

LLM as Clinical Graph Structure Refiner: Enhancing Representation Learning in EEG Seizure Diagnosis

Lincan Li¹, Zheng Chen², Yushun Dong^{1*}

¹Department of Computer Science, Florida State University

²SANKEN, The University of Osaka

{ll24bb,yushun.dong}@fsu.edu, chenz@sanken.osaka-u.ac.jp

Abstract

Electroencephalogram (EEG) signals are vital for automated seizure detection, but their inherent noise makes robust representation learning challenging. Existing graph construction methods, whether correlation-based or learning-based, often generate redundant or irrelevant edges due to the noisy nature of EEG data. This significantly impairs the quality of graph representation and limits downstream task performance. Motivated by the remarkable reasoning and contextual understanding capabilities of large language models (LLMs), we explore the idea of using LLMs as graph edge refiners. Specifically, we propose a two-stage framework: we first verify that LLM-based edge refinement can effectively identify and remove redundant connections, leading to significant improvements in seizure detection accuracy and more meaningful graph structures. Building on this insight, we further develop a robust solution where the initial graph is constructed using a Transformer-based edge predictor and multilayer perceptron, assigning probability scores to potential edges and applying a threshold to determine their existence. The LLM then acts as an edge set refiner, making informed decisions based on both textual and statistical features of node pairs to validate the remaining connections. Extensive experiments on TUSZ dataset demonstrate that our LLM-refined graph learning framework not only enhances task performance but also yields cleaner and more interpretable graph representations.

1 Introduction

Accurate epileptic seizure detection is a critical task in both clinical and research fields. Epilepsy affects millions of people worldwide, and accurate identification of seizure events is essential for timely diagnosis, effective treatment, and improved patient quality of life [Begley *et al.*, 2022]. Electroencephalogram (EEG) signals are the primary data source for monitoring brain activities, playing a central role in epilepsy

diagnosis and treatment [Benbadis *et al.*, 2020]. EEG records electrical activity across multiple brain regions using an array of electrodes, providing valuable information about both normal and abnormal brain states. However, analyzing EEG data remains a significant challenge, as it is usually contaminated by noise, artifacts, and inter-patient variability [Saba-Sadiya *et al.*, 2021; Aldahr *et al.*, 2022]. Furthermore, seizure events vary greatly in duration, frequency, and spatial manifestation across individuals. Traditional EEG analysis methods, such as manual inspection or basic signal processing techniques, are time-consuming and prone to human error [Rani *et al.*, 2024; Li *et al.*, 2026]. While recent deep learning approaches have improved performance, they usually struggle to effectively model the rich semantic context and higher-order relationships between EEG channels. These limitations make it difficult to distinguish meaningful patterns from noise or redundancy, highlighting the need for more advanced techniques with stronger reasoning and contextual understanding capabilities, which large language models possess.

Specifically, existing graph learning-based EEG seizure detection methods mainly fall into two categories: those that construct graphs using pairwise correlation or similarity between EEG channels, and those that adopt dynamic or data-driven approaches to learn graph structures from data [Ho and Armanfard, 2023; Tang *et al.*, 2022; Tang *et al.*, 2023b]. In the former, the graph is typically predefined based on statistical measures such as normalized cross-correlation or mutual information between channel pairs, and then kept fixed throughout model training [Ho and Armanfard, 2023; Tang *et al.*, 2022]. In the latter, more recent works attempt to learn or adapt graph connectivity directly from the data using neural network-based mechanisms, such as attention modules or state space models [Tang *et al.*, 2023b; Dong and Kluger, 2023]. However, both categories suffer from the same fundamental limitation: the noisy and variable nature of EEG signals makes it difficult to construct accurate and robust graphs. Whether based on correlation, similarity, or data-driven learning, the resulting graphs often contain redundant, spurious, or even erroneous edges, which can mislead graph neural networks (GNNs) and degrade both representation learning and seizure detection performance. Although some methods have attempted to alleviate this issue, most still lack explicit and principled mechanisms to refine graph connectivity and remove task-irrelevant connections.

*Yushun Dong is the corresponding author.

In contrast, LLM-based graph structure learning offers powerful reasoning and contextual understanding, enabling more intelligent and interpretable edge refinement that addresses the core weaknesses of previous approaches.

Therefore, improving the quality of graph structures becomes essential for advancing EEG seizure detection. An effective graph should capture true neural interactions while suppressing the redundant or noisy connections that mislead graph representation learning [Yu *et al.*, 2026]. Nevertheless, most existing approaches lack the structural learning and semantic understanding abilities needed to refine these connections. Large language models (LLMs) have demonstrated remarkable capabilities in reasoning, contextual understanding, and integrating diverse information sources [Cui *et al.*, 2024]. These strengths make LLMs well-suited for graph edge refinement, as they can more effectively distinguish between meaningful and irrelevant connections based on both statistical and semantic evidence. By leveraging LLMs for graph structure learning, it becomes possible to construct graphs that are not only more accurate and robust, but also more interpretable and aligned with underlying physiological mechanisms. This motivates our work to explore LLM-based edge refinement as a promising direction for addressing the key limitations of existing methods in EEG seizure detection. Our main contributions are summarized in the following:

- We propose a novel framework that leverages large language models (LLMs) as tools for clinical graph structure refinement, overcoming the limitations of existing methods in clinical graph representation learning and contextual understanding.
- We establish a comprehensive benchmark for LLM-based clinical graph structure refinement, systematically evaluating a range of state-of-the-art general-purpose LLMs as structural judges. This benchmark reveals the varying reasoning capabilities, consistency, and robustness of different LLMs in clinical graph reasoning.
- Comprehensive experimental evaluations are conducted on TUSZ benchmark, demonstrating that our method significantly improves both seizure detection performance and the quality of learned graph representations.

The remainder of this article is organized as follows: Section 2 reviews related works, including graph learning-based EEG seizure detection, and LLM applications in epilepsy diagnosis and treatment. Section 3 introduces the proposed two-stage framework for graph structure learning and edge refinement. In Section 4, we conduct comprehensive experiments to evaluate seizure detection performance, and establish a benchmark for LLM-based clinical graph structure refinement. Finally, Section 5 concludes this paper.

2 Related Works

2.1 Graph Representation Learning for EEG Seizure Detection

Graph representation learning has emerged as a powerful approach for EEG-based seizure detection, as it enables the modeling of complex, non-Euclidean spatial dependencies

among EEG electrodes [Bhandari *et al.*, 2024]. Early research primarily adopted static graph representations, where the connectivity between electrodes is predefined according to physical locations or simple statistical criteria [Klepl *et al.*, 2024]. For example, [Zhao *et al.*, 2021] constructs static graphs by calculating Pearson correlation matrices for EEG channel pairs and applying thresholds to define edges between highly correlated channels. [Wang *et al.*, 2020] transforms EEG signals into the frequency domain using Fast Fourier Transform, treating each electrode as a node and establishing connections based on the Visibility Graph rule. Despite their simplicity, static graph-based approaches struggle to reflect the dynamic and evolving relationships among EEG channels, as they rely on fixed graph structures. This limitation has motivated the development of dynamic graph representation learning methods that more effectively capture the spatial-temporal nature of EEG networks.

Depending on how dynamic connectivity is defined and updated, existing dynamic graph learning strategies for EEG signals can be grouped into three categories: (i) Correlation-based Graph Construction, (ii) Learnable Graph Construction, and (iii) Generative Graph Construction. *Correlation-based methods* [Tang *et al.*, 2022; Tao *et al.*, 2022; Xu *et al.*, 2024] build dynamic graphs by computing channel correlations (e.g., Pearson or cross-correlation) over sliding time windows, but these only reflect surface-level statistical relationships and are easily affected by noise, resulting in unstable graph structures. *Learnable graph construction methods* [Huang *et al.*, 2025; Li *et al.*, 2024] enable adaptive graph connectivity by leveraging attention mechanisms or mask learning during training, which increases modeling flexibility. However, these methods typically lack explicit physiological grounding and treat graph evolution as a byproduct of optimization, making it difficult to accurately model the temporal dynamics of neural activity during seizures or to integrate domain knowledge about seizure propagation. *Generative graph construction methods* embed graph generation modules within neural networks, dynamically synthesizing graph structures from latent EEG representations. Notable examples include self-supervised learning [Ho and Armanfard, 2023; Weng *et al.*, 2024] and diffusion modeling approaches [Chen *et al.*, 2022; Shu *et al.*, 2024; Xu *et al.*, 2026], which aim to model the spatial and temporal dynamics of EEG networks. While such data-driven methods are highly flexible and adaptive, they often lack physical interpretability and physiological plausibility, and may produce neural connections that are not biologically meaningful. Moreover, these approaches frequently overlook the inherent directionality and spatiotemporal aspects of neural signal propagation in epileptic seizures, thus limiting their ability to fully capture the underlying pathological mechanisms.

2.2 LLM Applications in Epilepsy Diagnosis

Existing LLM-based EEG seizure diagnosis methods can be broadly categorized into two types. The first category is *LLM-enhanced Clinical Text Analysis*, which encompasses the majority of existing studies. These works require large amounts of natural language as training corpus, such as electronic medical records and expert diag-

nostic narratives. For instance, one early study leveraged LLMs to predict seizure recurrence risk in children by analyzing clinical notes [Beaulieu-Jones *et al.*, 2023]; subsequent research introduced specialized models to interpret seizure semiology text for epileptogenic zone localization [Yang *et al.*, 2024]; other efforts explored automatic classification of seizure types from narrative case reports [Kerr *et al.*, 2025], as well as the extraction of seizure frequency and related attributes from clinical documentation using fine-tuned language models [Abeyasinghe *et al.*, 2025]. More recent work assessed the performance of general-purpose LLMs like GPT series in distinguishing seizure types based on patient self-reported descriptions [Ford *et al.*, 2025; Mishra *et al.*, 2025]. These approaches, however, demand not only EEG sequence data but also extensive auxiliary datasets and detailed expert annotations, which pose significant challenges for clinical implementation. Collecting such data and granular labels is costly and time-consuming, and the resulting large input size can considerably increase algorithmic inference time.

The second category encompasses a smaller set of research, which directly focuses on **EEG signal modeling with large-scale language models**. In this line of work, researchers have proposed to use self-supervised pretraining [Jiang *et al.*, 2024b] to obtain EEG signal large models, developed neural tokenization methods to map EEG data into discrete embeddings compatible with LLMs [Kim *et al.*, 2024], and designed frameworks that integrate EEG feature extraction with prompt-based or token-based LLM inference [Jiang *et al.*, 2024a; Barmpas *et al.*, 2025]. While these approaches show promise, their main limitation is that the EEG data does not directly contribute to graph structure learning or optimizing the underlying graph representations. Instead, these methods primarily rely on tokenization and sequential modeling, missing opportunities to exploit the inherent spatial-temporal relationships within EEG signals.

3 Methodology

3.1 Problem Definition and Overall Framework

In this work, we propose a robust and interpretable approach for EEG-based seizure detection. The objective is to automatically determine whether a seizure event occurs within a signal window by capturing physiologically meaningful interactions among brain regions and filtering out noisy or redundant connections.

Figure 1 (left panel) shows our proposed two-stage LLM-based clinical graph generation and refinement framework. The first stage employs a Transformer-based edge predictor to construct the preliminary graph structure from EEG data, modeling pairwise dependencies among channels. The second stage utilizes LLMs to perform context-aware edge refinement, ensuring that the resulting graph is sparse, consistent, and neurophysiologically plausible. The right panel illustrates the complete clinical workflow for EEG seizure diagnosis, in which the refined EEG graph is subsequently processed by a graph representation learning model for downstream seizure detection task, providing both enhanced accuracy and improved interpretability.

3.2 Basic Graph Representation Learning using Transformer-based Edge Predictor

In the first stage of our framework, we aim to construct an initial graph that captures the potential relationships between EEG channels. Given the sequential EEG data from N channels, denoted as $\mathbf{X}_\tau = \{x_\tau^1, x_\tau^2, \dots, x_\tau^N\}$, where $x_\tau^i \in \mathbb{R}^\tau$ represents the EEG time series segment of channel i at time τ , we employ a Transformer encoder to extract informative feature representations for each channel.

Feature Encoding with Transformer: Each EEG channel’s time series x_τ^i is first passed through a Transformer encoder, which models both local and global temporal dependencies within the signal. The output is an embedding vector for each node (channel), denoted as:

$$h_\tau^i = \text{Transformer}(x_\tau^i), \quad h_\tau^i \in \mathbb{R}^d \quad (1)$$

where d is the dimension of the learned embedding.

Edge Probability Prediction: For each pair of channels (i, j) , we concatenate their embeddings h_τ^i and h_τ^j to form a joint representation z_τ^{ij} :

$$z_\tau^{ij} = [h_\tau^i; h_\tau^j] \in \mathbb{R}^{2d} \quad (2)$$

This concatenated feature is then fed into a two-layer MLP to predict the probability that an edge connection exists between channel i and channel j :

$$p_\tau^{ij} = \sigma(\mathbf{w}_2^\top \phi(\mathbf{w}_1 z_\tau^{ij} + \mathbf{b}_1) + b_2) \quad (3)$$

where $\sigma(\cdot)$ is the sigmoid activation function, $\phi(\cdot)$ is the ReLU activation, $\mathbf{w}_1$ and $\mathbf{w}_2$ are the weight matrices for the 1st and 2nd layers, $\mathbf{b}_1$ and b_2 are the corresponding biases, and $p_\tau^{ij} \in [0, 1]$ is the predicted edge existence probability between nodes i and j .

Initial Edge Set Selection: We apply a threshold φ to determine the presence of an edge:

$$e_\tau^{ij} = \begin{cases} p_\tau^{ij}, & \text{if } p_\tau^{ij} > \varphi \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

The resulting binary adjacency matrix $\mathbf{E}_\tau^{ini} = [e_\tau^{ij}]$ defines the initial edge set for the constructed graph.

In summary, the first stage generates a data-driven initial graph structure by leveraging the temporal modeling capabilities of Transformer and the discriminative power of MLP for edge prediction. The initial graph will be further optimized in the next stage using an LLM-based refinement mechanism.

3.3 LLM-based Graph Structure Refinement

The second stage of the proposed framework aims to further improve the quality and interpretability of EEG graph structure by refining the edge set generated in the first step using large language models. The motivation for this approach lies in the ability of LLMs to reason over both statistical and semantic information, allowing them to make informed decisions about whether an edge between two EEG channels is meaningful or redundant. This process addresses the limitations of existing methods, which typically lack explicit mechanisms for context-aware edge selection and noise reduction.

For each candidate edge (i, j) in the initial graph constructed in the first step, we design proprietary prompts to

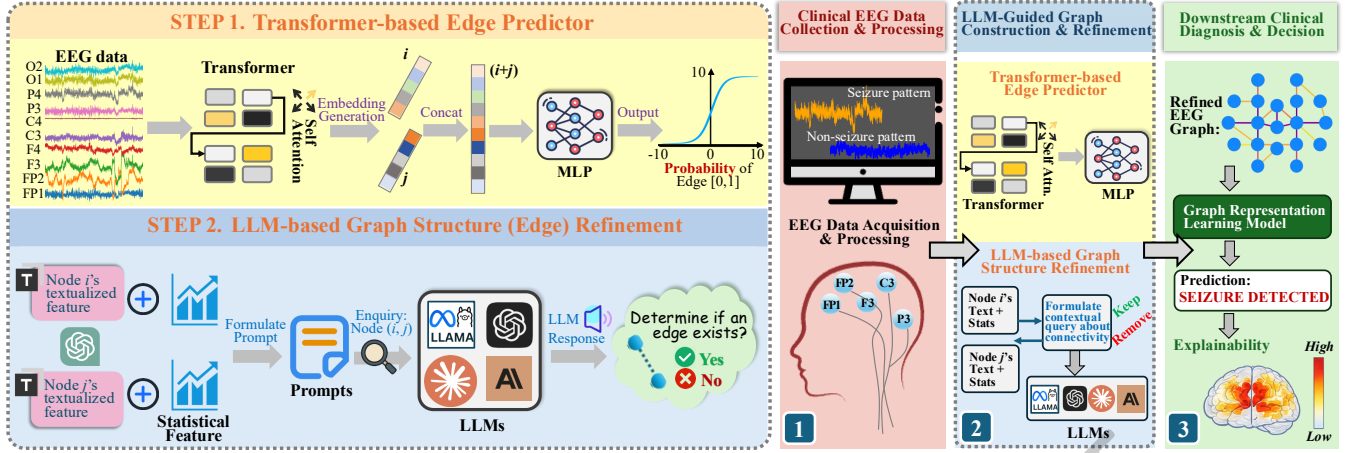


Figure 1: (Left) The proposed two-stage LLM-based clinical graph generation and refinement framework. (Right) The overall clinical EEG seizure diagnosis pipeline, in which the proposed framework corresponds to Step 2 for constructing and refining the EEG graph before downstream seizure prediction and interpretation.

query the LLM. These prompts incorporate both the textualized and statistical features of nodes i and j , ensuring that the model considers both context and data-driven evidence. The proprietary prompt consists of two parts: (a) the textualized description of the EEG data at the current time window using natural language. (b) summarized statistical features of the studied EEG data at the current time window.

Textualized EEG Feature Construction: For each node, we employ GPT-4o to describe its relevant characteristics of the EEG signal as natural language. For node i , the textualized feature includes: (a) The channel label (e.g., “F3”). (b) The anatomical location (e.g., “frontal lobe”). (c) High-level observations from the EEG data sequence (e.g., “shows sharp spikes”, “background is normal”), and the changing trend of the EEG data sequence. A sample textual description for node i is given here:

Example of Textualized EEG Feature Construction

Channel F3, located in the frontal lobe, shows moderate amplitude with intermittent sharp spikes. The EEG signal exhibits an increasing trend in amplitude over the observed time window.

EEG Statistical Feature Construction: Following the common practice in EEG signal statistics that are crucial and most discriminative for seizure detection [Zhang *et al.*, 2022; Shankar *et al.*, 2021; Sharmila and Geethanjali, 2020], we extract the the statistical features listed below using Python: (a) *Mean amplitude*: The average value of the EEG signal over the selected time window, reflecting the baseline activity level. (b) *Standard deviation*: The variability or dispersion of the EEG signal amplitudes, indicating the spreading of the signals. (c) *Dominant frequency*: The most prominent frequency component in the EEG signal, showing the primary oscillatory activity. (d) *Minimum*: The lowest amplitude observed in the EEG segment, capturing extreme negative fluctuations. (e) *Maximum*: The highest amplitude observed in the EEG segment, capturing extreme positive fluctuations. (f)

Median: The middle value of the sorted EEG amplitudes, representing the central tendency. (g) *(q25)*: The value below which 25% of the EEG data points fall, reflecting the lower distribution. (h) *(q75)*: The value below which 75% of the EEG data points fall, indicating the upper distribution. (i) *Skewness*: A measure of the asymmetry in the amplitude distribution of the EEG signal. (j) *Kurtosis*: A measure of the peakedness or flatness in the amplitude distribution of the EEG signal. (k) *Energy*: The total power of the EEG signal, calculated as the sum of squared amplitudes. (l) *Zero-crossing rate*: The number of times the EEG signal crosses the zero axis, often related to its frequency content.

These statistics are computed over the relevant EEG time window for each channel and presented in a standardized textual format. For example:

Example of EEG Statistical Feature Construction

Mean amplitude: 12.3 μV , Std: 4.5 μV , Dominant frequency: 8.5 Hz, Min: -6.2 μV , Max: 22.7 μV , Median: 10.1 μV , q25: 7.8 μV , q75: 14.5 μV , Skewness: 0.32, Kurtosis: 2.1, Energy: 1532.6, Zero-crossing rate: 42.

Prompt Design for Edge Refinement: The full prompt for a candidate edge (i, j) is constructed by concatenating the textual and statistical features of both nodes and then forming a specific inquiry. Suppose EEG channel F3 and T4 to be node i and j , respectively. Then the LLM-based edge refinement prompt is given in the following:

LLM-generated contextual prompt

Node i : Channel F3, located in the frontal lobe, shows moderate amplitude with intermittent sharp spikes. The EEG signal exhibits an increasing trend in amplitude over the observed time window. Mean amplitude: 12.3 μV , Std: 4.5 μV , Dominant frequency: 8.5 Hz, Min: -6.2 μV , Max: 22.7 μV , Median: 10.1 μV , q25: 7.8 μV , q75: 14.5 μV , Skewness: 0.32, Kurtosis: 2.1, Energy: 1532.6, Zero-

crossing rate: 42.

Node j : Channel T4, located in the temporal lobe, shows high amplitude with pronounced rhythmic slowing. The EEG signal displays a stable pattern with occasional bursts throughout the observed time window. Mean amplitude: $18.2 \mu V$, Std: $5.1 \mu V$, Dominant frequency: 4.2 Hz, Min: $-3.1 \mu V$, Max: $25.8 \mu V$, Median: $17.4 \mu V$, q25: $15.0 \mu V$, q75: $20.2 \mu V$, Skewness: -0.15, Kurtosis: 2.7, Energy: 2018.3, Zero-crossing rate: 33.

Question: Does a meaningful functional connection exist between Node i and Node j in this EEG segment? Answer yes or no, based on the context above.

This prompt is submitted to the LLM (e.g., GPT-5-mini), which returns a binary response ("yes" or "no") to indicate whether the edge should be retained. By iterating this process over all candidate edges, we obtain the final refined adjacency matrix where only edges validated by the LLM are preserved. This LLM-based refinement process enables more intelligent and interpretable graph construction, filtering out spurious and redundant edges based on a combination of contextual reasoning and data-driven evidence. Algorithm 1 illustrates the overall procedure for LLM-based edge refinement.

4 Experiments

In this section, we aim to answer the following research questions: **RQ1**: How do different graph construction and refinement strategies influence downstream EEG seizure detection performance? **RQ2**: How do various general-purpose LLMs perform as structural judges in graph reasoning? **RQ3**: What are the interpretability and consistency characteristics of LLM-based graph reasoning in clinical EEG contexts?

4.1 Dataset and Experimental Setup

Dataset. We conduct experiments on the publicly available EEG seizure diagnosis benchmark: Temple University Hospital EEG Seizure Corpus (TUSZ) v1.5.2 [Shah *et al.*, 2018]. TUSZ is currently one of the largest clinical EEG corpora for seizure diagnosis, comprising 5,612 EEG recordings and 3,050 seizure annotations. It includes 19 EEG channels recorded using the standard 10-20 system, covering a broad range of subjects across diverse clinical conditions. For interpretability analysis, we leverage the seizure onset-offset labels and event type annotations provided with TUSZ dataset.

Baseline Methods. To ensure a fair and comprehensive comparison, all models are built upon the same *modified GraphS4mer* backbone [Tang *et al.*, 2023a], where the original internal graph structure learning module is disabled and replaced by externally supplied graph structures. Only graph inputs vary; all downstream GraphS4mer components remain unchanged. We consider two categories of baselines: (1) *Conventional graph construction methods*, including (a) *Self-Correlation Graph*, which constructs graphs based on pairwise channel correlations; (b) *Distance-based Graph*, which connects channels according to their physical distance; (c) *Attention-based Graph*, which utilizes attention mechanisms to learn adaptive edge weights; (d) *K-Nearest Neighbor (KNN) Graph*, where each node is connected to its k most

similar channels based on signal similarity; (e) *Generative Module-based Graph*, which learns graph structures through generative neural network modules. (2) *LLM-based graph structure refinement methods*, where we benchmark a range of state-of-the-art general-purpose LLMs as structural judges, including Mistral 7B, Mistral 8x7B, Llama 3.1-70B, Llama 3.2-90B, Gemini-2.5-Pro, Gemini-2.5-Flash, GPT-4o, GPT-4.1, GPT-5, GPT5-mini. Each model independently refines the graph produced by the Transformer-based edge predictor.

Algorithm 1 Two-stage LLM-based Graph Structure Learning for EEG Seizure Detection

Require: Multichannel EEG data $\mathbf{X}_\tau = \{x_\tau^1, x_\tau^2, \dots, x_\tau^N\}$; threshold φ for edge selection; large language model $\mathcal{M}_{LLM}$;
Ensure: Final refined weighted adjacency matrix $\mathbf{E}_\tau^*$;
1: **Step 1: Initial Graph Construction**
2: **for** each channel i **do**
3: Obtain node embedding $h_\tau^i = \text{Transformer}(x_\tau^i)$
4: **end for**
5: **for** each channel pair (i, j) **do**
6: Concatenate embeddings: $z_\tau^{ij} = [h_\tau^i; h_\tau^j]$
7: Compute edge probability: $p_\tau^{ij} = \sigma(\text{MLP}(z_\tau^{ij}))$
8: **if** $p_\tau^{ij} > \varphi$ **then**
9: Set $e_\tau^{ij} = p_\tau^{ij}$ (edge weight in initial graph)
10: **else**
11: Set $e_\tau^{ij} = 0$ (no edge connection)
12: **end if**
13: **end for**
14: Obtain initial adjacency matrix $\mathbf{E}_\tau^{ini}$ from all e_τ^{ij}
15: **Step 2: LLM-based Edge Refinement**
16: Initialize $\mathbf{E}_\tau^* \leftarrow \mathbf{E}_\tau^{ini}$
17: **for** each node pair (i, j) with $e_\tau^{ij} > 0$ in $\mathbf{E}_\tau^{ini}$ **do**
18: Construct textualized and statistical features for node i and node j
19: Formulate prompt Q_τ^{ij} from features of node i and node j
20: Submit Q_τ^{ij} to $\mathcal{M}_{LLM}$ and obtain binary response R_τ^{ij}
21: **if** R_τ^{ij} is "yes" **then**
22: Set $e_\tau^{ij} = p_\tau^{ij}$
23: **else if** R_τ^{ij} is "no" **then**
24: Set $e_\tau^{ij} = 0$
25: **end if**
26: **end for**
27: **return** $\mathbf{E}_\tau^*$

4.2 Results and Performance Comparison

Experimental Setup. To evaluate the influence of different graph construction and refinement strategies on downstream seizure detection, we employ GraphS4mer as a unified backbone across all experiments, ensuring that performance variations arise solely from graph structural differences. The upstream graph inputs include two categories: (1) Non-LLM-based graph construction methods and (2) LLM-based graph structure refinement methods, where various LLMs serve as structural judges to refine the initial constructed graphs by Transformer model. The performance is evaluated using three widely adopted metrics: **F1-score**, **Accuracy**, and **Recall**.

Results and Analysis. Table 1 summarizes the seizure detection results across all baselines. Overall, models employing LLM-based graph refinement consistently outperform all

Model	F1	Accuracy	Recall
<i>None LLM-based Graph Construction Methods</i>			
Distance-based $\mathcal{G}$	0.6508	0.8225	0.7049
Self-Correlation $\mathcal{G}$	0.6682	0.8453	0.7158
KNN-based $\mathcal{G}$	0.7156	0.8762	0.7362
Attention-based $\mathcal{G}$	0.7338	0.8894	0.7537
Generative Graph Learner $\mathcal{G}$	0.7351	0.9039	0.7625
<i>LLM-based Graph Structure Refinement Methods</i>			
Mistral 7B + Transformer $\mathcal{G}$	0.7458	0.9087	0.7705
Mistral 8x7B + Transformer $\mathcal{G}$	0.7493	0.9056	0.7784
Llama 3.1-70B + Transformer $\mathcal{G}$	0.7522	0.9196	0.7786
Llama 3.2-90B + Transformer $\mathcal{G}$	0.7587	0.9173	0.7815
Gemini 2.5-Flash + Transformer $\mathcal{G}$	0.7613	0.9207	0.7836
Gemini 2.5-Pro + Transformer $\mathcal{G}$	0.7685	0.9261	0.7879
GPT-4o + Transformer $\mathcal{G}$	0.7726	0.9278	0.7947
GPT-4.1 + Transformer $\mathcal{G}$	0.7792	0.9236	0.7933
GPT-5-mini + Transformer $\mathcal{G}$	0.7859	0.9275	0.8024
GPT-5 + Transformer $\mathcal{G}$	0.7907	0.9315	0.8058

*All models share the same GraphS4mer backbone.

Table 1: Seizure detection performance under different graph construction / refinement methods.

non-LLM graph construction methods, demonstrating the effectiveness of leveraging language-driven reasoning to refine EEG connectivity structures. Among traditional approaches, Generative Graph Learner achieves the highest F1, indicating that learning-based graph formation offers moderate benefits over handcrafted correlations. In contrast, integrating different LLMs yields significant performance improvements, with GPT-5 + Transformer $\mathcal{G}$ achieving the best results (F1 = 0.7907, Accuracy = 0.9315, Recall = 0.8058). This performance gain highlights that larger, instruction-tuned models possess stronger reasoning and consistency when determining clinically meaningful connections between EEG channels. The steady improvement from smaller models (e.g., Mistral 7B) to larger ones (e.g., Llama 3.2-90B, GPT-5) further suggests that graph reasoning capability scales with model capacity, validating the efficacy of our proposed LLM-based graph refinement paradigm.

4.3 LLM Clinical Graph Reasoning Benchmark

Experimental Setup. To systematically evaluate the reasoning and structural refinement capabilities of different LLMs, we establish a unified benchmark named LLM Clinical Graph Refinement Benchmark. In this setting, each LLM independently serves as a structural judge that refines the graph generated by Transformer-based edge predictor under identical conditions, ensuring fairness and reproducibility. We benchmark a diverse set of representative models across different scales and architectures, including Mistral 7B, Mistral 8x7B, Llama 3.1-70B, Llama 3.2-90B, Gemini 2.5-Flash, Gemini 2.5-Pro, GPT-4o, GPT-4.1, GPT-5, and GPT-5-mini. Beyond the downstream seizure detection metrics (F1, Accuracy, and Recall), we further compute graph-level statistics: (I) **Graph Sparsity** [Cini *et al.*, 2023] and (II) **Jensen-Shannon Divergence(JSD)** [Wang and Yuan, 2025] to quantify each model’s reasoning consistency and structural diversity.

Results and Analysis. Table 2 reports the benchmark results across all models. Overall, the proposed benchmark re-

Model	Graph Sparsity	JSD
Mistral 7B + Transformer $\mathcal{G}$	0.4213	0.0576
Mistral 8x7B + Transformer $\mathcal{G}$	0.4167	0.0532
Llama 3.1-70B + Transformer $\mathcal{G}$	0.4019	0.0442
Llama 3.2-90B + Transformer $\mathcal{G}$	0.3984	0.0407
Gemini 2.5-Flash + Transformer $\mathcal{G}$	0.3931	0.0369
Gemini 2.5-Pro + Transformer $\mathcal{G}$	0.3875	0.0335
GPT-4o + Transformer $\mathcal{G}$	0.3826	0.0314
GPT-4.1 + Transformer $\mathcal{G}$	0.3778	0.0306
GPT-5-mini + Transformer $\mathcal{G}$	0.3758	0.0289
GPT-5 + Transformer $\mathcal{G}$	0.3741	0.0265

Table 2: Graph-level reasoning characteristics of different LLMs.

veals distinct reasoning behaviors and scaling trends among the evaluated LLMs. Smaller models such as Mistral 7B and Mistral 8x7B achieve moderate seizure detection accuracy but exhibit relatively high graph diversity, indicating less stable structural reasoning. In contrast, mid-sized open-source models like Llama 3.1-70B and Llama 3.2-90B improve both F1 and structural consistency, reflecting more coherent edge judgment. Gemini 2.5-Pro and GPT-series models demonstrate the strongest reasoning reliability and overall detection performance, with GPT-5 achieving the highest F1 score, edge agreement rate, and lowest divergence. These results suggest that reasoning capability in graph refinement scales with model capacity and training diversity. Moreover, we observe a clear trade-off: smaller models tend to explore more novel edge connections (higher diversity but lower stability), whereas larger models yield more clinically plausible and consistent graph structures. This benchmark thus provides a standardized lens through which to compare LLMs’ reasoning quality in clinical graph-based tasks.

4.4 Interpretability & Consistency Analysis

To further understand how different LLMs perform structural reasoning during EEG graph refinement, we conduct interpretability and consistency analysis. We visualize node-level importance distributions and assess the physiological plausibility of the refined graphs with respect to clinical EEG knowledge and ground-truth annotations.

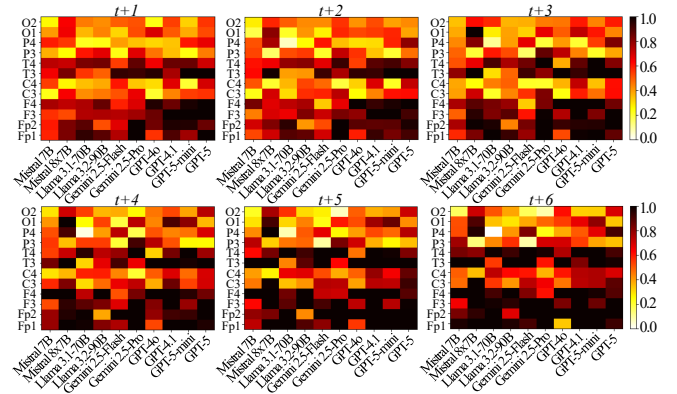


Figure 2: Node feature importance distributions across different LLM-based graph refiners in continuous time steps, illustrating seizure onset in frontal region and propagation toward temporal area.

Figure 2 illustrates the temporal evolution of node-level importance distributions across different LLM-based graph refiners during a seizure episode. As shown from $t+1$ to $t+6$, the activation pattern exhibits a physiologically consistent propagation from frontal (Fp1, Fp2, F3, F4) to temporal (T3, T4) regions, reflecting the typical onset and spread of epileptic activity. Among all models, the GPT family demonstrates more stable and spatially focused activations over key cortical areas, indicating stronger interpretability and better alignment with clinically meaningful EEG dynamics. In contrast, models such as Mistral and Llama display scattered and less consistent node activations. Notably, GPT-5 achieves the most coherent and biologically plausible feature patterns, followed by GPT-5-mini and GPT-4.1, corroborating the superior reasoning and refinement capability of advanced LLMs in capturing seizure-related brain connectivity.

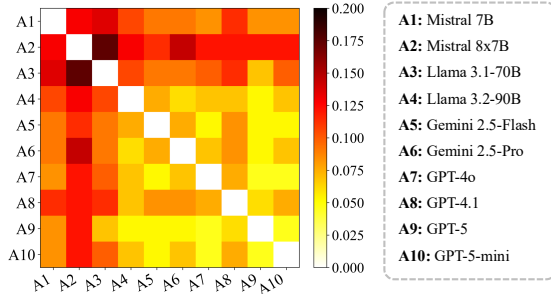


Figure 3: Edge Difference Heatmap. It shows the pairwise structural discrepancies among the refined graphs from different LLMs.

Figure 3 shows the pairwise edge difference heatmap of LLM-based graph refinement. Brighter regions (bottom-right) indicate higher structural consistency, while darker tones (upper-left) reveal greater divergence across models. Specifically, GPT model series (GPT-4o, GPT-4.1, GPT-5, GPT-5-mini) exhibit the highest internal consistency, as indicated by the lighter color in the lower-right corner, suggesting that these models generate highly similar and stable graph structures. This consistency implies that GPT models share a coherent reasoning pattern when inferring physiologically meaningful brain connectivity. In contrast, the Mistral and Llama series show substantially higher pairwise differences, reflecting less stable and more stochastic refinement behaviors. The Gemini models lie in between, demonstrating moderate consistency with the GPT family. Overall, the gradual reduction in structural discrepancy from Mistral to GPT-5 aligns with the progression of reasoning capability, indicating that more advanced LLMs achieve not only higher accuracy in edge refinement but also improved physiological plausibility in representing seizure-related brain networks.

Figure 4 presents a case study comparing the graphs refined by different LLMs across four time windows of a standard EEG signal sequence. The EEG signals exhibit a clear propagation pattern from frontal to temporal, central, and finally parietal-occipital regions, which reflects the typical spatial evolution of seizure activity. Consistent with this physiological trend, GPT series (GPT-4o and GPT-5) produce structured and spatially coherent graphs that capture this progres-

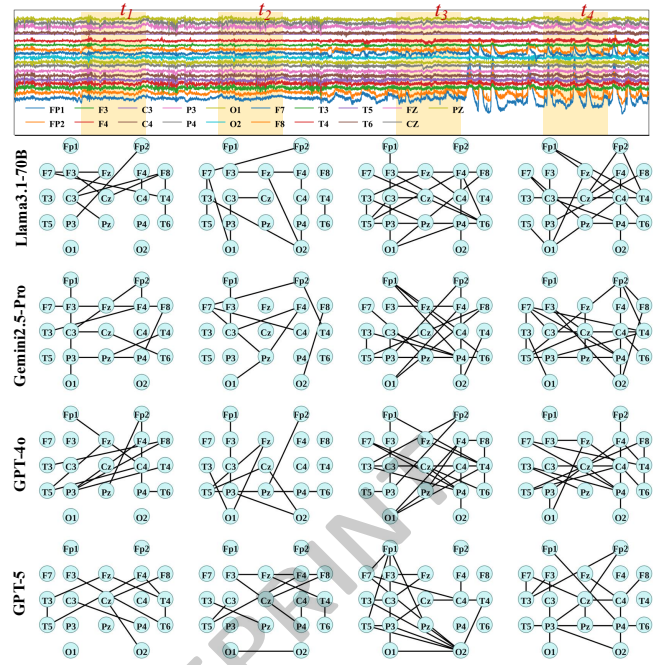


Figure 4: Visualization of the spatiotemporal evolution of LLM-guided EEG graphs.

sive propagation while maintaining sparse and physiologically plausible edge connectivity. Gemini 2.5-Pro demonstrates moderate consistency but introduces redundant inter-regional links, suggesting moderate precision in edge reasoning. Although graphs refined by Llama 3.1-70B reflect the seizure propagation pattern, it yields more dense and noisy connections. Overall, these results highlight that higher-capacity LLMs not only infer more accurate and stable graph topologies but also better align with the underlying neurophysiological dynamics observed in real EEG data.

5 Conclusions

In this paper, we propose a novel two-stage framework for EEG seizure detection that leverages the reasoning and contextual understanding capabilities of Large Language Models (LLMs) for clinical graph structure refinement. By integrating a Transformer-based edge predictor with an LLM-driven edge refinement module, our approach effectively identifies and removes redundant or irrelevant graph connections, improving graph quality and interpretability. We further establish an LLM-based clinical graph structure reasoning benchmark evaluating multiple general-purpose LLMs as graph structure judges. Extensive experiments on TUSZ benchmark demonstrate that our framework achieves superior seizure detection performance and generates more meaningful graph representations compared to existing methods. These findings highlight the potential of LLMs to bring new intelligence and interpretability to graph learning in neurophysiological signal analysis. Future work will explore domain-adaptive prompting and multimodal integration to further enhance interpretability, robustness, and generalization across broader biomedical graph learning tasks.

Acknowledgments

This material is, in part, supported by the National Science Foundation (NSF) under awards OAC-2530786 and GEO-2536578. We would also like to thank the anonymous reviewers for their constructive feedbacks.

References

- [Abeysinghe *et al.*, 2025] Rashmie Abeysinghe, Shiqiang Tao, Samden D Lhatoo, Guo-Qiang Zhang, and Licong Cui. Leveraging pretrained language models for seizure frequency extraction from epilepsy evaluation reports. *npj Digital Medicine*, 8(1):208, 2025.
- [Aldahr *et al.*, 2022] Raghdah Saem Aldahr, Munid Alanazi, and Mohammad Ilyas. Addressing inter-patient variability in eeg: Diversity-enhanced data augmentation and few-shot learning-based epilepsy detection. In *2022 International Conference on Healthcare Engineering (ICHE)*, pages 1–7, 2022.
- [Barmpas *et al.*, 2025] Konstantinos Barmpas, Na Lee, Alexandros Koliouisis, Yannis Panagakakis, Dimitrios A Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Neurorvq: Multi-scale eeg tokenization for generative large brainwave models. *arXiv preprint arXiv:2510.13068*, 2025.
- [Beaulieu-Jones *et al.*, 2023] Brett K Beaulieu-Jones, Mauricio F Villamar, Phil Scordis, Ana Paula Bartmann, Waqar Ali, Benjamin D Wissel, Emily Alsentzer, Johann de Jong, Arijit Patra, and Isaac Kohane. Predicting seizure recurrence after an initial seizure-like episode from routine clinical notes using large language models: a retrospective cohort study. *The Lancet Digital Health*, 5(12):e882–e894, 2023.
- [Begley *et al.*, 2022] Charles Begley, Ryan G Wagner, Annette Abraham, Ettore Beghi, Charles Newton, Churl-Su Kwon, David Labiner, and Andrea S Winkler. The global cost of epilepsy: a systematic review and extrapolation. *Epilepsia*, 63(4):892–903, 2022.
- [Benbadis *et al.*, 2020] Selim Benbadis, Sándor Beniczky, Edward Bertram, Stephanie MacIver, and Solomon L Moshé. The role of eeg in patients with suspected epilepsy. *Epileptic Disorders*, (2):143–155, 2020.
- [Bhandari *et al.*, 2024] Harish C Bhandari, Yagya R Pandeya, Kanhaiya Jha, Sudan Jha, and Sultan Ahmad. Exploring non-euclidean approaches: A comprehensive survey on graph-based techniques for eeg signal analysis. *Journal of Advances in Information Technology*, 15(10):1089–1105, 2024.
- [Chen *et al.*, 2022] Junru Chen, Yang Yang, Tao Yu, Yingying Fan, Xiaolong Mo, and Carl Yang. Brainnet: Epileptic wave detection from seeg with hierarchical graph diffusion learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2741–2751, 2022.
- [Cini *et al.*, 2023] Andrea Cini, Daniele Zambon, and Cesare Alippi. Sparse graph learning from spatiotemporal time series. *Journal of Machine Learning Research*, 24(242):1–36, 2023.
- [Cui *et al.*, 2024] Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-gpt: Towards a foundation model for eeg. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2024.
- [Dong and Kluger, 2023] Mingze Dong and Yuval Kluger. Towards understanding and reducing graph structural noise for gnns. In *International Conference on Machine Learning*, pages 8202–8226. PMLR, 2023.
- [Ford *et al.*, 2025] Joseph Ford, Nathan Pevy, Richard Grunewald, Stephen Howell, and Markus Reuber. Can artificial intelligence diagnose seizures based on patients’ descriptions? a study of gpt-4. *Epilepsia*, 2025.
- [Ho and Armanfard, 2023] Thi Kieu Khanh Ho and Narges Armanfard. Self-supervised learning for anomalous channel detection in eeg graphs: Application to seizure analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 7866–7874, 2023.
- [Huang *et al.*, 2025] Zhentao Huang, Yuyao Yang, Yahong Ma, Qi Dong, Jianyun Su, Hangyu Shi, Shanwen Zhang, and Liangliang Hu. Eeg detection and recognition model for epilepsy based on dual attention mechanism. *Scientific Reports*, 15(1):9404, 2025.
- [Jiang *et al.*, 2024a] Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals. *arXiv preprint arXiv:2409.00101*, 2024.
- [Jiang *et al.*, 2024b] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.
- [Kerr *et al.*, 2025] Wesley T Kerr, Katherine N McFarlane, Gabriela Figueiredo Pucci, Danielle R Carns, Alex Israel, Lianne Vighetti, Page B Pennell, John M Stern, Zongqi Xia, and Yanshan Wang. Supervised machine learning compared to large language models for identifying functional seizures from medical records. *Epilepsia*, 66(4):1155–1164, 2025.
- [Kim *et al.*, 2024] Jonathan W Kim, Ahmed Alaa, and Danilo Bernardo. Eeg-gpt: exploring capabilities of large language models for eeg classification and interpretation. *arXiv preprint arXiv:2401.18006*, 2024.
- [Klepl *et al.*, 2024] Dominik Klepl, Min Wu, and Fei He. Graph neural network-based eeg classification: A survey. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:493–503, 2024.
- [Li *et al.*, 2024] Yang Li, Yang Yang, Qinghe Zheng, Yunxia Liu, Hongjun Wang, Shangling Song, and Penghui Zhao. Dynamical graph neural network with attention mechanism for epilepsy detection using single channel eeg. *Medical & Biological Engineering & Computing*, 62(1):307–326, 2024.

- [Li *et al.*, 2026] Lincan Li, Rikuto Kotoge, Xihao Piao, Zheng Chen, and Yushun Dong. Optimizing eeg graph structure for seizure detection: An information bottleneck and self-supervised learning approach. *arXiv preprint arXiv:2604.01595*, 2026.
- [Mishra *et al.*, 2025] Abhijit Mishra, Shreya Shukla, Jose Torres, Jacek Gwizdka, and Shounak Roychowdhury. Thought2text: text generation from eeg signal using large language models (llms). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3747–3759, 2025.
- [Rani *et al.*, 2024] N Geetha Rani, Nayakanti Praveen Kumar, et al. Attcnnet: Attention based cnn network to detect seizures from eeg subjects. In *2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 800–804, 2024.
- [Saba-Sadiya *et al.*, 2021] Sari Saba-Sadiya, Eric Chantland, Tuka Alhanai, Taosheng Liu, and Mohammad M Ghassemi. Unsupervised eeg artifact detection and correction. *Frontiers in digital health*, 2:608920, 2021.
- [Shah *et al.*, 2018] Vinit Shah, Eva Von Weltin, Silvia Lopez, James Riley McHugh, Lillian Veloso, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. The temple university hospital seizure detection corpus. *Frontiers in neuroinformatics*, 12:83, 2018.
- [Shankar *et al.*, 2021] A. Shankar, S. Dandapat, and S. Barma. Classification of seizure types based on statistical variants and machine learning. *2021 IEEE 18th India Council International Conference*, pages 1–6, 2021.
- [Sharmila and Geethanjali, 2020] A. Sharmila and P. Geethanjali. Evaluation of time domain features on detection of epileptic seizure from eeg signals. *Health and Technology*, 10:711–722, 2020.
- [Shu *et al.*, 2024] Kai Shu, Le Wu, Yuchang Zhao, Aiping Liu, Ruobing Qian, and Xun Chen. Data augmentation for seizure prediction with generative diffusion model. *IEEE Transactions on Cognitive and Developmental Systems*, 2024.
- [Tang *et al.*, 2022] Siyi Tang, Jared Dunnmon, Khaled Kamal Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel Rubin, and Christopher Lee-Messer. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. In *International Conference on Learning Representations*, 2022.
- [Tang *et al.*, 2023a] Siyi Tang, Jared Dunnmon, Liangqiong Qu, et al. Modeling multivariate biosignals with graph neural networks and structured state space. In *ICLR 2023 Workshop on Time Series Representation Learning for Health*, 2023.
- [Tang *et al.*, 2023b] Siyi Tang, Jared A Dunnmon, Qu Liangqiong, Khaled K Saab, Tina Baykaner, Christopher Lee-Messer, and Daniel L Rubin. Modeling multivariate biosignals with graph neural networks and structured state space models. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 50–71, 2023.
- [Tao *et al.*, 2022] Tian-li Tao, Liang-hu Guo, Qiang He, Han Zhang, and Lin Xu. Seizure detection by brain-connectivity analysis using dynamic graph isomorphism network. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2302–2305, 2022.
- [Wang and Yuan, 2025] Zuo Wang and Ye Yuan. Jensen-shannon divergence message-passing for rich-text graph representation learning. *arXiv preprint arXiv:2512.20094*, 2025.
- [Wang *et al.*, 2020] Jialin Wang, Shen Liang, Dake He, Ye Wang, Yingpei Wu, and Yanchun Zhang. A sequential graph convolutional network with frequency-domain complex network of eeg signals for epilepsy detection. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 785–792. IEEE, 2020.
- [Weng *et al.*, 2024] Weining Weng, Yang Gu, Shuai Guo, Yuan Ma, Zhaohua Yang, Yuchen Liu, and Yiqiang Chen. Self-supervised learning for electroencephalogram: A systematic survey. *arXiv preprint arXiv:2401.05446*, 2024.
- [Xu *et al.*, 2024] Jie Xu, Shasha Yuan, Junliang Shang, Juan Wang, Kuiting Yan, and Yankai Yang. Spatiotemporal network based on gcn and bigru for seizure detection. *IEEE Journal of Biomedical and Health Informatics*, 28(4):2037–2046, 2024.
- [Xu *et al.*, 2026] Rongchao Xu, Kunlin Cai, Lin Jiang, Zhiqing Hong, Yuan Tian, and Guang Wang. Geogen: A two-stage coarse-to-fine framework for fine-grained synthetic location-based social network trajectory generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.
- [Yang *et al.*, 2024] Shihao Yang, Yaxi Luo, Meng Jiao, Neel Fotedar, Vikram R Rao, Xinglong Ju, Shasha Wu, Xiaochen Xian, Hai Sun, Ioannis Karakis, et al. Episemollm: A fine-tuned large language model for epileptogenic zone localization based on seizure semiology with a performance comparable to epileptologists. *MedRxiv*, pages 2024–09, 2024.
- [Yu *et al.*, 2026] Dahai Yu, Lin Jiang, Rongchao Xu, and Guang Wang. Healthmamba: An uncertainty-aware spatiotemporal graph state space model for effective and reliable healthcare facility visit prediction. *arXiv preprint arXiv:2602.05286*, 2026.
- [Zhang *et al.*, 2022] Shang Zhang, Guangda Liu, Ruolan Xiao, Wenjie Cui, Jing Cai, Xinlei Hu, Yubing Sun, Jiqing Qiu, and Yuan Qi. A combination of statistical parameters for epileptic seizure detection and classification using vmd and nltwsvm. *Biocybernetics and Biomedical Engineering*, 2022.
- [Zhao *et al.*, 2021] Yanna Zhao, Changxu Dong, Gaobo Zhang, Yaru Wang, Xin Chen, Weikuan Jia, Qi Yuan, Fangzhou Xu, and Yuanjie Zheng. Eeg-based seizure detection using linear graph convolution network with focal loss. *Computer Methods and Programs in Biomedicine*, 208:106277, 2021.