Original Article

Generative pre-trained transformer (GPT)-4 support for differential diagnosis in neuroradiology

Vera Sorin[1,2,3,4], Eyal Klang[1,2,3,4], Tamer Sobeh[1,2], Eli Konen[1,2], Shai Shrot[1,2], Adva Livne[1,2,3], Yulian Weissbuch[1,2], Chen Hoffmann[1,2#], Yiftach Barash[1,2,3#]

[1]Department of Diagnostic Imaging, Chaim Sheba Medical Center, Ramat Gan, Israel; [2]The Faculty of Medicine, Tel-Aviv University, Tel Aviv-Yafo, Israel; [3]DeepVision Lab, Chaim Sheba Medical Center, Ramat Gan, Israel; [4]Sami Sagol AI Hub, ARC, Sheba Medical Center, Ramat Gan, Israel

Contributions: (I) Conception and design: V Sorin, C Hoffmann, Y Barash; (II) Administrative support: E Konen, A Livne; (III) Provision of study materials or patients: E Klang, T Sobeh, S Shrot, Y Weissbuch, Y Barash; (IV) Collection and assembly of data: Y Barash; (V) Data analysis and interpretation: V Sorin, E Klang, Y Barash; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Vera Sorin, MD. Department of Diagnostic Imaging, Chaim Sheba Medical Center, Emek Haela St. 1, Ramat Gan 52621, Israel; The Faculty of Medicine, Tel-Aviv University, Tel Aviv-Yafo, Israel; DeepVision Lab, Chaim Sheba Medical Center, Ramat Gan, Israel; Sami Sagol AI Hub, ARC, Sheba Medical Center, Ramat Gan, Israel. Email: verasrn@gmail.com.

Background: Differential diagnosis in radiology relies on the accurate identification of imaging patterns. The use of large language models (LLMs) in radiology holds promise, with many potential applications that may enhance the efficiency of radiologists' workflow. The study aimed to evaluate the efficacy of generative pre-trained transformer (GPT)-4, a LLM, in providing differential diagnoses in neuroradiology, comparing its performance with board-certified neuroradiologists.

Methods: Sixty neuroradiology reports with variable diagnoses were inserted into GPT-4, which was tasked with generating a top-3 differential diagnosis for each case. The results were compared to the true diagnoses and to the differential diagnoses provided by three blinded neuroradiologists. Diagnostic accuracy and agreement between readers were assessed.

Results: Of the 60 patients (mean age 47.8 years, 65% female), GPT-4 correctly included the diagnoses in its differentials in 61.7% (37/60) of cases, while the neuroradiologists' accuracy ranged from 63.3% (38/60) to 73.3% (44/60). Agreement between GPT-4 and the neuroradiologists, and among the neuroradiologists was fair to moderate [Cohen's kappa (kw) 0.34–0.44 and kw 0.39–0.54, respectively].

Conclusions: GPT-4 shows potential as a support tool for differential diagnosis in neuroradiology, though it was outperformed by human experts. Radiologists should remain mindful to the limitations of LLMs, while harboring their potential to enhance educational and clinical work.

Keywords: Large language models (LLMs); generative pre-trained transformer (GPT); differential diagnosis; neuroradiology

Submitted Jan 30, 2024. Accepted for publication Jul 25, 2024. Published online Sep 23, 2024.
doi: 10.21037/qims-24-200
View this article at: https://dx.doi.org/10.21037/qims-24-200

Introduction

Natural language processing (NLP) has become an essential tool in radiology, with applications spanning voice recognition, information extraction, labeling, protocol generation, and the creation of research cohorts (1). Recently, transformer models have emerged as the state-of-the-art NLP, demonstrating abilities to process large texts and comprehend context (2,3). Large language models (LLMs), such as generative pre-trained transformer (GPT), are built upon this architecture, exhibiting impressive capabilities in text processing and generation, applicable across various medical domains (4-10).

The use of LLMs in radiology holds promise, with many potential applications that may enhance the efficiency of radiologists' workflow. These include assistance with radiology referrals review, generation of structured and comprehensive radiology reports, simplifying radiology reports for patient communication, question answering for both patients and clinicians, and supporting differential diagnosis. The models can also enhance discussions in complex case reviews, such as tumor boards, by providing rapid, data-driven insights. Research related tasks include assistance with extraction of information from articles, identification of trends and patterns in large datasets or literature and indexing large amounts of text data.

Differential diagnosis in radiology relies on the accurate identification of imaging patterns. This process requires extensive medical knowledge along with radiological expertise. Such expertise is indispensable for proficient radiologists, enabling them to discern potential diagnoses and thereby guide subsequent treatment strategies. LLMs are capable of rapidly processing large amounts of data. Although lacking radiological practical expertise, some models may have broader literature knowledge than physicians. Consequently, LLMs may occasionally offer contextual information that radiologists might not consider at the time (7-9). The aim of our study was to assess GPT-4 as an assistive tool in providing differential diagnoses in neuroradiology. We present this article in accordance with the GRRAS reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-24-200/rc).

Methods

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by institutional review board of Sheba Medical Center (No. 0143-23-SMC), and individual consent was waived due to retrospective nature of this study.

We selected radiology reports for 60 representative neuroradiology pathologies, prioritizing conditions that typically require differential diagnosis based on imaging appearance. Specifically, a board-certified radiologist (T.S.) identified 20 diagnoses featured in recent radiology oral board examinations, that were also representative of a spectrum of pathophysiologies relevant to clinical practice. For each of the selected diagnoses, we systematically extracted three consecutive reports from our radiology information system (RIS), ensuring a balanced representation. Cases were randomly selected and only those with confirmative pathology results or supported by substantial clinical evidence were included.

Each radiology report was initially reviewed by T.S. to remove identifiable demographic information, ensuring patient confidentiality. Reports in our center are inherently non-structured and exhibit variability, as each radiologist employs a unique reporting style, resulting in heterogenous data. The relevant descriptions from each report were translated into English. T.S reviewed the reports and extracted only pertinent sections that directly related to imaging findings necessary for generating differential diagnoses. These were then input into GPT-4 by Y.B., who requested the model to provide the three most likely differential diagnoses, based on the imaging descriptions provided. This was aimed to ensure that GPT-4's analysis was focused. GPT-4 was accessed on July 10th, 2023, for the first 20 cases, and again on April 24th, 2024 for the remaining 40 cases, and all responses were obtained on these dates. The exact prompt was: "Can you please read this imaginary clinical brain MRI/CT findings and give the 3 most probable differential diagnoses in a descending probability order".

Three board-certified neuroradiologists (S.S., Y.W. and C.H.) with at least 10 years of experience each, who were blinded to the diagnoses, were presented with identical texts (only the pertinent sections that were extracted from the reports) and provided the three most likely diagnoses for each case. Neither the neuroradiologists nor GPT were privy to any clinical or demographic details on the patients. It should be noted that the neuroradiologists in this study provided their differential diagnoses based solely on the pattern descriptions, without access to the actual images.

Diagnostic accuracy was assessed by determining whether the actual diagnosis was included in the differential diagnosis provided by GPT and the three radiologists. These accuracies were subsequently compared. The agreement between each radiologist and GPT, as well as among the radiologists themselves, was quantified using the linear weighted Cohen's kappa (kw) coefficient. The interpretation of the kw coefficient was categorized as follows: <0: poor agreement; 0.01–0.20: slight agreement; 0.21–0.40: fair agreement; 0.41–0.60: moderate agreement; 0.61–0.80: substantial agreement; 0.81–0.99: almost perfect agreement.

Results

Patient demographics

The study included 60 patients, the mean age was 47.8 years, and 65% (13/20) were female. An overview of each patient's diagnosis and clinical presentation is provided in *Table 1*.

Diagnostic accuracy

GPT-4's inclusion of the correct diagnosis within the

Table 1 Patient clinical characteristics and diagnosis

Case number	Gender	Age (years)	Indication	Modality	Diagnosis
1	Male	28	Right sided hypertonia, following a viral illness	MRI	ADEM
2	Female	75	Confusion	MRI	Amyloid neuropathy
3	Female	52	Follow-up	MRI	Cavernoma
4	Female	64	Decreased consciousness and increased spasticity	MRI	CJD
5	Male	77	Acute confusion	CT	Colloid cyst
6	Female	22	Severe headache	MRI	Ependymoma
7	Female	54	Suspected cholesteatoma	MRI	Epidermoid
8	Female	74	Confusion	MRI	GBM
9	Male	25	Follow-up	MRI	Herpes encephalitis
10	Female	–	Seizures	MRI	Hypoglycemia
11	Female	48	Seizures	CT	Meningioma
12	Female	19	Suspected demyelinating lesion in brain MRI	MRI	NMO
13	Female	44	Irritability	MRI	Osmotic demyelination syndrome
14	Female	50	Recurrent falls	MRI	Pineocytoma
15	Female	73	Acute confusion	MRI	PRES
16	Male	84	Aphasia	CT	Stroke
17	Male	64	Right hemiplegia	MRI	Subependymoma
18	Male	32	Follow-up	MRI	Tuberous sclerosis
19	Female	34	Head trauma	MRI	Vestibular schwannoma
20	Male	89	Follow-up	MRI	Melanoma metastases
21	Female	76	Follow-up	MRI	Vestibular schwannoma
22	Female	76	Vertigo	MRI	Vestibular schwannoma
23	Female	77	Vertigo	MRI	Cavernoma
24	Male	77	Confusion	MRI	Cavernoma
25	Female	43	Papilledema	MRI	Colloid cyst
26	Female	68	Follow-up	MRI	GBM

Table 1 (*continued*)

Table 1 (*continued*)

Case number	Gender	Age (years)	Indication	Modality	Diagnosis
27	Female	72	Follow-up	MRI	Colloid
28	Female	77	Seizures	MRI	Meningioma
29	Female	50	Gait disturbances	MRI	CJD
30	Female	32	Follow-up	MRI	Meningioma (atypical)
31	Female	75	Follow-up	MRI	GBM
32	Female	63	Cognitive impairment	MRI	CJD
33	Male	57	Confusion	MRI	Infarct
34	Male	71	Dysarthria	MRI	Infarct
35	Female	71	Follow-up	MRI	Melanoma metastases
36	Female	67	Follow-up	MRI	Melanoma metastases
37	Female	18	Seizures	MRI	Hypoglycemia
38	Female	88	Increased muscle tone	MRI	Hypoglycemia
39	Female	57	Follow-up	MRI	Subependymoma
40	Female	26	Right sided weakness	MRI	Subependymoma
41	Male	58	Decreased vision	MRI	ADEM
42	Male	24	Confusion	MRI	ADEM
43	Male	63	Dysarthria	MRI	Amyloid neuropathy
44	Female	39	Follow-up	MRI	Amyloid neuropathy
45	Female	51	Follow-up	MRI	Ependymoma (myxopapillary)
46	Male	68	Headache	MRI	Ependymoma
47	Female	44	Follow-up	MRI	Epidermoid
48	Female	44	Follow-up	MRI	Epidermoid
49	Female	26	Seizures	MRI	Herpes encephalitis
50	Male	–	Seizures	MRI	Herpes encephalitis
51	Male	73	Gait disturbances	MRI	NMO
52	Female	69	Follow-up	MRI	NMO
53	Female	54	Cerebellar symptoms	MRI	Osmotic demyelination syndrome
54	Female	67	Follow-up	MRI	Osmotic demyelination syndrome
55	Female	50	Follow-up	MRI	Pineocytoma
56	Female	33	Follow-up	MRI	Pineocytoma
57	Male	21	Seizures	MRI	PRES
58	Male	51	Seizures	MRI	PRES
59	Female	20	Seizures	MRI	Tuberous sclerosis
60	Female	31	Follow-up	MRI	Tuberous sclerosis

MRI, magnetic resonance imaging; ADEM, acute disseminated encephalomyelitis; CJD, Creutzfeldt-Jakob disease; CT, computed tomography; GBM, glioblastoma multiforme; NMO, neuromyelitis optica; PRES, posterior reversible encephalopathy syndrome.

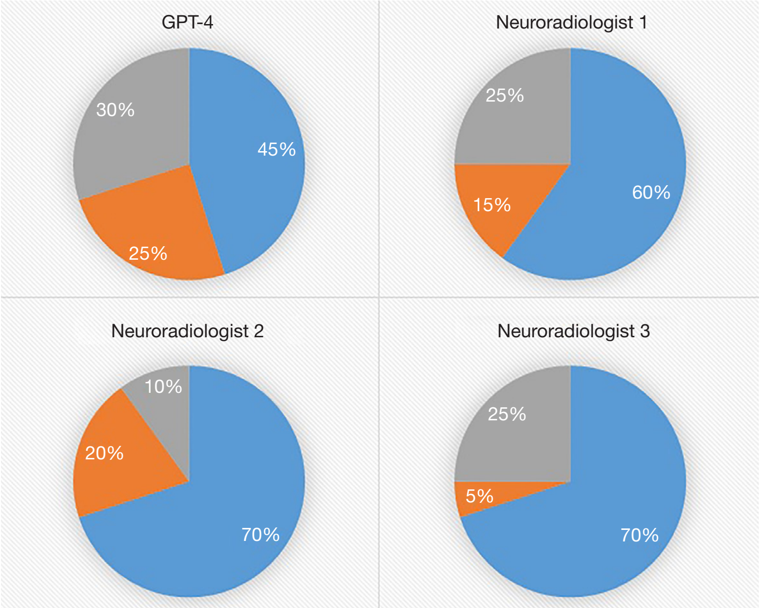

Figure 1 GPT-4 and neuroradiologists differential diagnoses success rates. Pie charts describing the performance of GPT-4 and each of the neuroradiologists separately in listing the correct diagnosis in the differential, based on the actual dagnoses of patients. In blue, the final diagnosis is in the first place of the differential diganosis; in orange, the correct diagnosis is listed in the second or third places of the differential diagnosis; and in grey, the correct diagnosis is not listed at all. GPT, generative pre-trained transformer.

differential diagnoses was observed in 61.7% (37/60) of cases. In contrast, the neuroradiologists achieved accuracy rates of 63.33% (38/60), 70% (42/60), and 73.33% (44/60), respectively. A comparison of diagnostic accuracy rates between GPT and the neuroradiologists relative to the actual diagnoses is depicted in *Figure 1*. Specific examples of cases are illustrated in *Figures 2,3*.

Agreement analysis

Agreement between GPT and the neuroradiologists on the first differential diagnosis was fair (kw 0.39, 0.34, 0.36, respectively). Among the radiologists, the agreement was fair to moderate (kw 0.42, 0.54, 0.39, respectively). A detailed breakdown of these agreement rates is presented in *Table 2*.

Qualitative analysis

In two cases, GPT-4 included the correct diagnosis as the primary differential, whereas the neuroradiologists did not list the correct diagnosis as the first differential. The diagnoses of these cases were herpes encephalitis and pineocytoma. In none of the cases did the LLM include the correct diagnosis in the top three differential when it was omitted by all radiologists. In three instances, involving cases of Creutzfeldt-Jakob disease, epidermoid cyst, and tuberous sclerosis, all three radiologists included the diagnoses in their differential, while the LLM omitted them.

In one case of herpes encephalitis, only one radiologist listed the diagnosis among the top three differentials, placing it third, while the other two did not include it at all. The radiology report described the findings as follows: “In the temporal lobes, areas of damage are demonstrated, more prominent in the right temporal lobe, but also on the lateral side of the left temporal lobe”. The radiologists’ differential diagnosis included old infarcts, chronic traumatic injury, malacia or postoperative changes. The LLM’s differential included herpes simplex encephalitis, Alzheimer’s disease and temporal lobe epilepsy. The LLM’s response is detailed in the Appendix 1 (Case #1).

There were 13 cases where all readers and the LLM included the correct diagnosis in the differential list. One example is a case of cavernoma, described in the report as: “A mottled parenchymal structure in the medial temporal lobe on the left side. A hypointense hemosiderin ring is present around the perimeter. There is a susceptibility effect

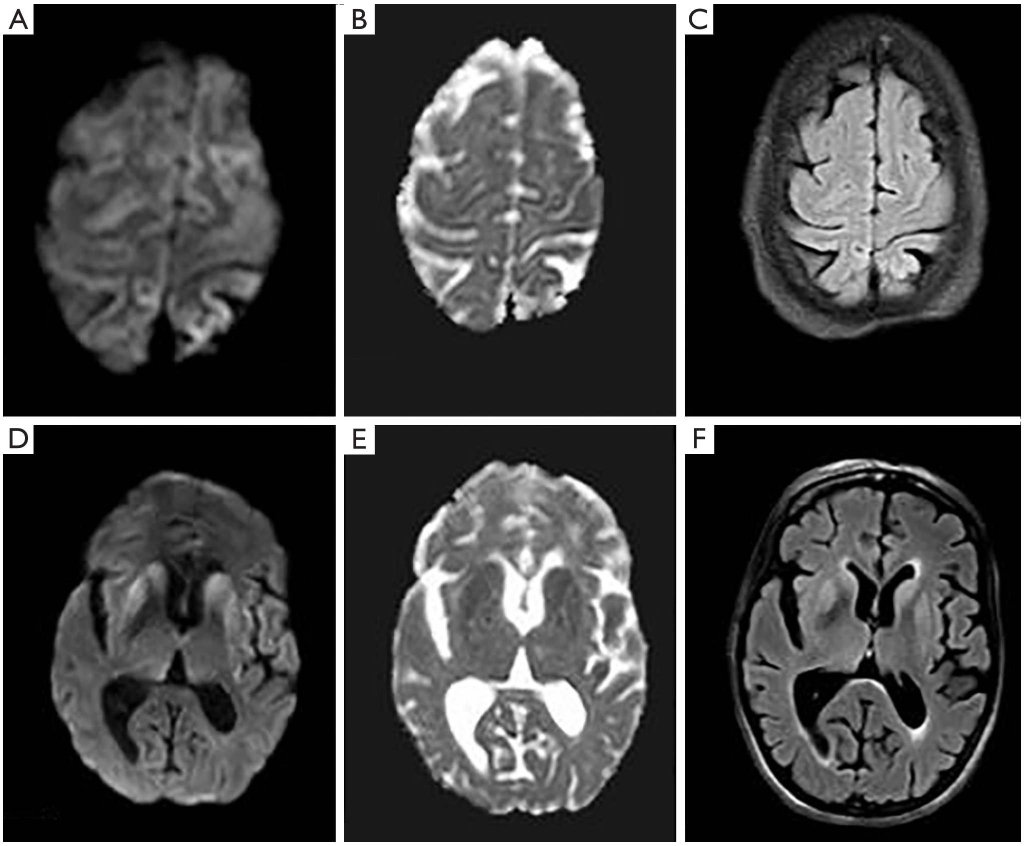

Figure 2 An example of MRI of a patient with CJD (patient #4). The MRI report described: diffusion restriction in the caudate and putamen nuclei bilaterally, the pulvinar nucleus of the thalamus on both sides, and the paramedian cortex in fronto-parietal areas bilaterally, left insula, bilateral occipital lobes, bilateral cingulate gyri, and the right hippocampus. There is also hyperintensity signal in the caudate and putamen nuclei bilaterally on FLAIR sequence, and increased signals in the parietal and occipital cortex. The images show a high DWI signal in the left fronto-parietal cortex (A), as well as the caudate, putamen and thalami bilaterally (D) accompanied by a corresponding low ADC signal (B,E), and an elevated signal on the FLAIR series (C,F) within the same regions. GPT-4 omitted CJD from its differential diagnosis, whereas all three neuroradiologists included CJD in their differential diagnosis. MRI, magnetic resonance imaging; CJD, Creutzfeldt-Jakob disease; FLAIR, fluid-attenuated inversion recovery; DWI, diffusion-weighted imaging; ADC, apparent diffusion coefficient.

(blooming) on SWI sequences, and also slightly amorphous internal and peripheral enhancement". The radiologists' differential lists included cavernoma, hemorrhagic stroke, a resolving hematoma, hemorrhage within a tumor, and metastases. The LLM's differential included glioma, cerebral amyloid angiopathy and cavernoma. The response is detailed in Appendix 1 (Case #2).

Discussion

This study evaluated GPT's performance in generating differential diagnoses in neuroradiology based on descriptions from radiology reports. The results show that GPT included the correct diagnosis in 62% of cases, while experienced neuroradiologists had higher accuracy rates. The agreement between GPT and radiologists on the first suggested diagnosis was only slight, and the agreement among radiologists ranged from fair to moderate.

The discrepancies observed between the radiologists and the LLM may be influenced by several factors. One key factor is that the original reports were written by different radiologists with varying reporting styles. Additionally, in cases involving complex pathologies, the LLM may theoretically have an advantage due to its broader general medical knowledge of rare cases compared to neuroradiologists. Conversely, neuroradiologists possess

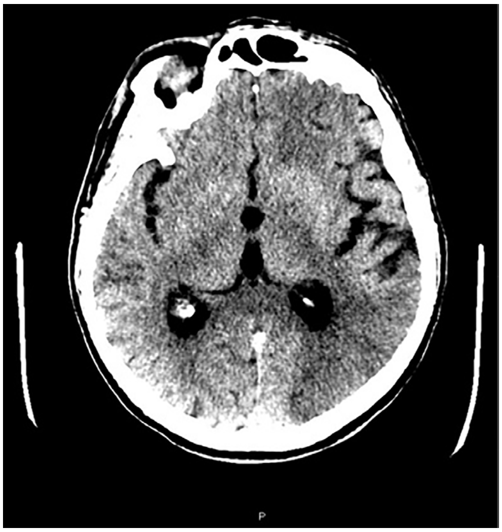

Figure 3 An example of non-contrast head CT of a patient with an acute stroke (patient #16). The CT report described extensive cytotoxic edema in the fronto-parietal, posterior temporal, and occipital areas on the left. There are isolated isodense foci within the edema. The image shows hypodensity in the left parieto-occipital area, accompanied by blurring of the white-gray matter border and effacement of the sulci. ChatGPT and all three neuroradiologists identified acute stroke as the primary differential diagnosis. CT, computed tomography; GPT, generative pre-trained transformer.

Table 2 Agreement rates based on kw coefficient

Comparison	First diagnosis	First to third diagnoses
GPT–Neuroradiologist 1	0.387	0.407
GPT–Neuroradiologist 2	0.344	0.375
GPT–Neuroradiologist 3	0.358	0.439
Neuroradiologist 1–Neuroradiologist 2	0.422	0.395
Neuroradiologist 1–Neuroradiologist 3	0.545	0.540
Neuroradiologist 2–Neuroradiologist 3	0.386	0.344

kw, Cohen's kappa; GPT, generative pre-trained transformer.

clinical experience and intuition that the LLM lacks. Our qualitative analysis did not reveal any specific patterns in the discrepancies observed. Therefore, a larger study is warranted to determine if any patterns can be identified.

LLMs, such as GPT, offer a new avenue to support various language tasks. Since ChatGPT was first introduced by Open-artificial intelligence (AI) at the end of 2022, other LLMs such as LLaMA, GPT-4, Gemini and Claude have emerged. Numerous studies have explored their applications in healthcare, including clinical decision support, referrals generation, patient question-answering, and research (2,8,11). Our study evaluates the potential of GPT-4 in generation of differential diagnoses in neuroradiology.

LLMs can significantly enhance radiology practice through various potential use-cases (12). These include automated report generation, reducing time radiologists spend on manual reporting while maintaining reporting accuracy and relevancy (13-15). The models can assist in creating structured templates from free text (16). They can also be used to support clinical decision-making, by summarizing research, generating differential diagnoses from textual descriptions (17), and enhancing discussions in complex case reviews such as tumor boards (18). In addition, the models can also be used for generating and reviewing radiology referrals and for protocols' generation (9). LLMs can be used to improve patient communication and education, by simplifying medical information into understandable language (19). LLMs also serve as conversational agents, providing a user-friendly platform for patient and clinician inquiries, thus improving patient engagement and operational efficiency in everyday clinical work (20). Moreover, in research, LLMs may aid in extracting information from articles, identifying trends in large datasets or literature, and indexing vast amounts of text data (21). All these use cases may contribute to the enhanced efficiency of radiologists' workflows.

Recent advancements have enabled some LLMs, including GPT-4, to analyze data from both text and images simultaneously. These multimodal models are particularly transformative in healthcare, where medical diagnosis relies on interpreting both textual and visual information. In radiology, this development suggests that the textual descriptions of findings are redundant when using the model, and that the model can analyze images alongside clinical information for a comprehensive patient analysis. However, preliminary research indicates that the current performance of GPT-4 in analyzing radiology images does not yet meet the standards required for clinical applications (22).

Looking ahead, LLMs could enhance patient care and streamline radiologists' workflow. The multimodal abilities are expected to improve and support studies interpretations while simultaneously generating reports drafts. Furthermore, integrating these models could improve communication between patients and radiologists, thereby enhancing patent education, understanding, and active involvement in care decisions. However, significant challenges remain. These include inherent biases in training data, navigating ethical concerns around AI use in patient care (23), and data security (24). Addressing these limitations while being mindful of the vast potential applications in clinical practice is crucial to developing reliable and applicable tools that can be integrated into routine clinical practice.

The literature generally discusses LLMs as tools to enhance and support healthcare providers, not replace them. Our findings align with this view, as neuroradiologists outperformed GPT in diagnostic accuracy. These models could be valuable for educating radiology residents and supporting less experienced radiologists, especially under time pressure and high workload, which are timely concerns in neuroradiology (25).

Despite the promising potential of LLMs, these models have limitations and risks. They can generate incorrect information, requiring careful validation (26). This is critical in healthcare, where false information can have serious consequences. LLMs may also reinforce existing biases and disparities (27,28). Data security, privacy, and legal responsibility must be considered when using patient data (29). Finally, legal issues related to decisions based on LLM outputs also need careful consideration (26).

This study has several limitations. First, we did not include clinical correlation in the differential diagnosis generation for both GPT and radiologists, focusing only on imaging pattern description. This is because in some cases, incorporating the clinical presentation was highly suggestive of the diagnosis (for instance, stroke). The small sample size and the retrospective nature of the study may also introduce selection bias. Only the pertinent sections of the reports, selected by T.S., were presented to the LLM and the radiologists. T.S. may have been influenced by prior knowledge of the diagnoses, which introduced bias affecting both the LLM and the radiologists. Additionally, we did not evaluate how different prompts, radiologists or report styles affected the performance of the LLM and radiologists. These are potential directions for future studies with larger datasets. We only evaluated a single LLM (GPT-4) and did not compare the performance of different LLMs on this task. In this study, neuroradiologists provided their differential diagnoses based solely on pattern descriptions, without viewing the actual images. Had the neuroradiologists been given access to the studies, it's plausible their diagnostic accuracy would have improved. Additionally, the level of agreement among them would likely have been higher. Finally, since conducting the study, multimodal capabilities have emerged in LLMs, enabling models like GPT-4 to analyze images in addition to text (24). However, as our research specifically focused on the textual analysis capabilities of GPT-4 and differential diagnosis derived from text-based pattern recognition, we did not incorporate image analysis. Preliminary data suggest that the model's current performance in radiology image analysis is not yet adequate for clinical applications (22).

Conclusions

In conclusion, GPT showed the ability to provide differential diagnoses in neuroradiology based on pattern descriptions from radiology reports. The performance of neuroradiologists surpassed the LLM. While LLMs have potential to enhance daily work and education in radiology, their limitations must be recognized, and they must be used responsibly.

Acknowledgments

The authors acknowledge the use of ChatGPT, an AI language model developed by OpenAI, solely for generating differential diagnoses in this study. The AI-generated diagnoses were compared with those provided by a qualified radiologist to assess their accuracy and reliability. No personal or identifiable patient data was used in this process; all data were anonymized to protect privacy. The authors did not use ChatGPT or any other AI tool for the writing of this manuscript. The authors declare no conflicts of interest and received no funding from OpenAI or related organizations.
Funding: None.

Footnote

Reporting Checklist: The authors have completed the GRRAS reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-24-200/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-200/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by institutional review board of Sheba Medical Center (No. 0143-23-SMC), and individual consent was waived due to retrospective nature of this study.

References

1. Sorin V, Barash Y, Konen E, Klang E. Deep Learning for Natural Language Processing in Radiology-Fundamentals and a Systematic Review. J Am Coll Radiol 2020;17:639-48.
2. Sorin V, Barash Y, Konen E, Klang E. Deep-learning natural language processing for oncological applications. Lancet Oncol 2020;21:1553-6.
3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Part of Advances in Neural Information Processing Systems 30 (NIPS 2017).
4. Sorin V, Barash Y, Konen E, Klang E. Large language models for oncological applications. J Cancer Res Clin Oncol 2023;149:9505-8.
5. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A. Language models are few-shot learners. Advances in Neural Information Processing Systems 33 (NeurIPS 2020) 2020;33:1877-901.
6. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620:172-80.
7. Sorin V, Klang E, Sklair-Levy M, Cohen I, Zippel DB, Balint Lahat N, Konen E, Barash Y. Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer 2023;9:44.
8. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel) 2023;11:887.
9. Barash Y, Klang E, Konen E, Sorin V. ChatGPT-4 Assistance in Optimizing Emergency Department Radiology Referrals and Imaging Selection. J Am Coll Radiol 2023;20:998-1003.
10. Gebrael G, Sahu KK, Chigarira B, Tripathi N, Mathew Thomas V, Sayegh N, Maughan BL, Agarwal N, Swami U, Li H. Enhancing Triage Efficiency and Accuracy in Emergency Rooms for Patients with Metastatic Prostate Cancer: A Retrospective Analysis of Artificial Intelligence-Assisted Triage Using ChatGPT 4.0. Cancers (Basel) 2023;15:3717.
11. Hu Y, Hu Z, Liu W, Gao A, Wen S, Liu S, Lin Z. Exploring the potential of ChatGPT as an adjunct for generating diagnosis based on chief complaint and cone beam CT radiologic findings. BMC Med Inform Decis Mak 2024;24:55.
12. Grewal H, Dhillon G, Monga V, Sharma P, Buddhavarapu VS, Sidhu G, Kashyap R. Radiology Gets Chatty: The ChatGPT Saga Unfolds. Cureus 2023;15:e40135.
13. Bosbach WA, Senge JF, Nemeth B, Omar SH, Mitrakovic M, Beisbart C, Horváth A, Heverhagen J, Daneshvar K. Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and integrated AO classifier. Curr Probl Diagn Radiol 2024;53:102-10.
14. Gertz RJ, Dratsch T, Bunck AC, Lennartz S, Iuga AI, Hellmich MG, Persigehl T, Pennig L, Gietzen CH, Fervers P, Maintz D, Hahnfeldt R, Kottlors J. Potential of GPT-4 for Detecting Errors in Radiology Reports: Implications for Reporting Accuracy. Radiology 2024;311:e232714.
15. Sun Z, Ong H, Kennedy P, Tang L, Chen S, Elias J, Lucas E, Shih G, Peng Y. Evaluating GPT4 on Impressions Generation in Radiology Reports. Radiology 2023;307:e231259.
16. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, Bressem KK. Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. Radiology 2023;307:e230725.
17. Kottlors J, Bratke G, Rauen P, Kabbasch C, Persigehl T, Schlamann M, Lennartz S. Feasibility of Differential Diagnosis Based on Imaging Patterns Using a Large Language Model. Radiology 2023;308:e231167.
18. Sorin V, Glicksberg BS, Artsi Y, Barash Y, Konen E, Nadkarni GN, Klang E. Utilizing large language models

in breast cancer management: systematic review. J Cancer Res Clin Oncol 2024;150:140.

19. Schmidt S, Zimmerer A, Cucos T, Feucht M, Navas L. Simplifying radiologic reports with natural language processing: a novel approach using ChatGPT in enhancing patient understanding of MRI results. Arch Orthop Trauma Surg 2024;144:611-8.
20. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. Radiology 2023;307:e230582.
21. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, Heußel CP, Kauczor HU, Weber TF. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. Radiology 2023;308:e231362.
22. Brin D, Sorin V, Barash Y, Konen E, Nadkarni G, Glicksberg BS, Klang E. Assessing GPT-4 Multimodal Performance in Radiological Image Analysis. medRxiv 2023. doi: 10.1101/2023.11.15.23298583.
23. Sorin V, Brin D, Barash Y, Konen E, Charney A, Nadkarni G, Klang E. Large language models (LLMs) and empathy-a systematic review. medRxiv 2023. doi: 10.1101/2023.08.07.23293769.
24. Sorin V, Kapelushnik N, Hecht I, Zloto O, Glicksberg BS, Bufman H, Barash Y, Nadkarni GN, Klang E. GPT-4 Multimodal Analysis on Ophthalmology Clinical Cases Including Text and Images. medRxiv 2023. doi: 10.1101/2023.11.24.23298953.
25. Chen JY, Lexa FJ. Baseline Survey of the Neuroradiology Work Environment in the United States with Reported Trends in Clinical Work, Nonclinical Work, Perceptions of Trainees, and Burnout Metrics. AJNR Am J Neuroradiol 2017;38:1284-91.
26. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med 2023;388:1233-9.
27. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. Lancet Digit Health 2023;5:e333-5.
28. Sorin V, Klang E. Artificial Intelligence and Health Care Disparities in Radiology. Radiology 2021;301:E443.
29. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019;363:1287-9.

Cite this article as: Sorin V, Klang E, Sobeh T, Konen E, Shrot S, Livne A, Weissbuch Y, Hoffmann C, Barash Y. Generative pre-trained transformer (GPT)-4 support for differential diagnosis in neuroradiology. Quant Imaging Med Surg 2024;14(10):7551-7560. doi: 10.21037/qims-24-200