

ChatGPT Passes Radiology Board Exam

Released: May 16, 2023

At A Glance

- ChatGPT (GPT-4) passed a radiology board-style exam, correctly answering 81% of questions.
- GPT-4 demonstrated marked improvement in advanced reasoning.
- ChatGPT used confident language consistently, even when incorrect, making it dangerous as a sole information source, especially for novices who may not recognize confident responses as inaccurate.

OAK BROOK, Ill. — The latest version of ChatGPT passed a radiology board-style exam, highlighting the potential of large language models but also revealing limitations that hinder reliability, according to two new research studies published in *Radiology*, a journal of the Radiological Society of North America (RSNA).

ChatGPT is an artificial intelligence (AI) chatbot that uses a deep learning model to recognize patterns and relationships between words in its vast training data to generate human-like responses based on a prompt. But since there is no source of truth in its training data, the tool can generate responses that are factually incorrect.

"The use of large language models like ChatGPT is exploding and only going to increase," said lead author Rajesh Bhayana, M.D., FRCPC, an abdominal radiologist and technology lead at University Medical Imaging Toronto, Toronto General Hospital in Toronto, Canada. "Our research provides insight into ChatGPT's performance in a radiology context, highlighting the incredible potential of large language models, along with the current limitations that make it unreliable."

ChatGPT was recently named the fastest growing consumer application in history, and similar chatbots are being incorporated into popular search engines like Google and Bing that physicians and patients use to search for medical information, Dr. Bhayana noted.

To assess its performance on radiology board exam questions and explore strengths and limitations, Dr. Bhayana and colleagues first tested ChatGPT based on GPT-3.5, currently the most commonly used version. The researchers used 150 multiple-choice questions designed to match the style, content and difficulty of the Canadian Royal College and American Board of Radiology exams.

[download photo](#)



Rajesh Bhayana, M.D., FRCPC

The questions did not include images and were grouped by question type to gain insight into performance: lower-order (knowledge recall, basic understanding) and higher-order (apply, analyze, synthesize) thinking. The higher-order thinking questions were further subclassified by type (description of imaging findings, clinical management, calculation and classification, disease associations).

The performance of ChatGPT was evaluated overall and by question type and topic. Confidence of language in responses was also assessed.

The researchers found that ChatGPT based on GPT-3.5 answered 69% of questions correctly (104 of 150), near the passing grade of 70% used by the Royal College in Canada. The model performed relatively well on questions requiring lower-order thinking (84%, 51 of 61), but struggled with questions involving higher-order thinking (60%, 53 of 89). More specifically, it struggled with higher-order questions involving description of imaging findings (61%, 28 of 46), calculation and classification (25%, 2 of 8), and application of concepts (30%, 3 of 10). Its poor performance on higher-order thinking questions was not surprising given its lack of radiology-specific pretraining.

GPT-4 was released in March 2023 in limited form to paid users, specifically claiming to have improved advanced reasoning capabilities over GPT-3.5.

In a follow-up study, GPT-4 answered 81% (121 of 150) of the same questions correctly, outperforming GPT-3.5 and exceeding the passing threshold of 70%. GPT-4 performed much better than GPT-3.5 on higher-order thinking questions (81%), more specifically those involving description of imaging findings (85%) and application of concepts (90%).

The findings suggest that GPT-4's claimed improved advanced reasoning capabilities translate to enhanced performance in a radiology context. They also suggest improved contextual understanding of radiology-specific terminology, including imaging descriptions, which is critical to enable future downstream applications.

"Our study demonstrates an impressive improvement in performance of ChatGPT in radiology over a short time period, highlighting the growing potential of large language models in this context," Dr. Bhayana said.

GPT-4 showed no improvement on lower-order thinking questions (80% vs 84%) and answered 12 questions incorrectly that GPT-3.5 answered correctly, raising questions related to its reliability for information gathering.

"We were initially surprised by ChatGPT's accurate and confident answers to some challenging radiology questions, but then equally surprised by some very illogical and inaccurate assertions," Dr. Bhayana said. "Of course, given how these models work, the inaccurate responses should not be particularly surprising."

ChatGPT's dangerous tendency to produce inaccurate responses, termed hallucinations, is less frequent in GPT-4 but still limits usability in medical

education and practice at present.

Both studies showed that ChatGPT used confident language consistently, even when incorrect. This is particularly dangerous if solely relied on for information, Dr. Bhayana notes, especially for novices who may not recognize confident incorrect responses as inaccurate.

"To me, this is its biggest limitation. At present, ChatGPT is best used to spark ideas, help start the medical writing process and in data summarization. If used for quick information recall, it always needs to be fact-checked," Dr. Bhayana said.

"Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations" and "GPT-4 in Radiology: Improvements in Advanced Reasoning." Collaborating with Dr. Bhayana were Satheesh Krishna, M.D., and Robert R. Bleakney, M.D.

In 2023, *Radiology* is celebrating its 100th anniversary with 12 [centennial issues](#), highlighting Radiology's legacy of publishing exceptional and practical science to improve patient care.

Radiology is edited by Linda Moy, M.D., New York University, New York, N.Y., and owned and published by the Radiological Society of North America, Inc. (<https://pubs.rsna.org/journal/radiology>)

RSNA is an association of radiologists, radiation oncologists, medical physicists and related scientists promoting excellence in patient care and health care delivery through education, research and technologic innovation. The Society is based in Oak Brook, Illinois. ([RSNA.org](https://www.rsna.org))

For patient-friendly information on professions in radiology, visit [RadiologyInfo.org](https://www.radiologyinfo.org).

Images (JPG, TIF):

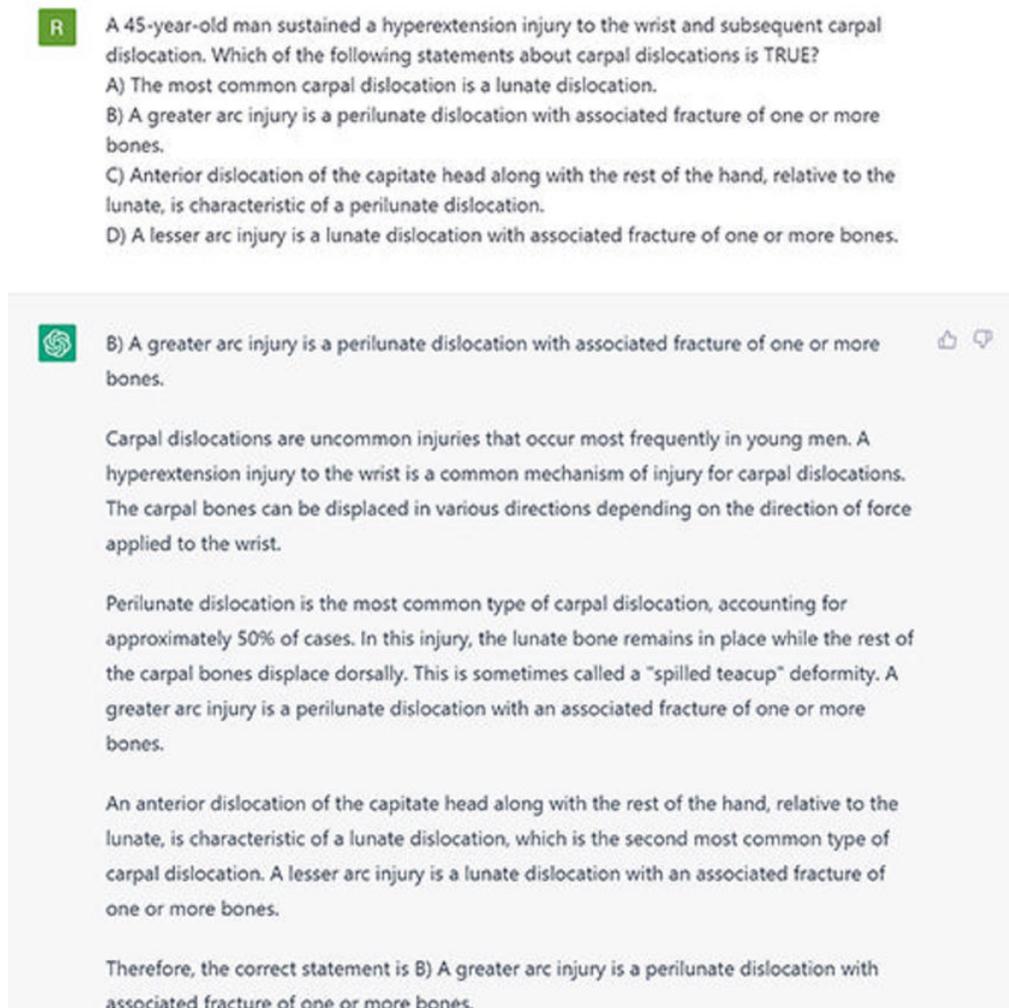


Figure 1. ChatGPT response to a question on carpal dislocations, primarily assessing lower-order thinking (knowledge recall, basic understanding). ChatGPT picked the correct answer (option B). However, its explanation included inaccuracies: perilunate dislocations are most common and are two to three times more common than lunate dislocations, not 50% of cases; anterior dislocation of the capitate is seen with perilunate dislocation, but characteristically the dislocation is posterior; lesser arc injuries are pure ligamentous disruptions without fracture. [High-res \(TIF\) version](#)

- R** A 64-year-old man undergoes a chest CT scan while in hospital for persistent pneumonia. A 3-cm adrenal nodule is incidentally noted, and a subsequent adrenal washout study is ordered. On the non-contrast imaging series, region of interest (ROI) measurement of the nodule provides a reading of 27 HU. On the contrast-enhanced series that follows, this measurement rises to 88 HU. Delayed images through the adrenal glands at 15 minutes show a reading of 45 HU. What is the absolute washout for this lesion?
- A) 31%
 - B) 49%
 - C) 59%
 - D) 70%

 The formula for absolute percentage washout (APW) is: 👍 🗨️

$(APW) = [(HU \text{ unenhanced} - HU \text{ delayed}) / (HU \text{ unenhanced} - HU \text{ enhanced})] \times 100\%$

Using the values provided in the question, we can calculate the absolute washout for the lesion as follows:

$(APW) = [(27 - 45) / (27 - 88)] \times 100\%$
 $(APW) = (-18) / (-61) \times 100\%$
 $(APW) = 0.295 \times 100\%$
 $(APW) = 29.5\%$

Therefore, the absolute washout for this lesion is 29.5%, which rounds to 31%, so the correct answer is A) 31%.

[🔄 Regenerate response](#)

Figure 2. ChatGPT response to a higher-order thinking question involving calculation of absolute washout in an adrenal nodule. The ChatGPT response included an inaccurate absolute washout formula. Subsequent calculation and answer (option A) were incorrect. The correct formula for absolute washout is as follows: (portal venous phase attenuation – delayed phase attenuation)/(portal venous phase attenuation – nonenhanced attenuation) x 100. When the correct values from the question are included, the calculation is as follows: (88 – 45)/(88 – 27) x 100. The absolute washout in the nodule described is 70% (correct answer is D). [High-res \(TIF\) version](#)

R

What is the TI-RADS score of the following lesion: solid, hypoechoic, macrocalcifications present, no microcalcifications, wider than it is tall, and no other suspicious features?

- A) TI-RADS 2
- B) TI-RADS 3
- C) TI-RADS 4
- D) TI-RADS 5



Based on the provided characteristics of the thyroid nodule, the TI-RADS score would be B) TI-RADS 3.



The Thyroid Imaging Reporting and Data System (TI-RADS) is a risk stratification system used for thyroid nodules seen on ultrasound. The system assigns a score based on the presence and degree of various suspicious features, with higher scores indicating a higher likelihood of malignancy.

In this case, the nodule is solid, hypoechoic, wider than it is tall, and has macrocalcifications. While macrocalcifications can be a suspicious feature, the absence of microcalcifications and other suspicious features suggest a low risk of malignancy. Therefore, the nodule would receive a TI-RADS score of 3, which indicates a low to intermediate risk of malignancy and suggests that the nodule should be monitored with follow-up ultrasound rather than biopsied.

Figure 3. ChatGPT response to a classification question involving the Thyroid Imaging Reporting and Data System (TI-RADS). The model selected the incorrect answer (option B, TI-RADS 3). Since the lesion is solid (2 points), hypoechoic (2 points), and has macrocalcifications (1 point), this corresponds to a TI-RADS 4 lesion (correct answer is option C).

[High-res \(TIF\) version](#)

R MRI images through the uterus demonstrate an appearance of two endometrial cavities separated by myometrial tissue. The internal indentation has an acute angle and the external fundal contour is preserved without indentation. Which of the following does this best describe?
A) Arcuate
B) Bicornuate
C) Didelphys
D) Septate

Based on the description provided, the most likely diagnosis is D) Septate uterus.

A septate uterus is a congenital anomaly in which the uterus has a septum that divides the uterine cavity into two, resulting in two endometrial cavities separated by myometrial tissue. On MRI images, a septate uterus may appear as an internal indentation with an acute angle and a preserved external fundal contour without indentation.

Arcuate uterus is a different congenital anomaly in which there is a mild degree of uterine septation, resulting in a shallow internal indentation with a rounded contour of the fundus.

Bicornuate uterus is another congenital anomaly in which the uterus has two horns that are separated by a septum. The septum may be partial or complete and can extend from the fundus to the cervix. On MRI images, a bicornuate uterus may appear as two endometrial cavities separated by myometrial tissue, with an internal indentation that has an acute angle and a preserved external fundal contour without indentation.

Didelphys uterus is a congenital anomaly in which there are two completely separate uteri and cervixes, with no communication between them.

Therefore, based on the provided information, the most likely diagnosis is a septate uterus (option D).

Figure 4. ChatGPT response to a question predominantly featuring a description of imaging findings. The question describes the classic appearance of a septate uterus. ChatGPT selected the correct answer (option D). The explanations are largely accurate, but its description of bicornuate uterus is inaccurate. Specifically, it indicates that the bicornuate uterus has a “preserved external fundal contour without indentation.” Bicornuate uterus is best differentiated from septate uterus by identifying an external fundal indentation greater than 1 cm.

[High-res \(TIF\) version](#)

Resources:

[Editorial](#)

[Study abstract](#)