HealthManagement
Promoting Management and Leadership

Vision-Based GPT-4 for Radiology Examinations: Performance and Challenges

Artificial Intelligence (AI) has been making significant strides in various fields, including healthcare, where its role is evolving rapidly. Large language models (LLMs), such as OpenAI's GPT-4, have demonstrated impressive capabilities in natural language processing, solving complex tasks like summarisation, translation, and question answering. Recently, with the introduction of vision capabilities in GPT-4, the model has extended its functionality to analyse and interpret images, opening the door to numerous applications in fields like radiology. This article evaluates the performance of GPT-4 with vision (GPT-4V) on radiology in-training examination questions, a domain where image interpretation is critical, and explores its strengths and limitations in handling text- and image-based questions.

Baseline Performance of GPT-4 with Vision on Radiology Examinations

A study was conducted to assess GPT-4V's performance using retired diagnostic radiology in-training examination (DXIT) questions, a common benchmark for evaluating the knowledge of radiology residents. The dataset used in this evaluation comprised text- and image-based questions, allowing for a detailed assessment of the model's capabilities in interpreting textual and visual data. GPT-4V's overall accuracy was 65.3%, with a significant disparity between its performance on text-based (81.5%) and image-based (47.8%) questions. This difference highlights the model's competence in understanding and processing text-based medical information while exposing its challenges in interpreting radiologic images accurately.

The variation in accuracy also extended to different radiology subspecialties. GPT-4V performed exceptionally well in domains without image-based questions, such as physics and general radiology, where it achieved 87% and 83% accuracy, respectively. However, the model's performance on image-based questions was significantly lower in subspecialties that relied heavily on image interpretation, like nuclear medicine and paediatric radiology. These findings underscore that while GPT-4V has mastered textual comprehension to a significant degree, its visual diagnostic capabilities remain limited.

The Impact of Prompting Techniques on GPT-4V's Accuracy

Prompt engineering, or the design of input instructions given to AI models, plays a critical role in determining the accuracy and efficiency of their responses. In this study, five distinct prompting styles were analysed to observe their impact on GPT-4V's performance. These prompts ranged from basic, short, and long instructions to chain-of-thought prompts, each varying in detail and guidance provided to the model.

Chain-of-thought prompting, which encourages the model to reason step by step through the information, emerged as the most effective technique for text-based questions. It outperformed other prompting styles, including the original prompt, by improving accuracy by 8.9% in some instances. However, no significant improvements were observed when this technique was applied to image-based questions, indicating that the primary challenges for GPT-4V in radiology stem from its image interpretation skills rather than how it processes or approaches the questions.

Additionally, GPT-4V sometimes declined to answer questions, particularly those with images, due to built-in safety protocols designed to prevent it from making uncertain or harmful conclusions. While such safeguards are essential for ensuring reliability and safety in a clinical setting, they also highlight a limitation in GPT-4V's ability to handle diagnostic tasks autonomously, as the model may refrain from making critical decisions when the information is incomplete or ambiguous.

Limitations in Image-Based Radiology Interpretation

The most notable finding from evaluating GPT-4V in radiology was its underperformance in image-based questions. In radiology, interpreting visual data is at the core of diagnostic processes, where physicians must identify subtle abnormalities or lesions to arrive at accurate conclusions.

While GPT-4V demonstrated the ability to analyse and understand complex medical terminology, its ability to correctly identify and diagnose conditions based on radiologic images was limited.

One significant challenge identified was GPT-4V's tendency to provide hallucinatory responses—where the model confidently gave an incorrect interpretation of an image, sometimes locating lesions in incorrect organs. For example, in one case, GPT-4V mistakenly placed a lesion on the opposite side of the body but still arrived at the correct diagnosis. These hallucinations raise concerns about GPT-4V's reliability in clinical settings where incorrect interpretations could have profound implications for patient outcomes.

Moreover, the model's performance in image-based questions varied significantly across subspecialties, with better accuracy observed in chest and genitourinary radiology but poor results in nuclear medicine, where only 20% of image-based questions were answered correctly. This variability suggests that the model's visual understanding is far from generalised, and its effectiveness is highly dependent on the complexity of the images and the specific radiologic subspecialty.

Conclusion

The integration of vision capabilities into GPT-4 marks a significant advancement in AI's potential role in medical fields like radiology. However, this study's results indicate that while GPT-4V demonstrates commendable performance on text-based diagnostic tasks, its image-based interpretation abilities fall short of the accuracy required in clinical practice. The model's limitations in visual analysis, particularly its tendency to hallucinate findings and its varying performance across subspecialties, emphasise the need for further research and development to enhance its diagnostic capabilities.

Moreover, prompt engineering is pivotal in maximising the model's text-processing abilities, with chain-of-thought prompting offering the best results for text-based radiology questions. However, these improvements do not translate to visual tasks, indicating that the model's current visual capabilities are relatively unresponsive to different prompting strategies.

In conclusion, GPT-4V holds promise as an assistive tool in radiology, particularly for text-heavy tasks such as report generation and knowledge retrieval. However, its application in the critical area of image interpretation remains limited, necessitating future iterations of the model to address these gaps. Until then, GPT-4V should be viewed as a supplementary aid rather than a standalone diagnostic tool in radiology. Further refinements, including specialised training in radiologic image analysis, may help bridge the gap between text and image interpretation, unlocking the full potential of AI in the radiology domain.

Source: Radiology

Image Credit: iStock

Published on : Mon, 16 Sep 2024