Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences

Araştırma makalesi / Research article

www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

Comparison of the accuracy performances of the Gemini Advanced, the GPT-4, the Copilot, and the GPT-3.5 models in medical imaging systems: A Zero-shot prompting analysis

Tıbbi görüntüleme sistemlerinde Gemini Advanced, GPT-4, Copilot ve GPT-3.5 modellerinin doğruluk performanslarının karşılaştırılması: Sıfır atışlı yönlendirme analizi

Alpaslan Koç[1*], **Ayşe Betül Öztiryaki**[2]

[1,2] *Samsun University, Biomedical Engineering Department, 55420, Samsun, Türkiye*

Abstract

Large Language Models (LLMs) have gained popularity across healthcare and attracted the attention of researchers of various medical specialties. Determining which model performs well in which circumstances is essential for accurate results. This study aims to compare the accuracy of recently developed LLMs for medical imaging systems and to evaluate the reliability of LLMs in terms of correct responses. A total of 400 questions were divided into four categories: X-ray, ultrasound, magnetic resonance imaging, and nuclear medicine. LLMs' responses were evaluated with a zero-prompting approach by measuring the percentage of correct answers. McNemar tests were used to evaluate the significance of differences between models, and Cohen kappa statistics were used to determine the reliability of the models. Gemini Advanced, GPT-4, Copilot, and GPT-3.5 resulted in accuracy rates of 86.25%, 84.25%, 77.5%, and 59.75%, respectively. There was a strong correlation between Gemini Advanced and the GPT-4 compared with other models, K=0.762. This study is the first that analyzes the accuracy of responses of recently developed LLMs: Gemini Advanced, GPT-4, Copilot, and GPT-3.5 on questions related to medical imaging systems. And a comprehensive dataset with three question types was created within medical imaging systems, which was evenly distributed from various sources.

Keywords: Large language models, Medical imaging systems, Generative ai, Comparison of the accuracy, Foundation models

Öz

Büyük dil modelleri (LLM'ler) sağlık hizmetlerinde popülerlik kazanmış ve çeşitli tıbbi uzmanlık alanlarındaki araştırmacıların ilgisini çekmektedir. Doğru sonuçlar için hangi modelin hangi koşullarda iyi performans gösterdiğini belirlemek önemlidir. Bu çalışma, yeni geliştirilen büyük dil modellerinin tıbbi görüntüleme sistemleri için doğruluklarını karşılaştırmayı ve bu modellerin verdikleri doğru yanıtlar açısından birbirleri arasındaki uyumluluklarını değerlendirmeyi amaçlamaktadır. Bu değerlendirme için toplam 400 soru X-ray, ultrason, manyetik rezonans görüntüleme ve nükleer tıp görüntüleme olarak dört kategoriye ayrılmıştır. Büyük dil modellerinin yanıtları, doğru yanıtların yüzdesi ölçülerek sıfır-atışlı yönlendirme yaklaşımıyla değerlendirilmiştir. Modeller arasındaki farkların anlamlılığını değerlendirmek için McNemar testi, modellerin güvenilirliğini belirlemek için ise Cohen kappa istatistiği kullanılmıştır. Gemini Advanced, GPT-4, Copilot ve GPT-3.5 için sırasıyla %86.25, %84.25, %77.5 ve %59.75 doğruluk oranları elde edilmiştir. Diğer modellerle karşılaştırıldığında Gemini Advanced ve GPT-4 arasında güçlü bir korelasyon bulunmuştur, K=0,762. Bu çalışma, yakın zamanda geliştirilen Gemini Advanced, GPT-4, Copilot ve GPT-3.5'in tıbbi görüntüleme sistemleriyle ilgili sorulara verdiği yanıtların doğruluğunu analiz eden ilk çalışmadır. Ayrıca bu çalışma ile tıbbi görüntüleme sistemleri ile ilgili çeşitli kaynaklardan üç soru tipinden oluşan kapsamlı bir veri seti oluşturulmuştur.

Anahtar kelimeler: Büyük dil modelleri, Tıbbi görüntüleme sistemleri, Üretken yapay zeka, Doğruluğun karşılaştırılması, Alt yapı modelleri

1 Introduction

A recent development in large language models (LLMs) has led to increased interest in LLMs that enable them to recognize, comprehend, analyze, and generate content. Through transformer-based architectures trained on a variety of text data, articles, websites, and books, the processes of learning, understanding, and generating text can be achieved [1]. Several LLMs have been developed to obtain more desired responses. OpenAI released GPT-3.5 (ChatGPT) and GPT-4 (ChatGPT Plus) based on the Generative Pretrained Transformer (GPT) architecture in the last two years [2, 3]. Furthermore, Microsoft made a new investment in OpenAI and developed the Bing Chat model, which is similar to

* Sorumlu yazar / Corresponding author, e-posta / e-mail: alpaslan.koc@samsun.edu.tr (A. Koç)
Geliş / Recieved: 29.05.2024 Kabul / Accepted: 30.07.2024 Yayımlanma / Published: .15.10.2024
doi: 10.28948/ngumuh.1492129

ChatGPT and known as Copilot [4]. Similarly, in December 2023, Google DeepMind released a new multimodal model called Gemini to replace Bard [5]. Today, these models have gained more attraction and are preferred across multiple disciplines due to their higher potential on providing responses, insights, suggestions, and even designing proposals.

Higher-order thinking requires advanced cognitive processes such as creating, analyzing, evaluating, and criticizing. The human brain processes and transfers deep information, which results in these processes. Researchers are trying to use deep and reinforcement learning methods; and attempt to build a basis for high-level thinking in machines and software. Investigating LLMs is a continuing concern within comprehensive implications in various fields of medicine and biomedical literature [6], including studies on clinical decision support (CDS) in radiology [7], medical and/or pathology examinations [8-13], symptoms [14], for the interpretation of radiological images [15].

One example of LLMs in medicine is to compare the performances of the GPT-4, the GPT-3.5, and the Med-PaLM on medical exams and benchmark datasets [8]. Nori et al. [8] evaluated the results for the zero-shot and the five-shot responses on text- and visual media-related questions and found that the GPT-4 performed significantly better than the GPT-3.5 and the Med-PALM. In another study, GPT-4 has been used to assess 52 radiological images obtained from Computed Tomography (CT), X-Ray, and Ultrasound (US) [15]. Brin et al. [15], however, found that GPT-4 did not provide reliable results for the interpretation of radiological images despite its potential uses in non-medical images. Huh et al. [12] compared the ChatGPT with medical students on the pathology examination in Korea. They observed that the performance of the ChatGPT was lower than that of the medical students. There was a 60.8% accuracy rate in the ChatGPT compared to 90.8% in the medical students' accuracy rate. Wang et al. [13] compared the knowledge ability of the ChatGPT with medical students using the Chinese National Medical Licensing Examination (NMLE) which belongs to the years 2020 to 2022. They found that the ChatGPT's performance was lower than that of the medical students. Gilson et al. [9] evaluated the ChatGPT compared with GPT-3 and InstructGPT on multiple choice questions from the United States Medical Licensing Examination (USMLE). The ChatGPT outperforms the other two methods by an 8.15% improvement with a 64.4% accuracy rate. Kung et al. [10] compared the performance of the ChatGPT on two different question models, the multiple-choice question, and the open-ended question, obtained from the USMLE. The ChatGPT model succeeded in three exams without reinforcement for the 60% threshold level. In another study, Sinha et al. [11] evaluated the ChatGPT's performance on 100 reasoning-type questions classified as higher-order knowledge in pathology. They categorized answers into five different scales and the ChatGPT's accuracy was around 80%. Efficient prompting techniques have been needed to reach more desired and correct outputs. However, LLMs' accuracy performances are varying and affected by inter- and intra-user variability. Therefore, zero-shot prompting analysis has been conducted in this study on recent large language models to avoid variability and subjectivity. The differences between the related works cited above and our study are summarized in Table 1. Knowing which model performs well in which situations and using it in education and research is essential to obtain accurate results. Although extensive research has been carried out on ChatGPT, no single study exists that compares the performance of the Gemini Advanced with other popular chatbots; GPT-4, Copilot, and GPT-3.5 on medical imaging systems in biomedical literature.

Our study contributes to the literature by addressing the following issues:

1) This study is the first that analyzes the accuracy of responses of recently developed LLMs such as Gemini Advanced, GPT-4, Copilot, and GPT-3.5 on questions related to medical imaging systems.
2) The performance of LLM models among various imaging modalities was also calculated for different question types.The compatibility of the LLM model was demonstrated for various question types in medical imaging systems.
3) A comprehensive dataset including three question types is created from various sources of medical imaging books.

2 Material and methods

2.1 Dataset

The dataset consists of four groups of questions and each group includes 100 questions covering medical imaging systems such as X-rays, US, MRIs, and nuclear medicine imaging systems. There are three subcategories of questions in each group: open-ended (OE), multiple-choice (MC), and computational questions (CQ). In each of these subgroups, the proportions of questions are arranged to have approximately equal weights using various sources of medical imaging systems [16-31]. The following are three different types of questions asked to LLMs. The Supplementary_file_dataset presents all questions with their references used in the study.

Open-ended questions are generally based on knowledge and interpretation.

Sample question: "*What determines the highest energy of x-ray photons emitted from an x-ray tube?*" The correct answer: "*The highest energy is determined by the peak x-ray tube voltage. For example, if the peak voltage is p kV, then the peak x-ray energy will be p keV*" [16].

Questions with at least two answer options are categorized as multiple-choice groups.

Table 1. Comparison of the state-of-the-art studies with our method

Related Works	**LLMs**	**Tasks**	**Results**	**Comparisons with our studies**
Şahin et al. (2024) [14]	ChatGPT, Ernie, Bard, Bing Copilot	To compare readability and quality of chatbots' responses	Bard is better than the other methods for readability	• Bard is replaced by a new model Gemini Advanced. • A relatively small number of questions were included in their study. • Our study is related to chatbots' accuracy on medical imaging. • Our study uses a new version GPT-4 and Gemini Advanced.
Brin et al. (2023) [15]	GPT-4	To assess radiological images	GPT-4 did not provide reliable results for the interpretation of radiological images.	• The study is related only radiological images. • A small number of images. • Lack of other LLMs: Gemini Advanced, Copilot, GPT-3.5
Huh et al. (2023) [12]	ChatGPT (the version cannot be defined)	To compare the ChatGPT and medical students on pathology examinations	The performance of the ChatGPT was lower than the medical students.	• A relatively small number of questions. • Lack of other LLMs: Gemini Advanced, Copilot, and GPT-4. • The study is related with the pathology exams.
Gilson et al. (2023) [9]	ChatGPT, GPT-3, InstructGPT	They compared 3 LLMs on USMLE	The ChatGPT outperforms the others.	• The previous version of ChatGPT is used. • A relatively small number of questions. • No various question types are tested. • Lack of other LLMs: Gemini Advanced, Copilot.
Kung et al. (2023) [10]	ChatGPT	To observe the performance of the ChatGPT on USMLE questions.	The ChatGPT model succeeded three exams without reinforcement.	• Lack of other LLMs: Gemini Advanced, Copilot, GPT-4. • Computational questions are not included.
Nori et al. (2023) [8]	GPT-4, GPT-3.5, Med-PALM	To compare the performance of the LLMs on medical challenge problems	the GPT-4 performed significantly better than the others.	• Lack of other LLMs: Gemini Advanced, Copilot. • The study is related to the medical challenge problems.
Wang et al. (2023) [13]	ChatGPT	To compare the knowledge ability of the ChatGPT with medical students using Chinese NMLE	The ChatGPT does not perform at the same level as students.	• Only the ChatGPT model is used. • Lack of other LLMs: Gemini Advanced, Copilot, GPT-4 • Variation of the question types is not considered.
Sinha et al. (2023) [11]	ChatGPT	To evaluate the performance of the ChatGPT on solving questions.	The ChatGPT's accuracy was around 80%.	• The study concerned only reasoning-type questions. • Lack of other LLMs: Gemini Advanced, Copilot. • Variation of the question types is not considered.
Rao et al. (2023) [7]	GPT-3.5 GPT-4	To evaluate GPT-3.5 and GPT-4 performance for clinical decision support (CDS) in radiology for breast cancer imaging and breast pain	ChatGPT-4 performed higher performances on radiology clinical decision-making tasks as compared with ChatGPT-3.5.	• The study only related to the breast cancer imaging devices and breast brain. • Lack of other LLMs: Gemini Advanced, Copilot. • Variation of the question types is not considered.
Our study	ChatGPT including GPT-3.5, GPT-4, Copilot, Gemini Advanced	To evaluate the performance of the various LLMs on various types of 400 questions related with medical imaging systems.	The accuracy rates of the LLMs were 86.25%, 84.25%, 77.5%, and 59.75% for the Gemini Advanced, the GPT-4, Copilot, and the GPT-3.5, respectively.	

Sample question: "*Of the digital radiographic detectors currently available, which is based on the use of photostimulable phosphors for image production?*
a. Computed radiography detectors
b. Indirect flat-panel detectors
c. Direct flat-panel detectors
d. Charge coupled device"
The correct answer: "*Computed radiography detectors*" [17].

In computational questions, simple equations and mathematical expressions such as "*sin (x")* or "*2x-5*" are written directly as text. The LaTeX script was used for more complex mathematical expressions.

Sample question: "Show that the decay factor DF is related to the half-life by DF = e^{\frac{0.693t}{T_{1/2}}}." The correct answer: $DF = e^{-\lambda t}, \frac{N(t)}{N(0)} = \frac{1}{2} = e^{-\lambda t_{1/2}}, ln2 = \lambda t_{1/2}, \lambda = \frac{0.693}{t_{1/2}}, DF = e^{-\frac{0.693}{t_{1/2}}t}$ obtained from the reference [16].

2.2 *Large Language Models*

LLMs use enormous data for training and successively manage text to produce coherent and context-sensitive responses within a conversational framework. The accuracy of the answers in response to specific tasks is an essential factor that increases competition between the chatbots. The answer accuracy performance of the latest models (GPT-3.5, GPT-4, Copilot, and Gemini Advanced) was compared and as a means of avoiding subjectivity, the zero-shot performance of the models was measured. We input each question separately to the GPT-3.5, GPT-4, Copilot, and Gemini Advanced language models. All answers given were compared with the correct answers, which were recorded as 1 if correct or 0 if incorrect. In order to prevent errors in checking answers, both authors used the cross-checking method. While asking questions, to prevent response-by response variation and cascading effects, each question was prompted three times with restarting the model, and highest number of identical responses were recorded as the answer of the model. A more creative conversation style was used for Copilot, whereas standard parameter settings were used for other models. Figure 1 presents an overview of the study protocol.

2.3 *Performance measurement and statistical analysis*

Each LLM's performance is evaluated by determining the percentage of correct answers given by the model to the total number of questions asked. Equation (1) defines the accuracy percentage (Acc%) of the model. The McNemar test was used to determine the significance of the difference between the groups, and the Cohen kappa (κ) statistic was used to assess their compatibility. All analyses were performed using SPSS version 23 [32].

$$\text{Acc\%} = \frac{\text{Number of correct answers}}{\text{Number of total questions in the group}} \text{x100} \qquad (1)$$

3 Results

A dataset consisting of 400 questions in total was produced, with three types of questions (OE, MC, and CQ) and four groups of 100 questions each (x-ray imaging, US, MRI, and nuclear medicine). Supplementary_file_results provides evaluations of all responses. Our first aim was to reveal the performance of chatbots on questions related to medical imaging systems. As seen in Table 2, the Gemini Advanced model showed the best performance in all groups, while GPT-3.5 showed the lowest performance. Gemini Advanced and GPT-4 performances are very close to each other.

Table 2. Comparison table for accuracies of LLMs.

	LLMs' Accuracies			
	GPT-3.5	**GPT-4**	**Copilot**	**Gemini Advanced**
X-RAY	62% (62/100)	82% (82/100)	80% (80/100)	84% (84/100)
US	59% (59/100)	87% (87/100)	77% (77/100)	89% (89/100)
MRI	60% (60/100)	82% (82/100)	76% (76/100)	83% (83/100)
NM	58% (58/100)	86% (86/100)	77% (77/100)	87% (87/100)
Total (Mean±SD)	59.75% ± 1.71% (239/400)	84.25% ± 2.63% (337/400)	77.5% ± 1.73% (310/400)	86.25% ± 2.75% (345/400)

US: Ultrasound, MRI: Magnetic resonance imaging, NM: Nuclear medicine.

Additionally, we measured the performance of LLMs on answering computational questions and compared them to other question types as shown in Table 3. The best accuracy performance is achieved in open-ended questions for all model types. However, the lowest performance was observed for computational questions. Specifically, GPT-3.5 and Copilot language models perform considerably worse than other methods on computational questions. As far as the accuracy of the Gemini Advanced language model and GPT-4 is concerned, both are superior to the others.

Table 3. LLMs accuracies for various question types in medical imaging systems

	LLMs' Accuracies			
	GPT-3.5	**GPT-4**	**Copilot**	**Gemini Advanced**
OE questions	76.26% (106/139)	92.81% (129/139)	87.77% (122/139)	**94.96% (132/139)**
MC questions	66.67% (88/132)	87.88% (116/132)	85.61% (113/132)	**89.39% (118/132)**
Computational questions	33.88% (45/129)	71.32% (92/129)	58.14% (75/129)	**73.64% (95/129)**
Total (Mean±SD)	59.75% ± 22.22% (239/400)	84.25% ± 11.26% (337/400)	77.5% ± 16.52% (310/400)	**86.25% ± 11.06% (345/400)**

OE: Open-ended, MC: Multiple-choice

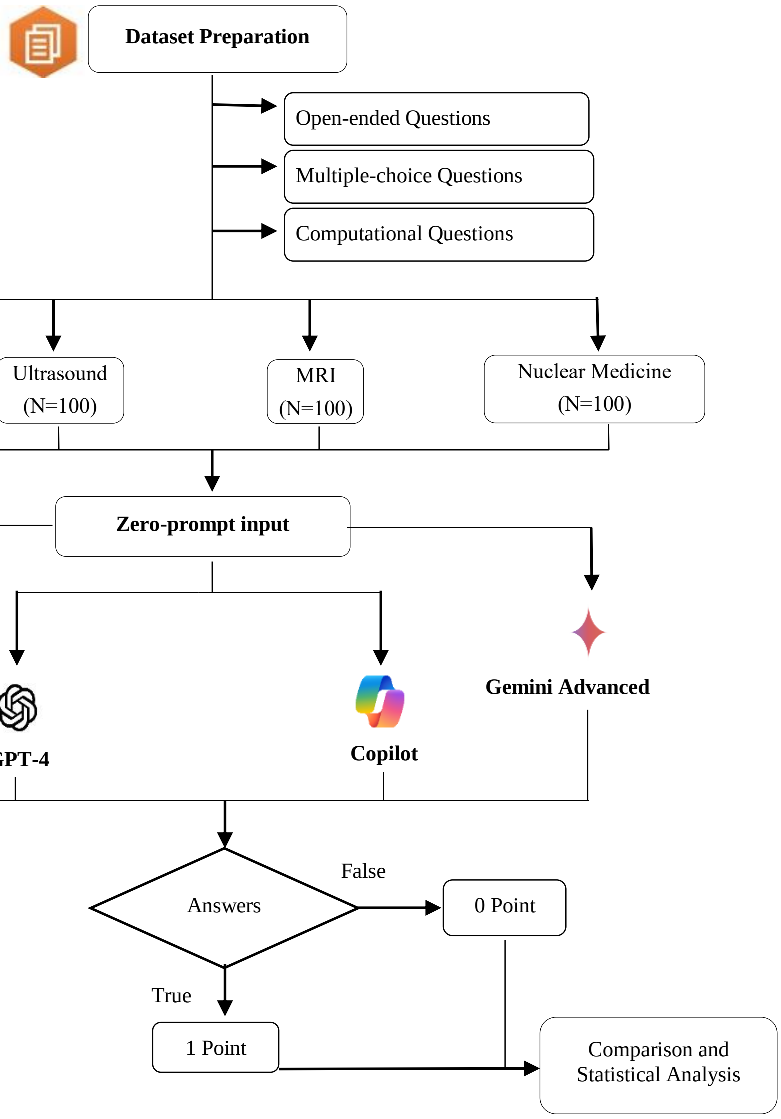

Figure 1. A schematic overview of the study protocol.

The significance level between the accuracy performances of the LLM models were analyzed in pairs using the McNemar test with a total data set of 400 questions. As shown in Table 4, statistically significant differences were observed for all groups (p<0.001) except GPT-4 vs Gemini Advanced (p=0.152). The highest difference, χ^2=104.009 and p<0.001, was obtained for the GPT-3.5 vs Gemini Advanced group. The lowest difference, χ^2=14.488 and p<0.001, was found for the GPT-4 vs Copilot group. The null hypothesis is rejected at a 5% significance level. According to the accuracy performances of the four models, Gemini Advanced, GPT-4, Copilot, and GPT-3.5 are arranged in decreasing order, but Gemini Advanced and GPT-4 have no statistically significant differences (p = 0.152).

Finally, we used Cohen's Kappa coefficient to determine how closely LLMs fit each other. A kappa value Ꝁ changes from 0 to 1, and 0 indicates there is no correlation, 1 represents almost a perfect relationship between the two groups [33]. Table 4 presents kappa values for six dyads based on four language models. The result showed that Gemini Advanced and GPT-4 are the most compatible and mainly produce correct answers to the same questions with similar accuracy (Ꝁ=0.762).

Table 4. Statistical analysis results for compared pair of groups.

	McNemar Test		Cohen's Kappa statistics (Ꝁ)
	χ^2	p	
GPT-3.5 vs GPT-4	96.01	.000	.434
GPT-3.5 vs Copilot	69.014	.000	.602
GPT-3.5 vs Gemini Advanced	**104.009**	.000	**.383**
GPT-4 vs Copilot	**14.488**	.000	.671
GPT-4 vs Gemini Advanced	-	**.152**	**.762**
Copilot vs Gemini Advanced	26.884	.000	.642

4 Discussion

Gemini Advanced and GPT-4 are the highest-performing models in all medical imaging systems, as shown in Figure 2. Despite having similar accuracy rates, the Gemini Advanced has a higher accuracy percentage ranging from 83% to 90% as opposed to the GPT-4 model's 82% to 87%. Gemini Advanced, however, did not provide the correct answer to eight questions including five computational and three multiple-choice questions, whereas GPT-4 did, as shown in

Table 5. As a result of Gemini Advanced's performance, 16 questions were correctly answered, including eight computations, five multiple-choice, and three open-ended questions, while GPT-4's performance required

assistance to get the right answers. It is difficult to determine the exact reason for this discrepancy due to differences in training data, model architecture, or other factors may have contributed to the difference in performance between the two versions.

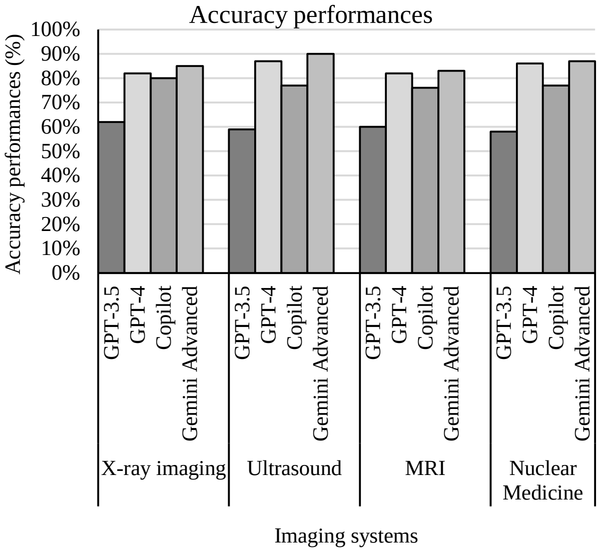

Figure 2. Comparison of accuracies of LLMs on various medical imaging fields

Previous studies cited in the introduction section have noted the importance of LLMs usage for educational, research, and teaching purposes in medicine and biomedical literature [7-15]. Our research included Google's newly developed model and other chatbots, and various question types were used as model inputs for medical imaging systems. The results showed that Gemini Advanced ranked first at 86.25%, followed by GPT-4 at 84.25%, Copilot at 77.50%, and GPT-3.5 at 59.75%. Moreover, according to Tables 2 and 3, the standard deviation of accuracy percentages varies from 1.71% to 2.75% for imaging modalities, and from 11.06% to 22.22% for question types. These results prove that question types significantly impact accuracy percentages, regardless of imaging modality types.

LLM models' compatibility with each other were evaluated and the result showed that the Gemini Advanced and the GPT-4 were more compatible, K = 0.762. In contrast, the lowest compatibility was observed between the Gemini Advanced and the GPT-3.5, K = 0.383. The compatibility of Copilot with other models was also evaluated and the highest kappa value was observed between Copilot and GPT-4 for the binary groups with Copilot, K = 0.671. The fact that Copilot uses the GPT-4 model proves that the two models are highly compatible. Compared to the studies in the introduction section, our study also contributes to the literature by investigating the compatibility between Gemini Advanced and other different models in terms of correct responses to the questions.

Table 5. 2x2 contingency tables indicating performances of Gemini Advanced and GPT-4

	Total (N=400)			**Computational questions (N=129)**	
	GPT-4 correct answer	GPT-4 incorrect answer		GPT-4 correct answer	GPT-4 incorrect answer
Gemini Advanced correct answer	329	16	Gemini Advanced correct answer	87	8
Gemini Advanced incorrect answer	8	47	Gemini Advanced incorrect answer	5	29
	Open-ended questions (N=139)			**Multiple-choice questions (N=132)**	
	GPT-4 correct answer	GPT-4 incorrect answer		GPT-4 correct answer	GPT-4 incorrect answer
Gemini Advanced correct answer	129	3	Gemini Advanced correct answer	113	5
Gemini Advanced incorrect answer	0	7	Gemini Advanced incorrect answer	3	11

Contrary to these advancements, we observed that Copilot and Gemini Advanced had longer response times than ChatGPT models. The ability to respond to questions quickly is an essential consideration for users. Due to its earlier release, GPT-3.5 is also free and has more users than other models. For students and researchers to make an informed decision regarding which model to use, they must be aware of LLMs' performance to determine and use the correct model based on the relevant tasks.

Our study has some limitations. We only used the zero-shot prompting technique to avoid subjectivity. The LLMs were not trained with any prompt inputs, and all models were restarted before each question. Another limitation is that the questions are only provided to the models as text input. Radiological images play an essential role in medical imaging systems, in addition to text questions. Brin et al. conducted a study on the interpretation of radiological images using GPT-4 and stated that GPT-4 did not give satisfactory results [15]. We therefore excluded radiological images and image-based questions from our study.

5 Conclusion

A recent development in LLMs has led to increased interest in foundation models that enable them to recognize, comprehend, analyze, and generate content. Determining which model performs well in which circumstances is important for better productivity. We determine the accuracy performance of LLM models on various questions related to medical imaging systems and investigate the reliability of the LLM models in their accurate response to questions for a specified task. The results showed that the Gemini Advanced model performed superior in all groups, while GPT-3.5 showed the lowest performance. Gemini Advanced and GPT-4 accuracy performances are very close to each other.

Additionally, the performance of LLMs for various types of questions was also evaluated. The best accuracy performance was observed in open-ended questions for all model types. However, the lowest performance was observed with computational questions. We also evaluated LLM models' compatibility with each other and found that the Gemini Advanced and the GPT-4 were more compatible. In the future, a comparative experiment will be conducted on humans using the dataset of this study.

Conflict of interest

Authors declear that there is no conflict of interest.

Similarty rate (iThenticate): 11%

References

[1] S. R. Bowman, Eight things to know about large language models, arXiv preprint arXiv:2304.00612, 2023. https://doi.org/10.48550/arXiv.2304.01964

[2] ChatGPT. https://chat.openai.com/ Accessed 27 Feb. 2024.

[3] GPT-4. https://openai.com/research/gpt-4, Accessed 27 Feb. 2024.

[4] Bing Chat: how to use Microsoft’s own version of ChatGPT Digital Trends. https://www.digitaltrends.com/computing/how-to-use-microsoft-chatgpt-bing-edge/, Accessed 27 Feb. 2024.

[5] Gemini - Google DeepMind. https://deepmind.google/technologies/gemini/#gemini-1.0, Accessed 28 Feb. 2024.

[6] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, Large language models in medicine, Nature medicine, vol. 29, no. 8, pp. 1930–1940, 2023. https://doi.org/10.1038/s41591-023-02448-8

[7] A. Rao, J. Kim, M. Kamineni, M. Pang, W. Lie, K. J. Dreyer, M. D. Succi, Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast ımaging pilot, Journal of the American College of Radiology, vol. 20, no. 10, pp. 990–997, 2023. https://doi.org/10.1016/j.jacr.2023.05. 003

[8] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, Capabilities of gpt-4 on medical challenge problems, arXiv preprint arXiv:2303.13375, 2023. https://doi.org/10.48550/arXiv.2303.13375

[9] A.Gilson, CW. Safranek, T. Huang, V. Socrates, L. Chi, RA. Taylor, D. Chartash, How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment, JMIR Medical Education, vol. 9, no. 1, p. e45312, 2023. doi:10.2196/45312

[10] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, V. Tseng, Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models, PLoS digital health, vol. 2, no. 2, p. e0000198, 2023. https://doi.org/10.1371/journal.pdig.0000198

[11] R. K. Sinha, A. D. Roy, N. Kumar, H. Mondal, and R. Sinha, Applicability of ChatGPT in assisting to solve higher order problems in pathology, Cureus, vol. 15, no. 2, 2023. doi: 10.7759/cureus.35237

[12] S. Huh, Are ChatGPT’s knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: A descriptive study,” Journal of educational evaluation for health professions, vol. 20, 2023. https://doi.org/10.3352/jeehp.2023.20.1

[13] X.Wang, Z. Gong, G. Wang, J. Jia, Y. Xu, J. Zhao, Q. Fan, S. Wu, W. Hu, X. Li, ChatGPT performs on the Chinese national medical licensing examination, Journal of Medical Systems, vol. 47, no. 1, p. 86, 2023. https://doi.org/10.1007/s10916-023-01961-0

[14] M. F. Şahin, H. Ateş, A. Keleş, Ç. Doğan, M. Akgül, C. M. Yazıcı, R. Özcan, Responses of five different artificial ıntelligence chatbots to the top searched queries about erectile dysfunction: A comparative analysis, Journal of Medical Systems, vol. 48, no. 1, p. 38, 2024. https://doi.org/10.1007/s10916-024-02056-0

[15] D. Brin, V. Sorin, Y. Barash, E. Konen, B. S. Glicksberg, G. N. Nadkarni, E. Klang, Assessing GPT-4 multimodal performance in radiological ımage analysis, medRxiv, pp. 2023–11, 2023. https://doi.org/10.1007/s00330-024-11035-5

[16] J. L. Prince and J. M. Links, Medical imaging signals and systems, vol. 37. Pearson Prentice Hall Upper Saddle River, 2006.

[17] E. Seeram, Medical Imaging Informatics, Digital Radiography: Review Questions, pp. 85–95, 2021.

[18] K. H. Ng, J. H. D. Wong, and G. Clarke, Problems and solutions in medical physics: Diagnostic Imaging Physics. CRC Press, 2018.

[19] W. R. Hendee and E. R. Ritenour, Medical imaging physics. John Wiley & Sons, 2003.

[20] G. Sawhney, Fundamental of biomedical engineering. New Age International, 2007.

[21] A. P. Dhawan, Medical image analysis. John Wiley & Sons, 2011.

[22] B. H. Brown, R. H. Smallwood, D. C. Barber, P. Lawford, and D. Hose, Medical physics and biomedical engineering. CRC Press, 2017.

[23] J. A. Miller, Review Questions for Ultrasound: A Sonographer’s Exam Guide. Routledge, 2018.

[24] C. K. Roth and W. H. Faulkner Jr, Review questions for MRI, 2013.

[25] S. C. Bushong and G. Clarke, Magnetic resonance imaging: physical and biological principles. Elsevier Health Sciences, 2003.

[26] H. Azhari, J. A. Kennedy, N. Weiss, and L. Volokh, From Signals to Image. Springer, 2020.

[27] W. A. Worthoff, H. G. Krojanski, and D. Suter, Medical physics: exercises and examples. Walter de Gruyter, 2013.

[28] M. Chappell, Principles of Medical Imaging for Engineers. Springer, 2019.

[29] E. Mantel, J. S. Reddin, G. Cheng, and A. Alavi, Nuclear Medicine Technology: Review Questions for the Board Examinations. Cham: Springer International Publishing, 2023. https://link.springer.com/10.1007/978-3-031-26720-8, Accessed 20 Mar. 2024.

[30] K. H. Ng, C. H. Yeong, and A. C. Perkins, Problems and Solutions in Medical Physics: Nuclear Medicine Physics, 1st ed. CRC Press, 2019. https://www.taylorfrancis.com/books/9780429629129, Accessed 20 Mar. 2024.

[31] D. D. Feng, Biomedical information technology. Academic Press, 2011.

[32] IBM SPSS Statistics for Windows. IBM Corp., Armonk, NY, Released 2015.

[33] M. L. McHugh, Interrater reliability: the kappa statistic, Biochemia medica, vol. 22, no. 3, pp. 276–282, 2012.