical education. First, in this exploratory prospective study, all LLMs would have passed the exam. Second, GPT-4 exhibited significantly better performance than its predecessors GPT-3.5, Perplexity AI, and the medical students with 88 % of the questions answered correctly. Third, GPT-4 maintained the best performance across all topics and categories compared to the medical students, GPT-3.5, and Perplexity AI. Fourth, the performance improvement was particularly pronounced for higher-order questions, where GPT-4 demonstrated the most significant improvement over the other GPT models and the students. Fifth, GPT-4 demonstrated the highest performance in the gastrointestinal

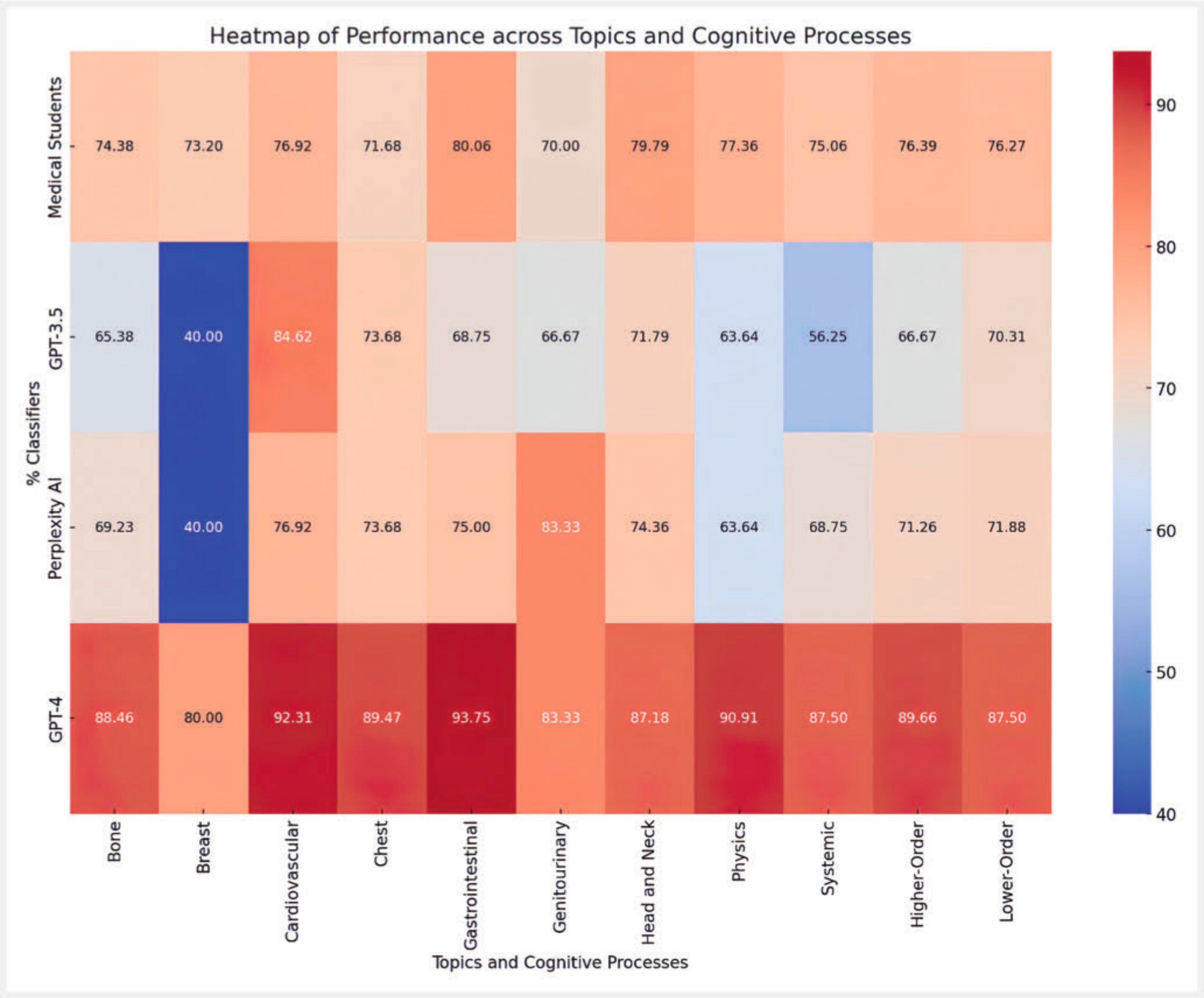


▶ **Fig. 5** Performance comparison in higher- and lower-order tasks: medical students vs. GPT models.

category with an accuracy of 93.75 %. The prevalence of gastrointestinal content in training datasets may have contributed to the model's enhanced performance in this domain.

Despite the ability of Perplexity AI to search the internet, it demonstrated the weakest performance with regard to knowledge. Internet searches can yield information from a wide range of sources, including those that are not peer-reviewed or scientifically accurate. Without a sophisticated mechanism to filter and prioritize high-quality, reliable sources, the model might incorporate inaccurate or outdated information. GPT-4's superior performance may be attributed to the fact that GPT-4 benefits from advanced model enhancements, including a deeper architecture and extensive training.

ChatGPT has demonstrated good performance in a wide range of professional examinations, including those in the medical field, even without the need for specialized domain pretraining [10, 11, 12, 13]. For instance, it was applied to the USMLE, where ChatGPT achieved accuracy rates exceeding 50 % across all examinations and surpassing 60 % in certain analyses [11].



▶ **Fig. 6** Performance heatmap across medical topics and cognitive functions.

Despite the absence of radiology-specific training, ChatGPT performed commendably. When new LLMs with radiology-specific pretraining and the ability to process images become publicly available, it will be interesting to see what results can be achieved.

As LLM technology continues to advance, radiologists will need to gain comprehensive understanding of the performance and reliability of these models and of their evolving role in radiology. The development of applications built on LLMs holds promise for further enhancing radiological practice and education, ultimately benefiting both current and future healthcare professionals. However, ChatGPT is designed to discern patterns and associations among words within its training data. Consequently, we anticipate limitations in cases requiring understanding of the context of specialized technical language or specific details and specialized knowledge, such as radiological terminology used in imaging descriptions, calculations, and classification systems.

Furthermore, ChatGPT consistently employs confident language in its responses, even when those responses are incorrect. This tendency is a well-documented limitation of LLMs [15]. Even when the most probable available option may be incorrect, ChatGPT tends to generate responses that sound convincingly human-like. Interestingly, increased human likeness in chatbots is associated with a higher level of trust [16]. Consequently, ChatGPT's inclination to produce plausible yet erroneous responses presents a significant concern when it serves as the sole source of information [17]. This concern is particularly critical with regard to individuals who may lack the expertise to discern inaccuracies in its assertions, notably novices. As a result, this behavior currently restricts the practicality of employing ChatGPT in medical education.

To prevent a future where LLMs influence the outcome of medical and radiological exams, several measures can be taken. These include designing exam questions that necessitate critical thinking and the application of knowledge rather than mere recall, integrating practical components or simulations that cannot be easily answered by LLMs, ensuring robust exam proctoring and monitoring procedures to detect any suspicious behavior, and continually updating exam formats and content to stay ahead of

▶ **Table 2** Odds ratio analysis of the performance of LLMs and medical students.

	Comparison	Odds ratio	p-value
All questions	GPT-4 vs. GPT-3.5	3.4	0.00002
	GPT-4 vs. Medical Students	2.2	0.006
	GPT-4 vs. Perplexity AI	2.9	0.0003
Higher-order	GPT-4 vs. GPT-3.5	4.3	0.0004
	GPT-4 vs. medical students	2.7	0.03
	GPT-4 vs. Perplexity AI	3.5	0.004
Lower-order	GPT-4 vs. GPT-3.5	2.9	0.03
	GPT-4 vs. Perplexity AI	2.7	0.047
	GPT-4 vs. medical students	2.2	0.11

potential cheating methods involving LLMs. Additionally, emphasizing the importance of genuine learning and skill acquisition can help maintain the integrity of medical exams amidst technological advancements.

Furthermore, we identified inconsistencies in ChatGPT's responses. In a subsequent evaluation, GPT-3.5 yielded different answers for five questions, while GPT-4 provided six different answers, but there were no significant differences in accuracy between the two models. These inconsistencies can be partially mitigated by adjusting parameters such as temperature, top-k, and top-p settings. Temperature controls the randomness of the model's responses; a lower temperature makes the output more focused and deterministic, while a higher temperature increases variability. Top-k limits the model to considering only the top k most likely next words, thus reducing the chance of less probable words being selected. Top-p adjusts the probability mass, allowing the model to consider the smallest possible set of words whose cumulative probability exceeds a certain threshold p, thereby balancing diversity and coherence.

However, this adjustment cannot be made directly through the web interface but can be done, for instance, in the OpenAI playground. Without a nuanced understanding of the influence of these parameters, there's a risk of overestimating or underestimating LLM capabilities, potentially leading to misleading conclusions about their effectiveness in educational settings. Moreover, the variability introduced by different parameter settings may result in significant fluctuations in LLM performance, thus challenging the generalizability of findings to real-world applications. Future research should prioritize comprehensive analyses of the impact of LLM settings on responses to radiology exam questions to ensure accurate assessments and to optimize LLM configurations for educational use in specialized fields.

Furthermore, it is essential to acknowledge certain limitations. First, we excluded questions containing images, which are typically integral to a radiology examination, due to ChatGPT's inability to process visual content at the time of this study. To thoroughly assess the performance of the LLMs presented in a real-world scenario, including all question types, further studies are necessary.

Second the pass/fail threshold we applied is an approximation, as normally a passing score of 60 % or above is standard for all written components, including those featuring image-based questions. Furthermore, the relatively small number of questions in each subgroup within this exploratory study has limited the statistical power available for subgroup analyses.

In conclusion, our study underscores the potential of LLMs like ChatGPT as a new and readily accessible knowledge source for medical students. Even without radiology-specific pretraining, ChatGPT demonstrated remarkable performance, achieving a passing grade on a radiology examination for medical students that did not include images. The model excelled with respect to higher-order as well as lower-order thinking questions. It is crucial for radiologists to be aware of ChatGPT's limitations, including its tendency to confidently generate inaccurate responses. Presently, it cannot be solely relied upon for clinical practice or educational purposes. However, ChatGPT presents an exciting opportunity as a new and readily accessible knowledge source for medical students, offering them a valuable tool to supplement their learning and understanding of radiology concepts.

Declarations

We disclose that the manuscript was proofread by ChatGPT. All sections proofread by ChatGPT were meticulously reviewed. Additionally, we adhered to data protection regulations, ensuring that only anonymized data was uploaded.

Statistical analysis was performed using Python (version 3.11). ChatGPT was utilized to understand and debug the Python code and adjust the graphics (▶ **Fig. 4**, ▶ **Fig. 5**, ▶ **Fig. 6**). Specifically, the diagrams were created using the Python code.

Informed Consent: Not applicable.

Data availability statement: All data and information used are included in this manuscript.

Conflict of Interest

C.B. received speaking fees from Siemens Healthineers. The other authors have no potential conflict of interest to disclose.

References

[1] Introducing ChatGPT [Internet]. [zitiert 6. September 2023]. Verfügbar unter: https://openai.com/blog/chatgpt

[2] Sallam M. The Utility of ChatGPT as an Example of Large Language Models in Healthcare Education, Research and Practice: Systematic Review on the Future Perspectives and Potential Limitations [Internet]. medRxiv; 2023 [zitiert 6. September 2023]. S. 2023.02.19.23286155. Verfügbar unter: https://www.medrxiv.org/content/10.1101/2023.02.19.23286155v1

[3] Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023; 5 (3): e107–e108

[4] Hosny A, Parmar C, Quackenbush J et al. Artificial intelligence in radiology. Nat Rev Cancer 2018; 18 (8): 500–510

[5] Shen Y, Heacock L, Elias J et al. ChatGPT and Other Large Language Models Are Double-edged Swords. Radiology 2023; 307 (2): e230163

[6] Wang S, Zhao Z, Ouyang X et al. ChatCAD: Interactive Computer-Aided Diagnosis on Medical Image using Large Language Models [Internet]. arXiv; 2023 [zitiert 6. September 2023]. Verfügbar unter: http://arxiv.org/abs/2302.07257

[7] Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. J Med Internet Res 2023; 25: e48568

[8] Rao A, Kim J, Kamineni M et al. Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making [Internet]. medRxiv; 2023 [zitiert 6. September 2023]. S. 2023.02.02.23285399. Verfügbar unter: https://www.medrxiv.org/content/10.1101/2023.02.02.23285399v1

[9] Jeblick K, Schachtner B, Dexl J et al. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports [Internet]. arXiv; 2022 [zitiert 6. September 2023]. Verfügbar unter: http://arxiv.org/abs/2212.14882

[10] Choi JH, Hickman KE, Monahan A et al. ChatGPT Goes to Law School [Internet]. Rochester, NY; 2023 [zitiert 6. September 2023]. Verfügbar unter: https://papers.ssrn.com/abstract=4335905

[11] Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models | PLOS Digital Health [Internet]. [zitiert 6. September 2023]. Verfügbar unter: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000198

[12] Gilson A, Safranek CW, Huang T et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ 2023; 9: e45312

[13] 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

[14] Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiology: Artificial Intelligence 2020. doi:10.1148/ryai.2020200029

[15] Xiao Y, Wang WY. On Hallucination and Predictive Uncertainty in Conditional Language Generation. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume [Internet]. Online: Association for Computational Linguistics; 2021 [zitiert 15. September 2023]. S. 2734–44. Verfügbar unter: https://aclanthology.org/2021.eacl-main.236

[16] Lu L, McDonald C, Kelleher T et al. Measuring consumer-perceived humanness of online organizational agents. Computers in Human Behavior 2022; 128: 107092

[17] Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus 2023; 15 (2): e35179