EpiSemoGPT: A Fine-tuned Large Language Model for Epileptogenic Zone Localization Based on Seizure Semiology with a Performance Comparable to Epileptologists

Shihao Yang[1], Meng Jiao[1], Yaxi Luo[2], Feng Liu[1,3*]

[1]Department of Systems and Enterprises, Stevens Institute of Technology, Hoboken, 07030, NJ, USA.
[2]Department of Computer Science, Stevens Institute of Technology, Hoboken, 07030, NJ, USA.
[3]Semcer Center for Healthcare Innovation, Stevens Institute of Technology, Hoboken, 07030, NJ, USA.

*Corresponding author(s). E-mail(s): fliu22@stevens.edu;

Abstract

Significance: Seizure semiology, which refers to the study of the clinical manifestations during a seizure episode, contains valuable information on inferring the epileptogenic zones (EZs). Given its descriptive nature and recent advances of large language models (LLMs), it is important to design a fine-tuned LLM specially for prediction of EZs by interpreting seizure semiology. In this study, the first fined-tuned LLM, termed as *EpiSemoGPT*, is introduced based on the Mistral-7b-instruct as the foundational LLM model.

Method: A total of 865 cases with descriptions of seizure semiology paired with validated EZs were derived from 189 publications. We used the training dataset of those semiology records and the corresponding EZs to fine-tune the foundational LLM to improve the prediction performance about the most likely EZs. To test the performance of the fine-tuned EpiSemoGPT, 100 well-defined cases are evaluated by analyzing the responses from the EpiSemoGPT and a panel of 5 epileptologists. The responses from EpiSemoGPT and epileptologists were graded based on the rectified reliability score (rRS) and regional accuracy rate (RAR). In addition, the performance of EpiSemoGPT is also compared with its

backbone model Mistral-7b-instruct, as well as different versions of ChatGPT as the representative LLMs.

Result: The EpiSemoGPT can provide valuable presurgical evaluations by identifying the most likely EZs provided with the description of seizure semiology. For comparison between EpiSemoGPT and the panel of epileptologists, the RAR score achieved by EpiSemoGPT in each general region with zero-shot prompt is 53.57% for the frontal lobe, 75.00% for the temporal lobe, 57.89% for the occipital lobe, 62.50% for the parietal lobe, 55.56% for the insula cortex, and 0.00% for the cingulate cortex. Comparatively, the RAR score achieved by epileptologists is 64.83% for the frontal lobe, 52.22% for the temporal lobe, 60.00% for the occipital lobe, 42.50% for the parietal lobe, 46.00% for the insular cortex, and 8.57% for the cingulate cortex. The fine-tuned EpiSemoGPT outperformed its foundational LLM Mistral-7b-instruct and ChatGPT especially with EZs in the insular cortex.

Conclusion: EpiSemoGPT demonstrates comparable performance to epileptologists in EZs inference and its value in the presurgical assessment given the patient's seizure semiology. EpiSemoGPT outperformed epileptologists on interpreting seizure semiology with EZs originating from the temporal and parietal lobes as well as insula cortex, whereas epileptologists outperformed EpiSemoGPT in the frontal and occipital lobes as well as cingulate cortex. Its better performance than the foundational model showing the effectiveness of leveraging high-quality, highly domain-specific samples on fine-tuning of LLMs.

Keywords: Epilepsy Semiology, Large Language Model, Epileptogenic Zones Localization, ChatGPT, Low-Rank Adaption Fine-tuning

1 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]. At present, the mainstay of epilepsy treatment is through antiseizure medications therapy, which helps eliminate or reduce seizure attacks for about two-thirds of epilepsy patients [3]. For those with drug-resistant focal epilepsy, surgical resection of the epileptogenic zones (EZs) may render a curative solution. Seizure semiology, which refers to the description of signs and symptoms exhibited and experienced by a patient during episodes of epileptic seizures [4], yields valuable clues for the localization of the EZs [5]. In order to achieve optimal surgical outcomes, accurately interpreting the seizure semiology plays a crucial role during the preoperative assessment workup.

Recently, large language models (LLMs) have showcased their capabilities across a wide range of natural language processing (NLP) tasks. As a representative example of

LLMs, ChatGPT, trained on a diverse database of text across almost all domains [6], can resemble human experts equipped with cross-disciplinary knowledge. In the field of medical informatics, ChatGPT exhibits an advanced proficiency in processing and interpreting extensive textual data, making it a valuable tool for information retrieval, clinical decision support, and medical report generation [7–10].

The increasing application of LLM in diagnosing various diseases inspired the study of utilizing it to interpret seizure semiology and map its association with the underlying EZs, which may provide a valuable presurgical evaluation tool for patients with epilepsy [11, 12]. However, several previous studies have raised various concerns about the responses from LLM, which may include incorrect or limited information [13, 14], given the limited data sources in specific and highly professional domain. In order to empower LLM the capabilities of generating more accurate and professional responses in those highly professional and specific domains, such as neurology, radiology, neuroscience, etc., fine-tuning of LLM on locally collected and well-annotated data is a commonly used approach with low computational requirement. Examples include ChatDoctor, which is developed by Li et al., is a fine-tuned LLM with Llama as the foundation model leveraging a large dataset of 100,000 patient-doctor dialogues. The ChatDoctor can accurately understand patient's needs and provide informative and professional advice [15]. Xie et al. fine-tuned Llama using an instruct-tuning medical dataset, which is termed as Me-LlaMA, and the Me-LlaMA outperformed many medical LLMs in both general and medical tasks [16]. Labrack et al. proposed BioMistral, which is based on the Mistral foundation LLM model and fine-tuned on medical data collected from high-quality PubMed Central resources, revealed the state-of-the-art performance in medical question-answering tasks comparing to other LLMs [17].

The descriptive nature of seizure semiology makes the LLM a natural fit for its interpretation and predicting its associated EZs which underpins the observed seizure semiology. The semiotic texts can be used to query responses from LLMs regarding the most likely locations of the EZs. In this study, we proposed to use our collected seizure semiology descriptions paired with validated EZs to fine-tune the LLM and improving its performance on predicting the EZs. Mistral-7B, as an emerging lightweight LLM, has been proven to outperform larger LLMs such as llama2 13B in various aspects, and it is easy to fine-tune on some domain specific tasks [18]. Specifically, the Mistral-7B-instruct (v2.0) model is chosen as the foundation model for the EZ prediction given the input of seizure semiology descriptions. The fine-tuned model is termed as EpiSemoGPT. It is also important to understand the advantages and limitations of the developed EpiSemoGPT, which motivates us to further compare its performance

systematically with human experts, and other predominant LLMs such as different versions of ChatGPT.

2 Methods

2.1 Seizure semiology and EZ data collection and annotation

We compiled an epilepsy-specific database from published articles indexed by PubMed. By searching for "seizures", "clinical semiology", and "epilepsy", etc, we identified 189 articles [19]. 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 [20]. 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 excluded. 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. 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 is comprised of 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 samples include both single and multiple EZ locations. In documenting these locations, a majority of articles merely delineate the general region without providing detailed information on the specific regions (anatomical brain regions under lobe level) of EZs for individual epilepsy cases. Consequently, there were more data on general regions (lobe level) than on specific regions. The reported cases present a straightforward anatomic classification for general regions: frontal lobe, parietal lobe, temporal lobe, occipital lobe, cingulate cortex, and insular cortex. It is worth noting that the delineation of specific regions within these six broader areas is more complicated. 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 from the descriptions of specific EZ regions across various studies, we have referenced a classification system based on FreeSurferWiki and LCN-CortLobes [21]. This system organizes the data into six broad regions and over twenty uniquely identifiable specific regions, as detailed in reference [22].

2.2 Fine-tuning of LLM

In this study, the collected semiology records in our self-compiled database were leveraged to investigate the capabilities of fine-tuned LLM in identifying the most likely location of the EZs. We used 756 pairs of semiology descriptions and validated EZs for the fine-tuning of the foundational backbone model, which is the Mistral-7B-instruct due to its good performance at a variety of tasks [18]. The remaining 100 records are used as the testing dataset for the evaluation of EpiSemoGPT, ChatGPT and a panel of epileptologists.

The model was fine-tuned in accordance with the self-instructed format dataset as Stanford Alpaca [23]. To minimize the impact of fine-tuning on the overall LLM performance and improve the fine-tuning efficiency, we applied Low-Rank Adaption (LoRA) [24], which has been integrated into the Hugging Face Parameter Efficient Fine-Tuning (PEFT) library, to optimize the fine-tuning process. Instead of comprehensively fine-tuning the huge weight matrix of the pre-trained LLM, LoRA fine-tunes two smaller matrices that approximate the larger weight matrix. These matrices constitute the LoRA adapter, and the adapter, after fine-tuning, can be easily loaded into pre-trained models and used for inference. The key hyperparameters for fine-tuning are 2 epochs, learning rate of 3×10^{-5}, maximum sequence length of 512 tokens, weight decay of 0.001, warmup ratio of 0.03, LoRA alpha of 32, and LoRA rank of 16. The prompt rendering can be obtained with two configurations, one is the zero-shot prompting fine-tuning [25], where the model was asked to identify the general EZs according to the semiology report. Another is the few-shot prompting fine-tuning [26], which is different from zero-shot prompt fine-tuning in that three samples were given as a hint or prior knowledge in the prompt to assist model training.

2.3 Semiology Interpretation Data Collection

To conduct the comparison between LLMs (both foundation models and fine-tuned ones) and epileptologists on 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 used the 100 hold-out semiology records that spanned all six general regions from our self-compiled database to seek the epileptologists' opinions on the most likely EZs. 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 focused 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, during which 70 survey invitations were sent out to doctors specializing in epilepsy identified from the National Association Epilepsy Center and American Epilepsy Society, and International League Against Epilepsy (ILAE). Among which five epileptologists completed the survey in full and three others completed it partially.

In addition to comparing the fine-tuned LLM performance with the epileptologists, we also took the results from non-fine-tuned pre-trained LLM, including Mistral-7B-instruct itself, ChatGPT 3.5 and ChatGPT 4.0, as the baseline models. While for ChatGPT prediction, 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.

2.4 Statistical Analysis

All responses from EpiSemoGPT, foundation LLMs (ChatGPT v3.5 and v4.0, Mistral-7B-instruct) and the panel of epileptologists underwent thorough review, with the addressed EZ locations being systematically cataloged and summarized. By evaluating and comparing responses from EpiSemoGPT, baseline foundation LLMs, and epileptologists, an in-depth understanding of the strengths and limitations of AI-generated medical information can be achieved which provides insights of its future applications in biomedical and healthcare domain.

To provide a quantitative assessment of the responses provided by the LLMs and epileptologists, two statistical performance metrics were introduced: the rectified Reliability Score (rRS) and the Regional Accuracy Rate (RAR).

The introduced rRS quantifies the accuracy of responses, where a score of 1 indicates 100% correct identification of EZ locations, aligning perfectly with ground truth. A score between 0 and 1 suggests partially accurate, while a score less than 0 indicate misleading responses, potentially complicating preoperative evaluations by neurologists. The rRS is calculated as follows:

$$rRS = \frac{\alpha N_c + \beta(N_i + N_m)}{N_T} \tag{1}$$

where N_T denotes the total count of different EZs identified in both the ground truth and the model predictions for a given semiology case. N_c and N_i respectively represent the counts of correct and incorrect predictions made by LLMs or epileptologists, given the ground truth of each epilepsy case, while N_m is the number of EZs in the ground truth that were not identified by either LLM or epileptologists. Correct predictions (N_c) positively contributes with a weight of $\alpha = 1$, while an incorrect ones (N_i) and missed ones (N_m) carries a negative weight of $\beta = -0.5$ to account for its potential to mislead epileptologists in determining the correct EZ location.

The weighting factors ($\alpha = 1, \beta = -0.5$) are chosen to balance the impact of correct and incorrect responses. With $\alpha = 1$ and $\beta = 0$, the rRS would be greater than or equal to zero, which simply represents the proportion of correct answers but could not reflect the potentially negative impact of incorrect and misleading answers. Conversely, with $\alpha = 1$ and $\beta = -1$, equal weight is assigned to both positive and negative impacts. However, this would result in an rRS of zero when the count of correct responses matches that of incorrect ones for a given semiology, failing to convey the positive impact of correct identifications in narrowing down the diagnoses of EZ. In our study, we set $\alpha = 1, \beta = -0.5$ to prevent the complete negation of a correct answer by an incorrect one, thus ensuring that the rRS accounts for the influence of both positive and negative responses on clinical decision-making.

For example, if we consider a semiology case where the ground truth for the EZ is the frontal, parietal, and occipital lobes, but LLM's response identifies the frontal, temporal, and parietal lobes as the potential EZs. In this scenario, two correct identification ($N_c = 2$), one incorrect identification ($N_i = 1$), and one missing prediction ($N_m = 1$), with the ground truth and identification specifying four different EZ ($N_T = 4$), result in an rRS of 0.25.

The RAR is introduced as a region-specific metric designed to evaluate the precision of EZ localization for specific regions. It calculate the region level localization performance by LLMs or epileptologists against the ground truth. The RAR values range from 0% to 100%. When the response from the LLMs 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:

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

and

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

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

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 LLM 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, and 0 otherwise. $\sum_{i=1}^{N} \mathbb{I}_{G_i}(x)$ 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. For instance, if we consider calculating the RAR for the frontal lobe within a dataset of 100 semiology-EZ pairs. In this dataset, the ground truth may include both cases with a single EZ located in the frontal lobe and cases where the frontal lobe is one of multiple EZs. Suppose there are 50 instances where the ground truth indicates the frontal lobe as an EZ, the value of $\sum_{i=1}^{N} \mathbb{I}_{G_i}(Frontal\ Lobe)$ would be 50. If only 30 responses – whether from LLM or an epileptologist- correctly match the ground truth, the value of $\sum_{i=1}^{N} \mathbb{I}_{R_i}(Frontal\ Lobe)$ would be 30. Consequently, the $RAR(Frontal\ Lobe)$ would be calculated as 60%.

3 Results

3.1 Evaluation of responses from EpiSemoGPT

In this study, the EpiSemoGPT was fine-tuned on 765 records and evaluated from a hold-out 100 records as stated in 2.2. The pre-trained LLMs (LLM without fine-tuning), specifically Mistral-7B, as well as ChatGPT 3.5 and 4.0 were taken as baseline models for comparison.

In the general region inference, EpiSemoGPT using ZSP demonstrated the highest rRS: 33% of responses were scored as 1 (completely correct), 19% of responses were scored as 0.25 to 1 (partially correct), 48% of responses were scored less than 0.25 (almost or completely wrong), while the Mistral-7B has rRS with 31% responses were scored as 1 (completely correct), 20% of responses were scored as 0.25 to 1 (partially correct), 49% of responses were scored less than 0.25 (almost or completely wrong), and the best results from ChatGPT models is ChatGPT 4.0 with 35% of responses were scored as 1 (completely correct), 27% of responses were scored as 0.25 to 1 (partially correct), 38% of responses were scored less than 0.25 (almost or completely wrong).

For the six general regions, the RAR for EpiSemoGPT with ZSP was 53.57% for the frontal lobe, 75.00% for the temporal lobe, 57.89% for the occipital lobe, 62.50% for the parietal lobe, 55.56% for the insular cortex, and 0.00% for the cingulate cortex, and the RAR for EpiSemoGPT with FSP was 60.71% for the frontal lobe, 58.33% for the temporal lobe, 63.16% for the occipital lobe, 62.50% for the parietal lobe, 55.56% for the insular cortex, and 14.29% for the cingulate cortex. By contrast, the Mistral-7B with ZSP has RAR of 42.86% for the frontal lobe, 86.11% for the temporal lobe, 58.33% for the occipital lobe, 42.11% for the parietal lobe, 0.00% for the insular cortex, and 0.00% for the cingulate cortex; and with FSP has RAR of 35.71% for the frontal lobe, 80.56% for the temporal lobe, 45.83% for the occipital lobe, 36.84% for the parietal lobe, 0.00% for the insular cortex, and 0.00% for the cingulate cortex, highlighting the importance of fine-tuning with highly domain specific datasets.

The best result from ChatGPT family is ChatGPT 4.0 with ZSP has RAR of 79.31% for the frontal lobe, 69.44% for the temporal lobe, 57.89% for the occipital lobe, 54.17% for the parietal lobe, 0.00% for the insular cortex, and 0.00% for the cingulate cortex; and with FSP has RAR of 79.31% for the frontal lobe, 63.89% for the temporal lobe, 63.16% for the occipital lobe, 45.83% for the parietal lobe, 0.00% for the insular cortex, and 14.29% for the cingulate cortex.

The results show that EpiSemoGPT performs well not only in interpreting seizure semiology for the most common frontal and temporal lobe epilepsies but also in the

less common occipital and parietal cases. Even in some very rare areas, such as the insula cortex, an accurate inference can still be obtained compared to baseline models that performed extremely poorly.

3.2 Comparison of responses from EpiSemoGPT and epileptologists

An online survey comprising of 100 questions regarding EZ locations and the 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 conducted 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 localization, epileptologist E4 demonstrated the highest rRS: 28% of responses were scored as 1 (completely correct), 26% of responses were scored as 0.25 to 1 (partially correct), 46% of responses were scored less than 0.25 (almost or completely wrong) (Fig.1,Fig.2). For six general regions, the mean RAR was 64.83% for the frontal lobe, 52.22% for the temporal lobe, 60.00% for the occipital lobe, 42.50% for the parietal lobe, 46.00% for the insular cortex, and 8.57% for the cingulate cortex, while the best possible RAR from all epileptologists' responses was 82.76% for the frontal lobe, 58.33% for the temporal lobe, 68.42% for the occipital lobe, 45.83% for the parietal lobe, 60.00% for the insular cortex, and 28.57% for the cingulate cortex (Fig.3,Fig.4). Notably, EpiSemoGPT demonstrated comparable or superior accuracy to best possible answers from all epileptologists at interpreting seizure semiology related to the frontal, temporal, parietal, and occipital lobes, as well as a rare region, the insula cortex. However, epileptologists showed significantly better performance in interpreting seizure semiology associated with the cingulate cortex, where EZs are the least commonly identified (Fig.3,Fig.4). A group t-test based on the bootstrap method was applied to identify the significance of the difference in RAR performance between the averaged epileptologists and EpiSemoGPT with different types of prompting is listed in Fig.5 and 6, which shows EpisemoGPT is comparable to or even outperform epileptologists in temporal, parietal, occipital and insular regions. The epileptologists' diagnosis of EZs in the frontal lobe is more accurate compared to EpiSemoGPT, even though the gap between the two groups is not significant in this case. While in the cingulate cortex identification, the EpiSemoGPT with few-shot prompting is slightly

better than epileptologists. A visualization of RAR distribution on different brain general regions is also shown in Figure Fig.7.

4 Discussion

The first fine-tuned LLM on predicting seizure onset zone by interpreting seizure semiology is proposed, alongside a systematic comparative analysis with a cohort of board-certified epileptologists and foundation LLMs including ChatGPT and Mistral. The fine-tuned EpiSemoGPT, which is trained from a meticulous collected cohort of seizure semiology-EZs pairs from published literatures, showed improved performance than the original foundation model. The EpiSemoGPT model is evaluated using sample-level rectified reliability score (rRS) and regional accuracy rate. The rRS evaluation revealed that EpiSemoGPT demonstrated comparable performance to epileptologists. Moreover, it outperformed its original version, Mistral-7B-instruct, highlighting the importance of using fine-tuning technique using valuable domain specific dataset. The results also showcase using zero-shot prompting gave more robust and accurate results. The reason for this phenomenon is that the few-shot prompting have a certain orientation in model training, and the model will give biased results for samples similar to the given cases when the number of training samples is not sufficient. Compared with the heavyweight ChatGPT, there is a big gap in both knowledge reserve from pre-training and model inference capabilities. According to the assessment of RAR, EpiSemoGPT exceeds all baseline models in inference accuracy in most epileptogenic general regions and achieved significant improvements in rare areas, especially in the insular cortex. Compared with the results of epileptologists, EpiSemoGPT achieved a performance that is comparable to the average level of epileptologists. However, there is still a gap between EpiSemoGPT's performance in the cingulate cortex and that of epileptologists. The discrepancy in epilepsy manifestations originating from cingulate cortex can be attributed to insufficient data samples as there are only 29 in the training samples. At the same time, the seizure semiology of these samples has similarities with other common areas. It is worth noting that even for experts, the localization accuracy in the cingulate cortex is the lowest among all brain regions, with an average RAR of less than 10%. Therefore, it is reasonable for EpiSemoGPT to achieve its current performance when the samples are insufficient and the results are difficult to identify.

Our results align with findings from previous studies assessing LLM's performance in epilepsy-related inquiries. Specifically, Kim et al. assessed the reliability of responses of LLM 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 [27]. Wu et al. evaluated the performance of LLM to a total of 378 questions related to epilepsy and 5 questions related to emotional support. Statistics indicated that LLM provided "correct and comprehensive" answers to 68.4% of the questions. However, when answering "prognostic questions", LLM performed poorly, with only 46.8% of answers rated comprehensive [28]. Moreover, the LLM performance improvement with fine-tuning is consistent with previous works. L1, et al. fine-tuned Llama with patient-doctor dialogues and significantly enhanced providing reliable medical information [15]. Wu et al. performed instruct tuning on Llama with medical question-answering, and fine-tuned LLM exhibited superior performance, even surpassing more complex state of the art models such as ChatGPT [29].

While this study provided valuable insights into the capability of EpiSemoGPT to offer reliable interpretations of seizure semiology, it still has several limitations. Firstly, the training sample volume is limited for some specific regions and unbalanced among regions. In addition, different EZ locations may share similar semiology, which could result in great difficulties in fine-tuning. Unlike ChatGPT, which prefers to give relatively conservative answers, EpiSemoGPT will bravely give the most likely inference, even if these identifications could be incorrect and the basis for its decision is obtained from fine-tuning samples. Specifically, the overlap of manifestations in the EZ locations made EpiSemoGPT learn more possibilities. In summary, EpiSemoGPT tends to give more choices in inferring the most likely EZ locations based on seizure semiology rather than ignoring those that are likely to be the right answer.

The above-mentioned issues are expected to be solved in future work. The training samples can be further improved to enrich the knowledge learned by the model. Moreover, A more complex LLM with more advanced architectures and larger weights can be applied to pursue a better overall performance after fine-tuning [30]. Lastly, the EpiSemoGPT in this version is based on textual information only. Thus, future studies might explore the feasibility of using semiology and video EEG for EZ localization, which could offer a novel method for EZ localization in preoperative assessments.

5 Conclusion

In this cross-sectional study of seizure semiology interpretation, EpiSemoGPT, an fine-tuned LLM, exhibited a satisfactory performance comparable to epileptologists. The model achieved better performance in regions where EZs are commonly located and significantly improved in some rare EZ locations. Overall, our results demonstrate

that EpiSemoGPT could be a valuable tool to assist in the preoperative assessment of epilepsy surgery. In conclusion, our results indicate that EpiSemoGPT could be a useful tool for aiding in the preoperative evaluation for epilepsy surgery. However, it is important to recognize that the information provided by EpiSemoGPT may not always be supported by reliable sources, which poses a challenge for verifying EpiSemoGPT-generated responses. Furthermore, medical professionals, including epileptologists and neurosurgeons, must fully recognize the limitations of EpiSemoGPT and exercise caution when utilizing its responses. This study is an important reference for employing EpiSemoGPT in seizure semiology interpretation while underscoring its present constraints. Presumably, with the continuous development of LLM and more training samples available, the reliability and accuracy of EpiSemoGPT will continue to improve in the foreseeable future.

Acknowledgement

We acknowledge all epileptologists who participated this study and all the insightful discussion.

References

[1] Singh, A., Trevick, S.: The epidemiology of global epilepsy. Neurologic clinics **34**(4), 837–847 (2016)

[2] Gavvala, J.R., Schuele, S.U.: New-onset seizure in adults and adolescents: a review. Jama **316**(24), 2657–2668 (2016)

[3] Chen, Z., Brodie, M.J., Liew, D., Kwan, P.: 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] Jan, M.M., Girvin, J.P.: Seizure semiology: value in identifying seizure origin. Canadian Journal of Neurological Sciences **35**(1), 22–30 (2008)

[5] Tufenkjian, K., Lüders, H.O.: Seizure semiology: its value and limitations in localizing the epileptogenic zone. Journal of clinical neurology **8**(4), 243–250 (2012)

[6] Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., Tang, Y.: 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] Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. Nature medicine **29**(8), 1930–1940 (2023)

[8] Zakka, C., Shad, R., Chaurasia, A., Dalal, A.R., Kim, J.L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., *et al.*: Almanac—retrieval-augmented language models for clinical medicine. NEJM AI **1**(2), 2300068 (2024)

[9] Hirosawa, T., Kawamura, R., Harada, Y., Mizuta, K., Tokumasu, K., Kaji, Y., Suzuki, T., Shimizu, T.: Chatgpt-generated differential diagnosis lists for complex case–derived clinical vignettes: Diagnostic accuracy evaluation. JMIR Medical Informatics **11**, 48808 (2023)

[10] Benary, M., Wang, X.D., Schmidt, M., Soll, D., Hilfenhaus, G., Nassir, M., Sigler, C., Knödler, M., Keller, U., Beule, D., *et al.*: Leveraging large language models for decision support in personalized oncology. JAMA Network Open **6**(11), 2343689–2343689 (2023)

[11] Ueda, D., Mitsuyama, Y., Takita, H., Horiuchi, D., Walston, S.L., Tatekawa, H., Miki, Y.: Diagnostic performance of chatgpt from patient history and imaging findings on the diagnosis please quizzes. Radiology **308**(1), 231040 (2023)

[12] Mehnen, L., Gruarin, S., Vasileva, M., Knapp, B.: Chatgpt as a medical doctor? a diagnostic accuracy study on common and rare diseases. medRxiv, 2023–04 (2023)

[13] Liu, J., Wang, C., Liu, S.: Utility of chatgpt in clinical practice. Journal of Medical Internet Research **25**, 48568 (2023)

[14] Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., Chang, S., Berkowitz, S., Finn, A., Jahangir, E., et al.: Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model. Research square (2023)

[15] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., Zhang, Y.: Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. Cureus **15**(6) (2023)

[16] Xie, Q., Chen, Q., Chen, A., Peng, C., Hu, Y., Lin, F., Peng, X., Huang, J., Zhang, J., Keloth, V., et al.: Me llama: Foundation large language models for medical applications. arXiv preprint arXiv:2402.12749 (2024)

[17] Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., Dufour, R.: Biomistral: A collection of open-source pretrained large language models for medical domains. arXiv preprint arXiv:2402.10373 (2024)

[18] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)

[19] Canese, K., Weis, S.: Pubmed: the bibliographic database. The NCBI handbook **2**(1) (2013)

[20] Alim-Marvasti, A., Romagnoli, G., Dahele, K., Modarres, H., Pérez-García, F., Sparks, R., Ourselin, S., Clarkson, M.J., Chowdhury, F., Diehl, B., *et al.*: Probabilistic landscape of seizure semiology localizing values. Brain Communications **4**(3), 130 (2022)

[21] Computational Neuroimaging, L.: FreeSurferWiki. https://surfer.nmr.mgh.harvard.edu/fswiki Accessed 2024-05-20

[22] Klein, A., Tourville, J.: 101 labeled brain images and a consistent human cortical labeling protocol. Frontiers in neuroscience **6**, 33392 (2012)

[23] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model (2023)

[24] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)

[25] Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)

[26] Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., Wang, L.: An empirical study of gpt-3 for few-shot knowledge-based vqa. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 3081–3089 (2022)

[27] Kim, H.-W., Shin, D.-H., Kim, J., Lee, G.-H., Cho, J.W.: 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)

[28] Wu, Y., Zhang, Z., Dong, X., Hong, S., Hu, Y., Liang, P., Li, L., Zou, B., Wu, X., Wang, D., *et al.*: Evaluating the performance of the language model chatgpt in responding to common questions of people with epilepsy. Epilepsy & Behavior **151**, 109645 (2024)

[29] Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Pmc-llama: Further finetuning llama on medical papers. arXiv preprint arXiv:2304.14454 (2023)

[30] Kamble, K., Alshikh, W.: Palmyra-med: Instruction-based fine-tuning of llms enhancing medical domain performance (2023)

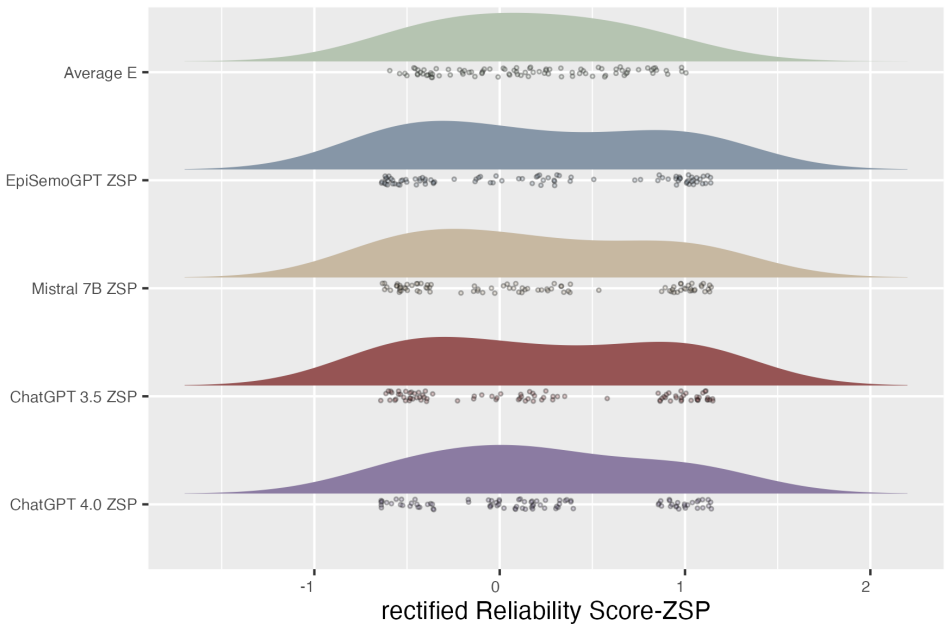

Fig. 1 Distribution of averaged epileptologists and different LLMs with zero-shot prompting performance according to rRS metrics.

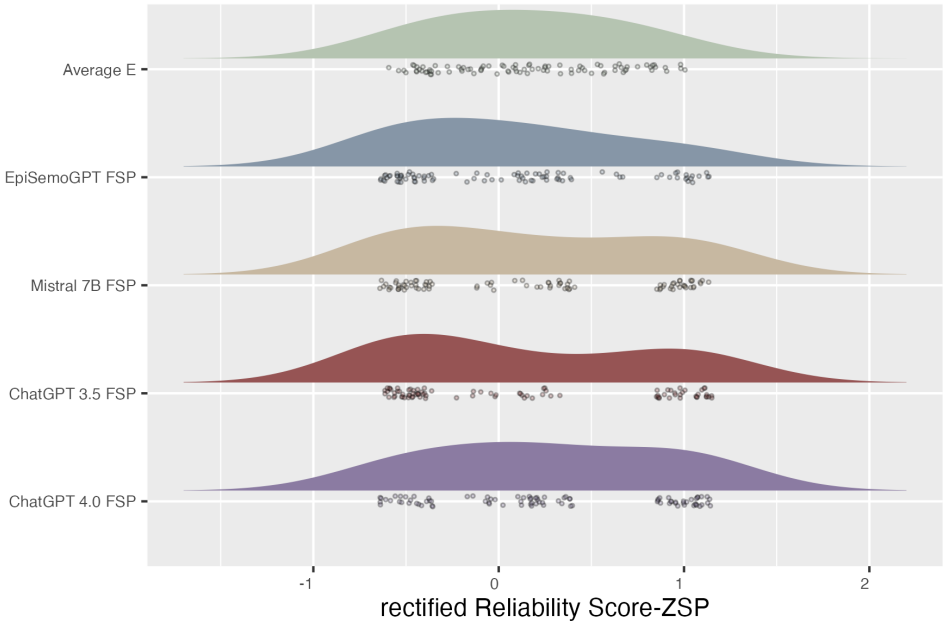

Fig. 2 Distribution of averaged epileptologists and different LLMs with few-shot prompting performance according to rRS metrics.

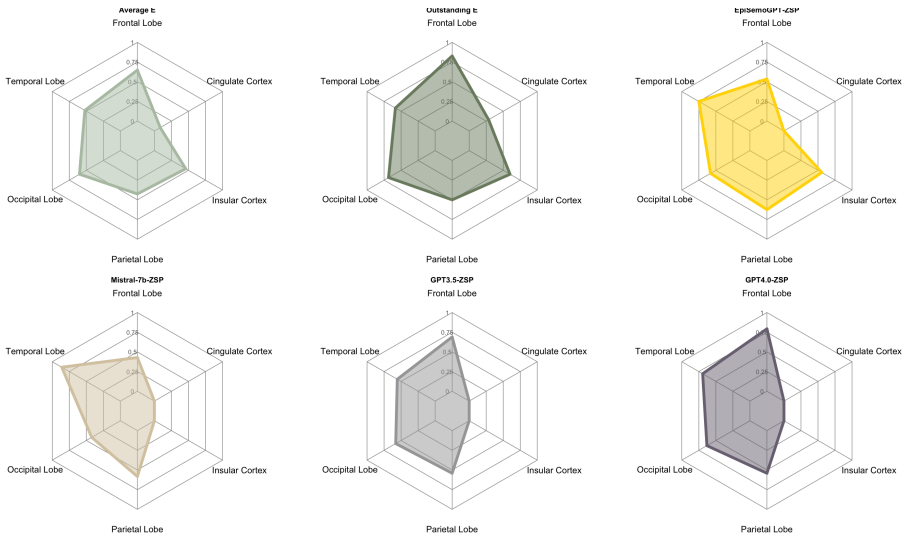

Fig. 3 RAR score for averaged epileptologists and different LLM with zero-shot prompting performance.

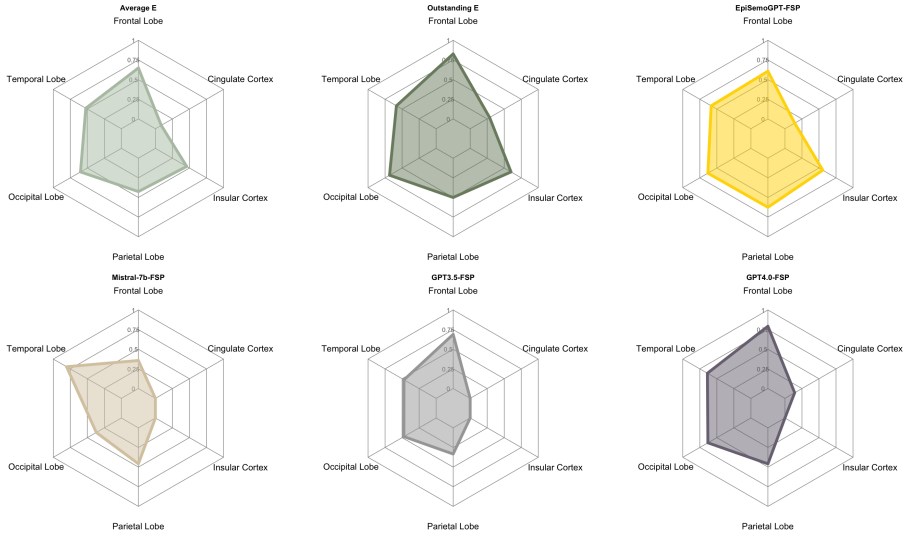

Fig. 4 RAR score for averaged epileptologists and different LLM with few-shot prompting performance.

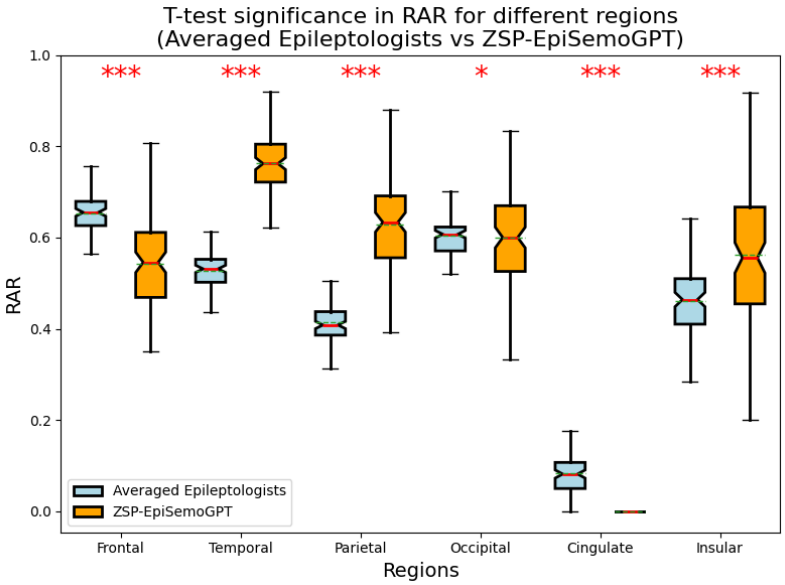

Fig. 5 T-test significance between epileptologists and EpiSemoGPT with zero-shot prompting for RAR on different regions based on the bootstrap method. The star signs on the top of the figure show the significance level of the difference, where * represents p-value<0.1, ** represents p-value<0.01, and *** represents p-value<0.001.

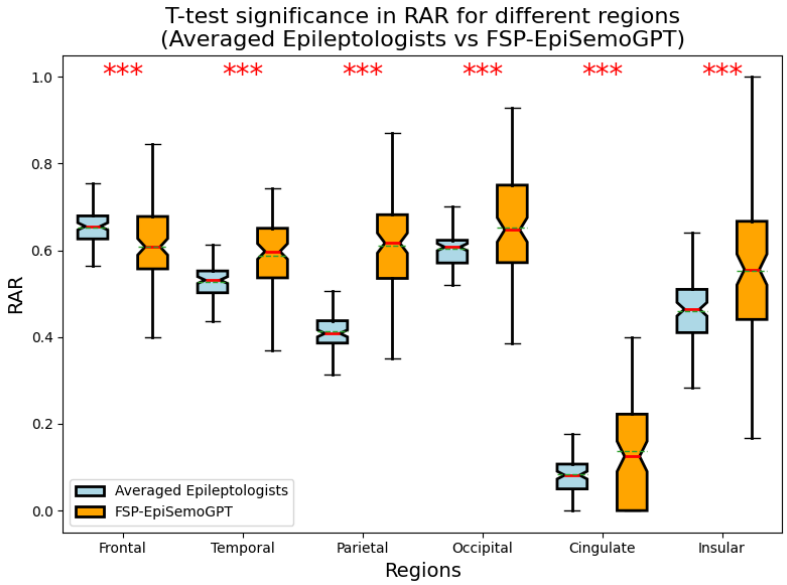

Fig. 6 T-test significance between epileptologists and EpiSemoGPT with few-shot prompting for RAR on different regions based on the bootstrap method. The star signs on the top of the figure show the significance level of the difference, where * represents p-value<0.1, ** represents p-value<0.01, and *** represents p-value<0.001.

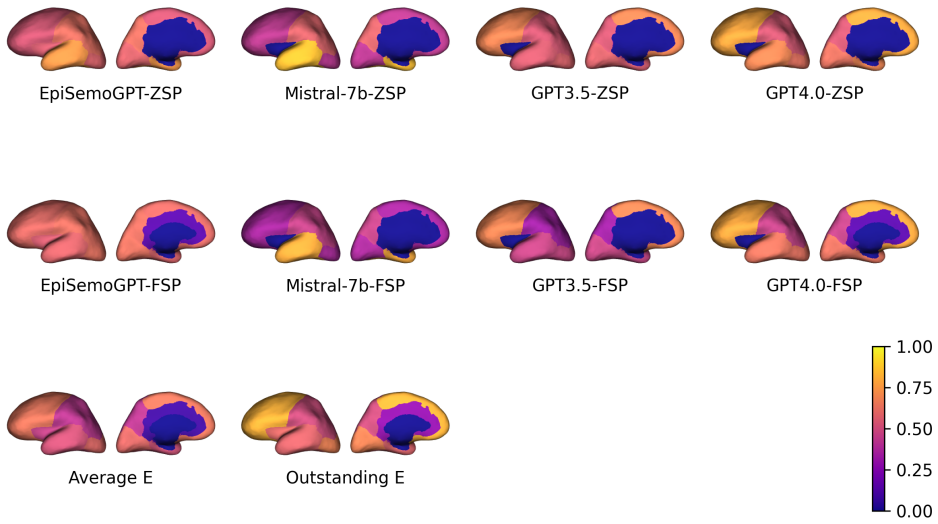

Fig. 7 RAR distribution on different brain general regions.