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

Supplementary

Appendix 1 Multiple Choice Questions and Answers

1. What is the main objective of O-RADS in relation to ovarian cancer?
   A) To diagnose ovarian cancer at an early stage
   B) To assess the response to chemotherapy
   C) To stratify patients based on their genetic predisposition
   D) To guide management decisions while minimizing unnecessary surgical procedures

2. According to the O-RADS US system, which category indicates an 'almost certainly benign' lesion with less than 1% risk of malignancy?
   A) O-RADS 0
   B) O-RADS 1
   C) O-RADS 2
   D) O-RADS 3

3. What does O-RADS 3 category signify in terms of risk of malignancy?
   A) High risk, more than 50% likelihood
   B) Intermediate risk, between 10% and 50% likelihood
   C) Low risk, between 1% and 10% likelihood
   D) Incomplete evaluation

4. Which feature is included in the description of typical features of a hemorrhagic cyst according to the O-RADS US lexicon?
   A) Multilocularity with solid components
   B) High color score in Doppler studies
   C) Retracting clot with concave margins
   D) Presence of significant solid tissue

5. What is the management recommendation for an O-RADS 1 category lesion, which represents a normal ovary in a premenopausal patient?
   A) Immediate surgical evaluation
   B) Routine follow-up in 6 months
   C) No additional imaging or follow-up necessary
   D) Repeat ultrasound in 12 weeks

6. When was the Ovarian-Adnexal Reporting and Data System (O-RADS) lexicon for US published?
   A) 2010
   B) 2015
   C) 2018
   D) 2020

7. Which US feature is most likely to be found in an O-RADS 2 category lesion?
   A) Irregular solid lesion with high color score
   B) Simple cyst less than 10 cm in diameter
   C) Multilocular cyst with solid components
   D) Ascites with peritoneal nodules

8. What does an O-RADS 0 category indicate in the O-RADS US system?
   A) Normal premenopausal ovary
   B) High risk of malignancy
   C) Incomplete evaluation
   D) Lesion with low risk of malignancy

https://dx.doi.org/10.21037/qims-24-1142

9. In the O-RADS US system, which category requires additional imaging study such as MRI?
   A) O-RADS 1
   B) O-RADS 0
   C) O-RADS 5
   D) O-RADS 3

10. What is the risk of malignancy associated with O-RADS 5 category lesions?
   A) Less than 1%
   B) 1% to less than 10%
   C) 10% to less than 50%
   D) 50% or more

11. Which management strategy is recommended for a lesion classified as O-RADS 4?
   A) No follow-up necessary
   B) Routine imaging follow-up
   C) Consultation with a gynecologic oncologist
   D) Only clinical follow-up

12. What is the definition of a simple cyst according to the O-RADS US lexicon?
   A) A cyst with solid components and septations
   B) A cyst that is anechoic with a smooth wall and posterior enhancement
   C) A cyst with irregular walls and increased vascularity
   D) A multilocular cyst with low color score

13. O-RADS 2 category lesions are described as having what level of risk for malignancy?
   A) High risk, greater than 50%
   B) Low risk, 1% to less than 10%
   C) Almost certainly benign, less than 1%
   D) Very high risk, greater than 70%

14. Which of the following is a characteristic of a typical endometrioma as per the O-RADS lexicon?
   A) Multilocular cyst with solid components
   B) Homogenous low-level internal echoes
   C) High color score with irregular solid components
   D) Simple cyst with posterior shadowing

15. What is the management recommendation for a lesion categorized under O-RADS 5?
   A) Immediate referral to gynecologic oncology
   B) No follow-up or routine follow-up
   C) Annual follow-up with MRI
   D) Ultrasound follow up 12 weeks later

16. Which US feature is NOT typically associated with a hemorrhagic cyst according to O-RADS?
   A) Retracting clot with concave margins
   B) High color score and vascularity
   C) Reticular pattern within the cyst
   D) No flow at color Doppler

17. For which O-RADS category is the presence of ascites particularly concerning for malignancy?
   A) O-RADS 1
   B) O-RADS 2
   C) O-RADS 4
   D) O-RADS 5

https://dx.doi.org/10.21037/qims-24-1142

18. O-RADS 4 category lesions have which level of risk for malignancy?
    A) Less than 1%
    B) 1% to less than 10%
    C) 10% to less than 50%
    D) 50% or more

19. What is a characteristic feature of lesions in the O-RADS 1 category?
    A) Papillary projections
    B) Multilocular cyst
    C) Corpus luteum <3cm
    D) İrregular thick cyst wall

20. What is typical feature of peritoneal inclusion cyst according to O-RADS?
    A) Contour that follows the adjacent pelvic organs and peritoneum
    B) Reticular pattern
    C) Hyperechoic component with acoustic shadowing
    D) Homogenous low level echoes

21. In the context of O-RADS, what does a lesion categorized as O-RADS 4 typically signify about its characteristics?
    A) It is likely benign and requires no follow-up.
    B) It poses intermediate risk of malignancy and may need gynecologic oncology referral.
    C) It is almost certainly benign.
    D) Evaluation is incomplete and needs further imaging.

22. Which scenario best applies to O-RADS 3 classification for management of ovarian lesions?
    A) Routine surgical intervention is necessary.
    B) Follow-up imaging or clinical follow-up may be sufficient.
    C) Direct referral to gynecologic oncology is mandatory.
    D) No imaging follow-up is needed.

23. What is the recommended management for premenopausal woman with O-RADS 2 lesion that is smaller than 3cm, non-simple unilocular and have smooth inner margin cyst?
    A) No imaging follow-up is needed
    B) Immediate surgical removal
    C) Further evaluation by MRI
    D) Specialist consultation

24. Which O-RADS category suggests a high risk of malignancy with a recommendation for direct management by a gynecologic oncologist?
    A) O-RADS 1
    B) O-RADS 2
    C) O-RADS 4
    D) O-RADS 5

25. For a lesion evaluated as O-RADS 4, what is the primary next step in management?
    A) Repeat ultrasound or alternate imaging study
    B) No follow-up necessary
    C) Immediate referral to gynecologist
    D) Routine follow-up in one year

https://dx.doi.org/10.21037/qims-24-1142

26. Which feature is NOT part of the O-RADS US lexicon for describing typical dermoid cysts?
    A) Hyperechoic lines and dots
    B) Reticular pattern with clot retraction
    C) Floating fat levels
    D) Hyperechoic component with acoustic shadowing

27. In the O-RADS system, how is a simple cyst in a postmenopausal woman typically managed if it is less than 3 cm in diameter?
    A) Follow-up imaging within three months
    B) Annual follow-up for five years
    C) No further imaging required
    D) Immediate surgical evaluation

28. What criterion differentiates an O-RADS 3 from an O-RADS 4 lesion?
    A) The presence of solid components with CS= 2-3
    B) The size of the lesion being over 10 cm
    C) Loculation
    D) The presence of septations

29. Which type of cyst is typically classified under O-RADS 2 due to its benign nature?
    A) Endometriomas
    B) Hemorrhagic cysts
    C) Complex cysts with septations
    D) Simple cysts

30. What management strategy is generally for typical hydrosalpinx?
    A) Regular imaging follow-ups
    B) Surgical intervention
    C) Gynecologist consultation
    D) MRI

Answers

1. D	2. C	3. C	4. C	5. C
6. C	7. B	8. C	9. B	10. D
11. C	12. B	13. C	14. B	15. A
16. B	17. D	18. C	19. C	20. A
21. B	22. B	23. A	24. D	25. C
26. B	27. C	28. A	29. D	30. C

https://dx.doi.org/10.21037/qims-24-1142

Appendix 2 The dataset of study

Number of Question	ChatGPT 3.5	ChatGPT 4	ChatGPT 4.o	Gemini 1.5pro	Gemini 1.0	Perplexity	Perplexity pro	Mistral large	Llama 3 70b	Claude 3 Opus	Claude Sonnet	General radiologist
1.	1	1	1	1	1	1	1	1	1	1	1	1
2.	1	1	1	1	1	1	1	1	0	1	0	1
3.	1	1	1	1	1	1	1	1	1	1	1	1
4.	1	1	1	1	1	1	1	1	1	1	1	1
5.	1	1	1	1	1	1	1	1	1	1	1	1
6.	1	1	1	1	1	1	1	1	1	1	0	1
7.	1	1	1	1	1	1	1	1	1	1	1	1
8.	1	1	1	1	1	1	1	1	1	1	1	1
9.	0	0	0	0	0	0	1	1	0	1	0	0
10.	1	1	1	1	1	1	1	1	1	1	1	1
11.	1	1	1	1	1	1	1	1	1	1	1	1
12.	1	1	1	1	1	1	1	1	1	1	1	1
13.	0	1	1	1	0	1	1	1	1	1	0	1
14.	1	1	1	0	1	1	1	1	1	1	1	1
15.	1	1	1	1	1	1	1	1	1	1	1	1
16.	0	1	1	1	0	0	1	1	1	1	1	1
17.	1	1	1	1	1	1	1	1	1	1	1	1
18.	1	1	1	1	1	1	1	1	1	1	1	1
19.	1	1	1	1	1	1	1	1	1	1	1	1
20.	1	1	1	1	1	1	1	1	1	1	1	1
21.	1	1	1	1	1	1	1	1	1	1	1	1
22.	1	1	1	1	1	1	1	1	1	1	1	1
23.	1	1	1	0	0	0	1	1	0	1	1	1
24.	1	1	1	1	1	1	1	1	1	1	1	1
25.	1	0	1	0	0	0	0	1	0	0	1	1
26	1	1	1	1	1	0	1	1	1	1	1	0
27	1	1	1	1	1	1	1	1	1	1	1	1
28	0	1	1	1	1	1	1	1	1	1	1	1
29	1	1	1	1	1	1	0	0	1	1	1	0
30	0	0	0	0	0	1	0	1	1	0	0	1
Accuracy	83%	90%	93%	83%	80%	83%	90%	96.66%	86.60%	93%	83%	90%
True:1 False: 0												

https://dx.doi.org/10.21037/qims-24-1142