

GPT-4 Performance for Neurologic Localization

Jung-Hyun Lee, MD, Eunhee Choi, MD, Robert McDougal, MD, and William W. Lytton, MD

Neurology: Clinical Practice 2024;14:e200293. doi:10.1212/CPJ.0000000000200293

Correspondence

Dr. Lee

jung-hyun.lee@downstate.edu

Abstract

Background and Objectives

In health care, large language models such as Generative Pretrained Transformers (GPTs), trained on extensive text datasets, have potential applications in reducing health care disparities across regions and populations. Previous software developed for lesion localization has been limited in scope. This study aims to evaluate the capability of GPT-4 for lesion localization based on clinical presentation.

Methods

GPT-4 was prompted using history and neurologic physical examination (H&P) from published cases of acute stroke followed by questions for clinical reasoning with answering for “single or multiple lesions,” “side,” and “brain region” using Zero-Shot Chain-of-Thought and Text Classification prompting. GPT-4 output on 3 separate trials for each of 46 cases was compared with imaging-based localization.

Results

GPT-4 successfully processed raw text from H&P to generate accurate neuroanatomical localization and detailed clinical reasoning. Performance metrics across trial-based analysis for specificity, sensitivity, precision, and F1-score were 0.87, 0.74, 0.75, and 0.74, respectively, for side; 0.94, 0.85, 0.84, and 0.85, respectively, for brain region. Class labels within the brain region were similarly high for all regions except the cerebellum and were also similar when considering all 3 trials to examine metrics by case. Errors were due to extrinsic causes—inadequate information in the published cases, and intrinsic causes—failures of logic or inadequate knowledge base.

Discussion

This study reveals capabilities of GPT-4 in the localization of acute stroke lesions, showing a potential future role as a clinical tool in neurology.

Introduction

A Generative Pretrained Transformer (GPT) is a large language model (LLM) trained on extensive text datasets across multiple domains of knowledge (ChatGPT is a user-friendly chatbot application that uses GPT).¹ GPT has shown remarkable capabilities in generating free-text responses to text-based queries.¹ Among its diverse applications, the potential for use in health care is of particular significance due to the great need for improved health care in underserved areas across multiple countries; notably, GPT-4 can function across different languages.²⁻⁵ Previous medicine-related studies have explored LLM functionality in answering

RELATED ARTICLE

Editorial

Exploring the Potential of Large Language Models in Neurology, Using Neurologic Localization as an Example

Page e200311

Department of Neurology (J-HL, WWL), State University of New York Downstate Health Sciences University; Department of Neurology (J-HL, WWL), Kings County Hospital; Department of Neurology (J-HL), Maimonides Medical Center, Brooklyn; Department of Internal Medicine (EC), Lincoln Medical Center, Bronx, NY; Department of Biostatistics (RM), Yale School of Public Health; Program in Computational Biology and Bioinformatics (RM); Wu-Tsai Institute (RM); Section of Biomedical Informatics and Data Science (RM), Yale School of Medicine, Yale University, New Haven, CT; and Department of Physiology and Pharmacology (WWL), State University of New York Downstate Health Sciences University, Brooklyn, NY.

Funding information and disclosures are provided at the end of the article. Full disclosure form information provided by the authors is available with the full text of this article at [Neurology.org/cp](https://www.neurology.org/cp).

The Article Processing Charge was funded by the authors.

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND), which permits downloading and sharing the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

free-form physician questions and medical board examination questions, in grading Glasgow coma score from a written neurologic examination, and in summarizing history of present illness or medical abstracts.⁶⁻¹³ Prior neurologic examination-based localization software has been limited to a few hypothetical cases.¹⁴ Our study explored the utility of GPT-4 in facilitating neuroanatomical localization in acute stroke scenarios using history and neurologic physical examination (H&P) from published case reports.

Methods

Study Design

We used *BMC Neurology* case reports as our primary source to data-mine due to the ease of handling its standardized case report format. Initial *PubMed Central* search was performed with query (“BMC Neurology” [Journal] AND “case” [Title]) AND (“stroke” [MeSH Terms] OR “infarction” [MeSH Terms] OR “ischemia” [MeSH Terms]). Exclusion criteria were as follows: mislabeled as case reports, not an acute stroke, pediatric, without imaging, focus on treatment failures or complications, insufficient clinical description, preexisting focal deficits, diencephalic strokes presenting with hypersomnia, and duplicates. Due to pulling too few spinal cord cases, an additional query (“case” [Title]) AND (“spinal” [Title]) AND (“stroke” [MeSH Terms] OR “infarction” [MeSH Terms] OR “ischemia” [MeSH Terms]) was used, yielding 30 cases of which 8 were useable. From the final selection of cases, we manually extracted unedited text of H&P and deleted any localization information that had been placed in these sections (except for replacing dermatome descriptions that revealed localization, e.g., “below the T4 level” to “below the nipple”).¹⁵ Each case input was presented to GPT-4 in 3 separate trials to identify the

variability in response due to internal GPT-4 processing randomization.

Case reports from *PubMed Central* were accessed in August 2023 and analyzed in September 2023. The study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline.

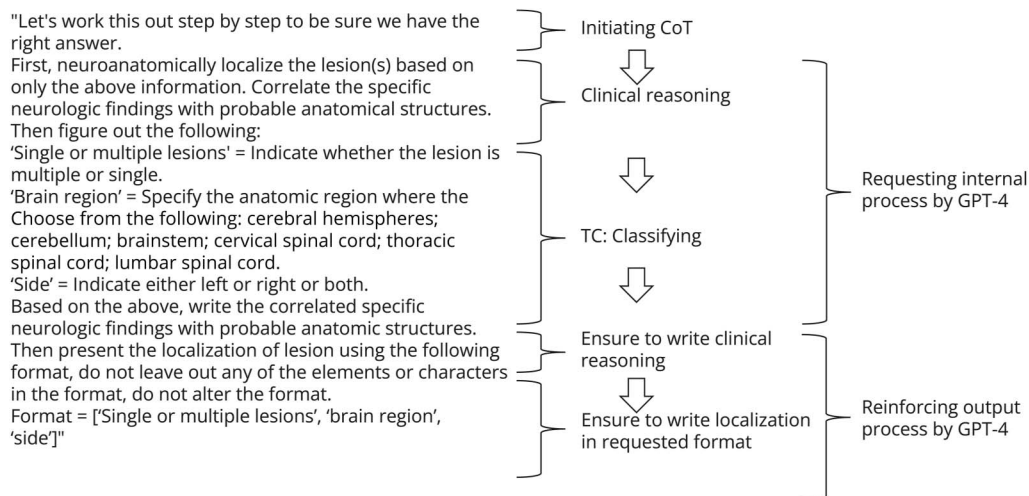
Prompt Engineering

We used 2 prompt engineering methods in tandem to generate our prompt: 1. *Zero-Shot Chain-of-Thought (CoT)* prompting; 2. *Text Classification (TC)* prompting (Figure 1). CoT prompting uses a sequence of concise sentences that provide a logical framework; here, we input, “Let’s work this out step-by-step to ensure accuracy,” to initiate a reasoning chain.¹⁶ The chain requested information for single or multiple lesions, brain region, and finally, side.¹⁶ Within the chain, TC prompting gives choices by providing class label examples.^{17,18}

Statistical Analysis

Most previous LLM studies in medicine have used measures of sensitivity and specificity, familiar from medical diagnostics where tests give yes/no binary truth values.⁶⁻⁹ In this study, we used GPT-4 as a classifier—it could provide many different answers for localization. We therefore used a wider range of performance metrics: specificity (proportion of negative samples of a given category correctly classified), sensitivity (the fraction of a given category correctly identified, equivalent to recall), precision (fraction of predictions that are correct, equivalent to positive predictive value), and F1-score (harmonic mean of precision and sensitivity).¹⁹ The overall statistical analysis was conducted on both trial-based and case-based assessments.

Figure 1 Prompt Design



Sequencing through chain-of-thought (CoT) with classification through text classification (TC). The final line identifies the final answer format.

Table Demonstration of Lesion Localization Provided by GPT-4 With a Sample Case (ID: PMC8686214)²⁹

Case H&P	A 55-year-old man presented with sudden onset of dizziness, diplopia, difficulty of closing eyes, and trouble of chewing after lifting heavy goods 3 d before admission. He had a history of mild hypertension for 1 y with treatment of amlodipine 2.5 mg/d and poorly controlled asthma. On examination, he showed completed bilateral horizontal gaze palsy, which was uncorrected by vestibulo-ocular reflex. Gazed-evoked [sic] upbeat nystagmus (UBN) was observed on attempted upward gaze but not on straight-ahead gaze position. In addition, he had bilateral peripheral facial paralysis with predominance on the left. Examination on other cranial nerves, including facial sensation, taste, hearing, and pharyngeal reflex, was normal. Mild ataxia was noticed on the left upper extremity when performing finger-to-nose test. His muscle strength was 5 on 4 limbs, and he had normal pinprick sensation and brisk tendon reflexes.
Prompt	“Let’s work this out step by step to be sure we have the right answer. First, neuroanatomically localize the lesion(s) based on only the above information. Correlate the specific neurologic findings with probable anatomical structures. Then figure out the following: ‘Single or multiple lesions’ = Indicate whether the lesion is multiple or single. ‘Brain region’ = Specify the anatomical region where the lesion is located. Choose from the following: cerebral hemispheres; cerebellum; brainstem; cervical spinal cord; thoracic spinal cord; lumbar spinal cord. ‘Side’ = Indicate either left or right or both. Based on the above, write the correlated specific neurologic findings with probable anatomical structures. Then present the localization of lesion using the following format, do not leave out any of the elements or characters in the format, do not alter the format. Format = [‘Single or multiple lesions’, ‘Brain region’, ‘Side’]”
GPT-4 output	This patient’s presentation is consistent with multiple lesions in the brainstem affecting both sides. The specific neurologic findings can be correlated with the following anatomical structures: (1) The bilateral horizontal gaze palsy that could not be corrected by vestibulo-ocular reflex suggests involvement of the pontine gaze centers (paramedian pontine reticular formation) on both sides. (2) Gaze-evoked upbeat nystagmus (UBN) on attempted upward gaze implies potential lesion or dysfunction of the midbrain. (3) Bilateral peripheral facial paralysis, with predominance on the left, is suggestive of lesions in the facial nuclei or the facial nerves. This combined with the bilateral gaze palsy further suggests a central pattern of facial palsy (Moebius syndrome). (4) Mild ataxia in the left upper extremity, as demonstrated by the finger-to-nose test, suggests involvement of cerebellar pathways, possibly linked to the inferior cerebellar peduncle located in the medulla. [‘Multiple lesions’, ‘Brainstem’, ‘Both sides’]
Imaging result	MRI: Diffusion-weighted image 1 d after disease onset revealed increased restricted diffusion signal in the right middle cerebellar peduncle and bilateral dorsal pontine tegmentum. Answer: [‘Multiple lesions’, ‘Brainstem’, ‘Both sides’]

The case report is taken verbatim with English errors uncorrected.

The trial-based assessment was performed across 3 trials for each case (46 cases × 3 trials = 138 trials). Statistical analysis involving multilabel classifications precluded computing performance metrics from a single confusion matrix.²⁰ To address this, the binary relevance method was used, transforming multilabel classifications into a series of binary classifications and generating distinct confusion matrices for each class label.²⁰ For example, for the category “side,” separate confusion matrices were generated for each class label “right,” “left,” and “both.” For “brain region,” the confusion matrix and performance metrics of class labels “cerebral hemisphere,” “brainstem,” “cerebellum,” and “spinal cord” were generated with the abovementioned method. The category “spinal cord” included cervical and thoracic cases, which were individually analyzed but then lumped in the final analysis due to the paucity of spinal cord cases. We used microaveraging, aggregating confusion matrices of each class label for individual categories to compute performance metrics to take into consideration the uneven distribution of cases across different class labels.²⁰

Case-based assessment was performed by combining all 3 trials of each individual case to determine localization performance, accounting for the variability in GPT-4 response to identical inputs. This was performed by identifying the majority answer (2 or all 3 trials) for each case for brain region and side (omitting multiple or single lesions) as the output. The abovementioned statistical method was then used to

construct confusion matrices across class labels for left or right for each brain region, again aggregating specificity, sensitivity, precision, and F1-score for each category using microaveraging.

Standard Protocol Approvals, Registrations, and Patient Consents

Data used for this study are publicly available, and no patients were involved in our study; therefore, Institutional Review Board approval was not required.

Data Availability

We have ensured that all the essential data necessary for replicating our results are included in our Supplementary file.

Results

An initial *PubMed Central* search yielded 116 articles; 78 were excluded (eFigure 1). Including 8 additional spinal cord cases brought us to 46 total cases. Based on the 3 categories: “single or multiple lesion,” “brain region,” and “side,” cases were allocated to class labels in each category for performance evaluation. The study was conducted under a rate limit of 200 requests per minute and 10,000 tokens per minute (100 tokens equivalent to approximately 75 words). Of the 138 trials, there was 1 indeterminate, no blank, and no unexpected outputs.

Confabulated contents were seen in only 2% of trials (3 of 138). Two of these 3 still produced correct results; 1 produced incorrect localization. Of the 46 cases examined, consistency in outputs across all 3 trials was observed in 35 (76%) cases for single or multiple lesions, 40 (87%) for brain region, and 38 (83%) for side. The output of each trial was compared with the answer manually extracted from the imaging findings reported in the original case report. Nineteen (41%) of the cases were found to be accurate across all 3 trials in all 3 categories.

Sample Demonstration

A sample case report demonstrates a relatively complex case requiring sustained neurologic reasoning (Table). As in all cases, GPT-4 received unstructured, verbatim H&P input taken from the journal, followed by the prompt. GPT-4 then generated clinical reasoning along with neuroanatomical localization, including structured output for the answer.

The long-form output of GPT-4's step-by-step analysis showed neurologic reasoning skills, correlating the specific neurologic findings with individual anatomical structures (paramedian pontine reticular formation, midbrain, facial nerve nuclei, and cerebellum). The reasoning skill of GPT-4 may be considered comparable with an early neurology trainee; for example, localization reasoning could have been more thorough regarding upward nystagmus if GPT-4 had considered alternative structures, including superior cerebellar peduncles, caudal medulla, and cerebellar vermis while narrowing down the location of the lesion.²¹ GPT-4 provided specific structures but did not always provide thorough anatomical and clinical reasoning for each finding. GPT-4 finally accurately summarized the localization in the

requested format (multiple lesions, brainstem, and both sides).

Statistical Analysis of Localization by GPT-4

Across trials, GPT-4 identified lesion side well, showing performance metrics of specificity (SP) 0.87, sensitivity (SE) 0.74, precision (P) 0.75, and F1-score (F) 0.74. Brain region classification was slightly better: SP 0.94, SE 0.85, P 0.84, and F 0.85. Breakdown across brain regions demonstrated strong performance for all regions except the cerebellum (Figure 2, eTable 1). Detection for single or multiple lesions yielded performance metrics of SP 0.73, SE 0.73, P 0.73, and F 0.73 (all the same). Individual class labels showed an F1-score of 0.33 for multiple lesions and 0.83 for single lesions, demonstrating that the principal difficulty was in identifying multiple lesions.

Case-based analysis (across the 3 trials of each case) was based on linked side and brain region localization. Overall performance was SP 0.95, SE 0.75, P 0.76, and F 0.75 with individual region performance metrics comparable with the trial-based analysis (Figure 3, eTable 2).

GPT-4 Clinical Reasoning

GPT-4 showed good localization with the exception of poorer results in the cerebellum (Figures 2 and 3, eTable 3). We identified errors due to extrinsic causes, limitations or ambiguities in the inputs, vs those due to intrinsic causes—errors from GPT-4 (eTables 4 and 5).

Extrinsic causes of errors were often due to limited description, confounding symptoms, and rare anatomical variant cases. This problem largely stemmed from using case

Figure 2 Trial-Based Performance Metrics Include Specificity, Sensitivity, Precision, and F1-Score Across Individual Brain Regions

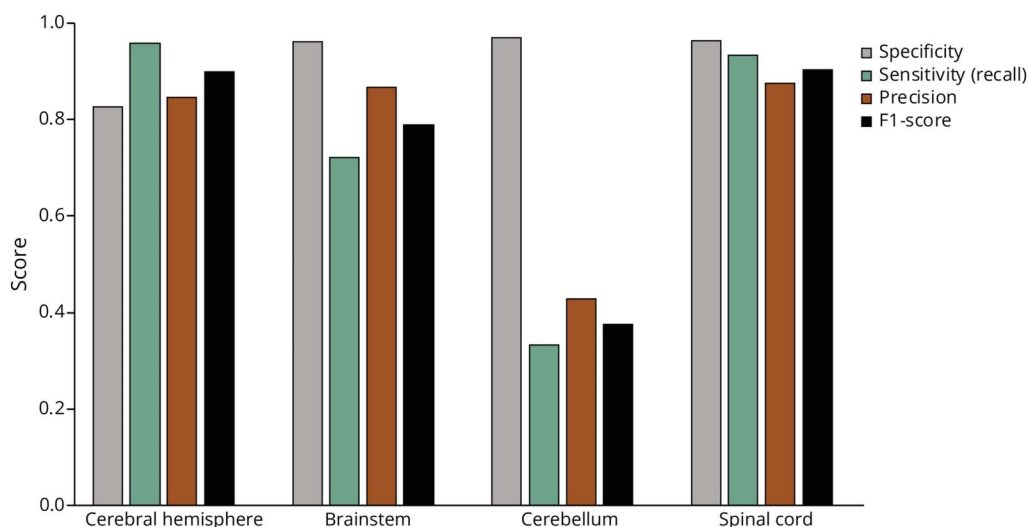
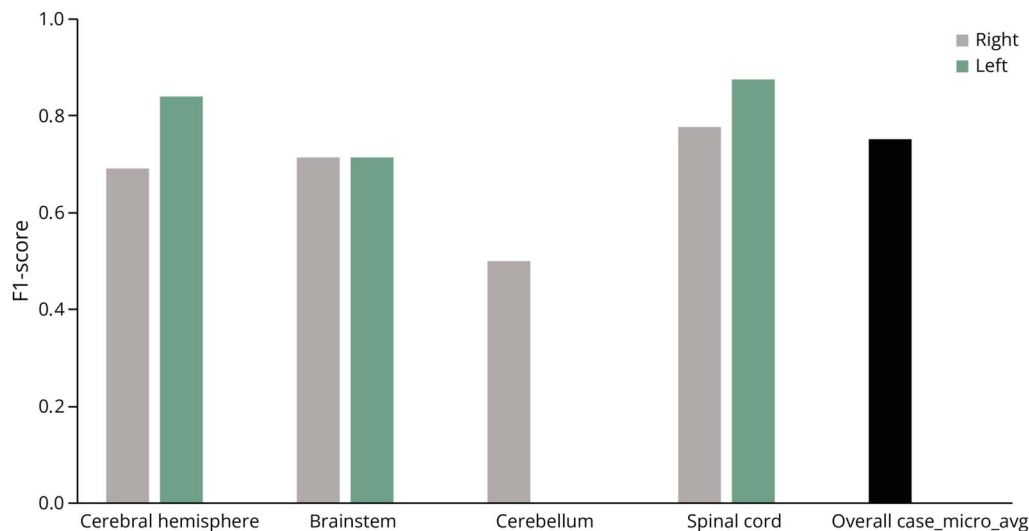


Figure 3 Case-Based F1-Score for Localization Across Individual Side-Brain Region Class Labels



The value for the left cerebellum was zero.

reports rather than sequential clinical presentations from a hospital. Case reports are unusual presentations that may be misleading from the start and are often the subject of chief rounds where they are puzzled over. Examples included an unusual case with a congenital lack of pyramidal decussation; a case where seizures related to a prior cerebral bleed obscured an acute brainstem lesion; a case of corticobasal degeneration whose symptoms were discovered in the setting of a pontine stroke; a subcortical infarction that presented with symptoms and signs of a peripheral neuropathy. We did not pick and choose cases to eliminate these conundrums. However, we did remove a set of diencephalic cases where presentations were ambiguous to the point of not even clearly localizing to the nervous system. Another cause of extrinsic error arose from ambiguities in brainstem vs cerebellar damage: a patient presenting with dysarthria and nystagmus could have either or both.

Errors originating from intrinsic causes stemmed from ignoring neurologic findings, confabulation (the tool, not the patient), and inadequate or inaccurate neurologic knowledge or faulty reasoning (not generally possible to distinguish). Ignored finding: in a case of cervical cord lesion, the upper extremity weakness noted in H&P was ignored in 2 of the 3 trials, leading to mistaken localization in the thoracic cord (eTable 5). Confabulation: in the brainstem case with dysarthria and nystagmus mentioned earlier, GPT-4 added a symptom of gait instability in 1 of 3 trials for that case—the addition of a frequently associated symptom resulted in attribution to a cerebellar lesion. Logical error or inadequate anatomical knowledge: wrong-side error for hemiplegia ranged from multiple errors across 3 cervical spinal cord lesion cases—GPT-4 seemed inconsistently aware of the corticospinal decussation being above the cord. Similarly,

1 of 3 trials incorrectly labeled a right hemispheric lesion case as left hemispheric. In the cervical lesion case mentioned earlier with arm weakness, in 1 of 3 trials, GPT-4 acknowledged the clinical finding of arm weakness; however, it still chose thoracic. There were also intrinsic errors due to inconsistencies in copying from the long-form to the final answer and lack of thoroughness of clinical reasoning as described earlier.

GPT-4 exhibited more extrinsic contributions to errors in identifying the brain region and side (eTable 4). Performance for the brain region was best in hemispheric and spinal cord regions, where clear localization by cortical sign or sensory level was typical and straightforward in most of the cases. For the brainstem, errors were generally due to confounding symptoms. Localization was worst in the cerebellum, where descriptions and number of cases were limited, and where cases had multiple lesions.

Discussion

Our study underscores the potential of GPT-4 in lesion localization in neurologic cases, demonstrating high performance in identifying brain region (F1-score 0.85) and side (0.74). Case-based analysis combining region and side across trials gave similarly good results (0.75). This relatively good performance is remarkable given that our results were achieved using published case reports of atypical cases, some lacking adequately detailed H&P, and working with a LLM without medical training or fine-tuning.

For the casual user of ChatGPT, LLM application research may seem to be simply a mindless task of dumping information into a funnel and pulling information out from the

bottom. However, 2 main tasks confront the researcher. The first task is the manipulation of inputs; in our study, we did very little other than stripping any direct localization information from the raw text. This lack of structuring of inputs was desirable based on the expectation of using similarly raw, incomplete inputs from busy emergency departments or clinic health workers in future applications.

The second essential task is prompt design. We combined CoT and TC prompting, both used extensively in GPT studies (Figure 1).¹⁶⁻¹⁸ Prompt engineering involves selecting the prompt pattern that fits with the task objective and then expanding the individual prompt questions.²² We used CoT to give GPT-4 step-by-step suggestions, asking to correlate the specific neurologic findings with probable anatomical structures, to figure out single or multiple, brain region, and side and to present lesion localization using the format with reinforcing statements to produce the correct output. Using TC, we provided specific class label samples for each category. Once the prompt pattern was determined, we developed the individual prompt queries. We investigated both zero-shot, a simpler direct question, and few-shot—an example question-answer set preceding the question.²² We did not achieve any significant improvement with few-shot; this seemed to addle GPT-4, which added irrelevant information to the answers. Therefore, we used the more straightforward zero-shot procedure. Our systematic prompt engineering produced improved results compared with simple instructional prompting methods.²³

One of the concerns of using an LLM in a clinical setting is the ethical component of transparency to both health care providers and patients. Our study displays a way to systematically scrutinize the GPT-4 outputs by assessing the clinical reasonings for localization, which could potentially address transparency issues. We specifically asked GPT-4 to correlate an anatomical structure to each neurologic finding to dissect the logical process of GPT-4. Moreover, we identified which errors arose from extrinsic causes, largely insufficient information, vs intrinsic causes, errors made by the tool. This process can be used to identify problem areas in future LLM research to improve performance.

Our study suggests potential for future application of GPT not only in stroke localization but also more broadly in neurologic diagnosis. The recent surge in telemedicine usage during and post pandemic highlights the importance of neurologic expertise for stroke intervention, particularly in remote hospitals lacking resources for on-call neurologic care—initial CT has only 31% stroke sensitivity.^{24,25} Although our study was focused on localization rather than on diagnosis, the importance of our study is more in its ability to form conclusions from plain-text H&P, a “skill” that would be particularly valuable for neurologic disorders where imaging is not diagnostic and where other biomarkers are not yet available. For example, Parkinson disease largely involves a clinical suspicion over a prolonged time course, which may

TAKE-HOME POINTS

- GPT-4, a large language model (LLM) with no specific medical training, was able to accurately localize stroke.
- Natural language processing will permit neurologic computer applications to interpret clinical free text.
- Prompt engineering (telling the program what you want) is crucial for obtaining useful results from LLMs in our domain, as in others.
- Our analysis of GPT-4 errors showed that many errors were due to inadequate or contradictory inputs in the cases, while some errors were “logical” or knowledge base errors from GPT-4.
- Our study provides an initial template for future development of LLMs for clinical use, but medicine as a field will need to address multiple issues including patient privacy and safety and ensuring greater LLM accuracy and consistency.

prompt referrals through orthopedics, rheumatology, and other specialties in its early and sometimes even late stages.²⁶

A major limitation of this study was the use of published case reports, which reduced the applicability of the study to everyday clinical practice. As noted earlier, case reports are published because they are unusual, hence not typical of cases that are encountered in the clinic. In addition, a published H&P will not be complete and will not follow the format of a full clinical admission note. This lack of completeness also meant that the case reports we used did not generally provide sufficient information for more detailed localization, particularly lacking pertinent negatives. Hence, we could only localize to broad anatomical regions. Another study limitation is that published H&Ps provide carefully vetted, detailed evaluations by neurologists. Evaluations conducted by other medical professionals, and performed on the first patient encounter, will lack this high level of neurologic detail and accuracy and will include more general medical information—potential distractors.

Other study limitations are specific to the tool itself. GPT-4 has not been trained or fine-tuned for neurology or neurologic diseases and thus is not illustrative of the future performance capabilities of a bespoke AI tool. Finally, there are many difficulties and further research to be conducted to transition from LLM literature analysis to a viable utilization in clinical settings: logistical, legal, and patient privacy issues; patient safety; integration with hospital workflows; and

ensuring accuracy and consistency are just some of the many challenges.²⁷ Our study represents an early exploration in this domain and highlights the need for continued research and development to address these multifaceted challenges effectively.

Our method of developing a prompt and subsequently analyzing GPT-4's errors provides a foundation for developing LLM for medical use. Future clinical use of LLM with fine-tuning and training focused on neurologic diseases and neuroanatomy would particularly assist non-neurologist health care providers with limited training in localization and diagnosis of nervous system disease.²⁸ LLM could provide timely decision-support in assessing the need for additional imaging studies or neurologist consultation. Furthermore, this study provides a basis for a broader role of LLMs in neurology, where we aim to encourage the involvement of diverse groups of health care participants in addition to medical professionals: researchers, patients, and policymakers.

Study Funding

The authors report no targeted funding.

Disclosure

The authors report no relevant disclosures. Full disclosure form information provided by the authors is available with the full text of this article at [Neurology.org/cp](https://www.neurology.org/cp).

Publication History

Received by *Neurology: Clinical Practice* December 21, 2023. Accepted in final form January 23, 2024. Submitted and externally peer reviewed. The handling editor was Editor Luca Bartolini, MD, FAAN, FAES.

Appendix Authors

Name	Location	Contribution
Jung-Hyun Lee, MD	Department of Neurology, State University of New York Downstate Health Sciences University; Department of Neurology, Kings County Hospital; Department of Neurology, Maimonides Medical Center, Brooklyn, NY	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data; study concept or design; and analysis or interpretation of data
Eunhee Choi, MD	Department of Internal Medicine, Lincoln Medical Center, Bronx, NY	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data; study concept or design; and analysis or interpretation of data
Robert McDougal, MD	Department of Biostatistics, Yale School of Public Health; Program in Computational Biology and Bioinformatics; Wu-Tsai Institute; Section of Biomedical Informatics and Data Science, Yale School of Medicine, Yale University, New Haven, CT	Drafting/revision of the article for content, including medical writing for content; analysis or interpretation of data

Appendix (continued)

Name	Location	Contribution
William W. Lytton, MD	Department of Neurology, State University of New York Downstate Health Sciences University; Department of Neurology, Kings County Hospital; Department of Physiology and Pharmacology, State University of New York Downstate Health Sciences University, Brooklyn, NY	Drafting/revision of the article for content, including medical writing for content; study concept or design; and analysis or interpretation of data

References

1. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Physical Syst.* 2023;3:121-154. doi:10.1016/j.iotcps.2023.04.003
2. Wang X, Sanders HM, Liu Y, et al. ChatGPT: promise and challenges for deployment in low- and middle-income countries. *Lancet Reg Health West Pac.* 2023;41:100905. doi:10.1016/j.lanwpc.2023.100905
3. Crema C, Attardi G, Sartiano D, Redolfi A. Natural language processing in clinical neuroscience and psychiatry: a review. *Front Psychiatry.* 2022;13:946387. doi:10.3389/fpsy.2022.946387
4. Sahni NR, Carrus B. Artificial intelligence in U.S. health care delivery. *N Engl J Med.* 2023;389(4):348-358. doi:10.1056/NEJMra2204673
5. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health.* 2023;5(3):e107-e108. doi:10.1016/S2589-7500(23)00021-3
6. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ.* 2023;9:e45312. doi:10.2196/45312
7. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurgery.* 2023;93(5):1090-1098. doi:10.1227/neu.0000000000002551
8. Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK neurology specialty certificate examination. *BMJ Neurol Open.* 2023;5(1):e000451. doi:10.1136/bmjno-2023-000451
9. Nayak A, Alkatis MS, Nayak K, Nikolov M, Weinfurt KP, Schulman K. Comparison of history of present illness summaries generated by a chatbot and senior internal medicine residents. *JAMA Intern Med.* 2023;183(9):1026-1027. doi:10.1001/jamainternmed.2023.2561
10. Tang L, Sun Z, Idray B, et al. Evaluating large language models on medical evidence summarization. *NPJ Digit Med.* 2023;6(1):158. doi:10.1038/s41746-023-00896-7
11. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and reliability of chatbot responses to physician questions. *JAMA Netw Open.* 2023;6(10):e2336483. doi:10.1001/jamanetworkopen.2023.36483
12. Chen TC, Kaminski E, Koduri L, et al. Chat GPT as a neuro-score calculator: analysis of a large language model's performance on various neurological exam grading scales. *World Neurosurg.* 2023;179:e342-e347. doi:10.1016/j.wneu.2023.08.088
13. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine.* 2023;95:104770. doi:10.1016/j.ebiom.2023.104770
14. Kamo H, Kiriya Y, Mizoe A, et al. A computer support system for neurological anatomical diagnosis. *Clin Neurol Neurosurg.* 2008;110(10):973-978. doi:10.1016/j.clineuro.2008.05.022
15. Lee MWL, McPhee RW, Stringer MD. An evidence-based approach to human dermatomes. *Clin Anat.* 2008;21(5):363-373. doi:10.1002/ca.20636
16. Zhou Y, Muresanu AJ, Han Z, et al. Large language models are human-level prompt engineers. *arXiv [cs.LG]. arXiv [cs.LG].* Published online November 3, 2022. <https://openreview.net/pdf?id=92gvk82DE->
17. Schick T, Schütze H. Exploiting cloze-questions for few-shot text classification and natural language inference. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume.* Association for Computational Linguistics; 2021:255-269.
18. Han X, Zhao W, Ding N, Liu Z, Sun M. PTR: prompt tuning with rules for text classification. *arXiv [cs.CL].* Published online May 24, 2021. <http://arxiv.org/abs/2105.11259>
19. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep.* 2022;12(1):5979. doi:10.1038/s41598-022-09954-8
20. Zhang ML, Zhou ZH. A review on multi-label learning algorithms. *IEEE Trans Knowl Data Eng.* 2014;26(8):1819-1837. doi:10.1109/tkde.2013.39
21. Pierrot-Deseilligny C, Milea D. Vertical nystagmus: clinical facts and hypotheses. *Brain.* 2005;128(Pt 6):1237-1246. doi:10.1093/brain/awh532
22. Wang J, Shi E, Yu S, et al. Prompt engineering for healthcare: methodologies and applications. *arXiv [cs.AI].* Published online April 28, 2023. doi:10.48550/ARXIV.2304.14670
23. Galetta K, Meltzer E. Does GPT-4 have neurophobia? Localization and diagnostic accuracy of an artificial intelligence-powered chatbot in clinical vignettes. *J Neurol Sci.* 2023;453:120804. doi:10.1016/j.jns.2023.120804
24. Faheem F, Zafar Z, Razzak A, Kalia JS. Implementing virtual care in neurology—challenges and pitfalls. *J Cent Nerv Syst Dis.* 2022;14:11795735221109745. doi:10.1177/11795735221109745

25. Latchaw RE, Alberts MJ, Lev MH, et al. Recommendations for imaging of acute ischemic stroke: a scientific statement from the American Heart Association. *Stroke*. 2009;40(11):3646-3678. doi:10.1161/STROKEAHA.108.192616
26. Saunders-Pullman R, Wang C, Stanley K, Bressman SB. Diagnosis and referral delay in women with Parkinson's disease. *Genet Med*. 2011;8(3):209-217. doi:10.1016/j.genm.2011.05.002
27. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med*. 2023;3(1):141-148. doi:10.1038/s43856-023-00370-1
28. Meyer FB, Hoehne SN, Murthy VD, et al. Perception of challenges in management of neurological cases in the emergency room. *J Vet Emerg Crit Care*. 2023;33(1):38-46. doi:10.1111/vec.13258
29. Zhuang S, Xie W, Mao C. Bilateral facial colliculus syndrome caused by pontine tegmentum infarction: a case report. *BMC Neurol*. 2021;21(1):492. doi:10.1186/s12883-021-02524-x

How to cite this article: Lee J-H, Choi E, McDougal R, Lytton WW. GPT-4 performance for neurologic localization. *Neurol Clin Pract*. 2024;14(3):e200293. doi: 10.1212/CPJ.0000000000200293.