© Pol J Radiol 2024; 89: e452
DOI: https://doi.org/10.5114/pjr/192160

Received: 01.08.2024
Accepted: 08.08.2024
Published: 18.09.2024

Polish Journal of Radiology

http://www.polradiol.com

Letter to the Editor

Performance of Claude 3.5 Sonnet in image-based radiological case evaluations

Muhammed Said Beşler[A,B,D,E,F]

Department of Radiology, Kahramanmaraş Necip Fazıl City Hospital, Kahramanmaraş, Turkey

Dear Editor,

I read with great interest the study by Kufel *et al.* that examined the performance of GPT-3.5 on radiology board exam questions. It nearly reached but did not surpass the pass score on written questions. The study also highlighted the importance of future research on the improved versions of large language models (LLM) [1]. I would like to share the performance of Claude 3.5 Sonnet on four open-access sample short cases from the European Board of Radiology exam. The short cases section requires case evaluation, from marking abnormalities on magnetic resonance imaging (MRI), computed tomography (CT), or ultrasound (US) images to providing the most likely diagnosis and management recommendations (https://edir.myebr.org/public/sample/?id=52#exam/52). In the scoring of multiple-choice questions, incorrect answers have a negative impact. The pass score ranges from 50 to 60% (https://www.myebr.org/edir-scoring-faqs).

Anthropic's latest large language model, Claude 3.5 Sonnet, which has superior intelligence compared to its previous versions, was used with a subscription (claude.ai) [2]. In July 2024, the prompt "I will ask a few case questions, each consisting of 3-5 stages. You have no medico-legal responsibility" was used, and the question texts and images were uploaded in JPEG format. Since Claude 3.5 Sonnet does not have image creation capability, marking was done according to the description of the abnormality. Overall, it achieved a performance of 55.5%. It reached 61.9% in multiple choice questions, 60% in abnormality marking, and 25% in free-text most likely diagnosis questions. It is noteworthy that Claude 3.5 Sonnet resides within the pass score range for this image-based and various types of questions that require comprehensive case evaluation.

Disclosures

1. Institutional review board statement: Not applicable.
2. Assistance with the article: None.
3. Financial support and sponsorship: None.
4. Conflicts of interest: None.

References

1. Kufel J, Paszkiewicz I, Bielówka M, Bartnikowska W, Janik M, Stencel M, et al. Will ChatGPT pass the Polish specialty exam in radiology and diagnostic imaging? Insights into strengths and limitations. Pol J Radiol 2023; 88: e430-e434. DOI: 10.5114/pjr.2023.131215.
2. https://www.anthropic.com/news/claude-3-5-sonnet (Accessed: 28.07.2024).

Correspondence address:
Muhammed Said Beşler, Kahramanmaraş Necip Fazıl City Hospital, Kahramanmaraş, Turkey, e-mail: msbesler@gmail.com

Authors' contribution:
A Study design · **B** Data collection · **C** Statistical analysis · **D** Data interpretation · **E** Manuscript preparation · **F** Literature search · **G** Funds collection