This CVPR Findings paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

BLEG: LLM Functions as Powerful fMRI Graph-Enhancer for Brain Network Analysis

Rui Dong Zitong Wang Jiaxing Li Weihuang Zheng Youyong Kong*
School of Computer Science and Engineering, Southeast University
{dongrui_0427,220242336,jiaxing_li,zhengweihuang,kongyouyong}@seu.edu.cn

Abstract

Graph Neural Networks (GNNs) have been widely used in diverse brain network analysis tasks based on preprocessed functional magnetic resonance imaging (fMRI) data. However, their performances are constrained due to high feature sparsity and inherent limitations of domain knowledge within uni-modal neurographs. Meanwhile, large language models (LLMs) have demonstrated powerful representation capabilities. Combining LLMs with GNNs presents a promising direction for brain network analysis. While LLMs and MLLMs have emerged in neuroscience, integration of LLMs with graph-based data remains unexplored. In this work, we deal with these issues by incorporating LLM's powerful representation and generalization capabilities. Considering great cost for directly tuning LLMs, we instead ***function LLM as enhancer*** *to boost GNN's performance on downstream tasks. Our method, namely* ***BLEG****, can be divided into three stages. We firstly prompt LLM to get augmented texts for fMRI graph data, then we design a "LLM-LM" instruction tuning method to get enhanced textual representations at a relatively lower cost. GNN is trained together for coarsened alignment. Finally we finetune an adapter after GNN for given downstream tasks. Alignment loss between LM and GNN logits is designed to further enhance GNN's representation. Extensive experiments on different datasets confirmed BLEG's superiority. Code can be available at https://github.com/KamonRiderDR/BLEG.*

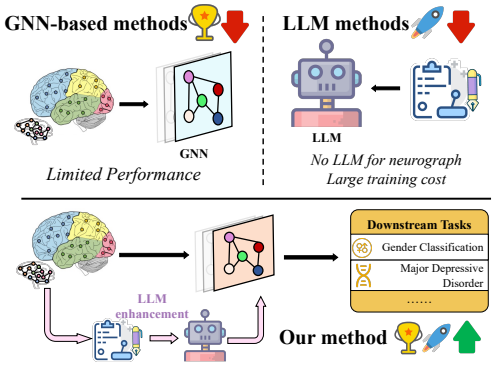


Figure 1. A illustration of our method. GNN-based methods have limited performance. LLM methods require great training cost. Our method aims to enhance GNN's performance with much less training cost.

1. Introduction

Brain network analysis holds significant importance for investigating intrinsic mechanisms of human brains and diagnosing neurological diseases. Deep learning-based methods have emerged as the predominant approach for brain network analysis, demonstrating superior performance for different tasks (gender classification [35], major depressive disorder diagnosis [6], autism spectrum disorder diagnosis [13], etc). Among these approaches, graph

*Corresponding author

neural networks (GNNs) have achieved remarkable success [10, 17, 36]. By operating message passing mechanism on fMRI-derived brain networks, GNN can effectively exploit their intrinsic topological features [3, 26]. Current GNN approaches mainly focus on two directions: (a) designing powerful modules to capture finer features (DTN [42], A-GCL [46], BrainNPT [15], etc). (b) enhancing model interpretability (BrainGNN [20], IBGNN [5], ContrastPool [41], etc).

Despite their success, performances of these methods are constrained by inherent limitations in brain graph data. Limited sample size [45] and sparse feature for preprocessed neurograph hinder further capability for data-driven deep learning methods [39, 43]. Meanwhile, brain network data are inherently confined to certain neuroimaging data, which lacks domain knowledge that is not explicitly encoded in imaging data. These features are further sparsified by preprocessing pipeline, leading to certain information loss; both factors jointly constrain GNN-based brain network analysis. In short, These inherent data limitations together pose challenges for more accurate GNN-based brain network analysis.

Meanwhile, the emergence of large language models (LLMs) has achieved remarkable success in Natural Language Processing (NLP) domain, for their exceptional representation reasoning and generalization capabilities [1, 9]. Existing LLM methods for neuroscience and brain network analysis mainly directly utilize LLMs' capabilities for different tasks which focus on single text modality (e.g. BioGPT [25], BioBERT [18]). They either employ multimodal language models for data from different modalities to help diagnosis [2, 29], including medical images (e.g. MedBLIP [2], LLaVA-Med [19]), electrophysiological recordings (e.g. BrainBERT [37]) and structured clinical data [16, 48]. However, cases are more complex when dealing with graph data, which contains both node and structure features. Currently, exploration of LLM applications into graph-based brain network analysis remains unexplored, and a combination of LLMs and brain GNNs deserves further research.

Hence, in this paper, we pioneer the integration of LLMs with GNN-based neuroscience tasks. Instead of using LLM as a decoder, we regard LLM as an enhancer, hoping to utilize its embeddings to enhance GNN's representation learning (shown in Fig. 1). However, directly tuning for LLM is costly. Thus, we design a novel framework which realizes **L**anguage-**E**nhanced **G**raph Neural Network for **B**rain Network Analysis (**BLEG**). Our BLEG can be divided into three stages: (1) We prompt LLM to generate augmented text description data for input graph. Each brain graph is prompted as text format. (2) We tune a smaller LM based on the text-graph dataset from previous stage. GNN encoder is also trained for coarse alignment. (3) We conduct supervised fine-tuning on GNN for different downstream tasks. Logits from tuned LM is utilized for fine-grained alignment.

Besides, we provide a theoretical analysis to demonstrate effectiveness of BLEG. By using LLM and LM as an enhancer, GNN can learn better representation beneficial for given downstream tasks. Extensive experiments on various real-world datasets to illustrate BLEG's superior performance on different tasks (ASD diagnosis, MDD diagnosis, etc). To the best of our knowledge, this is the first attempt to improve GNN brain networks' performance by exploring LLM methods. Notably, here only fMRI data is used while our BLEG is data-agnostic and model-agnostic, and we believe that it provides new insight for both research and real-scene applications.

2. Related Works

GNN-based Brain network analysis. Graph neural networks (GNNs) follow message passing mechanism which aggregates its neighborhood nodes before updating the value of its node feature (GCN [17], GAT [36], GraphSAGE [10]). GNN-based methods in neuroscience can be mainly divided into two categories: (a) Designing modules to capture deeper features within brain graphs [34, 42]. A-GCL employs adversarial module to enhance GNN's performance [46]. BrainNPT uses transformer as backbone to capture long-range features [15]. (b) Enhancing model interpretability. BrainGNN is a classical interpretable brain

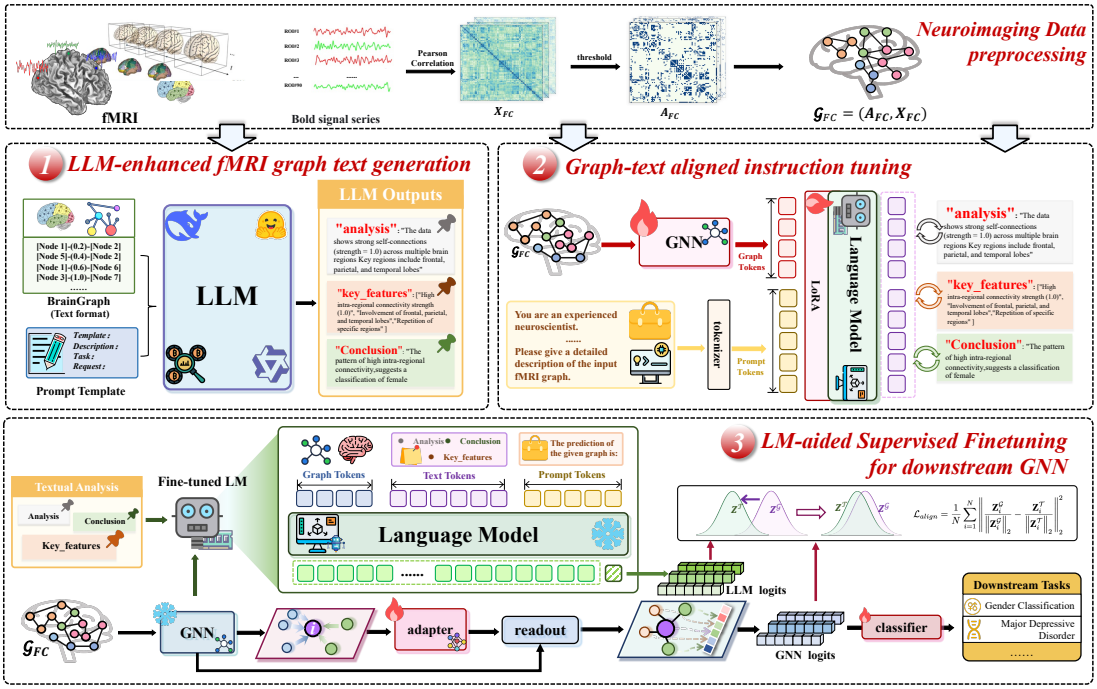


Figure 2. The overall framework of BLEG. (1) We prompt LLM to generate augmented text data for input graph. (2) A smaller LM is trained through instruction tuning based on generated textual data. GNN is trained together for coarsened alignment. (3) GNN tuning for certain tasks. Logit from LM is utilized to boost GNN's representations.

graph neural network [20]. IBGNN employs GNN backbone and generator for explanation [5]. ContrastPool is a differentiable graph pooling method which can realize explainable classification.

LLM methods in neuroscience. Based on pretrained LMs (GPT, BERT), methods like BioGPT [25], BioBERT [18] train a medical language model from medical datasets like Pubmed. Meanwhile, the advent of multimodal LLMs (MLLMs) can utilze information from other modalities to capture complementary features. LLaVA-Med realizes a large Language-and-Vision assistant for biomedicine [19]. MMGPL [29] designs graph prompt for fMRI images for better text-image fusion. Other methods consider using more modalities (video [31], structured clinical data [48]) and new model architectures (Mixture-of-Experts, MoE) for scalability [16].

3. Methodology

In this section, we will discuss more details of BLEG. As is shown in Fig. 2, BLEG can be divided into three stages: (a) We firstly **conduct augmented text-graph dataset** by prompting LLM for every fMRI brain graph. (b) Then we finetune a smaller LM based on the conducted dataset through **instruction tuning**. We also train a GNN model for coarse-grained text-graph alignment. And finally (c) we perform **supervised fine-tuning** for given downstream task on pretrained GNN. We add a trainable adapter while keeping GNN weights frozen. Alignment loss is designed between GNN and LM logits for fine-grained alignment.

3.1. Notations

Brain graph datasets are preprocessed from functional magnetic resonance imaging (fMRI, also write as FC). Pipeline for Data preprocessing is

shown in Fig. 2. AAL template is employed for FC data, and finally we denote brain network as $\mathcal{G}_i = \{X_i, A_i, \mathcal{V}_i, \mathcal{E}_i\}$, where nodes $\mathcal{V}_i$ represents different brain regions in the brain while edges $\mathcal{E}_i$ represents the connectivity between two brain regions. X_i denotes node features for each region and A_i denotes adjacency matrix. The dataset can be written as $\mathcal{D} = \{\mathcal{G}_i, y_i\}_{i=1}^N$, where $y_i \in \{0, 1\}$ which stands for different tasks (gender classification, MDD diagnosis, ASD diagnosis, etc). In general, Brain network analysis can be seen as a graph classification task:

$$\mathcal{F}(\mathcal{G}_i) = \text{GNN}(\mathcal{G}_i) \rightarrow \hat{y}_i \tag{1}$$

$\mathcal{F}$ can be any deep learning method and we hope to learn an optimal function to predict y. Here we use GNN as our $\mathcal{F}(\cdot)$, whose message passing function can be written as Eq. 2, where h_i^l is the representation for node v_i in l-th layer, **AGG**$(\cdot)$ stands for the aggregation of its neighborhood nodes $\mathcal{N}_i$, and **UPDATE**$(\cdot)$ means the update operation of h_i.

$$h_i^{l+1} = \textbf{UPDATE}(h_i^l,\ \textbf{AGG}(\{h_j^l | j \in \mathcal{N}_i\})) \tag{2}$$

3.2. LLM-enhanced text data generation

The core idea of this stage is to leverage LLM's capability for downstream GNN. However, for neuroscience, text data is limited where data forms as graph. Meanwhile, regular manual process for creating such text data for each brain network is time-consuming and requires too much human labor. Thus, we propose to fully utilize existing LLMs to generate augmented textual information for each brain network graph.

Prompt design. Compared to other graphs like social networks, fMRI-based FC graphs have more medical meanings, where each node stands for specific brain regions while edge weights represent connection strengths between two brain regions. Thus, we design a prompt suited for input graph $\mathcal{G}_i$. Formally, for given $\mathcal{G}_i$, our designed prompt consists of three parts: description $\mathcal{P}_i^D$, graph data ($\mathcal{P}_i^G$) and query ($\mathcal{P}_i^Q$). $\mathcal{P}_i^D$ depicts basic medical information from given brain network, including which dataset it belongs to, preprocess template, names of each brain region in neuroscience and possible downstream tasks. Design for prompts of input graph is challenging which may consider topological information during graph-text transformation [32, 47]. Unlike other graph learning tasks, brain networks have more concrete medical meanings, where $\mathcal{E}_{ij}$ stands for connection strength between region i and region j. On the other hand, brain graph is sparser compared to other graph data, with a limited number of edges. Thus, for its structure prompt, we consider modifying graph as "Node[i]-x-Node[j]" format for each $\mathcal{E}_{ij}$, where x means its connection strength. For node feature, we consider its mean value ($X^d \rightarrow X^1$). Structure prompt and feature prompt together make $\mathcal{P}_i^G$. For query $\mathcal{P}_i^Q$, we require LLM output json format data containing analysis, key features and conclusion for input FC graph.

These different prompts together make our prompt: $(\mathcal{P}_i^D, \mathcal{P}_i^G, \mathcal{P}_i^Q) \rightarrow \mathcal{P}_i$, which is used as LLM's input. Here we select Deepseek-v3 [21] as our LLM API. For each $\mathcal{G}_i$, we feed corresponding $\mathcal{P}_i$ into LLM and get response ($\mathcal{T}_i$):

$$\mathcal{T}_i = \textbf{LLM}(\mathcal{P}_i) \tag{3}$$

Data verification and refinement. For each brain network $\mathcal{G}_i$, we get its LLM-enhanced text data. To further improve quality of LLM generation, we utilize LLM (QwQ-32B [33]) for judgment and refinement. Quality of $\mathcal{T}_i$ is measured and scored from different dimensions (professional expressions, content relevance, generation repetition). $\mathcal{T}_i$ with lower scores are then refined by LLM. Remaining generations are corrected via professional medical experts. Here that we don't focus on design of data verification process, which can be left as future work. Finally, we curate a high-quality text-enhanced graph dataset: $\mathcal{D}_i' \leftarrow (\mathcal{G}_i, \mathcal{T}_i)$.

3.3. Graph-text aligned instruction tuning

Directly training for LLM is extremely challenging, which requires great expense for training. Meanwhile, some LLMs like GPT4, Deepseek-v3 only provides generated text, while in some cases LLM embeddings are necessary. Here, we choose to tune a smaller language model (LM) instead of directly tuning LLM [12]. For enhanced graph-text

data-pair $(\mathcal{G}_i, \mathcal{T}_i)$, we feed them into LM for tuning. However, there exists certain modality gap between graph and text representations on high-dimensional manifold space. Encouraged by LLaVA [22], we use GNN's embeddings and textual embeddings together as LM's input. As is shown in Fig. 2, we keep both GNN encoder and LM trainable to achieve coarsened alignment between graph and text embeddings.

$$\mathbf{H}_i = \mathbf{LM}(\mathbf{X}_i^{\mathcal{G}}, \mathbf{X}_i^{\mathcal{T}}) = \mathbf{LM}(f_\phi(\mathcal{G}_i), \mathbf{X}_i^{\mathcal{T}}) \quad (4)$$

For tuning method for GNN and LM, here we use instruction tuning. LM generates answers $(\mathbf{X}^{\mathcal{A}})$ for given graph data $(\mathbf{X}^{\mathcal{G}})$ and question $(\mathbf{X}^{\mathcal{Q}})$. Questions are queries requiring LM to make detailed description on given brain network. The answers are from the augmented text generated by LLM from previous stage, Generally, the input embeddings from Eq. 4 can be written as follows:

$$\mathbf{X}_i = \text{concat}(f_\phi(\mathcal{G}_i), \mathbf{X}_i^{\mathcal{Q}}, \mathbf{X}_i^{\mathcal{A}}) \quad (5)$$

LM instruction tuning is trained through auto-regressive loss, and we only compute loss function and optimize models based on answer tokens, which can be formatted as Eq. 6:

$$p(\mathbf{X}^{\mathcal{A}}|\mathbf{X}^{\mathcal{G}}, \mathbf{X}^{\mathcal{Q}}) = \prod_{i=1}^{L} p_\theta(x^i|\mathbf{X}^{\mathcal{G}}, \mathbf{X}^{\mathcal{Q},<i}, \mathbf{X}^{\mathcal{A},<i}) \quad (6)$$

Through graph-text aligned LM instruction tuning, we can obtain textual representations for given brain graph at a much smaller training cost.

3.4. LM-aided finetuning for GNN

The final stage of BLEG is LM-aided supervised finetuning for different downstream tasks. LM logit is utilized to assist downstream GNN for better representation. To be specific, we save weights of GNN encoder and LM after instruction tuning and keep them frozen in this stage. As is shown in Fig. 2, we add a trainable adapter after frozen GNN, which is a two-layer FFN (denote as g_φ) for implementation. The embeddings of graph embeddings can be formatted as:

$$\mathbf{Z}_i = f_\phi \circ g_\varphi(\mathcal{G}_i) \quad (7)$$

For $\mathbf{Z}_i \in \mathbb{R}^{N\times d}$, we use READOUT$(\cdot)$ function to get graph-level logits $(\mathbf{Z}_i^{\mathcal{G}})$. The READOUT function incorporates residual connection along with batch normalization (Eq. 8).

$$\mathbf{Z}_i^{\mathcal{G}} = \text{READOUT}(\text{Norm}(\mathbf{Z}_i + \mathbf{X}_i^{\mathcal{G}})) \quad (8)$$

Finally $\mathbf{Z}_i^{\mathcal{G}}$ is fed into a trainable classification head for prediction, with cross-entropy loss used for model optimization $(\mathcal{L}_{CE})$.

To further enhance GNN's ability to capture text-augmented representation, we introduce an auxiliary alignment loss $(\mathcal{L}_{align})$ between text $(\mathbf{Z}_i^{\mathcal{T}})$ and graph logits $(\mathbf{Z}_i^{\mathcal{G}})$. $\mathbf{Z}_i^{\mathcal{T}}$ is obtained through tuned LM, the input is the same format as Eq. 4, where $\mathbf{X}_i^{\mathcal{Q}}$ is about give the prediction result of the input brain network. We add a `cls` token at the end of each input sequence whose output logit is used as $\mathbf{Z}_i^{\mathcal{T}}$ for fine-grained graph-text alignment. For implementation of $\mathcal{L}_{align}$, we use MSE$(\cdot)$ for alignment at high manifold dimension (Eq. 9).

$$\mathcal{L}_{align} = \frac{1}{N}\sum_{i=1}^{N}\left\| \frac{\mathbf{Z}_i^{\mathcal{G}}}{\|\mathbf{Z}_i^{\mathcal{G}}\|_2} - \frac{\mathbf{Z}_i^{\mathcal{T}}}{\|\mathbf{Z}_i^{\mathcal{T}}\|_2} \right\|_2^2 \quad (9)$$

The overall loss function is composed of $\mathcal{L}_{CE}$ and $\mathcal{L}_{align}$, weighted by coefficient $\alpha \in (0, 1)$:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha \cdot \mathcal{L}_{align} \quad (10)$$

4. Theoretical Analysis

The main idea of theoretical analysis is that through LLM and LM as enhancer, the performance of GNN will be improved for given downstream tasks as its representations contains augmented text information which is useful for downstream classification.

Theorem 4.1 ***(Complementary Representations from LM for GNN)*** *We define representation from original GNN $(\boldsymbol{X}^{\mathcal{G}})$ and fine-tuned LM $(\boldsymbol{X}^{\mathcal{T}})$, LM-distilled GNN representation is denoted as $\boldsymbol{X}^{\mathcal{G}'}$. Downstream label representation is denoted as Y. Given above assumptions, we have: $\left\| I(\boldsymbol{X}^{\mathcal{G}'}; Y) - I(\boldsymbol{X}^{\mathcal{G}}, \boldsymbol{X}^{\mathcal{T}}; Y) \right\| \leq C \cdot \epsilon$, where C is a constant and $\epsilon > 0$. Thus for $I(\boldsymbol{X}^{\mathcal{G}'}; Y)$ and $I(\boldsymbol{X}^{\mathcal{G}}; Y)$, $I(\boldsymbol{X}^{\mathcal{G}'}; Y) > I(\boldsymbol{X}^{\mathcal{G}}; Y)$.*

Theorem. 4.1 shows that through BLEG, GNN can capture complementary information from LM, which enhances its capability for downstream tasks.

5. Experiments

5.1. Experimental settings

Datasets. Our experiments were performed on four public real-world brain network datasets: Autism Brain

Table 1. Comparison results on public datasets. We run 10 times for each model and record the corresponding average acc ± std (%). The best results are marked **bold**, and second best underline.

Methods	HCP		ADHD		MDD		ABIDE	
	ACC	SEN	ACC	SEN	ACC	SEN	ACC	SEN
GCN	64.03 ± 1.21	55.51 ± 2.88	66.74 ± 1.47	31.83 ± 2.33	62.38 ± 0.37	70.88 ± 1.94	69.14 ± 0.84	64.07 ± 2.20
GAT	65.72 ± 0.67	60.91 ± 3.26	66.56 ± 0.30	32.31 ± 1.56	63.05 ± 0.35	69.30 ± 2.16	68.79 ± 0.74	63.83 ± 3.23
GraphSAGE	66.87 ± 0.32	61.32 ± 2.71	67.80 ± 0.37	31.52 ± 2.54	63.29 ± 0.27	69.48 ± 1.53	70.74 ± 0.89	65.52 ± 2.81
GraphTrans	67.46 ± 0.55	62.71 ± 3.17	67.00 ± 0.37	31.85 ± 2.54	63.37 ± 0.27	68.31 ± 1.93	68.41 ± 1.20	62.79 ± 2.75
BrainGNN	66.46 ± 2.12	62.92 ± 2.45	67.16 ± 2.01	32.75 ± 3.31	63.25 ± 1.06	68.91 ± 2.76	70.03 ± 1.69	62.80 ± 4.19
IBGNN	64.72 ± 1.04	55.98 ± 4.01	65.59 ± 0.49	30.84 ± 2.92	63.07 ± 0.29	68.38 ± 2.47	66.02 ± 1.18	61.31 ± 3.61
BrainNPT	67.78 ± 1.53	60.67 ± 1.34	67.84 ± 2.11	32.18 ± 2.19	63.81 ± 0.77	69.55 ± 3.07	68.80 ± 1.10	59.52 ± 3.05
THFCN	67.29 ± 1.74	59.82 ± 2.84	66.73 ± 1.40	33.01 ± 3.69	62.41 ± 0.67	66.59 ± 2.17	66.93 ± 1.29	62.37 ± 2.27
ContrastPool	68.10 ± 1.74	63.59 ± 1.21	65.10 ± 0.82	35.92 ± 4.21	64.05 ± 0.47	66.22 ± 3.68	69.89 ± 0.88	66.71 ± 2.79
TAPE	69.32 ± 1.41	63.80 ± 2.93	67.41 ± 1.06	36.59 ± 2.11	64.12 ± 0.82	68.50 ± 2.79	70.43 ± 0.95	66.64 ± 2.95
OFA	71.00 ± 0.82	63.04 ± 2.42	68.22 ± 0.70	39.72 ± 1.94	62.97 ± 1.21	68.55 ± 2.40	69.72 ± 1.48	65.93 ± 3.11
BLEG (Ours)	**71.21 ± 0.91**	**68.27 ± 3.39**	**69.41 ± 0.53**	**41.11 ± 2.10**	**65.63 ± 0.48**	**71.12 ± 2.97**	**72.21 ± 0.80**	**67.16 ± 2.30**
Δ GCN	**7.18** ↑	**12.76** ↑	**2.67** ↑	**9.28** ↑	**3.25** ↑	**0.24** ↑	**3.07** ↑	**3.09** ↑

Table 2. Comparison results on private dataset. We run 10 times for each model and record the corresponding average acc ± std (%). The best results are marked **bold**, and second best underline.

Methods	GCN	GraphTrans	BrainGNN	BrainNPT	ContrastPool	BLEG (medium)	BLEG (large)	Qwen3-8B
ACC	71.79 ± 1.07	70.50 ± 0.73	71.06 ± 1.05	70.25 ± 1.55	71.11 ± 0.73	75.38 ± 0.63	75.21 ± 1.01	**75.82 ± 0.77**
SEN	85.70 ± 2.46	86.07 ± 1.84	84.00 ± 1.81	83.00 ± 1.59	86.85 ± 1.94	**87.03 ± 0.89**	86.77 ± 1.27	**87.03 ± 1.58**
SPE	50.60 ± 5.41	50.92 ± 3.03	48.77 ± 2.12	49.21 ± 3.30	52.55 ± 3.91	58.31 ± 2.72	58.50 ± 1.93	**58.81 ± 2.00**
F1	78.50 ± 0.67	78.43 ± 0.49	77.18 ± 1.92	77.27 ± 0.74	78.43 ± 0.49	79.35 ± 1.66	**80.78 ± 1.72**	79.81 ± 1.58
AUC	68.15 ± 1.71	65.70 ± 1.16	66.39 ± 0.61	67.53 ± 1.03	68.70 ± 1.16	70.77 ± 1.16	**70.78 ± 1.24**	69.43 ± 2.03

Imaging Data Exchange (**ABIDE**, 618 subjects) [7], Human Connectsome Project (**HCP**, 1039 subjects) [35], Attention Deficit Hyperactivity Disorder (**ADHD**, 938 subjects) [4] and Rest-meta-MDD (**MDD**, 2165 subjects) dataset. We also use one private dataset **zhongdaxinxiang** (short as **ZDXX**, 520 subjects), which is collected from Zhongda Hospital of Southeast University, the Second Affiliated Hospital of Xinxiang Medical University and Hangzhou Hospital. Due to data privacy concerns, in first and second stage, we only use public datasets for text generation and LM instruction tuning.

Baselines. We select nine representative baselines for comparison, which can mainly be divided into two types: **(a) GNNs based methods**, including GCN [17], GAT [36], GraphSAGE [10] and GraphTrans [40]. **(b) Brain Network based methods**, including classical methods (BrainGNN [20] and IBGNN [5]) and latest brain network methods (BrainNPT [15], THFCN [34] and ContrastPool [41]). **(c) LLM-GNN methods**: TAPE [12] and OFA [23]. Code implementations of all methods are taken from their original papers.

Experimental settings. For LLM, we select Deepseek-v3 to generate augmented text data. We select BioGPT-base as our instruction tuning LM, whose parameter is 347M in total. For GNN encoder. We choose a 3-layer GCN. Our model is implemented in PYG and trained on RTX Titans with 24GB memory. For instruction tuning, we tune our LM and GNN on four public datasets whose number of total sample is 4760 and tuning epoch is from 3 to 5. For supervised fine-tuning, the total training epoch is 150 with 50 as early stopping.

We evaluate models' performance on five metrics: accuracy (**ACC**), sensitivity (**SEN**), specificity (**SPE**), f1 score (**F1**) and ROC-AUC (**AUC**), where higher value means better performance. We record ACC and SEN for public datasets, while all five metrics on zhongdaxinxiang. For evaluation on all the methods, we use 10-fold cross validation on ten random runs and record mean value and standard deviation.

5.2. Comparison results

Comparison results on public datasets. Comparison results on public four datasets are presented in Tab. 1. The results show that our BLEG outperforms all other methods on all the datasets. Compared to GCN which also works as our backbone, the maximum improvement

of ACC can be 7.18 ↑ on HCP dataset. It also exceeds SOTA brain network analysis method (ContrastPool) at 3.11 ↑ on ACC.

Comparison results on private dataset. Comparison results on private dataset zhongdaxinxiang are presented in Tab. 2. Here we select two additional LMs with larger parameters for tuning: BioGPT-1.5B and Qwen3-8B. We employ Low-Rank Adaptation (LoRA) during training [14]. BLEG also achieves best results on all evaluation metrics, although we did not use it for augmented-text generation and instruction tuning. Its maximum accuracy improvement over vanilla GCN reaches 4.03 ↑.

5.3. Few-shot experiment results

LLMs have shown remarkable performance in few-shot learning settings. Thus, to further illustrate capability of BLEG, we construct few-shot splitting of datasets and test BLEG's performance in few-shot cases.

We first set a train ratio to gradually decrease number of training samples for given dataset. As is shown in Fig. 4, we set training ratio from 10% to 70%. We set a fixed validation size (10%) and the rest data is for test. Results from Fig. 4 show that compared to other methods (GNN-based BrainGNN and transformer-based BrainNPT), BLEG can always show a leading advantage.

We also conduct k-shot experiments on BLEG. For given dataset, we randomly select k samples from each label as training set ($k \in \{1, 2, 5\}$). For testing set, we select 50, 100 and rest data ($L_{\mathcal{D}}$) for each label respectively. As is shown in Tab. 3, BLEG outperforms other methods under extreme few-shot cases. The results demonstrate BLEG's superiority under few-shot scenes.

5.4. Ablation studies & Sensitivity analyses

We conduct ablation studies to analyze whether each submodule of BLEG works. We directly train our GCN on downstream tasks to verify effectiveness of $\mathcal{L}_{align}$. We also test performance for vanilla BioGPT to verify if instruction tuning stage works. The results in Fig. 4 (f) shows that **w/o align loss**, accuracy of BLEG will witness a great decrease. Meanwhile, $\mathcal{L}_{align}$ on BioGPT **w/o tuning** will also influence model's performance.

For sensitivity analyses, we record BLEG's accuracy under different align loss coefficient ($\alpha \in [0.2, 0.7]$) and plot following figures (Fig. 4 (a) - (b)). The results demonstrate that the optimal value of α for achieving highest accuracy varies across different datasets. On the other hand, despite variations in different α, BLEG consistently outperforms other methods (BrainGNN, BrainNPT) in most cases, further demonstrating the effectiveness of LM-aided representation enhancement.

Table 3. k-shot few-shot experiments. We run ten times and record average accuracy.

Methods	N_{test} \ k	HCP 1	HCP 2	HCP 5	zhongdaxinxiang 1	zhongdaxinxiang 2	zhongdaxinxiang 5
BrainGNN	50	55.00	57.00	56.30	52.00	49.00	53.00
	100	51.00	54.50	57.00	52.00	48.50	50.00
	$L_{\mathcal{D}}$	47.88	48.75	50.81	45.56	55.62	53.43
BrainNPT	50	57.00	56.00	59.00	53.00	52.50	54.50
	100	56.00	52.50	55.50	52.40	51.00	50.00
	$L_{\mathcal{D}}$	56.41	56.40	57.46	48.76	55.63	55.13
BLEG	50	**61.00**	**59.00**	**60.30**	**58.00**	**58.50**	**60.00**
	100	**58.50**	**57.50**	**59.00**	**54.50**	**54.00**	**56.00**
	$L_{\mathcal{D}}$	**56.81**	**56.64**	**57.53**	**52.23**	**56.01**	**56.52**

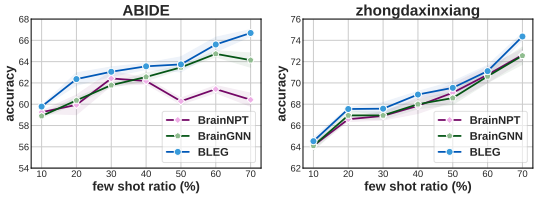


Figure 4. Few shot results on different ratios.

5.5. Empirical studies

We conduct biomarker visualization of brain regions for empirical studies. We average embeddings after GNN encoder for all samples and statistically analyze and visualize the top 10 brain regions. The results are shown in Fig. 3. Visualizations for top 10 regions in ABIDE dataset are: Precuneus (L), Precuneus (R), Inferior parietal, but supramarginal and angular gyri, which are consistent with findings in [27], Superior frontal gyrus and medial orbital which show additional abnormalities between HC and ASD patients [38], Amygdala (L) and Amygdala (R) whose atypical activation can occur between patients [28], Heschl gyrus which is positively related to ASD symptoms [44], Anterior cingulate, paracingulate gyri and Parahippocampal gyrus, which are consistent with studies in [11]. For ZDXX dataset, brain regions like Posterior cingulate gyrus, Precuneus, Anterior cingulate and paracingulate gyri which are among top-10 values indicate their relations to excessive fluctuations of FC in DMN-related regions [24]. Other top-10 regions are consistent with prior findings regarding identification of salient brain regions [8, 30].

5.6. Text generation for LM

Here we mainly utilize LLM as enhancer which functions in embedding level. Yet we also select Qwen3-8B

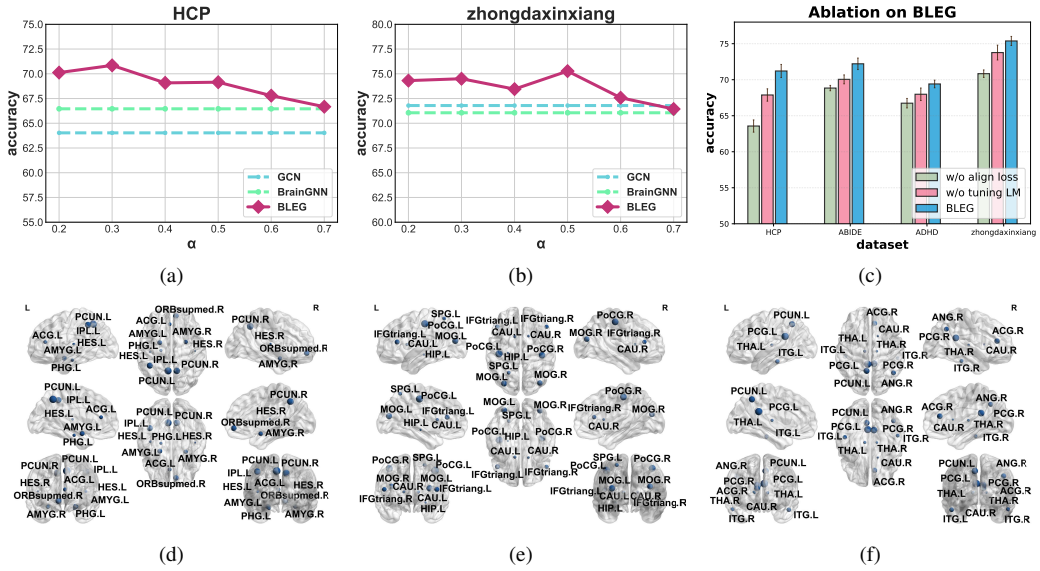


Figure 3. (a)–(b) k-shot experiments on different datasets. (c) Ablation studies on different datasets. (d) Biomarker visualizations on ABIDE, (e) ADHD and (f) zhongdaxinxiang datasets.



Figure 5. Text generation for ZDXX dataset from LM.

for tuning which has great capability of instruction following. The aim is to explore possibility of interpretable brain network analysis. As is shown in Fig. 5, compared to vanilla Qwen, tuned LLM can generate more professional analysis on unseen private data, with more confidence for brain disease diagnosis judgment (MDD). It suggests a promising future of BLEG for further explainable and generalizable brain network analysis. Modern LM have the ability to capture deeper domain-specific knowledge from public datasets which can lead to better performance on private dataset.

6. Conclusion

In this work, we propose BLEG, a novel method that functions LLM as a powerful enhancer to boost GNN's performance in brain network analysis. Instead of directly training LLMs, we adopt a LLM-LM paradigm to leverage enhanced textual representations more efficiently. LM-aided SFT further enhances GNN's capability for downstream tasks. Extensive experiments on five datasets demonstrate effectiveness of BLEG, including different few-shot experiments which confirmed BLEG's great generalization. Our BLEG is a first attempt to leverage LLM with GNN-based brain network analysis and we think it can provide new insight for both research and real-world medical diagnose applications.

Acknowledgement

This work is supported by National Natural Science Foundation of China (Grant No.62471133). This work is also supported by the Big Data Computing Center of Southeast University.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Qiuhui Chen and Yi Hong. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. In *Proceedings of the Asian Conference on Computer Vision*, pages 2404–2420, 2024. 2

[3] Jonathan D Cohen, Nathaniel Daw, Barbara Engelhardt, Uri Hasson, Kai Li, Yael Niv, Kenneth A Norman, Jonathan Pillow, Peter J Ramadge, Nicholas B Turk-Browne, et al. Computational approaches to fmri analysis. *Nature neuroscience*, 20 (3):304–313, 2017. 2

[4] ADHD-200 consortium. The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in systems neuroscience*, 6:62, 2012. 6

[5] Hejie Cui, Wei Dai, Yanqiao Zhu, Xiaoxiao Li, Lifang He, and Carl Yang. Interpretable graph neural networks for connectome-based brain disorder analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 375–385. Springer, 2022. 2, 3, 6

[6] Lulu Cui, Shu Li, Siman Wang, Xiafang Wu, Yingyu Liu, Weiyang Yu, Yijun Wang, Yong Tang, Maosheng Xia, and Baoman Li. Major depressive disorder: hypothesis, mechanism, prevention and treatment. *Signal transduction and targeted therapy*, 9(1):30, 2024. 1

[7] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6): 659–667, 2014. 6

[8] Selene Gallo, Ahmed El-Gazzar, Paul Zhutovsky, Rajat M Thomas, Nooshin Javaheripour, Meng Li, Lucie Bartova, Deepti Bathula, Udo Dannlowski, Christopher Davey, et al. Functional connectivity signatures of major depressive disorder: machine learning analysis of two multicenter neuroimaging studies. *Molecular Psychiatry*, 28(7):3013–3022, 2023. 7

[9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[10] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1024–1034, 2017. 2, 6

[11] Teruo Hashimoto, Susumu Yokota, Yutaka Matsuzaki, and Ryuta Kawashima. Intrinsic hippocampal functional connectivity underlying rigid memory in children and adolescents with autism spectrum disorder: A case–control study. *Autism*, 25 (7):1901–1912, 2021. 7

[12] Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. Harnessing explanations: Llm-to-lm interpreter for enhanced text-attributed graph representation learning. *arXiv preprint arXiv:2305.19523*, 2023. 4, 6

[13] Tomoya Hirota and Bryan H King. Autism spectrum disorder: a review. *Jama*, 329(2):157–168, 2023. 1

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 7

[15] Jinlong Hu, Yangmin Huang, Nan Wang, and Shoubin Dong. Brainnpt: Pre-training transformer networks for brain network classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024. 2, 6

[16] Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. *arXiv preprint arXiv:2404.10237*, 2024. 2, 3

[17] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017. 2, 6

[18] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. 2, 3

[19] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Nau-

mann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 2, 3

[20] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, 74:102233, 2021. 2, 3, 6

[21] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 4

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 5

[23] Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks, 2024. 6

[24] Yicheng Long, Hengyi Cao, Chaogan Yan, Xiao Chen, Le Li, Francisco Xavier Castellanos, Tongjian Bai, Qijing Bo, Guanmao Chen, Ningxuan Chen, et al. Altered resting-state dynamic functional brain networks in major depressive disorder: Findings from the rest-meta-mdd consortium. *NeuroImage: Clinical*, 26:102163, 2020. 7

[25] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022. 2, 3

[26] Andrea I Luppi, Helena M Gellersen, Zhen-Qi Liu, Alexander RD Peattie, Anne E Manktelow, Ram Adapa, Adrian M Owen, Lorina Naci, David K Menon, Stavros I Dimitriadis, et al. Systematic evaluation of fmri data-processing pipelines for consistent functional connectomics. *Nature Communications*, 15(1):4745, 2024. 2

[27] Aarthi Padmanabhan, Charles J Lynch, Marie Schaer, and Vinod Menon. The default mode network in autism. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2(6):476–486, 2017. 7

[28] Mahie Patil, Nofel Iftikhar, and Latha Ganti. Neuroimaging insights into autism spectrum disorder: Structural and functional brain. *Health Psychology Research*, 12:123439, 2024. 7

[29] Liang Peng, Songyue Cai, Zongqian Wu, Huifang Shang, Xiaofeng Zhu, and Xiaoxiao Li. Mmgpl: Multimodal medical data analysis with graph prompt learning. *Medical Image Analysis*, 97: 103225, 2024. 2, 3

[30] Daniel Porta-Casteràs, Marta Cano, Joan A Camprodon, Colleen Loo, Diego Palao, Carles Soriano-Mas, and Narcís Cardoner. A multimetric systematic review of fmri findings in patients with mdd receiving ect. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 108:110178, 2021. 7

[31] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024. 3

[32] Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 491–500, 2024. 4

[33] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, 2025. Accessed:2025-09-22. 4

[34] Yingzhi Teng, Kai Wu, Jing Liu, Yifan Li, and Xiangyi Teng. Constructing high-order functional connectivity networks with temporal information from fmri data. *IEEE Transactions on Medical Imaging*, 2024. 2, 6

[35] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013. 1, 6

[36] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018. 2, 6

[37] Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. *arXiv preprint arXiv:2302.14367*, 2023. 2

[38] Charles SE Weston. Four social brain regions, their dysfunctions, and sequelae, extensively explain autism spectrum disorder symptomatology. *Brain sciences*, 9(6):130, 2019. 7

[39] Peter Wills and François G Meyer. Metrics for graph comparison: a practitioner's guide. *Plos one*, 15(2):e0228728, 2020. 2

[40] Zhanghao Wu, Paras Jain, Matthew Wright, Azalia Mirhoseini, Joseph E Gonzalez, and Ion Stoica. Representing long-range context for graph neural networks with global attention. *Advances in neural information processing systems*, 34:13266–13279, 2021. 6

[41] Jiaxing Xu, Qingtian Bian, Xinhang Li, Aihu Zhang, Yiping Ke, Miao Qiao, Wei Zhang, Wei Khang Jeremy Sim, and Balázs Gulyás. Contrastive graph pooling for explainable classification of brain networks. *IEEE Transactions on Medical Imaging*, 2024. 2, 6

[42] Xi Yang, Yan Jin, Xiaobo Chen, Han Zhang, Gang Li, and Dinggang Shen. Functional connectivity network fusion with dynamic thresholding for mci diagnosis. In *Machine Learning in Medical Imaging: 7th International Workshop, MLMI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Proceedings 7*, pages 246–253. Springer, 2016. 2

[43] Jia Yi, Huilin Jiang, Xiaoyong Wang, and Yong Tan. A comprehensive review on sparse representation and compressed perception in optical image reconstruction. *Archives of Computational Methods in Engineering*, 31(5):3197–3209, 2024. 2

[44] Anyi Zhang, Lin Liu, Suhua Chang, Le Shi, Peng Li, Jie Shi, Lin Lu, Yanping Bao, and Jiajia Liu. Connectivity-based brain network supports restricted and repetitive behaviors in autism spectrum disorder across development. *Frontiers in psychiatry*, 13:874090, 2022. 7

[45] Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, et al. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*, 2024. 2

[46] Shengjie Zhang, Xiang Chen, Xin Shen, Bohan Ren, Ziqi Yu, Haibo Yang, Xi Jiang, Dinggang Shen, Yuan Zhou, and Xiao-Yong Zhang. A-gcl: Adversarial graph contrastive learning for fmri analysis to diagnose neurodevelopmental disorders. *Medical Image Analysis*, 90:102932, 2023. 2

[47] Yizhuo Zhang, Heng Wang, Shangbin Feng, Zhaoxuan Tan, Xiaochuang Han, Tianxing He, and Yulia Tsvetkov. Can llm graph reasoning generalize beyond pattern memorization? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2289–2305, 2024. 4

[48] Wenhao Zheng, Liaoyaqi Wang, Dongshen Peng, Hongxia Xu, Yun Li, Hongtu Zhu, Tianfan Fu, and Huaxiu Yao. Multimodal clinical trial outcome prediction with large language models. *arXiv preprint arXiv:2402.06512*, 2024. 2, 3