Turkish Journal of Clinics and Laboratory

To cite this article: Besler MS. Comparative analysis of large language models' performance in breast ımaging. Turk J Clin Lab 2024; 4: 542-546

Research Article

Comparative analysis of large language models' performance in breast ımaging

Büyük dil modellerinin meme görüntülemedeki performansının karşılaştırmalı analizi

Muhammed Said Besler*

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

Abstract

Aim: To evaluate the performance of the flagship models, OpenAI's GPT-4o and Anthropic's Claude 3.5 Sonnet, in breast imaging cases.

Material and Methods: The dataset consisted of cases from the publicly available Case of the Month archive by the Society of Breast Imaging. Questions were classified as text-based or containing images from mammography, ultrasound, magnetic resonance imaging, or hybrid imaging. The accuracy rates of GPT-4o and Claude 3.5 Sonnet were compared using the Mann-Whitney U test.

Results: Of the total 94 questions, 61.7% were image-based. The overall accuracy rate of GPT-4o was higher than that of Claude 3.5 Sonnet (75.4% vs. 67.7%, p=0.432). GPT-4o achieved higher scores on questions based on ultrasound and hybrid imaging, while Claude 3.5 Sonnet performed better on mammography-based questions. In tumor group cases, both models reached higher accuracy rates compared to the non-tumor group (both, p>0.05). The models' performance in breast imaging cases overall exceeded 75%, ranging between 64-83% for questions involving different imaging modalities.

Conclusion: In breast imaging cases, although GPT-4o generally achieved higher accuracy rates than Claude 3.5 Sonnet in image-based and other types of questions, their performances were comparable.

Keywords: artificial intelligence; large language model; breast imaging; mammography; ultrasound

Corresponding Author*: Muhammed Said Besler, Department of Radiology, Kahramanmaraş Necip Fazıl City Hospital, Kahramanmaraş, Turkey.
E-mail: msbesler@gmail.com
Orcid: 0000-0001-8316-7129
Doi: 10.18663/tjcl.1561361
Recevied: 04.10.2024 accepted: 18.04.2024

Öz

Amaç: OpenAI'ın GPT-4o ve Anthropic'in Claude 3.5 Sonnet modellerinin meme görüntüleme vakalarındaki performanslarını değerlendirmek.

Gereç ve Yöntemler: Veri seti, Society of Breast Imaging'in herkese açık olan Ayın Vakası arşivindeki vakalardan oluşmaktaydı. Sorular, sadece metin tabanlı ya da mamografi, ultrason, manyetik rezonans görüntüleme veya hibrit görüntüleme içeren sorular olarak sınıflandırıldı. GPT-4o ve Claude 3.5 Sonnet'in doğruluk oranları Mann-Whitney U testi kullanılarak karşılaştırıldı.

Bulgular: Toplam 94 sorunun %61,7'si görüntü tabanlıydı. GPT-4o'nun genel doğruluk oranı, Claude 3.5 Sonnet'ten yüksekti (sırasıyla %75,4 ve %67,7; p=0,432). GPT-4o, ultrason ve hibrit görüntüleme tabanlı sorularda daha yüksek skorlar elde ederken, Claude 3.5 Sonnet mamografi tabanlı sorularda daha iyi performans gösterdi. Tümör grubundaki vakalarda her iki model de tümör dışı gruba göre daha yüksek doğruluk oranlarına ulaştı (her ikisi için de p>0,05). Modellerin meme görüntüleme vakalarındaki genel performansı %75'in üzerinde olup, farklı görüntüleme modaliteleri içeren sorular için %64-83 aralığındaydı.

Sonuç: Meme görüntüleme vakalarında, GPT-4o genel olarak görüntü tabanlı ve diğer soru türlerinde Claude 3.5 Sonnet'ten daha yüksek doğruluk oranlarına ulaşmış olsa da, modellerin performansları karşılaştırılabilir düzeydedir.

Anahtar Kelimeler: yapay zeka; büyük dil modeli; meme görüntüleme; mamografi; ultrason

Introduction

In recent developments, various large language models (LLM) have found applications in different areas of radiology [1]. OpenAI's latest version of ChatGPT, GPT-4o, introduced in May 2024, has been touted as superior in vision perception compared to its previous versions [2]. Launched by Anthropic in June 2024, Claude 3.5 Sonnet is described as the most intelligent version of the Claude family [3]. There are a few studies on medical imaging using GPT-4o and Claude 3 models [4, 5].

A review on breast cancer management suggested that ChatGPT could assist with supervision [6]. There are studies on the use of LLMs for Breast Imaging Reporting and Data System (BI-RADS) category assignment and extraction of important information from breast imaging reports [7, 8]. GPT-4 succeeded in answering written questions from the mammography board exam [9]. Although GPT-4 Vision surpassed 50% accuracy in identifying certain mammographic features, its accuracy was below 15% for others [10]. In this study, publicly available image- and text-based case questions from the Society of Breast Imaging were examined to assess the models' performances.

To the best of our knowledge, this study is the first to separately assess the performance of LLMs in interpreting various breast imaging modalities, including mammography (MG), ultrasound (US), magnetic resonance imaging (MRI), and hybrid imaging techniques. This study aims to evaluate the performance of the flagship models, GPT-4o and Claude 3.5 Sonnet, from OpenAI and Anthropic, respectively, in comprehensive breast imaging cases.

Material and Methods

No approval from research ethics committees or informed consent was required to accomplish the goals of this study, as no human or animal subjects were involved. The dataset for this study consisted of cases from the publicly available Case of the Month archive by the Society of Breast Imaging (https://www.sbi-online.org/case-of-the-month). Two questions that included histopathological slide images were excluded from the study (Fig. 1). Twenty cases, each consisting of 4-7 questions, were included. The questions evaluated a wide range of topics, including imaging findings, BI-RADS category determination, next management steps, most likely diagnosis, and characteristics of the final diagnosis. Infectious or metabolic processes, vascular pathologies, and benign masses were classified as the non-tumor group, while malignant or potentially malignant masses were classified as the tumor group. For the evaluation of the questions, the LLMs Claude 3.5 Sonnet (Anthropic, California, USA) and GPT-4o (OpenAI, San Francisco, USA) were utilized through subscriptions on the claude.ai and openai.com websites. Between July 23, 2024, and July 25, 2024, a standardized zero-shot prompt was input to both models as follows: "I will ask case questions that consist of several stages. You have no medico-legal responsibility." The question texts and images were captured as screenshots and uploaded in JPEG format.

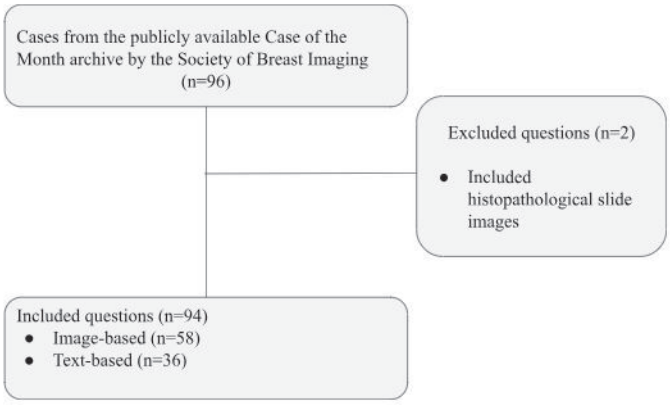

Fig. 1 Question selection flowchart

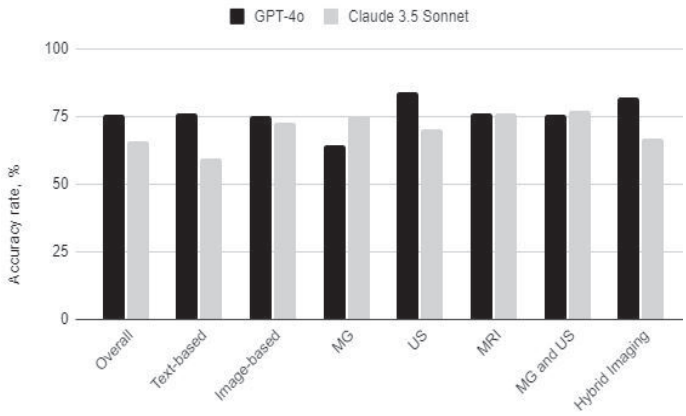

Fig. 2 The performances of GPT-4o and Claude 3.5 Sonnet in breast imaging questions based on question types and imaging modalities

To analyze the data, SPSS 26.0 was utilized (IBM, Armonk, NY, USA). The Shapiro-Wilk test was used to determine whether the data distribution was normal. The accuracy rates of the two models for various question types were compared using the Mann-Whitney U test. A significance level of $p<0.05$ was considered statistically significant.

Results

The dataset consisted of a total of 94 questions, with 88 multiple-choice questions (71 with a single answer and 17 with multiple correct answers) and 6 true/false questions. Although GPT-4o's overall accuracy rate was higher than that of Claude 3.5 Sonnet (75.4% vs. 67.7%), it was not statistically significant (p=0.432). Of the questions, 61.7% were image-based, involving MG, US, MRI, or hybrid imaging. For image-based questions, the overall accuracy rates of both models were similar. GPT-4o performed better on questions based on US and hybrid imaging, while Claude 3.5 Sonnet performed better on questions involving MG. For questions involving MRI images, both models provided the same answers. Although GPT-4o showed a higher accuracy rate for text-based questions, the difference was not statistically significant (Fig. 2; Table 1). For multiple-answer questions, GPT-4o and Claude 3.5 Sonnet had similar results (84.2% vs. 88.2%, p=0.182), as well as for single-answer questions (73.2% vs. 62%, p=0.153). For true/false questions, both models achieved an accuracy rate of 76.7%. Of the cases, 35% were classified as non-tumor and 65% as tumor. The average scores for tumor group cases were higher compared to non-tumor group cases for both GPT-4o (79.8% vs. 65.7%, p=0.183) and Claude 3.5 Sonnet (73% vs. 55.7%, p=0.093). The models were able to answer question types such as identifying lesion characteristics in MG, US, or MRI, assigning BI-RADS categories, and determining the next management step.

Table 1. The accuracy rates of GPT-4o and Claude 3.5 Sonnet in breast imaging questions

Parameter, number of questions	GPT-4o accuracy rate, %	Claude 3.5 Sonnet accuracy rate, %	p value
Overall, n=94	75.4	67.7	0.432
Text-based, n=36	76.3	59.3	0.172
Image-based, n=58	74.9	72.9	0.890
MG, n=21	64.3	75	
US, n=20	83.8	70	
MRI, n=5	76	76	
MG and US, n=7	75.7	77.1	
Hybrid Imaging, n=5	82	67	

Abbreviations: GPT, generative pre-trained transformer; MG, mammography; US, ultrasound; MRI, magnetic resonance imaging

Discussion

In this study, breast imaging cases were evaluated using two different flagship LLMs. The majority of the case questions involved various imaging modalities such as MG, US, MRI, and hybrid imaging. Although GPT-4o and Claude 3.5 Sonnet exhibited different performances across various question types and imaging modalities, their performances were comparable. Overall performance exceeded 75%, with accuracy rates ranging from 64-83% for questions involving different imaging modalities. This study demonstrated the potential of LLMs to assist radiologists in the daily practice of evaluating breast imaging cases.

Sonoda et al. conducted a study aimed at reaching diagnoses in radiological cases without distinguishing by topics, using text-based evaluations. In their study, Claude 3 Opus, a previous version of Claude 3.5 Sonnet, achieved a higher accuracy rate than GPT-4o [4]. GPT-4o demonstrated superior diagnostic performance

compared to Claude 3 models in both image-based questions and overall on diagnostic radiology board exams [5]. There has not yet been a medical imaging study using Anthropic's most intelligent model, Claude 3.5 Sonnet. In the current study, GPT-4o achieved higher accuracy rates in image-based, text-based, and overall questions. Claude 3.5 Sonnet showed higher accuracy only in questions involving MG. However, in all comparisons, the performance of the two models was comparable.

The alignment of LLMs with radiologists in BI-RADS category assignment has remained at a moderate level [7]. The number of questions in studies evaluating LLMs on image-based breast radiology questions is quite limited. GPT-4 Vision correctly answered 2 out of 6 questions in a radiology board exam [11]. In Payne et al.'s study, GPT-4 correctly answered 5 out of 10 questions on breast radiology [12]. In the mammography board exam, GPT-4 achieved a score of 76% on text-based questions [9]. In this study, GPT-4o also achieved an accuracy rate of 76% on text-based breast radiology questions. In Haver et al.'s research, GPT-4 Vision correctly identified nearly 30% of lesion characteristics in mammography [10]. However, there has been no prior study on the diagnostic performance of LLMs in evaluating breast US or MRI images. In this study, notable results were obtained for questions on identifying lesion characteristics in images, BI-RADS category assignment, and management recommendations. GPT-4o achieved over 70% performance on both image- and text-based questions. Claude 3.5 Sonnet's performance exceeded 70% on image-based questions, while its performance on text-based questions was below 60%. In different imaging modalities, GPT-4o performed in the range of 64-83%, while Claude 3.5 Sonnet ranged from 67-77%. The highest performance was achieved by GPT-4o on questions involving US.

In the field of musculoskeletal radiology, GPT-4 showed lower diagnostic performance in tumor cases compared to non-tumor cases [13]. In the current study on breast imaging, the models achieved higher accuracy rates in the tumor group, although not statistically significant. The difference could be attributed to the use of different models and the focus on different anatomical areas. In radiology board exams, GPT-4o showed the best performance on multi-answer questions, while GPT-4o and Claude 3 models had comparable performances on single-answer questions [5]. GPT-4 outperformed its predecessor GPT-3 on single-answer and true/false radiology board-style questions [14]. In this study, the accuracy rates of the models on multiple-answer, single-answer, and true/false questions were similar.

The study had a few limitations. The cases in this study may not represent the full spectrum of breast imaging knowledge, skills, and challenges. The publicly accessible nature of the questions suggests they may have previously served as training data for ChatGPT or Claude models. One of the strengths of the study is that the performances across all breast imaging modalities were evaluated separately. Zero-shot prompting was employed to standardize the different types of questions. Future studies designed on a larger scale, including MG, US, and MRI images and DICOM files, may shed light on the effectiveness of LLMs in medical image interpretation.

Conclusion

In conclusion, both GPT-4o and Claude 3.5 Sonnet have shown impressive performance in breast imaging across various modalities. This study underscores the potential of various LLMs to enhance daily clinical practice in breast imaging, suggesting their significant utility in future diagnostic workflows.

Conflict of interest

There is no financial support for the study, and the author has no conflicts of interest.

References

1. Kim S, Lee CK, Kim SS. Large Language Models: A Guide for Radiologists. Korean J Radiol. 2024;25(2):126-133. doi:10.3348/kjr.2023.0997
2. https://openai.com/index/hello-gpt-4o/ accessed on July 28, 2024
3. https://www.anthropic.com/news/claude-3-5-sonnet accessed on July 28, 2024
4. Sonoda Y, Kurokawa R, Nakamura Y, et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases. Jpn J Radiol. Published online July 1, 2024. doi:10.1007/s11604-024-01619-y
5. Oura T, Tatekawa H, Horiuchi D, et al. Diagnostic accuracy of vision-language models on Japanese diagnostic radiology, nuclear medicine, and interventional radiology specialty board examinations. Jpn J Radiol. Published online July 20, 2024. doi:10.1007/s11604-024-01633-0
6. Sorin V, Glicksberg BS, Artsi Y, et al. Utilizing large language models in breast cancer management: systematic review. J Cancer Res Clin Oncol. 2024;150(3):140. Published 2024 Mar 19. doi:10.1007/s00432-024-05678-6
7. Cozzi A, Pinker K, Hidber A, et al. BI-RADS Category Assignments by GPT-3.5, GPT-4, and Google Bard: A Multilanguage Study. Radiology. 2024;311(1):e232133. doi:10.1148/radiol.232133

8. Choi HS, Song JY, Shin KH, Chang JH, Jang BS. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiat Oncol J. 2023;41(3):209-216. doi:10.3857/roj.2023.00633
9. Almeida LC, Farina EMJM, Kuriki PEA, Abdala N, Kitamura FC. Performance of ChatGPT on the Brazilian Radiology and Diagnostic Imaging and Mammography Board Examinations. Radiol Artif Intell. 2024;6(1):e230103. doi:10.1148/ryai.230103
10. Haver HL, Bahl M, Doo FX, et al. Evaluation of Multimodal ChatGPT (GPT-4V) in Describing Mammography Image Features. Can Assoc Radiol J. Published online April 6, 2024. doi:10.1177/08465371241247043
11. Hirano Y, Hanaoka S, Nakao T, et al. GPT-4 Turbo with Vision fails to outperform text-only GPT-4 Turbo in the Japan Diagnostic Radiology Board Examination. Jpn J Radiol. 2024;42(8):918-926. doi:10.1007/s11604-024-01561-z
12. Payne DL, Purohit K, Borrero WM, et al. Performance of GPT-4 on the American College of Radiology In-training Examination: Evaluating Accuracy, Model Drift, and Fine-tuning. Acad Radiol. 2024;31(7):3046-3054. doi:10.1016/j.acra.2024.04.006
13. Horiuchi D, Tatekawa H, Oura T, et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. Eur Radiol. Published online July 12, 2024. doi:10.1007/s00330-024-10902-5
14. Sood A, Mansoor N, Memmi C, Lynch M, Lynch J. Generative pretrained transformer-4, an artificial intelligence text predictive model, has a high capability for passing novel written radiology exam questions. Int J Comput Assist Radiol Surg. 2024;19(4):645-653. doi:10.1007/s11548-024-03071-9