Is ChatGPT Better Than Epileptologists at Interpreting Seizure Semiology?

Jun-En Ding, Feng Liu

Submitted to: Journal of Medical Internet Research
on: November 23, 2024

Table of Contents

Is ChatGPT Better Than Epileptologists at Interpreting Seizure Semiology?

Jun-En Ding[1*]; Feng Liu[1*] PhD

[1]Stevens Institute of Technology New Jersey US
[*]these authors contributed equally

Corresponding Author:
Feng Liu PhD
Stevens Institute of Technology
1 Castle Point Terrace, Hoboken
New Jersey
US

Abstract

Background: This study evaluates the clinical utility of ChatGPT in interpreting seizure semiology for epileptogenic zone (EZ) localization in focal epilepsy presurgical assessment. We analyzed two datasets: 852 semiology-EZ pairs from 193 peer-reviewed publications and 184 pairs from Far Eastern Memorial Hospital (FEMH), Taiwan. ChatGPT's performance was tested using zero-shot and few-shot prompting methods, and compared against eight epileptologists' interpretations of 100 randomly selected cases. Performance was measured using regional sensitivity (RSens), weighted sensitivity (WSens), and net positive inference rate (NPIR). Results showed ChatGPT achieved >80% sensitivity for frontal and temporal lobes, ~40% for occipital lobe, 20-30% for parietal lobe, 20% for insular cortex, and 0% for cingulate cortex across both datasets. Compared to epileptologists, ChatGPT demonstrated superior performance in frontal and temporal lobe localization, comparable accuracy in occipital and parietal regions, but underperformed in insular and cingulate cortices. Both ChatGPT and epileptologists showed similar WSens and NPIR values. These findings suggest ChatGPT could serve as a valuable clinical tool in epilepsy presurgical workup, with potential for further improvement as language model technology advances.

Objective: This study aims to evaluate the clinical value of representative large language models (LLMs), namely ChatGPT, on interpreting seizure semiology to localize epileptogenic zones (EZs) for presurgical assessment in patients with focal epilepsy.

Methods: We compiled two data cohorts through public sources and a private database respectively. The data cohort compiled from public sources consists of 852 semiology-EZ pairs derived from 193 peer-reviewed journal publications. The private database includes 184 semiology-EZ pairs collected from the Far Eastern Memorial Hospital (FEMH) in Taiwan. ChatGPT was asked to generate the most likely EZ locations based on the semiology records from both cohorts with two prompting methods: Zero-shot prompting (ZSP) and Few-shot prompting (FSP). To evaluate the ChatGPT's performance compared to epileptologists, a panel of eight epileptologists were recruited for an online survey to provide their interpretations on 100 randomly selected semiology records. The responses from ChatGPT and epileptologists were compared using three metrics: regional sensitivity (RSens), weighted sensitivity (WSens) and net positive inference rate (NPIR).

Results: In the evaluation of interpreting seizure semiology, ChatGPT achieved over 80% sensitivity for the frontal and temporal lobes, approximately 40% for the occipital lobe, 20-30% for the parietal lobe, 20% for the insular cortex, and 0% for the cingulate cortex consistently in both data cohorts. By analyzing the responses from epileptologists, ChatGPT-4 outperformed epileptologists in localizing the frontal and temporal lobes, exhibited similar accuracy for the occipital and parietal lobes, but underperformed in the insular and cingulate cortices. Both ChatGPT and epileptologists demonstrated comparable value for WSens and mean of NPIR.

Conclusions: In this cross-sectional study of seizure semiology interpretation, ChatGPT-generated responses outperformed or matched the responses from epileptologists in regions where EZs are commonly located, including the frontal lobe and the temporal lobe. However, epileptologists provided more accurate responses in regions where EZs are rarely located, such as the insula and the cingulate cortex. Overall, our results demonstrate that ChatGPT might serve as a valuable tool to assist in the preoperative assessment for epilepsy surgery. However, it must be acknowledged that the information provided by ChatGPT may not always be backed by reliable sources, posing a challenge to the verification of ChatGPT-generated responses.
Furthermore, medical professionals, including epileptologists and epilepsy surgeons, must fully recognize the limitations of ChatGPT and exercise caution when utilizing its responses. This study serves as an important reference for employing ChatGPT

in seizure semiology interpretation while underscoring its present constraints.

(JMIR Preprints 23/11/2024:69173)
DOI: https://doi.org/10.2196/preprints.69173

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Title: Is ChatGPT Better Than Epileptologists at Interpreting Seizure Semiology?

Authors: Meng Jiao[1,16], Yaxi Luo[2,16], Neel Fotedar[4], Jun-En Ding[1], Ioannis Karakis[5,6], Vikram R. Rao[7], Melissa Asmar[8], Xiaochen Xian[9], Orwa Aboud[10], Yuxin Wen[11], Jack J. Lin[8], Fang-Ming Hung[12,13], Hai Sun[14], Felix Rosenow[15], Feng Liu[1,3*]

Author Affiliations:

[1]Department of Systems and Enterprises, Schaefer School of Engineering & Science, Stevens Institute of Technology, Hoboken, NJ 07030, United States.

[2]Department of Computer Science, Schaefer School of Engineering & Science, Stevens Institute of Technology, Hoboken, NJ 07030, United States.

[3]Semcer Center for Healthcare Innovation, Stevens Institute of Technology, Hoboken, NJ, 07030, United States

[4]Department of Neurology, University Hospitals Cleveland Medical Center, School of Medicine at Case Western Reserve University, Cleveland, OH, 44106, United States.

[5]Department of Neurology, Emory University School of Medicine GA 30322, United States.

[6]Department of Neurology, University of Crete School of Medicine, Giofirakia 71500, Greece

[7]Department of Neurology and Weill Institute for Neurosciences, University of California San Francisco, San Francisco, CA 94143, United States.

[8]Department of Neurology, University of California Davis, Davis, CA 95616, United States.

[9]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, United States.

[10]Department of Neurology and Neurological Surgery, University of California Davis, Davis, CA 95616, United States.

[11]Fowler School of Engineering, Chapman University, Orange, CA 92866, United States.

[12]Center of Artificial Intelligence, Far Eastern Memorial Hospital, New Taipei City, Taiwan

[13]Surgical Trauma Intensive Care Unit, Far Eastern Memorial Hospital, New Taipei City, Taiwan

[14]Department of Neurosurgery, Rutgers Robert Wood Johnson Medical School, Rutgers University, New Brunswick, NJ 08901, United States.

[15]Goethe-University Frankfurt, Epilepsy Center Frankfurt Rhine-Main, Department of Neurology, Frankfurt am Main, 60590, Germany.

[16]Meng Jiao and Yaxi Luo contributed equally.

[*]**Corresponding Author(s).** Email(s): fliu22@stevens.edu.

Manuscript word count: 3264

Acknowledgments: The authors are grateful to the epileptologists who completed the survey.

Competing interest declaration*:* "All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf."

Dr. Aboud has served on the advisory board for Servier and is supported in part by the UC Davis Paul Calabresi Career Development Award for Clinical Oncology as funded by the National Cancer Institute/National Institutes of Health through grant #2K12CA138464-11. Dr. Rosenow has received research support from the Federal State of Hesse, specifically at the Center for Personalized Translational Epilepsy Research from 2018 to 2022. Dr. Rosenow has received research support from Chaja-Foundation Frankfurt, focusing on establishing and evaluating the ketogenic diet in an institution. Dr. Rosenow received research support from Reiss-Foundation Frankfurt, mainly for the research on the ketogenic diet in GLUT1-DS. Dr. Rosenow received research support from German Ministry of Education, focusing on the ERAPerMed Raise-Genic.

IRB approval: The approval of this secondary data analysis study was exempted by the Institutional Review Board (IRB) at Stevens Institute of Technology under protocol 2024-039 (N).

Abstract

Objective:

This study aims to evaluate the clinical value of representative large language models (LLMs), namely ChatGPT, on interpreting seizure semiology to localize epileptogenic zones (EZs) for presurgical assessment in patients with focal epilepsy.

Method:

We compiled two data cohorts through public sources and a private database respectively. The data cohort compiled from public sources consists of 852 semiology-EZ pairs derived from 193 peer-reviewed journal publications. The private database includes 184 semiology-EZ pairs collected from the Far Eastern Memorial Hospital (FEMH) in Taiwan. ChatGPT was asked to generate the most

likely EZ locations based on the semiology records from both cohorts with two prompting methods: Zero-shot prompting (ZSP) and Few-shot prompting (FSP). To evaluate the ChatGPT's performance compared to epileptologists, a panel of eight epileptologists were recruited for an online survey to provide their interpretations on 100 randomly selected semiology records. The responses from ChatGPT and epileptologists were compared using three metrics: regional sensitivity (RSens), weighted sensitivity (WSens) and net positive inference rate (NPIR).

Results:

In the evaluation of interpreting seizure semiology, ChatGPT achieved over 80% sensitivity for the frontal and temporal lobes, approximately 40% for the occipital lobe, 20-30% for the parietal lobe, 20% for the insular cortex, and 0% for the cingulate cortex consistently in both data cohorts. By analyzing the responses from epileptologists, ChatGPT-4 outperformed epileptologists in localizing the frontal and temporal lobes, exhibited similar accuracy for the occipital and parietal lobes, but underperformed in the insular and cingulate cortices. Both ChatGPT and epileptologists demonstrated comparable value for WSens and mean of NPIR.

Significance:

ChatGPT was shown as a clinically valuable tool to assist the decision making in the epilepsy preoperative workup. With ongoing advancements in LLMs, it is anticipated that the reliability and accuracy of ChatGPT will continue to improve in the future.

Keywords: Epilepsy Semiology, Epileptogenic Zones Localization, Large Language Model

Introduction

Epilepsy is one of the most common neurological diseases affecting more than 70 million people worldwide [1], with approximately 50.4 per 100,000 people developing new-onset epilepsy each year [2, 3]. For patients with drug-resistant focal epilepsy (DRE), surgical resection of the epileptogenic zone (EZ) offers an effective means to control seizure attack. Seizure semiology, which describes signs and symptoms exhibited and experienced by a patient during epileptic seizures [4], yields valuable clues for the localization of the EZs [5]. To achieve optimal post-surgical outcomes, accurately interpreting the seizure semiology plays a crucial role in the presurgical decision making phase.

Recently, large language models (LLMs), especially chatbots, have showcased their capabilities across a wide range of natural language processing (NLP) tasks. As a representative example of LLM, ChatGPT developed by OpenAI [6] holds a dominant position due to its exceptional natural language processing capabilities, achieved by training on enormous amounts of textual information using a combination of supervised learning and reinforcement learning from human feedback. In medical informatics, the descriptive nature of health records and doctor's notes makes LLMs well-suited to assisting clinical prediction and diagnoses. ChatGPT exhibits advanced proficiency in processing and interpreting extensive textual data with input prompt, making it a potent tool for information retrieval, clinical decision support, and medical report generation [7, 8, 9]. A study in February 2023 reported that ChatGPT has successfully passed the United States Medical Licensing Examination (USMLE) [1010], demonstrating its potential as a reliable source of medical information.

The increasing application of ChatGPT in diagnosing various diseases inspired us to utilize it to interpret seizure semiology and localize the EZs, potentially being used as an AI tool for the presurgical evaluation for patients with epilepsy [11, 12]. Therefore, to assess the clinical value of ChatGPT on interpreting seizure semiology, we evaluated ChatGPT's performance by comparing to a

panel of Board-certified epileptologists [13] based on three metrics: regional sensitivity (RSens), weighted sensitivity (WSens) and net positive inference rate (NPIR), using both public and private data cohorts. All the participating epileptologists were employed at different epilepsy centers during the time of survey, with their years of practice ranging from a minimum of 7 to a maximum of 35 years. To obtain the semiology interpretation, we provide the semiology descriptions to both ChatGPT and the epileptologists and both respond with the most likely EZs.

For the LLM model selection, we focused on GPT-4 due to its significantly better performance compared to GPT-3.5. Detailed results from GPT-3.5 have been included in the appendix for reference. By evaluating and comparing the responses from GPT-4 and epileptologists, this study offers an in-depth discussion of the strengths and limitations of decision making rendered by advanced AI tools and suggests directions for future research.

Methods

● Public and Private Data Cohorts

To evaluate the performance of ChatGPT, we compiled a seizure semiology database from publicly available studies published peer-reviewed journals. To avoid potential inclusion of the data records that may have been used for the training of ChatGPT, we also created a separate private data cohort based on electronic health records (EHR) from Far Eastern Memorial Hospital (FEMH) hospital in Taiwan.

To create the data cohort from published cases in peer reviewed journals, we identified 309 publications by searching keywords in PubMed, including “seizures,” “clinical semiology,” and “epilepsy,” etc. [14] The selected publications documented nearly 900 epilepsy cases, providing detailed descriptions of seizure semiology across various surgically validated EZ locations. These EZ locations, regarded as the ground truth, were identified through stereoelectroencephalography (sEEG) findings combined with postoperative outcomes, with seizure-free status determined according to ILAE criteria “Class I: Completely seizure free; no auras” [15] or Engel “Class I: Seizure free or no more than a few early, non-disabling seizures” [16]. All findings were validated by the authors of the respective papers [17]. We labeled the lobes where EZs is located using the LCN-CortLobes classification system (FreeSurferWiki, LCN-CortLobes) [18]. It offers a general anatomical classification for brain regions, grouping them into six general regions: frontal lobe, parietal lobe, temporal lobe, occipital lobe, cingulate cortex, and insular cortex. Each EZ location was mapped to a general region based on this mapping criteria. In cases where multiple EZs were identified, multiple general regions were assigned accordingly.

We then excluded 116 studies that presented uncertain EZs, such as those only specifying hemisphere-level EZs (e.g., right hemispherectomy or left subtotal hemispherectomy). Additionally, 43 cases from the remaining 193 studies were excluded during the semiology extraction process for reasons such as the use of nonspecific terms (e.g., "non-specific aura"), descriptions shorter than two

words, or aggregations of large patient cohorts without providing detailed semiology for individual cases. As a result, the final database contained 852 semiology-EZ pairs from the public peer-reviewed studies.

The private data cohort compiled from FEMH in Taiwan is based on the EHR records from 2017 to 2021. This HER dataset includes the biographies, medical diagnoses, and laboratory test results of 40,749 records. After excluding duplicates and records from unrelated departments, 590 records were identified as relevant to epilepsy. To locate the EZs in the EHR database, we referenced attached laboratory results, such as EEG or MRI findings [19, 20]. The EZs were determined based on EEG or MRI results, while semiology was extracted from notes documented by physicians. The methodology for compiling both data cohorts is illustrated in **Fig. 1** according to the PRISMA guidelines.

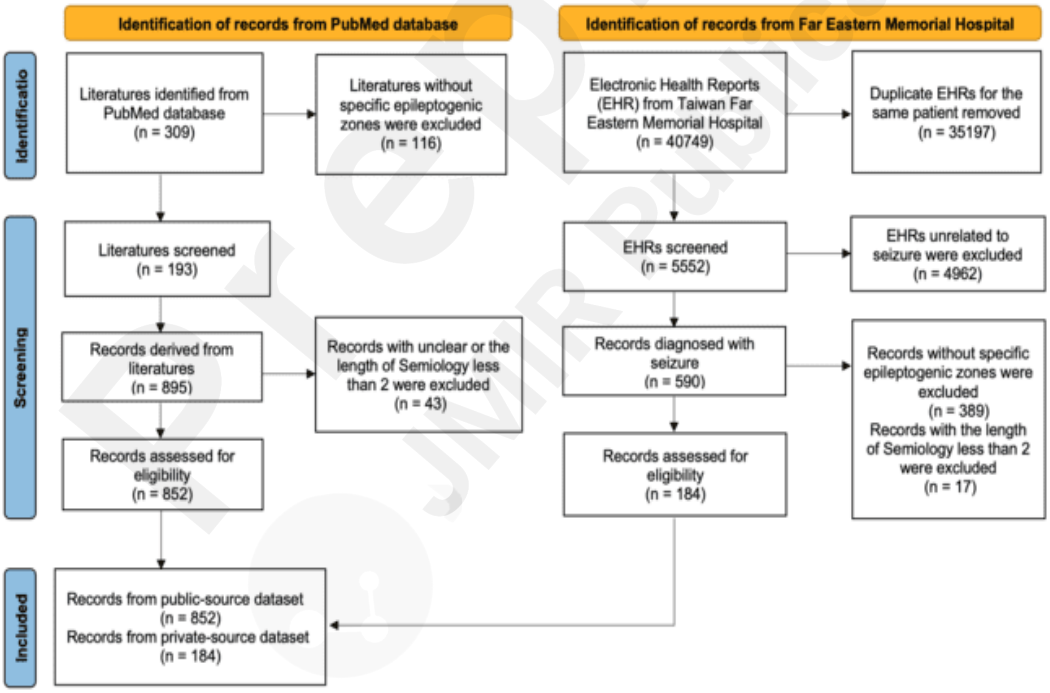

Figure 1. PRISMA Flow Diagram of Database Construction.

- **Demographics of patients**

The public-source data cohort comprised 852 semiology-EZ pairs from 852 patients. The demographic distribution included 320 females, 404 males, and 128 patients with undisclosed gender. Of all 852 patients, 134 were right-handed, 22 left-handed, 3 ambidextrous, and 706 had

unspecified handedness. The age range of the patients spanned from newborn to 77 years, with 310 individuals under the age of 18 and 335 adults.

The FEMH data cohort consisted of 184 semiology-EZ pairs. This group included 44 female and 46 male patients, with an age range from newborn to 87 years. Among them, 37 were under 18 years old, and 50 were adults.

- **Response Generation with ChatGPT**

In this study, we utilized ChatGPT-4.0 through the "gpt-4-turbo" APIs, incorporating two distinct prompt configurations: zero-shot prompting (ZSP) [21] and few-shot prompting (FSP) [22], to assess their impact on ChatGPT's performance. In the ZSP configuration, ChatGPT received no prior information, whereas in the FSP configuration, the input included three semiology-EZ pair examples to guide the responses more closely to the ground truth. To streamline the comparison of responses, we defined a specific output format for ChatGPT, limiting it to the EZ location. Specific query examples are provided in the Appendix (Table 1).

- **Response Collection from Epileptologists**

To compare ChatGPT's seizure semiology interpretation with that of epileptologists, a panel of eight epileptologists, each with an average of 10 years of experience in treating epilepsy patients, was invited to participate in an online survey (https://survey.zohopublic.com/zs/NECl0I). In this survey, we randomly selected a subset of 100 semiology records from our self-compiled database, covering all six general brain regions, and asked the epileptologists to identify the most likely EZ location. The final subset used in the survey met the following criteria: (1) selected semiology records contained comprehensive and explicit descriptions of seizure symptoms; (2) the distribution of EZs spanned all six general regions, rather than focusing on one region; and (3) the records were chosen to capture the widest possible range of seizure symptoms.

We invited doctors specializing in epilepsy through the National Association Epilepsy Center and the American Epilepsy Society, sending over 70 survey invitations globally. Ultimately, 5

epileptologists completed the survey in full, while the remaining participants completed it partially. Responses were collected from January 2024 to July 2024.

- **Statistical Analysis**

The inference of EZ location is determined using the six-lobe classification criteria. To evaluate the responses from ChatGPT and epileptologists, we used three metrics: Regional Sensitivity (*RSens*), Weighted Sensitivity (*WSens*) and Net Positive Inference Rate (*NPIR*) [23].

Specifically, *RSens* measured the accuracy of ChatGPT or epileptologists in identifying the correct region and is defined as follows:

$$RSens_i = \frac{TP_i}{TP_i + FN_i}$$

where *i* refers to the index corresponding to six general regions, $RSens_i$ represents the sensitivity value for region *i*, TP_i (True Positive) is the number of correctly identified EZs, and FN_i (False Negative) is the number of EZs that were not correctly identified.

Additionally, given the unbalanced distribution of EZs across the six general regions, we addressed the class imbalance issue by using *WSens* to provide a more accurate performance assessment. *WSens* evaluates overall accuracy by considering the RSens of each region and its corresponding weight in the dataset, which is calculated as follows:

$$WSens = \frac{1}{N}\sum_{i=1}^{k}(N_i \times RSens_i)$$

where k represents the total number of regions, *i* is the index corresponding to each region; N is the total number of regions in the dataset, and N_i is the count of instances for the *i*-th region.

For each semiology-EZ pair, the inferred EZ locations based on epileptic seizure semiology consist of two parts: regions containing true EZs and regions excluding EZs. To assess the reliability of these inferences, we introduced the *NPIR*, which is based on *RSens*. This metric treats correctly

inferred regions as positive inferences, while incorrectly inferred regions incur a penalty. The *NPIR* for an individual response, either from ChatGPT or epileptologists, is calculated as follows:

$$NPIR = \frac{TP - FP}{TP + FN}$$

where *TP* (True Positive) is the number of correctly identified regions, *FP* (False Positive) is the number of regions incorrectly identified, and *FN* (False Negative) is the number of regions that were part of the ground truth but not identified.

The *NPIR* value reflects the reliability of the inference results. An *NPIR* of 1 indicates a completely correct inference of the EZ location. A value below 1 suggests the inference is partially incorrect or contains omissions. An *NPIR* below 0 indicates that the inference is unreliable and could mislead physicians during preoperative assessments for epilepsy surgery.

Results

● Evaluation of Responses from ChatGPT on Public-Source Cohort

In this section, we evaluated the performance of ChatGPT-4 using ZSP (abbreviated as GPT-4 ZSP), and ChatGPT-4 using FSP (GPT-4 FSP) in interpreting seizure semiology based on the public-source cohort. The evaluation results for ChatGPT-4 are presented in **Fig 2**, while those for ChatGPT-3.5 are provided in **Appendix**.

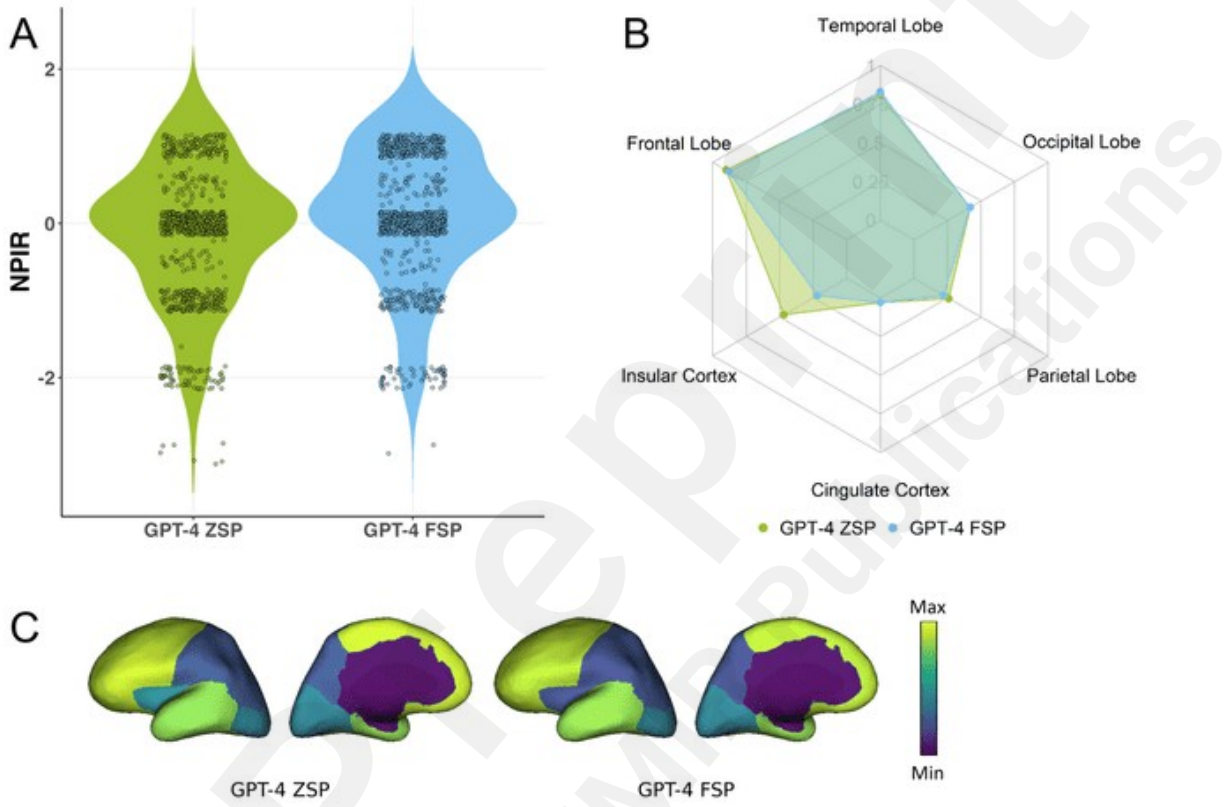

Figure 2: (A) Net Positive Inference Rate Distribution generated by ChatGPT-4 with zero-shot prompting (GPT-4 ZSP) and few-shot prompting (GPT-4 FSP). (B) and (C), regional sensitivity generated by ChatGPT with different prompt configurations (GPT-4 ZSP, GPT-4 FSP).

As shown in **Fig. 2A**, both prompting methods of ChatGPT-4 demonstrated similar performance in interpreting seizure semiology across different prompt configurations. GPT-4 ZSP achieved a mean *NPIR* of -0.21 and *WSens* of 0.69, whereas GPT-4 FSP showed a slightly higher *NPIR* level with a mean of 0.03 but a relatively lower *WSens* of 0.67. Additionally, the *RSens* values for each region were calculated and presented in **Figure 2B-C**. GPT-4 ZSP achieved *RSens* values of 0.9 for the frontal lobe, 0.81 for the temporal lobe, 0.42 for the occipital lobe, 0.26 for the parietal

lobe, 0.47 for the insular cortex, and 0.03 for the cingulate cortex. Similarly, GPT-4 FSP achieved *RSens* values of 0.88 for the frontal lobe, 0.83 for the temporal lobe, 0.42 for the occipital lobe, 0.22 for the parietal lobe, 0.22 for the insular cortex, and 0.03 for the cingulate cortex.

These results highlight ChatGPT's proficiency in reliably interpreting seizure semiology related to frontal and temporal lobe epilepsies, while also revealing its limitations in accurately interpreting semiology associated with the parietal, occipital, cingulate, and insular regions, which are relatively less common.

- **Evaluation of Responses from ChatGPT with Private-Source Cohort**

Given that all papers used to compile the public-source database are available online, some may have been included in the training corpus of ChatGPT, potentially making the evaluation results less objective and convincing. To address this concern, we employed a database with a private source for external validation of ChatGPT's performance. The evaluation results for this database are presented in **Fig. 3**.

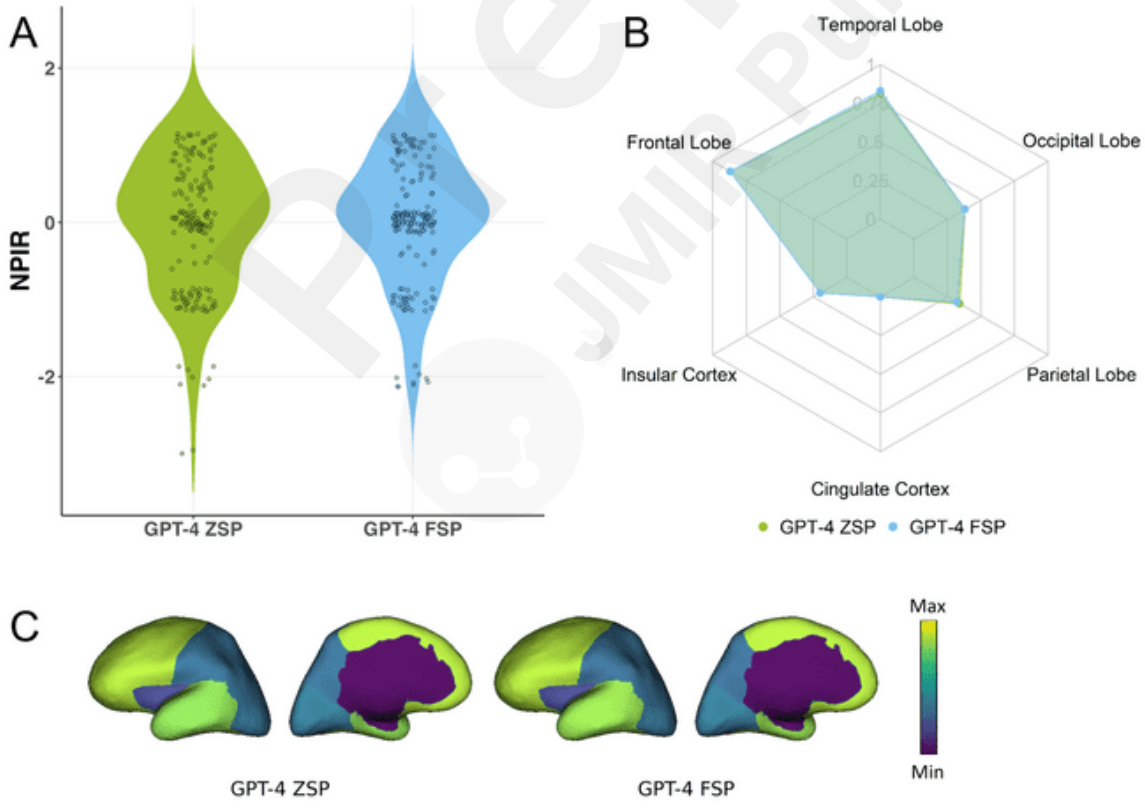

Figure 3: (A) Net Positive Inference Rate Distribution generated by ChatGPT-4 with zero-shot prompting (GPT-4 ZSP) and few-shot prompting (GPT-4 FSP). (B) and (C), regional sensitivity generated by ChatGPT with different

prompt configurations (GPT-4 ZSP, GPT-4 FSP).

As shown in **Figure 3A**, GPT-4 ZSP achieved a mean *NPIR* of -0.2 and *WSens* of 0.73. GPT-4 FSP achieved a mean *NPIR* of -0.12 and *WSens* of 0.74. Additionally, the *RSens* values for each region were calculated and presented in **Fig. 3B-C**. GPT-4 ZSP achieved *RSens* values of 0.87 for the frontal lobe, 0.81 for the temporal lobe, 0.38 for the occipital lobe, 0.34 for the parietal lobe, 0.2 for the insular cortex, and 0 for the cingulate cortex. Similarly, GPT-4 FSP achieved *RSens* values of 0.87 for the frontal lobe, 0.83 for the temporal lobe, 0.38 for the occipital lobe, 0.32 for the parietal lobe, 0.2 for the insular cortex, and 0 for the cingulate cortex.

These evaluation results based on the private-source cohort are consistent with those from the public-source cohort, further confirming the variation in ChatGPT's performance when interpreting seizure semiology across different brain regions.

- **Comparison of Responses from ChatGPT and Epileptologists**

In this section, we compared the performance of ChatGPT with that of epileptologists. Eight board-certified epileptologists participated in the online survey comprising 100 randomly selected questions regarding all EZ locations and detailed seizure semiologies and six of them completed it. Consequently, the analysis focused on the fully completed responses from five epileptologists (EP-1, EP-2, EP-3, EP-4, EP-5). The comparison results are shown in **Fig. 4**.

As illustrated in **Fig. 4A**, ChatGPT demonstrated comparable or, in some regions, superior performance compared to the epileptologists in interpreting seizure semiology. GPT-4 ZSP achieved a mean *NPIR* of -0.14 and *WSens* of 0.61. GPT-4 FSP achieved a mean *NPIR* of -0.02 and *WSens* of 0.63. Unlike ChatGPT's consistent metrics value in two prompting methods, the performance of epileptologists showed significant variation. EP-5 achieved the highest performance with a mean *NPIR* of -0.08 and *WSens* of 0.51, while EP-2 had the lowest, with a mean *NPIR* of -0.13 and *WSens* of 0.49.

When comparing *RSens* values across regions, ChatGPT-4 *outperformed* epileptologists in

interpreting seizure semiology for the frontal (ZSP: 0.73, FSP: 0.76, EPs: 0.57–0.73) and temporal lobes (ZSP: 0.76, FSP: 0.93, EPs: 0.44–0.61). The model also demonstrated comparable performance in the parietal (ZSP: 0.39, FSP: 0.32, EPs: 0.29–0.57) and occipital lobes (ZSP: 0.63, FSP: 0.63, EPs: 0.58–0.79). However, epileptologists *outperformed* ChatGPT in interpreting seizure semiology associated with the cingulate (0–0.5) and insular cortex (0.44–0.67), compared to ChatGPT's performance for the cingulate (ZSP/FSP: 0.12) and insular cortex (ZSP: 0.56, FSP: 0.22) (**Fig. 4B-C**).

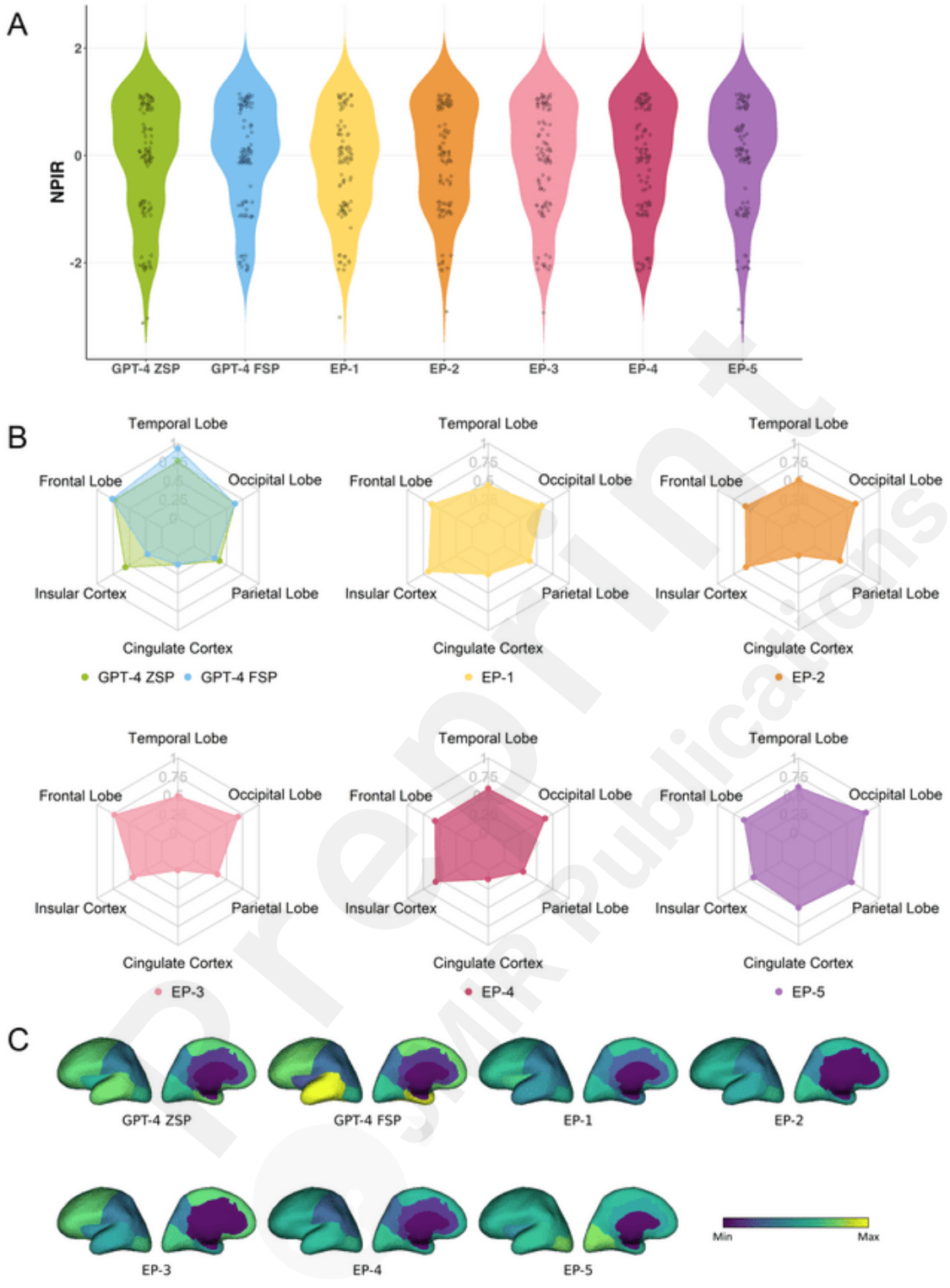

Figure 4: (A) Net Positive Inference Rate Distribution generated by ChatGPT-4 with zero-shot prompting (GPT-4 ZSP) and few-shot prompting (GPT-4 FSP). (B) and (C), regional sensitivity generated by ChatGPT-4 with different prompt configurations (GPT-4 ZSP, GPT-4 FSP).

Discussion

We evaluated the capability of ChatGPT-4 on interpreting seizure semiology to localize the epileptogenic zone (EZ) using both a public-source cohort with 852 EZ-semiology pairs and a private-source cohort with 184 pairs. The evaluation was conducted using three metrics: *RSens*, *WSens*, and *NPIR*. These same metrics were applied to the responses from a 100-question survey completed by epileptologists. The results from all three sets of responses—ChatGPT on the public-source and private-source databases, as well as the epileptologist survey—were analyzed and compared using these metrics.

For the analysis of *RSens*, it revealed that ChatGPT-4 significantly outperformed epileptologists in the frontal and temporal lobes. However, the influence of different prompting techniques on the responses from both versions was minimal. Notably, ChatGPT-4 demonstrated comparable, and in some cases superior performance to that of epileptologists, particularly in identifying EZ locations that are more commonly found. The responses from ChatGPT-4 were comprehensive and surprisingly valuable for reference, with well-founded reasoning regarding EZ locations (**See Appendix: Table 1**). Nonetheless, our *RSens* analysis highlighted that for seizure semiology indicating less common EZ locations, such as the cingulate and insular cortex, interpretations and localization from epileptologists remained more precise and reliable. The discrepancy in epilepsy manifestations in less common regions can be attributed to an insufficient data volume, which hampers the training of language models to achieve predictive reliability comparable to that of epileptologists.

Our results align with findings from previous studies assessing ChatGPT's performance in epilepsy-related inquiries [24, 25]. Specifically, Kim et al. assessed the reliability of responses of ChatGPT to 57 commonly asked epilepsy questions, and all responses were reviewed by two epileptologists. The results suggested that these responses were either of “sufficient educational value” or “correct but inadequate” for almost all questions [24]. Wu et al. evaluated the performance

of ChatGPT to a total of 378 questions related to epilepsy and 5 questions related to emotional support. Statistics indicated that ChatGPT provided “correct and comprehensive” answers to 68.4% of the questions. However, when answering “prognostic questions”, ChatGPT performed poorly with only 46.8% of answers rated comprehensive [25].

Limitations

Although this study offered an important reference on the capability of ChatGPT to interpret the descriptions of seizure semiology to localize EZs, there are still several limitations. First, the number of articles involved in compiling the semiology-EZ database can be further expanded and updated. These semiology records were mainly related to seizures originating from the frontal lobe and temporal lobe, while for the remaining regions, especially the cingulate cortex and insular cortex, the collected semiology descriptions were inadequate and incomplete. Future studies could include more semiology corresponding to regions where EZs are rarely found. Second, when inferring the EZ location according to semiology, the identified area is referred to as the Symptomatogenic Zone (SZ), which is the region responsible for the observed seizure symptoms but may not fully align with the actual EZ which is a theoretical definition given by Dr. Hans Luders [20] and whose removal will make the patients seizure free. This will result in limited precision of predicting EZs. Additionally, epileptic seizures often involve abnormal activities across multiple brain regions, with certain symptoms arising from the propagation of activity to regions beyond the EZ, which may lead to potential misjudgments. Third, the limited number of participating epileptologists inherently restricts the sample size for comparative analysis, making it more challenging to detect true differences between ChatGPT and human responses. Furthermore, insights from a small group of epileptologists may not adequately reflect the broader expertise and perspectives of the global specialist community, thereby limiting the generalizability of the findings about interpreting seizure semiology to localize EZs. Moreover, ChatGPT was trained on the Common Crawl corpus, which encompasses a wide array of general knowledge from books, articles, web pages, etc., limiting the ability of ChatGPT to

generate responses with a specific focus on the medical domain. Future research could explore the application of LLMs fine-tuned on epilepsy-specific corpora for improved seizure semiology interpretation and EZ localization.

Conclusions

In this cross-sectional study of seizure semiology interpretation, ChatGPT-generated responses outperformed or matched the responses from epileptologists in regions where EZs are commonly located, including the frontal lobe and the temporal lobe. However, epileptologists provided more accurate responses in regions where EZs are rarely located, such as the insula and the cingulate cortex. Overall, our results demonstrate that ChatGPT might serve as a valuable tool to assist in the preoperative assessment for epilepsy surgery. However, it must be acknowledged that the information provided by ChatGPT may not always be backed by reliable sources, posing a challenge to the verification of ChatGPT-generated responses.

Furthermore, medical professionals, including epileptologists and epilepsy surgeons, must fully recognize the limitations of ChatGPT and exercise caution when utilizing its responses. This study serves as an important reference for employing ChatGPT in seizure semiology interpretation while underscoring its present constraints.

Acknowledgments: The authors are grateful to the epileptologists who completed the survey. Research reported in this publication was partially supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health, United States under Award Number R21NS135482 (PI: Liu). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions:

Jiao (data curation, formal analysis, software, Visualization , Writing - Original Draft), Luo (data curation, formal analysis, software, Visualization ,Writing - Original Draft), Fotedar (Conceptualization, Methodology, Validation, Investigation, Writing - Review & Editing), Karakis (Validation, Investigation, Writing - Review & Editing), Aboud (Validation, Investigation, Writing - Review & Editing), Rao (Validation, Investigation, Writing - Review & Editing), Asmar (Validation, Investigation, Writing - Review & Editing), Xian (Statistical Analysis, Validation, Writing - Review & Editing), Wen(Statistical Analysis, Validation, Writing - Review & Editing), Ding (Validation, Writing - Review & Editing), Lin (Resources, Validation, Writing - Review & Editing), Rosenow (Validation, Investigation, Writing - Review & Editing), Sun (Conceptualization, Methodology, Validation, Investigation, Writing - Review & Editing), Liu (Conceptualization, Supervision, Formal analysis, Funding acquisition, Resources, Methodology, Writing - Original Draft)

Competing interest declaration*:* "All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf."

Dr. Aboud has served on the advisory board for Servier and is supported in part by the UC Davis Paul Calabresi Career Development Award for Clinical Oncology as funded by the National Cancer Institute/National Institutes of Health through grant #2K12CA138464-11. Dr. Rosenow has received research support from the Federal State of Hesse, specifically at the Center for Personalized Translational Epilepsy Research from 2018 to 2022. Dr. Rosenow has received research support from Chaja-Foundation Frankfurt, focusing on establishing and evaluating the ketogenic diet in

institution. Dr. Rosenow received research support from Reiss-Foundation Frankfurt, mainly for the research on the ketogenic diet in GLUT1-DS. Dr. Rosenow received research support from German Ministry of Education, focusing on the ERAPerMed Raise-Genic.

IRB approval: The approval of this secondary data analysis study was exempted by the Institutional Review Board (IRB) at Stevens Institute of Technology under protocol 2024-039 (N).

References

1. Anuradha Singh and Stephen Trevick. The epidemiology of global epilepsy. *Neurologic clinics*, 34(4):837–847, 2016.
2. Jay R Gavvala and Stephan U Schuele. New-onset seizure in adults and adolescents: a review. *JAMA*, 316(24):2657–2668, 2016.
3. Zhibin Chen, Martin J Brodie, Danny Liew, and Patrick Kwan. Treatment outcomes in patients with newly diagnosed epilepsy treated with established and new antiepileptic drugs: a 30-year longitudinal cohort study. *JAMA neurology*, 75(3):279–286, 2018.
4. Mohammed MS Jan and John P Girvin. Seizure semiology: value in identifying seizure origin. *Canadian Journal of Neurological Sciences*, 35(1):22–30, 2008.
5. Krikor Tufenkjian and Hans O Lüders. Seizure semiology: its value and limitations in localizing the epileptogenic zone. *Journal of Clinical Neurology*, 8(4):243–250, 2012.
6. Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/ CAA Journal of Automatica Sinica*, 10(5):1122–1136, 2023.
7. Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, 2023.
8. Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068, 2024.
9. Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Kn¨odler, Ulrich Keller, Dieter Beule, et al. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*, 6(11):e2343689–e2343689, 2023.

10. Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepa ño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198, 2023.

11. Daiju Ueda, Yasuhito Mitsuyama, Hirotaka Takita, Daisuke Horiuchi, Shannon L Walston, Hiroyuki Tatekawa, and Yukio Miki. Diagnostic performance of ChatGPT from patient history and imaging findings on the diagnosis please quizzes. *Radiology*, 308(1):e231040, 2023.

12. Lars Mehnen, Stefanie Gruarin, Mina Vasileva, and Bernhard Knapp. ChatGPT as a medical doctor? A diagnostic accuracy study on common and rare diseases. *medRxiv*, pages 2023–04, 2023.

13. Habib S, Butt H, Goldenholz SR, Chang CY, Goldenholz DM. Large Language Model Performance on Practice Epilepsy Board Examinations. *JAMA Neurol.* 2024;81(6):660–661. doi:10.1001/jamaneurol.2024.0676

14. Kathi Canese and Sarah Weis. PubMed: The bibliographic database. *The NCBI handbook*, 2(1), 2013.

15. Wieser HG, Blume WT, Fish D, Goldensohn E, Hufnagel A, King D, Sperling MR, Lüders H, Pedley TA; Commission on Neurosurgery of the International League Against Epilepsy (ILAE). ILAE Commission Report. Proposal for a new classification of outcome with respect to epileptic seizures following epilepsy surgery. Epilepsia. 2001 Feb;42(2):282-6. PMID: 11240604.

16. Engel J Jr, Levesque MF, Shields WD. Surgical treatment of the epilepsies: presurgical evaluation. Clin Neurosurg. 1992;38:514-34. PMID: 1537201.

17. Ali Alim-Marvasti, Gloria Romagnoli, Karan Dahele, Hadi Modarres, Fernando P´erez-Garc ´ıa, Rachel Sparks, S ´ebastien Ourselin, Matthew J Clarkson, Fahmida Chowdhury, Beate

Diehl, et al. Probabilistic landscape of seizure semiology localizing values. *Brain Communications*, 4(3):fcac130, 2022.

18. Arno Klein and Jason Tourville. 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in neuroscience*, 6:33392, 2012.
19. Jehi, Lara. "The epileptogenic zone: concept and definition." Epilepsy currents 18.1 (2018): 12-16.
20. Lüders, Hans O., Imad Najm, Dileep Nair, Peter Widdess-Walsh, and William Bingman. "The epileptogenic zone: general principles." Epileptic disorders 8 (2006): S1-S9.
21. Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
22. Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of GPT-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022.
23. Hossin M, Sulaiman MN. A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process. 2015 Mar 1;5(2):1.
24. Hyun-Woo Kim, Dong-Hyeon Shin, Jiyoung Kim, Gha-Hyun Lee, and Jae Wook Cho. Assessing the performance of ChatGPT’s responses to questions related to epilepsy: A cross-sectional study on natural language processing and medical information retrieval. *Seizure: European Journal of Epilepsy*, 114:1–8, 2024.
25. YuXin Wu, Zaiyu Zhang, Xinyu Dong, Siqi Hong, Yue Hu, Ping Liang, Lusheng Li, Bin Zou, Xuanxuan Wu, Difei Wang, et al. Evaluating the performance of the language model ChatGPT in responding to common questions of people with epilepsy. *Epilepsy & Behavior*, 151:109645, 2024.

Appendix

Table 1. Examples of Questions and Responses Generated by ChatGPT with Different Prompt Configurations.

	Zero-Shot Promoting	Few-Shot Promoting
User Input	Semiology	Semiology
Prompt	None	Case 1: - "Semiology: aura, loss of contact, motor." - "EZ: Occipital lobe." Case 2: …… Case 3: ……
Query	"Based on the user's semiology, list the most likely epileptogenic zones (EZ) in descending order of likelihood. The epileptogenic zones include frontal lobe, temporal lobe, parietal lobe, occipital lobe, cingulate, insular cortex. Provide the answer in this format: 'EZ1, EZ2, ...'. If there is only one likely EZ, list only that one. Do not include any explanations."	"Based on the user's semiology, list the most likely epileptogenic zones (EZ) in descending order of likelihood. The epileptogenic zones include frontal lobe, temporal lobe, parietal lobe, occipital lobe, cingulate, insular cortex. Provide the answer in this format: 'EZ1, EZ2, ...'. If there is only one likely EZ, list only that one. Do not include any explanations."
Response	EZ1, EZ2, ...	EZ1, EZ2, ...

Figure 1: (A) Net Positive Inference Rate (NPIR) distribution for GPT-3.5 with zero-shot prompting (ZSP) and few-shot prompting (FSP) on the public-source database. (B) Regional sensitivity (RSens) for GPT-3.5 ZSP and FSP on the public-source database. (C) NPIR distribution on the private-source database. (D) RSens on the private-source database. (E) NPIR distribution on the 100-question survey. (F) RSens on the 100-question survey.