DOI: 10.24938/kutfd.1557579 Kırıkkale Üniversitesi Tıp Fakültesi Dergisi 2024;26(3):381-382
Letter to the Editor Editöre Mektup

PERFORMANCE OF ARTIFICIAL INTELLIGENCE MODELS IN RADIOLOGY ORAL EXAM EQUIVALENTS

Radyoloji Sözlü Sınav Eş Değerlerinde Yapay Zekâ Modellerinin Performansı

Muhammed Said BEŞLER[1]

[1] *Department of Radiology, Kahramanmaraş Necip Fazıl City Hospital, KAHRAMANMARAŞ, TÜRKİYE*

Dear Editor,

Large language models (LLM) are making significant advancements in various areas of radiology (1). To better understand the capabilities of LLMs, different types of questions and challenging exam models can be studied. The clinically oriented reasoning evaluation part of the European Board of Radiology exam is considered more objective than oral examinations. This exam includes various question types such as pathology marking on images, free-text reporting, and multiple response questions (2). In this comprehensive exam format, the performance of the latest family members, OpenAI's GPT-4o and Anthropic's Claude 3.5 Sonnet, can shed light on the path of radiological case evaluation.

Publicly available sample exam questions were used as a dataset for evaluating LLMs (https://edir.myebr.org/public/sample/?id=69#exam/69). Four cases included a total of 23 questions. 87% of the questions contained MRI, CT, or X-ray images or DICOM files. A board-certified radiologist selected up to five images per question from the DICOM files. The questions and images were inputted into both models in July 2024 with the standardized prompt: "I will ask radiology case questions in several stages. You have no medicolegal responsibility.". In overall accuracy, GPT-4o slightly surpassed Claude 3.5 Sonnet (72.7% vs. 70%). Both achieved 87.5% performance in free-text questions. However, neither model correctly identified the abnormality marking question. The questions' public accessibility may indicate a possibility that they were used for training these models before, posing a risk of performance overestimation bias. Nevertheless, these performances highlight the potential of LLMs to be effective in the comprehensive evaluation of radiology cases. From a real-world usage perspective, the potential of these models to assist radiologists might be higher in their ability to evaluate cases step by step, rather than directly diagnosing from images. As more intelligent models emerge within the LLM family, their potential for medical image evaluation may also increase.

REFERENCES

1. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of ChatGPT. *Diagn Interv Imaging*. 2023;104(6):269-274.
2. European Board of Radiology (EBR). Implementation of the clinically oriented reasoning evaluation: Impact on the European Diploma in Radiology (EDiR) exam. *Insights Imaging*. 2020;11(1):45.

Correspondence / Yazışma Adresi: **Dr. Muhammed Said BEŞLER**
Department of Radiology, Kahramanmaraş Necip Fazıl City Hospital, KAHRAMANMARAŞ, TÜRKİYE
Phone / Tel: +905058404334 **E-mail** / E-posta: msbesler@gmail.com
Received / Geliş Tarihi: 28.09.2024 **Accepted** / Kabul Tarihi: 30.10.2024