AJNR

Toward Improved Radiologic Diagnostics: Investigating the Utility and Limitations of GPT-3.5 Turbo and GPT-4 with Quiz Cases

Tomohiro Kikuchi, Takahiro Nakao, Yuta Nakamura, Shouhei Hanaoka, Harushi Mori and Takeharu Yoshikawa

Toward Improved Radiologic Diagnostics: Investigating the Utility and Limitations of GPT-3.5 Turbo and GPT-4 with Quiz Cases

Tomohiro Kikuchi, Takahiro Nakao, Yuta Nakamura, Shouhei Hanaoka, Harushi Mori, and Takeharu Yoshikawa

ABSTRACT

BACKGROUND AND PURPOSE: The rise of large language models such as generative pretrained transformers (GPTs) has sparked considerable interest in radiology, especially in interpreting radiologic reports and image findings. While existing research has focused on GPTs estimating diagnoses from radiologic descriptions, exploring alternative diagnostic information sources is also crucial. This study introduces the use of GPTs (GPT-3.5 Turbo and GPT-4) for information retrieval and summarization, searching relevant case reports via PubMed, and investigates their potential to aid diagnosis.

MATERIALS AND METHODS: From October 2021 to December 2023, we selected 115 cases from the "Case of the Week" series on the *American Journal of Neuroradiology* website. Their Description and Legend sections were presented to the GPTs for the 2 tasks. For the Direct Diagnosis task, the models provided 3 differential diagnoses that were considered correct if they matched the diagnosis in the diagnosis section. For the Case Report Search task, the models generated 2 keywords per case, creating PubMed search queries to extract up to 3 relevant reports. A response was considered correct if reports containing the disease name stated in the diagnosis section were extracted. The McNemar test was used to evaluate whether adding a Case Report Search to Direct Diagnosis improved overall accuracy.

RESULTS: In the Direct Diagnosis task, GPT-3.5 Turbo achieved a correct response rate of 26% (30/115 cases), whereas GPT-4 achieved 41% (47/115). For the Case Report Search task, GPT-3.5 Turbo scored 10% (11/115), and GPT-4 scored 7% (8/115). Correct responses totaled 32% (37/115) with 3 overlapping cases for GPT-3.5 Turbo, whereas GPT-4 had 43% (50/115) of correct responses with 5 overlapping cases. Adding Case Report Search improved GPT-3.5 Turbo's performance ($P = .023$) but not that of GPT-4 ($P = .248$).

CONCLUSIONS: The effectiveness of adding Case Report Search to GPT-3.5 Turbo was particularly pronounced, suggesting its potential as an alternative diagnostic approach to GPTs, particularly in scenarios where direct diagnoses from GPTs are not obtainable. Nevertheless, the overall performance of GPT models in both direct diagnosis and case report retrieval tasks remains not optimal, and users should be aware of their limitations.

ABBREVIATIONS: AJNR = *American Journal of Neuroradiology*; API = Application Programming Interface; AI = artificial intelligence; COW = Case of the Week; GPT = generative pretrained transformer; JSON = JavaScript Object Notation; LLM = large language model

Advancements in artificial intelligence (AI), particularly the emergence of large language models (LLMs), have ushered in a new era in the medical field.[1,2] Among these, ChatGPT and its underlying model, the generative pretrained transformer (GPT), have been widely used and recognized for demonstrating remarkable performance in zero-shot learning.[3,4] Several studies have explored the potential applications of this innovative technology in the medical field.

A hot topic among these is deducing diagnoses from patient histories and image findings.[5-7] By using LLMs to interpret imaging findings that lead to the formulation of differential diagnoses, these models can assist radiologists effectively, potentially elevating the quality and efficiency of medical diagnostics. However, existing research on LLMs has primarily focused on their direct diagnostic capabilities, with less attention paid to exploring alternative approaches for cases in which direct diagnosis does not yield correct answers. Addressing this gap by exploring additional functionalities may expand the scope of LLM application.

SUMMARY

PREVIOUS LITERATURE: Recent advances in AI, particularly with large language models such as ChatGPT, have revolutionized medical research. These models have been applied across various medical tasks, notably in diagnosing based on patient histories and imaging. They have proved particularly useful in aiding radiologists by interpreting diagnostic images to help formulate differential diagnoses, enhancing both the quality and efficiency of medical diagnostics. However, most research has focused on their direct diagnostic abilities, neglecting the potential of these AI models to improve diagnosis through alternative methods, such as searching for relevant case reports.

KEY FINDINGS: GPT-4 showed superior direct diagnosis capability over GPT-3.5 Turbo, with success rates of 41% and 26%, respectively. Case Report Search tasks had lower success for both models, but notably, adding this search improved GPT-3.5 Turbo's performance, not that of GPT-4.

KNOWLEDGE ADVANCEMENT: The study demonstrates the utility of combining direct diagnostics with case report searches in improving GPT-3.5 Turbo's accuracy. This suggests a nuanced application of AI in diagnostics, emphasizing the strengths and limitations of current GPT technology.

In this context, we propose using LLMs to search for case reports and other online teaching cases in the interpretation of radiologic descriptions. Case reports and other online teaching cases serve as crucial resources in clinical medicine, disseminating unique patient experiences and findings often absent in other publications.[8,9] If LLMs could efficiently generate search queries from radiologic descriptions, their results could offer an alternative pathway for diagnosis, different from directly asking LLMs for diagnosis.

Through this investigation, we aimed to test whether GPTs (GPT-3.5 Turbo and GPT-4) can accurately identify diseases from image findings in quiz cases at a level that would educate radiologists and whether they can extract appropriate keywords for effective search queries. Additionally, we assessed whether integrating the results from both tasks increased the overall accuracy of the correct diagnosis. Our study utilizes the "Case of the Week (COW)" (https://www.ajnr.org/cow/by/year) feature from the *American Journal of Neuroradiology* (AJNR) to assess the capabilities of GPT-3.5 Turbo and GPT-4 in interpreting textual descriptions of diagnostic images.

MATERIALS AND METHODS

Overview

Ethical approval was not required at our institution because this study was based on publicly available literature. This article follows the Standards for Reporting Diagnostic Accuracy reporting guidelines.[10] This study explored the response capabilities of the OpenAI GPT-3.5 Turbo and GPT-4 models by using the OpenAI Application Programming Interface (API) in response to diagnostic quizzes from the COW section of AJNR. Two distinct scenarios were investigated: 1 required a direct diagnosis, and the other required the creation of suitable keywords for case report searching. Experimental results were obtained on the same day, February 20, 2024. Fig 1 illustrates the experimental overview of a single session of the COW. After the textual information was extracted from the Description and Legend sections, it was
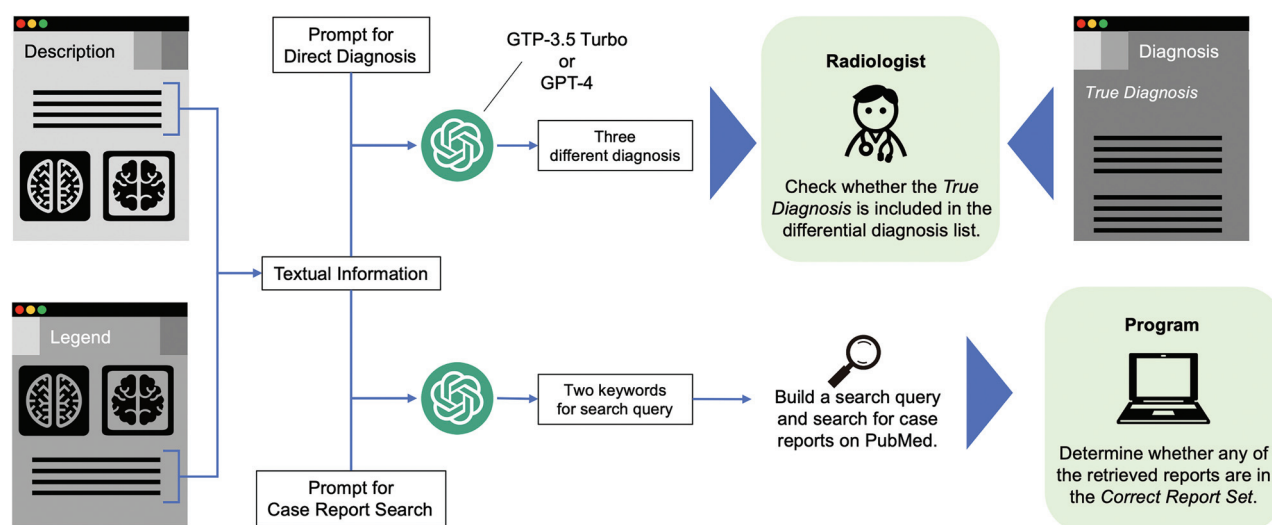
FIG 1. Overview of the experiment for a single session of COW. Textual information is extracted from the Description and Legend sections. The subsequent upper section constitutes the Direct Diagnosis task, wherein it is determined whether the differential diagnosis output by the GPTs includes the True Diagnosis. The lower section involves Case Report Search, in which case reports are searched in PubMed by using keywords generated by GPTs. It is then assessed whether these reports correspond to a predefined Correct Report Set.

Table 1: Prompts given to the GPTs

Task	Prompt
Direct diagnosis	You are an experienced neuroradiologist participating in a diagnostic imaging quiz. The patient information is provided under the "Description" section for the patient's background and the "Legends" for the imaging findings. Based on the information given, please list 3 differential diagnoses in order of most to least suspected.
Case report search	You are an experienced neuroradiologist participating in a diagnostic imaging quiz. The patient information is provided under the "Description" section for the patient's background and the "Legends" for the imaging findings. At present, you are considering several diseases for differential diagnosis. You may search PubMed for case reports to assist in your diagnostic process. What 2 keywords would you suggest for the search (each within 3 words)? Your search will be structured as (keyword1) AND (keyword2) AND ("Computed Tomography"[title/abstract] OR "CT"[title/abstract] OR "MR imaging"[title/abstract] OR "MR imaging"[title/abstract]) AND "case reports"[pt]. The response should be in the subsequent JSON format: "{Keywords:{"1":keyword1, "2":keyword2}}."

utilized for 2 subsequent tasks: Direct Diagnosis and Case Report Search.

Data Set

The COW portal of AJNR has featured educational neuroradiology imaging diagnostic quizzes approximately once a week from 2007 to the present date of October 2023. We extracted 121 cases from September 2021 to December 2023, ensuring they did not overlap with the training data set period for GPTs. The cases from September 2021 were used as the development set for determining prompts. The remaining cases after October 2021, within a time period that did not overlap with the GPTs' training data extraction period, were used as test cases for performance evaluation. Cases without a confirmed diagnosis in the Diagnosis section were excluded from the analysis. Textual information was extracted from the Description and Legends sections. Additionally, the word sequence presented at the top of the diagnosis section was defined as a "True Diagnosis." Because our study focused on radiologic evaluations, any histopathologic findings presented in the legends were removed. Furthermore, for the Case Report Search task mentioned later, case reports extracted from PubMed with the following search query were automatically designated as the "correct report set": "True Diagnosis"[title/abstract] AND ("Computed Tomography"[title/abstract] OR "CT"[title/abstract] OR "MR imaging"[title/abstract] OR "MR imaging"[title/abstract]) AND "case reports"[pt].

Prompts. Using the development set, we determined the prompts to ensure the GPT-3.5 Turbo and GPT-4 models yielded responses. The prompts were structured to prompt the LLMs to recognize themselves as "experienced neuroradiologists," reflecting the typical participant profile of AJNR's COW. First, a prompt for the Direct Diagnosis task in Table 1 was created and validated in the development set. Because both GPT models returned the 3 differential diagnoses as we intended, we determined the prompt for the Direct Diagnosis task. For the Case Report Search task, we used the first part of the Direct Diagnosis prompt and added the following sentences: "At present, you are considering several diseases for differential diagnosis. You may search PubMed for case reports to assist in your diagnostic process. What 2 keywords would you suggest for the search?" To avoid overly lengthy keyword extraction and to ensure a standardized format for responses, keywords should be limited to 3 words, and the output format should be in JavaScript Object Notation (JSON). This facilitates the automated creation of PubMed search queries. The final prompts are shown in Table 1.

The GPT models were accessed through OpenAI's API with the temperature parameter set to zero to ensure deterministic responses.

Direct Diagnosis. Cases were presented to the GPT-3.5 Turbo and GPT-4 models by using a structured prompt that instructed the model to provide 3 differential diagnoses in the order of likelihood based on patient information from the Description and Legends sections. A radiologist with 7 years of experience evaluated the AI-generated differential diagnoses. If any of the 3 proposed differential diagnoses were consistent with the True Diagnosis in each case, the response was classified as correct. From May 2023, ChatGPT-4 has been equipped with a web browsing function. Consequently, it was considered beneficial to evaluate its capability for the Direct Diagnosis task of ChatGPT-4 by using this function. However, it should be noted that there is a discrepancy in the data period used for training between the GPTs (GPT-3.5 Turbo [API], GPT-4 [API], and ChatGPT-4 [website]), and that web browsing function may allow ChatGPT to reference AJNR's COW page or related documents directly to generate responses. Therefore, we confined our evaluation of this function to a preliminary survey (Online Supplemental Data).

Case Report Search. The case information provided was the same as that in the Direct Diagnosis task, and the models were instructed to generate 2 keywords for the Case Report Search. Based on the GPT responses, a search query was generated and executed as follows: (keyword1) AND (keyword2) AND ("Computed Tomography"[title/abstract] OR "CT"[title/abstract] OR "MR imaging"[title/abstract] OR "MR imaging"[title/abstract]) AND "case reports"[pt]. We extracted up to 3 case reports in the order of relevance and verified whether they included the correct report set defined in the data set section. If the search query identified a single case report in the correct report set, it was considered a correct response.

Evaluation and Statistical Analysis. The percentage of correct responses was calculated for each model-task combination. The McNemar test was used to evaluate the differences in performance across the models for each task. We also used the McNemar test to examine whether there were differences in accuracy between cases involving Direct Diagnosis alone and those that included both Direct Diagnosis and Case Report Search. Statistical analyses were performed by using JMP Pro 17.0.0 software (JMP Statistical Discovery). Statistical significance was set at $P < .05$. Additionally, previous research has compiled the accuracy rates of ChatGPT for each etiology.[7] The aggregate

results based on these findings are also presented in the Online Supplemental Data.

RESULTS

Fig 2 presents a flow chart illustrating the selection process for these cases. One case was excluded from the test set because the diagnosis was not listed in the Diagnosis section, and the remaining 115 cases were designated as the test set for this study. A histogram of the number of correct reports is shown in Fig 3; 33 cases did not have any hits on PubMed for case reports qualifying as the correct report set. Note that these are cases for which there are no correct answers in the Case Report Search task. Table 2 presents the cases in which the correct response was obtained for each combination of model and task. For the Direct Diagnosis task, the correct response rate for GPT-3.5 Turbo was 30 out of 115 cases (26%), whereas that for GPT-4 was 47 out of 115 cases (41%). Regarding the Case Report Search task, GPT-3.5 Turbo achieved a correct response in 11 out of 115 cases (10%) and GPT-4 in 8 out of 115 cases (7%). When utilizing GPT-4, the performance on the Direct Diagnosis task was significantly better compared to the GPT-3.5 Turbo ($P < .001$), while there was no significant difference in the Case Report Search task ($P = .579$).

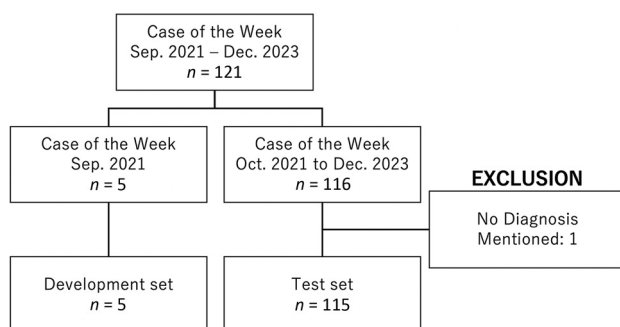Fig 4 shows the performance changes and overlap of correctly diagnosed cases resulting from adding Case Report Search to Direct Diagnosis. GPT-3.5 Turbo and GPT-4 overlapped in 4 and 5 cases, respectively. While adding Case Report Search improved the performance of GPT-3.5 Turbo ($P = .023$), no such improvement was observed for GPT-4 ($P = .248$). Results, including web search and aggregated tables for each etiology, are presented in Online Supplemental Data.

DISCUSSION

This study proposes Case Report Search, in addition to Direct Diagnosis, as a method of using GPTs to support radiologic interpretation. In the Direct Diagnosis task, GPT-3.5 Turbo had a correct response rate of 26% (30/115 cases), and GPT-4 had a rate of 41% (47/115 cases). In the Case Report Search task, GPT-3.5 Turbo achieved a correct response rate of 10% (11/115 cases), and GPT-4 achieved a correct response rate of 7% (8/115 cases). Correct responses totaled 32% (37/115) with 4 overlapping cases for GPT-3.5 Turbo, whereas GPT-4 had 43% (50/115) of correct responses with 5 overlapping cases. The significance of adding Case Report Search task was confirmed in GPT-3.5 Turbo.

With the widespread public availability of GPTs in medicine, there has been a surge of research exploring the various tasks to which these models can be applied, and this is no exception in the interpretation and conversion of radiologic descriptions.[11,12] An assessment explored the potential of GPTs in generating radiologic descriptions from concise imaging findings.[13] Another study harnessed the capabilities of GPT-4 for the post hoc transformation of free-text radiologic descriptions into structured formats.[14] Additionally, the feasibility of using ChatGPT to translate radiologic descriptions into plain language has been studied.[15] Furthermore, research has leveraged ChatGPT to fit report descriptions into existing disease classifications and grading systems.[15,16] Of course, there has also been a considerable focus on using GPTs to derive diagnoses from radiologic descriptions.[6,7,17] Such advent and evolution of LLMs have been notable aids in interpreting radiologic descriptions, yet their use may require caution. LLMs sometimes have the drawback of fabricating nouns, concepts, or information and producing answers that, although seemingly credible, lack verifiable evidence (hallucinations).[3,18–20] One of the most serious problems is citation fabrication (see Online Supplemental Data for an example).[21,22] Our proposed task, the Case Report Search, which allows verification against established literature in PubMed, might offer a way to avoid LLMs' citation fabrication explicitly.

We observed that combining the correct responses from Direct Diagnosis with those from Case Report Search enhanced the overall accuracy of GPT-3.5 Turbo. However, this combined approach did not yield a similar improvement in GPT-4 accuracy. Possible reasons for this relate to the ability to extract information. The GPT-3.5 Turbo might

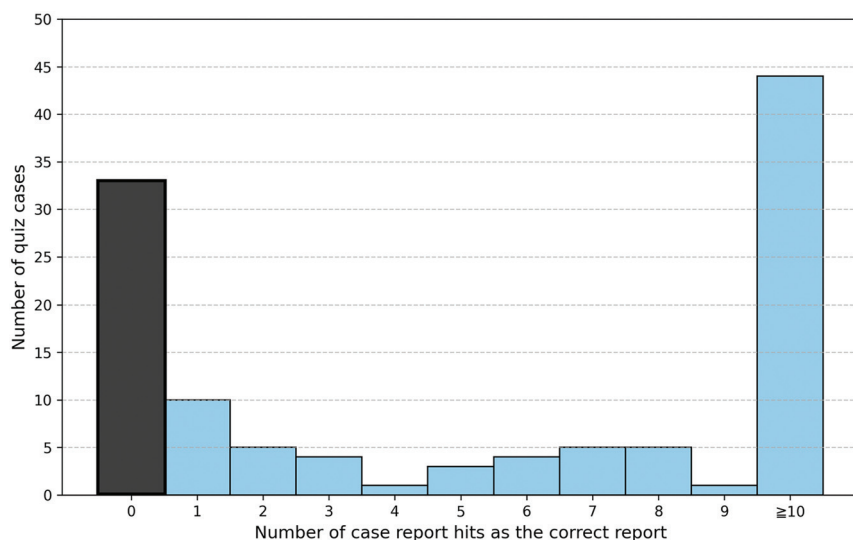**FIG 2.** Flow chart for development and test set determination.

FIG 3. Histogram of the number of correct reports per case. Among the 115 cases being analyzed, 33 cases yielded zero case report hits (*black bars*).

be less adept at identifying key details from the input than GPT-4, leading to a broader, less precise diagnostic performance. This could mean GPT-3.5 uses wider conceptual keywords for searches. On the other hand, GPT-4's enhanced skill in pinpointing specific information could make it more likely to focus on particular diagnoses. Yet, this precise focus might narrow the search too much, possibly missing out on relevant case reports. Conversely, the direct diagnostic performance of GPT-4 may be inherently higher, resulting in more overlaps between both tasks in GPT-4. Additionally, the performance of the Case Report Search task was not as high as the Direct Diagnosis, suggesting that there is potential for improvement in the task design itself. For example, the frequency of diseases reported in COW may differ from those found in case reports in PubMed, reflecting varying types of conditions typically published in each platform. Therefore, attempting to extract answers for COW cases only from PubMed might underestimate the capabilities of GPTs. Additionally, because the diagnosis section of COW is somewhat flexible in its writing, it might be better to modify these terms in the search to define the correct report set (in our setting, the correct report set could not be defined for 33 cases). Furthermore, in recent years, the practice of writing case reports with strict peer review has often shifted toward sharing fresher information on web services. COW is one of the typical examples of this. As many educational cases are not listed in PubMed, there is a need to explore frameworks that utilize a broader range of educational resources, not just searching PubMed case reports. In this sense, the use of web browsing functionality is considered useful, and preliminary investigations have been conducted (Online Supplemental Data).

LLMs are evolving at an incredible pace, not just through knowledge updates but also by expanding functionalities and integrations. Custom instructions and retraining for individual users are examples of such advancements. It is essential to continue to validate these technological advances in the field of radiology. For example, this course includes evaluating the performance of GPT models that have been retrained or fine-tuned by using COW contents. Additionally, while our current investigation was limited to textual analysis, LLMs are now becoming capable of handling multimodal inputs, and their ability to address medical quizzes by using a combination of medical images and descriptive texts is beginning to be explored.[23] As a resource for validating these technological advancements, the high-quality content and specialization in a fixed format, akin to COW, will remain crucial assets.

This study has several limitations. First, there is no guarantee that the prompts used for the GPT in this study were optimal. Depending on the prompt's content and manner of expression, the resulting outputs can vary, which presents a challenge when using LLMs.[24] Second, we have not been able to verify the full content of the "correct report set" in the case studies. However, based on our search methodology, it can be inferred that these reports are likely to contain useful information for an accurate diagnosis. Third, we did not conduct comparisons with other LLM services. Notably, some LLM services, including ChatGPT, can utilize web search results to create responses to queries. While this capability represents a notable improvement in explainability, it likely does not resolve all known issues, such as the generation of confident-seeming incorrect answers or the creation of fictitious terms. Furthermore, because the scope of this research was not to provide a comprehensive comparison of which LLM services were superior, these platforms were not included in the analysis. Additionally, the design of the Case Report Search, which involved searching for case reports within PubMed, may have had a restricted scope by not fully incorporating other forms of online teaching cases. Future studies should expand the search criteria to include various types of educational content across multiple search engines, not limited to PubMed.

Table 2: Number of correct responses by models and tasks

Task	GPT-3.5 Turbo	GPT-4	*P* Value
Direct diagnosis	30/115 (26%)	47/115 (41%)	<.001
Case report search	11/115 (10%)	8/115 (7%)	.579
Total	37/115 (32%)	50/115 (43%)	.009
(Overlap)	(4)	(5)	

FIG 4. Changes when adding case report search to direct diagnosis. For GPT-3.5 Turbo, there are 4 overlapping instances between the 2 tasks, while for GPT-4, there are 5 overlapping cases (highlighted in gray).

CONCLUSIONS

This study assessed the direct diagnostic capability and case report search query generation proficiency of GPT-3.5 Turbo and GPT-4, utilizing radiologic reports and findings. While GPT-4 demonstrated superior Direct Diagnostic capabilities, adding the Case Report Search improved the performance of GPT-3.5 Turbo. These findings suggest that when using GPT-3.5
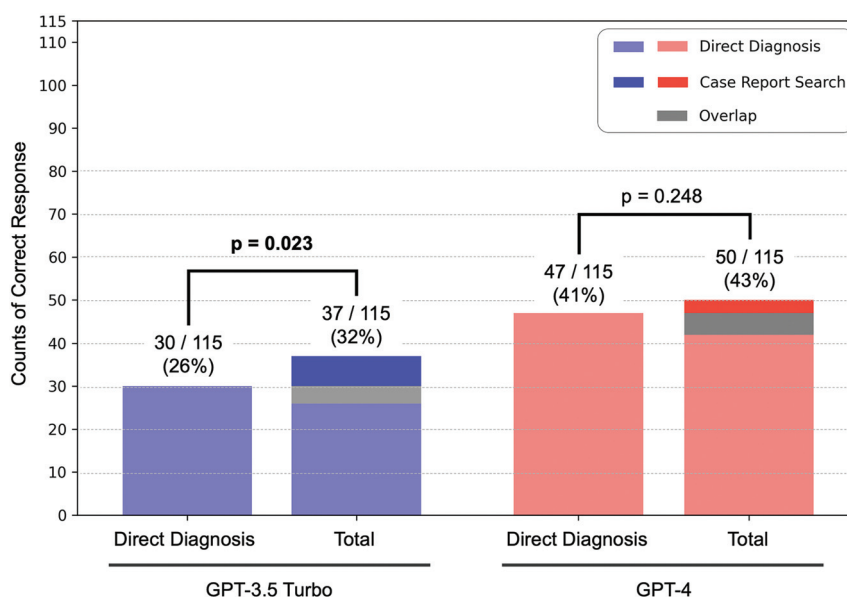
Turbo, a case report search may be considered to derive diagnostic information from the radiologic descriptions in addition to directly asking for a diagnosis. However, the final results obtained through this method have not achieved optimal performance, indicating a need for awareness of the current capabilities, strengths, and limitations in the appropriate use of GPTs.

Disclosure forms provided by the authors are available with the full text and PDF of this article at www.ajnr.org.

REFERENCES

1. Youssef A, Ng MY, Long J, et al. **Organizational factors in clinical data sharing for artificial intelligence in health care.** *JAMA Netw Open* 2023;6:e2348422 CrossRef Medline
2. Clusmann J, Kolbinger FR, Muti HS, et al. **The future landscape of large language models in medicine.** *Commun Med (Lond)* 2023; 3:141 CrossRef Medline
3. Shaikh O, Zhang H, Held W, et al. **On second thought, let's not think step by step! Bias and toxicity in zero-shot reasoning.** Presented at: *Annual Meeting of the Association for Computational Linguistics*; July 2023; Toronto, Canada. https://aclanthology.org/2023.acl-long.244/ CrossRef
4. OpenAI. **GPT-4 Technical Report.** arXiv [csCL] 2023 Mar 15; epub ahead of print.
5. Suthar PP, Kounsal A, Chhetri L, et al. **Artificial intelligence (AI) in radiology: a deep dive into ChatGPT 4.0's accuracy with the American Journal of Neuroradiology's (AJNR) "Case of the Month.** *Cureus* 2023;15:e43958 CrossRef Medline
6. Ueda D, Mitsuyama Y, Takita H, et al. **ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes.** *Radiology* 2023;308:e231040 CrossRef Medline
7. Horiuchi D, Tatekawa H, Shimono T, et al. **Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases.** *Neuroradiology* 2024;66:73–79 CrossRef Medline
8. Albrecht J, Meves A, Bigby M. **Case reports and case series from Lancet had significant impact on medical literature.** *J Clin Epidemiol* 2005;58:1227–32. CrossRef Medline
9. Nissen T, Wynn R. **The clinical case report: a review of its merits and limitations.** *BMC Res Notes* 2014;7:264 CrossRef Medline
10. Cohen JF, Korevaar DA, Altman DG, et al. **STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration.** *BMJ Open* 2016;6:e012799 CrossRef Medline
11. Grewal H, Dhillon G, Monga V, et al. **Radiology gets chatty: the ChatGPT saga unfolds.** *Cureus* 2023;15:e40135 CrossRef Medline
12. Kim S, Lee CK, Kim SS. **Large language models: A guide for radiologists.** *Korean J Radiology* 2024;25:126–33 CrossRef Medline
13. Nakaura T, Yoshida N, Kobayashi N, et al. **Preliminary assessment of automated radiology report generation with generative pretrained transformers: comparing results to radiologist-generated reports.** *Jpn J Radiology* 2024;42:190–200 CrossRef Medline
14. Adams LC, Truhn D, Busch F, et al. **Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: A multilingual feasibility study.** *Radiology* 2023;307:e230725 CrossRef Medline
15. Lyu Q, Tan J, Zapadka ME, et al. **Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential.** *Vis Comput Ind Biomed Art* 2023;6:9 CrossRef Medline
16. Nakamura Y, Kikuchi T, Yamagishi Y, et al. **ChatGPT for Automating Lung Cancer Staging: Feasibility Study on Open Radiology Report Data set.** medRxiv CrossRef
17. Gunes YC, Cesur T. **A Comparative Study: Diagnostic Performance of ChatGPT. Google, Microsoft Bing, and Radiologists in Thoracic Radiology cases.** *bioRxiv:3.5.* CrossRef
18. Schramowski P, Turan C, Andersen N, et al. **Large pre-trained language models contain human-like biases of what is right and wrong to do.** *Nat Mach Intell* 2022;4:258–68 CrossRef
19. Pal A, Umapathi LK, Sankarasubbu M. **Med-HALT: Medical Domain Hallucination Test for Large Language Models.** arXiv: 2307.15343 [csCL] CrossRef
20. Karabacak M, Margetis K. **Embracing large language models for medical applications: opportunities and challenges.** *Cureus* 2023; 15:e39305 CrossRef Medline
21. Walters WH, Wilder EI. **Fabrication and errors in the bibliographic citations generated by ChatGPT.** *Sci Rep* 2023;13:14045 CrossRef Medline
22. McGowan A, Gui Y, Dobbs M, et al. **ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search.** *Psychiatry Res* 2023;326:115334 CrossRef Medline
23. Meskó B. **The impact of multimodal large language models on health care's future.** *J Med Internet Res* 2023;25:e52865 CrossRef Medline
24. Meskó B. **Prompt engineering as an important emerging skill for medical professionals: tutorial.** *J Med Internet Res* 2023;25:e50638 CrossRef Medline