Original Investigation

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

Authors: Meng Jiao, MS, School of Systems and Enterprises, Stevens Institute of Technology; **Yaxi Luo**, BS, Department of Computer Science, Schaefer School of Engineering & Science, Stevens Institute of Technology; **Neel Fotedar**, MD, Department of Neurology, University Hospitals Cleveland Medical Center, School of Medicine at Case Western Reserve University; **Ioannis Karakis**, MD, PhD, Department of Neurology, Emory University School of Medicine and Department of Neurology, University of Crete School of Medicine; **Vikram R. Rao**, MD, PhD, Department of Neurology and Weill Institute for Neurosciences, University of California San Francisco; **Melissa Asmar**, MD, Department of Neurology, University of California Davis; **Xiaochen Xian**, PhD, Department of Industrial & Systems Engineering, University of Florida; **Orwa Aboud**, MD, PhD, Department of Neurology and Neurological Surgery, University of California Davis; **Yuxin Wen**, PhD, Fowler School of Engineering, Chapman University; **Jack J. Lin**, MD, Department of Neurology, University of California Davis; **Felix Rosenow**, MD, MHBA, Goethe-University Frankfurt, Epilepsy Center Frankfurt Rhine-Main, Department of Neurology; **Hai Sun**, MD, PhD, Department of Neurosurgery, Rutgers Robert Wood Johnson Medical School, Rutgers University; **Feng Liu**, PhD, School of Systems and Enterprises, Stevens Institute of Technology.

Corresponding Author: Feng Liu, PhD, School of Systems and Enterprises, Stevens Institute of Technology, 07030 Hoboken, United States. Tel: 201-216-8009, Fax: 201-216-5541, Email: fliu22@stevens.edu.

Word count: [3000 words]

Abstract

Background: Utilizing large language models (LLMs), primarily ChatGPT, to interpret the seizure semiology with focal epilepsy could yield valuable data for presurgical assessment. Assessing the reliability and comparability of LLM-generated responses with those from well-trained neurologists, especially epileptologists, is crucial for ascertaining the value of LLMs in the presurgical evaluation.

Methods: A total of 865 descriptions of seizure semiology and validated epileptogenic zone (EZ) pairs were derived from 189 public papers. These semiology records were utilized as input of ChatGPT to generate responses on the most likely locations of EZ. Additionally, a panel of 5 epileptologists was recruited to complete an online survey by providing responses on EZ locations based on 100 well-defined semiology records. All responses from ChatGPT and epileptologists were graded for their reliability score (RS) and regional accuracy rate (RAR).

Results: In evaluating responses to semiology queries, the highest RARs in each general region from ChatGPT-4.0 were 89.28% for the frontal lobe and 71.39% for the temporal lobe. However, the RAR was lower for the occipital lobe at 46.24%, the parietal lobe at 31.01%, the insular cortex at 8.51%, and the cingulate cortex at 2.78%. Comparatively, the RAR achieved by epileptologists was 82.76% for the frontal lobe, 58.33% for the temporal lobe, 68.42% for the occipital lobe, 50% for the parietal lobe, 60% for the insular cortex, and 28.57% for the cingulate cortex.

Conclusions: In this study of seizure semiology interpretation, ChatGPT-4.0 outperformed epileptologists in interpreting seizure semiology originating in the frontal and temporal lobes, whereas epileptologists outperformed ChatGPT-4.0 in the occipital and parietal lobes, and significantly outperformed in the insular cortex and cingulate cortex. ChatGPT demonstrates the potential to assist in the preoperative

assessment for epilepsy surgery. Presumably, with the continuous development of LLM, the reliability of ChatGPT will be strengthened in the foreseeable future.

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 those with drug-resistant focal epilepsy (DRE), surgical resection of the epileptogenic zone (EZ) may be curative. Seizure semiology, which refers to the description of signs and symptoms exhibited and experienced by a patient during epileptic seizures [4], yields valuable clues for the localization of the EZ [5]. In order to achieve optimal surgical outcomes, accurately interpreting the seizure semiology plays a crucial role in the preoperative assessment process.

Recently, large language models (LLMs), especially chatbots, have showcased their capabilities across a wide range of natural language processing (NLP) tasks. As a specific instance of LLM, ChatGPT developed by OpenAI [6] posited a dominant position in the current market for chatbot applications. ChatGPT, trained on a diverse database of text across different subjects, can resemble an expert equipped with a broad range of knowledge. In the field of medical informatics, ChatGPT exhibits an advanced proficiency in processing and interpreting extensive textual data upon any 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 achieved success in passing the United States Medical Licensing Examination (USMLE) [10], demonstrating its potential as a reliable source of medical information.

The increasing application of ChatGPT in diagnosing various diseases inspired the idea of utilizing it to interpret seizure semiology, which may pioneer a novel tool for the surgical treatment of epilepsy [11,12]. However, since several previous studies have raised various concerns about the responses from ChatGPT may include incorrect or limited information [13,14], we, therefore, assess the reliability and accuracy of ChatGPT in seizure semiology interpretation by comparing its performance against that of a panel of Board-certified epileptologists. The descriptive nature of seizure semiology makes LLMs a natural fit for interpreting the associated clinical diagnosis. The semiotic texts were used to generate queries regarding the most likely locations of the EZs. All responses from ChatGPT and epileptologists underwent thorough review, with the addressed EZ locations being systematically cataloged and summarized. By evaluating and comparing responses from ChatGPT and epileptologists, this study provides an in-depth discussion of the strengths and limitations of AI-generated medical information and proposes directions for future research.

Methods

Database Source

We compiled an epilepsy-specific database by actively utilizing PubMed, a free search engine primarily accessing the MEDLINE database, to source 189 articles by searching for “seizures”, “clinical semiology”, and “epilepsy”, etc [15]. These selected articles documented over 900 epilepsy cases, detailing seizure semiology across various surgically validated EZ locations. These EZ locations (i.e., ground truth) were labeled based on postoperative outcomes (defined as seizure freedom after surgery), concordance of imaging and neurophysiology, or available

stereoelectroencephalography (sEEG) findings [16]. Cases with indeterminate EZs or nonspecific semiology were excluded. For example, cases indicating only hemisphere-level EZs, like right hemispherectomy or left subtotal hemispherectomy, without general or specific regional details, were omitted. Additionally, cases described with nonspecific terms, like 'non-specific aura' or those aggregating a large patient cohort without providing detailed semiology for each individual, were excluded. The methodology for the data compilation is depicted in **Figure 1**. In addition, a thorough collection of patient data was compiled, encompassing demographics, seizure semiology, imaging and/or sEEG findings (if provided), and surgical results.

The resultant database comprised 865 patient-derived semiology-EZ pairs, with a demographic spread of 134 right-handed, 22 left-handed, 3 ambidextrous, and 706 unspecified handedness individuals, aged from newborn to 77 years. These semiology-EZ pairs indicate either single or multiple EZ locations. In documenting these locations, the majority of articles merely delineate the general region without providing detailed information on the specific regions of EZs for individual epilepsy cases. Consequently, there were more data on general regions than on specific regions. Besides, these articles present a straightforward anatomic classification for general regions: frontal lobe, parietal lobe, temporal lobe, occipital lobe, cingulate cortex, and insular cortex.

Nevertheless, the delineation of specific regions within these six broader areas presents complexities. First, despite the orthographic differences in EZ location terminology, such terms often refer to the same brain region. For example, 'medial temporal' and 'mesial temporal' both denote the same specific area within the temporal lobe. Second, specific regions within an individual general region cannot be readily

identifiable by their name alone. For example, the frontal lobe includes regions such as the Precentral and Paracentral gyri; and within the parietal lobe lie the Supramarginal, Postcentral, and Precuneus.

In summary, to address the issues of inconsistent terminology and the ambiguity of region names that do not directly indicate their locations in descriptions of specific EZ (Epileptogenic Zone) regions across various studies, we have referenced a classification system from a website (FreeSurferWiki, LCN-CortLobes). This system organizes the data into six broad regions and over twenty uniquely identifiable specific regions, as detailed in reference [17].

- **Response Generation with ChatGPT**

In this study, all 865 semiology records in our self-compiled database were leveraged to investigate the capabilities of ChatGPT in accurately identifying the most probable location of the EZ using seizure semiology as inputs. The evaluation encompassed two versions of ChatGPT: ChatGPT-3.5 and ChatGPT-4.0. Additionally, two distinct prompt configurations: zero-shot prompting (ZSP) [18] and few-shot prompting (FSP) [19] were incorporated to examine their influence on the performance of ChatGPT. In the ZSP configuration, ChatGPT received no preliminary information; in the FSP configuration, the input consisted of three semiology-EZ pair examples intended to steer the responses to align with the ground truth (**Table 1**) more accurately.

Given that consecutive interactions within the same chat session might influence the responses of ChatGPT due to its stateful nature, we adopted a stringent methodology where each query was entered in a separate “New Chat” session to mitigate any potential bias or interference. The query section was conducted over a period extending from December 2023 to January 2024.

- **Response Collection from Epileptologists**

To conduct the comparison between ChatGPT and epileptologists in seizure semiology interpretation, a panel of eight epileptologists with an average of 10 years of experience in treating epilepsy patients were recruited to complete an online survey on semiology interpretation (https://survey.zohopublic.com/zs/NECl0I). In this survey, we selected a representative subset of 100 semiology records that spanned all six general regions from our previously mentioned self-compiled database to seek the epileptologists' opinions on the most likely EZ. The selection of semiology records fulfilled the following criteria that have been reviewed by epileptologists: (1) selected semiology records provided comprehensive and explicit descriptions of seizure symptoms; (2) the distribution of EZs corresponding to selected semiology records involved all six general regions, rather than being concentrated in a particular region; (3) the selection of semiology records was intended to capture the broadest possible range of seizure symptoms.

All responses for the survey were collected during the period from January 2024 to February 2024. We found doctors specializing in epilepsy from the National Association Epilepsy Center and American Epilepsy Society and sent out more than 70 survey invitations across the world. Consequently, five epileptologists completed the survey in full and three others completed it partially.

- **Statistical Analysis**

To facilitate a precise and logical assessment of the responses provided by ChatGPT and epileptologists, two bespoke statistical metrics were introduced: the Reliability Score (RS) and the Regional Accuracy Rate (RAR).

First, the RS quantifies the accuracy of responses, where a score of 1 indicates 100% correct identification of EZ locations, aligning perfectly with ground truth. Scores below 1 suggest partial accuracy by ChatGPT or epileptologists, while those

under 0 may represent misleading information, potentially complicating preoperative evaluations by neurologists. The RS is calculated as:

$$RS = \frac{\alpha N_c + \beta N_i}{N_T} \tag{1}$$

where N_T denotes the total count of EZs identified in the ground truth for a given semiology case. N_c and N_i respectively represent the counts of correct and incorrect identifications made by ChatGPT or epileptologists, relative to the ground truth of each epilepsy case. A correct identification (N_c) positively contributes with a weight of $\alpha = 1$, while an incorrect one (N_i) carries a negative weight of $\beta = -0.5$ to account for its potential to mislead epileptologists in determining the correct EZ location.

Second, the RAR is a region-specific metric designed to evaluate the precision of EZ location identification for individual regions. It compares the identification of the epileptogenic general region from ChatGPT or epileptologists against the ground truth. The RAR values range from 0% to 100%. When the response from the ChatGPT or epileptologist perfectly matches the ground truth, the RAR is 100%; otherwise, less than 100%.

For a general region x - which is one of the following regions: frontal lobe, parietal lobe, temporal lobe, occipital lobe, cingulate cortex, and insular cortex - the RAR_x is calculated as:

$$RAR_x = \frac{\sum_{i=1}^{N} \mathbb{I}_{R_i}(x)}{\sum_{i=1}^{N} \mathbb{I}_{G_i}(x)} \times 100\% \tag{2}$$

with

$$\mathbb{I}_{R_i}(x) = \begin{cases} 1, & if\ x \in G \cap R \\ 0, & if\ x \notin G \cap R \end{cases} \tag{3}$$

$$\mathbb{I}_{G_i}(x) = \begin{cases} 1, & if\ x \in G \\ 0, & if\ x \notin G \end{cases} \tag{4}$$

where N represents the number of semiology-EZ pairs; G and R represent the sets of EZ locations from the ground truth and the responses from ChatGPT or epileptologists, respectively. $\mathbb{I}_{G_i}(x)$ is an indicator function that scores 1 if the general region x is included in the ground truth G of a semiology-EZ pairing, otherwise 0. $\mathbb{I}_{R_i}(x)$ is a similar indicator function that scores 1 if region x is correctly identified in the intersection of the response R and the ground truth G for a semiology-EZ pairing, otherwise 0. $\sum_{i=1}^{N} \mathbb{I}_{G_i}(x)$ quantifies the aggregate presence of the general region x within the ground truth in N semiology-EZ pairs. $\sum_{i=1}^{N} \mathbb{I}_{R_i}(x)$ quantifies the intersection of the response R and the ground truth G for a semiology-EZ pairing, in N semiology-EZ pairs, regardless of whether a single region or multiple regions for each pair.

Results

Evaluation of responses from ChatGPT

In this study, the performance of ChatGPT-3.5 and ChatGPT-4.0 was evaluated in interpreting seizure semiology from a database of 865 records. The ChatGPT responses generated with ZSP and FSP configurations were compared on the wordiness and reliability regarding EZ locations. The median length (with interquartile range [IQR]) of responses indicated that ZSP yielded significantly longer responses compared to FSP with ChatGPT-3.5 (ZSP: 180 [135~230] words; FSP: 111 [83~155] words) and ChatGPT-4.0 (ZSP: 246 [217~277] words; FSP: 231 [201~270] words).

Considering that ChatGPT's identification encompasses both general and specific regions, our assessment was organized into two separate levels: one for general regions and another for specific regions.

In the general region level, ChatGPT-4.0 using ZSP demonstrated the highest RS: 260 responses (30.06%) were scored as 1 (completely correct), 251 responses (29.02%) were scored as 0.5 to 0.75 (partially correct), and 354 responses (40.92%) were scored less than 0.5 (partially or completely wrong). For the six general regions, the RAR was 89.28% for the frontal lobe, 71.39% for the temporal lobe, 46.24% for the occipital lobe, 31.01% for the parietal lobe, 8.51% for the insular cortex, and 0% for the cingulate cortex (**Figure 2**). These results highlight ChatGPT's proficiency in reliably interpreting seizure semiology for frontal and temporal lobe epilepsies, while underscoring its limitations in accurately interpreting semiology associated with the parietal, occipital, cingulate, and insular epilepsies, which are relatively less common.

At the specific region level, our available collation was limited to 347 records, as some articles merely delineate the general region without detailing specific regions of EZs for individual epilepsy cases. Based on these records, ChatGPT-3.5 with FSP demonstrated the highest RS with 63 responses (18.16%) scored as 1 (completely correct), 25 responses (7.2%) scored as 0.5 to 0.75 (partially correct), and 259 responses (74.64%) scored less than 0.5 (partially or completely wrong). The median [IQR] reliability scores were as follows: ChatGPT-3.5 with ZSP (-0.12 [-0.5~0.11]), ChatGPT-4.0 with ZSP (-0.25 [-1~0.26]), ChatGPT-3.5 with FSP (-0.1 [-0.5~0.5]), ChatGPT-4.0 with FSP (-0.11 [-0.5~0.5]).

- **Comparison of responses from ChatGPT and epileptologists**

An online survey comprising 100 questions regarding EZ locations and corresponding seizure semiologies was formed to collect responses from eight Board-certified epileptologists. Out of these, five participants completed the survey entirely while three others completed it partially. Consequently, the analysis was confined to the fully completed responses of five epileptologists (E1, E2, E3, E4, E5). After

excluding ambiguous characterization of EZs from the literature, the analysis was further narrowed down to 90 questions for the comparison analysis between ChatGPT and epileptologists.

In the general region level, regarding the evaluation of seizure semiology interpretation, epileptologist E4 demonstrated the highest RS: 25 responses (27.78%) were scored as 1 (completely correct), 23 responses (25.56%) were scored as 0.5 to 0.75 (partially correct), and 52 responses (57.78%) were scored less than 0.5 (basically or completely wrong) (**Figure 3**). For six general regions, the RAR was 58.62% for the frontal lobe, 55.56% for the temporal lobe, 63.16% for the occipital lobe, 29.17% for the parietal lobe, 50% for the insular cortex, and 14.29% for the cingulate cortex (**Figure 4**). Notably, ChatGPT demonstrated comparable or superior accuracy to that of epileptologists in interpreting seizure semiology related to the frontal, temporal, parietal, and occipital lobes. However, epileptologists showed significantly better performance in interpreting seizure semiology associated with the cingulate cortex and insular cortex, where EZs are less commonly identified (**Figure 4**).

At the specific region level, our available collation was limited to 43 records, as some articles merely delineate the general region without detailing specific regions of EZs for individual epilepsy cases. Given most of the answers to the specific regions are not provided by epileptologists E3 and E4, only responses from E1, E2, and E5 were analyzed for the specific EZ locations. Compared to ChatGPT, Epileptologist E5 exhibited the highest RS: 3 responses (6.98%) were scored as 1 (completely correct), 8 responses (18.6%) were scored as 0.5 to 0.75 (partially correct), and 32 responses (74.42%) were scored less than 0.5 (basically or completely wrong). The median [IQR] reliability scores were as follows: ChatGPT-3.5 with ZSP (-0.25 [-0.5~0]),

ChatGPT-4.0 with ZSP (-0.5 [-1~0.21]), ChatGPT-3.5 with FSP (-0.25 [-0.5~0]), ChatGPT-4.0 with FSP (-0.12 [-0.62~0.5]), E1 (-0.25 [-0.75~0.5]), E2 (0 [-0.5~0.33]), and E5 (0 [-0.5~0.38]).

Discussion

We conducted the first assessment evaluating the capability of ChatGPT in interpreting seizure semiology, alongside a comparative analysis with Board-certified epileptologists. A total of 865 seizure semiology and EZ location pairs were extracted from 189 articles to assess the responses generated by ChatGPT. A panel of epileptologists was invited to complete an online survey comprising 100 questions to collect their insights on EZ locations based on reported seizure semiology.

The analysis of reliability scores revealed that ChatGPT-4.0 significantly outperformed its predecessor, ChatGPT-3.5. However, the influence of different prompting techniques on the responses from both versions was minimal. ChatGPT-4.0 displayed a tendency to suggest a broader array of potential EZ locations, which, despite increasing the scope of analysis, occasionally led to the inclusion of incorrect locations, thereby slightly reducing its reliability score compared to ChatGPT-3.5. Notably, ChatGPT 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 were comprehensive and surprisingly valuable for reference, with well-founded reasoning regarding EZ locations (**Table 1**). Nonetheless, our accuracy analysis highlighted that for seizure semiology indicating less common EZ locations, such as the cingulate and insular cortex, interpretations 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 [20][21]. 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 [20]. 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 [21].

Limitations

Although this study offered an important reference on the ability of ChatGPT to deliver reliable insights into interpreting seizure semiology, 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, the small number of human epileptologists limited the power of comparative analysis between ChatGPT and human responses.

Furthermore, 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.

Conclusions

In this cross-sectional study of seizure semiology interpretation, ChatGPT-generated responses outperformed or matched responses from epileptologists in regions where EZs are commonly located, such as 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.

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.

Corresponding Author: Feng Liu, PhD, School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ 07030, United States; (fliu22@stevens.edu).

Author Affiliations:

School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ 07030, United States (Jiao, Liu);

Department of Computer Science, Schaefer School of Engineering & Science, Stevens Institute of Technology, Hoboken, NJ 07030, United States (Luo);

Department of Neurology, University Hospitals Cleveland Medical Center, School of Medicine at Case Western Reserve University, Cleveland, OH, 44106, United States (Fotedar);

Department of Neurology, Emory University School of Medicine and Department of Neurology, University of Crete School of Medicine, GA 30322, United States (Karakis);

Department of Neurology and Neurological Surgery, University of California Davis, Davis, CA 95616, United States (Aboud);

Department of Neurology and Weill Institute for Neurosciences, University of California San Francisco, San Francisco, CA 94143, United States (Rao);

Department of Neurology, University of California Davis, Davis, CA 95616, United States (Asmar);

Department of Industrial & Systems Engineering, University of Florida, Gainesville, FL 32611, United States (Xian);

Fowler School of Engineering, Chapman University, Orange, CA 92866, United States (Wen);

Department of Neurology, University of California Davis, Davis, CA 95616, United States (Lin);

Goethe-University Frankfurt, Epilepsy Center Frankfurt Rhine-Main, Department of Neurology, Frankfurt am Main, 60590, Germany (Rosenow);

Department of Neurosurgery, Rutgers Robert Wood Johnson Medical School, Rutgers University, New Brunswick, NJ 08901, United States (Sun).

Semcer Center for Healthcare Innovation, Stevens Institute of Technology, Hoboken, NJ, 07030, United States (Liu)

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), 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)

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˜no, 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. Jialin Liu, Changyu Wang, and Siru Liu. Utility of ChatGPT in clinical practice. *Journal of Medical Internet Research*, 25:e48568, 2023.

14. Douglas Johnson, Rachel Goodman, J Patrinely, Cosby Stone, Eli Zimmerman, Rebecca Donald, Sam Chang, Sean Berkowitz, Avni Finn, Eiman Jahangir, et al. Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the Chat-GPT model. *Research square*, 2023.

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

16. 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.

17. Arno Klein and Jason Tourville. 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in neuroscience*, 6:33392, 2012.

18. 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.

19. 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.

20. 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.

21. 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.

Figures

Figure 1. Pipeline of Database Construction.

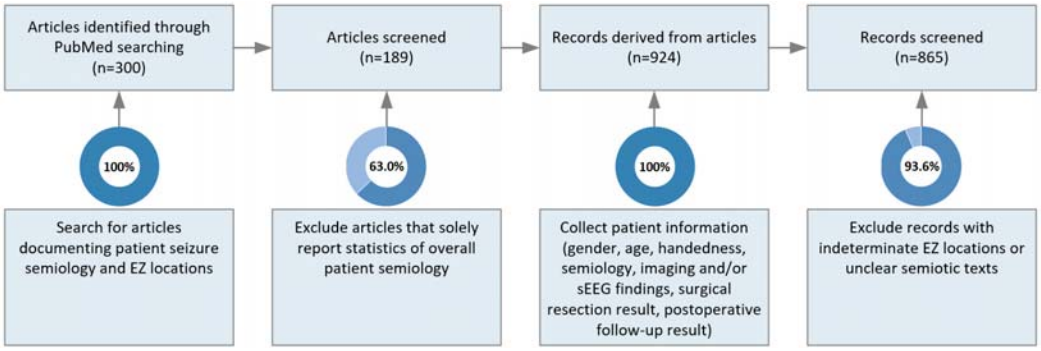
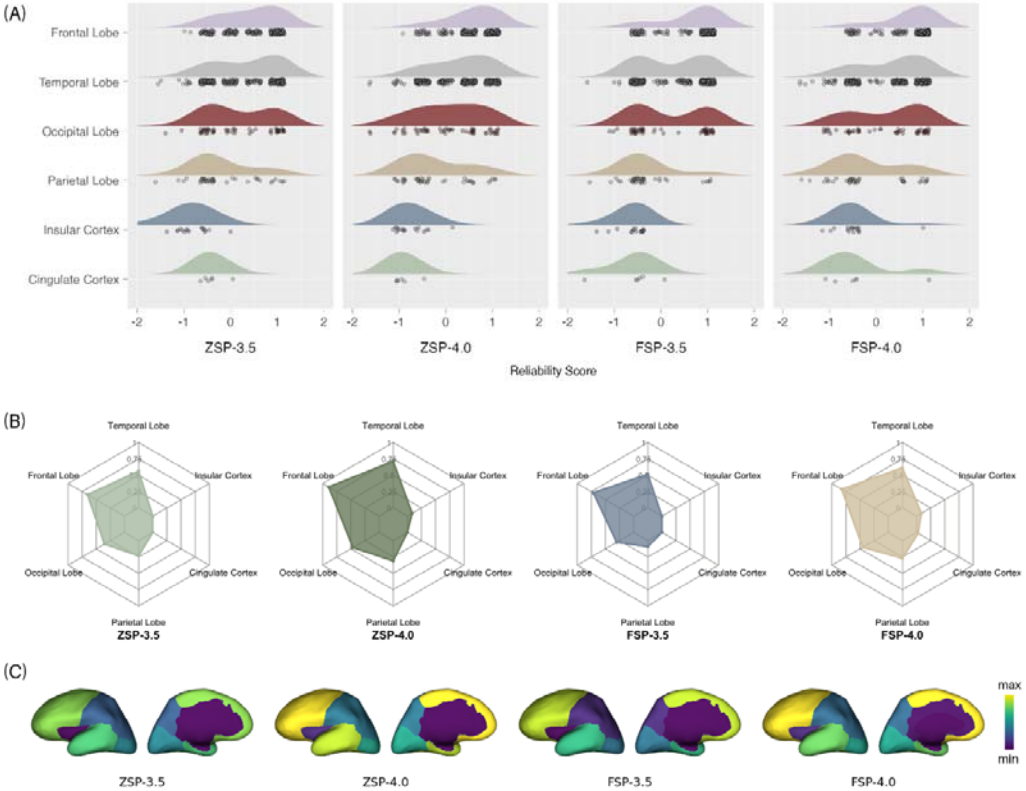
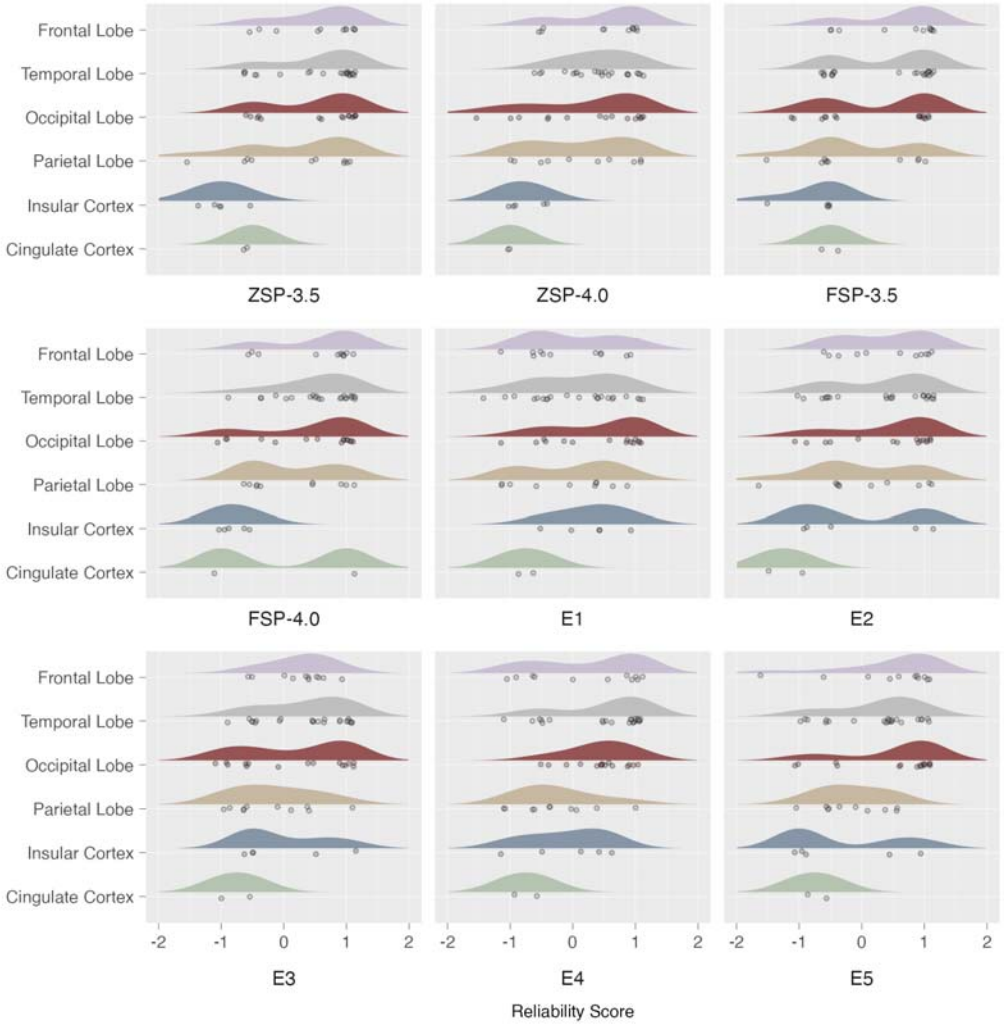


Figure 2. Evaluation of ChatGPT-Generated Answers.



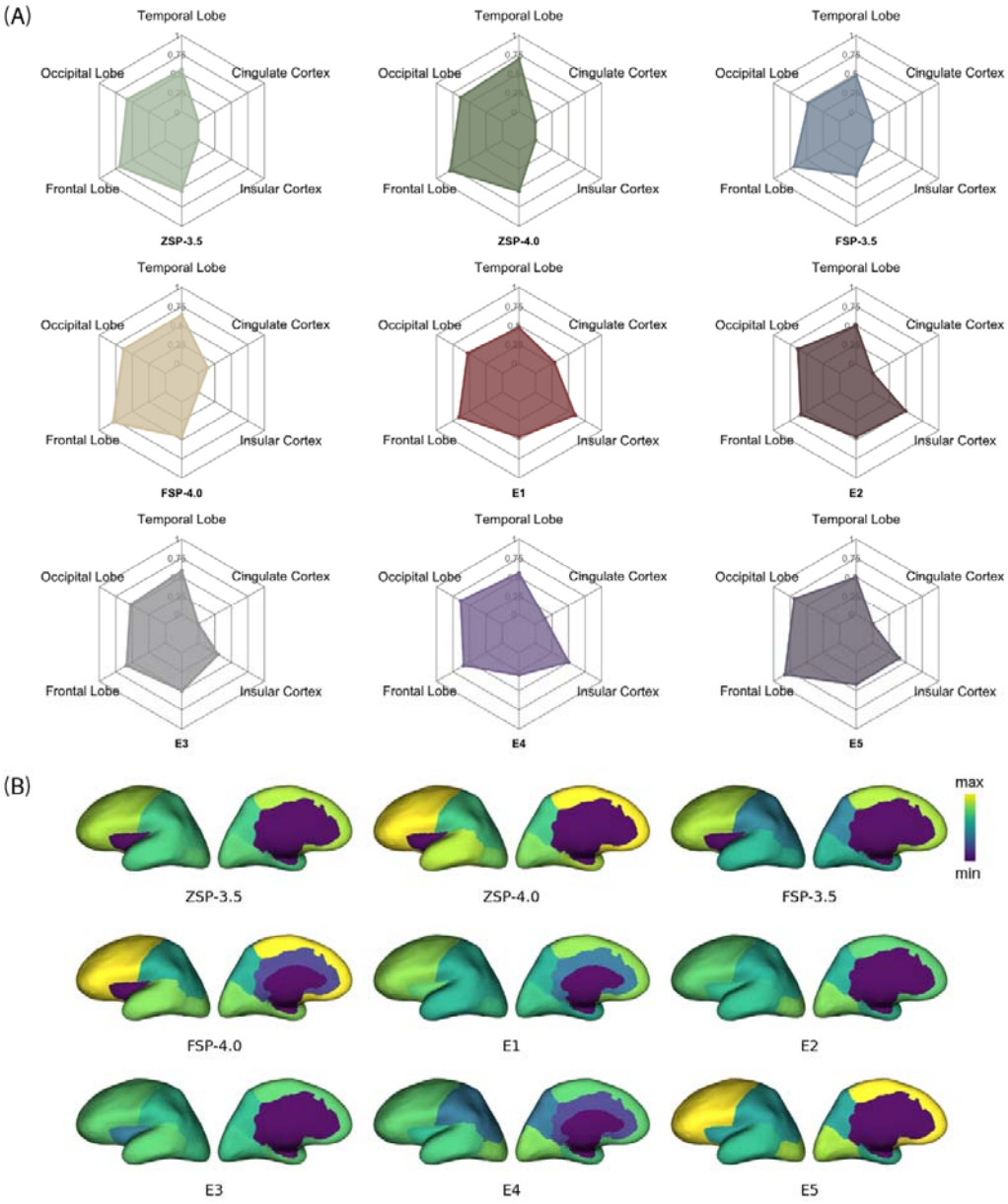
(A), Reliability scores of responses generated by ChatGPT-3.5 with zero-shot prompting (ZSP-3.5), ChatGPT-4.0 with zero-shot prompting (ZSP-4.0), ChatGPT-3.5 with few-shot prompting (FSP-3.5), and ChatGPT-4.0 with few-shot prompting (FSP-4.0). (B) and (C), regional accuracy rate of responses generated by ChatGPT with different prompt configurations (ZSP-3.5, ZSP-4.0, FSP-3.5, FSP-4.0).

Figure 3. Reliability Score Comparison of Answers from ChatGPT and Epileptologist Panel.



Reliability scores of responses generated by ChatGPT with different prompt configurations (ZSP-3.5, ZSP-4.0, FSP-3.5, FSP-4.0) and epileptologists (E1, E2, E3, E4, E5).

Figure 4. Regional Accuracy Rates Comparison of Answers from ChatGPT and Epileptologist Panel.



Tables

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

Question Type	ChatGPT-3.5-Generated Answer		ChatGPT-4.0-Generated Answer	
Zero-Shot Prompting	**Original Answer**	**Extracted Region**	**Original Answer**	**Extracted Region**
The patient is a 54-year-old right-handed male. He has a 2.5-month duration of epilepsy. His seizure semiology includes SPS: frequent episodes of olfactory hallucinosis, head pressure, dysarthria, and then fatigue. According to the patient's reported semiology, specify the general region of the lobe and the specific region inside each lobe that are related to the reported semiology. General Region of the lobe: Specific Region inside each lobe:	Based on the patient's reported seizure semiology, it appears that the seizures involve the following features: 1. Olfactory Hallucinosis: This typically refers to experiencing hallucinations related to the sense of smell. 2. Head Pressure: ..... 3. Dysarthria: ..... 4. Fatigue: ..... Given these features, it's possible that the general region of the lobe related to this semiology is the **Temporal Lobe**. The olfactory hallucinations and dysarthria, in particular, may suggest involvement of the temporal lobe. However, it's important to note that the specific region inside the temporal lobe associated with these symptoms can vary, and further diagnostic evaluation, such as neuroimaging and EEG, would be needed to pinpoint the exact location within the temporal lobe. Additionally, the sensation of head pressure and fatigue may have multifactorial causes, and their association with a specific brain region may be less clear.	General Region: Temporal Lobe Specific Region: NA	The seizure semiology for this 54-year-old right-handed male with a 2.5-month history of epilepsy suggests involvement of specific brain regions: 1. Simple Partial Seizures (SPS) with Olfactory Hallucinosis: - General Region of the Lobe: Temporal Lobe, specifically the medial aspects. - Specific Region inside each Lobe: The uncus or the parahippocampal gyrus, which are parts of the medial temporal lobe associated with olfaction. Olfactory hallucinations typically arise from these areas. 2. Head Pressure: ..... 3. Dysarthria (Difficulty with Speech Articulation): ..... 4. Fatigue (Postictal Symptom): ..... Given these symptoms, the primary focus appears to be in the **medial temporal lobe**, particularly areas related to olfactory processing. The dysarthria points to involvement of the **frontal lobe**.	General Region: Temporal Lobe, Frontal Lobe Specific Region: Medial Temporal
Few-Shot Prompting				
Case 1 The patient is a 24-year-old right-hand male. Her age at seizure onset was 4, and her reported semiology is a slightly painful sensation of tightness in the right shoulder and the region of the right sternocleidomastoid muscle. General Region: Frontal lobe Specific Region: Medial Frontal (include medial premotor and its constituents as its subsets) Case 2 ..... Case 3 ..... According to upper cases, specify the general and specific regions related to the below-reported semiology: The patient is a 54-year-old right-handed male. He has a 2.5-month duration of epilepsy. His seizure semiology includes SPS: frequent episodes of olfactory hallucinosis, head pressure, dysarthria, and then fatigue. General Region: Specific Region:	Based on the reported seizure semiology of the 54-year-old male, let's identify the general and specific regions associated with each aspect: General Region: **Temporal Lobe** - The semiology includes frequent episodes of olfactory hallucinosis, head pressure, dysarthria, and fatigue, which are suggestive of temporal lobe involvement. Specific Region: - SPS (frequent episodes of olfactory hallucinosis, head pressure, dysarthria, and then fatigue): Simple partial seizures with symptoms like olfactory hallucinosis and head pressure are often associated with the temporal lobe. These symptoms can be indicative of an aura that precedes a seizure. Dysarthria (difficulty speaking) may also be associated with temporal lobe activity. It's important to note that the specific regions mentioned here are based on the symptoms described but would typically be determined by a neurologist or epileptologist through clinical assessment and diagnostic tests, including EEG (electroencephalogram) data and neuroimaging. Epilepsy and seizure semiology can vary among individuals, and accurate localization requires a thorough evaluation.	General Region: Temporal Lobe Specific Region: NA	For the 54-year-old right-handed male patient with a 2.5-month duration of epilepsy and seizure semiology of olfactory hallucinosis, head pressure, dysarthria, and fatigue, the general and specific brain regions involved can be identified as follows: General Region: **Temporal Lobe** and **Frontal Lobe** 1. Temporal Lobe: Olfactory hallucinosis (smelling odors that are not present) typically implicates the temporal lobe, particularly areas involved in olfactory processing. 2. Frontal Lobe: ..... Specific Region: **Mesial Temporal Structures** and **Broca's Area** 1. Mesial Temporal Structures (Temporal Pole): The olfactory hallucinosis points towards involvement of the mesial temporal structures, such as the uncus, part of the parahippocampal gyrus, which is involved in olfaction. 2. Broca's Area (Frontal Lobe): ..... It's important to note that the combination of olfactory hallucinations and dysarthria is somewhat unusual and might suggest complex seizure activity involving multiple brain regions. Further diagnostic evaluation, including EEG and brain imaging studies, would be important to more accurately localize these seizures and guide treatment. The interplay between the temporal and frontal lobes, as indicated by this semiology, highlights the complexity and potentially multifocal nature of the patient's epilepsy.	General Region: Temporal Lobe, Frontal Lobe Specific Region: Mesial Temporal Structures (Temporal Pole), Broca's Area (Frontal Lobe)