SemioLLM: Assessing Large Language Models for Semiological Analysis in Epilepsy Research

Meghal Dani [1 2] **Muthu Jeyanthi Prakash** [1] **Zeynep Akata** [3 2] **Stefanie Liebe** [4]

Abstract

Large Language Models have shown promising results in their ability to encode general medical knowledge in standard medical question-answering datasets. However, their potential application in clinical practice requires evaluation in domain-specific tasks, where benchmarks are largely missing. In this study *semioLLM*, we test the ability of state-of-the-art LLMs (GPT-3.5, GPT-4, Mixtral 8x7B, and Qwen-72chat) to leverage their internal knowledge and reasoning for epilepsy diagnosis. Specifically, we obtain likelihood estimates linking unstructured text descriptions of seizures to seizure-generating brain regions, using an annotated clinical database containing 1269 entries. We evaluate the LLM's performance, confidence, reasoning, and citation abilities in comparison to clinical evaluation. Models achieve above-chance classification performance with prompt engineering significantly improving their outcome, with some models achieving close-to-clinical performance and reasoning. However, our analyses also reveal significant pitfalls with several models being overly confident while showing poor performance, as well as exhibiting citation errors and hallucinations. In summary, our work provides the first extensive benchmark comparing current SOTA LLMs in the medical domain of epilepsy and highlights their ability to leverage unstructured texts from patients' medical history to aid diagnostic processes in health care.

[1]University of Tübingen, Tübingen, Germany [2]Helmholtz Munich, MCML, Munich, Germany [3]Technical University of Munich, Munich, Germany [4]University Hospital Tübingen, Department of Neurology and Epileptology, Tübingen, Germany. Correspondence to: Meghal Dani <meghal.dani@uni-tuebingen.de>.

AI4Science Workshop at the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Epilepsy is a chronic neurological disorder currently affecting 70 million people worldwide (Thijs et al., 2019). It is characterized by a predisposition of the central nervous system to unpredictably generate seizures. About two thirds of patients suffer from focal epilepsies, which produce distinct seizure-related changes in sensation and behavior depending on the seizure onset zone (SOZ) in the brain (Beniczky et al., 2022; Lüders et al., 2006). Patient reports on seizure symptoms - so-called seizure semiology - are therefore routinely recorded by clinicians and used as one crucial source of information for localization of the SOZ. This is particularly important for patients with drug-resistant epilepsies, for which surgical resection of the SOZ is the only potentially curable therapy option (Wiebe et al., 2001; Sisodiya & Goldstein, 2007). For these patients, the clinicians' task is to determine a confident and accurate estimate of the SOZ, as only then a recommendation for epilepsy surgery can be made.

The advent of Large Language Models (LLMs) has sparked interest in their potential to leverage medical knowledge (Singhal et al., 2023; Savage et al., 2024; Sarvari et al., 2024). However, there is a lack of systematic evaluation of LLMs' understanding of specific clinical domains. Addressing this requires large-scale annotated text-datasets, systematic investigation of prompt designs, exploration of in-context learning strategies - all in comparison to problem evaluation technique followed by healthcare professionals in real world. Here, we address several of these questions and show a first comprehensive investigation for benchmarking currently accessible LLMs in the domain of epilepsy. For this task, we leverage a large-scale annotated database Semio2Brain (Alim-Marvasti et al., 2022) linking unstructured text reports of seizure symptoms to seizure onset zones. We highlight the contributions and key insights of this paper as follows:

- First, we measure the correctness of the SOZ localization outputs generated by the LLMs. Second, we approximate the confidence in their outputs using entropy. Third, we perform a human evaluation to assess the LLMs' understanding, reasoning, and source retrieval abilities specific to the epilepsy domain.

- We estimate the impact of prompt strategies on the task employing five prompt styles: zero-shot, few-shot, Chain of Thought (CoT), few-shot CoT, and self-consistency (SC). Performance substantially increases with more sophisticated prompting strategies, highlighting the models' ability to leverage contextual information and domain-specific knowledge in epilepsy.
- With respect to model comparison, GPT-4 emerges as the top-performing model across all evaluation metrics. Mixtral8x7B, while competitive with GPT-4 in performance, exhibits tendencies to hallucinate in source citations and provides incomplete and partially incorrect reasoning.Notably, GPT-3.5 and Qwen-72B exhibit higher confidence levels in their outputs, albeit with reduced correctness.

2. Related Work

LLMs in medicine: Recent advancements in large language models have led to impressive performances across a variety of tasks such as report summarization (Yu et al., 2023) or question answering (Singhal et al., 2023; Liévin et al., 2023; Li et al., 2024; Brown et al., 2020; Bubeck et al., 2023; Achiam et al., 2023). Furthermore, recent advancements in prompting techniques (Brown et al., 2020; Dong et al., 2022; Wei et al., 2022; Wang et al., 2022) have shown promising results in utilizing existing LLMs without any weight updates. This opens up exciting possibilities for leveraging pre-trained LLMs to go beyond question-answering and analyze unstructured text obtained from medical histories to aid the diagnostic process in specific medical domains.
NLP in epilepsy: The potential of Natural Language Processing (NLP) and LLMs in the field of epilepsy has only very recently been discussed (van Diessen et al., 2024; Boßelmann et al., 2023), with only one study investigating basic knowledge on epilepsy in a question-answering scheme and comparing LLMs performance to experienced epileptologists (Kim et al., 2024). In our study, we go beyond this general approach and harvest seizure descriptions obtained from medical histories of epilepsy patients which provide a rich and unique source of information relevant for important diagnostic decisions. Moreover, we are able to test model performance not only against clinicians' performance, but also the ground-truth data obtained from the annotated dataset.

3. Database and data curation

We utilize the publicly available Semio2Brain (Alim-Marvasti et al., 2022) database, which maps seizure semiologies to brain regions based on a meta-analysis of seizure descriptions from 4,643 patients. Each entry in the database includes a description of a seizure symptom—either a behavioral or sensational phenomenon that occurred during a seizure—and which is assigned seven potential major brain regions: temporal lobe, frontal lobe, cingulate gyrus, parietal lobe, occipital lobe, insula, and hypothalamus. The assignment of brain regions to seizure descriptions is based on two types of information:

1. Post-operative Seizure Freedom: Knowledge about seizure freedom after resection of the brain region.
2. Seizure Activity: Seizure patterns recorded from intracranial EEG located within the brain region.

Both types of information serve as potential ground truths linking seizure semiology to SOZ in clinical practice. For our task, we focused on cases based on post-operative seizure freedom, as this is considered the gold standard for post-hoc evaluation of successfully identifying the SOZ. Finally, we performed several data preprocessing steps, including expanding abbreviations in the semiology descriptions, correcting spelling errors, and removing uninformative words or keywords (for details see supplementary material). This refinement process resulted in a final dataset of 1,269 entries.

4. Methodology

4.1. Task Formulation

The LLM's task is to predict the next probable token T given the input prompt $\widehat{P}$ consisting of persona P, the user query Q, and the instruction format I as shown in Fig. 1. The prediction is based on maximizing the likelihood of the next token. The likelihood function L can be defined as:

$$L(T|P,Q,I) = \prod_{i=1}^{n} P(T_i|P,Q,I), \tag{1}$$

where $P(T_i|P,Q,I)$ is the probability of the token T_i given the persona P, the user query Q, and the instruction format I.

We emulate a clinical task exactly how it occurs in practice, where epileptologists evaluate seizure semiologies in order to predict brain regions that are likely involved in producing the seizure symptoms. Importantly, there is no ground truth available for clinicians when making this judgement (as a potential resection has not occurred yet). Thus, for the clinician several brain regions might be possible as potential seizure onset zones, especially if seizure symptoms are complex. Hence, for the SOZ localization task, in the user query, we ask for the percentage likelihood of brain regions given a semiology behavior s. We consider 7 main brain regions(R) in our case such that R={"Temporal Lobe", "Frontal Lobe", "Cingulate Gyrus", "Parietal Lobe", "Occipital Lobe", "Insula", "Hypothalamus"}. The instruction for the output is to obtain a dictionary D where each key is a brain region

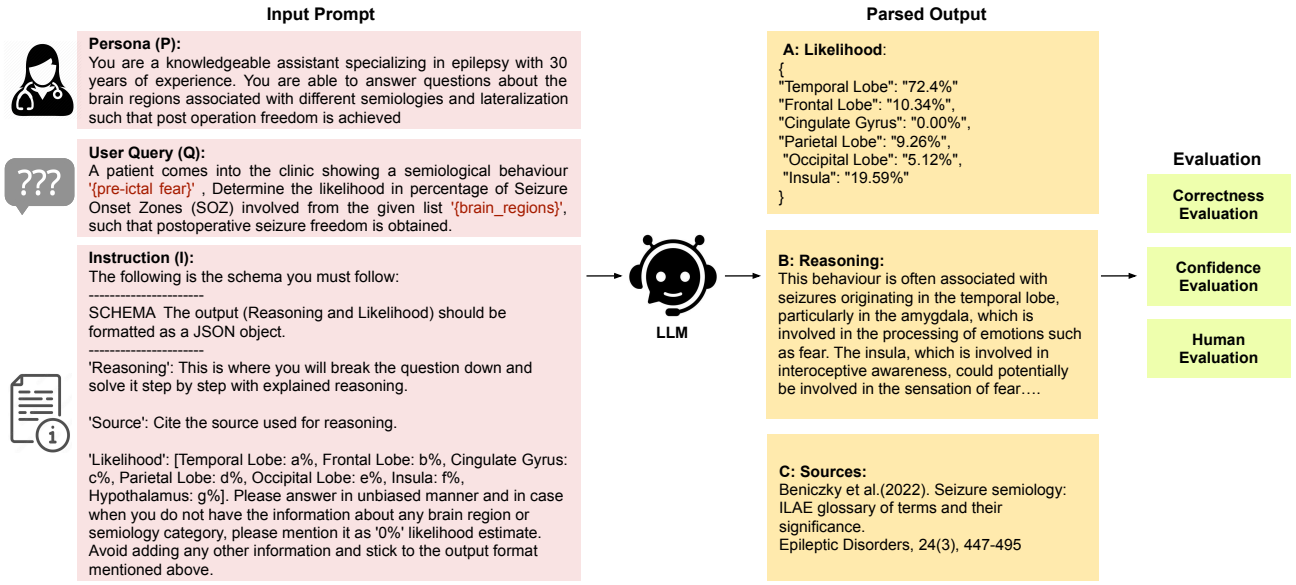

Figure 1. Overview of the experimental framework: The input prompt to the LLM is structured into three components (pink box): (1) Persona (P), which defines the role the model should assume while responding; (2) User Query (Q), which includes the semiology to be analyzed; and (3) Output Format Instruction (I), which specifies the desired format of the response. The model's response (orange box) includes the 'likelihood' of the seizure onset zone (SOZ) for seven major brain regions, along with 'reasoning'** and 'sources'** considered to give the output.All model responses are evaluated for correctness and confidence (uncertainty estimation). To clinically validate the results human evaluation is performed. (**obtained only for few-shot CoT)

$r \in R$ and the value is the likelihood $\ell'(r|s)$ as shown in the equation below:

$$D = \{r : \ell'(r|s)\}, \quad r \in R. \tag{2}$$

We obtain the clinical reasoning and supporting evidence from the LLM for the likelihood output it generates. Specifically, the LLM is expected to provide a detailed rationale and cite relevant medical literature or sources that it has leveraged to arrive at the predicted likelihoods for the potential SOZs.

4.2. LLMs

For the localization task, we compare the performance of four popular LLMs: GPT-3.5 (specifically "gpt-3.5-turbo-1106"), GPT-4 ("gpt-4-turbo-preview" in our case), Mixtral-8x7B Mixture of Experts (MoE) instruct version and Qwen-72B chat model.

4.3. Prompting Techniques

The task we undertake is challenging and requires rigorous prompt engineering to define our final prompt template. As shown in Fig.1, our base prompt template consists of three parts: (i) persona that the LLM should assume i.e. of an expert epileptologist, (ii) user query where we add the semiology we want to inquire about and the exact question, and (iii) the instruction which guides the model to output the response according to the required schema.

Zero-Shot prompting: This structured approach is particularly effective for zero-shot prompting, where the model is expected to perform tasks based solely on its pre-existing knowledge without any specific examples or prior training on the task.

Few-shot prompting: We carefully curate a handful of examples guided by a clinician, which are added as prompt text in the input context. The key idea of this in-context learning is to demonstrate the input and output structure required to the model without any finetune or weight update (Brown et al., 2020; Dong et al., 2022). The context provided typically have $K = 5$ examples of queries and answers, and then one user query for which the model is expected to provide the response.

Chain-of-Thought(CoT) prompting: Chain-of-thought prompting is a technique to ask the model to think *step-by-step* and provide intermediate reasoning and sources used to get to the final answer. This technique mimics a human cognitive process, to break complex problems into small, manageable steps. It is helpful where a straightforward answer may not be trivial (Wei et al., 2022). To do this, we add two more keys to the schema of the base prompt namely, "Reasoning" and "Sources". We later assess the reasoning for correctness, completeness, and reliability of the sources cited.

Few-shot COT: We also use a hybrid technique to combine few-shot and chain-of-thought prompting.Specifically, we provide the model with examples (curated by expert clinicians) demonstrating how an expert reasons through the problem and arrives at the final conclusion. The model is then expected to learn and mimic this reasoning process in-context and generate its output in the form of a dictionary with keys 'reasoning' and 'sources', as illustrated in Fig. 1. This allows us to gain insights into the model's decision-making process, assess the validity of its reasoning, and evaluate the relevance of the sources it cites.

Self Consistency (SC): Wang et al. (2022) introduced this concept of "self-consistency" which involves generating multiple response paths and then choosing the most consistent response as the final answer. This is done via a majority voting technique. By cross-referencing different outputs, we can ensure that the final response is robust and dependable. In our problem statement, we get likelihoods of 7 brain regions at a time. We use median majority voting to get the winner or the most consistent output. For each brain region r_i, compute the median of its likelihoods over the 5 iterations L_{ij} for $j \in \{1, 2, 3, 4, 5\}$, and then subtract this median from the likelihoods of each iteration to obtain the adjusted likelihoods A_{ij} as follows:

$$A_{ij} = |L_{ij} - median(L_{i1}, L_{i2}, L_{i3}, L_{i4}, L_{i5})| \quad (3)$$

Now, sum of these adjusted likelihoods for each iteration is obtained and the iteration with the minimum sum is the winner W as it is most close to the median:

$$W = \arg\min_j \sum_{i=1}^{7} A_{ij} \quad (4)$$

The equations 3 and 4 capture the entire process of majority voting using median activation. The rationale behind using the median is its robustness to outliers and its representation of the central tendency of the data. Unlike the mean and mode, where the former could be skewed by anomalies in the data and the latter which identifies the most frequent value and could lead to no winner in case all outputs are different.

5. Experiments and Results

5.1. Correctness Measure

Performance is assessed using a multi-class evaluation emulating the fact that the actual ground truth (post-surgical seizure freedom) can only be true for 1 of the 7 brain regions. The predicted SOZ is determined by selecting the class with the highest likelihood values. In certain cases when model assigns equal high probabilities to more than one brain region, we resolve the tie with a simple deterministic approach np.argmax(). Using these labels and the ground truth labels, we compute precision, recall, and F1 score for each class.

LLM	Prompt Strategy				
	Zero-Shot	Few-Shot	CoT	FewShot-CoT	SC
GPT-3.5	38.13	48.36	49.32	51.65	51.84
GPT-4.0	**52.33**	**50.96**	**52.64**	52.11	**53.78**
Mixtral8x7B	23.07	42.26	48.97	**52.72**	52.29
Qwen-72B	39.23	44.64	45.47	48.17	45.75
Clinician	49.07				

Table 1. Comparison results for four state-of-the-art LLMs with different prompting techniques: zero-shot (ZS), few-shot, chain-of-thought (CoT), few-shot CoT and Self-Consistency(SC) prompting techniques. The table reports standard weighted F1 scores, where higher values indicate better performance. The last row presents the evaluation of clinician responses for comparison.

To account for potential class imbalances, we calculate the weighted average precision, recall, and F1 score across all classes. The weighting is based on the support (number of instances) for each class. This approach ensures that classes with more instances contribute proportionally to the overall metric, mitigating the impact of class imbalance. To estimate a lower bound performance expected by chance, we compute the minimum precision and recall. We take the ratio of positive instances (class support) to the total number of instances (total support), assuming a naive baseline classifier that predicts all instances as positive. Using the support values, the weighted precision, recall, and F1 score for this naive baseline is 39.34%. (see Supplement for full calculation). Finally, we asked a clinician to provide their input, similar to how we query the LLMs, and compute the F1 score, which is 49.07%.

To estimate the variability and uncertainty associated with the F1 scores, we perform bootstrapping (Tibshirani & Efron, 1993), a resampling technique that provides robust estimates of the metric's distribution. Specifically, we resample 10% of the full dataset 999 times, and calculate the F1 score for each sub-sample. We report the mean and standard deviation for each model in the supplementary material. Similarly, we subsample the responses of the clinician in order to obtain variance estimate as shown in Fig. 2(B).

We report our localization results in Tab.1. For zero-shot prompting, only GPT4 shows significantly higher performance than the lower bound. With better prompting techniques, however, all models achieve substantially higher, and significantly better performance than expected by chance and is even comparable to clinical evaluation (see Fig. 2 (A)). For almost all prompting styles GPT-4.0 outperforms the other models and achieves the highest F1 score of 53.78% using the self-consistency prompt-style. For few-shot CoT Mixtral-8x7B shows the best performance with an F1 score of 52.72%. Importantly, the models' performance is competitive with the clinician's response. Since there is no benchmark to compare our results with, we further assess the significance of our model's F1 scores using a permu-

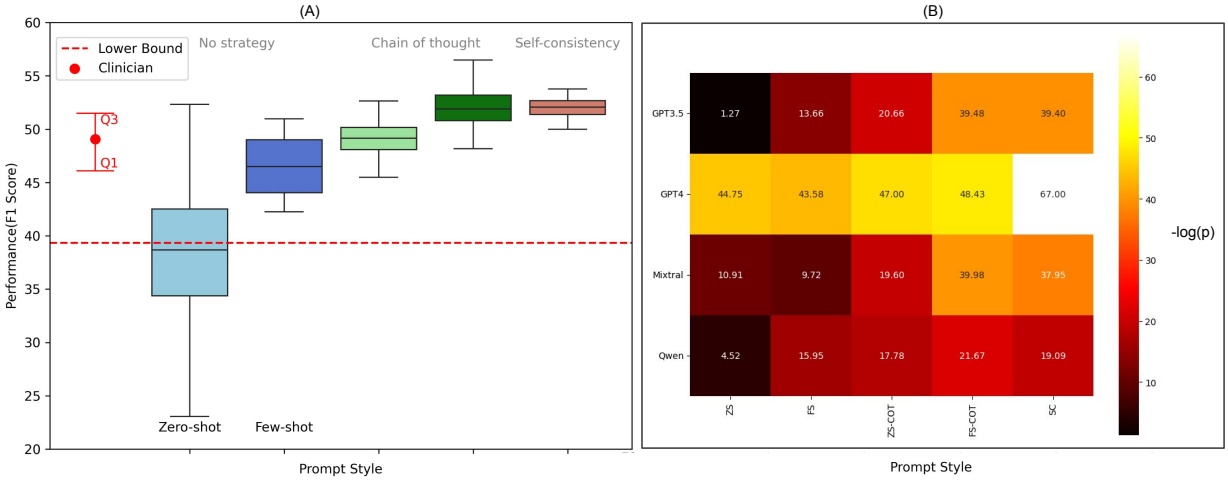

Figure 2. Correctness measure results.(A) This panel illustrates the impact of prompt engineering on SOZ localization performance. The light blue and green box plots represent zero-shot and zero-shot CoT prompting, respectively, while the dark blue and green box plots represent few-shot and few-shot CoT prompting, respectively. Each box plot is generated using the F1 scores from all four models under evaluation. The random chance lower bound is indicated at 39.34%, and clinician performance is shown at 49.07% F1 score, with bootstrapped quantiles Q1 and Q3 denoted. (B) This panel presents the performance difference of each experiment compared to its random bootstrapped distribution. The heatmap displays the negative log(p) values, indicating that the main results in Table 1 are significantly different from random performance (except GPT-3.5 and Qwen-72B zero-shot)

tation test. For each model, we generate random outputs by shuffling the model's predictions 999 times, creating a random distribution of F1 scores. To quantify the deviation of our model's actual F1 score from the random distribution, we perform a Z-score normalization. This transformation allows us to express the difference between the actual and random scores in terms of standard deviations, providing a standardized measure of significance as follows:

$$z = \frac{(x - \mu)}{\sigma} \tag{5}$$

where, x is the actual F1 score, μ is the mean of the random distribution and σ is the standard deviation of the random distribution. We calculate the associated p-value of the resulting Z-score and report the negative log(p) for illustration purposes in the Fig. 2(B). Values larger than 4.60 are significantly different considering a conservative critical p-value of 0.01. Using this criterion, our results demonstrate that, except for GPT-3.5 and Qwen at Zero-shot, all values are highly significantly different from random (for all the values refer the supplementary material).

5.2. Confidence measure / Uncertainty Estimate

The likelihood output from the LLMs is more informative than a single class prediction, as it allows for understanding which classes the model considers plausible, and to what degree, rather than just which class it considers the "winner". We leverage this feature to approximate a confidence/uncertainty measure. Specifically, we employ Shannon entropy for this purpose, which is a fundamental concept in information theory that quantifies the "fuzziness" or uncertainty of a system's state. Given a discrete random variable X with possible outcomes $x_1, x_2, ..., x_n$, each with probability $P(x_i)$, the Shannon entropy $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i) \tag{6}$$

The minimum entropy value is 0, which occurs when the model assigns a likelihood of 100% to one brain region and 0% to all others. The maximum entropy value is achieved when the model assigns equal likelihood to all seven brain regions, i.e. $\frac{1}{7} * 100 = 14.28$. We normalize the $H(X)$, which effectively "calibrate" the metric, enabling a fair assessment of the relative gain achieved by each model in comparison. We eventually report the loss entropy, defined as $1 -$ normalized entropy. If loss entropy trends towards 1, it means that the model is confident or less uncertain and vice versa if it shows a tendency towards 0.

In Fig. 3(A) we can see a clear trend that the model becomes more confident when given some support examples in the case of Few-shot ICL (No strategy) and Few-Shot Chain of Thought. Evidently, it is least certain in zero-shot and most certain in self-consistency. Note the outlier in the self-consistency prompt style, which is attributed to the low confidence of the Mixtral8x7B model. For exact values of loss entropy, please refer to the supplementary material. We also compute model-wise entropy for all prompting techniques as shown in Fig. 3(B). It is crucial to note that GPT-3.5 and Qwen-72B did not perform well in terms of correctness evaluation but are much more confident about

their responses. While the responses from the two best performing models GPT-4 and Mixtral8x7B are more grounded which makes them more trustworthy to be used in a clinical setting.

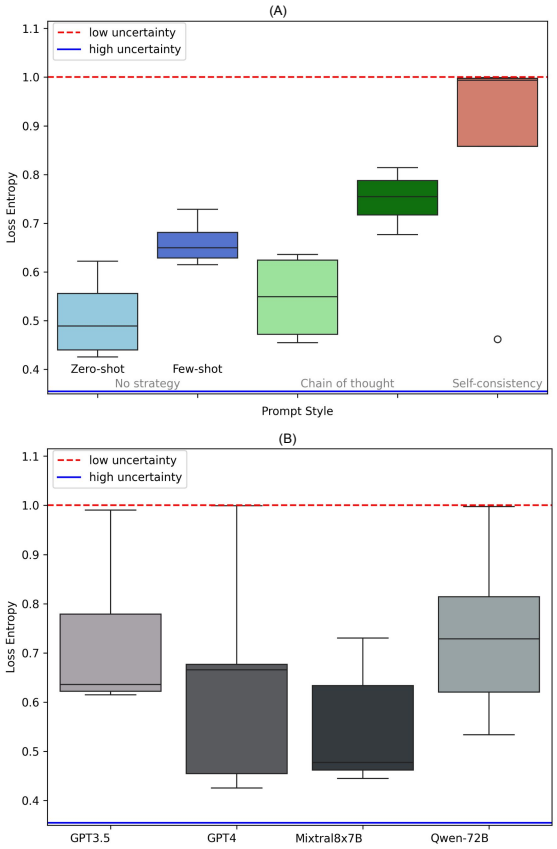

Figure 3. Confidence Measure Analysis: The higher the box-plot towards loss entropy=1, the more confidence in the output. Part (A) depicts the impact of prompt style on confidence. The dark blue and dark green show the few-shot and few-shot CoT prompt styles, respectively. Part (B) shows model-wise loss entropy result.

5.3. Clinical Evaluation

In a rigorous clinical evaluation, we validate the reasoning, completeness, and source citation accuracy of our best-performing language models, GPT-4 and Mixtral8x7B. The clinician (disclosure: one author of this paper) specializes in epileptology and is presented with the models' responses to a diverse set of fifty eight queries. The task is to assess the correctness and completeness of the provided reasoning on a three-point scale (absolutely correct/complete, somewhat correct/complete, and not correct/complete) as shown in Fig. 6. Additionally, two authors of the paper verify whether the sources cited by the models are correct or not. We only consider the source to be accurate if the author list, and exact title match verbatim.

GPT-4 cited sources correctly for 83.33% of the queries presented, while Mixtral8x7B achieves an average accuracy of 18.33% as the latter hallucinates its sources by combining authors and titles from multiple research works. We also computed Cohen's kappa to assess the consensus between the two evaluators. As shown in Fig. 4, for GPT-4 and Mixtral8x7B kappa values are 0.88 and 0.78 respectively. We also note that GPT-4 comes up with correct and complete reasoning for 55.15% and 63.79% of the examples respectively while Mixtral8x7B is somewhat correct for 44.82% of cases or complete in its reasoning for 36.20% of the examples given as query (Refer Fig. 5).

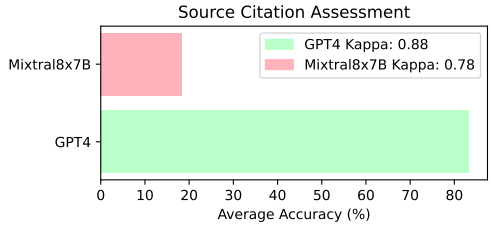

Figure 4. Source citation assessment by two evaluators

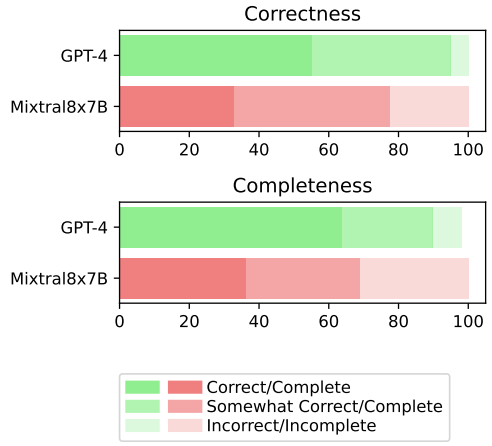

Figure 5. Clinical evaluation of reasoning capabilities: (a) correctness and (b) completeness of reasoning provided by two best performing models

6. Conclusion

Our study provides the first investigation of whether current SOTA LLMs leverage their internal knowledge for epilepsy diagnosis. Specifically, we focus on a human-centric approach by emulating a characteristic use case in clinical practice, where clinicians evaluate the relationship between seizure symptoms and the likely underlying seizure onset zone. Importantly, we utilize a large-scale ground-truth annotated dataset and provide a first benchmark in this domain. We could show, that foundation large language models are able to correctly identify the Seizure Onset Zone (SOZ) significantly above chance from unstructured text descriptions of seizure behavior. Importantly, prompt engineering leads to substantial improvements in performance, achieving

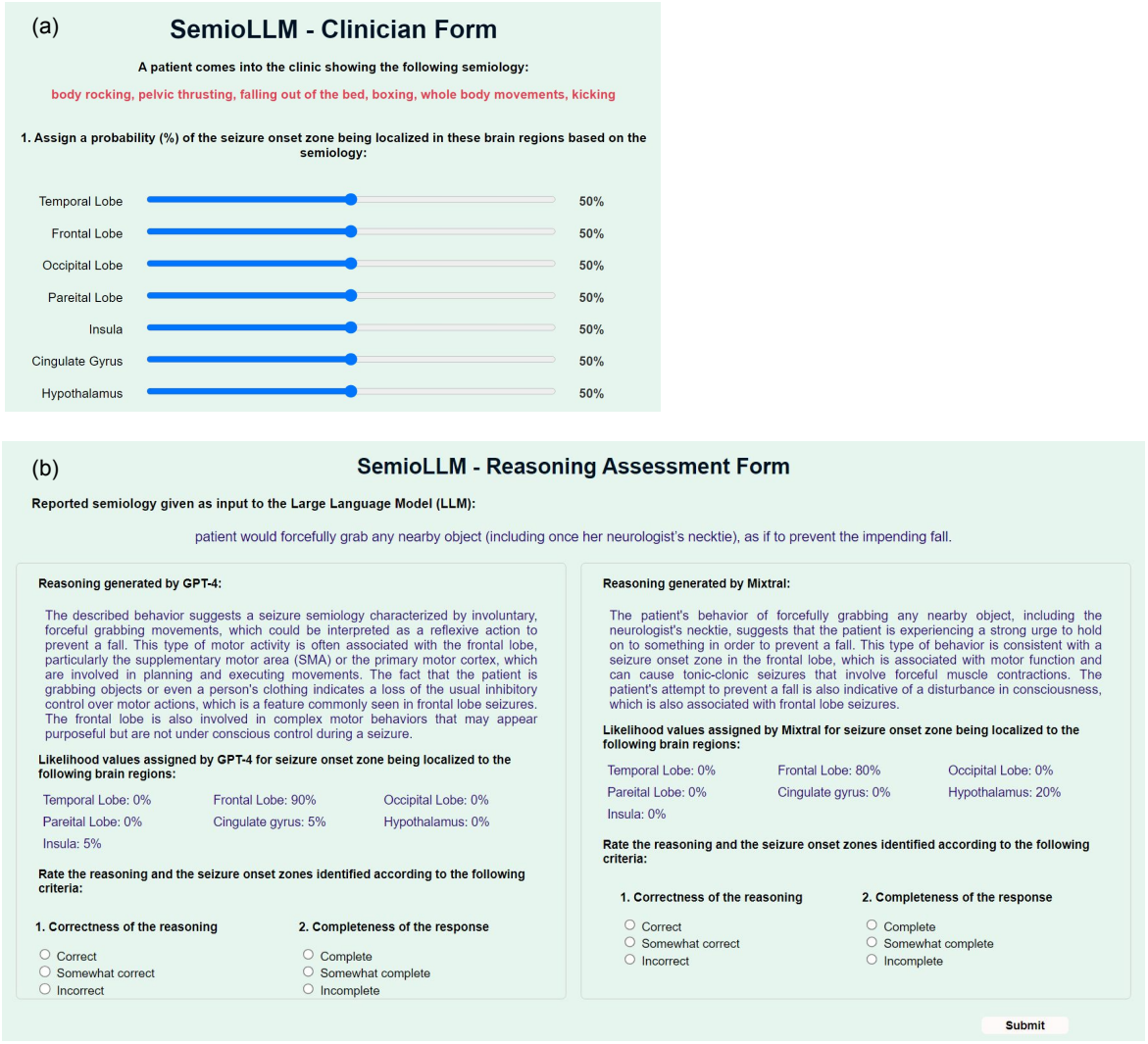

Figure 6. Clinical evaluation form illustration. (a) SOZ localization form for clinicians to provide likelihood estimates for 7 major brain regions based on 1269 semiological queries, similar to LLMs. (b) Snapshot of reasoning assessment form, where clinicians evaluate the correctness and completeness of reasoning provided by state-of-the-art LLMs for a subset of 58 semiological queries.

close to clinical performance in some models. We further provide a method to approximate uncertainty from the models' response distributions, which can be used to evaluate their confidence. This is important as models are required to perform correctly - and be confident about it. Overall our results suggest, that models achieve the highest performance and certainty using the self-consistency prompting technique. Comparing models, GPT-4 and Mixtral8x7B perform best, while being more grounded in their responses. GPT3.5 and Qwen 72, however, are more confident in their responses despite not fairing well in performance, which is not desirable. Lastly, we also asked the clinicians to rate the reasoning and assessed citations from these models. In source citation, both evaluators had a high consensus in responses with an agreement that Mixtral8x7B show hallucination tendencies and gives only partially correct and incomplete reasonings in the majority of the cases.

In summary, our analyses provide evidence for the potential applicability of LLMs outside of the vastly used use case of standard medical QA datsets. In the domain of epilepsy, we show their potential contribution to improving the efficiency of diagnostic processes and treatment planning. Our approach can be extended to other domains, where unstructured text in medical history provides medical diagnostic information that can be verified using ground-truth datasets. Overall, our work contributes to the ongoing development and refinement of reliable and trustworthy language models for domain-specific applications, facilitating a comprehensive understanding of the models' performance and their potential for real-world use cases in the medical domain.

7. Acknowledgments

We gratefully acknowledge the support received for this research. MD is supported by the Else Kröner Medical Scientists Kolleg Clinbrain: Artificial Intelligence for Clinical Brain Research. MD and ZA are supported by the ERC (853489-DEXIM). SL and MJP are supported by DFG, German Research Foundation - SPP 2241 - Project Number 520287829. MD, MJP and SL are supported by the Machine Learning Cluster of Excellence, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 39072764. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting MD. We extend our sincere gratitude to Prof. Dr. rer. nat. Jakob H. Macke and his lab members for their insightful discussions and feedback, with special appreciation to Jaivardhan Kapoor. We are also thankful to Dr. A. Sophia Koepke and Shyamgopal Karthik for their valuable review of the manuscript.

8. Author Contribution

MD: Conceptualization, Methodology, Validation, Formal Analysis, Data Curation, Visualization, Writing - Original Draft, Review and Editing. MJP: Methodology (creation of clinical study form), Formal Analysis (clinical response compilation and evaluation of source citation by models), Visualization, Writing - Review and Editing. ZA: Supervision, Funding acquisition SL: Conceptualization, Formal Analysis (additionally, clinical evaluation), Project Administration, Supervision, Funding acquisition, Writing - Review and Editing.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alim-Marvasti, A., Romagnoli, et al. Probabilistic landscape of seizure semiology localizing values. *Brain Communications*, 4(3):fcac130, 2022.

Beniczky, S., Tatum, W. O., Blumenfeld, H., Stefan, H., Mani, J., Maillard, L., Fahoum, F., Vinayan, K. P., Mayor, L. C., Vlachou, M., et al. Seizure semiology: Ilae glossary of terms and their significance. *Epileptic Disorders*, 24 (3):447–495, 2022.

Boßelmann, C. M., Leu, C., and Lal, D. Are ai language models such as chatgpt ready to improve the care of individuals with epilepsy? *Epilepsia*, 64(5):1195–1199, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Kim, H.-W., Shin, D.-H., Kim, J., Lee, G.-H., and 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.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

Liévin, V., Hother, C. E., Motzfeldt, A. G., and Winther, O. Can large language models reason about medical questions? *Patterns*, 2023.

Lüders, H. O., Najm, I., Nair, D., Widdess-Walsh, P., and Bingman, W. The epileptogenic zone: general principles. *Epileptic disorders*, 8:S1–S9, 2006.

Sarvari, P., Al-fagih, Z., Ghuwel, A., and Al-fagih, O. A systematic evaluation of the performance of gpt-4 and palm2 to diagnose comorbidities in mimic-iv patients. *Health Care Science*, 2024.

Savage, T., Nayak, A., Gallo, R., Rangan, E., and Chen, J. H. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20, 2024.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Sisodiya, S. M. and Goldstein, D. B. Drug resistance in epilepsy: more twists in the tale. *Epilepsia*, 48(12):2369–2370, 2007.

Thijs, R. D., Surges, R., O'Brien, T. J., and Sander, J. W. Epilepsy in adults. *The lancet*, 393(10172):689–701, 2019.

Tibshirani, R. J. and Efron, B. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1):1–436, 1993.

van Diessen, E., van Amerongen, R. A., Zijlmans, M., and Otte, W. M. Potential merits and flaws of large language models in epilepsy care: A critical review. *Epilepsia*, 2024.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wiebe, S., Blume, W. T., Girvin, J. P., and Eliasziw, M. A randomized, controlled trial of surgery for temporal-lobe epilepsy. *New England Journal of Medicine*, 345(5): 311–318, 2001.

Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., Fonseca, E. K. U. N., Lee, H. M. H., Abad, Z. S. H., Ng, A. Y., et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.

A. Additional Dataset Details

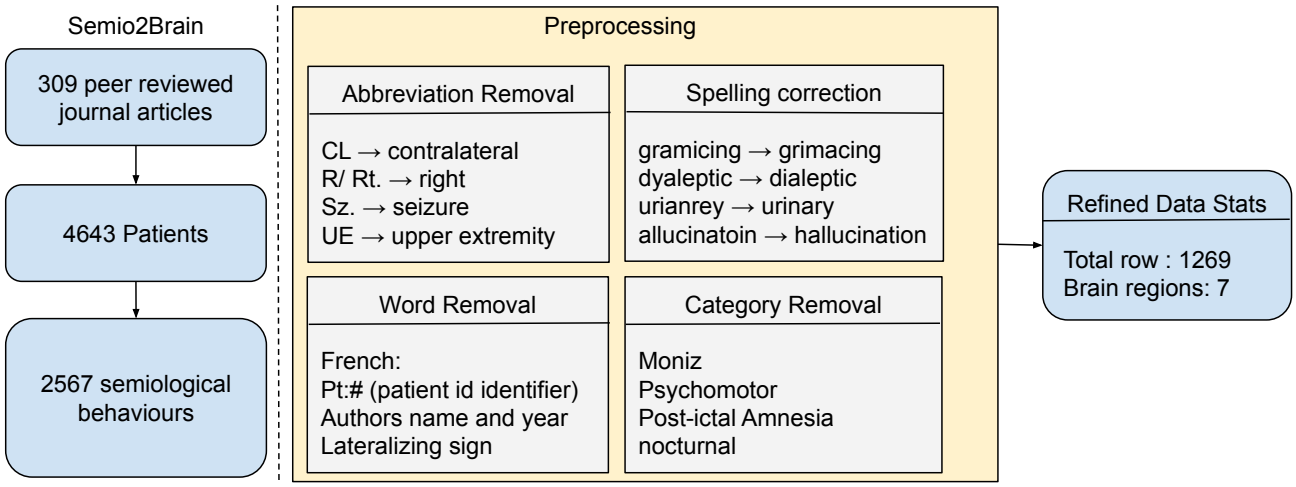

Figure 7. Data preprocessing pipeline: We use Semio2Brain (Alim-Marvasti et al., 2022) dataset which is a collection of 2567 semiologies spread across 7 major brain regions. Steps in preprocessing this data include abbreviation removal and replacing them with their respective full-forms, correction of spelling errors present in the data, removing uninformative words and semiology categories. This result in overall 1269 rows we finally use for our analysis.

In order to efficiently utilize the Semio2Brain dataset for our study, we employ several pre-processing steps. The semiological descriptions in the original dataset were classified into 43 semiological categories (36 ictal, 7 postictal and 1 absence of reported semiology) using a dictionary of regular expressions called SemioDict. The dictionary with accompanying descriptions of categories was made publicly available with the database by the curators. We add additional categories like Simple Motor and Aura to the original SemioDict and remove a few categories based on clinician's suggestion. We further retain only the semiologies belonging to ictal categories and with datapoints from patients who achieved post-operative seizure freedom. Using regular expressions, we replace abbreviations with full forms, remove spelling errors and eliminate keywords defining seizure type. The final version of the dataset includes 1269 reported semiologies spread across 36 ictal categories. We calculate the likelihood values for the seizure onset zone being localized to a brain region for each semiology by normalization (dividing each datapoint by the sum of datapoints across the 7 brain regions), followed by conversion to percentage values. We intend to release the data curation code in future, making it available for public use and further research in this domain.

B. Lower bound computation

The chance baseline for a multi-label classification problem is the ratio of positive instances to the total number of instances. Our data consists of 1269 rows and imbalanced class ratio with support for each class as $Support = 682, 398, 35, 78, 49, 18, 9$. Thus random precision and recall for each class is given as:

$$Prec_i \ , \ Rec_i = \frac{Support_i}{\sum_{i=1}^{7} Support_i} \tag{7}$$

and weighted precision and recall is given as:

$$Prec_{wtd} \ , \ Rec_{wtd} = \frac{\sum_{i=1}^{7} Prec_i * Support_i}{\sum_{i=1}^{7} Support_i} \tag{8}$$

Using this equation we get weighted precision and recall value. We then compute F1 Score, which is the harmonic mean of precision and recall to obtain the random chance lower bound of 39.34.

C. Additional Results

C.1. Bootstrapping for correctness measure

To perform bootstrapping (Tibshirani & Efron, 1993), we resample 10% of the full dataset 999 times to get a distribution of F1 scores to estimate the possible variability in the output. We report the mean and standard deviation for each model in Tab. 2 below:

LLM	Prompt Strategy				
	Zero-Shot	Few-Shot	CoT	FewShot-CoT	SC
GPT-3.5	38.19±0.13	48.15±3.38	49.14±3.72	51.52±3.67	51.14±3.71
GPT-4.0	**52.15±3.51**	50.60±3.59	**52.33±3.58**	51.81±3.79	**53.57±3.58**
Mixtral-8x7B	22.98±2.81	41.95±3.94	48.68±3.52	**52.64±3.69**	51.54±3.58
Qwen-72B	39.24±1.24	44.54±3.97	45.35±3.11	47.87±3.59	45.62±3.08
Clinician	48.83±3.85				

Table 2. Localization results on full data with boostrapping. We report mean±standard deviation of the F1 scores for bootstrapped distribution for combination of each model and prompt style. CoT and SC are the abbreviations for Chain-of-Thought and Self-Consistency prompting techniques respectively

C.2. Comparison of correctness measure with random distribution

LLM	Prompt Strategy	Mean	S.D.	z-score	p-value
GPT-3.5	Zero-shot	38.28	0.13	-1.08	0.27
	Few-shot	42.77	1.15	4.86	1.17e-06
	CoT	42.39	1.13	6.09	1.06e-09
	Few-shot CoT	40.89	1.24	8.61	7.16e-18
	SC	42.18	1.06	8.60	7.70e-18
GPT-4	Zero-shot	41.59	1.16	9.19	3.68e-20
	Few-shot	39.20	1.29	9.07	1.18e-19
	CoT	41.32	1.19	9.43	3.89e-21
	Few-shot CoT	39.30	1.33	9.58	9.26e-22
	SC	40.71	1.15	11.34	7.99e-30
Mixtral8x7B	Zero-shot	26.40	0.77	-4.28	1.82e-05
	Few-shot	36:60	1.41	4.01	5.99e-05
	CoT	42.26	1.13	5.92	3.06e-09
	Few-shot CoT	42.04	1.23	8.66	4.33e-18
	SC	42.03	1.14	8.43	3.30e-17
Qwen-72B	Zero-shot	38.49	0.28	2.54	0.01
	Few-shot	38.35	1.18	5.29	1.18e-07
	CoT	40.73	0.84	5.62	1.90e-08
	Few-shot CoT	41.18	1.11	6.25	3.87e-10
	SC	40.76	0.85	5.84	5.10e-09

Table 3. Random bootstrapping distribution statistics. We report mean, standard deviation, z-score obtained with respect to actual F1 score for each model and prompt strategy and finally p-value.

C.3. Loss entropy results

LLM	Prompt Strategy	Mean	S.D.	Entropy (Norm)	Loss Entropy
GPT-3.5	Zero-shot	1.06	0.32	0.37	0.62
	Few-shot	1.08	0.56	0.38	0.61
	CoT	1.02	0.63	0.36	0.63
	Few-shot CoT	0.62	0.54	0.22	0.77
	SC	0.02	0.16	0.01	0.99
GPT-4	Zero-shot	1.61	0.71	0.57	0.42
	Few-shot	0.93	0.63	0.33	0.66
	CoT	1.52	0.50	0.54	0.45
	Few-shot CoT	0.90	0.50	0.32	0.67
	SC	0.00	0.05	0.00	0.99
Mixtral8x7B	Zero-shot	1.55	0.43	0.55	0.44
	Few-shot	1.02	0.61	0.36	0.63
	CoT	1.46	0.53	0.52	0.47
	Few-shot CoT	0.75	0.69	0.26	0.73
	SC	1.51	0.53	0.53	0.46
Qwen-72B	Zero-shot	1.30	0.45	0.46	0.53
	Few-shot	0.76	0.70	0.27	0.72
	CoT	1.06	0.69	0.37	0.62
	Few-shot CoT	0.52	0.56	0.18	0.81
	SC	0.00	0.08	0.00	0.99

Table 4. Uncertainty measure results including mean, standard deviation (S.D.), normalized entropy value with minimum(0) and maximum (2.807) shannon entropy and loss entropy

C.4. Clinical evaluation results

LLM	Evaluator 1 Score	Evaluator 2 Score	Average Score	Cohen's Kappa
GPT4	83.33	83.33	83.33	0.88
Mixtral	21.66	15.00	18.33	0.78

Table 5. Lay User Evaluation of Source Correctness. This table presents the evaluation of the correctness of sources cited by the models in their responses. Two evaluators assessed the accuracy of the title and author list of the cited publications. The table reports the individual scores from each evaluator, the average score, and Cohen's kappa coefficient, which measures the inter-rater agreement between the evaluators

LLM	Correctness Measure			Completeness Measure		
	Correct	Somewhat Correct	Incorrect	Complete	Somewhat Complete	Incomplete
GPT-4	55.17	39.65	5.17	63.79	25.86	8.6
Mixtral8x7B	32.75	44.82	22.41	36.20	32.75	31.03

Table 6. Clinical evaluation result on reasoning correctness and completeness

D. User Study and Reasoning Output

We introduce clinicians in the benchmarking pipeline using two online surveys. In the first survey, we ask clinicians to assign likelihood values for SOZ being localized to the 7 brain regions for 58 semiologies chosen from diverse semiological categories. The form consists of the semiology and sliders next to the seven brain regions for clinicians to choose from a continuous range (0-100%) of likelihood values (Fig). This layout ensures similarity with the model's output format. We utilize a threshold (50%) to obtain the clinician's prediction of SOZ for comparison with the ground truth. We assess the relative performance of clinicians and models providing insights about the reliability of using these models in a clinical setting. A major advantage of LLMs is their capacity to generate reasoning that leads to seizure foci predictions. We ask clinicians to evaluate the reasoning generated by the models in the second survey. The form consists of the semiology, reasoning and likelihood values from the best-performing models (GPT-4 and Mixtral8x7B), along with questions to score correctness and completeness across three levels (Fig. 6). This enables us to effectively quantify the utility of long-form reasoning generated by models in epilepsy diagnosis.

On the next page, in Tab. 7 we show snippets of the output obtained from GPT-4 model for few-shot CoT prompt strategy on two semiologies. Each output contains three results including 'Reasoning', 'Likelihood' and 'Source(s)'.

1. Semiology:
he had hypertony of all four limbs
Reasoning:
The symptom described as ”hypertony of all four limbs” refers to increased muscle tone affecting the arms and legs. This increased muscle tone, or hypertonia, can be a feature of seizures originating in several different brain regions. However, the fact that all four limbs are involved suggests a more generalized seizure activity or a seizure onset that quickly involves bilateral motor pathways. The frontal lobe is a common site for seizures that result in tonic or clonic movements, as it contains the primary motor cortex which controls voluntary muscle movements. Seizures originating in the frontal lobe can lead to bilateral tonic posturing due to the spread of seizure activity through motor pathways. While other regions like the parietal lobe can be involved in sensory aspects of seizures, the motor manifestations such as hypertonia are more typically associated with frontal lobe seizures. Given the lack of specificity in the symptom description, the likelihood percentages are estimates based on common seizure semiologies associated with these regions.
Likelihood values:
TL: 0% FL:80% Cing:10% PL:5% OL:0% Insula: 5% Hyp: 0%
Source(s):
Tatum, W. O. (2012). Ellen R. Grass Lecture: Extraordinary EEG. Neurodiagnostic Journal, 52(4), 333-353. doi: 10.1080/21646821.2012.11079899

---

2. Semiology:
visuospatial manifestation: the patient reported her impairment as being able to know “intellectually and visually” when a person or object is near, but sensing otherwise. She said this created a tension between her sensation and cognition. She further detailed a gap in her left visual field that was “constantly opening up”, forcing her to rationalize her position in space rather than rely on sensory input. Although uncertain for a while, the patient grew assured in the fact that this was not a visual issue because even with her eyes closed, she would struggle to locate her place in a room.
Reasoning:
The patient’s description indicates a complex visuospatial impairment that involves both visual processing and spatial awareness. The ability to k̈now intellectually and visuallyšuggests that the patient’s primary visual processing (i.e., the ability to see and recognize objects and people) is intact, but there is a disconnect with the spatial perception and sensation of proximity. The g̈ap in her left visual fieldsuggests a problem with the right visual field representation, which is processed by the occipital lobe, particularly the right side, given the contralateral nature of visual field representation in the brain. The fact that the patient experiences difficulty in locating her position in space even with her eyes closed points towards a deficit in spatial awareness and proprioception, which is often associated with the parietal lobe, particularly the right parietal lobe for spatial attention and representation. The tension between sensation and cognition and the need to rationalize position in space may also implicate the involvement of the posterior parietal cortex, which integrates sensory information and is involved in spatial cognition and navigation. Given the bilateral nature of the symptoms and the involvement of spatial awareness, it is likely that the seizure onset zone could be in the right parietal lobe or at the junction of the parietal and occipital lobes, possibly affecting the right hemisphere more due to the left visual field deficit described.
Likelihood values:
TL: 0% FL:0% Cing:0% PL:70% OL:30% Insula: 0% Hyp: 0%
Source(s):
Kolb, B., & Whishaw, I. Q. (1990). Fundamentals of human neuropsychology (3rd ed.). New York: W.H. Freeman and Company

Table 7. Examples of reasoning and responses generated by GPT-4