GPT-4 Performance on Diagnostic Radiology Boards-Style Examination: Image vs Text with Confidence Analysis

Sydney Stewart, Randall Aziz, Rebecca Lipscomb, Karim Hanna

GPT-4 performs **better** on **text-based** radiology board-style questions than image-based ones and reports **higher confidence** with correct answers.

INTRODUCTION

- Large language models (LLMs) have the potential to aid physicians and patients through means of education and productivity.
- Despite their potential, the breadth of their applicability to radiology is still underexplored.
- This study aims to assess the accuracy and confidence level of GPT-4.0 with radiology board-style questions and compare its performance between image and text-based questions.

METHODS

- This study used 346 retired questions (184 image-based and 162 text-only) from the American College of Radiology Diagnostic Radiology In-Training Examinations to assess GPT-4.0's performance.
- The questions were categorized by medical subspecialty (breast imaging, neuroradiology, etc.) and question type (image-based or text only).
- The questions were presented to GPT-4 in three iterations.
- Chi-square test was used to compare the question type and the accuracy of GPT-4 for each iteration.
- Mann-Whitney test was used to compare the question type and accuracy with the confidence level of GPT-4 for each iteration.
- SPSS version 30 was used for analysis.
- P-values of less than .05 were considered statistically significant.

RESULTS

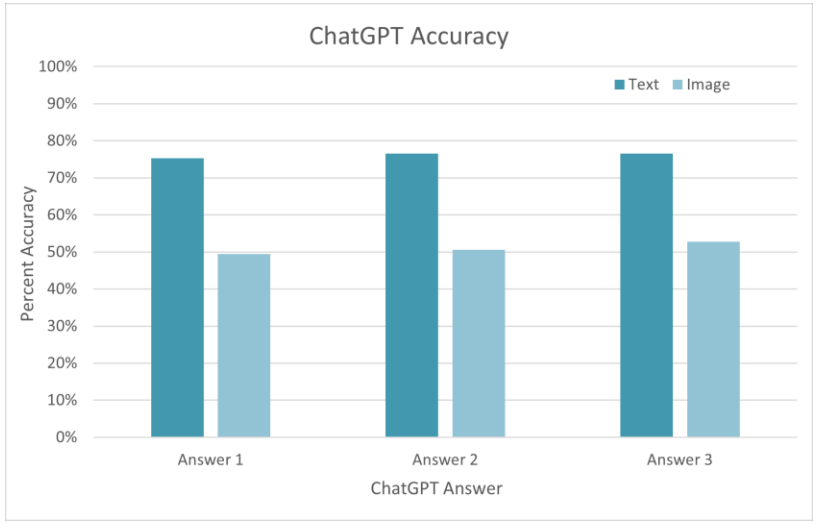


Figure 1. ChatGPT Answer Accuracy between image- and text-based questions. ChatGPT significantly more accurate answering text only questions than image-based in all three iterations (all p<.001).

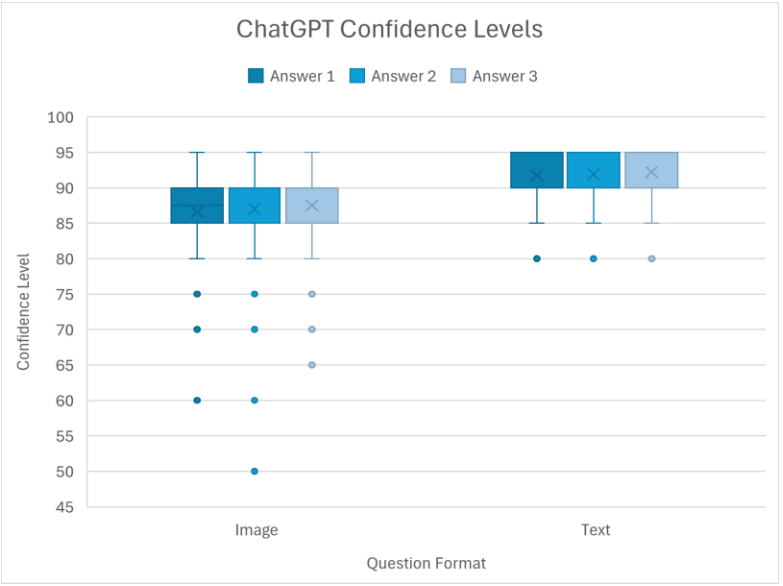


Figure 2. ChatGPT Answer Confidence Levels between image- and text-based questions. ChatGPT significantly more confident answering text only questions than image-based in all three iterations (p<.001). Please note y-axis scale does not start at 0.

- Mean confidence level of GPT-4 was higher for correct answers than incorrect answers in the first iteration (90.63±5.089 vs. 86.47±6.536, p<0.001), second iteration (90.55±5.521 vs. 87.21±6.184, p<0.001), and third iteration (90.88±5.091 vs. 87.64±6.184, p<0.001).
- There was no significant association between accuracy and topic for all three iterations.

CONCLUSION

- This study highlights the limits of radiological analysis by GPT-4, suggesting better performance with text-only inquiries.
- Future studies analyzing performance across different LLMs could provide more information regarding their applicability to radiology.

Take a picture for references and abstract