A Radiological Turing Test: Can Large Language Models Answer Board-Style Questions?

Thomas P. Reith MD, Michael P. D'Alessandro MD

Department of Radiology, University of Iowa

Background

- For an AI system to replace a radiologist, it must first pass the ABR Qualifying (Core) exam and the USMLE Step 3 exam.
- ChatGPT (OpenAI Inc[1]) and Bing Chat (Microsoft[2]) are newly-released implementations of a type of AI called a large language model. Both products were trained on billions of words and can produce human-like text in response to user input.
- At the time of this study, **ChatGPT was powered by GPT-3.5**, trained on online text existing before Q4 2021. **Bing Chat is powered by GPT-4**, a newer, more powerful release with real-time access to the Internet.

Materials/Methods

- Sample USMLE Step 3 questions were collected from the USMLE website. Sample ABR Core Exam questions were collected from publicly-available ABR materials, the Aunt Minnie Board Review tool, and the RadPrimer and Calisi Physics question banks.
- Questions containing visual media were discarded. The remainder were submitted verbatim to ChatGPT and Bing Chat.

Results

- After exclusions, 136 radiology questions and 124 USMLE Step 3 questions were submitted to both ChatGPT and Bing Chat.
- Percentages of correctly answered questions are listed in **Table 1**.
- Both models' superior performance on USMLE Step 3 questions was statistically significant, as was Bing Chat's superior performance over ChatGPT.

Conclusions

- Since the typical USMLE Step 3 passing threshold is 60% correct,[3] **both models met or exceeded passing level on a general medical licensing exam.**
- Both models fared worse on radiology questions, but **Bing Chat may already be at or close to passing level for written radiology board questions.**
- Since OpenAI's new implementation of GPT-4 has the ability to recognize image inputs, it seems plausible that **large language models may soon have the potential to pass the ABR Core exam.**

Table 1: Percentage of questions answered correctly by ChatGPT and Bing Chat

	ChatGPT	Bing Chat	
USMLE Step 3	62.9%	86.3%	$p = .0035$*
Radiology	44.9%	67.6%	$p < .001$*
	$p < .001$*	$p < .001$*	

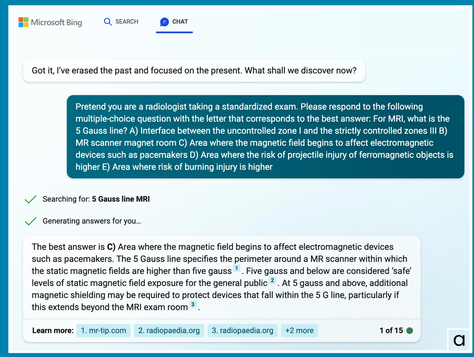
*Statistical significance was evaluated using Chi-Square tests.

References

1. www.openai.com/blog/chatgpt
2. www.bing.com/new
3. Scoring and Score Reporting. United States Medical Licensing Examination. https://www.usmle.org/bulletin-information/scoring-and-score-reporting

ChatGPT and Bing Chat performed at or above passing level on USMLE Step 3 questions, but fared worse on radiology questions.

Figure 2: Screenshot of ChatGPT (a) correctly answering a USMLE Step 3 question and (b) incorrectly answering a radiology question. In the latter case, the correct answer is choice D.



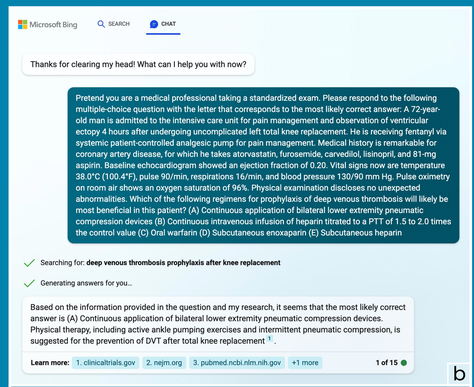


Figure 3: Screenshot of Bing Chat (a) correctly answering a radiology question and (b) incorrectly answering a USMLE Step 3 question. In the latter case, the correct answer is choice D.

Learn more about ChatGPT

Learn more about Bing Chat

thomas-reith@uiowa.edu

www.linkedin.com/in/thomaspreith/

University of Iowa Hospitals & Clinics
University of Iowa Health Care