BrainEC-LLM: Brain Effective Connectivity Estimation via Multiscale Mixing LLM

Wen Xiong, Junzhong Ji, Jinduo Liu *
Beijing University of Technology
xiongwen@emails.bjut.edu.cn, {jinduo, jjz01}@bjut.edu.cn

Abstract

Pre-trained Large language models (LLMs) have shown impressive advancements in functional magnetic resonance imaging (fMRI) analysis and causal discovery. Considering the unique nature of the causal discovery field, which focuses on extracting causal graphs from observed data, research on LLMs in this field is still at an early exploratory stage. As a subfield of causal discovery, effective connectivity (EC) has received even less attention, and LLM-based approaches in EC remain unexplored. Existing LLM-based approaches for causal discovery typically rely on iterative querying to assess the causal influence between variable pairs, without any model adaptation or fine-tuning, making them ill-suited for handling the cross-modal gap and complex causal structures. To this end, we propose BrainEC-LLM, the first method to fine-tune LLMs for estimating brain EC from fMRI data. Specifically, multiscale decomposition mixing module decomposes fMRI time series data into short-term and long-term multiscale trends, then mixing them in bottom-up (fine to coarse) and top-down (coarse to fine) manner to extract multiscale temporal variations. And cross attention is applied with pre-trained word embeddings to ensure consistency between the fMRI input and pre-trained natural language. The experimental results on simulated and real resting-state fMRI datasets demonstrate that BrainEC-LLM can achieve superior performance when compared to state-of-the-art baselines. The code is available at https://github.com/XiongWenXww/BrainEC-LLM.

1 Introduction

Brain effective connectivity (EC) estimation has attracted significant scientific attention and has been widely used in clinical studies involving Alzheimer's disease [9, 65], schizophrenia [4], depression [52], and autism spectrum disorders [37, 70]. Based on application needs, the most commonly used neuroimaging modality is functional magnetic resonance imaging (fMRI).

In recent years, traditional machine learning (ML) and deep learning (DL) have made significant progress in estimating brain EC using fMRI data [31]. Traditional ML methods are often highly interpretable, and their relatively simple model structures lead to low computational complexity [61]. However, this simplicity also means that they require extensive feature engineering, have limited capacity to handle complex relationships, and often necessitate different algorithms for different problems [44]. In contrast, DL methods, with their more complex model structures, do not require manual feature extraction [57, 41]. They can automatically extract, process, and make decisions based on features from raw data [2, 8]. However, DL methods require separate training for specific tasks, a large amount of labeled data, and costly model training [33].

*Corresponding author: Jinduo Liu (jinduo@bjut.edu.cn)

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

While numerous studies have explored the application of LLMs to fMRI data, most focus on analysis tasks such as decoding [11]. Meanwhile, EC estimation is a causal discovery task that aims to infer a directed graph representing causal relationships between brain regions, and typically lacks ground truth. Existing approaches applying LLMs to causal discovery typically rely on iterative querying to assess the causal influence between variable pairs, without any model adaptation or fine-tuning, limiting their capacity to capture complex causal structures [78].

To address these challenges, we propose BrainEC-LLM, a novel approach that leverages LLMs for the first time to estimate brain EC from fMRI time series data through fine-tuning, rather than relying solely on inference. Specifically, prompts generation (PG) module produces prompts in the form of fMRI dataset description, prior knowledge, and task description to guide LLM, enhancing its ability to utilize temporal information and cross-brain connectivity from fMRI data. Next, to capture intricate fMRI multiscale temporal variations and reduce cross-modal disparities, we propose multiscale decomposition mixing (MDM) module. This module performs bottom-up (fine to coarse) and top-down (coarse to fine) mixing of short-term and long-term multiscale fMRI trends, and then aligns multiscale fMRI features with pre-trained word embeddings using cross attention. Then prompts are served as prefixes and concatenated with the multiscale fMRI embeddings before being delivered into LLM. Finally, the multiscale features output by the LLM are fed into multiscale reconstruction mixing (MRM) module to achieve the fusion of multiscale information, and the brain EC network is estimated using self attention. Our key contributions can be summarized as follows:

- We propose BrainEC-LLM, the first method to fine-tune LLMs for brain EC estimation, in contrast to prior inference-only causal discovery methods.
- We propose PG to generate task description, dataset description and prior knowledge, which allows LLM to comprehend temporal dependencies and cross-brain connectivity.
- We present MDM, which disentangles complex fMRI multiscale temporal variations from a new perspective, enhancing LLM's understanding of these dynamics.
- Extensive experiments on both simulated and real resting-state fMRI datasets demonstrate that BrainEC-LLM outperforms the current state-of-the-art baselines.

2 Related Works

2.1 Brain Effective Connectivity Methods

Brain effective connectivity (EC) estimation seeks to uncover causal graphs that characterize the influence patterns among different brain regions using fMRI data. Since real fMRI datasets lack ground truth EC, autoregressive models have been widely adopted in this field. These models exploit the temporal autocorrelation within fMRI time series to predict the data itself, allowing the model to learn more representative and informative connectivity patterns.

Existing EC estimation approaches can generally be categorized into two groups: traditional machine ML methods and DL methods. Traditional ML methods for estimating brain EC networks include DCM [7, 20], SEM [17] and GC [15]. These methods are advantageous for their interpretability. However, their model structures are typically predefined and sensitive to noise and high-dimensional data. In contrast, DL methods have shown greater capability in estimating brain EC networks, particularly when processing high-dimensional and complex data. Examples of these methods include ACOCTE [40], RL-EC [42], CR-VAE [36], MetaCAE [30], MetaRLEC [75] and CUTS+ [12]. While these models achieve good performance on brain EC estimation tasks, they require separate training for different tasks.

2.2 Large Language Models

Numerous research efforts have shown that pre-trained LLMs can be effectively adapted to unseen tasks through fine-tuning [14, 62, 74]. While some works raise concerns about their robustness [60, 73], others have demonstrated that, with proper encoding or architectural tuning, LLMs can achieve strong performance across diverse time series tasks [24, 32]. As a domain-specific branch of causal discovery, EC estimation aims to infer causal graphs that describe the directed influence among brain regions based on fMRI time series data. Although LLMs have been initially explored in

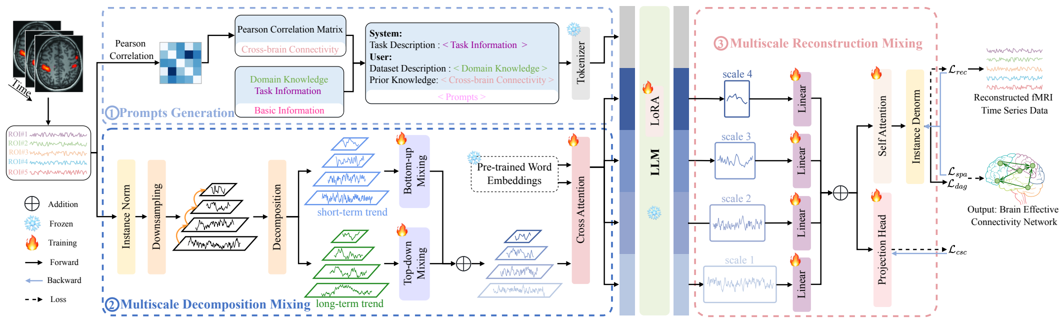


Figure 1: The overview of BrainEC-LLM framework. First, we enhance the inference capability of the LLM using ① prompts generation module. Next, we decomposition the complex variations in fMRI time series with ② multiscale decomposition mixing module, aligning pre-training word information of LLM with fMRI time series data. Finally, the multiscale features produced by LLM are combined across different scales using ③ multiscale reconstruction mixing module.

both fMRI analysis and causal discovery, their application to EC estimation for fMRI data has not yet been investigated.

fMRI Analysis. LLMs have been increasingly explored in the context of fMRI time series data, such as decoding linguistic information from brain signals [11, 25], reconstructing language from non-invasive brain recordings [68], and modeling the alignment between text and brain activity [19]. However, directly applying LLMs to fMRI time series poses challenges due to the complex spatial-temporal patterns and modality gap between fMRI time series and language representations.

Causal Discovery. LLMs have demonstrated strong capabilities for zero-shot inference of causal structures and identifying complex dependency relationships [5, 59, 47]. However, current approaches require LLMs to iteratively evaluate variable pairs for causal determination [72, 6], leading to computational inefficiency while failing to capture global dependencies and introducing model biases.

3 Notation and Problem Statement

In this paper, we utilize lowercase font (e.g., v_i) to indicate the (i-th) brain region, adopt uppercase font (e.g., A) to denote matrix, use uppercase calligraphic letters (e.g., $\mathcal{X}$) to represent three-dimensional tensor, and math bold italic letters (e.g., $\boldsymbol{Z}$) are employed to signify four-dimensional tensor.

The problem of brain EC estimation can be formulated as learning $\mathcal{G}$ from brain data (e.g., fMRI). We then introduce the definition of brain effective connectivity (EC). Brain EC can be represented as a directed graph $\mathcal{G} =< v, A >$, where v stands for the set of nodes, and each node $v_i \in v$ represents a brain region or region of interest (ROI). A denotes brian EC adjacency matrix, where A_{ij} symbolizes the effective connectivity from brain region v_i to v_j, indicating that v_i has a causal influence on v_j.

4 Methodology

BrainEC-LLM consists of three key components as shown in Figure 1. A prompt generation module first guides pre-trained LLMs (e.g., Llama 3, Mistral) by generating prompts tailored to the input fMRI time series. Then, a multiscale decomposition mixing module performs short-term and long-term mixing, followed by cross attention with pre-trained word embeddings for alignment. Finally, a reconstruction module fuses the multiscale features output by the LLM.

4.1 Prompts Generation

Prompts serve as a straightforward and effective technique for enhancing the reasoning capabilities of LLMs [69]. To harness the powerful capabilities of LLMs, we propose a comprehensive prompt generation module. First, the identity of LLM is established through system messages. Following this, a specific task description is provided, clearly outlining the objectives for the LLM. Next, the module describes the fMRI dataset using a user message. This description includes detailed information about the dataset, such as the generation and dimension of fMRI time series data. Additionally, the user message incorporates prior knowledge about the relationships across brain regions, derived from Pearson's correlation coefficient, helping the model focus on potentially relevant region pairs. Our prompt design is intended to provide the LLM with a structured representation of inter-regional brain relationships by explicitly encoding the Pearson correlation matrix, where each row and column corresponds to a specific brain region. To preserve a consistent and interpretable mapping between brain regions and their associated fMRI ROI time series, numerical indexing is incorporated into the prompt along with corresponding explanations, ensuring the model can clearly associate each index with a specific brain region. Examples of prompts and their mapping from Pearson correlation matrix to fMRI time series are provided in the Appendix A. The LLM tokenizer then converts the prompts above into a sequence of tokens, enabling LLM to process and understand the prompt efficiently.

4.2 Multiscale Decomposition Mixing

fMRI time series exhibit distinct patterns at different temporal sampling scales. For instance, high-resolution fMRI data recorded per second (lower-level scale) provide higher spatial and temporal resolution, enabling the capture of subtle changes and local neural dynamics. In contrast, low-resolution fMRI data recorded hourly (higher-level scale) tend to reflect the overall or global activity patterns of brain regions. These observations naturally motivate the adoption of a multiscale analysis paradigm to disentangle the complex temporal structures embedded in fMRI signals. A multi-scale perspective enables the model to separate and capture diverse components of brain activity across different timescales, which is crucial for accurately modeling temporal variations.

Specifically, for short-term trends, where fine-grained temporal fluctuations such as rapid neural responses are more prominent, we adopt a bottom-up mixing strategy that emphasizes detailed local information. Conversely, for long-term trends, which emphasize broader patterns like sustained functional connectivity, we employ a top-down mixing strategy to incorporate high-level global structures.

Decomposition. The fMRI time series is initially normalized using reversible instance normalization [34] to achieve zero mean and unit standard deviation, addressing the issue of distribution shift in fMRI time series data. Different phenomena and patterns may emerge at various time scales, with fine scales capturing detailed patterns and coarse scales emphasizing broader changes [46]. By integrating information from multiple scales, LLMs can more accurately predict behaviors or trends, reducing errors in the process. Concretely, we employ downsampling to decompose the complex patterns [51], resulting in M time series at different scales:

$$\mathcal{X}_{i+1} = AvgPool(\mathcal{X}_i), i \in \{1, ..., M-1\}, \tag{1}$$

where $\mathcal{X}_i \in \mathbb{R}^{\lfloor \frac{T}{2^{i-1}} \rfloor \times N}$ indicates the fMRI time series of i-th scale. The lowest level sequence $\mathcal{X}_0$ represents the input fMRI time series $\mathcal{X}$, containing subtle temporal variations, while the highest level sequence $\mathcal{X}_{M-1}$ captures macroscopic variations.

As illustrated in Figure 1, the multiscale fMRI time series are then decomposed to short-term multiscale trends $\mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_M\}$ and long-term multiscale trends $\mathcal{T} = \{\mathcal{T}_1, ..., \mathcal{T}_M\}$ using series decomposition block [64]:

$$\mathcal{T}_i = AvgPool(Padding(\mathcal{X}_i)), \mathcal{S}_i = \mathcal{X}_i - \mathcal{T}_i. \tag{2}$$

Subsequently, a feedforward network is applied to each trend independently. Due to the distinct variations contained in the short-term and long-term multiscale fMRI time series, they require separate processing to manage complex time variations.

Bottom-up Mixing. For short-term multiscale trends, lower-level detailed fMRI time series (e.g.$\mathcal{S}_1$) offers higher spatial and temporal resolution, capturing small changes and local dynamics in brain activity. In contrast, higher-level coarse fMRI time series (e.g.$\mathcal{S}_M$) integrate a broader range of brain activity, reflecting more holistic trends and patterns. Therefore, we adopt a bottom-up approach from lower-level fine-scale fMRI time series upward to provide additional information for higher-level coarser-scale fMRI time series:

$$\mathcal{S}_i = \mathcal{S}_i + \Phi(\mathcal{S}_{i-1}), \tag{3}$$

where $\Phi(\cdot)$ signifies ModernTCN block [43]. To better utilize temporal features and cross-brain connectivity in fMRI time series, we use a ModernTCN block to connect fMRI features across different scales, and the detailed framework of ModernTCN block is illustrated in Figure 2.

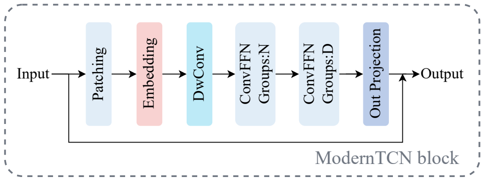


Figure 2: The architecture of ModernTCN block.

Unlike in computer vision, ModernTCN performs patch embedding [48] to preserve correlations among variables rather than embedding the N variables to dimension D. Denote patch length as P and stride as S. Through this patching process, we obtain patches of length P and the total number of patches $N_p = \lfloor \frac{T-P}{S} \rfloor + 2$. For the short-term fMRI time series trends at i-th scale, we obtain $\boldsymbol{P}_i \in \mathbb{R}^{N \times N_p \times P}$. Then, one-dimensional convolution is applied to these patches, resulting in $\boldsymbol{E}_i \in \mathbb{R}^{N \times N_p \times D}$. Next, ModernTCN uses depthwise convolution to extract the temporal dependencies of short-term fMRI patches, convolutional feed forward network (ConvFFN) with group set to N to extract feature representations of different brain regions, and ConvFFN with group set to D to extract correlations between brain regions. Among them, ConvFFN includes two pointwise convolutions with intermediate GELU activation function. Finally, these patches are projected and aligned with fMRI features at a different scale.

Top-down Mixing. For long-term multiscale trends, unlike short-term multiscale trends, small changes can introduce noise when capturing macro trends in brain activity, while higher-level coarse fMRI time series (e.g.$\mathcal{T}_M$) typically represent the overall activity of brain regions and are less affected by noise and short-term fluctuations. This relatively stable trend provides a solid foundation for lower-level fine-grained fMRI time series (e.g.$\mathcal{T}_1$), making detailed analysis at lower levels more reliable. Therefore, a top-down approach is adopted:

$$\mathcal{T}_i = \mathcal{T}_i + \Phi(\mathcal{T}_{i+1}). \tag{4}$$

We then obtained multiscale fMRI by summing the two multiscale fMRI trends.

Cross Attention. In order to adaptively extract local semantic information, we first perform patching on multiscale fMRI time series data to aggregate similar tokens and preserve the localization of fMRI time series. However, since pre-trained LLM is trained on textual data and struggles to interpret fMRI time series data, we apply cross attention between the obtained multiscale fMRI patches and the pre-trained word embeddings to align two modalities. Here, the pre-trained word embeddings refer to vector representations of all the words in the corpus that are used to train LLM. Given the vast size of corpus, directly using all these vector representations would introduce a significant amount of task-irrelevant information into the LLM. To address this issue, we employ a linear mapping matrix to reduce the vocabulary size, effectively merging related tokens. This transformation changes the dimension of the pre-trained word embeddings from $S \times D_{llm}$ (E) to $S' \times D_{llm}$ (E'), where S' is significantly smaller than S. Here, S represents the vocabulary size, and D_{llm} denotes the hidden dimensions of LLM model. Next, cross attention is applied and defined as:

$$\begin{aligned} \boldsymbol{Q}_i^h &= \boldsymbol{H}_i^h W_Q^h, K^h = E'^h W_K^h, V^h = E'^h W_V^h, \\ \boldsymbol{O}_i^h &= dropout(softmax(\frac{\boldsymbol{Q}_i^h {K^h}^\top}{\sqrt{d}}))V^h, \end{aligned} \tag{5}$$

where $\boldsymbol{Q}_i^h$ represents the fMRI patches at the i-th scale and the h-th head, $\boldsymbol{H}_i^h \in \mathbb{R}^{N \times N_p \times D}$, $W_Q^h \in \mathbb{R}^{D \times \frac{D}{H}}$ and W_K^h, $W_V^h \in \mathbb{R}^{D_{llm} \times \frac{D}{H}}$. By concatenating $\boldsymbol{O}_i^h$ across different heads, we derive $\boldsymbol{O}_i$. This result is then aligned with the hidden dimensions of the LLM model using a linear mapping, producing the final output $\boldsymbol{O}_i' \in \mathbb{R}^{N \times N_p \times D_{llm}}$.

This alignment of fMRI time series with natural language helps achieve more natural representations for the LLM. The resulting prompt embeddings as prefix, and multiscale fMRI embeddings are concatenated and serve as the input for the pre-trained LLM, which is then fine-tuned using LoRA.

4.3 Multiscale Reconstruction Mixing

As shown on the right side of Figure 1, we discard the prefix part to obtain multiple scales of fMRI time series data. Subsequently, linear mapping is performed to generate patches, followed by flattening and additional linear mapping to produce the predicted multiscale fMRI time series data $\mathcal{X}'$. We further split $\mathcal{X}'$ and express it in concatenation form:

$$\mathcal{X}' = \{\mathcal{X}_1', ..., \mathcal{X}_M'\}, \tag{6}$$

where $\mathcal{X}' \in \mathbb{R}^{M \times T \times N}$. To fully utilize the multiscale information, we first align time dimensions of fMRI time series data across multiple scales to T through linear mapping, and then summed them:

$$\mathcal{Y}'_i = \mathcal{W}_i \mathcal{X}'_i + bias_i, i = \{1, ..., M\}, Y' = \sum_{i=1}^{M-1} \mathcal{Y}'_i, \tag{7}$$

where $\mathcal{W}_i$ and $bias_i$ are learnable parameters. In the absence of ground truth in real fMRI datasets, we employ an autoregressive model that leverages autocorrelation (effective connectivity) to predict fMRI data itself. This strategy enables the model to learn more representative and informative connectivity structures. This is implemented using self attention, where the attention weights serve as brain EC, and the output of the self attention corresponds to the reconstructed fMRI time series data $Y \in \mathbb{R}^{T \times N}$ derived from the brain effective connectivity $A \in \mathbb{R}^{N \times N}$.

4.4 Overall Objective Function

Our final loss consists of three components: reconstruction loss $\mathcal{L}_{rec}$, sparsity loss $\mathcal{L}_{spa}$, directed acyclic loss $\mathcal{L}_{dag}$ and cross-scale contrastive loss $\mathcal{L}_{csc}$, which are as follows:

$$\mathcal{L} = \mathcal{L}_{rec} + \alpha_{spa}\mathcal{L}_{spa} + \alpha_{dag}\mathcal{L}_{dag} + \alpha_{csc}\mathcal{L}_{csc}, \tag{8}$$

where α_{spa}, α_{dag} and α_{csc} are weights coefficient.

The reconstruction loss, quantified as Euclidean distance between the reconstructed and actual fMRI time series data, reflects the accuracy of reconstruction. fMRI signals exhibit rich temporal autocorrelation and dynamic interactions between brain regions, both of which are indicative of the EC network. By training the model to reconstruct the BOLD signal (i.e., to predict fMRI data from itself), the model is encouraged to learn the causal relationships (effective connectivity) among brain regions, even without manual labels. Thus, minimizing the reconstruction loss between the predicted and actual fMRI signals enables model to discover effective connectivity in an unsupervised manner.

The sparsity loss is L1 regularization of Brain EC network A, which ensures brain EC derived from BrainEC-LLM is a sparse graph.

When brain EC network is directed acyclic graph (DAG), a directed acyclic loss [77] is necessary to constrain the estimated brain EC. This loss is optional and depends on whether the ground truth is DAG:

$$\mathcal{L}_{dag} = tr(\exp(A \odot A)) - N, \tag{9}$$

where $\odot$ indicates Hadamard product.

To maintain consistency in the information across different scales of fMRI time series, we employ cross-scale contrastive loss. Considering that in multiscale decomposition mixing module, the $(i+1)$-th scale is downsampled from i-th scale, maximizing their similarity (positive samples) ensures cross-scale consistency, while minimizing similarity with non-adjacent scales (negative samples) reduces irrelevant features. Prior to calculating cross-scale contrastive loss, a nonlinear projection head is required, which has proven to be effective [10]. Let $\mathbf{Z} \in \mathbb{R}^{B \times M \times N \times T}$ represent the multiscale fMRI time series data obtained after passing through the projection head. Then the contrastive loss is defined as follows:

$$\begin{gathered} g(\mathbf{Z}_i^b, \mathbf{Z}_{i+1}^b) = \exp(sim(\mathbf{Z}_i^b, \mathbf{Z}_{i+1}^b)/\tau), \\ \mathcal{L}_{csc}^{i,i+1} = \frac{1}{2B(M-1)} \sum_{b=1}^{B} \sum_{i=1}^{M-1} -\log \frac{g(\mathbf{Z}_i^b, \mathbf{Z}_{i+1}^b)}{\sum_{j \neq i} g(\mathbf{Z}_i^b, \mathbf{Z}_j^b)}, \\ \mathcal{L}_{csc} = \mathcal{L}_{csc}^{(i,i+1)} + \mathcal{L}_{csc}^{(i+1,i)}, \end{gathered} \tag{10}$$

where $\mathbf{Z}_i^b$ stands for the fMRI time series data at i-th scale of b-th sample, τ represents temperature coefficient and sim denotes cosine similarity.

Theorem 4.1. *The maximum lower bound for $\mathcal{L}_{csc}^{opt}$ is*

$$\mathcal{L}_{csc}^{opt} \geq 2\log(M-1) - (I(\mathbf{Z}_i^b, \mathbf{Z}_{i+1}^b) + I(\mathbf{Z}_{i+1}^b, \mathbf{Z}_i^b)), \tag{11}$$

where $I(\mathbf{Z}_i^b, \mathbf{Z}_{i+1}^b)$ stands for the mutual information between $\mathbf{Z}_i^b$ and $\mathbf{Z}_{i+1}^b$, $\mathcal{L}_{cc}^{opt}$ denotes the optimal contrastive loss.

The detailed proof is shown in Appendix B. By proving the bounds of $\mathcal{L}_{cc}^{opt}$, we can maximize a lower bound on mutual information between different scale representations of fMRI by minimizing $\mathcal{L}_{cc}^{opt}$. The algorithm description and pseudocode can be discovered in Appendix C.

5 Experiments

5.1 Experimental Setups

Simulated fMRI Dataset. The benchmark simulated datasets we use are Smith dataset [58] and Sanchez dataset [54] both generated by dynamic causal model. Specifically, Sanchez dataset offers higher temporal resolution and acquisition frequency compared to Smith dataset and minimally affects the non-Gaussianity of the BOLD signal, a configuration commonly observed in real brain networks [45].

Furthermore, we generate a simulated dataset using CDRL [79]. Considering that real resting-state fMRI dataset we use includes 7 ROIs and features cycles, the generated simulated dataset contains 5 and 7 nodes, named simulated complex network with node 5 (SCN5) and simulated complex network with node 7 (SCN7), respectively. Details on all benchmark simulated datasets used are provided in Appendix D.1.

Real Resting-state fMRI Dataset. To assess the performance of methods under real BOLD data conditions, we utilize high-resolution 7T human resting-state fMRI data [56] from medial temporal lobe. The real resting-state fMRI dataset comprises 23 healthy adults, each with an acquisition time of 1.0 seconds and a session duration of 7 minutes per subject, yielding fMRI time series of 421 data points. We consider seven ROIs from the medial temporal lobe, specifically cornu ammonis 1 (CA1), cornu ammonis 2, 3, and dentate gyrus (CA23DG), subiculum (SUB), entorhinal cortex (ERC), brodmann area 35 (BA35), brodmann area 36 (BA36), and parahippocampal Cortex (PHC). These regions are assigned the numbers 1 through 7, respectively. More details about the real resting-state dataset can be found in Appendix D.2.

Post Process and Evaluation. The BrainECLLM directly outputs a directed weighted graph representing the brain EC network. For evaluation, we apply a post processing step to generate a binary EC matrix that matches the ground truth in the simulated fMRI dataset. Similarly, for real fMRI data, we visualize the EC by indicating the presence or absence of directed edges between brain regions. This approach is consistent with common practices in the literature when true connectivity weights are unavailable [22, 18]. The details of post processing can be found in Appendix D.3.

To evaluate the effectiveness of methods, we employ the following five commonly used evaluation metrics: Precision, Recall, F1, Accuracy, and Structural Hamming Distance (SHD). See Appendix D.4 for specific calculations.

Implementation Details. We use Llama3-8B [16] as the default backbone model unless otherwise specified. All our experiments are repeated three times, and we report the average results. All experiments are conducted on a single Nvidia L20-48GB GPU. For Smith dataset, training BrainEC-LLM takes approximately 5.5 hours. Model specific time complexity analysis and efficiency analysis experiments are shown in Appendix D.5.

We train the model unsupervisedly in an autoregressive manner. Detailed training methods and hyperparameter settings can be found in Appendix D.6 and Appendix D.7, respectively.

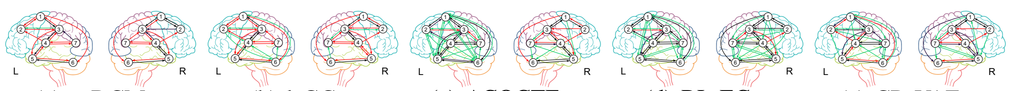
Baselines. Eight baseline methods are harnessed for comparison with the proposed method, and the baseline methods are as follows: spDCM [20], lsGC [15], ACOCTE [40], RL-EC [42], CR-VAE [36], MetaCAE [30], MetaRLEC [75], CUTS+ [12]. Extra descriptions of the baseline methods can be accessed in Appendix D.8.

5.2 Results on Simulated fMRI Dataset

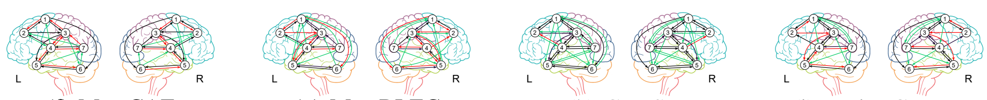
We run BrainEC-LLM and the 8 baseline methods three times (group analysis) for all subjects in each simulated dataset and calculate the mean and variance for these runs. It is important to note that some methods produce identical results across multiple runs, resulting in a variance of 0. Our main results are presented in Table 1, where BrainEC-LLM outperforms all baselines in most cases. Smith, Sanchez, and SCN5 all have EC networks with 5 nodes, but SCN5 features more edges, resulting in a

Table 1: The mean and variance of the nine methods on the simulated fMRI datasets. The best and second-best values are **highlighted** and underlined.

Datasets	Metrics	spDCM 2014	lsGC 2017	ACOCTE 2022	RL-EC 2022	CR-VAE 2023	MetaCAE 2024	MetaRLEC 2024	CUTS+ 2024	BrainEC-LLM (**Ours**)
Smith	Precision ↑	0.50±0.00	0.50±0.00	0.53±0.17	**0.58±0.12**	0.45±0.19	0.55±0.10	0.38±0.05	0.33±0.09	$\underline{0.57\pm0.06}$
	Recall ↑	0.40±0.00	0.60±0.00	0.40±0.16	0.47±0.09	0.73±0.25	0.79±0.15	0.60±0.12	0.42±0.12	**0.80±0.15**
	F1 ↑	0.44±0.00	0.55±0.00	0.45±0.17	0.52±0.10	0.56±0.21	$\underline{0.65\pm0.12}$	0.46±0.07	0.37±0.06	**0.67±0.11**
	Accuracy ↑	0.80±0.00	0.80±0.00	0.81±0.05	$\underline{0.83\pm0.04}$	0.76±0.11	$\underline{0.82\pm0.08}$	0.72±0.03	0.72±0.04	**0.84±0.06**
	SHD ↓	5.00±0.00	5.00±0.00	4.67±0.25	$\underline{4.33\pm0.94}$	6.00±2.83	$\underline{4.26\pm0.93}$	7.00±0.86	7.00±0.78	**4.02±0.84**
Sanchez	Precision ↑	0.57±0.00	0.60±0.00	0.76±0.06	**0.80±0.02**	0.50±0.04	0.50±0.07	0.80±0.04	0.64±0.03	$\underline{0.78\pm0.02}$
	Recall ↑	0.57±0.00	$\underline{0.86\pm0.00}$	0.57±0.00	0.57±0.07	0.76±0.13	0.29±0.11	0.57±0.14	0.86±0.00	**0.97±0.02**
	F1 ↑	0.57±0.00	$\underline{0.71\pm0.00}$	0.65±0.02	0.67±0.05	0.60±0.07	0.37±0.07	0.67±0.06	$\underline{0.74\pm0.02}$	**0.86±0.07**
	Accuracy ↑	0.76±0.00	0.80±0.00	0.83±0.02	$\underline{0.84\pm0.02}$	0.72±0.03	0.72±0.04	0.84±0.03	$\underline{0.83\pm0.02}$	**0.91±0.05**
	SHD ↓	6.00±0.00	5.00±0.00	4.33±0.47	$\underline{3.95\pm0.47}$	7.05±0.82	6.89±0.85	4.00±0.89	4.33±0.47	**2.10±0.75**
SCN5	Precision ↑	0.40±0.00	0.27±0.00	0.50±0.04	$\underline{0.67\pm0.05}$	0.40±0.03	0.36±0.02	0.50±0.01	0.45±0.02	**0.83±0.04**
	Recall ↑	**0.75±0.00**	0.38±0.00	0.13±0.02	$\underline{0.25\pm0.08}$	$\underline{0.67\pm0.06}$	0.58±0.06	0.25±0.04	0.63±0.03	0.63±0.05
	F1 ↑	0.52±0.00	0.32±0.00	0.20±0.04	0.36±0.04	$\underline{0.50\pm0.04}$	0.44±0.03	0.33±0.03	$\underline{0.53\pm0.03}$	**0.71±0.04**
	Accuracy ↑	0.56±0.00	0.48±0.00	$\underline{0.68\pm0.05}$	0.72±0.12	0.57±0.04	0.53±0.02	0.68±0.07	$\underline{0.64\pm0.09}$	**0.84±0.10**
	SHD ↓	11.00±0.00	13.00±0.00	$\underline{8.00\pm0.43}$	$\underline{6.82\pm0.85}$	10.67±0.94	11.67±0.47	8.24±0.45	9.06±0.67	**4.26±0.83**
SCN7	Precision ↑	$\underline{0.67\pm0.00}$	0.30±0.00	**0.77±0.02**	0.65±0.01	0.37±0.00	0.34±0.03	0.52±0.02	0.34±0.00	0.62±0.02
	Recall ↑	$\underline{0.13\pm0.00}$	0.60±0.00	0.22±0.03	0.49±0.03	$\underline{0.80\pm0.14}$	**0.84±0.08**	0.47±0.13	0.58±0.02	0.61±0.05
	F1 ↑	0.22±0.00	0.40±0.00	0.34±0.04	$\underline{0.56\pm0.03}$	$\underline{0.50\pm0.03}$	0.48±0.05	0.48±0.04	0.43±0.03	**0.61±0.03**
	Accuracy ↑	0.71±0.00	0.45±0.00	0.74±0.01	$\underline{0.75\pm0.01}$	0.52±0.03	0.44±0.05	0.69±0.05	0.52±0.02	**0.76±0.05**
	SHD ↓	14.00±0.00	27.00±0.00	12.67±0.47	**11.67±0.47**	23.67±1.70	27.33±2.49	14.94±0.82	23.33±0.94	$\underline{11.86\pm0.91}$



(a) spDCM (b) lsGC (c) ACOCTE (d) RL-EC (e) CR-VAE



(f) MetaCAE (g) MetaRLEC (h) CUTS+ (i) BrainEC-LLM

Figure 3: Brain EC estimated by nine methods on real resting-state fMRI dataset. The **black** lines represent correct connections, the **green** lines stand for spurious connections, and the **red** lines indicate missing connections.

more complex EC network. SCN7, on the other hand, has 7 nodes with a complex EC network. We observe that most methods perform poorly on the Smith dataset, a finding that aligns with existing literature [39]. Baseline methods such as lsGC, ACOCTE, and RL-EC perform relatively well on the first two datasets, but their performance declines as the network complexity increases. CR-VAE shows a more consistent performance across all four datasets. BrainEC-LLM excels on the first three datasets; however, its performance diminishes on the fourth dataset, likely due to the increased number of nodes. Overall, BrainEC-LLM frequently outperforms the baseline methods across the four different datasets.

To assess the significant differences between BrainEC-LLM and other baseline methods, ***t*-test analysis** and **Wilcoxon test with Holm correction** are performed, with results presented in Appendix E.1. The analysis reveals that, except for RL-EC, which shows no significant difference in the evaluation metric SHD, BrainEC-LLM is significantly different from the baseline methods in all other cases.

5.3 Results on real resting-state fMRI Dataset

Since no established brain EC network is available for existing real resting-state fMRI data, we evaluate methods by referencing [54]. We conduct experiments on the left and right hemispheres using BrainEC-LLM and other baseline methods. The resulting EC networks are displayed in Figure 3. In the left hemisphere, BrainEC-LLM correctly identifies 11 connections, generates 6 spurious connections, and misses 6 correct connections. Similarly, in the right hemisphere, BrainEC-LLM identifies 12 correct connections, produces 6 spurious connections, and misses 5 correct connections. In contrast, spDCM and lsGC generated relatively sparse graphs, detecting fewer edges and consequently missing many correct connections. Although CUTS+ misses fewer connections, its higher SHD indicates that BrainEC-LLM achieves superior overall accuracy in EC estimation, as SHD penalizes both missing and spurious connections. Overall, BrainEC-LLM performs best in left hemisphere and ranks second

Table 2: Zero-shot learning results of BrainEC-LLM.

Datasets	Precision↑	Recall↑	F1↑	Accuracy↑	SHD↓
Smith→SCN5	0.83	0.63	0.71	0.84	4.0
SCN5→Smith	0.68	0.51	0.58	0.76	5.0

Table 3: Classification results on ABIDE I.

Methods	Precision(%)↑	Recall(%)↑	F1(%)↑	Accuracy(%)↑
spDCM	64.15±9.89	62.88±5.10	63.51±7.73	64.74±4.21
lsGC	65.32±8.92	63.01±4.81	64.14±6.80	65.82±3.46
ACOCTE	55.34±9.51	55.67±3.66	55.50±6.41	57.83±3.76
RL-EC	55.76±9.98	53.74±5.69	54.73±7.22	56.62±4.03
CR-VAE	63.12±9.03	63.33±3.36	63.22±5.89	64.35±3.75
MetaCAE	62.34±9.51	61.67±4.31	62.00±5.93	64.18±4.39
MetaRLEC	62.35±8.71	65.97±3.88	64.14±5.01	66.84±4.63
CUTS+	67.82±8.34	65.41±4.13	67.52±6.80	69.14±3.85
BrainEC-LLM	**69.20±7.98**	**81.11±8.19**	**72.31±5.11**	**71.11±3.20**

Table 4: Classification results on ADHD.

Methods	Precision(%)↑	Recall(%)↑	F1(%)↑	Accuracy(%)↑
spDCM	69.15±8.60	63.84±5.31	66.38±6.56	66.73±5.97
lsGC	**70.87**±**5.62**	60.95±5.24	65.53±5.42	64.23±5.33
ACOCTE	63.94±4.17	62.13±2.91	62.75±3.42	61.29±3.41
RL-EC	61.86±4.88	60.42±3.43	61.13±4.02	59.72±3.78
CR-VAE	66.39±7.71	63.84±4.42	64.04±5.61	64.61±4.84
MetaCAE	64.59±7.92	63.93±4.14	64.20±5.43	64.31±4.97
MetaRLEC	63.72±8.23	62.80±4.82	63.21±5.98	64.39±4.83
CUTS+	66.50±7.52	65.62±5.32	66.01±4.76	65.56±4.25
BrainEC-LLM	67.45±6.23	**72.38**±**4.91**	**69.82**±**5.57**	**67.87**±**5.02**

only to spDCM in right hemisphere. A case study for visualization of LLM inputs is detailed in Appendix E.2.

5.4 Zero-shot Learning

This task evaluates the transferability of BrainEC-LLM from source domain to target domain, specifically assessing how well the model performs on dataset A (without any training data from A) when already trained on dataset B. Table 2 demonstrates that our method performs well, even when it has never encountered another dataset. The diverse zero-shot performance reflects differences in network complexity between Smith and SCN5 datasets. SCN5, with 5 nodes, 8 arcs, and 4 cycles introducing quadratic nonlinear causality, represents a more complex structure than Smith (5 nodes, 5 arcs, 0 cycles). When transferring from Smith to SCN5, the model maintains stable F1 performance, as SCN5's complexity allows leveraging multiscale features learned from simpler Smith dynamics. Conversely, SCN5 $\rightarrow$ Smith transfer shows performance degradation, likely due to overfitting to SCN5's nonlinear patterns, making adaptation to Smith's simpler acyclic structure suboptimal. This indicates better model transferability to complex structures than simpler ones in zero-shot scenarios.

5.5 Downstream Tasks (Brain Disease Classification using EC networks)

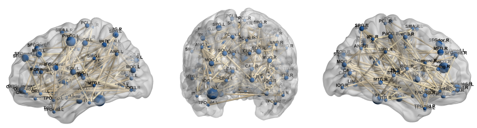
To further validate the discriminative power of BrainEC-LLM for EC estimated on real fMRI dataset without ground truth, we apply it to the ABIDE I dataset[2] and ADHD dataset[3] to estimate brain EC network. The resulting network is then used as input to SVM classifier for downstream classification tasks. The 10-fold cross validation results of this analysis are presented in Table 3, which demonstrate that BrainEC-LLM not only effectively estimates EC but also yields strong performance in downstream classification tasks.

Figure 4 illustrates a comparative analysis of brain EC networks in healthy controls (HCs) versus individuals with autism spectrum disorder (ASD), based on AAL-90 atlas (90 regions) [13]. To analyze the most critical effective connectivity between brain regions, we preserve the top 5% highest-scoring edges for visualization. Previous literature has consistently reported that individuals with ASD exhibit increased connectivity in right middle temporal gyrus (MTG) [71], decreased connectivity in temporal pole (TPO) [23] and supplementary motor area (SMA) [63], and asymmetric connectivity alterations in the middle frontal gyrus (MFG), with reduced connectivity in the left hemisphere and increased connectivity in the right hemisphere [23, 28]. The connectivity patterns observed in figure align with these previous findings. More details can be found in Appendix E.3.

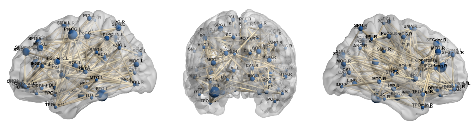
5.6 Model Analysis

Ablation Study. We perform ablation studies on model backbone, model module, and the loss function, with the results presented in Table 5. "w/o LLM" refers to removing the pre-trained LLM entirely from the BrainEC-LLM framework. Our results indicate that including $\mathcal{L}_{csc}$ provides a slight

[2]http://preprocessed-connectomes-project.org/abide/
[3]https://fcon_1000.projects.nitrc.org/indi/adhd200/

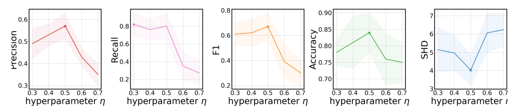


(a) HCs. (b) HCs. (c) HCs.

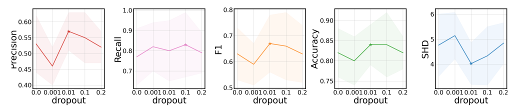


(d) ASD. (e) ASD. (f) ASD.

Figure 4: Comparison of brain EC network between healthy controls and ASD patients on the ABIDE I dataset Using BrainEC-LLM.



(a) Influence of soft threshold hyperparameter η.



(b) Influence of the dropout.

Figure 5: Hyperparameter analysis on the Smith dataset (the starred results are the best results).

improvement over not including it. For other components, such as the prompts generation module and multiscale mixing module (includes both multiscale decomposition and mixing, without loss $\mathcal{L}_{csc}$), the improvements are more pronounced in BrainEC-LLM. Additionally, the results using Llama 3 as the backbone are manifestly outperforms Mistral. More results can be found in the Appendix E.4.

Table 5: Ablations on Smith dataset.

Variant	Precision↑	Recall↑	F1↑	Accuracy↑	SHD↓
Llama3-8B (Default)	0.57±0.06	**0.80±0.15**	**0.67±0.11**	**0.84±0.06**	**4.02±0.84**
Mistral-7B	0.29±0.07	0.40±0.10	0.33±0.13	0.68±0.08	7.65±1.36
w/o Prompts Generation	0.38±0.04	0.60±0.15	0.46±0.14	0.72±0.08	7.10±0.93
w/o Multiscale Mixing	0.20±0.08	0.21±0.12	0.20±0.13	0.68±0.06	7.88±0.95
w/o Cross Attention	0.23±0.12	0.32±0.13	0.27±0.12	0.53±0.06	11.80±0.93
w/o LLM	0.33±0.11	0.42±0.14	0.35±0.14	0.72±0.08	7.14±1.20
w/o $\mathcal{L}_{csc}$	**0.58±0.04**	0.70±0.16	0.57±0.09	0.78±0.07	4.84±1.26

Hyperparameter Analysis. We conduct experiments on Smith datasets to evaluate the parameter sensitivity of BrainEC-LLM. Figures 5 (a) and 5 (b) display the experimental results for soft threshold hyperparameters and dropout variations, respectively. When the soft threshold is set to 0.5, all metrics are optimal except for Recall, which remains sub-optimal, a pattern also observed with dropout. Consequently, the soft threshold is ultimately set to 0.5, and dropout is set to 0.01. Refer to Appendix E.6 for further details.

6 Conclusion

We present BrainEC-LLM, the first work to fine-tune pre-trained LLMs for estimating brain effective connectivity (EC) from fMRI data. BrainEC-LLM introduces multiscale decomposition mixing module that downsamples and decomposes fMRI time series data to capture short-term and long-term multiscale trends, mixing complex multiscale temporal variations in both bottom-up and top-down manner. These multiscale fMRI sequences are then aligned with natural language embeddings through cross attention using pre-trained word embeddings. Finally, the multiscale features generated by LLM are input into multiscale reconstruction mixing module to integrate the information across different scales. The brain EC network is then estimated using self attention. Our comprehensive empirical study demonstrates the effectiveness of BrainEC-LLM. One potential limitation that warrants further investigation is that existing LLMs are primarily trained on textual data, whereas fMRI time series data are numerical. LLMs with enhanced numerical reasoning capabilities may yield better results in this context.

7 Acknowledgements

The authors thank anonymous reviewers for their valuable comments and helpful suggestions. This work was sponsored by Beijing Nova Program (20240484635), supported by National Natural Science Foundation of China (62106009, 62276010), and R&D Program of Beijing Municipal Education Commission (KM202210005030, KZ202210005009).

References

[1] Daniel Alcalá-López, Jonathan Smallwood, Elizabeth Jefferies, Frank Van Overwalle, Kai Vogeley, Rogier B Mars, Bruce I Turetsky, Angela R Laird, Peter T Fox, Simon B Eickhoff, et al. Computing the social brain connectome across systems and states. *Cerebral cortex*, 28(7): 2207–2232, 2018.

[2] Neena Aloysius and M Geetha. A review on deep convolutional neural networks. In *2017 international conference on communication and signal processing (ICCSP)*, pages 0588–0592. IEEE, 2017.

[3] Galia Avidan, Michal Tanzer, Fadila Hadj-Bouziane, Ning Liu, Leslie G Ungerleider, and Marlene Behrmann. Selective dissociation between core and extended regions of the face processing network in congenital prosopagnosia. *Cerebral cortex*, 24(6):1565–1578, 2014.

[4] Sara Bagherzadeh, Mohsen Sadat Shahabi, and Ahmad Shalbaf. Detection of schizophrenia using hybrid of deep learning and brain effective connectivity image from electroencephalogram signal. *Computers in Biology and Medicine*, 146:105570, 2022.

[5] Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, and Huanhuan Chen. Causal structure learning supervised by large language model. *arXiv preprint arXiv:2311.11689*, 2023.

[6] Kamyar Barakati, Aleksander Molak, Chris Nelson, Xiaohang Zhang, Ichiro Takeuchi, and Sergei V Kalinin. Causal discovery from data assisted by large language models. *Applied Physics Letters*, 127(12), 2025.

[7] A Bastos. Dcm for complex-valued data: cross-spectra, coherence and phase-delays. *Neuroimage*, 59(1):439–455, 2012.

[8] Nitin Kumar Chauhan and Krishna Singh. A review on conventional machine learning vs deep learning. In *2018 International conference on computing, power and communication technologies (GUCON)*, pages 347–352. IEEE, 2018.

[9] Dongdong Chen and Lichi Zhang. Fe-stgnn: Spatio-temporal graph neural network with functional and effective connectivity fusion for mci diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 67–76. Springer, 2023.

[10] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[11] Xiaoyu Chen, Changde Du, Che Liu, Yizhe Wang, and Huiguang He. Open-vocabulary auditory neural decoding using fmri-prompted llm. *arXiv preprint arXiv:2405.07840*, 2024.

[12] Yuxiao Cheng, Lianglong Li, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. Cuts+: High-dimensional causal discovery from irregular time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11525–11533, 2024.

[13] Zhengjia Dai, Chaogan Yan, Zhiqun Wang, Jinhui Wang, Mingrui Xia, Kuncheng Li, and Yong He. Discriminative analysis of early alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (m3). *Neuroimage*, 59(3):2187–2195, 2012.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[15] Adora M DSouza, Anas Z Abidin, Lutz Leistritz, and Axel Wismüller. Exploring connectivity with large-scale granger causality on resting-state functional mri. *Journal of neuroscience methods*, 287:68–79, 2017.

[16] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[17] Nico Eisenhauer, Matthew A Bowker, James B Grace, and Jeff R Powell. From patterns to causal understanding: structural equation modeling (sem) in soil ecology. *Pedobiologia*, 58 (2-3):65–72, 2015.

[18] Sam Ereira, Sheena Waters, Adeel Razi, and Charles R Marshall. Early detection of dementia with default-mode network effective connectivity. *Nature Mental Health*, 2(7):787–800, 2024.

[19] Ebrahim Feghhi, Nima Hadidi, Bryan Song, Idan A Blank, and Jonathan C Kao. What are large language models mapping to in the brain? a case against over-reliance on brain scores. *arXiv preprint arXiv:2406.01538*, 2024.

[20] Karl J Friston, Joshua Kahan, Bharat Biswal, and Adeel Razi. A dcm for resting state fmri. *Neuroimage*, 94:396–407, 2014.

[21] Le Gao, Shuang Qiao, Yigeng Zhang, Tao Zhang, Huibin Lu, and Xiaonan Guo. Parsing the heterogeneity of brain structure and function in male children with autism spectrum disorder: a multimodal mri study. *Brain Imaging and Behavior*, pages 1–14, 2025.

[22] Matthieu Gilson, Ruben Moreno-Bote, Adrián Ponce-Alvarez, Petra Ritter, and Gustavo Deco. Estimation of directed effective connectivity from fmri functional connectivity hints at asymmetries of cortical connectome. *PLoS computational biology*, 12(3):e1004762, 2016.

[23] Enrico Glerean, Raj K Pan, Juha Salmi, Rainer Kujala, Juha M Lahnakoski, Ulrika Roine, Lauri Nummenmaa, Sami Leppämäki, Taina Nieminen-von Wendt, Pekka Tani, et al. Reorganization of functionally connected brain subnetworks in high-functioning autism. *Human brain mapping*, 37(3):1066–1079, 2016.

[24] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36: 19622–19635, 2023.

[25] Youssef Hmamouche, Ismail Chihab, Lahoucine Kdouri, and Amal El Fallah Seghrouchni. A multimodal llm for the non-invasive decoding of spoken text from brain recordings. *arXiv preprint arXiv:2409.19710*, 2024.

[26] Amelia Rizky Idhartono, Nurul Hidayati, Alsyah Safarina Subekti, and Margareta Vernanda Moi. -strategi pembelajaran di tingkat sma untuk siswa dengan autism spectrum disorder (asd). *Kanigara*, 4(2):129–139, 2024.

[27] Takashi Itahashi, Takashi Yamada, Hiromi Watanabe, Motoaki Nakamura, Haruhisa Ohta, Chieko Kanai, Akira Iwanami, Nobumasa Kato, and Ryu-ichiro Hashimoto. Alterations of local spontaneous brain activity and connectivity in adults with high-functioning autism spectrum disorder. *Molecular autism*, 6:1–14, 2015.

[28] Takashi Itahashi, Takashi Yamada, Hiromi Watanabe, Motoaki Nakamura, Haruhisa Ohta, Chieko Kanai, Akira Iwanami, Nobumasa Kato, and Ryu-ichiro Hashimoto. Alterations of local spontaneous brain activity and connectivity in adults with high-functioning autism spectrum disorder. *Molecular autism*, 6:1–14, 2015.

[29] Shruti Japee, Kelsey Holiday, Maureen D Satyshur, Ikuko Mukai, and Leslie G Ungerleider. A role of right middle frontal gyrus in reorienting of attention: a case study. *Frontiers in systems neuroscience*, 9:23, 2015.

[30] Junzhong Ji, Zuozhen Zhang, Lu Han, and Jinduo Liu. Metacae: Causal autoencoder with meta-knowledge transfer for brain effective connectivity estimation. *Computers in Biology and Medicine*, page 107940, 2024.

[31] Yanchun Jiang, Yanbo Zhang, Liluo Nie, Huihua Liu, and Jinou Zheng. Identification and effective connections of core networks in patients with temporal lobe epilepsy and cognitive impairment: Granger causality analysis and multivariate pattern analysis. *International Journal of Neuroscience*, 133(9):935–946, 2023.

[32] Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. Position: What can large language models tell us about time series analysis. In *Forty-first International Conference on Machine Learning*, 2024.

[33] Katikapalli Subramanyam Kalyan. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, page 100048, 2023.

[34] Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *10th International Conference on Learning Representations, ICLR 2022*, 2022.

[35] Tracey A Knaus, Claire Burns, Jodi Kamps, and Anne L Foundas. Atypical activation of action-semantic network in adolescents with autism spectrum disorder. *Brain and cognition*, 117:57–64, 2017.

[36] Hongming Li, Shujian Yu, and Jose Principe. Causal recurrent variational autoencoder for medical time series generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 8562–8570, 2023.

[37] Lei Li, Changchun He, Taorong Jian, Xiaonan Guo, Jinming Xiao, Ya Li, Heng Chen, Xiaodong Kang, Huafu Chen, and Xujun Duan. Attenuated link between the medial prefrontal cortex and the amygdala in children with autism spectrum disorder: Evidence from effective connectivity within the "social brain". *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 111:110147, 2021.

[38] Di Liang, Shengxiang Xia, Xianfu Zhang, and Weiwei Zhang. Analysis of brain functional connectivity neural circuits in children with autism based on persistent homology. *Frontiers in Human Neuroscience*, 15:745671, 2021.

[39] Jinduo Liu, Junzhong Ji, Guangxu Xun, Liuyi Yao, Mengdi Huai, and Aidong Zhang. Ec-gan: inferring brain effective connectivity via generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4852–4859, 2020.

[40] Jinduo Liu, Junzhong Ji, Guangxu Xun, and Aidong Zhang. Inferring effective connectivity networks from fmri time series with a temporal entropy-score. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10):5993–6006, 2022.

[41] Jinduo Liu, Lu Han, and Junzhong Ji. Mcan: multimodal causal adversarial networks for dynamic effective connectivity learning from fmri and eeg data. *IEEE Transactions on Medical Imaging*, 43(8):2913–2923, 2024.

[42] Yilin Lu, Jinduo Liu, Junzhong Ji, Han Lv, and Mengdi Huai. Brain effective connectivity learning with deep reinforcement learning. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1664–1667. IEEE, 2022.

[43] Donghao Luo and Xue Wang. Moderntcn: A modern pure convolution structure for general time series analysis. In *The Twelfth International Conference on Learning Representations*, 2024.

[44] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9(1):381–386, 2020.

[45] Falko Mecklenbrauck, Marius Gruber, Sophie Siestrup, Anoushiravan Zahedi, Dominik Grotegerd, Marco Mauritz, Ima Trempler, Udo Dannlowski, and Ricarda I Schubotz. The significance of structural rich club hubs for the processing of hierarchical stimuli. *Human Brain Mapping*, 45(4):e26543, 2024.

[46] MC Mozer. Induction of multiscale temporal structure. In *Neural Information Processing Systems*, volume 4, pages 275–282. Morgan Kaufmann, 1992.

[47] Izzy Newsham, Luka Kovačević, Richard Moulange, Nan Rosemary Ke, and Sach Mukherjee. Large language models for zero-shot inference of causal structures in biology. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, 2025.

[48] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

[49] Alessandra M Pereira, Brunno M Campos, Ana C Coan, Luiz F Pegoraro, Thiago JR De Rezende, Ignacio Obeso, Paulo Dalgalarrondo, Jaderson C Da Costa, Jean-Claude Dreher, and Fernando Cendes. Differences in cortical structure and functional mri connectivity in high functioning autism. *Frontiers in neurology*, 9:539, 2018.

[50] David C Plaut and Marlene Behrmann. Response to susilo and duchaine: beyond neuropsychological dissociations in understanding face and word representations. *Trends in cognitive sciences*, 17(11):546, 2013.

[51] Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. *Advances in neural information processing systems*, 31, 2018.

[52] Dipanjan Ray, Dmitry Bezmaternykh, Mikhail Mel'nikov, Karl J Friston, and Moumita Das. Altered effective connectivity in sensorimotor cortices is a signature of severity and clinical course in depression. *Proceedings of the National Academy of Sciences*, 118(40):e2105730118, 2021.

[53] Siti Hajar Mohd Roffeei, Noorhidawati Abdullah, and Siti Khairatul Razifah Basar. Seeking social support on facebook for children with autism spectrum disorders (asds). *International journal of medical informatics*, 84(5):375–385, 2015.

[54] Ruben Sanchez-Romero, Joseph D Ramsey, Kun Zhang, Madelyn RK Glymour, Biwei Huang, and Clark Glymour. Estimating feedforward and feedback effective connections from fmri time series: Assessments of statistical methods. *Network Neuroscience*, 3(2):274–306, 2019.

[55] Letten F Saugstad. Infantile autism: a chronic psychosis since infancy due to synaptic pruning of the supplementary motor area. *Nutrition and health*, 20(3-4):171–182, 2011.

[56] Preya Shah, Danielle S Bassett, Laura EM Wisse, John A Detre, Joel M Stein, Paul A Yushkevich, Russell T Shinohara, John B Pluta, Elijah Valenciano, Molly Daffner, et al. Mapping the structural and functional network architecture of the medial temporal lobe using 7t mri. *Human Brain Mapping*, 39(2):851–865, 2018.

[57] Pramila P Shinde and Seema Shah. A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, pages 1–6. IEEE, 2018.

[58] Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

[59] Yuni Susanti and Michael Färber. Can llms leverage observational data? towards data-driven causal discovery with llms. *arXiv preprint arXiv:2504.10936*, 2025.

[60] Mingtian Tan, Mike Merrill, Vinayak Gupta, Tim Althoff, and Tom Hartvigsen. Are language models actually useful for time series forecasting? *Advances in Neural Information Processing Systems*, 37:60162–60191, 2024.

[61] Mohammad Mustafa Taye. Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*, 12(5):91, 2023.

[62] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[63] Yonglu Wang, Lingxi Xu, Hui Fang, Fei Wang, Tianshu Gao, Qingyao Zhu, Gongkai Jiao, and Xiaoyan Ke. Social brain network of children with autism spectrum disorder: characterization of functional connectivity and potential association with stereotyped behavior. *Brain Sciences*, 13(2):280, 2023.

[64] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

[65] Wen Xiong, Jinduo Liu, Junzhong Ji, and Fenglong Ma. Brain effective connectivity estimation via fourier spatiotemporal attention. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 1657–1668, 2025.

[66] Jinping Xu, Chao Wang, Ziyun Xu, Tian Li, Fangfang Chen, Kai Chen, Jingjing Gao, Jiaojian Wang, and Qingmao Hu. Specific functional connectivity patterns of middle temporal gyrus subregions in children and adults with autism spectrum disorder. *Autism Research*, 13(3): 410–422, 2020.

[67] Shilin Xu, Xin Wang, Linling Shen, Xiaohui Yan, Guoyan Feng, and Fan Cao. Brain functional differences during irony comprehension in adolescents with asd. *Cerebral Cortex*, 35(2): bhaf003, 2025.

[68] Ziyi Ye, Qingyao Ai, Yiqun Liu, Maarten de Rijke, Min Zhang, Christina Lioma, and Tuukka Ruotsalo. Generative language reconstruction from brain recordings. *Communications Biology*, 8(1):346, 2025.

[69] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), November 2024. ISSN 2053-714X. doi: 10.1093/nsr/nwae403. URL `http://dx.doi.org/10.1093/nsr/nwae403`.

[70] Minqi Yu, Jinduo Liu, and Junzhong Ji. Causal invariance-aware augmentation for brain graph contrastive learning. In *Forty-second International Conference on Machine Learning*.

[71] Yaxu Yu, Xiaoqin Wang, Junyi Yang, and Jiang Qiu. The role of the mtg in negative emotional processing in young adults with autistic-like traits: a fmri task study. *Journal of Affective Disorders*, 276:890–897, 2020.

[72] Khadija Zanna and Akane Sano. Fairness-driven llm-based causal discovery with active learning and dynamic scoring. *arXiv preprint arXiv:2503.17569*, 2025.

[73] Xinyu Zhang, Shanshan Feng, and Xutao Li. From text to time? rethinking the effectiveness of the large language model for time series forecasting. *arXiv preprint arXiv:2504.08818*, 2025.

[74] Zhengxin Zhang, Dan Zhao, Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Qing Li, Yong Jiang, and Zhihao Jia. Quantized side tuning: Fast and memory-efficient tuning of quantized large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–17, 2024.

[75] Zuozhen Zhang, Junzhong Ji, and Jinduo Liu. Metarlec: Meta-reinforcement learning for discovery of brain effective connectivity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10261–10269, 2024.

[76] Xiaoxin Zhao, Shuyi Zhu, Yang Cao, Peipei Cheng, Yuxiong Lin, Zhixin Sun, Wenqing Jiang, and Yasong Du. Abnormalities of gray matter volume and its correlation with clinical symptoms in adolescents with high-functioning autism spectrum disorder. *Neuropsychiatric Disease and Treatment*, 18:717, 2022.

[77] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

[78] Xiaofan Zhou, Liangjie Huang, Pinyang Cheng, Wenpen Yin, Rui Zhang, Wenrui Hao, and Lu Cheng. Accelerating causal network discovery of alzheimer disease biomarkers via scientific literature-based retrieval augmented generation. *arXiv preprint arXiv:2504.08768*, 2025.

[79] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. In *International Conference on Learning Representations*, 2020.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Refer to the sentence "Considering the unique nature of... the first method to fine-tune LLMs for estimating brain EC from fMRI data" of the abstract and the last paragraph of section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Refer to Appendix F.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Refer to Appendix B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Refer to "Implementation Details" in section 5.1, Appendix C and Appendix D.7.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Refer to the last sentence of abstract, Appendix D.1 and Appendix D.2.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Refer to "Post Process and Evaluation" and "Implementation Details" in section 5.1, Appendix D.7.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Refer to Table 1 and Appendix E.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Refer to "Implementation Details" in section 5.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The datasets used in the paper are publicly available datasets, refer to "Simulated fMRI Dataset" and "Real Resting-state fMRI Dataset" in section 5.1.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There are no privacy or security concerns and therefore no negative societal impacts, refer to first paragraph of section 1 for positive societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The source of the data can be found at section 5.1. We have properly cited all LLM-related papers whose models, code, or methods are used or referenced in our work, which can be found in section 4. All datasets, pre-trained models, and libraries used in our experiments are publicly available and used in accordance with their respective licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Our work proposes the first method to fine-tune LLMs for estimating brain effective connectivity from fMRI data in causal discovery field.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Examples of Prompts

Our prompt design is intended to provide the LLM with a structured representation of inter-regional brain relationships by explicitly encoding the Pearson correlation matrix, where each row and column corresponds to a specific brain region. To preserve a consistent and interpretable mapping between brain regions and their associated fMRI ROI time series, numerical indexing is incorporated into the prompt along with corresponding explanations, ensuring the model can clearly associate each index with a specific brain region. Examples of prompts are illustrated in Figure 6, which clearly illustrates how the LLM associates brain regions in the Pearson correlation matrix with those in the fMRI data. The purple, green, and pink prompts correspond to task description, dataset description, and prior knowledge, as depicted in Figure 1, respectively.

```
{"role":"system", "content":"You are a data analysis
expert. Your task is to help users analyze fMRI time series
data to extract brain effective connectivity network."},
{"role":"user", "content":"Dataset description: This
simulated fMRI time series dataset is generated through
dynamic causal models, with dimensions B × T × N, where B
is the batch size, T is the data points, and N is the
number of brain regions.
The following is a matrix of Pearson correlation
coefficients for 5 fMRI time series, with the number i
denoting the i-th brain region:"
"Prior knowledge: 1, 2, 3, 4, 5
1: 0.00, 0.31, 0.14, 0.06, 0.30
2: 0.31, 0.00, 0.35, 0.12, 0.15
3: 0.14, 0.35, 0.00, 0.32, 0.17
4: 0.06, 0.12, 0.32, 0.00, 0.33
5: 0.30, 0.15, 0.17, 0.33, 0.00"}
```

(a) Prompt example of Smith dataset.

```
{"role":"system", "content":"You are a data analysis expert. fMRI is a
neuroimaging technique that measures brain activity by tracking changes in
blood oxygenation.
If we use simulated fMRI data, the time series are directly generated, while
real fMRI data consists of time series that have been preprocessed by external
sources.
We use these preprocessed fMRI time series to estimate brain effective
connectivity (EC), which refers to the directed influence between brain
regions, capturing causal interactions rather than just correlations."},
{"role":"user", "content":"Dataset description: This simulated fMRI time series
dataset is generated through dynamic causal models, with dimensions B × T × N,
where B is the batch size, T is the data points, and N is the number of brain
regions.
The following is a matrix of Pearson correlation coefficients for 5 fMRI time
series, with the number i denoting the i-th brain region:"
"Prior knowledge: 1, 2, 3, 4, 5
1: 0.00, 0.31, 0.14, 0.06, 0.30
2: 0.31, 0.00, 0.35, 0.12, 0.15
3: 0.14, 0.35, 0.00, 0.32, 0.17
4: 0.06, 0.12, 0.32, 0.00, 0.33
5: 0.30, 0.15, 0.17, 0.33, 0.00"}
```

(b) Complex prompt example of Smith dataset.

Figure 6: Prompt example of Smith dataset.

B Proof of Theorem 1

For positive samples $\{\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b\}$, they are drawn from the joint distribution $p_1(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b) = p(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b)$. In contrast, the negative samples $\{\boldsymbol{Z}_i^b, \boldsymbol{Z}_j^b\}, j \neq i, j \neq i+1$ are drawn from the marginal distribution $p_2(\boldsymbol{Z}_i^b, \boldsymbol{Z}_j^b) = p(\boldsymbol{Z}_i^b)p(\boldsymbol{Z}_j^b)$. Given a set of samples, the probability of correctly identifying a positive sample is

$$\begin{aligned}
&p(\{\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b\} | \{\{\boldsymbol{Z}_i^b, \boldsymbol{Z}_j^b\}, j = 1, ..., M\}) \\
=&\frac{p_1(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b) \prod_{j\neq i,(i+1)} p_2(\boldsymbol{Z}_i^b, \boldsymbol{Z}_j^b)}{\sum_{k\neq i} p_1(\boldsymbol{Z}_i^b, \boldsymbol{Z}_k^b) \prod_{j\neq k} p_2(\boldsymbol{Z}_i^b, \boldsymbol{Z}_j^b)} \\
=&\frac{p(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b) \prod_{j\neq i,(i+1)} p(\boldsymbol{Z}_i^b)p(\boldsymbol{Z}_j^b)}{\sum_{k\neq i} p(\boldsymbol{Z}_i^b, \boldsymbol{Z}_k^b) \prod_{j\neq k,i} p(\boldsymbol{Z}_i^b)p(\boldsymbol{Z}_j^b)} \\
=&\frac{p(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b) \prod_{j\neq i,(i+1)} p(\boldsymbol{Z}_i^b)p(\boldsymbol{Z}_j^b)}{\sum_{k\neq i} p(\boldsymbol{Z}_i^b, \boldsymbol{Z}_k^b) \frac{\prod_{j\neq i} p(\boldsymbol{Z}_i^b)p(\boldsymbol{Z}_j^b)}{p(\boldsymbol{Z}_i^b)p(\boldsymbol{Z}_k^b)}} \\
=&\frac{p(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b) \prod_{j\neq i,(i+1)} p(\boldsymbol{Z}_i^b)p(\boldsymbol{Z}_j^b)}{\prod_{j\neq i} p(\boldsymbol{Z}_i^b)p(\boldsymbol{Z}_j^b) \sum_{k\neq i} \frac{p(\boldsymbol{Z}_i^b, \boldsymbol{Z}_k^b)}{p(\boldsymbol{Z}_i^b)p(\boldsymbol{Z}_k^b)}} \\
=&\frac{\frac{p(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b)}{p(\boldsymbol{Z}_i^b)p(\boldsymbol{Z}_{i+1}^b)}}{\sum_{k\neq i} \frac{p(\boldsymbol{Z}_i^b, \boldsymbol{Z}_k^b)}{p(\boldsymbol{Z}_i^b)p(\boldsymbol{Z}_k^b)}}.
\end{aligned} \tag{12}$$

By comparing Eq. 12 and the definition of $\mathcal{L}_{csc}^{i,i+1}$, we find that the optimal function $g^{opt}(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b)$ is proportional to $\frac{p(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b)}{p(\boldsymbol{Z}_i^b)p(\boldsymbol{Z}_{i+1}^b)}$. Given that $\frac{p(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b)}{p(\boldsymbol{Z}_i^b)p(\boldsymbol{Z}_{i+1}^b)} = \frac{p(\boldsymbol{Z}_i^b|\boldsymbol{Z}_{i+1}^b)}{p(\boldsymbol{Z}_i^b)}$, we can conclude:

$$g^{opt}(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b) \propto \frac{p(\boldsymbol{Z}_i^b|\boldsymbol{Z}_{i+1}^b)}{p(\boldsymbol{Z}_i^b)}. \tag{13}$$

Now, we substitute Eq. 13 into the optimal loss objective $\mathcal{L}_{csc}^{(i,i+1),opt}$ to obtain

$$\begin{aligned}
&\mathcal{L}_{csc}^{(i,i+1),opt} \\
&= \frac{1}{2B(M-1)} \sum_{b=1}^{B} \sum_{i=1}^{M-1} -\log \frac{g^{opt}(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b)}{\sum_{j\neq i} g^{opt}(\boldsymbol{Z}_i^b, \boldsymbol{Z}_j^b)} \\
&= -\mathop{\mathbb{E}}_{\mathbf{Z}} \log \left[\frac{g^{opt}(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b)}{\sum_{j\neq i} g^{opt}(\boldsymbol{Z}_i^b, \boldsymbol{Z}_j^b)} \right] \\
&= -\mathop{\mathbb{E}}_{\mathbf{Z}} \log \left[\frac{\frac{p(\boldsymbol{Z}_i^b|\boldsymbol{Z}_{i+1}^b)}{p(\boldsymbol{Z}_i^b)}}{\sum_{j\neq i} \frac{p(\boldsymbol{Z}_i^b|\boldsymbol{Z}_j^b)}{p(\boldsymbol{Z}_i^b)}} \right] \\
&= \mathop{\mathbb{E}}_{\mathbf{Z}} \log \left[1 + \frac{p(\boldsymbol{Z}_i^b)}{p(\boldsymbol{Z}_i^b|\boldsymbol{Z}_{i+1}^b)} \sum_{j\neq i,(i+1)} \frac{p(\boldsymbol{Z}_i^b|\boldsymbol{Z}_j^b)}{p(\boldsymbol{Z}_i^b)} \right] \\
&\approx \mathop{\mathbb{E}}_{\mathbf{Z}} \log \left[1 + \frac{p(\boldsymbol{Z}_i^b)}{p(\boldsymbol{Z}_i^b|\boldsymbol{Z}_{i+1}^b)} (M-2) \mathop{\mathbb{E}}_{\mathbf{Z}_j^b} \left[\frac{p(\boldsymbol{Z}_i^b|\boldsymbol{Z}_j^b)}{p(\boldsymbol{Z}_i^b)} \right] \right] \\
&= \mathop{\mathbb{E}}_{\mathbf{Z}} \log \left[1 + \frac{p(\boldsymbol{Z}_i^b)}{p(\boldsymbol{Z}_i^b|\boldsymbol{Z}_{i+1}^b)} (M-2) \right] \\
&\geq \mathop{\mathbb{E}}_{\mathbf{Z}} \log \left[\frac{p(\boldsymbol{Z}_i^b)}{p(\boldsymbol{Z}_i^b|\boldsymbol{Z}_{i+1}^b)} (M-1) \right] \\
&= \mathop{\mathbb{E}}_{\mathbf{Z}} [\log(M-1)] - \mathop{\mathbb{E}}_{\mathbf{Z}} \log \left[\frac{p(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b)}{p(\boldsymbol{Z}_i^b)p(\boldsymbol{Z}_{i+1}^b))} \right] \\
&= \log(M-1) - \\
&\qquad \mathop{\mathbb{E}}_{\{\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b\} \sim p(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b)} \log \left[\frac{p(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b)}{p(\boldsymbol{Z}_i^b)p(\boldsymbol{Z}_{i+1}^b))} \right] \\
&= \log(M-1) - I(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b).
\end{aligned} \tag{14}$$

Therefore, $\mathcal{L}_{csc}^{(i,i+1),opt} \geq \log(M-1) - I(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b)$. Similarly, for other case, we also obtain $\mathcal{L}_{csc}^{(i+1,i),opt} \geq \log(M-1) - I(\boldsymbol{Z}_{i+1}^b, \boldsymbol{Z}_i^b)$. Combining the above two equations with $\mathcal{L}_{csc}^{opt} = \mathcal{L}_{csc}^{(i,i+1),opt} + \mathcal{L}_{csc}^{(i+1,i),opt}$, we have

$$\mathcal{L}_{csc}^{opt} \geq 2\log(M-1) - (I(\boldsymbol{Z}_i^b, \boldsymbol{Z}_{i+1}^b) + I(\boldsymbol{Z}_{i+1}^b, \boldsymbol{Z}_i^b)) \tag{15}$$

C Algorithm Description

The BrainEC-LLM algorithm comprises three main components: prompts generation (PG), multiscale decomposition mixing (MDM), and multiscale reconstruction mixing (MRM) modules. The detailed description of the algorithm can be found in Algorithm 1. Initially, BrainEC-LLM employs PG to produce relevant prior prompts. Next, MDM downsamples and decomposes fMRI time series data to obtain short-term and long-term multiscale trends, which mix the multiscale information in a

Algorithm 1 BrainEC-LLM

Input: fMRI time series data
Parameter: Parameters of prompts generation, multiscale decomposition mixing and multiscale reconstruction mixing modules: Ψ_{pg}, Ψ_{mdm}, Ψ_{mrm}, the training epochs E
Output: Brain effective connectivity network A

for $epoch = 1\ to\ E$ **do**
    Prompts Generation: Generate task description, dataset description and prior knowledge prompts to bootstrap LLM;
    Multiscale Decomposition Mixing: Downsampling and decomposition of fMRI time series data are performed as described in Eq.(2) and Eq. (3);
    Fine and coarse multiscale sequences are blended using bottom-up and top-down approaches, respectively, as Eq. (4) and Eq. (5);
    Perform cross attention on multiscale fMRI time series and pre-trained word embeddings;
    Using prompts as prefixes and multiscale fMRI embeddings as input to the LLM, fine-tune with LoRA;
    Multiscale Reconstruction Mixing: Mixing multiscale fMRI time series and extracting brain effective connectivity networks A with self attention;
end for
Post-process;
return Brain effective connectivity A

bottom-up and top-down manner, respectively. These multiscale fMRI sequences are then aligned with natural language expressions by performing cross attention with pre-trained word embeddings. The prompts are used as prefixes and concatenated with the multiscale fMRI embeddings, which are then input into the LLM and fine-tuned using LoRA. Finally, the prefix part of the LLM output is discarded, and the reconstructed fMRI time series and brain effective connectivity network A are produced by inputting the data into self attention, fusing the information from multiple scales through the MRM.

D Experimental Settings

D.1 Simulated fMRI Dataset

The benchmark simulated datasets we use are Smith dataset[4] and Sanchez dataset[5], both generated by dynamic causal model. Specifically, Sanchez dataset offers higher temporal resolution and acquisition frequency compared to Smith dataset and minimally affects the non-Gaussianity of the BOLD signal. Each session for both datasets lasted 10 minutes. Additionally, Smith dataset has repetition time (TR) of 3.00 seconds, while Sanchez dataset has TR of 1.20 seconds, reflecting current acquisition protocols with higher temporal resolution. For Smith dataset, we select Simulation 1 due to the fact that Simulation 1 is widely used as a baseline for comparison with other datasets. In Sanchez dataset, we chose Simulation 2 because it features a simple structure with two closed loops sharing a single node. This configuration is common in real brain networks and helps validate the model's performance in handling highly connected regions, known as hubs [45].

To further validate the model's accuracy, we generate a simulated dataset using CDRL[6] [79]. Considering that real resting-state fMRI dataset we use includes 7 ROIs and features cycles, the simulated dataset contains 5 and 7 nodes, named simulated complex network with node 5 (SCN5) and simulated complex network with node 7 (SCN7), respectively. The generated simulated datasets exhibit quadratic nonlinear causality, and the corresponding EC network contains cycles and more complex relationships. Specifically, the ground truth in SCN5 contains 5 nodes, while the one in SCN7 contains 7 nodes. Each dataset includes 50 samples. Details regarding the individual simulated datasets are presented in Table 6.

[4]https://www.fmrib.ox.ac.uk/datasets/netsim/index.html
[5]https://github.com/cabal-cmu/feedbackdiscovery
[6]https://github.com/huawei-noah/trustworthyAI/tree/master/datasets

Table 6: Description of the benchmark simulation dataset.

Dataset	Subjects	Data Points	Nodes	Arcs	Cycles
Smith	50	200	5	5	0
Sanchez	60	500	5	7	2
SCN5	50	200	5	8	4
SCN7	50	200	7	15	10

D.2 Real Resting-state fMRI Dataset

To assess the algorithm under real BOLD data conditions, we utilize high-resolution 7T human resting-state fMRI data[7] from the medial temporal lobe. The real resting-state fMRI dataset was band-pass filter in the range of 0.008 to 0.08 Hz without spatial smoothing to prevent signal aliasing between neighboring regions. The resulting dataset comprises 23 healthy adults, each with an acquisition time of 1.0 seconds and a session duration of 7 minutes per subject, yielding fMRI time series of 421 data points. We focus on seven regions of interest within each hemisphere of the medial temporal lobe: perirhinal cortex, divided into Brodmann areas 35 and 36 (BA35 and BA36); parahippocampal cortex (PHC); entorhinal cortex (ERC); subiculum (SUB); cornu ammonis 1 (CA1); and a region comprising CA2, CA3 and dentate gyrus together (CA23DG). Averaging the signals from CA2, CA3, and CA23DG into a single regional signal helps mitigate potential issues in connectivity estimation caused by signal mixing between adjacent regions (Smith et al., 2011), a challenge particularly pronounced in regions that are difficult to segment, such as CA2, CA3, and the dentate gyrus.

D.3 Post Process

For the obtained brain EC matrix A, a threshold θ and maximum number of parent nodes $MaxPa$ is required to convert A into a binary matrix. Given that different fMRI datasets may exhibit varying signal characteristics and noise levels, distinct thresholds need to be applied. Therefore, we set the threshold θ to be adaptive, and it is calculated as follows:

$$\theta = min(|A|) + \eta \times (max(|A|) - min(|A|)), \tag{16}$$

where η is a hyperparameter used to adjust thresholds. The maximum parent node proportion is established for the brain EC matrix, with the maximum number of parent nodes $MaxPa$ being determined by multiplying the number of brain regions N with the maximum parent node proportion $MaxPr$. For values exceeding the threshold in each row and column, the top $MaxPa$ value is set to 1, while all other values are set to 0.

D.4 Evaluation Metrics

In order to assess the effectiveness of the mothods, we use the following evaluation metrics: Precision, Recall, F1, Accuracy and Structural Hamming distance (SHD). Among them, Recall and Precision are commonly used metrics in brain effective connectivity (EC) learning and other learning tasks. F1 is a harmonic mean that combines Precision and Recall, providing a balanced assessment of both metrics. Accuracy refers to the proportion of samples in the prediction results that are correctly predicted. SHD represents the difference between the learned brain EC and ground-truth brain EC. In general, the higher the Precision, Recall, F1, Accuracy, and the lower the SHD, the better the performance of the method. The Precision, Recall, F1, Accuracy and SHD can be calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{19}$$

[7]https://github.com/shahpreya/MTlnet

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{20}$$

$$SHD = EA + MA + RA \tag{21}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively. EA, MA, RA signify extra arcs, missing arcs and reverse arcs, respectively.

D.5 Time Complexity Analysis

The model's time complexity is as follows:

- Prompt generation module: $O(N^2 \times T)$, where N and T denote the number of brain regions and the number of data points, respectively.
- Multiscale decomposition mixing module: the time complexity of downsampling is $M \times O(N \times T)$, where M denotes the number of scales; the time complexity of both bottom-up and top-down mixing is $M \times O(T^2)$, and the time complexity of cross-attention is $M \times O(d_{llm} \times d_{model}^2)$, where d_{llm} denotes the vector size of tokens in the LLM embedding space and d_{model} represents the size of the model embedding space vector.
- Multiscale reconstruction mixing: the time complexity of the linear layer is $O(T^2)$ and the time complexity of self attention is $O(N^2 \times T)$.

Therefore, the time complexity of the whole model is $O(N^2 \times T) + M \times O(N \times T) + M \times O(T^2) + M \times O(d_{llm} \times d_{model}^2) + O(llm_finetuning) + M \times O(T^2) + O(N^2 \times T) = M \times O(N^2 \times T + T^2 + llm_finetuning + d_{llm} \times d_{model}^2)$, whereas LLM is mainly made by stacking the encoder or decoder of the transformer, the number of parameters is huge and the time complexity is also very high, so it can be seen that the time complexity of our proposed module is insignificant compared to the time complexity of LLM. We conducted experiments with and without LLM on a single Nvidia L20-48GB GPU, and their efficiency results are presented in Table 7. The results indicate that the method's time complexity is primarily driven by the LLM's time complexity.

Table 7: Efficiency analysis on Smith dataset.

Methods	w/o LLM	Default
Mem.(MiB/subject)	3584	24291
Speed(s/epoch)	1.563	89.324

D.6 Training Methods

Given the limited availability of fMRI data, segmentation would further reduce the dataset size and potentially compromise the reliability of results. Moreover, since EC lacks ground truth on real fMRI datasets, we opt not to segment the dataset but instead directly train model on the available fMRI data in an autoregressive manner, obtaining both the reconstructed fMRI data and brain EC. This approach leverages the inherent autocorrelation (i.e., the brain effective connectivity network) within the fMRI time series to predict the fMRI data itself, enabling the model to learn more representative and informative connectivity structures. As a result, the model is trained by minimizing the difference between the predicted and actual fMRI signals. The EC network is generated as the outcome of this optimization process, and evaluation metrics are directly computed based on the predicted EC network without the need for a separate test set.

D.7 Hyperparameter Settings of BrainEC-LLM

The key parameter settings for BrainEC-LLM across different datasets are presented in Table 8. The llm layers differ due to the limitations of the machine's video memory capacity; they are set to the maximum number of layers that the system can support in that scenario. Although Sanchez dataset includes cycles, the relationships between the nodes are relatively simple. By incorporating acyclic constraints $\mathcal{L}_{dag}$, the complexity of BrainEC-LLM can be reduced, which helps to minimize overfitting and improve the model's generalization ability.

The choice of $\eta = 0.5$ represents a balanced trade-off between generating overly sparse or dense effective connectivity networks, which we found to work robustly across multiple fMRI datasets in our experiments. While η can indeed be tuned for specific applications, we adopted this universal value. And to ensure fair comparison, we maintain consistent evaluation criteria across all methods by applying the same threshold parameter ($\eta = 0.5$) to convert all estimated EC matrices into binary networks.

Table 8: Hyperparameters settings of BrainEC-LLM in the aforementioned experiments.

Hyperparameters	Smith	Sanchez	SCN5	SCN7
llm layers	24	18	24	12
α_{dag}	100	100	0	0
batch size	2	2	2	2
η	0.5	0.5	0.5	0.5
dropout	0.01	0.01	0.01	0.01
d model	32	32	32	32

D.8 Baseline Methods Parameter Setting

Eight baseline methods are harnessed for comparison with the proposed method, which can be categorized into two groups: machine learning methods and deep learning methods. The baseline methods are as follows: spDCM [20], lsGC [15], ACOCTE [40], RL-EC [42], CR-VAE [36], MetaCAE [30], MetaRLEC [75], CUTS+ [12]. The parameter settings for these baseline methods are shown in Table 9.

Table 9: Parameter settings of seven baseline methods

Methods	Years	Parameters
spDCM	2014	$\mathrm{nonlinear} = 0, \mathrm{two\_state} = 0, \mathrm{stochastic} = 1,$ $\mathrm{centre} = 1, \mathrm{induced} = 1, \mathrm{maxit} = 10$
lsGC	2017	$\mathrm{cmp} = 5, \mathrm{ARorder} = 2, \mathrm{normalize} = 1$
ACOCTE	2022	$\alpha = 1.0, \beta = 2.0, q_0 = 0.98, \rho = 0.2$
RL-EC	2022	$\mathrm{nh} = 256, \mathrm{heads} = 16, \mathrm{stacks} = 6, \mathrm{nh}_{\mathrm{decoder}} = 16$
CR-VAE	2023	$\mathrm{context} = 20, \lambda = 0.1, \mathrm{lr} = 0.05, \mathrm{nh} = 64$
MetaCAE	2024	$\mathrm{nh} = 64, \alpha = 0.05, \beta = 20.0, \mathrm{k} = 3, \mathrm{d} = 4, \mathrm{lr}_1 = 0.02,$ $\mathrm{lr}_2 = 0.02, \mathrm{lr}_3 = 0.001, \mathrm{lr}_{\mathrm{main}} = 0.002$
MetaRLEC	2024	$\mathrm{heads}_{\mathrm{encoder}} = 8, \mathrm{blocks}_{\mathrm{encoder}} = 3, \mathrm{dropout}_{\mathrm{encoder}} = 0.1,$ $\mathrm{lr}_{\mathrm{actor}} = 10^{-4}, \mathrm{lr}_{\mathrm{critic}} = 10^{-3}$
CUTS+	2024	$\mathrm{layers}_{\mathrm{gru}} = 1, \mathrm{lr}_{\mathrm{stage1}} = 10^{-3} \to 10^{-4}, \mathrm{lr}_{\mathrm{stage2}} = 10^{-2} \to 10^{-3}$

E Experiments

E.1 t-test Analysis

To assess the significant differences between BrainEC-LLM and other baseline methods, t-test analysis and Wilcoxon test with Holm correction are performed, with results presented in Table 10 and 11. p-values less than 0.05 indicates a significant difference at the 95% confidence level. The analysis in t-test reveals that, except for RL-EC, which shows no significant difference in the evaluation metric SHD, BrainEC-LLM is significantly different from the baseline methods in all other cases. The Wilcoxon test results are consistent with those of the t-test, confirming the absence of a significant difference in SHD between BrainEC-LLM and RL-EC. In addition, the Wilcoxon test indicates no significant difference in recall between BrainEC-LLM and CR-VAE. However, significant differences are observed in the other evaluation metrics, further demonstrating the superior performance of our method.

Table 10: p-values obtained from the t-test for BrainEC-LLM and seven baseline methods. Underlined values indicate no significant difference at the 95% confidence level.

Methods	Precision	Recall	F1	Accuracy	SHD
spDCM	1.30×10^{-3}	2.34×10^{-3}	2.87×10^{-5}	1.63×10^{-3}	4.23×10^{-2}
lsGC	2.48×10^{-5}	4.85×10^{-3}	1.27×10^{-5}	6.88×10^{-4}	2.75×10^{-2}
ACOCTE	1.63×10^{-3}	2.80×10^{-4}	4.61×10^{-6}	1.51×10^{-3}	4.45×10^{-2}
RL-EC	9.07×10^{-3}	3.95×10^{-5}	4.58×10^{-6}	3.17×10^{-3}	$\underline{6.75 \times 10^{-2}}$
CR-VAE	1.62×10^{-3}	1.34×10^{-3}	5.16×10^{-6}	9.52×10^{-4}	4.17×10^{-2}
MetaCAE	3.47×10^{-4}	2.93×10^{-3}	1.88×10^{-6}	6.18×10^{-4}	3.47×10^{-2}
MetaRLEC	3.74×10^{-4}	1.16×10^{-3}	8.19×10^{-7}	4.77×10^{-4}	3.44×10^{-2}
CUTS+	1.40×10^{-4}	1.46×10^{-3}	8.85×10^{-7}	2.84×10^{-4}	3.07×10^{-2}

Table 11: p-values obtained from the Wilcoxon test for BrainEC-LLM and seven baseline methods. Underlined values indicate no significant difference at the 95% confidence level.

Methods	Precision	Recall	F1	Accuracy	SHD
spDCM	1.61×10^{-2}	6.84×10^{-3}	4.88×10^{-4}	2.44×10^{-3}	4.88×10^{-4}
lsGC	4.88×10^{-4}	1.47×10^{-3}	4.88×10^{-4}	9.77×10^{-4}	4.88×10^{-4}
ACOCTE	3.13×10^{-2}	4.88×10^{-4}	9.77×10^{-4}	1.61×10^{-2}	6.84×10^{-3}
RL-EC	1.20×10^{-2}	4.88×10^{-4}	3.42×10^{-3}	5.62×10^{-3}	$\underline{3.42 \times 10^{-2}}$
CR-VAE	9.77×10^{-4}	$\underline{8.98 \times 10^{-1}}$	1.46×10^{-3}	9.77×10^{-4}	9.77×10^{-4}
MetaCAE	3.42×10^{-3}	4.65×10^{-3}	3.42×10^{-3}	1.95×10^{-3}	3.42×10^{-3}
MetaRLEC	4.88×10^{-3}	4.88×10^{-4}	4.88×10^{-4}	4.88×10^{-4}	4.88×10^{-4}
CUTS+	4.88×10^{-4}	3.42×10^{-3}	4.88×10^{-4}	4.88×10^{-4}	4.88×10^{-4}

E.2 Visualization of LLM Inputs

Figure 7 illustrates the evolution of fMRI embeddings during training, demonstrating how BrainECLLM learns meaningful representations. Since prompts generation module is frozen, prompt tokens remain fixed throughout training, allowing us to focus solely on visualizing the fMRI multi-scale embeddings in LLM input. As shown in Figure 7, at the early stages of training (e.g., epoch 50), BrainEC-LLM has not yet fully learned, resulting in predominantly negative feature values. As training progresses and the number of epochs increases, the model gradually converges, and the feature representations shift towards more meaningful directions.

E.3 ABIDE I Visualization

Figure 8 illustrates a comparative analysis of brain effective connectivity (EC) networks in healthy controls (HCs) versus individuals with autism spectrum disorder (ASD), based on AAL-90 atlas (90 regions) [13]. To analyze the most critical effective connectivity between brain regions, we preserved the top 5% of edges with the highest scores for subsequent analysis and visualization. The six figures illustrate the binarized brain EC networks, derived by averaging the original EC networks of healthy controls and ASD patients from the ABIDE I dataset, respectively. The nodes represent brain regions, while the edges indicate the effective connectivity. The node size reflects the centrality or importance of each region within the network. The other six heatmaps display three types of brain EC networks for both healthy controls and ASD patients respectively: the original brain EC networks, the top 5% of the strongest connections in the original brain EC networks, and the top 5% of the strongest connections in the binarized brain EC networks.

Previous literature has consistently reported that individuals with Autism Spectrum Disorder (ASD) exhibit increased connectivity in the right middle temporal gyrus (MTG) [71], decreased connectivity in the temporal pole (TPO) [23] and supplementary motor area (SMA) [63], and asymmetric connectivity alterations in the middle frontal gyrus (MFG), with reduced connectivity in the left hemisphere

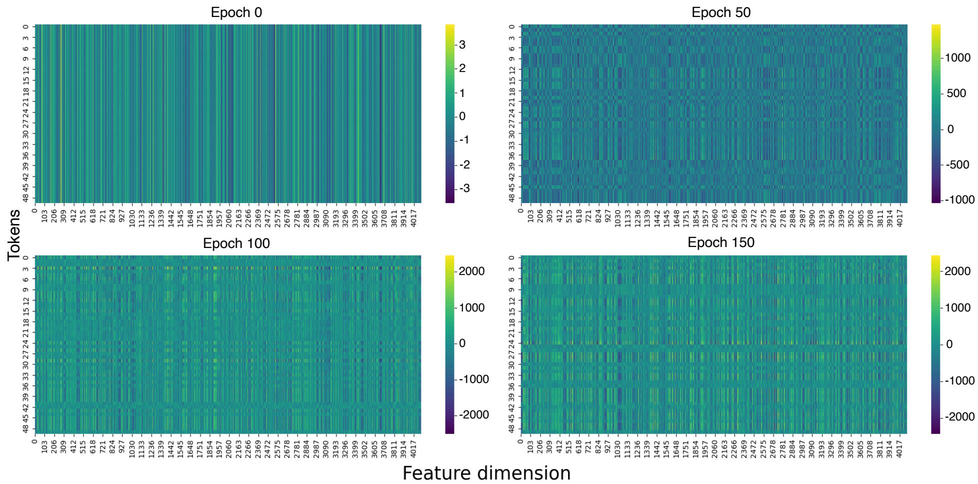


Figure 7: Visualization of fMRI multiscale embedding in LLM inputs.

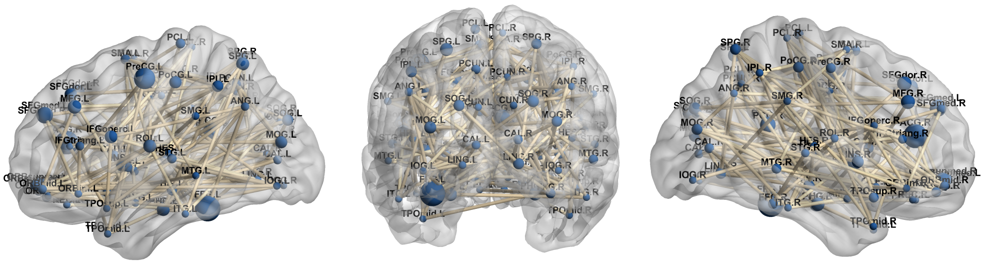
and increased connectivity in the right hemisphere [23, 28]. The connectivity patterns observed in the figure align with these previous findings, reinforcing the notion that these regions play critical roles in the neural underpinnings of ASD.

Middle Temporal Gyrus (MTG). The MTG is fundamentally involved in supporting a range of social and linguistic processes [67]. Evidence from a facial recognition study demonstrated that individuals with congenital prosopagnosia showed diminished neural activation and weaker connectivity within the anterior temporal lobe, which highlights the region's critical role in facial information processing [3]. Additionally, research has indicated that the recognition of both faces and words relies on high-resolution visual representations, shaped by the close functional coupling of visual and language-related brain regions and the neural tendency to preserve short inter-regional pathways [50].

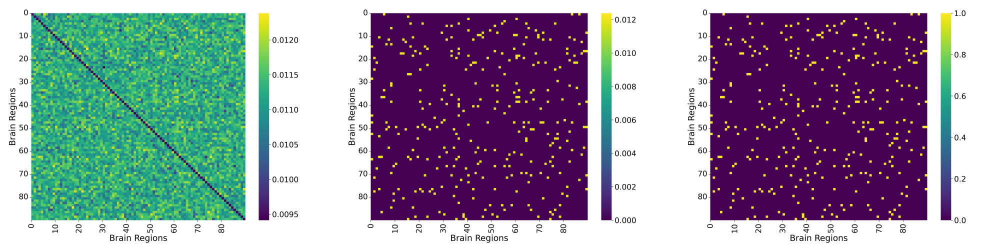
Temporal Pole (TPO). The TPO is a vital structure within the social brain network, is essential for the theory of mind (ToM), a fundamental cognitive skill that enables individuals to comprehend and infer the mental states of others, as well as anticipate their actions [1, 66]. A substantial body of research has highlighted that disruptions in both the functional and structural integrity of the TPO are likely linked to the social difficulties observed in individuals with ASD [49], who frequently face considerable difficulties when engaging in ToM-related tasks.

Supplementary Motor Area (SMA). The SMA, located on the medial aspect of the frontal lobe anterior to the primary motor cortex, is primarily associated with motor planning and execution. In children with ASD, reduced functional connectivity between the SMA and social-related brain regions has been observed, which may contribute to deficits in social interaction and communication [55]. Compared to healthy controls, individuals with ASD exhibit reduced functional connectivity in SMA neural circuits and diminished SMA activation [35, 38]. Such disruptions are potentially linked to abnormal semantic verb processing, which may hinder their ability to construct coherent discourse and develop adequate language proficiency, ultimately affecting behavioral patterns and the capacity for effective social engagement [26].

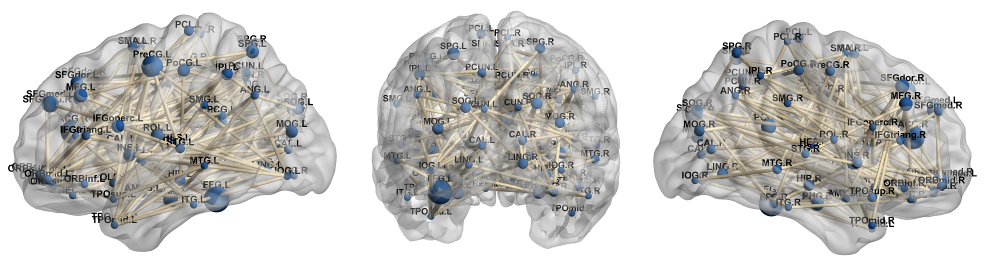
Middle Frontal Gyrus (MFG). Numerous neurocognitive and neuroimaging studies have demonstrated that MFG is associated with the pathophysiology of ASD [76, 21]. MFG is integral to executive functions, including working memory and attention control [29]. Research has identified the MFG as one of the abnormal brain regions within the occipital pole network that is associated with social and communication deficits in individuals with ASD [27]. The MFG is primarily involved in coordinating and integrating various types of information, playing a key role in executive function and cognitive control [29]. Prior studies have also reported that individuals with ASD often exhibit impairments in information integration and processing, which may contribute to difficulties in effective communication, particularly in public or socially demanding environments [53].



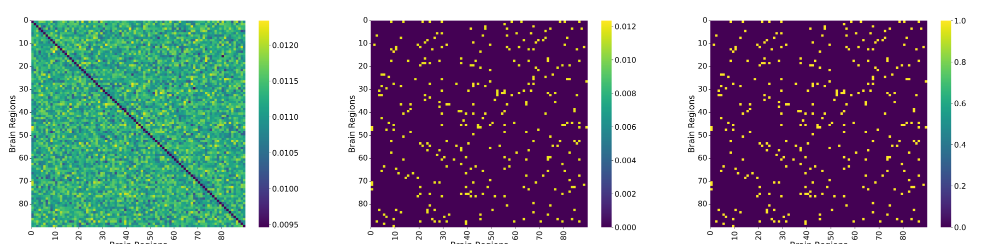
(a) Healthy controls. (b) Healthy controls. (c) Healthy controls.



(d) Original brain EC network heatmap (healthy controls). (e) Top 5% original brain EC heatmap (healthy controls). (f) Top 5% binarized brain EC heatmap (healthy controls).



(g) ASD. (h) ASD. (i) ASD.



(j) Original brain EC network heatmap (ASD). (k) Top 5% original brain EC heatmap (ASD). (l) Top 5% binarized brain EC heatmap (ASD).

Figure 8: Comparison of brain EC network between healthy controls and ASD patients on the ABIDE I dataset Using BrainEC-LLM.

E.4 Ablation Study

We perform ablation studies on model backbone, model module, and the loss function, with the results presented in Table 12. In the Table, "Transformer Decoder" indicates that the LLM is replaced with a conventional transformer decoder. "w/o LLM" refers to completely removing the pre-trained LLM from the BrainEC-LLM framework. Since the cross attention is specifically designed for the LLM to reduce cross-modal discrepancies, it is also removed. "Spearman Correlation Matrix" and "Kendall Correlation Matrix" refer to replacing the Pearson Correlation Matrix in the Prompt Generation module with the corresponding correlation matrix.

Table 12: Ablations on Smith and Sanchez dataset. The best and second-best values are **highlighted** and underlined.

Variant	Smith					Sanchez				
	Precision↑	Recall↑	F1↑	Accuracy↑	SHD↓	Precision↑	Recall↑	F1↑	Accuracy↑	SHD↓
Llama3-8B (Default)	0.57±0.06	**0.80±0.15**	**0.67±0.11**	**0.84±0.06**	**4.02±0.84**	**0.78±0.02**	0.97±0.02	**0.86±0.07**	**0.91±0.05**	**2.10±0.75**
Mistral-7B	0.29±0.07	0.40±0.10	0.33±0.13	0.68±0.08	7.65±1.36	0.48±0.05	0.43±0.03	0.46±0.12	0.72±0.07	7.13±1.24
Transformer Decoder	0.44±0.14	0.78±0.11	0.56±0.12	0.74±0.06	6.16±0.92	0.39±0.13	0.82±0.09	0.53±0.11	0.71±0.07	6.35±0.88
Spearman Correlation Matrix	0.43±0.09	0.62±0.13	0.51±0.17	0.76±0.08	6.14±0.91	0.39±0.07	0.67±0.11	0.47±0.18	0.72±0.06	6.45±0.88
Kendall Correlation Matrix	0.51±0.06	0.76±0.12	0.61±0.15	0.81±0.08	5.23±1.06	0.53±0.05	0.78±0.13	0.63±0.16	0.83±0.07	5.12±1.10
w/o Prompts Generation	0.38±0.04	0.60±0.15	0.46±0.14	0.72±0.08	7.10±0.93	0.52±0.06	0.57±0.04	0.54±0.11	0.72±0.06	6.92±0.86
w/o Multiscale Mixing(no $\mathcal{L}_{csc}$)	0.20±0.08	0.21±0.12	0.20±0.13	0.68±0.06	7.88±0.95	0.41±0.05	0.29±0.04	0.34±0.08	0.68±0.06	7.94±1.19
w/o Cross Attention	0.23±0.12	0.32±0.13	0.27±0.12	0.53±0.06	11.80±0.93	0.43±0.04	0.35±0.08	0.39±0.10	0.71±0.07	7.24±0.93
w/o LLM	0.33±0.11	0.42±0.14	0.35±0.14	0.72±0.08	7.14±1.20	0.50±0.07	0.55±0.06	0.52±0.13	0.76±0.06	7.04±1.16
w/o Task Description	0.33±0.07	0.25±0.13	0.28±0.11	0.76±0.04	6.32±0.97	0.35±0.06	0.27±0.12	0.30±0.10	0.74±0.05	6.15±1.02
w/o Dataset Description	0.15±0.13	0.20±0.12	0.17±0.18	0.62±0.07	10.31±1.42	0.08±0.14	0.28±0.11	0.12±0.17	0.73±0.06	8.95±1.45
w/o Pearson Correlation	0.43±0.08	0.63±0.09	0.51±0.13	0.76±0.06	6.11±1.26	0.32±0.07	0.71±0.10	0.42±0.12	0.85±0.05	7.25±1.30
w/o L_{spa}	0.53±0.06	0.62±0.15	0.57±0.14	0.81±0.05	5.28±0.94	0.45±0.07	0.70±0.14	0.50±0.13	0.88±0.04	6.15±0.90
w/o L_{dag}	0.34±0.12	0.22±0.14	0.26±0.17	0.77±0.08	6.42±1.23	0.28±0.11	0.15±0.13	0.18±0.15	0.85±0.07	7.35±1.20
w/o $\mathcal{L}_{csc}$	**0.58±0.04**	0.70±0.16	0.57±0.09	0.78±0.07	4.84±1.26	0.67±0.04	**0.98±0.03**	0.76±0.10	0.82±0.07	4.34±0.93

Our results indicate that including $\mathcal{L}_{csc}$ provides a slight improvement over not including it. For other components, such as the prompts generation module and multiscale mixing module (includes both multiscale decomposition and mixing, without loss $\mathcal{L}_{csc}$), the improvements are more pronounced in BrainEC-LLM. Additionally, the results using Llama 3 as the backbone are manifestly outperforms Mistral.

E.5 Prompt Complexity Analysis

We conduct experiments to examine the effect of prompt complexity, using both simplified and complex prompts. The results are presented in Table 13, and examples of complex prompt is shown in Figure 6 (b). These results demonstrate that all components of the prompt are crucial, with the absence of the task description and dataset description leading to a significant decline in performance metrics.

Table 13: Sensitivity analysis experiment on the simplicity or complexity of prompts.

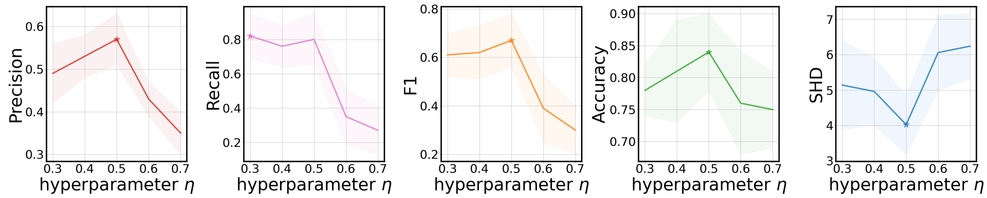
Variant	Smith					Sanchez				
	Precision↑	Recall↑	F1↑	Accuracy↑	SHD↓	Precision↑	Recall↑	F1↑	Accuracy↑	SHD↓
Llama3-8B (Default)	0.57±0.06	**0.80±0.15**	**0.67±0.11**	**0.84±0.06**	**4.02±0.84**	0.78±0.02	**0.97±0.02**	**0.86±0.07**	**0.91±0.05**	**2.10±0.75**
w/o Task Description	0.33±0.07	0.25±0.13	0.28±0.11	0.76±0.04	6.32±0.97	0.35±0.06	0.27±0.12	0.30±0.10	0.74±0.05	6.15±1.02
w/o Dataset Description	0.15±0.13	0.20±0.12	0.17±0.18	0.62±0.07	10.31±1.42	0.08±0.14	0.28±0.11	0.12±0.17	0.73±0.06	8.95±1.45
w/o Pearson Correlation	0.43±0.08	0.63±0.09	0.51±0.13	0.76±0.06	6.11±1.26	0.32±0.07	0.71±0.10	0.42±0.12	0.85±0.05	7.25±1.30
Complex Prompt	**0.62±0.10**	0.61±0.12	0.61±0.14	0.83±0.06	4.23±0.92	**0.83±0.06**	0.71±0.06	0.76±0.11	0.88±0.07	3.17±0.95

E.6 Hyperparameter Analysis

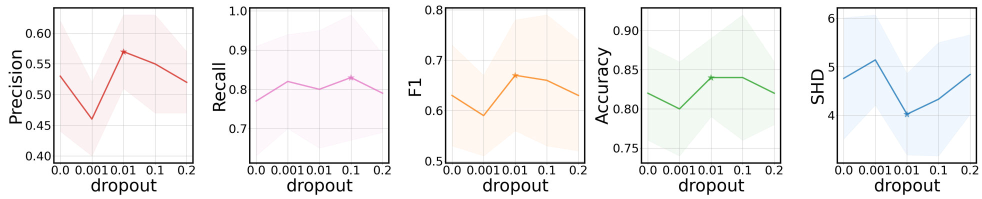
We conduct experiments on Smith datasets to evaluate the parameter sensitivity of BrainEC-LLM. Figures 9 (a) and 9 (b) display the experimental results for soft threshold hyperparameters and dropout variations, respectively. When the soft threshold is set to 0.5, all metrics are optimal except for Recall, which remains sub-optimal, a pattern also observed with dropout. Consequently, the soft threshold is ultimately set to 0.5, and dropout is set to 0.01.

As shown in Figure 9 (c), (d), and (e), we also conduct sensitivity analysis for the loss weight coefficients. Given the different magnitudes of each loss term, we carefully selected distinct ranges for the hyperparameter settings in our experiments. For instance, we set α_{spa} to {0.1, 1, 2, 10, 100} and α_{dag} to {1, 10, 100, 500, 1000}. For α_{spa}, performance peaks at 2, while the best results for α_{dag} and α_{csc} are achieved at 100 and 10, respectively. Both overly small and large values degrade performance, highlighting the need for balanced regularization. We therefore set $\alpha_{spa} = 2$, $\alpha_{dag} = 100$, and $\alpha_{csc} = 10$ in all experiments.

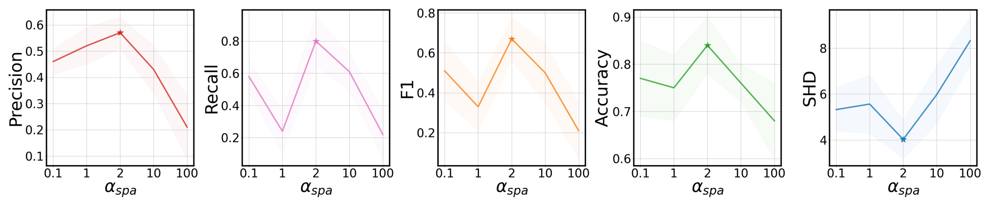
We also conduct additional experiments with a refined search around the optimal region of soft threshold hyperparameter η. Since the best performance occurs at $\eta = 0.5$ in Figure 9 (a), we perform



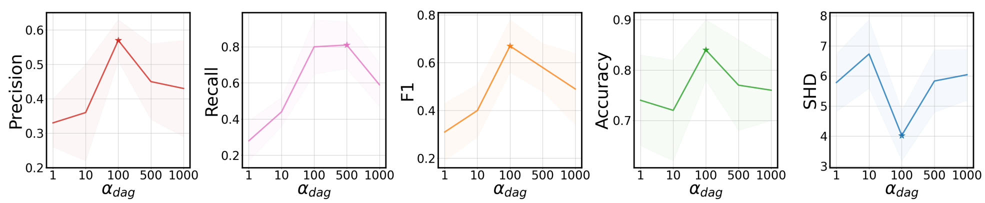
(a) Influence of the soft threshold hyperparameter η.



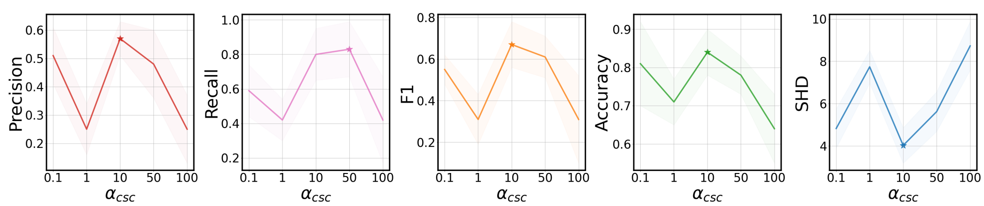
(b) Influence of the dropout.



(c) Influence of the loss weights coefficient α_{spa}.



(d) Influence of the loss weights coefficient α_{dag}.



(e) Influence of the loss weights coefficient α_{csc}.

Figure 9: Hyperparameter analysis on the Smith dataset, where the starred results are the best results.

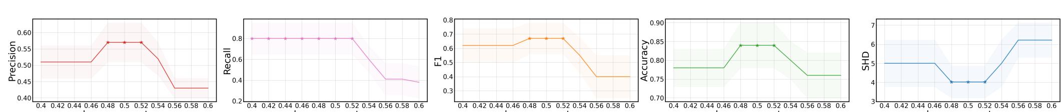


Figure 10: Influence of the soft threshold hyperparameter η with fine-grained step.

a fine-grained search from $\eta = 0.4$ to $\eta = 0.6$ with step 0.02. The results are presented in Figure 10. The refined search shows that optimal performance is consistently achieved at $\eta \in$ {0.48,0.50,0.52} across all metrics. The choice of $\eta = 0.5$ is particularly intuitive given the threshold formulation: $\theta = min(|A|) + \eta \times (max(|A|) - min(|A|))$. When $\eta = 0.5$, the threshold is set to the midpoint between the minimum and maximum absolute values in the adjacency matrix, which represents a natural balance point for binary classification.

F Limitation

One potential limitation that warrants further investigation is that existing LLMs are primarily pre-trained on massive textual corpora and therefore inherently optimized for processing and reasoning over natural language. In contrast, fMRI time series data are continuous, high-dimensional numerical signals that exhibit complex temporal and spatial dependencies. This modality gap introduces challenges in effectively aligning the statistical properties of fMRI signals with the natural language representations used in LLMs. As a result, current LLMs may not fully capture the fine-grained temporal dynamics and domain-specific patterns embedded in fMRI data.