Letter to the Editor

Comparison of the performance of large language models and general radiologist on Ovarian-Adnexal Reporting and Data System (O-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, Ministry of Health Kirikkale Yuksek Ihtisas Hospital, Kırıkkale, Türkiye

Correspondence to: Eren Çamur, MD. Department of Radiology, Ministry of Health Ankara 29 Mayis State Hospital, Aydınlar, Dikmen Cd No. 312, 06105 Çankaya/Ankara, Türkiye. Email: eren.camur@outlook.com.

Comment on: Wang Z, Zhang Z, Traverso A, Dekker A, Qian L, Sun P. Assessing the role of GPT-4 in thyroid ultrasound diagnosis and treatment recommendations: enhancing interpretability with a chain of thought approach. Quant Imaging Med Surg 2024;14:1602-15.

Submitted Jun 07, 2024. Accepted for publication Jun 25, 2024. Published online Jul 15, 2024.
doi: 10.21037/qims-24-1142
View this article at: https://dx.doi.org/10.21037/qims-24-1142

We read with great interest the study by Wang *et al.* about assessing the role of GPT-4 in thyroid ultrasound (US) diagnosis and treatment recommendations (1). This study provides valuable information and insights into the potential use of large language models (LLMs) in imaging. With the increasing number of studies investigating the radiological knowledge of LLMs and their benefits to radiology, we aimed to uncover the knowledge of LLMs about Ovarian-Adnexal Reporting and Data System (O-RADS), an important lexicon for ovarian and adnexial lesions diagnosis, reporting and follow-up to provide a new perspective on this field (2,3).

Radiological features play a critical role in the diagnosis and treatment decision of ovarian and adnexal lesions. The recommendations of the radiologist guide patient management. US have become indispensable in the radiological evaluation of these masses today, as they provide high-resolution images that provide very important information about the imaging features of ovarian and adnexal lesions without radiation exposure. Therefore, O-RADS lexicon for US was introduced in 2018, establishing a comprehensive and standardized terminology that encompasses all relevant descriptors and precise definitions pertaining to the characteristic ultrasonographic appearance of normal ovaries and ovarian or other adnexal lesions (4). This lexicon serves as a valuable resource for radiologists, promoting consistency and accuracy in the description and interpretation of these structures during ultrasonographic examinations.

Radiologist (E.Ç.) who obtained board certified (EDiR) prepared the 30 multiple-choice questions in this letter utilizing the information in O-RADS, thus eliminating the need for ethics committee approval. To ensure transparency and reproducibility, all the questions used in this letter and data are included in Appendices 1,2. 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 ovarian and adnexal masses. Give just letter of correct choice from the questions I will give you about O-RADS. Each question has only one correct answer". This prompt was tested in June 2024 on eleven different LLMs using the default settings. The testing included models from various developers: Anthropic's Claude 3 Opus and Sonnet (https://claude.ai.com), OpenAI's ChatGPT 3.5, ChatGPT 4 and ChatGPT 4o (https://chat.openai.com), Google Gemini 1.5 Pro (https://aistudio.google.com)

^ ORCID: Eren Çamur, 0000-0002-8774-5800; Turay Cesur, 0000-0002-2726-8045; Yasin Celal Güneş, 0000-0001-7631-854X.

and Gemini 1.0 (https://gemini.google.com), Mistral Large (https://mistral.ai), Meta Llama 3 70B (https://metaai.com), Perplexity and Perplexity pro (https://perplexity.ai). Also general radiologist (T.C.) board certified by EDiR and with 6 years of experience each, answered the same questions.

The results revealed that Mistral Large achieved the highest accuracy of 96.66% (29/30 questions), followed by ChatGPT 4o and Claude 3 Opus with 93% accuracy (28/30 questions). Following these ChatGPT 4 and Perplexity pro 90% (27/30 questions), Meta Llama 3 70 B at 86.6%, ChatGPT 3.5, Gemini 1.5 pro and Claude Sonnet at 83% (25/30 questions) and lastly Gemini 1.0 had accuracy of 80% (24/30 questions). General radiologist (T.C.) has accuracy of 90%.

Our findings show that most of the LLM models perform comparably to the general radiologist in handling questions related to the O-RADS lexicon, although there is some variability in performance. The observed differences in LLM performance can be attributed to the unique architectural and training characteristics of each model. These results underline the potential for specific LLM models to significantly increase our understanding and knowledge of O-RADS. However, it is clear that additional research is required to fully exploit the capabilities of these models in the context of ovarian and adnexal lesion characterisation using US.

Acknowledgments

The authors used ChatGPT, a language model based on the GPT-3.5 architecture (May 2024 Version; OpenAI; https://chat.openai.com/) to revise the grammar and English translation of this article.
Funding: None.

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-1142/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Ethical approval is not applicable to this study.

References

1. Wang Z, Zhang Z, Traverso A, Dekker A, Qian L, Sun P. Assessing the role of GPT-4 in thyroid ultrasound diagnosis and treatment recommendations: enhancing interpretability with a chain of thought approach. Quant Imaging Med Surg 2024;14:1602-15.
2. Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, Cuocolo R, Cannella R, Koçak B. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol 2024;30:80-90.
3. Çamur E, Cesur T, Güneş YC. Accuracies of large language models in answering radiation protection questions. J Radiol Prot 2024. doi: 10.1088/1361-6498/ad4b29.
4. Andreotti RF, Timmerman D, Strachowski LM, Froyman W, Benacerraf BR, Bennett GL, Bourne T, Brown DL, Coleman BG, Frates MC, Goldstein SR, Hamper UM, Horrow MM, Hernanz-Schulman M, Reinhold C, Rose SL, Whitcomb BP, Wolfman WL, Glanc P. O-RADS US Risk Stratification and Management System: A Consensus Guideline from the ACR Ovarian-Adnexal Reporting and Data System Committee. Radiology 2020;294:168-85.

Cite this article as: Çamur E, Cesur T, Güneş YC. Comparison of the performance of large language models and general radiologist on Ovarian-Adnexal Reporting and Data System (O-RADS)-related questions. Quant Imaging Med Surg 2024;14(9):6990-6991. doi: 10.21037/qims-24-1142

