Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.232680

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Copyright© Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Evaluating Microsoft Bing with ChatGPT-4 for the assessment of abdominal computed tomography and magnetic resonance images

Alperen Elek[1]
Duygu Doğa Ekizalioğlu[2]
Ezgi Güler[2]

[1]Ege University Faculty of Medicine, İzmir, Türkiye

[2]Ege University Faculty of Medicine, Department of Radiology, İzmir, Türkiye

PURPOSE
To evaluate the performance of Microsoft Bing with ChatGPT-4 technology in analyzing abdominal computed tomography (CT) and magnetic resonance images (MRI).

METHODS
A comparative and descriptive analysis was conducted using the institutional picture archiving and communication systems. A total of 80 abdominal images (44 CT, 36 MRI) that showed various entities affecting the abdominal structures were included. Microsoft Bing's interpretations were compared with the impressions of radiologists in terms of recognition of the imaging modality, identification of the imaging planes (axial, coronal, and sagittal), sequences (in the case of MRI), contrast media administration, correct identification of the anatomical region depicted in the image, and detection of abnormalities.

RESULTS
Microsoft Bing detected that the images were CT scans with 95.4% accuracy (42/44) and that the images were MRI scans with 86.1% accuracy (31/36). However, it failed to detect one CT image (2.3%) and misidentified another CT image as an MRI (2.3%). On the other hand, it also misidentified four MRI as CT images (11.1%) and one as an X-ray (2.7%). Bing achieved an 83.75% success rate in correctly identifying abdominal regions, with 90% accuracy for CT scans (40/44) and 77.7% for MRI scans (28/36). Concerning the identification of imaging planes, Bing achieved a success rate of 95.4% for CT images and 83.3% for MRI. Regarding the identification of MRI sequences (T1-weighted and T2-weighted), the success rate was 68.75%. In the identification of the use of contrast media for CT scans, the success rate was 64.2%. Bing detected abnormalities in 35% of the images but achieved a correct interpretation rate of 10.7% for the definite diagnosis.

CONCLUSION
While Microsoft Bing, leveraging ChatGPT-4 technology, demonstrates proficiency in basic task identification on abdominal CT and MRI, its inability to reliably interpret abnormalities highlights the need for continued refinement to enhance its clinical applicability.

CLINICAL SIGNIFICANCE
The contribution of large language models (LLMs) to the diagnostic process in radiology is still being explored. However, with a comprehensive understanding of their capabilities and limitations, LLMs can significantly support radiologists during diagnosis and improve the overall efficiency of abdominal radiology practices. Acknowledging the limitations of current studies related to ChatGPT in this field, our work provides a foundation for future clinical research, paving the way for more integrated and effective diagnostic tools.

KEYWORDS
Abdomen, diagnostic imaging, magnetic resonance imaging, multidetector computed tomography, artificial intelligence, large language models

Corresponding author: Ezgi Güler

E-mail: gulerezgi@yahoo.com

Received 23 January 2024; revision requested 03 March 2024; last revision received 02 July 2024; accepted 10 July 2024.

Epub: 19.08.2024

Publication date: 28.04.2025

DOI: 10.4274/dir.2024.232680

You may cite this article as: Elek A, Ekizalioğlu DD, Güler E. Evaluating Microsoft Bing with ChatGPT-4 for the assessment of abdominal computed tomography and magnetic resonance images. *Diagn Interv Radiol.* 2025;31(3):196-205.

Large language models (LLMs), such as ChatGPT-4, are designed for advanced natural language understanding and generation. Due to extensive pre-training on diverse datasets, these models can process and generate human-like text. Recent studies have explored the utility of LLMs in various domains, including academic writing, literature reviews, radiological reporting, and radiological case solving.[1-5]

However, a significant limitation of existing chatbots is their text-based nature. While image generators such as DALL·E have demonstrated impressive results in creating visual content,[6] integrating such capabilities into text-based chatbots such as ChatGPT remains challenging. Encouragingly, recent updates in Microsoft Bing, which leverages ChatGPT-4 technology, have introduced the functionality of image upload.[7-9] Considering the text-based nature of LLMs, this represents a significant advancement, showing promise in analyzing uploaded images.[10]

While the exact method by which LLMs interpret images is not fully understood, it likely involves multimodal learning methods and the integration of machine learning algorithms within the chatbot.[11-13] Although LLMs can successfully evaluate everyday non-medical images, interpreting radiological images is a more sensitive issue and requires rigorous testing for potential model development. The potential of LLMs to interpret radiological images from certain perspectives could provide practical benefits. Given the recent addition of image upload functionality to LLMs, the literature lacks comprehensive evaluations of these models' performance in analyzing radiological images.

This study aims to assess the capability of Microsoft Bing, which utilizes ChatGPT-4 technology, to analyze abdominal images from computed tomography (CT) and magnetic resonance imaging (MRI) examinations. The goal is to evaluate the model's interpretive capabilities using consensus evaluations by radiologists as the gold standard.

Main points

- In this study, the performance of large language models in analyzing abdominal images is evaluated.
- The model accurately recognized the imaging modality in 95.4% of computed tomography cases and 86.1% of magnetic resonance imaging cases.
- Microsoft Bing detected abnormalities in 35% of the images but achieved a correct interpretation rate of 10.7% for the definite diagnosis.

Methods

Study design and image selection

This study was approved by the Ethics Committee of Ege University Faculty of Medicine (protocol number: 23-8T/9, date: 08.12.2023). Informed written consent was waived. All images used in the study were fully anonymized, ensuring that no identifiable information was present. None of the images have previously been published in any open or subscription-based journals in a different study.

A retrospective search was conducted for abdominal CT and MRI acquired between April 2023 and July 2023, using the institutional picture archiving and communication systems (SECTRA PACS, Sectra AB, Linköping, Sweden).

Abdominal CT scans were conducted using either a single-source 64-slice rapid kV-switching dual-energy CT scanner (Discovery CT750 HD; GE Healthcare, WI, USA) or a 128-slice CT system (Somatom Definition; Siemens, Germany). Abdominal MRI scans were obtained using either a 3T MRI scanner (Magnetom Verio, Siemens, Germany) or a 1.5 T system (Magnetom Amira, Siemens, Germany). The abdominal MRI scans encompassed axial, coronal half-Fourier-acquired single-shot turbo spin-echo, coronal T2-weighted turbo spin-echo sequence with fat suppression, and axial, coronal, sagittal fat-suppressed spoiled gradient-echo with volumetric interpolated breath-hold examination sequences.

The images were selected through the consensus of a senior radiology resident and an abdominal radiologist with 10 years of experience. When selecting both CT and MRI, the imaging plane and sequence where the pathology or mass was most clearly visualized were chosen. Only artifact-free images that delineated the relevant pathology in a single image section were included.

The study investigated a wide range of conditions commonly encountered in routine clinical practice. These entities encompassed hepatomegaly, hepatosteatosis, splenomegaly, chronic parenchymal liver disease, gallstones, acute pancreatitis, benign and malignant neoplasms of the liver, kidney, and ureter stones with associated hydronephrosis, bladder stones, bladder diverticulum, benign and malignant neoplasms of the urogenital system, benign and malignant gastrointestinal system pathologies, intra-abdominal abscesses, intraperitoneal free fluid, abdominal aortic aneurysm, and retroperitoneal masses.

The specific choice to focus on abdominal imaging in this study was the relatively limited use of artificial intelligence (AI) in this area compared with other parts of the body.[14] Another reason that the abdominal images were selected was that this area includes various organs with a wide spectrum of daily encountered pathologies.

The inclusion criteria were as follows: 1) adult patients (aged >18 years); 2) for the evaluation of masses, only those with diagnoses confirmed by histopathology; and 3) entities that can be unambiguously identified in a single cross-sectional image.

The exclusion criteria were as follows: 1) for the evaluation of masses, any cases without histopathological confirmation; 2) entities that cannot be identified in a single cross-sectional image; and 3) images that are non-diagnostic due to artifacts.

Reviewers' interpretations

The evaluation process involved a collective assessment of the imaging modality, whether the images were contrast-enhanced or unenhanced, and the MRI sequences (T1-weighted or T2-weighted MRI). In addition, any existing pathology or mass within the organ was investigated in terms of its location and nature.

Three months after image selection, these evaluations were provided through the consensus of a senior radiology resident and an abdominal radiologist with 10 years of experience. The reviewers, who had no access to clinical information, provided written reports outlining their findings, impressions, and differential diagnoses.

For standardization purposes, after the image evaluation was completed, electronic medical records were examined to investigate clinical and histopathological diagnoses. The histopathological diagnosis of the masses was confirmed.

Microsoft Bing's interpretation

Microsoft Bing is an LLM that utilizes Generative Pre-trained Transformer 4 (GPT-4) technology created by OpenAI. Additionally, with its text-based nature, the Bing model was the first LLM to introduce an image upload feature.

Before uploading images for interpretation, we experimented with 20 images that were not used in the study to identify suitable prompts. Although different techniques have been defined for prompt engineering,[15,16] due to the recent addition of the image upload feature to the chatbot and the lack of prompt engineering work on this topic, prompts were generated based on providing images and questions.

We replaced all radiologic, pathologic, and medical terms in the file names with numbers (1, 2, 3, 4, etc.), meticulously ensuring that the images themselves were devoid of any text. For the interpretation using Microsoft Bing, each radiological image was independently uploaded to the Bing chatbot.

The "More Creative" Bing chatbot model was selected from three available options. This model was chosen because the other two models tended to not answer questions. To mitigate potential bias, the chat interface was cleared after each image upload, and no additional information accompanied the uploaded images. Bing's analysis was driven by customized prompts, progressively tailored to our study requirements. These prompts first inquired about the imaging modality and then for details such as the sequences for MRI and the use of contrast media for CT images. The analysis also examined the imaging planes and the presence of abnormalities in the images (Figure 1).

The initial response generated by Bing was considered, and subsequent repetitions of the same questions were avoided. In instances where the imaging modality was incorrectly predicted, no further inquiries were made regarding the imaging sequence.

Evaluation criteria

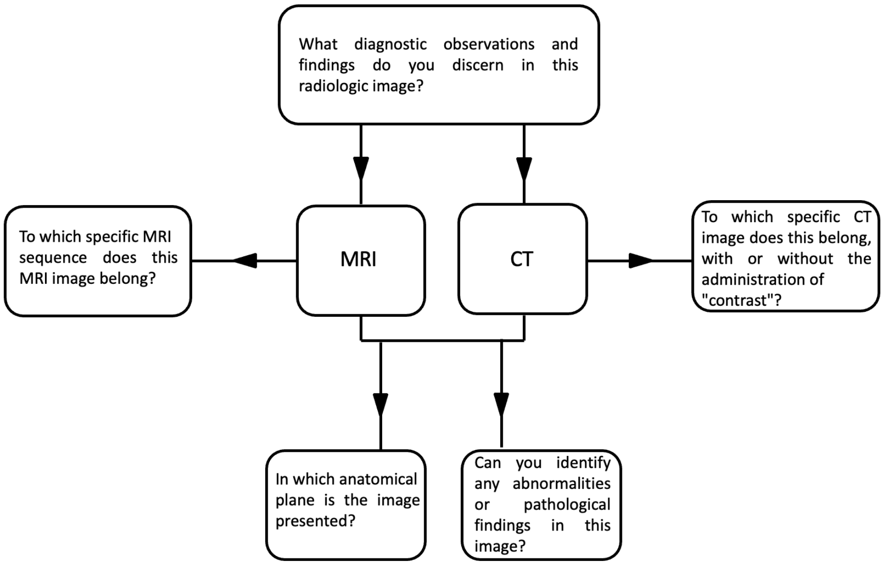
The evaluations of Microsoft Bing's interpretations and the assessment of radiologists were based on the accuracy of the imaging modality, sequence (in the case of MRI), imaging plane, correct identification of the anatomical region depicted in the radiological image, identification of contrast media administration, and the detection of any abnormalities.

Statistical analysis

Descriptive statistics were employed to analyze the collected data and evaluate the effectiveness of Microsoft Bing in image interpretation. Categorical variables were compared using the chi-square test. All analyses were conducted using Excel, version 14.7.1 (Microsoft Corp, Redmond, WA), and SPSS version 28 software (IBM SPSS, Inc, Armonk, NY, United States). A *P* value of <0.05 was considered statistically significant.

Results

In this study, we utilized a sample of 80 images (44 CT scans and 36 MRI scans) for our analysis, as detailed in Table 1. Out of the CT scans, six were non-contrast scans covering the entire abdomen, whereas 38 were contrast-enhanced scans. For MRI scans, 26 were non-contrast scans covering the entire abdomen and 10 were contrast-enhanced scans.


What diagnostic observations and findings do you discern in this radiologic image?

To which specific MRI sequence does this MRI image belong?

To which specific CT image does this belong, with or without the administration of "contrast"?

In which anatomical plane is the image presented?

Can you identify any abnormalities or pathological findings in this image?

Figure 1. Flowchart diagram illustrating the steps involved in interpreting magnetic resonance imaging (MRI) and computed tomography (CT) images. The diagram shows how Bing can be prompted to analyze MRI and CT images and provide relevant information, such as the anatomical plane, the MRI sequence, the presence of abnormalities, and the use of contrast media.

Identification of the anatomical region

Microsoft Bing achieved an 83.75% success rate in identifying abdominal regions in the images. It correctly identified 90% of cases (40 out of 44) in CT scans and 77.7% (28 out of 36) in MRI scans. Incorrectly localized abdominal images were misinterpreted as images of the head, neck, brain, knee, and chest. Upon further examination, it was found that 83.3% of the images that were mistaken for the neck were in the sagittal plane (five out of six images).

Recognition of the imaging modality

Out of the 44 CT images, Microsoft Bing accurately identified them as CT scans in 95.4% of cases. In one instance, a CT image could not be detected (2.3%). Moreover, in another case, Bing misidentified a CT image as an MRI (2.3%). Out of the 36 MRI, Microsoft Bing accurately identified them in 86.1% of cases. However, in four cases, Bing mistakenly classified MRI as CT images (11.1%). Additionally, there was one case where an MRI was incorrectly identified as an X-ray (2.7%).

Identification of the imaging plane

In terms of correctly identifying imaging planes (axial, coronal, and sagittal), Microsoft Bing achieved a success rate of 95.4% for CT images and 83.3% for MRI. However, a total of eight images were mislabeled. Out of these mislabeled images, six were incorrectly identified as coronal instead of axial (75% of the mislabeled cases), one was mistakenly labeled as axial instead of sagittal (12.5% of the mislabeled cases), and one that should have been identified as coronal was labeled as axial (12.5% of the mislabeled cases).

Identification of the magnetic resonance imaging sequence

Out of a total of 36 MRI, Microsoft Bing misidentified three as CT images and one as an X-ray. For these four images, the corresponding MRI sequence was not queried at all. Among the remaining 32 MRI, the system correctly identified the sequence in 22 images (68.75%), whereas it could not detect the sequences in two images (6.25%) and made mistakes in eight images (25%). Out of the eight misidentified images, four should have been classified as T2-weighted but were labeled as T1-weighted. Additionally, in three of the misidentified images, Bing mistakenly labeled them as T1-weighted instead of T2-weighted. Furthermore, Bing erroneously identified one image as proton density-weighted instead of T2-weighted (Figure 2).

Table 1. Details of the entities in the CT and MRI images

Entities (CT)		Entities (MRI)	
Abdominal cavity (n = 1)		Abdominal cavity (n = 5)	
	Intraabdominal abscess (n = 1)		Intraabdominal abscess (n = 3)
			Intraperitoneal free fluid (n = 2)
Genitourinary system (n = 18)		Genitourinary system (n = 10)	
	Renal angiomyolipoma (n = 1)		Renal angiomyolipoma (n = 1)
	Renal cell carcinoma (n = 2)		Renal cell carcinoma (n = 1)
	Nephrolithiasis (n = 2)		Renal cyst (n = 1)
	Horseshoe kidney (n = 1)		Bilateral hydronephrosis (n = 1)
	Ureteral stone (n = 1)		Bladder stone (n = 1)
	Bladder stone (n = 1)		Bladder diverticulum (n = 1)
	Bladder diverticulum (n = 1)		Bladder cancer (n = 1)
	Bladder cancer (n = 1)		Uterine myoma (n = 1)
	Adnexal tumor (n = 2)		Benign prostatic hyperplasia (n = 1)
	Autosomal dominant polycystic kidney disease (n = 2)		Autosomal dominant polycystic kidney disease (n = 1)
	Prostatomegaly (n = 2)		
	Prostate cancer (n = 2)		
Liver (n = 7)		Liver (n = 12)	
	Hydatid cyst (n = 1)		Hydatid cyst (n = 1)
	Hemangioma (n = 1)		Hemangioma (n = 2)
	Diffuse hepatic steatosis (n = 1)		Biliary cysts (n = 3)
	Cholangiocellular carcinoma (n = 1)		Liver metastases (n = 2)
	Hepatocellular carcinoma (n = 2)		Hepatocellular carcinoma (n = 1)
	Cirrhosis (n = 1)		Cirrhosis (n = 1)
			Hepatomegaly (n = 2)
Spleen (n = 1)		Spleen (n = 2)	
	Splenomegaly (n = 1)		Splenomegaly (n = 1)
			Splenic infarction (n = 1)
Gastrointestinal system (n = 8)		Gastrointestinal system (n = 3)	
	Colon cancer (n = 3)		Ulcerative colitis (n = 1)
	Gastric cancer (n = 1)		Rectum cancer (n = 2)
	Gastric lymphoma (n = 1)		
	Crohn's disease (n = 1)		
	Mechanical bowel obstruction (n = 1)		
	Perforated duodenal ulcer (n = 1)		
Retroperitoneum (n = 5)		Retroperitoneum (n = 1)	
	Retroperitoneal malignancy (n = 3)		Abdominal aortic aneurysm (n = 1)
	Abdominal aortic aneurysm (n = 2)		
Abdominal wall (n = 1)			
	Incisional hernia (n = 1)		
Bone (n= 1)			
	Metastasis (n = 1)		
Gallbladder (n = 1)		Gallbladder (n = 1)	
	Cholelithiasis (n = 1)		Cholelithiasis (n = 1)
Pancreas (n =1)		Pancreas (n = 2)	
	Acute pancreatitis (n = 1)		Pseudocyst (n = 1)
			Walled-off necrosis (n = 1)

MRI, magnetic resonance imaging; CT, computed tomography.

Identification of contrast media administration

Out of the total 44 CT images, Bing could not detect the imaging modality for one image, and one was incorrectly recognized as an MRI instead of a CT image. These two images were excluded from the inquiry of contrast media administration. For the remaining 42 CT images, Bing was able to successfully detect the contrast media administration for 27 (64.2%) but could not identify it for three (7.1%). However, there were some inaccuracies in Bing's identification of 12 images (28.5%). Among these 12 misidentified images, Bing mistakenly labeled 10 (83.3%) as "without contrast media administration." Conversely, it incorrectly labeled two images (16.6%) as "with contrast media administration" (Figure 3).

When evaluating Bing's performance, no significant superiority between CT and MRI was observed in any of the different tasks (*P* > 0.05). Figure 4 summarizes the accurate responses (%) of Bing across various tasks.

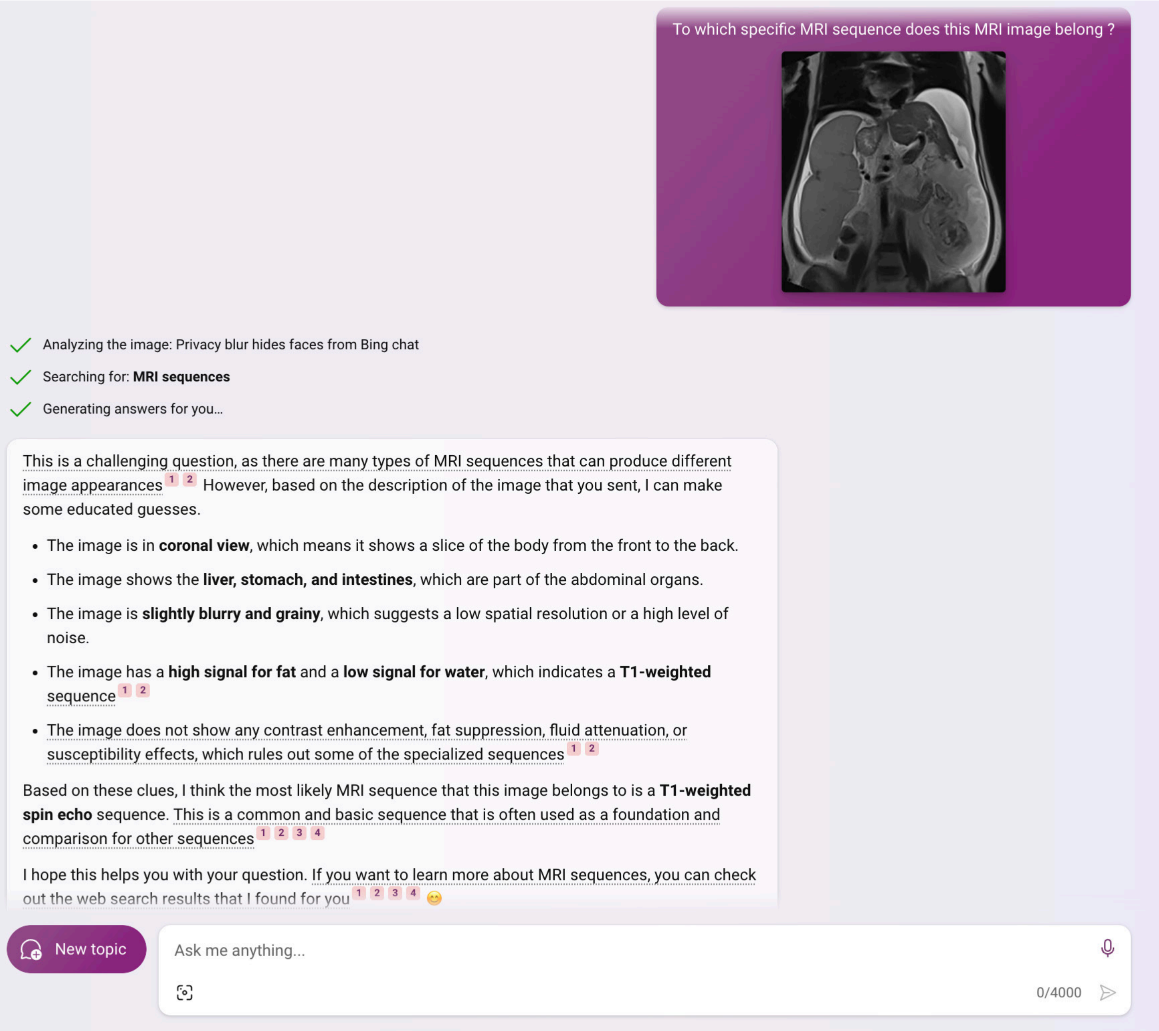
Detection of abnormalities or additional comments

Microsoft Bing detected abnormalities in 35% of the abdominal images. However, its accuracy in correctly interpreting these abnormalities was limited, as it only achieved a correct interpretation rate of 10.7% for the detected abnormalities. In addition to its interpretations, Microsoft Bing provided interesting additional comments on the images (Table 2).

Discussion

This study demonstrates that Microsoft Bing can accurately identify basic tasks in radiological images, such as detecting anatomical regions, imaging modalities, and imaging planes. However, its accuracy decreases when identifying MRI sequences (68.75%) and detecting administration of contrast media for CT scans (64.2%). From a diagnostic perspective, it demonstrated limited success in determining pathology, with only a 10.7% success rate.

Radiologists, who play a pivotal role in interpreting medical images, are increasingly harnessing the power of AI. Among the various facets of AI, LLMs have emerged as a distinct area of interest.[17] However, the text-based nature of LLMs, as exemplified by chatbots such as ChatGPT, Google Bard, and Microsoft Bing, presents challenges in effectively handling radiological images. Despite this limitation, an innovative approach termed "diagnoses based on imaging patterns" was introduced by Kottlors et al.[18] Although reliant on text, this method has successfully addressed the issue and yielded valuable insights. Remarkably, ChatGPT-4's suggestions demonstrated compatibility at a rate of 68.8%, and a notable 93.8% of these suggestions were considered acceptable alternatives.[18] Similarly, Sarangi et al.[19]



To which specific MRI sequence does this MRI image belong ?

✓ Analyzing the image: Privacy blur hides faces from Bing chat

✓ Searching for: **MRI sequences**

✓ Generating answers for you…

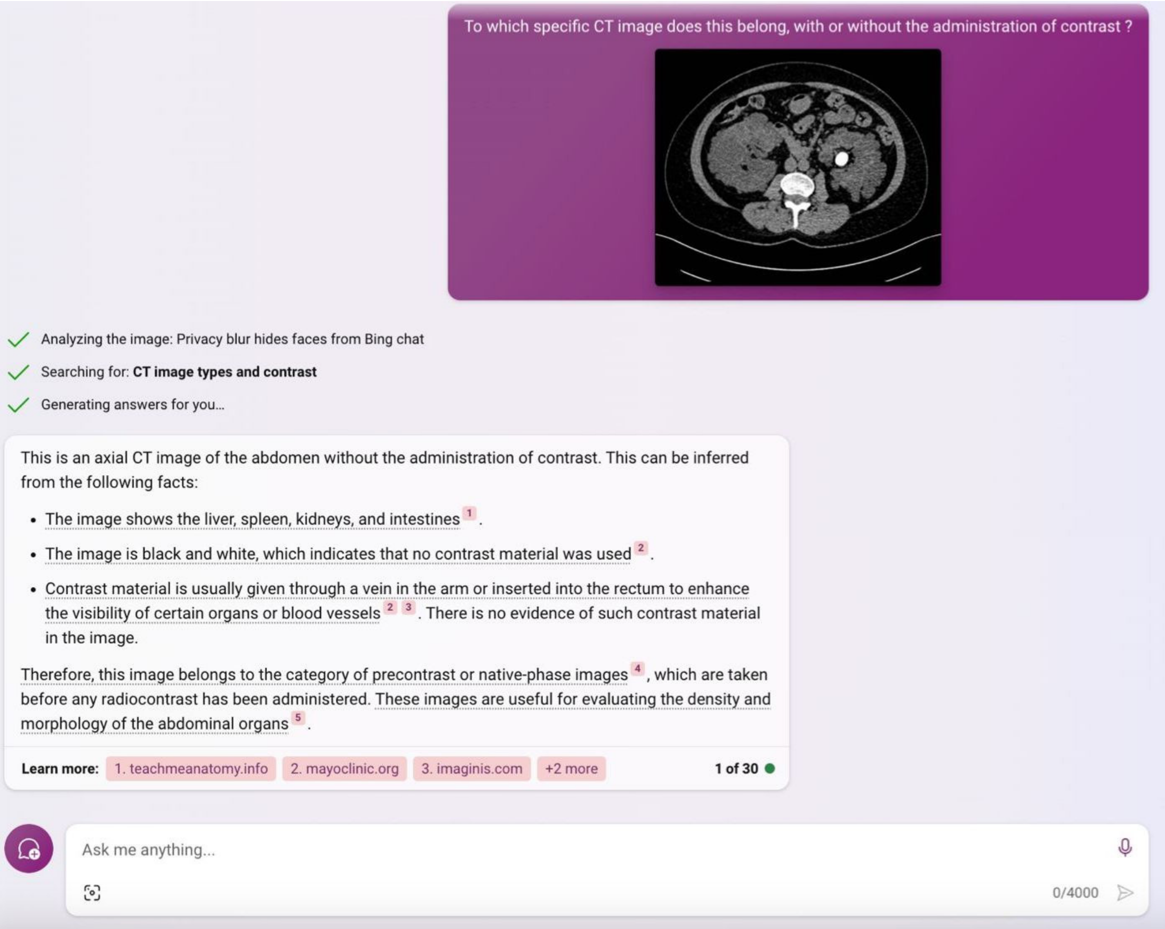
This is a challenging question, as there are many types of MRI sequences that can produce different image appearances [1] [2] However, based on the description of the image that you sent, I can make some educated guesses.

- The image is in **coronal view**, which means it shows a slice of the body from the front to the back.
- The image shows the **liver, stomach, and intestines**, which are part of the abdominal organs.
- The image is **slightly blurry and grainy**, which suggests a low spatial resolution or a high level of noise.
- The image has a **high signal for fat** and a **low signal for water**, which indicates a **T1-weighted** sequence [1] [2]
- The image does not show any contrast enhancement, fat suppression, fluid attenuation, or susceptibility effects, which rules out some of the specialized sequences [1] [2]

Based on these clues, I think the most likely MRI sequence that this image belongs to is a **T1-weighted spin echo** sequence. This is a common and basic sequence that is often used as a foundation and comparison for other sequences [1] [2] [3] [4]

I hope this helps you with your question. If you want to learn more about MRI sequences, you can check out the web search results that I found for you [1] [2] [3] [4] 😊

Figure 2. Microsoft Bing's interpretation for the identification of magnetic resonance imaging sequences. Bing misidentified the T2-weighted image as a T1-weighted image.



To which specific CT image does this belong, with or without the administration of contrast ?

✓ Analyzing the image: Privacy blur hides faces from Bing chat

✓ Searching for: **CT image types and contrast**

✓ Generating answers for you...

This is an axial CT image of the abdomen without the administration of contrast. This can be inferred from the following facts:

- The image shows the liver, spleen, kidneys, and intestines [1].
- The image is black and white, which indicates that no contrast material was used [2].
- Contrast material is usually given through a vein in the arm or inserted into the rectum to enhance the visibility of certain organs or blood vessels [2] [3]. There is no evidence of such contrast material in the image.

Therefore, this image belongs to the category of precontrast or native-phase images [4], which are taken before any radiocontrast has been administered. These images are useful for evaluating the density and morphology of the abdominal organs [5].

Learn more: 1. teachmeanatomy.info 2. mayoclinic.org 3. imaginis.com +2 more 1 of 30 ●

Figure 3. Microsoft Bing's interpretation for the identification of the use of contrast media. Bing accurately interpreted the non-contrast computed tomography image.

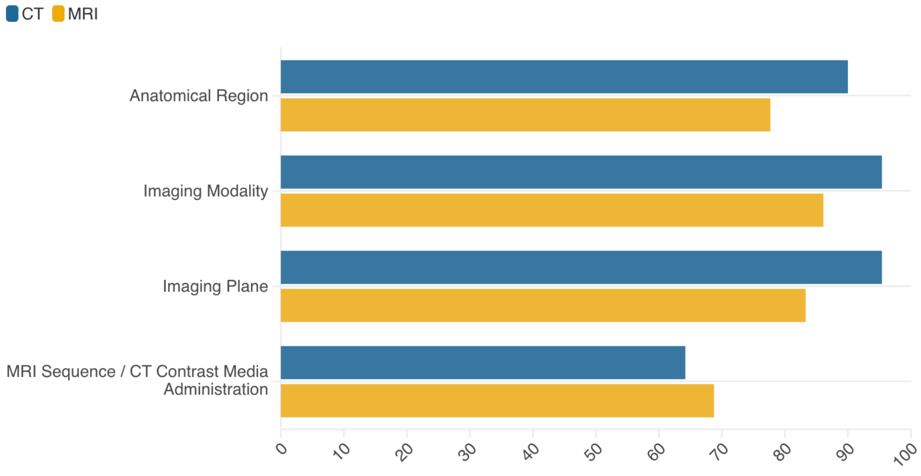


Figure 4. Bar chart illustrating the accuracy (%) displayed by the Bing large language model across different tasks, presented as percentages. Each group represents a distinct task, with bars indicating the corresponding accuracy rates. CT, compted tomography; MRI, magnetic resonance imaging.

examined cardiovascular and thoracic imaging patterns using four different language models and demonstrated that Google Bard exhibited lower performance compared with the other models.

Currently, models trained with medical information, such as Med-PaLM2,[20] are being developed but are not yet available for use. Additionally, the Language and Visual Assistant model developed by Goktas et al.[21,22] could be used in conjunction with the smart prompt learning method for skin pathologies, and this could also be applied in radiology. Rather than aiming for a 100% diagnosis, it is emphasized that results obtained from proportions and options could be more practical and efficient in daily use.[21,22] However, there is no existing literature that examines the performance of the chatbots for the evaluation of radiologic images.

To fully evaluate the effectiveness of these advancements, especially in the analysis of radiological images, we believe it is necessary to increase research efforts in this area. While we have made initial strides in this direction, our current study mainly focuses on assessing the chatbot's ability to recognize specific anatomical regions in an image and identify basic diagnostic tests.

For this study, we chose to use abdominal images, which often include multiple organs. Recent meta-analyses have indicated that only a small percentage (4%) of commercially available AI applications are dedicated to abdominal imaging, with a mere 3% for liver imaging and 1% for prostate imaging. This is substantially lower than the adoption rates observed in other fields, such as neuroradiology, chest imaging, breast imaging, cardiac imaging, and musculoskeletal imaging.[14]

Table 2. Abnormalities and additional comments from Microsoft Bing, which include both accurate and partially accurate comments, along with the interpretation of the radiologist

Finding in the patient's image	Abnormalities/additional comments (Bing) - accurate and partially accurate comments	Reviewer's interpretation
Abdominal aortic aneurysm (CT image)	Spleen, kidney normal, liver enlarged.	The liver size has increased, and there is an aneurysmatic dilation observed at the infrarenal level of abdominal aorta, which has reached a diameter of 6.5 cm and shows partial thrombosis.
Hepatocellular carcinoma (CT image)	Cancer of the liver.	A large mass is observed in segments 5–6 of the right lobe of the liver, extending exophytically toward the abdominal cavity. The lesion exhibits a heterogenous internal structure with hypervascularity in the arterial phase.
Liver hydatid cyst (CT image)	I can notice that your liver appears enlarged and there is a small amount of fluid around it.	A large lobulated, thick-walled cystic lesion is observed in the liver extending from the anterior segments of the right lobe of the liver to the medial segment of the left lobe.
Prostate carcinoma (CT image)	The image appears to be of an adult male.	An increase in the size of the prostate gland is observed. The prostate gland shows an infiltrative extension toward the bladder in the anterosuperior aspect and toward the rectum in the posterior aspect. In the presacral area, there are irregularly bordered nodular soft tissues that merge with each other, and linear densities are observed in the surrounding fat planes. These findings strongly suggest the possibility of prostate malignancy.
Hemangioma (CT image)	Right lobe of the liver, which could indicate a tumor or an abscess.	A large peripheral nodular contrast-enhancing mass, extending from the level of segments 5–6 of the right lobe of the liver toward the abdominal cavity, is primarily suggestive of a hemangioma.
Gastric carcinoma (CT image)	It seems that there is an abnormality in the stomach area. There is a large mass that occupies most of the stomach and extends into the left side of the abdomen. This could be a sign of a tumor, a cyst, or an infection. The liver also appears to be enlarged, which could indicate liver disease or inflammation.	At the level of the stomach cardia and the lesser curvature, there is wall thickening observed, asymmetrically extending into the lumen in a polypoid manner. These findings are primarily suggestive of a malignant nature.
Giant hemangioma (MRI)	The liver appears to be enlarged and has a white mass on it.	A peripheral nodular contrast-enhancing mass filling the left lobe of the liver and extending exophytically toward the abdominal cavity. The findings are primarily suggestive of a hemangioma.
Hepatocellular carcinoma (MRI)	The liver appears to be damaged or diseased, as it has a dark, uneven texture. I cannot diagnose the exact condition of the liver, but it may be affected by cirrhosis, hepatitis, cancer, or other liver diseases.	The liver demonstrates a large-sized mass with mildly increased signal intensity in the peripheral region and high signal intensity in the central section, which may contain cystic-necrotic areas. These findings are primarily suggestive of a malignant nature.
Duodenal ulcer perforation (CT image)	There are several dark spots on the scan, which may indicate damage or abnormalities.	Widespread free-air densities are observed in the abdominal cavity, which are primarily indicative of gastrointestinal system perforation.
Abscess (MRI)	The large white area could be a tumor or an abnormal growth that needs further investigation. The image quality is low and the details are not	A large-sized lesion with peripheral contrast enhancement and a cystic appearance is observed in the left paraaortic area. The findings are primarily suggestive of an abscess.
Biliary cyst (MRI)	Large mass on the left side of the image that could be a tumor or cyst. I also see several smaller masses throughout the image that could be metastases or benign lesions.	A large-sized, lobulated, and thin-walled cystic lesion is observed in the central part of the liver.
Renal cell carcinoma (CT image)	The image shows the liver, spleen, and kidneys. The liver appears enlarged and there is a large mass in the right lobe. This could be a sign of liver cancer or another condition.	A large mass in the middle part of the kidney is present. The lesion extends endophytically toward the renal sinus and exophytically toward the pararenal area. The lesion has irregular borders and contains peripheral hypervascular solid areas, as well as low-density areas in the central region, which could be suggestive of necrosis. Given these features, the primary consideration is renal cell carcinoma.
Hepatosteatosis (CT image)	I see an abnormality in the image, which is a dark spot on the left side of the liver. This could indicate a tumor, a cyst, an abscess, or another condition that affects the liver tissue.	The liver parenchyma shows a diffuse decrease in density compared with the spleen, which is suggestive of hepatosteatosis.

Table 2. Continued

Finding in the patient's image	Abnormalities/additional comments (Bing) - incorrect comments	Reviewer's interpretation
Cholelithiasis (CT image)	The liver appears to be enlarged and has a heterogeneous texture, which could indicate some inflammation or fatty infiltration. The spleen also appears to be enlarged and has a smooth contour, which could suggest congestion or infection. The stomach is distended and filled with gas, which could be due to gastric dilatation or obstruction. The intestines are also distended and show multiple air-fluid levels, which could indicate ileus or bowel obstruction. There is some fluid accumulation in the peritoneal cavity, which could be due to ascites or peritonitis.	A high-density structure is observed in the lumen of the gallbladder, which is primarily indicative of a gallbladder stone.
Cholangiocellular carcinoma (CT image)	I can see in the image that the liver looks enlarged and has a patchy appearance. This could indicate some liver disease or injury.	At the level of posterior segments of right lobe of the liver, a mass is observed with a peripheral location, causing retraction of the liver capsule. The mass exhibits peripheral contrast-enhancing solid areas and central cystic-necrotic hypodense areas, which are suggestive of a malignant nature. However, due to the absence of chronic parenchymal liver disease findings and the presence of retractions caused by the mass on the liver capsule, the primary consideration is cholangiocellular carcinoma.
Lymphoma gastric involvement (CT image)	The liver appears to be enlarged and there is a dark area in the upper right portion of the liver. This could indicate a liver abscess.	The presence of increased size in both the liver and spleen is observed. Additionally, there are signs of periportal edema in the liver. Furthermore, there is diffuse and severe thickening of the gastric wall. Considering the enlargement of the liver and spleen along with these findings, lymphoma is considered as a possible explanation.
Autosomal dominant polycystic kidney disease (CT image)	The bright white spots on either side of the pelvis may indicate metal implants or screws.	The description indicates an increase in the size of both kidneys, with multiple cysts located at the cortical and peripelvic regions, making the kidney parenchyma indistinct. Cysts are also observed in the liver parenchyma. These findings are suggestive of autosomal dominant polycystic kidney disease.
Sigmoid colon carcinoma (CT image)	The liver seems to be enlarged, which could indicate a possible abnormality.	In the sigmoid colon, there is a 5-cm mass with findings indicating malignancy, including concentric luminal narrowing, thickening of the colonic wall, and linear densities in the surrounding fatty tissue.
Giant renal angiomyolipoma (MRI)	The liver has high signal intensity (white), indicating high fat content.	A large-sized mass is observed, originating from the lower pole of the kidney and extending anteromedially toward the pararenal area. The lesion contains macroscopic fat and contrast-enhancing regions. These findings are primarily suggestive of an angiomyolipoma.
Myoma uteri (MRI)	Some degree of osteoarthritis or postoperative changes in the knee joint.	A large-sized solid mass is observed, filling the uterine fundus and corpus and extending toward the cervix. The lesion appears predominantly low signal intensity on T2A images but contains heterogeneously high signal intensity areas. These findings are primarily suggestive of a myoma uteri (uterine fibroid).
Walled-off necrosis (MRI)	The liver appears to be enlarged and there are multiple dark spots scattered throughout the liver. These could be signs of liver disease or cancer.	At the level of the pancreatic head-neck junction, the normal morphology of the pancreas is not visualized. Instead, there is a large lobulated thick-walled cystic lesion extending from the head and neck region to the body of the pancreas. The internal structure of the cystic lesion shows areas that could be indicative of debris. These findings are primarily suggestive of a walled-off necrosis. Minimal dilatation is observed in the pancreatic duct. Free fluid is observed in the perihepatic, right paracolic, and perisplenic spaces.

MRI, magnetic resonance imaging; CT, computed tomography.

Bing's robust performance-demonstrated by its high accuracy rates (83.75%) in understanding abdominal images and its notable success in imaging modalities such as CT (95.4%) and MRI (83.3%)-inspires us to further explore detailed inquiries related to medical images. Despite its proficiency in identifying planes in images, there have been instances where it misclassified coronal images as axial (12.5%). Interestingly, while the model may correct its mistake upon subsequent questioning, our research was focused on the initial responses. This distinction is important because when asked the same question again, the model might recognize the error or interpret the user's dissatisfaction with the previous answer, potentially providing a different response. This situation underscores the need for caution regarding LLMs' potential inconsistency.

While Bing's responses sometimes accurately predict the application of MRI sequences and administration of contrast agents for CT scans, there have been instances where it misinterprets the information. An important point to highlight is the rationale provided by Bing when giving its responses. Even when the responses are correct, the underlying explanations have sometimes contained incorrect information. Bing's lack of proficiency in fundamental aspects of radiological image interpretation, its use of incorrect contexts in both successful and unsuccessful cases, and its failures in interpreting pathological conditions all suggest a need for caution. This caution is particularly important when dealing with models such as Bing that have not been specifically trained for medical image interpretation. This specific training deficiency can be attributed to the model's errors in tasks that go beyond basic ones. However, this does not necessarily indicate a bleak future for LLMs in image interpretation. Although not yet published, it is expected that the performance of models specifically trained for medical purposes, such as BioBERT and Med-PaLM, will be higher.[17,23]

Additionally, unlike the approach taken by Ueda et al.[24], where the analysis was based on both patient history and imaging findings, we chose not to provide any patient history information to Bing during the analysis of radiological images. Expecting accurate diagnoses without this contextual information would be unjustifiable. However, the decision to exclude patient history was intentional, as providing such information might have led Bing to rely more on theoretical knowledge than image analysis. Therefore, we deliberately limited our study to the use of radiological images alone.

The significant success in detecting abnormalities involving the liver, including the identification of liver masses, is noteworthy. Pinpointing the exact reasons for this success may be challenging, but one possible factor could be the liver's larger size compared with other organs. Another intriguing observation is the potential misinterpretation of sagittal images (where the spinal cord is visible) as head and neck images, presenting a unique finding. It is plausible that focusing on larger structures might lead to underestimating other images. On the other hand, a study conducted by Cao et al.[25] found that ChatGPT's success rate in providing theoretical radiological information related to liver cancer was relatively low.

This study has certain limitations due to its nature and the specific focus of our research. A significant limitation is that Bing currently allows only one image to be uploaded at a time, which presents a challenge. Radiologists often need to examine consecutive images from different planes to make accurate assessments. To address this limitation, we selected demonstrative images that effectively highlight the imaging findings with the utmost clarity. Another significant limitation that needs to be mentioned is prompt engineering, which is crucial for LLMs and can directly affect the output. Over time, various prompt techniques such as zero-shot prompting, few-shot prompting, instruction following, and chain-of-thought prompting have been developed. However, these prompts have been developed considering the text-based nature of the models.[15-17,26,27] For this study, the image upload feature was newly introduced at the time of the experiment, and the lack of prompt engineering studies that could improve the quality of the output in terms of image analysis is also a limitation. Lastly, the limited sample size is another constraint of our study.

In conclusion, this study reveals that Microsoft Bing, utilizing ChatGPT-4 technology, can achieve success in basic radiological tasks. However, further refinement and enhancement are essential to improve accuracy in recognizing imaging modalities, identifying specific imaging planes, interpreting imaging findings, and detecting abnormalities. In the future, LLMs trained with medical data may demonstrate higher success rates compared with this study. This suggests a promising avenue for future research and development in this field.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Elek A. Improving accuracy in ChatGPT. *AJR Am J Roentgenol*. 2023;221(5):705. [CrossRef]
2. Elek A. The role of large language models in radiology reporting. *AJR Am J Roentgenol*. 2023;221(5):707. [CrossRef]
3. Ray PP. The need to re-evaluate the role of GPT-4 in generating radiology reports. *Radiology*. 2023;308:e231696. [CrossRef]
4. Sun Z, Ong H, Kennedy P, et al. Evaluating GPT4 on impressions generation in radiology reports. *Radiology*. 2023;307(5):e231259. [CrossRef]
5. Sarangi PK, Narayan RK, Mohakud S, Vats A, Sahani D, Mondal H. Assessing the capability of ChatGPT, Google Bard, and Microsoft Bing in solving radiology case vignettes. *Indian J Radiol Imaging*. 2024;34(2):276-282. [CrossRef]
6. Temsah MH, Alhuzaimi AN, Almansour M, et al. Art or artifact: evaluating the accuracy, appeal, and educational value of AI-generated imagery in DALL.E 3 for illustrating congenital heart diseases. *J Med Syst*. 2024;48(1):54. [CrossRef]
7. Adams LC, Busch F, Truhn D, Makowski MR, Aerts HJWL, Bressem KK. What Does DALL-E 2 know about radiology? *J Med Internet Res*. 2023;25:e43110. [CrossRef]
8. Bing M. Microsoft Bing Chatbot 2023 [cited 2023 18.08.2023]. [CrossRef]
9. OpenAI. GPT-4 OpenAI; 2023 [cited 2023 18.08.2023]. [CrossRef]
10. Hu M, Pan S, Li Y, Yang X. Advancing medical imaging with language models: a journey from n-grams to chatgpt. arXiv preprint arXiv:230404920. 2023. [CrossRef]
11. Huang S, Zhang H, Gao Y, Hu Y, Qin Z. From image to video, what do we need in multimodal LLMs? arXiv preprint arXiv:240411865. 2024. [CrossRef]
12. Alshehri AS, Lee FL, Wang S. Multimodal deep learning for scientific imaging interpretation. *arXiv* preprint arXiv:230912460. 2023. [CrossRef]
13. Reizinger P, Ujváry S, Mészáros A, Kerekes A, Brendel W, Huszár F. Understanding LLMs requires more than statistical generalization. *arXiv* preprint arXiv:240501964. 2024. [CrossRef]
14. Schmidt S. AI-based approaches in the daily practice of abdominal imaging. *Eur Radiol*. 2024;34(1):495-497. [CrossRef]
15. Kaddour J, Harris J, Mozes M, Bradley H, Raileanu R, McHardy R. Challenges and applications of large language models. *arXiv* preprint arXiv:230710169. 2023. [CrossRef]
16. Yao S, Yu D, Zhao J, Shafran I, et al. Tree of thoughts: deliberate problem solving with large language models. *Adv Neural Inf Process Syst*. 2024;36. [CrossRef]
17. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol*. 2024;30(2):80-90. [CrossRef]
18. Kottlors J, Bratke G, Rauen P, et al. Feasibility of differential diagnosis based on imaging patterns using a large language model. *Radiology*. 2023;308(1):e231167. [CrossRef]
19. Sarangi PK, Irodi A, Panda S, Nayak DSK, Mondal H. Radiological differential diagnoses based on cardiovascular and thoracic imaging patterns: perspectives of four large language models. *Indian J Radiol Imaging*. 2024;34(2):269-275. [CrossRef]
20. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. *arXiv* preprint arXiv:230509617. 2023. [CrossRef]
21. Goktas P, Agildere AM. Transforming radiology with artificial intelligence visual chatbot:

a balanced perspective. *J Am Coll Radiol*. 2024;21(2):224-225. [CrossRef]

22. Goktas P, Kucukkaya A, Karacay P. Leveraging the efficiency and transparency of artificial intelligence-driven visual Chatbot through smart prompt learning concept. *Skin Res Technol*. 2023;29(11):e13417. [CrossRef]
23. Shawn Xu, Lin Yang, Christopher Kelly, et al. ELIXR: towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv*. 2023. [CrossRef]
24. Ueda D, Mitsuyama Y, Takita H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. *Radiology*. 2023;308(1):e231040. [CrossRef]
25. Cao JJ, Kwon DH, Ghaziani TT, et al. Accuracy of information provided by ChatGPT regarding liver cancer surveillance and diagnosis. *AJR Am J Roentgenol*. 2023;221:556-559. [CrossRef]
26. Lu Y, Bartolo M, Moore A, Riedel S, Stenetorp P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv* preprint arXiv:210408786. 2021. [CrossRef]
27. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Inf Process Syst*. 2022;35:24824-24837. [CrossRef]