Letter to the Editor

eISSN 2005-8330
https://doi.org/10.3348/kjr.2024.0438
Korean J Radiol 2024;25(8):767-768

Korean Journal of Radiology
KJR

Comparative Evaluation of the Accuracies of Large Language Models in Answering VI-RADS-Related Questions

Eren Çamur[1], Turay Cesur[2], Yasin Celal Güneş[3]

[1]Department of Radiology, Ministry of Health Ankara 29 Mayis State Hospital, Ankara, Türkiye
[2]Department of Radiology, Ankara Mamak State Hospital, Ankara, Türkiye
[3]Department of Radiology, TC Saglik Bakanligi Kirikkale Yuksek Ihtisas Hastanesi, Kırıkkale, Türkiye

Keywords: Large language model; Artificial intelligence; Bladder; Cancer; VI-RADS

We read with great interest the letter by Kaba et al. [1] that examined the accuracy of large language models (LLMs) in answering questions related to the Korean Thyroid Imaging reporting and data system (RADS). This letter provides valuable information and insights into the potential use of LLMs in imaging and reporting systems. With the increasing number of studies investigating the radiological knowledge of LLMs and their benefits to radiology, we aimed to uncover LLMs' knowledge of vesical imaging (VI)-RADS, an important lexicon for bladder cancer (BC) reporting to provide a new perspective on this field [2-4].

Tumor grade, stage, and biological potential are pivotal for managing BC, and this essential information is best obtained through comprehensive clinical, histopathological, and radiological assessments. Multiparametric magnetic resonance imaging (mpMRI) has become indispensable for the radiological evaluation of BC as it delivers high-resolution images while avoiding radiation exposure. Therefore, standardizing mpMRI reports for BC has become imperative. In this context, the VI-RADS, published in 2018, was designed to define a standardized approach for mpMRI imaging and reporting for BC [5].

Radiologists (E.Ç.) who obtained the European Diploma in Radiology prepared the 25 multiple-choice questions in this letter utilizing the information in the VI-RADS, thus eliminating the need for ethics committee approval (Supplement). We initiated the input prompt as follows: "Act like a professor of radiology who has 30 years of experience in genitourinary radiology, especially studies on BC. Give just letter of correct choice from the questions I will give you about VI-RADS. Each question have only one correct answer." The LLMs were asked each question individually. This prompt was tested in May 2024 on nine different LLMs using the default settings. The testing included models from various developers: Claude 3 Opus by Anthropic (https://claude.ai), ChatGPT-3.5, ChatGPT-4 by OpenAI (https://chat.openai.com), Gemini 1.5 Pro by Google (https://aistudio.google.com) and Gemini 1.0 by Google (https://gemini.google.com), Microsoft Copilot (Balanced) (https://copilot.microsoft.com), Mistral Large (https://mistral.ai), Meta LLaMA 3 70B by Meta (https://metaai.com), and Perplexity (https://perplexity.ai).

The results revealed that Claude 3 Opus achieved the highest accuracy of 96% (24/25 questions), followed by ChatGPT-4, Mistral Large, and Meta LlaMA 3 70B with 92% accuracy (23/25 questions). Following these, Gemini 1.5 Pro at 88% (22/25 questions), ChatGPT-3.5 at 84% (21/25 questions), Perplexity and Gemini 1.0 at 80% (20/25 questions) and lastly Copilot had an accuracy of 68% (17/25 questions).

Our results show that although there are some variations, most LLM models exhibit significant adequacy in answering questions related to VI-RADS. The outstanding success of Claude 3 Opus raises the question of whether it can be a new game-changer in this field. The variations in the performance of LLMs result from their distinctive designs. These results illustrate that while certain LLM models can significantly enhance our comprehension and knowledge of VI-RADS, further investigation is necessary to fully realize their potential in this field.

Received: May 10, 2024 **Accepted:** May 15, 2024
Corresponding author: Eren Çamur, MD, Department of Radiology, Ministry of Health Ankara 29 Mayis State Hospital, Aydınlar, Dikmen Cd No:312, Çankaya/Ankara 06105, Türkiye
• E-mail: eren.camur@outlook.com

Supplement

The Supplement is available with this article at https://doi.org/10.3348/kjr.2024.0438.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Author Contributions

Conceptualization: Eren Çamur. Data curation: Eren Çamur. Formal analysis: Eren Çamur. Investigation: Eren Çamur, Yasin Celal Güneş. Methodology: Eren Çamur, Turay Cesur. Supervision: Turay Cesur, Yasin Celal Güneş. Validation: Eren Çamur. Writing—original draft: Eren Çamur. Writing—review & editing: Eren Çamur.

ORCID IDs

Eren Çamur
https://orcid.org/0000-0002-8774-5800
Turay Cesur
https://orcid.org/0000-0002-2726-8045
Yasin Celal Güneş
https://orcid.org/0000-0001-7631-854X

Funding Statement

None

REFERENCES

1. Kaba E, Hürsoy N, Solak M, Çeliker FB. Accuracy of large language models in thyroid nodule-related questions based on the Korean thyroid imaging reporting and data system (K-TIRADS). *Korean J Radiol* 2024;25:499-500
2. Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol* 2024;30:80-90
3. Gunes YC, Cesur T. A comparative study: diagnostic performance of ChatGPT 3.5, Google Bard, Microsoft Bing, and radiologists in thoracic radiology cases. medRxiv [Preprint]. 2024 [accessed on January 20, 2024]. Available at: https://doi.org/10.1101/2024.01.18.24301495
4. Kim K, Cho K, Jang R, Kyung S, Lee S, Ham S, et al. Updated primer on generative artificial intelligence and large language models in medical imaging for medical professionals. *Korean J Radiol* 2024;25:224-242
5. Panebianco V, Narumi Y, Altun E, Bochner BH, Efstathiou JA, Hafeez S, et al. Multiparametric magnetic resonance imaging for bladder cancer: development of VI-RADS (vesical imaging-reporting and data system). *Eur Urol* 2018;74:294-306