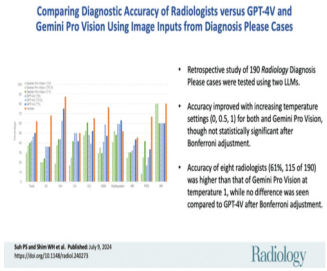
HealthManagement
Promoting Management and Leadership

Promise and Potential of AI in Radiology: GPT-4V and Gemini Pro Vision



Artificial intelligence (AI) has made remarkable strides in recent years, particularly with the development of large language models (LLMs) such as OpenAI's GPT-3.5 and GPT-4. These models, trained on extensive datasets using deep neural networks, have shown impressive capabilities in various fields, including medicine and radiology. GPT-4, a multimodal model, offers even greater accuracy and the ability to contextualise complex problems across different domains. This study aims to explore the diagnostic abilities of two advanced multimodal LLMs, GPT-4V by OpenAI and Gemini Pro Vision by Google DeepMind, using image inputs and varying temperature settings, and compare their performance to that of radiologists.

Advancements in AI for Medical Diagnostics

The recent advancements in AI, particularly in LLMs like GPT-4, have revolutionised the field of medical diagnostics. GPT-4, capable of processing both text and images, has shown potential in supporting clinical and diagnostic decision-making. Previous versions, such as GPT-3.5, demonstrated surprising proficiency in medical examinations and general radiology knowledge. However, these assessments were primarily text-based, lacking the integration of visual data crucial for accurate radiological diagnoses.

GPT-4V, an enhanced version with vision capabilities, addresses this gap by accepting image inputs in prompts. Similarly, Google's Gemini Pro Vision, another multimodal LLM, processes image data, marking a significant leap in AI-driven diagnostics. The ability to analyse patient history and medical images simultaneously enables these models to generate more accurate and contextually relevant differential diagnoses.

Study Design and Methodology

The study involved a comprehensive review of 318 cases from the Radiology journal's "Diagnosis Please" section, spanning from August 1998 to October 2023. After applying exclusion criteria, 190 cases with adequate image quality were included. The cases were categorised into eight radiology subspecialties, and inputs included patient history, original images, and figure legends extracted from PDF files.

Two LLMs, GPT-4V and Gemini Pro Vision, were evaluated for diagnostic accuracy. Prompts were crafted to minimise the AI systems' inherent biases while eliciting differential diagnoses from the models. Each LLM was instructed to present three potential disease candidates, considering the possibility of rare or unique cases. The likelihood of each disease was assessed on a scale from 1 to 10.

The temperature settings for the LLMs were adjusted to 0, 0.5, and 1, with responses generated five times for each case at each temperature. This approach aimed to balance determinism and variability, with higher temperatures expected to produce more creative outputs.

Results and Discussion

The study found that increasing the temperature settings generally improved the diagnostic accuracy of both GPT-4V and Gemini Pro Vision, although not significantly. GPT-4V achieved an overall accuracy of 49% at the highest temperature setting (T1), compared to 39% for Gemini Pro Vision. Radiologists, however, outperformed the AI models with an accuracy of 61%.

Interestingly, the performance of GPT-4V and Gemini Pro Vision varied across subspecialties. For instance, GPT-4V showed the highest accuracy in chest radiology (75%) and the lowest in pediatric radiology (33%). Gemini Pro Vision performed best in genitourinary radiology (61%) but struggled in gastrointestinal radiology (24%). These findings highlight the variability in AI performance across different medical domains.

A subgroup analysis comparing the first differential diagnosis accuracy revealed that radiologists significantly outperformed GPT-4V, achieving 48% accuracy compared to 15% for the AI model. This discrepancy underscores the current limitations of LLMs in providing precise initial diagnoses.

Future research should focus on evaluating the clinical decision support of LLMs when used in conjunction with radiologists, exploring other emerging multimodal models, and refining the AI algorithms to enhance their diagnostic accuracy. As AI continues to evolve, its role in medical diagnostics will likely become increasingly significant, offering valuable support to healthcare professionals and potentially improving patient outcomes.

This study demonstrates the potential of multimodal LLMs like GPT-4V and Gemini Pro Vision in supporting radiological diagnostics. While these AI models show promise, particularly with higher temperature settings, they still fall short of matching the accuracy of experienced radiologists. The integration of image inputs represents a significant advancement, but further improvements and studies are needed to realise the full potential of AI in clinical practice.

Source & Image Credit: Radiology

Published on : Wed, 10 Jul 2024