








Radiological Differential Diagnoses Based on Cardiovascular and Thoracic Imaging Patterns: Perspectives of Four Large Language Models

Pradosh Kumar Sarangi¹ Aparna Irodi² Swaha Panda³ Debasish Swapnesh Kumar Nayak⁴
Himel Mondal⁵

¹ Department of Radiodiagnosis, All India Institute of Medical Sciences, Deoghar, Jharkhand, India

² Department of Radiodiagnosis, Christian Medical College and Hospital, Vellore, Tamil Nadu, India

³ Department of Otorhinolaryngology and Head and Neck Surgery, All India Institute of Medical Sciences, Deoghar, Jharkhand, India

⁴ Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India

⁵ Department of Physiology, All India Institute of Medical Sciences, Deoghar, Jharkhand, India

Address for correspondence Himel Mondal, MBBS, MD, Department of Physiology, All India Institute of Medical Sciences, Deoghar, Jharkhand 814152, India (e-mail: himelmkg@gmail.com).

Indian J Radiol Imaging 2024;34:269–275.

Abstract

Background Differential diagnosis in radiology is a critical aspect of clinical decision-making. Radiologists in the early stages may find difficulties in listing the differential diagnosis from image patterns. In this context, the emergence of large language models (LLMs) has introduced new opportunities as these models have the capacity to access and contextualize extensive information from text-based input.

Objective The objective of this study was to explore the utility of four LLMs—ChatGPT3.5, Google Bard, Microsoft Bing, and Perplexity—in providing most important differential diagnoses of cardiovascular and thoracic imaging patterns.

Methods We selected 15 unique cardiovascular ($n = 5$) and thoracic ($n = 10$) imaging patterns. We asked each model to generate top 5 most important differential diagnoses for every pattern. Concurrently, a panel of two cardiothoracic radiologists independently identified top 5 differentials for each case and came to consensus when discrepancies occurred. We checked the concordance and acceptance of LLM-generated differentials with the consensus differential diagnosis. Categorical variables were compared by binomial, chi-squared, or Fisher's exact test.

Results A total of 15 cases with five differentials generated a total of 75 items to analyze. The highest level of concordance was observed for diagnoses provided by Perplexity (66.67%), followed by ChatGPT (65.33%) and Bing (62.67%). The lowest score was for Bard with 45.33% of concordance with expert consensus. The acceptance rate was highest for Perplexity (90.67%), followed by Bing (89.33%) and ChatGPT (85.33%). The lowest acceptance rate was for Bard (69.33%).

Keywords

- ▶ artificial intelligence
- ▶ cardiothoracic
- ▶ ChatGPT
- ▶ Google Bard
- ▶ Microsoft Bing
- ▶ perplexity
- ▶ differential diagnosis
- ▶ radiologists

article published online
December 28, 2023

DOI <https://doi.org/10.1055/s-0043-1777289>.
ISSN 0971-3026.

© 2023. Indian Radiological Association. All rights reserved.
This is an open access article published by Thieme under the terms of the Creative Commons Attribution-NonDerivative-NonCommercial-License, permitting copying and reproduction so long as the original work is given appropriate credit. Contents may not be used for commercial purposes, or adapted, remixed, transformed or built upon. (<https://creativecommons.org/licenses/by-nc-nd/4.0/>)
Thieme Medical and Scientific Publishers Pvt. Ltd., A-12, 2nd Floor, Sector 2, Noida-201301 UP, India

Conclusion Four LLMs—ChatGPT3.5, Google Bard, Microsoft Bing, and Perplexity—generated differential diagnoses had high level of acceptance but relatively lower concordance. There were significant differences in acceptance and concordance among the LLMs. Hence, it is important to carefully select the suitable model for usage in patient care or in medical education.

Introduction

Identifying imaging patterns from different radiological modalities and linking them to specific pathologies while taking into account clinical contexts and probabilities is a crucial aspect of radiological diagnosis for which radiologists have to possess a vast amount of knowledge. Radiologists in the early stages of their training often rely on seeking guidance from seniors and delving into relevant literature to validate or expand their list of potential diagnoses, which can be a time-consuming and resource-intensive endeavor.¹ However, the emergence of artificial intelligence (AI) and large language models (LLMs) has introduced new opportunities in this regard as these models have the capacity to access and contextualize extensive information present in their text-based training data.² The deep learning (DL) models serve as the foundation for the design of various LLMs. DL models use artificial neural networks yet operate on the same principles as the human brain. The foundation of any accessible LLM is comprised of these pretrained DL models.³

Cardiovascular and thoracic diseases present diverse and complex imaging patterns, often necessitating careful interpretation. The advent of LLMs like Open AI's ChatGPT, Google Bard (Experiment), Microsoft Bing (creative), and Perplexity AI introduces an intriguing prospect.⁴ These models, trained on extensive medical literature and data, possess the ability to comprehend complex diagnostic contexts that offer unique insights that can potentially assist in providing differential diagnoses from text-based description of imaging pattern.⁵

ChatGPT has been explored for an adjunct for radiologic decision-making and it was found to be feasible to use it for improving clinical workflow.⁶ In addition, ChatGPT performed well in radiology board-style examination without images. Hence, it has the capability to comprehend textual description of radiological question.⁷ However, in another study, it was reported that ChatGPT3.5 performed below the average student in written tasks.⁸ Kottlors et al used the latest version of the paid model ChatGPT4. They found that ChatGPT4 provides 68.8% concordant and 93.8% acceptable differential diagnosis in radiology.⁹ ChatGPT4 is a premium version of Open AI's chatbot. Users from developing countries may not have access to this version. Free chatbots like Google Bard (Experiment), Microsoft Bing (creative), and Perplexity are available for users.

The role of freely available chatbots in the domain of radiology in providing relevant differential diagnoses from text-based descriptions of image patterns remains unexplored. Hence, this study aimed to bridge this gap by investigating the potential of four important and widely used free

LLMs to provide relevant differential diagnosis from imaging pattern (cardiovascular and thoracic imaging). By comparing their generated differential diagnoses against expert consensus, the utility of LLMs in augmenting traditional diagnostic approaches is explored.

Methods

Study Design

This research employed a cross-sectional observational study design to explore the application of LLMs in suggesting most relevant differential diagnoses for cardiovascular and thoracic imaging patterns.

Imaging Pattern

We curated a dataset of 15 cardiovascular and thoracic imaging patterns sourced from a textbook (Chapman & Nakielny's Aids to Radiological Differential Diagnosis) and an online platform <https://radiopaedia.org>.⁹ The imaging patterns are shown in ►Table 1.

LLMs

We observed that various LLMs have been developed in recent years. According to the literature, there will more than 36 LLMs in the market by 2023.⁴ There are two different kinds of accessible LLMs: one is open source and available to all users for free and the other is subscription-based and requires a fee to use the advanced features. Based on their popularity, architecture, usefulness, and services to medical science, we chose four open source LLMs for this study. We used Open AI's ChatGPT3.5 (<https://chat.openai.com>) free research version, Google Bard (<https://bard.google.com>) Experiment, Microsoft Bing (<https://www.bing.com/>) Chat (Creative) based on GPT4, and Perplexity AI (<https://www.perplexity.ai>). Henceforth in this manuscript, we will refer to these as ChatGPT, Bard, Bing, and Perplexity. A summary of the four LLMs used in this study is shown in ►Table 2.

Model-Generated Differential Diagnoses

For each of the 15 imaging patterns, ChatGPT, Bard, Bing, and Perplexity were asked to generate top 5 most important differential diagnoses. These model-generated diagnoses were stored for further analysis. A brief of the study procedure is shown in ►Fig. 1.

Expert Consensus

An expert panel comprising two experienced radiologists specialized in cardiothoracic imaging independently

Table 1 Cardiothoracic imaging pattern used in the study

Thorax	Unilateral hyperlucent hemithorax on chest radiograph
	Nonresolving or recurrent lung consolidation
	Reticular pattern with honeycombing in the lungs
	Mosaic attenuation pattern in high-resolution computed tomography (HRCT) of the thorax
	Pulmonary nodules with cavitation
	Rib lesion with an adjacent soft-tissue mass
	Diffuse ground-glass nodules on HRCT of the thorax
	Pediatric mediastinal masses
	Mediastinal mass containing fat
	Cystic lung disease
Cardiac	Late gadolinium enhancement on cardiac magnetic resonance imaging (MRI)
	Pulmonary arterial enlargement
	Cardiac calcification
	Septal bounce sign on cardiac MRI
	Left ventricular hypertrophy

Table 2 Language models used in this study and their architectures

LLM	Developer	Launch date	Transformer/neural network architectures
Bard	Google AI	March 21, 2023	PaLM
Bing	Microsoft	February 2023	GPT 4
ChatGPT3.5	OpenAI	November 30, 2022	GPT 3.5
Perplexity	Perplexity AI	August 2022	GPT 3.5

Abbreviations: AI, artificial intelligence; GPT, generative pretrained transformer; LLM, large language models; PaLM, pathways language model.

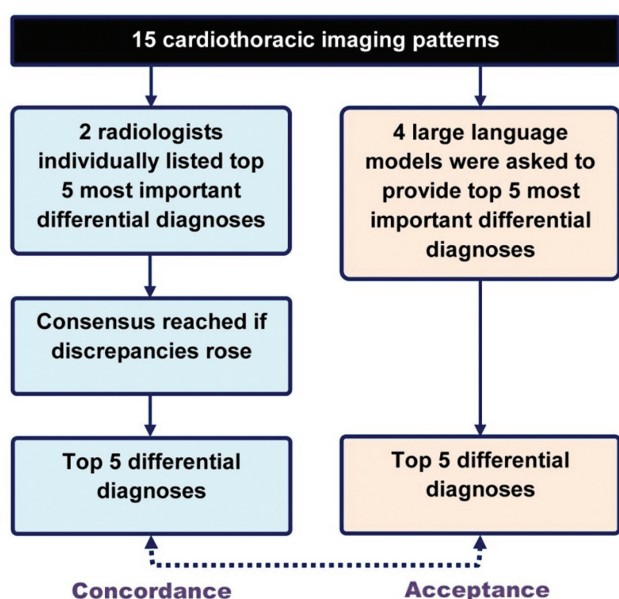


Fig. 1 Brief study procedure.

identified the top five most important differential diagnoses for each imaging pattern leveraging their clinical expertise, domain knowledge, and book references. Then a consensus was reached to generate a final list of five most important differential diagnoses for each imaging pattern.

Concordance and Acceptance Evaluation

To assess the performance of the LLMs, we evaluated two key metrics—concordance and acceptance. Concordance was the overlap between the differential diagnoses suggested by the LLMs and those determined by the expert consensus panel (i.e., matching differentials). Acceptance was determined by the proportion of model-generated diagnoses that were deemed acceptable alternatives by the experts including concordance. Experts had the liberty to utilize reference sources they considered suitable to validate their judgments, when needed, such as textbooks, publications, or online platforms.

Statistical Analysis

The results were presented in number and percentages. Categorical variables were compared statistically by the

chi-squared test or Fisher's exact test where frequency was less than 5. The statistically significant difference between yes and no categories was tested by binomial test where significance indicates that the occurrence was not by chance. We used Microsoft Excel 2010 for data storage and GraphPad Prism 9.5.0 (GraphPad Software, United States) for inferential statistics. A *p* value of less than 0.05 was considered statistically significant.

Ethical Considerations

The study did not use any identifiable patient data. The data generated by LLMs were also not presented in this study. Hence, according to the ethical guidelines, this study does not require institutional ethics committee clearance.

Results

A total of 15 cases with five differentials generated a total of 75 items to analyze. The highest level of concordance was observed for diagnoses provided by Perplexity (66.67%), followed by ChatGPT (65.33%) and Bing (62.67%). The lowest score was for Bard with 45.33% of concordance with expert consensus. The acceptance rate was highest for Perplexity (90.67%), followed by Bing (89.33%) and ChatGPT (85.33%). The lowest acceptance rate was for Bard (69.33%; ►Fig. 2). However, the acceptance and concordance percentages were not significantly different from each other ($p = 0.93$).

Domain-wise score of four LLMs are shown in ►Table 3. ChatGPT in cardiac, Bing in thorax, and Perplexity in thorax showed significance in concordance. However, all LLMs showed significantly higher acceptance. There was no statistically significant difference in the performance of LLMs in providing differential diagnosis in cardiac and thoracic cases.

The concordance among the four LLMs were significantly different (chi-squared test, $p = 0.002$) and the scores are

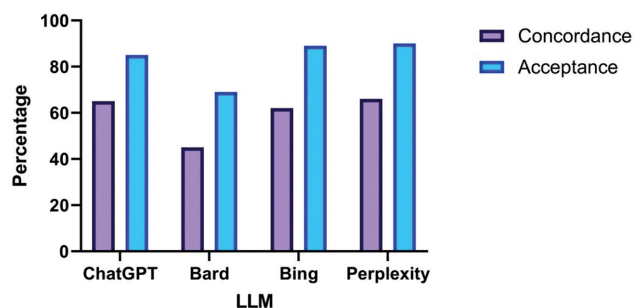


Fig. 2 Percentage of concordance and acceptance of diagnoses. LLM, large language model.

shown in ►Fig. 3. The acceptance also differed (chi-squared; $p = 0.03$) in the four LLMs as shown in ►Fig. 4.

Discussion

In terms of concordance with expert consensus, Perplexity emerged as the top performer, with a little lower performance by ChatGPT and Bing (all had >60% concordance). This suggests that the algorithm and training data used by the LLMs to generate diagnoses align closely with what experts would determine. These concordance rates are similar to the concordance rate of ChatGPT4 (69%) as reported by Kottlors et al.⁹ Bard, with the lowest concordance rate, likely employs an algorithm or training data that substantially diverge from expert consensus, leading to a lower level of agreement.

Examining the acceptance rates of the generated diagnoses, Perplexity once again came out on top with the highest acceptance rate and ChatGPT and Bing also had an acceptance rate greater than 85%. This suggests that the diagnoses generated by these three were more likely to be accepted by the evaluators. The acceptance rate was slightly lower than

Table 3 Domain wise concordance and acceptance of diagnoses provided by four large language models

LLM	Category	Concordance		<i>p</i> -Value (binomial)	Acceptance		<i>p</i> -Value (binomial)
		Yes	No		Yes	No	
		<i>n</i> (%)			<i>n</i> (%)		
ChatGPT	Thorax (<i>n</i> = 50)	29 (58)	21 (42)	0.26	41 (82)	9 (18)	< 0.0001 ^a
	Cardiac (<i>n</i> = 25)	20 (80)	5 (20)	0.004 ^a	23 (92)	2 (8)	< 0.0001 ^a
	<i>p</i> (chi-squared)	0.07		–	0.32		–
Bard	Thorax (<i>n</i> = 50)	23 (46)	27 (54)	0.67	34 (68)	16 (32)	0.02 ^a
	Cardiac (<i>n</i> = 25)	11 (44)	14 (56)	0.69	19 (76)	6 (24)	0.009 ^a
	<i>p</i> (chi-squared)	0.87		–	0.47		–
Bing	Thorax (<i>n</i> = 50)	32 (64)	18 (36)	0.06 ^a	45 (90)	5 (10)	< 0.0001 ^a
	Cardiac (<i>n</i> = 25)	15 (60)	10 (40)	0.33	22 (88)	3 (12)	0.0002 ^a
	<i>p</i> (chi-squared)	0.74		–	0.79		–
Perplexity	Thorax (<i>n</i> = 50)	33 (66)	17 (34)	0.03 ^a	47 (94)	3 (6)	< 0.0001 ^a
	Cardiac (<i>n</i> = 25)	17 (68)	8 (32)	0.08	21 (84)	4 (16)	0.009 ^a
	<i>p</i> (chi-squared)	0.86		–	0.16		–

^aStatistically significant *p*-Value of binomial test.

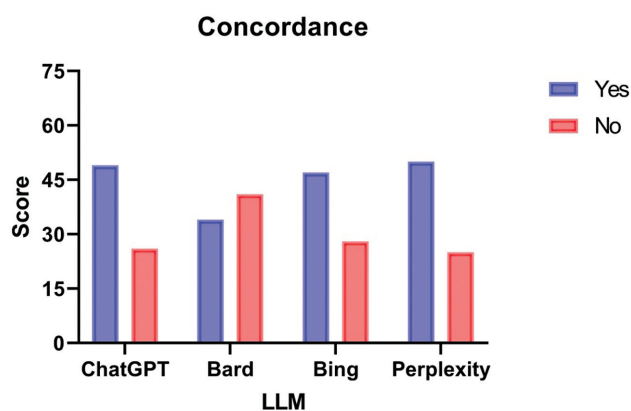


Fig. 3 Concordance scores of four large language models (LLMs) in overall cases (15 imaging patterns and 75 differential diagnoses).

that of ChatGPT4 (94%) as reported by Kottlors et al.⁹ Bard, with the lowest acceptance rate (69.33%), likely generated diagnoses that were less frequently deemed acceptable. This could be due to variation and limitations in algorithm or training data, resulting in diagnoses that were different across different LLMs. However, exploring the underlying cause was beyond the scope of this study.

In the context of concordance, only cardiac domain for ChatGPT, thorax domain for Bing, and thorax domain for Perplexity showed significance. However, in other instances, the score was not significantly different, which indicates that responses of the models are not necessarily generating differential diagnosis like human experts. However, all LLMs demonstrated significance in acceptance across the domains. This implies that regardless of the specific medical domain (cardiac or thorax), all of these language models produced diagnoses or assessments that were considered acceptable by evaluators or domain experts. This uniform significance in acceptance underscores the overall competence of these LLMs in generating diagnoses that are deemed suitable or credible in both the cardiac and thorax domains.

When we compared the overall score of concordance and acceptance, there was significant difference in concordance and acceptance rates among LLMs. The significant differ-

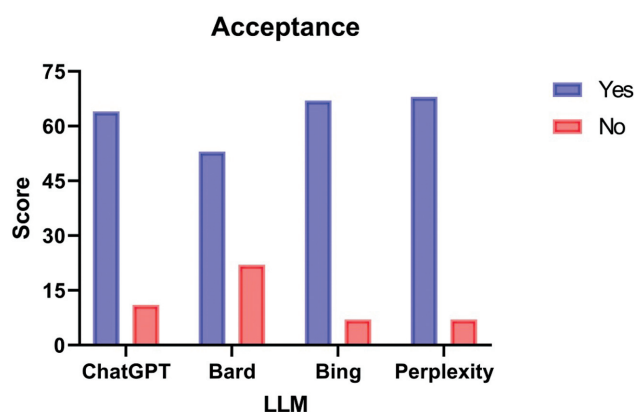


Fig. 4 Acceptance scores of four large language models (LLMs) in overall cases (15 imaging patterns and 75 differential diagnoses).

ences in acceptance rates and concordance levels among the four LLMs were likely the result of a combination of factors, including algorithm design, training data, model complexity, domain-specific knowledge, potential bias, postprocessing methods, and variability in evaluator perspectives. However, as a radiologist, for clinical purposes, a careful consideration of concordance and acceptance may be the priority for selecting the LLMs in generating the most significant differential diagnoses within the realm of cardiovascular and thoracic imaging patterns, and potentially other subspecialties in radiology. Concordance and acceptance are two important metrics to assess the performance of LLMs. A higher level of concordance indicates a closer alignment with expert consensus. Therefore, an LLM with higher concordance and acceptance is preferred.

While LLMs have been explored in various domains of medical decision-making,¹⁰⁻¹³ our study is an addition to the knowledge of AI in radiology, especially for the cardiothoracic imaging pattern. This study may be beneficial for radiologists in training who aim to strike a balance between clinical efficiency and ongoing knowledge acquisition. Moreover, the study holds the potential to guide the development of LLM-based decision support systems tailored for cardiovascular and thoracic imaging interpretation. Such systems could empower health care professionals to make more informed decisions, improving patient outcomes and contributing to the ongoing evolution of personalized medicine. In addition, this study holds importance in teaching radiology and facilitating self-directed learning for students in the medical field.¹⁴ They can utilize this information to make judicious decisions about which LLMs to integrate into educational resources, ensuring that students have access to materials with the highest concordance rates and acceptance levels.

However, in some instances, the LLMs may fail to provide a credible answer. For an example, **Fig. 5** shows the answer by ChatGPT3.5 to the question – “Think you are a radiologist. Please provide me with the top five differential diagnoses of diffuse ground-glass nodules on HRCT thorax.” The answer neither had concordance nor acceptance. Hence, users should be careful while using LLMs for patient care and always remember the limitation of AI.

Novelty and Limitation

To the best of our knowledge, this is the initial study evaluating the capabilities of four LLMs in generating appropriate lists of potential diagnoses based on textual descriptions of image patterns in cardiothoracic radiology. It is important to note that the results presented in this report are preliminary, and there is a need for more comprehensive research using a structured methodology. Nevertheless, it is essential to emphasize that the LLMs used in this initial analysis were not specifically designed and trained for the purpose of generating differential diagnoses from text-based descriptions of image patterns in radiology.

There are some limitations of the study. Only two radiologists made the list of differential diagnoses and four LLMs were tested. Furthermore, the study's findings may not fully

When interpreting diffuse ground-glass nodules on high-resolution computed tomography (HRCT) of the thorax, there are several important differential diagnoses to consider. Ground-glass nodules refer to areas of the lung parenchyma that appear hazy and less opaque than the surrounding lung tissue, and nodular opacities are present within this hazy appearance. Here are the top five differential diagnoses for diffuse ground-glass nodules:

1. **Interstitial Lung Disease (ILD):** Various types of interstitial lung diseases, such as idiopathic pulmonary fibrosis (UIP pattern), non-specific interstitial pneumonia (NSIP), and hypersensitivity pneumonitis, can present with diffuse ground-glass nodules as part of their imaging findings.
2. **Lymphangitic Carcinomatosis:** Lymphangitic carcinomatosis is a condition where cancer cells spread along the lymphatic vessels within the lung. It can cause diffuse ground-glass nodules due to lymphatic obstruction and inflammation.
3. **Cryptogenic Organizing Pneumonia (COP):** COP is characterized by inflammatory plugs within the small airways and alveoli. It can lead to ground-glass nodules, often with a peripheral and subpleural distribution.
4. **Acute Respiratory Distress Syndrome (ARDS):** ARDS is a severe lung condition often triggered by various insults. It can result in diffuse ground-glass nodules due to alveolar inflammation, hyaline membrane formation, and fibrosis.
5. **Pulmonary Alveolar Proteinosis (PAP):** PAP is a rare disorder characterized by the accumulation of surfactant-derived lipoproteinaceous material in the alveoli. It can manifest as diffuse ground-glass nodules on imaging.

Fig. 5 An example answer by ChatGPT-3.5 where the differential diagnoses were neither concordant nor acceptable.

generalize to real-world medical settings, where clinical judgment, patient history, and physical examinations play pivotal roles in diagnosis. In addition, LLMs are continuously evolving technologically.¹⁵ Hence, the result at this point of time may vary in the near future. Therefore, the results should be interpreted with caution, recognizing the limitations. In addition, we only used textual input to get response from the LLMs and did not feed any image. However, our study functions as a demonstration of the capability of LLMs to produce pertinent differential diagnoses tailored to distinct imaging patterns. Consequently, it underscores their potential in offering support for diagnostic decision-making.

Conclusion

This study sheds light on the varying performance of LLMs in predicting medical differential diagnoses from cardiothoracic imaging patterns. There was acceptance of differential diagnoses generated by LLMs, but their concordance with expert radiologists was low. Significant differences were also observed in acceptance rates and concordance levels among the LLMs. Hence, it is important to carefully select the suitable model for usage in patient care or in medical education. The four different LLMs tested here currently

hold great potential in providing relevant differential diagnoses from text-based descriptions of image patterns in cardiothoracic radiology.

Funding

None.

Conflict of Interest

None declared.

Acknowledgments

The corresponding author would like to thank Sarika Mondal and Ahana Aarshi for their sacrifice of family time during data analysis, interpretation, visualization, preparation, and handling of the manuscript on journal management system.

References

- 1 Hussain S, Mubeen I, Ullah N, et al. Modern diagnostic imaging technique applications and risk factors in the medical field: a review. *BioMed Res Int* 2022;2022:5164970
- 2 Alberts IL, Mercolli L, Pyka T, et al. Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? *Eur J Nucl Med Mol Imaging* 2023;50(06):1549–1552

- 3 De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health* 2023;11:1166120
- 4 Kumari A, Kumari A, Singh A, et al. Large language models in hematology case solving: a comparative study of ChatGPT-3.5, Google Bard, and Microsoft Bing. *Cureus* 2023;15(08):e43861
- 5 Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023;29(08):1930–1940
- 6 Rao A, Kim J, Kamineneni M, Pang M, Lie W, Dreyer KJ, Succu MD. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. *J Am Coll Radiol* 2023;20(10):990–997
- 7 Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology* 2023;307(05):e230582
- 8 Currie G, Singh C, Nelson T, Nabasenja C, Al-Hayek Y, Spuur K. ChatGPT in medical imaging higher education. *Radiography* 2023;29(04):792–799
- 9 Kottlors J, Bratke G, Rauen P, et al. Feasibility of differential diagnosis based on imaging patterns using a large language model. *Radiology* 2023;308(01):e231167
- 10 Davies SG. *Chapman & Nakielny's Aids to Radiological Differential Diagnosis*. 6th ed. Edinburg: Elsevier Saunders; 2014
- 11 Elkassem AA, Smith AD. Potential use cases for ChatGPT in radiology reporting. *AJR Am J Roentgenol* 2023;221(03):373–376
- 12 Schukow C, Smith SC, Landgrebe E, et al. Application of ChatGPT in routine diagnostic pathology: promises, pitfalls, and potential future directions. *Adv Anat Pathol* 2023 (e-pub ahead of print). Doi: 10.1097/PAP.0000000000000406
- 13 Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res* 2023;25:e48568
- 14 Mondal H, Mondal S, Podder I. Using ChatGPT for writing articles for patients' education for dermatological diseases: a pilot study. *Indian Dermatol Online J* 2023;14(04):482–486
- 15 Tsang R. Practical applications of ChatGPT in undergraduate medical education. *J Med Educ Curric Dev* 2023;10:238212 05231178449