Vision-based ChatGPT shows deficits interpreting radiologic images

September 3 2024

Credit: CC0 Public Domain

Researchers evaluating the performance of ChatGPT-4 Vision found that the model performed well on text-based radiology exam questions but struggled to answer image-related questions accurately. The study's

results were published in *Radiology*.

Chat GPT-4 Vision is the first version of the large language model that can interpret both text and images.

"ChatGPT-4 has shown promise for assisting radiologists in tasks such as simplifying patient-facing radiology reports and identifying the appropriate protocol for imaging exams," said Chad Klochko, M.D., musculoskeletal radiologist and artificial intelligence (AI) researcher at Henry Ford Health in Detroit, Michigan. "With image processing capabilities, GPT-4 Vision allows for new potential applications in radiology."

For the study, Dr. Klochko's research team used retired questions from the American College of Radiology's Diagnostic Radiology In-Training Examinations, a series of tests used to benchmark the progress of radiology residents. After excluding duplicates, the researchers used 377 questions across 13 domains, including 195 questions that were text-only and 182 that contained an image.

GPT-4 Vision answered 246 of the 377 questions correctly, achieving an overall score of 65.3%. The model correctly answered 81.5% (159) of the 195 text-only queries and 47.8% (87) of the 182 questions with images.

"The 81.5% accuracy for text-only questions mirrors the performance of the model's predecessor," he said. "This consistency on text-based questions may suggest that the model has a degree of textual understanding in radiology."

Genitourinary radiology was the only subspecialty for which GPT-4 Vision performed better on questions with images (67%, or 10 of 15) than text-only questions (57%, or four of seven). The model performed

better on text-only questions in all other subspecialties.

The model performed best on image-based questions in the chest and genitourinary subspecialties, correctly answering 69% and 67% of the image-containing questions, respectively. The model performed lowest on image-containing questions in the nuclear medicine domain, correctly answering only two of 10 questions.

The study also evaluated the impact of various prompts on the performance of GPT-4 Vision.

- Original: You are taking a radiology board exam. Images of the questions will be uploaded. Choose the correct answer for each question.

- Basic: Choose the single best answer in the following retired radiology board exam question.

- Short instruction: This is a retired radiology board exam question to gauge your medical knowledge. Choose the single best answer letter and do not provide any reasoning for your answer.

- Long instruction: You are a board-certified diagnostic radiologist taking an examination. Evaluate each question carefully and if the question additionally contains an image, please evaluate the image carefully in order to answer the question. Your response must include a single best answer choice. Failure to provide an answer choice will count as incorrect.

- Chain of thought: You are taking a retired board exam for research purposes. Given the provided image, think step by step for the provided question.

Although the model correctly answered 183 of 265 questions with a basic prompt, it declined to answer 120 questions, most of which contained an image.

"The phenomenon of declining to answer questions was something we hadn't seen in our initial exploration of the model," Dr. Klochko said.

The short instruction prompt yielded the lowest accuracy (62.6%).

On text-based questions, chain-of-thought prompting outperformed long instruction by 6.1%, basic by 6.8%, and original prompting style by 8.9%. There was no evidence to suggest performance differences between any two prompts on image-based questions.

"Our study showed evidence of hallucinatory responses when interpreting image findings," Dr. Klochko said. "We noted an alarming tendency for the model to provide correct diagnoses based on incorrect image interpretations, which could have significant clinical implications."

Dr. Klochko said his study's findings underscore the need for more specialized and rigorous evaluation methods to assess large language model performance in radiology tasks.

"Given the current challenges in accurately interpreting key radiologic images and the tendency for hallucinatory responses, the applicability of GPT-4 Vision in information-critical fields such as radiology is limited in its current state," he said.

More information: Performance of GPT-4 with Vision on Text- and Image-based ACR Diagnostic Radiology In-Training Examination Questions, *Radiology* (2024).

Provided by Radiological Society of North America