JOURNAL OF
CONTEMPORARY MEDICINE

DOI:10.16899/jcm.1626433
J Contemp Med 2025;15(5):245-249

Original Article / Orijinal Araştırma

A Comparative Analysis of The Diagnostic Efficacy of Diverse Artificial Intelligence (AI) Algorithms in Ultrasound-Based Cases

Ultrason Tabanlı Vakalarda Çeşitli Yapay Zeka (YZ) Algoritmalarının Tanısal Etkinliğinin Karşılaştırmalı Analizi

Başak Erdemli Gürsel, Gökhan Öngen, Dilek Sağlam

Bursa Uludağ University Faculty of Medicine, Department of Radiology, Bursa, Türkiye

Abstract

Aim: To evaluate the diagnostic performance of Large Language Models (LLM) (ChatGPT 3.5, ChatGPT 4, Gemini 1.0, and Gemini Advance) in Ultrasound (US) cases and their superiority over each other.

Material and Method: In this retrospective study, the data of 25 real cases with US examination and confirmed diagnoses were evaluated between 2020-2024. Clinical information, relevant laboratory data, and US findings of these cases were simultaneously presented to four Artificial Intelligence (AI) (ChatGPT 3.5, ChatGPT 4, Gemini 1.0, Gemini Advance). The correct response rates of the four AIs to the cases were compared. Two radiology experts in the US evaluated the answers.

Results: The correct response rates of ChatGPT 3.5, ChatGPT 4, Gemini 1.0, and Gemini Advance models in the cases were 92% (23/25), 92% (23/25), 76% (19/25), 84% (21/25), respectively, and with no statistically significant differences between them.

Conclusion: This is the first study about four AI performances in diagnosis in real US cases. The results suggest that no matter which AI we use, AIs have the potential to assist radiologists in diagnosis significantly. The fact that they are easy and fast to use can also significantly speed up the daily workflow. However, it should be remembered that they cannot yet completely replace a radiologist.

Keywords: Artificial intelligence, large language models, ChatGPT, Gemini, ultrasound

Öz

Amaç: Ultrason (US) vakalarında Geniş Dil Modellerinin (LLM) (ChatGPT 3.5, ChatGPT 4, Gemini 1.0 ve Gemini Advance) tanısal performansını ve birbirlerine göre üstünlüklerini değerlendirmek.

Gereç ve Yöntem: Bu retrospektif çalışmada, 2020-2024 yılları arasında US incelemesi yapılmış ve tanıları doğrulanmış 25 gerçek vakanın verileri değerlendirilmiştir. Bu vakaların klinik bilgileri, ilgili laboratuvar verileri ve US bulguları eş zamanlı olarak dört Yapay Zekaya (YZ) (ChatGPT 3.5, ChatGPT 4, Gemini 1.0, Gemini Advance) sunulmuştur. Dört YZ'nin vakalara doğru yanıt verme oranları karşılaştırılmıştır. Yanıtlar iki radyoloji uzmanı tarafından değerlendirmiştir.

Bulgular: ChatGPT 3.5, ChatGPT 4, Gemini 1.0 ve Gemini Advance modellerinin vakalardaki doğru yanıt oranları sırasıyla %92 (23/25), %92 (23/25), %76 (19/25), %84 (21/25) olup aralarında istatistiksel olarak anlamlı farklılık yoktur.

Sonuç: Bu çalışma, gerçek US vakalarıyla yapılmış, 4 YZ'nin tanı performanslarının değerlendirildiği ilk çalışmadır. Sonuçlar, hangi YZ'yi kullanırsak kullanalım, YZ'lerin radyologlara tanıda önemli ölçüde yardımcı olma potansiyeline sahip olduğunu göstermektedir. Kullanımlarının kolay ve hızlı olması da günlük iş akışını önemli ölçüde hızlandırabilir. Bununla birlikte, henüz gerçek bir radyoloğun yerini tamamen alamayacakları da unutulmamalıdır.

Anahtar Kelimeler: Yapay zeka, geniş dil modelleri, ChatGPT, Gemini, ultrason

Corresponding *(İletişim)*: Başak Erdemli Gürsel, Bursa Uludağ University Faculty of Medicine, Department of Radiology, Bursa, Türkiye
E-mail *(E-posta)*: basakgursel@uludag.edu.tr
Received *(Geliş Tarihi)*: 01.02.2025 **Accepted *(Kabul Tarihi)*:** 22.09.2025

INTRODUCTION

With the advancement of technology, artificial intelligence (AI) is increasingly prevalent in the medical sector, as in many other fields. Large Language Models (LLMs) are AI tools that can analyze and understand texts in natural language. These models are AI tools continuously trained with human input (such as books, articles, and web pages) and utilize deep learning techniques to create powerful predictive models. AI tools like ChatGPT, the pioneer of these models, can be utilized in various ways in radiology. In particular, they have the potential to help radiologists in many areas, such as determining radiation doses, creating imaging protocols and reports, interpreting radiological data, and even interpreting radiological images.[1] It is promising to improve workflow efficiency and accuracy of radiologic diagnoses by reducing interpretation variability and assessment errors among radiologists. However, there are some limitations, such as the need for high-quality data, further research and development to improve the performance and usability of the model, and ethical considerations.[2] Despite these challenges, ChatGPT can potentially impact radiology and medical imaging diagnosis significantly. However, despite this promising potential, it should be kept in mind that these models are not free from errors.[3,4]

In recent years, ChatGPT has been seen as the most used among LLMs in increasing AI studies. In 2023, Google's Gemini (formerly BARD), which uses the LaMDA language family, was ambitiously introduced with algorithms similar to ChatGPT. In recent years, there has been a notable increase in the use of ChatGPT, particularly in the context of AI studies, compared to other LLMs. In 2023, Google also launched Gemini (formerly known as BARD), which utilizes the LaMDA language family with algorithms similar to ChatGPT's, entering the market with significant ambition. However, a literature review indicates that the performance of Gemini or its predecessor, BARD, has not been sufficiently investigated in radiology. In our study, we aimed to evaluate the performance and superiority of ChatGPT 3.5, ChatGPT 4, Gemini 1.0, and Gemini Advance models by presenting real patient data from patients admitted to our clinic for various reasons.

MATERIAL AND METHOD

This retrospective study was approved by the institutional review board of our hospital (Decision number: 2024-4/6) and the investigations were carried out following the rules of the Declaration of Helsinki of 1975, which was revised in 2013. We investigated cases admitted to our hospital for various reasons and underwent ultrasound (US) examinations in the radiology clinic in 2020-2024. We selected 25 patients (n=25) >18 years old whose diagnoses were confirmed by histopathological or radiological/clinical follow-up or treatment. We choose cases commonly seen in daily practice and those that are less common and could be confusing. Cases were submitted to ChatGPT 3.5, ChatGPT 4, Gemini 1.0, and Gemini Advance within approximately one month (February-March). Prior to the submission of case information, all AI programs were instructed to evaluate the data from the perspective of an expert radiologist, and an expert in the field of ultrasound. The patients' demographics, complaints, examination findings, relevant laboratory data, and US findings were all presented simultaneously and in the same order. The laboratory data of the patients was extensive; however, only that which was indicative of differential diagnosis for the existing disease scenario was used. Laboratory data are given with reference ranges in parentheses. Ultrasound images of all patients were available in PACS (Picture Archiving and Communication Systems), but image uploading was not performed. Instead, US findings were described in a way that would not lead to any bias. As is known, different diseases can lead to similar complaints and clinical and radiologic findings. Therefore, a comprehensive evaluation of the cases is crucial for differential diagnosis. Presenting only the US images could result in complete and correct interpretations and a lengthy and complex list of differential diagnoses. Therefore, the cases' information was presented in its entirety. The question "What is your primary diagnosis?" was asked for all cases.

The responses from four different AIs were evaluated together by two expert radiologists with 15 and 12 years of experience in the US field, and a consensus was reached. In the presence of different opinions, a consensus was reached with the contribution of a senior expert. Answers were first categorized as true/false. The degree of accuracy of the answers was then scored using the Global Scale Criteria. Each response was scored individually. The Global Scale Criteria are as follows:

Score	Global Scale Criteria
1	Poor quality, poor flow of the site, most information missing, not at all useful for patients
2	Generally poor quality and poor flow, some information listed but many important topics missing, of very limited use to patients
3	Moderate quality, suboptimal flow, some important information is adequately discussed but others poorly discussed, somewhat useful for patients
4	Good quality and generally good flow, most of the relevant information is listed, but some topics not covered, useful for patients
5	Excellent quality and excellent flow, very useful for patients

Furthermore, three expert radiologists with at least 15 years of experience in ultrasound imaging classified the cases into two groups based on the degree of difficulty (easy/difficult) and the frequency of occurrence (common and rare). These groups were then used to compare the success rates of artificial intelligence in difficult and easy cases and common and rare cases.

Statistical Analysis

Statistical analyses were performed with SPSS version 28.0 software (IBM Corporation, Armonk, NY). Quantitative data was expressed in percentages. Fisher Freeman Halton test was used to compare quality scores of language models.

RESULTS

The correct response rates of ChatGPT 3.5, ChatGPT 4, Gemini 1.0, and Gemini Advance models in the cases were 92% (23/25), 92% (23/25), 76% (19/25), 84% (21/25), respectively. There was no statistically significant difference between the models regarding correct response rates. The Gemini Advanced, ChatGPT 4, ChatGPT 3, and Gemini 1.0 achieved excellent quality rates of 72% (12/25), 68% (17/25), 48% (11/25), and 44% (18/25), respectively. ChatGPT 3 and ChatGPT 4 revealed the lowest poor quality rates, which were both 8% (2/25), while the poor quality rates of Gemini 1.0 and Gemini Advanced were 12% (3/25), (Overall comparison p=0.194). The distribution of the models' responses according to the Global Scale Criteria is shown in **Table 1** and **Figure 1**.

Table 1. Distribution of responses of artificial intelligence models according to Global Scale Criteria

	ChatGPT3 N(%)	ChatGPT4 N (%)	Gemini N (%)	Gemini Ad N (%)	p
Poor quality	2 (20)	2 (20)	3 (30)	3 (30)	0.194
Generally poor quality	0 (0)	0 (0)	3 (75)	1 (25)	
Moderate quality	4 (44.4)	1 (11.1)	4 (44.4)	0 (0)	
Good quality	7 (36.8)	5 (26.3)	4 (21.1)	3 (15.8)	
Excellent quality	12 (20.7)	17 (29.3)	11 (19)	18 (31)	
Total	25 (25)	25 (25)	25 (25)	25 (25)	

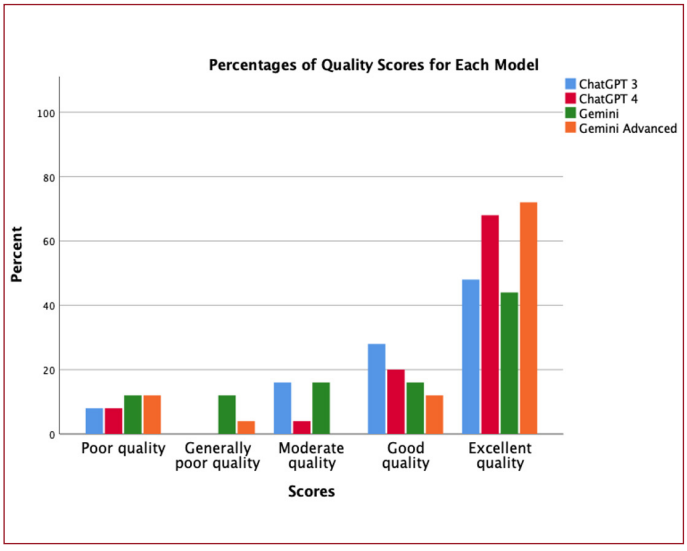

Figure 1. Distribution of responses of artificial intelligence models according to Global Scale

According to the difficulty level of the cases, 17 (68%) cases were grouped as easy and 8 (32%) as difficult. This classification found no significant differences in diagnostic performance among the four AI models in both easy and difficult cases (p=0.150 and p=0.580, respectively). The classification of the cases according to the degree of difficulty and the responses of the four AIs are shown in **Table 2**.

Table 2. Quality scores of each model according to degree of difficulty

	ChatGPT3	ChatGPT4	Gemini	Gemini Advanced	p
Easy (n=17)					**0.150**
Poor quality	1	2	2	2	
Generally poor quality	0	0	1	1	
Moderate quality	1	0	3	0	
Good quality	5	2	3	0	
Excellent quality	10	13	8	14	
Hard (n=8)					**0.580**
Poor quality	1	0	1	1	
Generally poor quality	0	0	2	0	
Moderate quality	3	1	1	0	
Good quality	2	3	1	3	
Excellent quality	2	4	3	4	

Based on the frequency of the cases, 12 (48%) cases were classified as common diseases, and 13 (52%) cases were classified as rare diseases. According to this classification, there was no significant difference in the success of the answers of the four artificial intelligences against easy and difficult cases (p=0.103, p=0.241, respectively). The quality scores of each model by frequency are shown in **Table 3**.

Table 3. Quality scores of each model according tofrequency

	ChatGPT3	ChatGPT4	Gemini	Gemini Adv.	p
Easy (n=17)					**0.103**
Poor quality	0	1	1	1	
Generally poor quality	0	0	0	0	
Moderate quality	1	0	2	0	
Good quality	3	1	4	0	
Excellent quality	8	10	5	11	
Hard (n=8)					**0.241**
Poor quality	2	1	2	2	
Generally poor quality	0	0	3	1	
Moderate quality	3	1	2	0	
Good quality	4	4	0	3	
Excellent quality	4	7	6	7	

DISCUSSION

The accuracy of diagnoses in radiology is increasing every year due to the development of technology. However, it is estimated that artificial intelligence models can prevent the loss of time in reaching the correct diagnosis in intensive clinical practice. Today, language-based AI models have introduced a new perspective to these resources. Due to the public accessibility, quick availability, and ease of use of AI models, their use by physicians in daily practice is steadily increasing.

There are a wide variety of studies and reviews in the literature evaluating the performance, limitations and future of ChatGPT in radiology.[5-9] Current applications of GPT-based models in radiology include report generation,

training support, clinical decision support, patient communication and data analysis. When we look at the radiological data and images used in these studies, most of them are cross-sectional methods such as CT and MRI. The number of studies using ultrasound data is relatively small. There are also studies focused on analyzing and simplifying US reports using ChatGPT.[10,11] Additionally, Allahqoli et al. reported that in their study of 30 cases prepared from an obstetrics and gynecology casebook, ChatGPT achieved a diagnostic accuracy rate of 90% (27/30 cases).[12] Wang et al. reported that on breast ultrasound diagnosis and reporting, artificial intelligence demonstrated comparable performance to an intermediate-level radiologist and could assist junior radiologists in BI-RADS classification.[13] Moro et al. reported that the published literature on artificial intelligence applied to ultrasound in benign gynecological disorders has primarily focused on creating classification models to distinguish between normal and pathological cases.[14] However, to the best of our knowledge, there is no other study in the literature that consists of only various US cases and compares the diagnostic performance of different artificial intelligence programs.

In our study, in real cases with different clinical information and US data, the correct response rates of ChatGPT 3.5, ChatGPT 4, Gemini 1.0, and Gemini Advance models were 92% (23/25), 92% (23/25), 76% (19/25), 84% (21/25), respectively, and with no statistically significant differences between them. In the literature, there are different studies evaluating the diagnostic performance of artificial intelligence (ChatGPT) in imaging findings, including patient information and different radiologic methods.[5,7] Ueda et al. found an overall accuracy rate of 54% (170/313 cases) in their study assessing the diagnostic performance of ChatGPT based on patient history and imaging findings.[5] Suthar et al., in their study evaluating the cases of the month published in AJNR (American Journal of Neuroradiology) between 2011 and 2023 with ChatGPT-4, reported an overall diagnostic accuracy rate of 57.86% (81/140 cases).[7] In this study, the clinical history was presented first, followed by imaging findings presented weekly for four weeks, as in the AJNR "Case of the Month" practice7. Our study results show that it is noteworthy that ChatGPT 3.5, ChatGPT 4, Gemini 1.0, and Gemini Advance had higher rates of correctly predicting cases (76-92%). This may be related to patient history, relevant laboratory data, and US imaging findings that are presented simultaneously. In daily radiology practice, when evaluating a patient, the more information we have about a patient, the higher the likelihood of reaching an accurate diagnosis. Therefore, when utilizing AI support, providing all relevant case data simultaneously and appropriately may increase the likelihood of the AI reaching an accurate diagnosis. In our study, there was no statistically significant difference between the success rates of the models. Therefore, we believe the important thing is the correct prompt and appropriate and sufficient patient data.

When the distribution of answers in difficult cases is analyzed in our study, it is noteworthy that ChatGPT 4 and Gemini Ad have a higher number of Moderate, Good, and Excellent quality answers compared to other models (8/8, 7/8, respectively). Classifying cases as easy/difficult and common/rare did not result in significant differences in the AI models' success rates. This suggests that AI has the potential to be a valuable assistant to radiologists, even in challenging or rare cases.

One of the limiting factors of our study is the relatively small number of cases, which could affect the results. Additionally, the fact that AI learns and develops from non-medical sources is another potential weakness. The fact that US images are not uploaded as images and are described can be considered a factor that may lead to bias. However, we attempted to present the findings as objectively as possible in a manner that even a less experienced radiologist could understand. We planned our study based on the description of the images, but different studies can be conducted in which images are uploaded to the artificial intelligence program.

CONCLUSION

These continuously generating models show great promise in interpreting radiological data and are being increasingly used. Our results indicate that regardless of which AI model is used, they are a valuable resource for radiologists. However, it is essential to remember that AI cannot replace an experienced radiologist, and the radiologist should always make the final diagnostic decision. As these models continue to progress and develop, the role and success of GPT-based models in the diagnostic process within radiology are likely to grow.

ETHICAL DECLARATIONS

Ethics Committee Approval: The research protocol was approved by the Bursa Uludağ University Health Sciences Research and Publication Ethics Committee (Date: 20.03.2024, Decision No: 2024-4/6).

Informed Consent: Because the study was designed retrospectively, no written informed consent form was obtained from patients.

Referee Evaluation Process: Externally peer-reviewed.

Conflict of Interest Statement: The authors have no conflicts of interest to declare.

Financial Disclosure: The author declared that this study has received no financial support.

Author Contributions: All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

REFERENCES

1. Biswas SS. Role of ChatGPT in radiology with a focus on pediatric radiology: proof by examples. Pediatr Radiol. 2023;53(5):818-22.
2. Srivastav S, Chandrakar R, Gupta S, et al. ChatGPT in Radiology: The Advantages and Limitations of Artificial Intelligence for Medical Imaging Diagnosis. Cureus. 2023;15(7):e41435.

3. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine. 2023;90:104512.
4. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023;11(6):887.
5. Ueda D, Mitsuyama Y, Takita H, et al. ChatGPT's Diagnostic Performance from Patient History and Imaging Findings on the Diagnosis Please Quizzes. Radiology. 2023;308(1):e231040.
6. Ueda, D., Walston, S.L., Matsumoto, T. et al. Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz. BMC Digit Health. 2024;2:4.
7. Suthar PP, Kounsal A, Chhetri L, Saini D, Dua SG. Artificial Intelligence (AI) in Radiology: A Deep Dive Into ChatGPT 4.0's Accuracy with the American Journal of Neuroradiology's (AJNR) "Case of the Month". Cureus. 2023;15(8):e43958.
8. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of ChatGPT. Diagn Interv Imaging. 2023;104(6):269-74.
9. Keshavarz P, Bagherieh S, Nabipoorashrafi SA, et al. ChatGPT in radiology: A systematic review of performance, pitfalls, and future perspectives. Diagn Interv Imaging. 2024;105(7-8):251-65.
10. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for Simplifying Radiology Reports. Radiology. 2023;309(2):e232561.
11. Li H, Moon JT, Iyer D, et al. Decoding radiology reports: Potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. Clin Imaging. 2023 Sep;101:137-41.
12. Allahqoli L, Ghiasvand MM, Mazidimoradi A, Salehiniya H, Alkatout I. Diagnostic and Management Performance of ChatGPT in Obstetrics and Gynecology. Gynecol Obstet Invest. 2023;88(5):310-3.
13. Wang J, Tian H, Yang X, et al. Artificial Intelligence in Breast US Diagnosis and Report Generation. Radiol Artif Intell. 2025;7(4):e240625.
14. Moro F, Giudice MT, Ciancia M et al. Application of artificial intelligence to ultrasound imaging for benign gynecological disorders: systematic review. Ultrasound Obstet Gynecol. 2025;65(3):295-302.