Uludağ Üniversitesi Tıp Fakültesi Dergisi
Journal of Uludağ University Medical Faculty
50 (3) 551-556, 2024
DOI: https://doi.org/10.32708/uutfd.1568479

ORIGINAL RESEARCH

Comparative Performance Evaluation of Multimodal Large Language Models, Radiologist, and Anatomist in Visual Neuroanatomy Questions

Yasin Celal GÜNEŞ[1], Mehmet ÜLKİR[2]

[1] Kirikkale Yuksek Ihtisas Hospital, Department of Radiology, Kirikkale, Türkiye.
[2] Hacettepe University Faculty of. Medicine, Department of Anatomy, Ankara, Türkiye.

ABSTRACT

This study examined the performance of four different multimodal Large Language Models (LLMs)—GPT4-V, GPT-4o, LLaVA, and Gemini 1.5 Flash—on multiple-choice visual neuroanatomy questions, comparing them to a radiologist and an anatomist. The study employed a cross-sectional design and evaluated responses to 100 visual questions sourced from the Radiopaedia website. The accuracy of the responses was analyzed using the McNemar test. According to the results, the radiologist demonstrated the highest performance with an accuracy rate of 90%, while the anatomist achieved an accuracy rate of 67%. Among the multimodal LLMs, GPT-4o performed the best, with an accuracy rate of 45%, followed by Gemini 1.5 Flash at 35%, ChatGPT4-V at 22%, and LLaVA at 15%. The radiologist significantly outperformed both the anatomist and all multimodal LLMs ($p<0.001$). GPT-4o significantly outperformed GPT4-V and LLaVA ($p<0.001$), but no significant difference was found between GPT-4o and Gemini 1.5 Flash ($p=0.123$). However, Gemini 1.5 Flash showed significant superiority over LLaVA ($p<0.001$) and also demonstrated a statistically significant difference compared to GPT4-V ($p=0.004$). This study highlights the significant performance gap between multimodal LLMs and medical professionals. While multimodal LLMs hold great potential in the medical field, they have not yet reached the level of accuracy of medical experts in correctly identifying neuroanatomical regions.

Keywords: Neuroanatomy. Large language models. GPT-4o. Gemini 1.5 Flash.

Çok Modlu Büyük Dil Modelleri, Bir Radyolog ve Bir Anatomistin Görsel Nöroanatomi Sorularındaki Karşılaştırmalı Performans Değerlendirmesi

ÖZET

Bu çalışma, dört farklı çok modlu Büyük Dil Modeli'nin (GPT4-V, GPT-4o, LLaVA, Gemini 1.5 Flash) görsel nöroanatomi çoktan seçmeli sorularındaki performansını, bir radyolog ve bir anatomistle karşılaştırarak incelemiştir. Kesitsel bir araştırma dizaynına dayanan çalışmada, Radiopaedia web sitesinden alınan 100 görsel soruya verilen yanıtlar değerlendirilmiştir. Yanıtların doğruluğu McNemar testi kullanılarak analiz edilmiştir. Sonuçlara göre, radyolog %90 doğruluk oranı ile en yüksek performansı sergilerken, anatomist %67 doğruluk oranı elde etmiştir. Çok modlu LLM'ler arasında en iyi performansı %45 doğruluk oranı ile GPT-4o göstermiştir; onu %35 ile Gemini 1.5 Flash, %22 ile ChatGPT4-V ve %15 ile LLaVA takip etmiştir. Radyolog, hem anatomiste hem de tüm çok modlu LLM'lere kıyasla anlamlı derecede üstün bir performans sergilemiştir ($p<0.001$). GPT-4o, GPT4-V ve LLaVA'ya kıyasla anlamlı derecede daha iyi bir performans göstermiş ($p<0.001$), ancak Gemini 1.5 Flash ile arasında anlamlı bir fark gözlenmemiştir ($p=0.123$). Bununla birlikte, Gemini 1.5 Flash, LLaVA'ya karşı anlamlı bir üstünlük sağlamış ($p<0.001$) ve GPT4-V ile karşılaştırıldığında da istatistiksel olarak anlamlı bir fark ortaya çıkmıştır ($p=0.004$). Bu çalışma, çok modlu LLM'ler ile tıbbi uzmanlar arasındaki belirgin performans farkını ortaya koymaktadır. Çok modlu LLM'ler tıp alanında büyük bir potansiyel vaat etse de, nöroanatomik bölgeleri doğru bir şekilde tanımlama konusunda henüz tıbbi uzmanların doğruluk seviyesine ulaşamamaktadırlar.

Anahtar Kelimeler: Nöroanatomi. Büyük dil modelleri. GPT-4o. Gemini 1.5 Flash.

Date Received: October 16, 2024
Date Accepted: January 02, 2025

Dr. Yasin Celal GÜNEŞ
Kırıkkale Yüksek İhtisas Hastanesi,
Radyoloji Department,
Kırıkkale, Türkiye.
Phone:+90 506 242 20 72
E-mail: gunesyasincelal@gmail.com

Authors' ORCID Information:
Yasin Celal GÜNEŞ: 0000-0001-7631-854X
Mehmet ÜLKİR: 0000-0001-5615-8913

Artificial Intelligence (AI) tools known as large language models (LLMs) are trained to process and generate text at a level that closely resembles human abilities. One of the competencies of LLMs is their ability to respond to inquiries, translate text, paraphrase, and summarize after processing various inputs[1]. The release of GPT-4 in March 2023 was significant for multimodal LLMs. GPT-4, also known as Generative Pre-Training Transformer-4th series

with Vision (GPT4-V), introduced advanced image evaluation capabilities[2].

GPT4-V and Large Language-and-Vision Assistant (LLaVA) are multimodal LLMs with image analysis capabilities that allow them to tackle more complex situations by combining language and visual information (3). The latest multimodal LLMs, GPT-4o and Gemini 1.5 Flash, were released in May 2024[4-5].

LLMs serve as a valuable resource for medical professionals, providing rapid access to comprehensive information on anatomy, surgical techniques, and postoperative care. Furthermore, LLMs can create interactive quizzes and educational tools that allow students to evaluate their skills and receive instant feedback[6]. Accurately identifying neuroanatomical landmarks is essential for radiologists to diagnose pathologies and for surgeons to perform neurosurgical and endovascular procedures effectively[7].

Recently, numerous articles have discussed the potential applications of LLMs in medical fields such as dermatology, pediatrics, radiology, anatomy, otolaryngology, and forensic science[8-13]. Most previous studies have focused on the integration of LLMs with only text-based capabilities. However, with the development of multimodal LLMs, visual data can now be evaluated to accurately diagnose pathologies in photos, interpret radiology images, and solve board examinations[14-16].

Despite advancements in multimodal LLMs, to our knowledge, there are no studies evaluating the performance of these models on visual neuroanatomy multiple choice questions (MCQs). The aim of this study is to investigate and compare the performance of radiologists, anatomists, and multimodal LLMs on visual neuroanatomy MCQs covering spatial anatomy and various radiological images.

Material and Method

Study design

This cross-sectional observational study compared multimodal LLMs (GPT4-V, GPT-4o, LLaVA, Gemini 1.5 Flash), and the responses of radiologists and anatomists in solving visual neuroanatomy MCQs. The study did not require ethics committee approval as it relied solely on open-access published online MCQs and and did not involve any human subjects or identifiable patient information. This study followed the Standards for Reporting Diagnostic Accuracy Studies (STARD) and the Checklist for Artificial Intelligence in Medical Imaging (CLAIM)[17-18].

Data collection

Radiopaedia provides publicly available multiple-choice questions (MCQs) assessing knowledge of cross-sectional anatomy on its website. In this study, we utilized a comprehensive dataset of 3,904 MCQs spanning various body systems. Each question had 4–6 choices with one correct answer and included both text-based and visual questions, available on the Radiopaedia website (Courtesy of Dr. Frank Gaillard; accessed September 2023; URL: https://radiopaedia.org/questions).

Among these, 964 questions were specifically related to the central nervous system (CNS), and within these CNS questions, 347 included associated images. From this subset, we identified 166 questions focusing on anatomical topics, particularly neuroanatomy. We randomly selected 100 anatomy-related questions from these 166 using a computer-generated random number sequence to ensure a representative sample. This selection aimed to include a balanced distribution of questions covering both spatial anatomy (visual questions involving anatomical structures without imaging modalities) and radiological image interpretation (requiring analysis of images from modalities such as MRI and CT).

Among the chosen 100 questions, 55 were non-contrast MRI scans (55%), 18 were non-contrast CT scans (18%), and 27 were spatial anatomy questions (27%). All selected questions are listed in Supplementary Material 1, and the workflow of the study is detailed in Figure 1.

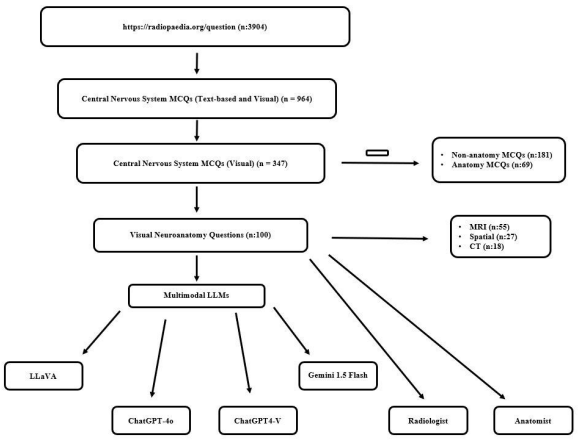

Figure 1.
Demonstration of workflow of study. MCQs: multiple-choice questions, GPT4-V: Generative Pre-Training Transformer-4th series with Vision, LLaVA: Large Language-and-Vision Assistant, GPT-4o: Generative Pre-Training Transformer-4th series omni, MRI: Magnetic resonance imaging, CT: Computed tomography

Prompt design and performance evaluation for multimodal LLMs

We initiated the input prompt in our study design as: "I am working on a neuroanatomy quiz and will provide you visual cases and multiple-choice questions. Act like a medical professor, please indicate the correct answer. There is only one correct answer." This prompt was presented in September 2024 on two distinct platforms with default hyperparameters, OpenAI's GPT4-V (https://chat.openai.com) and LLaVA (https://llava-vl.github.io) by radiologist (Y.C.G.). Subsequently, same prompt was presented in September 2024 on two distinct platform with identical parameters, OpenAI's GPT-4o (https://chat.openai.com) and Google's Gemini 1.5 Flash (https://deepmind.google/technologies/gemini/flash) (Figure 2).

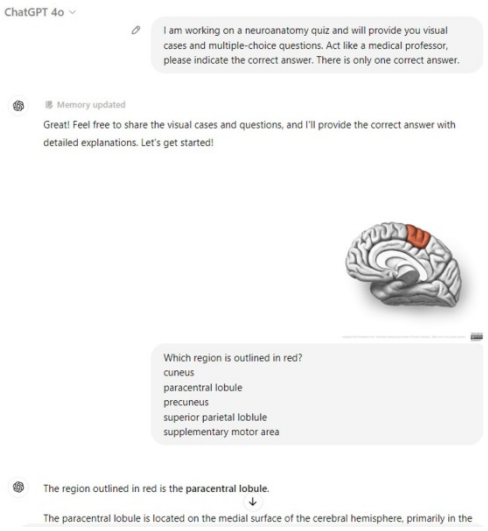

Figure 2.
Illustration of prompt and answer of LLM

The visual questions were sequentially added to the same chat session. Each multimodal LLMs was presented in 100 questions, and responses were recorded. Multimodal LLMs were not pretrained with a specific command or question set for this study. Each question was posed in a single chat session, without opening a new chat tab for individual inquiries. Radiologist (Y.C.G.) and anatomist (M.Ü) jointly evaluated the multimodal LLMs' answers according to the correct answer list provided by Radiopaedia either correct (1) or incorrect (0).

Radiologist and anatomist performance evaluation

Board-certified (EDiR) radiologist (Y.C.G.) and anatomist (M.Ü.), each with 6 years of experience, independently assessed the visual questions using their own computers. Upon completion of questions, they evaluated each other's answers according to the correct answer list provided by Radiopaedia either correct (1) or incorrect (0).

Statistical analysis

Basic descriptive statistics, including counts and percentages, were employed to analyze the performance of GPT4-V, GPT-4o, LLaVA, Gemini 1.5 Flash, radiologists, and anatomists. McNemar's test was utilized to compare the proportions of correct responses among these groups. All statistical analyses were conducted using SPSS 26.0, with statistical significance defined as $p<0.05$.

Results

A total of 100 visual neuroanatomy MCQs were included in the study. The radiologist correctly answered 90% (90/100 questions), surpassing the anatomist who scored 67% (67/100 questions). GPT-4o responded accurately to 45% (45/100 questions), followed by Gemini 1.5 Flash with 35% (35/100 questions), GPT4-V with 22% (22/100 questions), and LLaVA with 15% (15/100 questions) (Table I, Figure 3).

Table I. Diagnostic accuracy and classification by question types

	Accuracy (MRI)	Accuracy (CT)	Accuracy (Spatial)	Total Accuracy
Radiologist	96,3% (53/55)	100 (18/18)	70.4% (19/27)	90.0% (90/100)
Anatomist	63.6% (35/55)	44.4% (8/18)	88.9% (24/27)	67.0% (67/100)
GPT4-V	12.7% (7/55)	27.8% (5/18)	37.0% (10/27)	22.0% (22/100)
LLaVA	10.9% (6/55)	16.7% (3/18)	22.2% (6/27)	15.0% (15/100)
Gemini 1.5 Flash	36.3% (20/55)	44.4% (8/18)	25.9% (7/27)	35.0% (35/100)
GPT-4o	43.6% (24/55)	16.7% (3/18)	66.7% (18/27)	45% (45/100)

GPT4-V: Generative Pre-Training Transformer-4th series with Vision, LLaVA: Large Language-and Vision Assistant, GPT-4o: Generative Pre-Training Transformer-4th series omni, MRI: non-contrast magnetic resonance imaging, C+MRI: contrast enhanced magnetic resonance imaging, CT: non-contrast computed tomography, DSA: digital subtraction angiography.

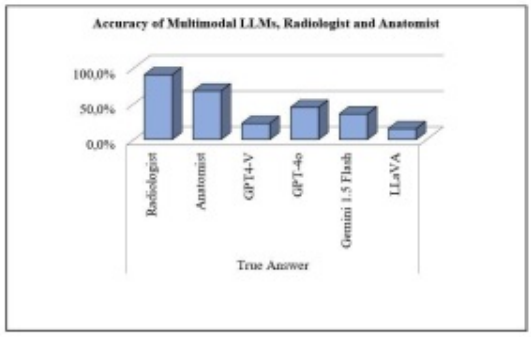

Figure 3.
Demonstration of accuracy of multimodal LLMs, radiologist and anatomist. LLMs: Large Language Models, GPT4-V: Generative Pre-Training Transformer-4th series with Vision, LLaVA: Large Language-and-Vision Assistant, GPT-4o: Generative Pre-Training Transformer-4th series omni.

Comparatively, the radiologist demonstrated significantly higher diagnostic accuracy than the anatomist (p=0.008). Both medical professionals outperformed the multimodal LLMs (p<0.05).

Among the multimodal LLMs, GPT-4o exhibited the highest rate of correct responses. Its performance significantly surpassed that of GPT4-V and LLaVA (p<0.001), while showing no significant difference compared to Gemini 1.5 Flash (p=0.123). Furthermore, the Gemini 1.5 Flash demonstrated a significant superiority over LLaVA and GPT4-V (p<0.05). GPT4-V correctly answered more questions than LLaVA, which was statistically significant (p=0.016) (Table II).

Table II. Comparison of diagnostic accuracy of multimodal LLMs, radiologist and anatomist with p-values obtained from McNemar's Test

	Radiologist	Anatomist	GPT4-V	LLaVA	GPT-4o	Gemini 1.5 Flash
Radiologist	-	<0.001	<0.001	<0.001	<0.001	<0.001
Anatomist	<0.001	-	<0.001	<0.001	0.015	0.002
GPT4-V	<0.001	<0.001	-	0.016	0.001	0.004
LLAVa	<0.001	<0.001	0.016	-	<0.001	<0.001
GPT-4o	<0.001	0.015	0.001	<0.001	-	0.123
Gemini 1.5 Flash	<0.001	0.002	0.004	<0.001	0.123	-

Radiologists demonstrated the highest performance across all question types except for spatial anatomy questions. In spatial anatomy questions, anatomists exhibited the highest accuracy rate (88.9%) compared to radiologists (70.4%). Among the LLMs, GPT-4o achieved the highest success rate (66.7%) in spatial anatomy questions.

Discussion and Conclusion

In our study, radiologists exhibited superior performance in answering visual neuroanatomy questions, significantly outperforming anatomists and multimodal LLMs (p<0.001). Anatomists showed better performance than multimodal LLMs (p<0.05). GPT-4o outperformed other multimodal LLMs (p<0.05), exception of Gemini 1.5 Flash (p=0.123). The superior performance of these two multimodal LLMs compared to GPT-4V and LLaVA may be attributed to their more recent training with larger and more advanced datasets. Although GPT-4V demonstrated the highest performance among the multimodal LLMs in our study, its accuracy of 45% indicates that these models currently lack sufficient proficiency in visual neuroanatomy. This underscores the need for further development and training of LLMs with specialized medical image datasets.

Notably, radiologists showed the highest performance in questions involving radiological evaluation, whereas anatomists provided more correct answers, particularly in spatial anatomy questions. The performance differences between radiologists and anatomists can be attributed to the fact that the majority of questions were related to sectional anatomy through radiological imaging methods. The higher performance of anatomists in spatial anatomy questions may be due to their exposure to a greater number of spatial and non-spatial anatomy questions during their training. This study suggests that anatomists should receive more training in sectional anatomy based on radiological imaging during their education.

There are studies in the literature evaluating the performance of large language models in anatomy questions. Bolgova et al. conducted a study assessing ChatGPT 3.5's performance in answering text-based multiple-choice questions (MCQs) across various anatomical regions[19]. Out of a total of 325 questions, ChatGPT 3.5 successfully answered 44.1% of them. Specifically focusing on neuroanatomy questions pertaining to the head and neck region, it achieved an approximate success rate of 48.8% out of 50 questions[19]. Ilgaz et al.'s study revealed that both ChatGPT 3.5 and Google Bard performed below 50% accuracy in answering text-based non-spatial anatomy questions. Furthermore, the study found no statistically significant difference in ChatGPT 3.5's performance between anatomy questions asked in Turkish and English[12].

Studies have highlighted the utility of LLMs in anatomy education. Lee indicated that integrating ChatGPT into anatomy education could improve

efficacy and students' engagement in the subject. However, concerns were raised regarding ChatGPT's tendency to generate hallucinations and provide inaccurate responses[20]. Mogaliet al. showcased ChatGPT's potential as an online anatomy tutor[21]. Similarly, Totlis et al. demonstrated the effectiveness of GPT 4 in generating and addressing various types of anatomy-related questions for learning purposes[22]. The low performance of the multimodal language models (LLMs) in recognizing neuroanatomical regions in our study precludes their consideration as a reliable standalone source for visual neuroanatomy education.

Recent advancements in multimodal LLMs, driven by the development of visual evaluation of images that have led to the creation of models tailored to the healthcare domain, such as LLaVA and CLIP[23]. However, most studies evaluating the performance of multimodal LLMs have primarily focused on X-rays[24]. The inclusion of images from different radiological modalities in our study may pose a challenge for multimodal LLMs in providing accurate answers. It is necessary to conduct studies utilizing different radiological modalities in order to demonstrate the efficacy of multimodal LLMs in clinical settings.

Zhu et al. demonstrated that GPT-4V was able to accurately diagnose medical conditions with a 77% accuracy rate when given visual USMLE-style questions[25]. However, when patient history was removed, the accuracy rate dropped to 19.54% This suggests that the model relies heavily on patient history to make accurate diagnoses. Node et al. found that the model's accuracy varied depending on the type of question, with image-based questions being more challenging in answering questions from the otolaryngology board certification exam[26]. The correct answer rate was 30.4% when only text was provided, but increased to 41.3% when images were also included[26].

Nakao et al. tested GPT4-V's ability to recognize images in the Japanese National Medical Licensing Examination[16]. The model was able to correctly answer 68% of image-based questions and 72% of text-based questions. It is noteworthy that there was no significant difference in the model's performance on image-based versus text-based questions. In contrast to Nakao et al.'s study, our study demonstrated that both GPT-4o (45.3%) and GPT4-V (22.6%) performed lower in visual neuroanatomy questions. We believe that the differences in clinical history and prompts may have caused these varying performances among studies. Overall, these studies suggest that multimodal LLMs like GPT4-V have the potential to be useful tools in radiology, but may require further development to reach their full potential in the future. Moreover, the Gemini 1.5 Flash demonstrated comparable performance to GPT models, which may indicate the remarkable potential for further development in this field.

There are few studies comparing the diagnostic performance of GPT4-V and LLaVA in visual images, and these studies are primarily related to melanoma. Cirone et al. demonstrated that GPT4-V outperformed LLaVA in all evaluated aspects, achieving an overall accuracy of 85%, whereas LLaVa achieved 45%. GPT4-V consistently provided detailed descriptions of relevant features of melanoma[14]. Similarly, Akrout et al. also found that GPT4-V performed better than LLaVA across all assessed features of melanoma[27]. Our study is also consistent with these studies regarding GPT4-V has better performance than LLaVA regarding image interpretation.

Limitations

Although our study makes a significant contribution to the comparison between multimodal LLMs and medical professionals, it has some limitations. Firstly, the number of visual questions in the study is limited, which may not fully capture the complexity of neuroanatomy. A larger set of questions could provide a more accurate assessment of multimodal LLMs' competence. Secondly, the use of different modalities in the study provides heterogeneous information about multimodal LLMs' performance, but future studies should test the performance of multimodal LLMs separately for each radiological modality and visual anatomy question to gain a more nuanced understanding. Thirdly, the small sample size, consisting of only one radiologist and one anatomist, may limit the generalizability of the findings. Including a larger cohort with varying levels of experience could provide more comprehensive insights. Lastly, the choice of prompt using the role-play technique may have influenced multimodal LLMs' performance. Prompts made using zero-shot and few-shot techniques could provide more detailed information about multimodal LLMs' performance in future studies.

In conclusion, this study provides valuable insights into the comparative performance of multimodal LLMs and medical professionals in visual neuroanatomy assessment. While multimodal LLMs demonstrate potential, they are not yet capable of accurately identifying neuroanatomical regions. Further research and development are necessary to bridge the gap between the capabilities of LLMs and human expertise regarding neuroanatomical knowledge.

Fikir ve tasarım: Y.C.G., M.Ü.; Veri toplama ve işleme: Y.C.G., M.Ü.; Analiz ve verilerin yorumlanması: Y.C.G., M.Ü.; Makalenin önemli bölümlerinin yazılması: Y.C.G., M.Ü.

Ethics Committee Approval Information:

Since this study was conducted using publicly available internet data and the images did not contain patient information, the study did not require an ethics committee. The study was conducted in accordance with the Standards for Reporting Studies on Diagnostic Accuracy (STARD) and the Checklist for Artificial Intelligence in Medical Imaging (CLAIM).

Researcher Contribution Statement:

Idea and design: Y.C.G., M.Ü.; Data collection and processing: Y.C.G., M.Ü.; Analysis and interpretation of data: Y.C.G., M.Ü.; Writing of significant parts of the article: Y.C.G., M.Ü.;

Support and Acknowledgement Statement:

The authors used ChatGPT 4o (September 2024 Release; OpenAI; https://chat.openai.com/) to review the grammar and English translation. The content of the publication is the sole responsibility of the authors, who reviewed and edited it as they deemed necessary.

The authors would like to thank Juliette Hancox, Image Licensing Manager at Radiopaedia.org, for permission to use the images on the website.

Conflict of Interest Statement:

The authors of the article have no conflict of interest declarations.

References

1. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, Löffler CML, Schwarzkopf SC, Unger M, Veldhuizen GP, Wagner SJ, Kather JN (2023) The future landscape of large language models in medicine. Commun Med (Lond) 3:141. https://doi.org/10.1038/s43856-023-00370-1
2. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. OpenAI.https://openai.com/gpt-4 / GPT-4V(ision) System Card. OpenAI. Accessed Date Accessed
3. Liu H, Li C, Wu Q, Lee YJ (2024) Visual instruction tuning. Adv Neural Inf Process Syst 36
4. https://deepmind.google/technologies/gemini/flash/. Accessed Date Accessed
5. https://openai.com/index/hello-gpt-4o/. Accessed Date Accessed
6. Kuang Y-R, Zou M-X, Niu H-Q, Zheng B-Y, Zhang T-L, Zheng B-W (2023) ChatGPT encounters multiple opportunities and challenges in neurosurgery. Int J Surg 109:2886-2891. https://doi.org/doi: 10.1097/JS9.0000000000000571
7. Gunes YC, Camur E, Cesur T (2024) Correspondence on 'Evaluation of ChatGPT in knowledge of newly evolving neurosurgery: middle meningeal artery embolization for subdural hematoma management'by Koester et al. J Neurointerv Surg
8. Andykarayalar R, Surapaneni KM (2024) ChatGPT in Pediatrics: Unraveling Its Significance as a Clinical Decision Support Tool. Indian Pediatr 61:357-358
9. Dinis-Oliveira RJ, Azevedo RM (2023) ChatGPT in forensic sciences: a new Pandora's box with advantages and challenges to pay attention. Forensic Sci Res 8:275-279. https://doi.org/doi: 10.1093/fsr/owad039
10. Elkassem AA, Smith AD (2023) Potential use cases for ChatGPT in radiology reporting. American Journal of Roentgenology 221:373-376. https://doi.org/10.2214/AJR.23.29198
11. Ferreira AL, Chu B, Grant-Kels JM, Ogunleye T, Lipoff JB (2023) Evaluation of ChatGPT dermatology responses to common patient queries. JMIR dermatol 6:e49280. https://doi.org/doi: 10.2196/49280
12. Ilgaz HB, Çelik Z (2023) The significance of artificial intelligence platforms in anatomy education: an experience with ChatGPT and google bard. Cureus 15:e45301. https://doi.org/doi: 10.7759/cureus.45301
13. Langlie J, Kamrava B, Pasick LJ, Mei C, Hoffer ME (2024) Artificial intelligence and ChatGPT: An otolaryngology patient's ally or foe? Am J Otolaryngol 45:104220. https://doi.org/doi: 10.1016/j.amjoto.2024.104220
14. Cirone K, Akrout M, Abid L, Oakley A (2024) Assessing the utility of multimodal large language models (GPT-4 vision and large language and vision assistant) in identifying melanoma across different skin tones. JMIR dermatol 7:e55508. https://doi.org/doi: 10.2196/55508
15. Deng J, Heybati K, Shammas-Toma M (2024) When vision meets reality: Exploring the clinical applicability of GPT-4 with vision. Clin Imaging 108:110101. https://doi.org/doi: 10.1016/j.clinimag.2024.110101
16. Nakao T, Miki S, Nakamura Y, Kikuchi T, Nomura Y, Hanaoka S, Yoshikawa T, Abe O (2024) Capability of GPT-4V (ision) in the Japanese National Medical Licensing Examination: Evaluation Study. JMIR Med Educ 10:e54393. https://doi.org/doi: 10.2196/54393
17. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, De Vet HC (2015) STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Radiology 277:826-832. https://doi.org/doi: 10.1136/bmj.h5527
18. Mongan J, Moy L, Kahn CE, Jr. (2020) Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol Artif Intell 2:e200029. https://doi.org/10.1148/ryai.2020200029
19. Bolgova O, Shypilova I, Sankova L, Mavrych V (2023) How Well Did ChatGPT Perform in Answering Questions on Different Topics in Gross Anatomy? EJMED 5:94-100. https://doi.org/doi: 10.24018/ejmed.2023.5.6.1989
20. Lee H (2023) The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ. https://doi.org/doi: 10.1002/ase.2270
21. Mogali SR (2024) Initial impressions of ChatGPT for anatomy education. Anat Sci Educ 17:444-447. https://doi.org/doi: 10.1002/ase.2261
22. Totlis T, Natsis K, Filos D, Ediaroglou V, Mantzou N, Duparc F, Piagkou M (2023) The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. Surg Radiol Anat 45:1321-1329
23. Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, Naumann T, Poon H, Gao J (2024) Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Adv Neural Inf Process Syst 36. https://doi.org/doi: 10.48550/arXiv.2306.00890
24. Monajatipoor M, Rouhsedaghat M, Li LH, Jay Kuo C-C, Chien A, Chang K-W (2022) Berthop: An effective vision-and-language model for chest x-ray disease diagnosis. International Conference on Medical Image Computing and Computer-Assisted Intervention:725-734. https://doi.org/doi: 10.48550/arXiv.2108.04938
25. Zhu L, Mou W, Lai Y, Chen J, Lin S, Xu L, Lin J, Guo Z, Yang T, Lin A (2024) Step into the era of large multimodal models: A pilot study on ChatGPT-4V (ision)'s ability to interpret radiological images. Int J Surg:10.1097. https://doi.org/doi: 10.1097/JS9.0000000000001359
26. Noda M, Ueno T, Koshu R, Takaso Y, Shimada MD, Saito C, Sugimoto H, Fushiki H, Ito M, Nomura A (2024) Performance of GPT-4V in Answering the Japanese Otolaryngology Board Certification Examination Questions: Evaluation Study. JMIR Med Educ 10:e57054. https://doi.org/doi: 10.2196/57054
27. Akrout M, Cirone KD, Vender R (2024) Evaluation of Vision LLMs GTP-4V and LLaVA for the Recognition of Features Characteristic of Melanoma. J Cutan Med Surg 28:98-99. https://doi.org/doi: 10.1177/12034754231220934