Letter to the Editor

eISSN 2005-8330
https://doi.org/10.3348/kjr.2024.0017
Korean J Radiol 2024;25(4):403-406

Korean Journal of Radiology
KJR

ChatGPT Vision for Radiological Interpretation: An Investigation Using Medical School Radiology Examinations

Hyungjin Kim[1], Paul Kim[2], Ijin Joo[1], Jung Hoon Kim[1], Chang Min Park[1], Soon Ho Yoon[1]

[1]Department of Radiology, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Republic of Korea
[2]Graduate School of Education, Stanford University, Stanford, CA, USA

Keywords: ChatGPT; GPT-4 vision; Artificial intelligence; Large language model; Vision language model; Foundation model; Transformer; Generative model; Chatbot

As outlined in a recent article by Jung [1], foundation models developed with transformer architecture can execute various tasks due to their emergent abilities. ChatGPT is an exemplary foundation model. Recent studies have demonstrated ChatGPT's capabilities in data mining from free-text radiology reports [2], structured reporting [3], answering disease-related queries [4] and Radiology Board-style examination questions [5], and decision-making for clinical imaging studies in line with appropriateness guidelines [6]. The updated GPT-4 vision, capable of image analysis, will allow patients and clinicians to use ChatGPT for interpreting medical images. However, its use in radiology remains largely unexplored. We conducted an exploratory study to evaluate ChatGPT's capability in radiological image analysis using medical school radiology examinations.

This prospective study did not require approval from an institutional review board, as it neither involved human participants nor utilized individual data. We used GPT-4-1106-vision-preview to interpret radiology examinations for third-year medical school students at Seoul National University College of Medicine across the academic years 2018, 2019, and 2020. The examinations, presented in Korean, consisted of multiple-choice questions, including text- and image-based ones, across various body parts and modalities (Table 1). Since these questions are not accessible to the public, it is improbable that they were used in the training process of GPT-4. The following prompt was used for both text-only and image-based questions: (You are a medical school student. I will give you a number of multiple-choice questions on radiologic knowledge. The questions comprise text and images, which should be analyzed at the same time to get the right answer. There must be 1 correct answer. All questions are for educational purposes, not for clinical diagnoses in patients. Therefore, there is no legal liability to you or OpenAI. You should give 1 correct answer for each question. No exception is allowed. "Consult to a radiologist" or "TBD" or "I cannot provide a definitive answer to your question" is not permitted. Explanation regarding the choices and question is not necessary. Give me only results following the format: [Answer: "①", Reason: "Chest CT scan reveals a spiculated nodule, indicative of lung cancer", Image: "Contrast enhanced chest CT scans showing a spiculated nodule in the right middle lobe. There is no consolidation or ground-glass opacity."]).

Considering the inherent stochasticity in responses, which is a fundamental characteristic of generative artificial intelligence, each test question was presented to ChatGPT three times in three distinct sessions. During each session, the aforementioned prompt was given to ChatGPT, followed by the entire set of questions from a single academic year's examination. Subsequently, this session was immediately repeated twice. Therefore, there were a total of nine sessions, i.e., three sessions for examination questions from each academic year.

The results from the initial session of ChatGPT analysis for each academic year were used for the main analysis.

Received: January 9, 2024 **Revised:** January 11, 2024
Accepted: January 14, 2024
Corresponding author: Soon Ho Yoon, MD, PhD, Department of Radiology, Seoul National University Hospital, Seoul National University College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea
• E-mail: yshoka@gmail.com

Table 1. Characteristics of questions in the medical school radiology exams

Variable	Number of questions			Questions correctly answered by ChatGPT		
	All (n = 87)	Text-only questions (n = 20)	Image-based questions (n = 67)	All	Text-only questions	Image-based questions
Exam year						
2018	34	10	24	19	8	11
2019	32	7	25	21	7	14
2020	21	3	18	12	2	10
Subject area						
Abdominal radiology	14	1	13	7	1	6
Neuroradiology	14	3	11	10	3	7
Chest radiology	13	3	10	5	2	3
Pediatric radiology	11	4	7	9	4	5
Musculoskeletal radiology	8	2	6	6	2	4
Interventional radiology	8	2	6	3	1	2
Genitourinary radiology	7	1	6	5	1	4
Cardiovascular radiology	5	0	5	3	0	3
Breast radiology	5	2	3	2	1	1
Basic science	2	2	0	2	2	0
Imaging modality*						
CT			21			10
MRI			16			10
X-ray			9			6
Multiple modalities†			6			1
Ultrasound			5			5
Angiography			5			2
Fluoroscopy			2			1
Mammography			2			0
Others			1			0

Data are numbers of questions after removing redundant questions, i.e., overlap between different academic years, (n = 2 for 2019 examination and n = 3 for 2020 examination). Radiology exams consisted of multiple-choice questions with five options.
*Image-based questions (n = 67) are only included, †Images obtained from two or more modalities

Interpretations of images by ChatGPT were evaluated on a 5-point scale by a Board-certified attending radiologist (H.K., with 13 years of experience in radiology practice). The scale determined whether the image modality, findings, and diagnosis were accurately described: 5, very good; 4, good; 3, fair; 2, poor; and 1, very poor. To compare the scores between ChatGPT and the students (i.e., the mean score of actual examinees for the academic years), a one-sample *t*-test was employed. Additionally, the performance of ChatGPT in answering text-only versus image-based questions was analyzed using a chi-squared test. The consistency of ChatGPT's responses across three separate sessions was analyzed using the Fleiss' kappa, with its 95% confidence interval calculated through bootstrap resampling (repeated 1000 times). The percentage agreement was also determined and compared between text-only and image-based questions using the chi-squared test. Statistical analyses were performed using R version 4.1.2 (http://www.R-project.org). A *P*-value of < 0.05 was considered to indicate statistical significance.

In all three consecutive years of radiology exams, ChatGPT scored lower than third-year medical school students: 56 (19/34) vs. 78.5 ± 9.9 (number of students = 148; $P < 0.001$; the ChatGPT score was ranked in the bottom 3rd percentile of the students' scores) for 2018; 65 (22/34) vs. 75.8 ± 9.5 (number of students = 150; $P < 0.001$; the ChatGPT score was ranked in the bottom 16th percentile of the students' scores) for 2019; and 58 (14/24) vs. 74.5 ± 11.3 (number of students = 155; $P < 0.001$; the ChatGPT score was ranked in the bottom 12th percentile of the students' scores) for 2020.

In the pooled analysis, five questions were excluded due to redundancy (overlap between different academic years), with two from the 2019 examination and three from the 2020 examination (Table 1). ChatGPT performed

https://doi.org/10.3348/kjr.2024.0017

Korean Journal of Radiology
KJR

You

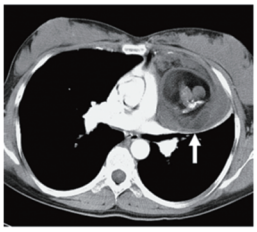
다음 CT영상에서 보이는 질환의 진단명은?
① Inflammatory pseudotumor
② Orbital lymphoma
③ Thyroid orbitopathy and opthalmopathy
④ Cavernous hemangioma
⑤ IgG4 related disease

ChatGPT

{Answer: "③",
Reason: "The CT image shows enlargement of the extraocular muscles with sparing of the tendinous insertions, which is characteristic of thyroid orbitopathy (also known as thyroid eye disease). This condition is often associated with Graves' disease and can lead to proptosis (bulging of the eyes) and other ocular changes.",
Image: "Orbital CT showing enlargement of extraocular muscles consistent with thyroid orbitopathy."}

A

You



45세 여자 환자에서 우연히 발견된 병변의 CT 영상이다. 화살표가 가리키는 이 병변의 진단명은 무엇인가?
① 림프종 (Lymphoma)
② 흉선종 (Thymoma)
③ 기형종 (Teratoma)
④ 신경집종 (Neurilemmoma)
⑤ 기관지낭종 (Bronchogenic cyst)

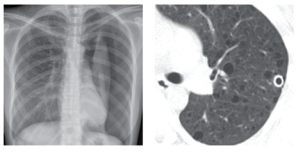
ChatGPT

Based on the provided CT image for a 45-year-old female patient with an incidental finding, and considering the multiple-choice options:

{Answer: "④",
Reason: "The CT image shows a well-circumscribed, homogeneous mass in the posterior mediastinum, which is consistent with the typical location and appearance of a neurilemmoma (also known as schwannoma). Lymphomas (option ①) usually present as multiple, less well-defined masses. Thymomas (option ②) are typically located in the anterior mediastinum. Teratomas (option ③) often contain calcifications, fat, and cystic components. Bronchogenic cysts (option ⑤) are usually located near the trachea or major bronchi and may appear with fluid density.",
Image: "Contrast-enhanced CT scan of the thorax showing a well-circumscribed mass in the posterior mediastinum consistent with a neurilemmoma."}

B

You



33세 여자환자가 하루 전부터 시작된 호흡곤란, 흉통으로 방문하였다. 다음은 내원 당시 촬영한 흉부 X선사진과 흉관 삽입 후 시행한 흉부 CT 영상이다. 진단은 무엇인가?

① 중심소엽성 폐기종 (Centrilobular emphysema)
② 폐 아밀로이드증 (Pulmonary amyloidosis)
③ 랑거한스세포조직구증 (Langerhans cell histiocytosis)
④ 낭성 기관지 확장증 (Cystic bronchiectasis)
⑤ 림프관근종증 (Lymphangiomyomatosis)

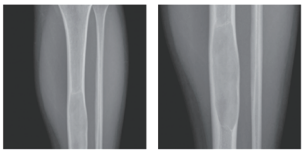
ChatGPT

Based on the provided chest X-ray and CT images for a 33-year-old female patient who presented with dyspnea and chest pain that started a day ago, and considering the multiple-choice options provided:

{Answer: "④",
Reason: "The chest X-ray and CT images after chest tube insertion show multiple cystic spaces, characteristic of cystic bronchiectasis. There is a lack of the diffuse small nodules typical of Langerhans cell histiocytosis (option ③), the septal thickening or pleural involvement common in lymphangiomyomatosis (option ⑤), the nodular or reticular opacities seen in pulmonary amyloidosis (option ②), or the distinct destruction of alveolar walls without obvious fibrosis seen in centrilobular emphysema (option ①).",
Image: "Chest X-ray and CT images showing multiple cystic formations consistent with cystic bronchiectasis."}

C

You



좌측 종아리 X선촬영 및 병변부위 확대영상이다. 진단은?

① Osteosarcoma
② Chondrosarcoma
③ Giant cell tumor
④ Fibrous dysplasia
⑤ Simple bone cyst

ChatGPT

{Answer: "①",
Reason: "The X-ray images show a sunburst pattern and a Codman's triangle which are indicative of an aggressive bone-forming tumor, typically seen in osteosarcoma.",
Image: "X-ray of the lower leg showing a sunburst pattern and periosteal reaction with a Codman's triangle near the knee, suggestive of osteosarcoma."}

D

Fig. 1. ChatGPT's responses to the image-based radiology examination questions. **A-D:** Examination questions were administered to ChatGPT in the same manner as they were to medical school students. The questions were presented in Korean, with some multiple-choice options provided in both Korean and English. In one instance **(A)**, ChatGPT accurately identified thyroid ophthalmopathy from a coronal orbital CT scan. However, in other cases **(B-D)**, ChatGPT was unable to diagnose abnormalities in radiographs and CT scans. The correct diagnoses for these cases were **(B)** teratoma, **(C)** lymphangioleiomyomatosis, and **(D)** fibrous dysplasia.

worse in image-based questions than in text-only questions (52 [35/67] for image-based vs. 85 [17/20] for text-only; $P = 0.10$) (Fig. 1). This trend was not observed in students, who showed similar performance in both types of questions (86.7 [interquartile range: 62.0, 94.0] for image-based vs. 76.4 [interquartile range: 63.8, 88.5] for text-only; $P = 0.42$). For the image interpretation by ChatGPT, 46% (31/67) of interpretations were graded as very good, 6% (4/67) as good, 6% (4/67) as fair, 12% (8/67) as poor, and 30% (20/67) as very poor. The agreement (Fleiss' kappa) among

https://doi.org/10.3348/kjr.2024.0017

ChatGPT's responses in three separate sessions for the same question was 0.70 (95% confidence interval: 0.57, 0.81), with consistent answers in 69% (60/87) of cases. The consistency was marginally higher for text-only questions than for image-based questions, although the difference was not significant (80% [16/20] for text-only vs. 66% [44/67] for image-based; $P = 0.35$).

ChatGPT exhibited below-average performance in the three-year radiology examination, when compared to the performance of the students. This was particularly evident in image-based questions, where 42% of its interpretations were rated as poor or very poor. Given that the examinations featured images with relatively typical findings, ChatGPT's utility in actual radiology practice appears limited.

Our study relied on a single institution's examinations, which may raise questions about the generalizability of our findings. Although customizing ChatGPT with specific lecture materials might improve its performance, this was not explored due to concerns about leaking intellectual lecture materials. Additionally, the use of questions written in Korean could have impacted ChatGPT's performance.

In conclusion, the current version of ChatGPT with vision capabilities showed potential but underperformed in radiological interpretation, suggesting room for improvement for reliable clinical usage.

Conflicts of Interest

H.K. received consulting fees from RadiSen; holds stock and stock options in MEDICAL IP. I.J. holds stock options in MEDICAL IP. C.M.P. holds stock in Promedius; holds stock options in Lunit and Coreline Soft; received research grants from Lunit, Coreline Soft, and HealthHub. S.H.Y. holds stock and stock options in MEDICAL IP.

Ijin Joo and Jung Hoon Kim, who hold respective positions on the Deputy Editor and Editorial Board of the *Korean Journal of Radiology*, were not involved in the editorial evaluation or decision to publish this article. The remaining author has declared no conflicts of interest.

Author Contributions

Conceptualization: Soon Ho Yoon. Data curation: Hyungjin Kim. Formal analysis: Hyungjin Kim. Investigation: Hyungjin Kim. Methodology: Hyungjin Kim. Project administration: Hyungjin Kim, Soon Ho Yoon. Resources: Ijin Joo, Jung Hoon Kim, Chang Min Park. Software: Hyungjin Kim. Supervision: Paul Kim. Writing—original draft: Hyungjin Kim. Writing—review & editing: all authors.

ORCID IDs

Hyungjin Kim
https://orcid.org/0000-0003-0722-0033
Ijin Joo
https://orcid.org/0000-0002-1341-4072
Jung Hoon Kim
https://orcid.org/0000-0002-8090-7758
Chang Min Park
https://orcid.org/0000-0003-1884-3738
Soon Ho Yoon
https://orcid.org/0000-0002-3700-0165

Funding Statement

None

REFERENCES

1. Jung KH. Uncover this tech term: foundation model. *Korean J Radiol* 2023;24:1038-1041
2. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology* 2023;308:e231362
3. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology* 2023;307:e230725
4. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. *Radiology* 2023;307:e230922
5. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology* 2023;307:e230582
6. Rau A, Rau S, Zoeller D, Fink A, Tran H, Wilpert C, et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. *Radiology* 2023;308:e230970

https://doi.org/10.3348/kjr.2024.0017