Dx-LLM: Two-layer Retrieval-Augmented Multilingual Diagnosis System

Anonymous ACL submission

Abstract

Automatic diagnosis (AD) represents a pivotal area in healthcare, where patient symptoms are analyzed for disease diagnosis. Traditional approaches depend on extracting features from symptoms and diseases within collected patient cases. However, real-life patient data collection poses challenges, often resulting in incomplete clinical datasets that can lead to misdiagnosis, especially for new diseases or unrecorded symptoms. Recently, retrieval-augmented large language models (RA-LLMs) have shown significant promise in addressing knowledge-intensive Natural Language Processing (NLP) tasks. To mitigate reliance on previously seen data, we propose a two-layer AD system, termed Dx-LLM, leveraging RA-LLMs. Dx-LLM first constructs a disease-symptom knowledge graph from an external dataset of disease symptom descriptions and conducts initial disease filtering to identify potential candidate diseases based on patient symptoms. Subsequently, in the second layer, we utilize the robust language understanding and generation capabilities of LLMs to re-rank these candidates, thereby producing refined diagnostic outcomes. This two-layer approach reduces the computational load on the second-layer LLM by narrowing down the disease candidates in the first layer. Our results demonstrate that Dx-LLM achieves hit@10 scores of 71.41% and 70.38% across 1058 diseases in English and Chinese datasets, consistently outperforming state-of-the-art baselines.

1 Introduction

The rapid development of artificial intelligence (AI) has been revolutionizing the healthcare industry and automating a wide spectrum of tasks. One notable AI-driven healthcare application is automatic diagnosis (AD), which applies machine learning algorithms to help doctors diagnose diseases based on patient symptoms. Despite substantial progress in this field, current AD methods depend heavily on the quality and quantity of training data. It limits their ability to generalize to public patients, where new diseases or unrecorded symptoms can be present. Existing AD methods can be divided into two main categories: graph-based and LLM-based approaches. Graph-based methods model health data into graphs and perform diagnoses through graph representation learning (Hosseini et al., 2018; Wang et al., 2021). LLM-based approaches design domain-specific LLMs to resolve sub-tasks in the medical domain (Shoham and Rappoport, 2023; Tu et al., 2024). However, the graph-based approaches are heavily restricted by training data, which results in reliable diagnosis only for seen diseases and recorded symptoms. The domain-specific LLMs, though have brought in rich reliable domain-related knowledge, are not fine-tuned specially for AD tasks, thus achieving an inferior performance.

The emergence of Retrieval-Augmented Large Language Models (RA-LLMs) has introduced new potential for AD tasks (Brown et al., 2020; Fan et al., 2024; Zhao et al., 2023), They leverage the strong language understanding and generation capabilities of LLMs, addressing issues such as hallucination and outdated information by retrieving reliable domain-specific knowledge. Additionally, rich external resources alleviate the limitations posed by relying solely on limited patient cases for accurate diagnosis. However, there are two challenges for directly applying RA-LLMs to AD tasks:

- **First**, directly applying RA-LLMs to raw disease and symptom information for ranking and inference is computationally prohibitive. Pre-processing and filtering are necessary to first curate a quality candidate disease set for efficient inference;
- **Second**, retrieving information that accounts for the varying importance of symptoms is

crucial but challenging. Current methods typically select relevant information based solely on semantic similarity, neglecting distinct importance levels of symptoms.

In this paper, we introduce Dx-LLM, a two-layer multilingual disease diagnosis system powered by RA-LLMs. Dx-LLM employs a two-tiered approach to identify relevant diseases based on patient symptoms. The first layer performs coarse-grained disease identification using a knowledge graph constructed from external disease symptom descriptions. The second layer re-ranks these candidates and refines the diagnostic outcomes using RA-LLMs. This two-layer mechanism, enhanced by a Heterogeneous Information Networks (HIN) module, is designed to tackle the aforementioned challenges. First, the HIN module represents the constructed knowledge graph via a variational graph auto-encoder (VGAE) (Kipf and Welling, 2016) and generates coarse-grained disease candidates, preventing RA-LLMs from processing all information without pre-processing. Second, the HIN embeddings enable RA-LLMs to thoroughly assess the relevance of various information, effectively addressing the different importance levels of symptoms. The framework of Dx-LLM is shown in Figure 1.

Our contributions can be summarized as follows:

- We applied RA-LLMs and constructed a disease-symptom HIN graph based on the external Mayo Clinic dataset for symptom and disease representation learning, which overcomes the reliance on the limited collection of real-life patient diagnosis data.
- We designed a two-stage diagnosis system that takes advantage of LLMs' language understanding and generation ability while restricting the inference time of LLMs by selecting high-quality candidate diseases after first-layer graph mining.
- Our proposed Dx-LLM system can realize multi-lingual diagnosis and can achieve 71.41% and 70.38% of hit@10 among 1058 diseases in English dataset and Chinese dataset correspondingly, and consistently outperforms other state-of-the-art (SOTA) baseline models.

2 Related Work

2.1 AI for Automatic Diagnosis

Recent AI-based Automatic Diagnosis (AD) methods include graph-based approaches and LLM-based approaches. Graph-based approaches resolve AD problems by converting the health data into graph structures. (Hosseini et al., 2018) proposed Heteromed, which models the high-dimensional Electronic Healthcare Records (EHRs) data with Heterogeneous Information Network (Shi et al., 2016) and applied Graph Convolutional Transformer (GCT) (Choi et al., 2019) and attention Graph Convolutional Networks (GCN) (Hosseini et al., 2019) to embed the nodes. (Wang et al., 2021) organized Electronic Healthcare Records (EHRs) into a heterogeneous graph that can model interactions among users, symptoms, and diseases to resolve the cold start problem in GCN. With the recent development of LLMs, LLM-based AD approaches have emerged. (Shoham and Rappoport, 2023) proposed CPLLM, which fine-tunes LLMs with historical diagnosis records, and demonstrates its superiority in clinical prediction tasks. (Wang et al., 2023) proposed Coad which introduced a disease and symptom collaborative generation framework. (Tu et al., 2024) proposed AMIC, which is a medical knowledge graph based on patients' transcription data. Different from existing approaches, Dx-LLM constructs a RA-LLM-based system, which leverages external knowledge and LLM's semantic understanding and generation ability for efficient and accurate diagnosis.

2.2 Retrieval-Augmented Large Language Models

Recently, LLMs have demonstrated great potential in language understanding and generation in various application fields (Brown et al., 2020; Fan et al., 2024; Zhao et al., 2023). However, they still suffer from challenging problems including lacking domain-specific knowledge, hallucination, and containing out-of-date information. To address this problem, Retrieval Augmented Generation (RAG) has been applied to LLMs and promoted a line of research around Retrieval-Augmented Large Language Models (RA-LLM). (Lewis et al., 2020) improved the pre-trained language model's performance in knowledge-intensive NLP tasks by introducing RAG models where the parametric memory is a pre-trained seq2seq model and the non-parametric memory is a dense vector index

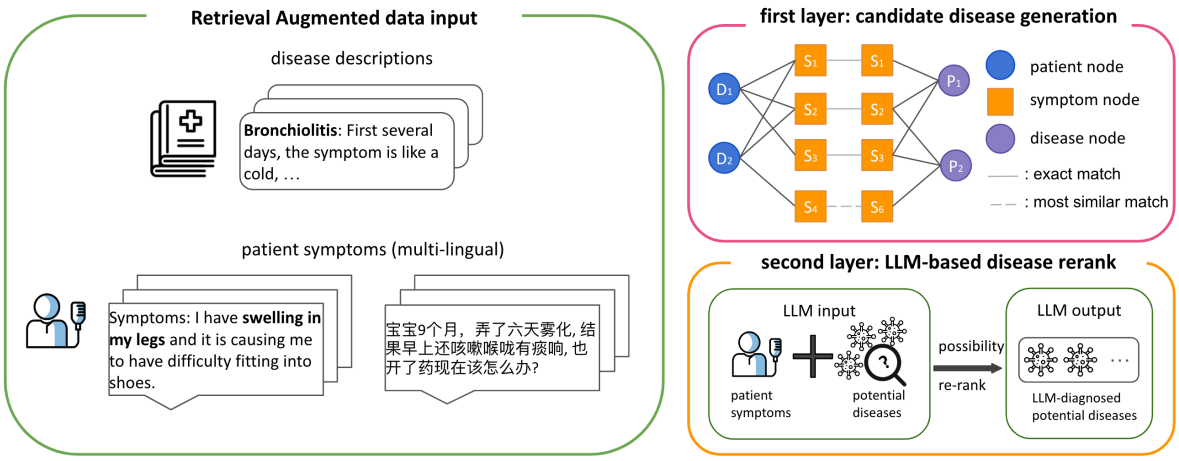

Figure 1: The overall framework of Dx-LLM. The content on the left shows the input data, which consists of the retrieval augmented disease descriptions and the patient symptoms to diagnose. The content at the upper right shows the first layer of graph mining to select the candidate diseases. The content at the bottom right shows the second layer of LLM-based diagnosis generation by re-ranking the candidate diseases.

of Wikipedia. (Ram et al., 2023) proposed an in-context Retrieval-Augmented Language Modeling (RALM) method that leaves the language model (LM) architecture unchanged and prepends grounding documents to the input. (Shi et al., 2023) introduced a retrieval-augmented language modeling framework that treats LM as a black box and prepends retrieved documents to the input for the frozen black-box LM. In our work, we retrieve disease symptoms knowledge from Mayo Clinics to avoid the disease knowledge insufficiency introduced by data scarcity.

3 Method

Dx-LLM consists of three parts: (1) RA-LLMs-based HIN graph generation (Section 3.1), (2) first layer: graph-based candidate disease generation (Section 3.2), (3) second layer: LLM-based disease re-rank (Section 3.3).

3.1 RA-LLMs-based HIN Graph Generation

In the module of RA-LLMs-based HIN Graph Generation, our objective is to construct a disease-symptom HIN graph and a patient-symptom HIN graph. The input data is a set of disease description passages M, and a set of patient symptom description passages N. Each item M_i in M is in the format of $D_i : R_i$, where D_i is the disease name and R_i corresponds to the disease symptom description passage. All D_i forms a disease set D, where each D_i represents a disease node. Each item N_i in N is in the format of $P_i : C_i$, where P_i is the patient id and C_i is the corresponding self-reported symptoms of patient i. All P_i will form a patient set P, where each P_i represents a patient node. In our design, we apply RAG to extract M from Mayo clinics to construct a global disease-symptom graph. Then we prompt LLMs to extract the symptom keywords from R_i and C_i. Each generated symptom subset S_{R_i} will be added to the symptom set S^D, where $S_i \in S^D$ represents a unique symptom node that appears in one or more diseases. S_{C_i} will be added to the symptom set S^P, where $S_i \in S^P$ represents a unique symptom node that appears in one or more patients' symptoms. Given the disease symptom description passage R_i and the patient's self-reported symptoms C_i, the extracted symptoms can be denoted as:

$$\begin{aligned} S_{R_i} &= LLM(R_i) \\ S_{C_i} &= LLM(C_i) \end{aligned} \tag{1}$$

We construct two graphs, namely, the disease-symptom graph G^D and the patient-symptom graph G_P. G^D comprises two sets of nodes, D and S^D, where D represents diseases and S^D represents extracted symptoms for diseases. G^P comprises two sets of nodes, P and S^P, where P represents patients and S^P represents extracted symptoms for patients. S^D and S^P can share common or different nodes. The edges in G^D are represented as E^D. For each $(D_i, S_j) \in E^D$, it indicates disease D_i has the symptom S_j, where $D_i \in D, S_j \in S^D$ Similarly, the edges in G^P are represented as E^P. For each $(P_i, S_j) \in E^P$, it indicates patient P_i shows the symptom S_j, where $P_i \in P, S_j \in S^P$. The

procedure of G^D and G^P generation is showcased in Figure 2.

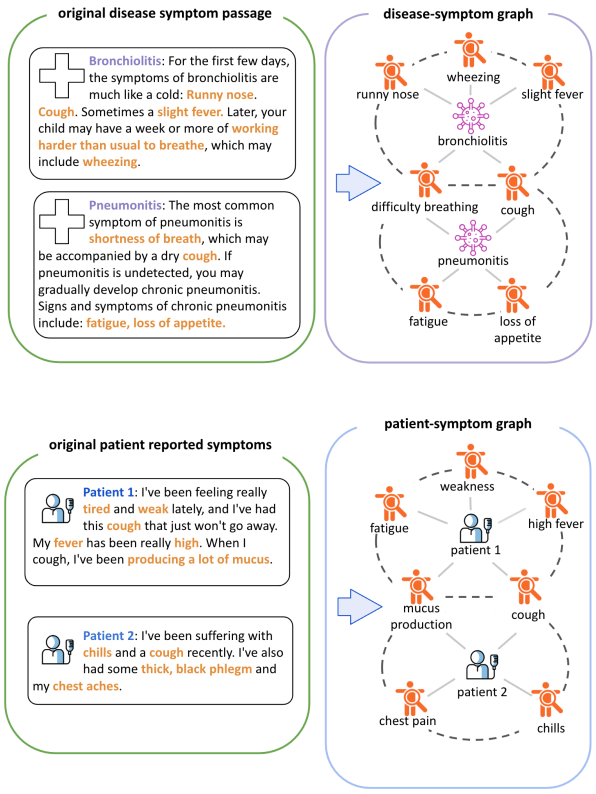

Figure 2: Knowledge graph generation process. The above figure shows an example of disease-symptom knowledge graph generation. The bottom graph shows an example of patient-symptom knowledge graph generation.

3.2 First Layer: Graph-based Candidate Disease Generation

We formalize the first layer candidate disease selection module as a similarity ranking task, where we select diseases with the most similar embedding with the patient as candidate diseases. The candidate disease generation consists of three steps: (1) symptom embedding generation module based on VGAE, (2) disease and patient embedding generation, (3) disease similarity rank. The overall process is shown in Figure 3.

VGAE-based symptom embedding generation. First, we generate the symptom embedding for S^D with a disease-symptom graph, which contains an HIN encoder and an HIN decoder.

HIN encoder. We use $\mathbf{A^D}$ to denote the adjacency matrix of G^D and $\mathbf{X^D} \in \mathbb{R}^{N\times F}$ to denote the feature matrix of G^D, where the initial $\mathbf{X^D}$ is the BERT (Devlin et al., 2018) embedding of the content in each node D_i or S_i. We use L layers GCN as the encoder. In the lth layer, the hidden state of GCN is denoted as $\mathbf{H}^{(l)} = GCN^{(l)}(\mathbf{A^D}, \mathbf{X^D})$, $\mathbf{H}^{(l)} \in \mathbb{R}^{N\times d_l}$, d_l represents the dimension of hidden state in the l_{th} layer. $\mathbf{H}^{(1)} = \mathbf{X}$. The l_{th} GCN layer is formulated as

$$\mathbf{H}^{(l)} = \gamma(\tilde{\mathbf{A}}\mathbf{H}^{(l-1)}\mathbf{W}^{(l-1)}), (2 \le l \le L-1), \tag{2}$$

where $\tilde{A} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ is the symmetrically normalized adjacency matrix of $\mathbf{A}$. $\mathbf{D}$ is the diagonal degree matrix, where $D_{k,k} = \sigma_{i=1}^{N}\mathbf{A}_{k_l}$. $W_l \in \mathbb{R}^{d_l\times d_{l+1}}$ is the weight matrix of the l_{th} layer. γ is the activation function. In our experiment, we use Rectified Linear Unit (ReLU) as the activation function.

Let $\mathbf{Z} \in \mathbb{R}^{N\times T}$ denote the latent matrix, where T is the dimension of node embedding, $\mathbf{z}_i$ is the latent embedding of the ith node. The inference model can be defined as

$$q(\mathbf{Z}|\mathbf{A}, \mathbf{X}) = \prod_{i=1}^{N} q(\mathbf{z_i}|\mathbf{A}, \mathbf{X}), \tag{3}$$

where $q(\mathbf{z_i}|\mathbf{A}, \mathbf{X}) \sim \mathcal{N}(\mathbf{z_i}|\mu_\mathbf{i}, \sigma_\mathbf{i}^\mathbf{2})$

HIN decoder. We utilize linear inner product as the HIN decoder, which is formulated as:

$$p(\mathbf{A}|\mathbf{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{N} p(\mathbf{A}_{i,j}|\mathbf{z}_i, \mathbf{z}_j), \tag{4}$$

where $p(\mathbf{A}_{i,j}|\mathbf{z}_i, \mathbf{z}_j) = \sigma(\mathbf{z}_i^T z_j)$, and $\sigma(\cdot)$ is the logistic sigmoid function.

Disease and patient embedding generation. The disease and patient embedding is based on symptom embedding. We extract the symptoms for each disease and patient. Then we average the embedding of contained symptoms to represent the disease or the patient. Since the symptoms are generated in the natural language format, not all symptoms of patients can have their exact match in the symptom set S. Therefore, we apply a fuzzy match and use the average embedding of symptoms with the BERT embedding cosine similarity larger than a threshold, to represent the unseen symptom of a patient. Suppose z_{S_i} is the embedding of symptom S_i, the embedding z_{D_i} and z_{P_i} for node D_i and P_i can be represented as

$$\begin{cases} S(D_i) = \{S_i \in S^D | (D_i, S_i) \in E^D\} \\ z_{D_i} = avg(\{z_{s_j} | s_j \in S(D_i)\} \\ S(P_i) = \{S_i \in S^P | (P_i, S_i) \in E^P\} \\ F(S(P_i)) = \{S_i \in S^D | S_j \in S^P, sim(S_i, S_j) \ge \beta\} \\ z_{P_i} = avg(\{z_{s_j} | s_j \in F(S(P_i))\} \end{cases} \tag{5}$$

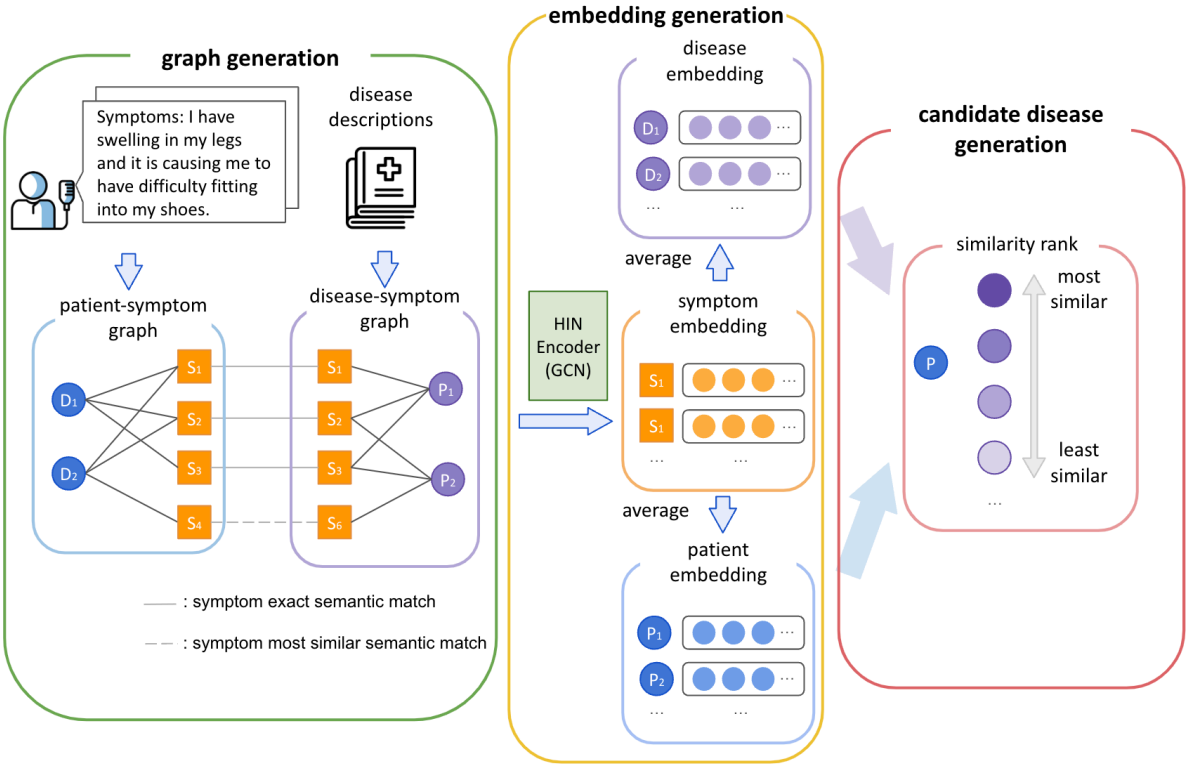

Figure 3: The graph-based candidate disease generation process. The left part is the graph construction part. The middle part is the VGAE-based graph auto-regression process. The right part is a similarity ranking task for candidate disease selection.

where $F(\cdot)$ represents the fuzzy match between symptoms in S^P and S^D, $sim(.)$ represents the cosine similarity between the BERT embedding of S_i and S_j.

Disease similarity rank. After we get the similarity of each $S_i \in S^D$, $D_i \in D$, and $P_i \in P$, we select the candidate disease for each patient P_i based on the cosine similarity rank of diseases, which is defined as

$$sim(P_i, D_j) = \frac{\mathbf{z}_{P_i}^T \cdot \mathbf{z}_{D_j}}{||\mathbf{z}_{P_i}|| \cdot ||\mathbf{z}_{D_j}||}. \quad (6)$$

3.3 Second Layer: LLM-based disease re-rank

With LLMs' superior capability in language understanding and generation, after selecting candidate diseases in the first layer, we feed the patient's information and the candidate disease list to LLMs and prompt LLMs to re-rank the possibility of potential diseases. The prompt of LLM-based disease re-rank is shown in Figure 4. The final output is a re-ranked diagnosis disease list from the highest possibility to the lowest possibility, which can be represented as:

$$D_{final} = LLM(symptom, candidates). \quad (7)$$

where $symptom$ refers to patients' symptoms, $candidates$ refers to the candidate diseases generated after the first layer filtering.

Second layer: disease re-rank prompt

Assume you are a doctor and you need predict the patient's potential disease. I will provide you with the patient's self-described symptoms and the possible diseases.
Patient's symptoms: {patient_symptom}
Candidate diseases: {candidate_disease}
Please do the following task: Re-rank the candidate diseases based on the possibility that the patient might catch.

Figure 4: The prompt for LLM to re-rank potential diagnosed diseases.

4 Experiment

4.1 Dataset

In our experiment, due to the limited budget of the LLMs prompt, we only select a subset of the public dataset for evaluation. For the Chinese dataset, we apply Google Translate API to translate the content

Table 1: Dx-LLM diagnose performance.

Model name	H@1	H@10	H@20	H@50	N@1	N@10	N@20	N@50
Chinese dataset								
BERT	0.1579	0.3158	0.3889	0.4678	0.1579	0.2366	0.2552	0.2706
Roberta	0.1520	0.2602	0.2690	0.2953	0.1520	0.2030	0.2051	0.2102
mpnet	0.1725	0.3450	0.3684	0.4415	0.1725	0.2489	0.2548	0.2703
Dx-LLM (1st layer)	0.0819	0.2924	0.4035	0.6345	0.0819	0.1783	0.2060	0.2515
Dx-LLM (Llama3)	0.2419	0.7064	0.7701	0.7754	0.2419	0.4320	0.4436	0.4452
Dx-LLM (GPT3.5)	0.2322	**0.7357**	0.**7711**	**0.7807**	0.2322	0.4044	0.4191	0.4331
Dx-LLM (GPT4)	**0.2760**	0.7038	0.7522	0.7665	**0.2760**	**0.4715**	**0.4847**	**0.4883**
English dataset								
BERT	0.0588	0.2677	0.3286	0.3972	0.0588	0.1555	0.1709	0.1846
Roberta	0.0783	0.2481	0.2884	0.3493	0.0783	0.1617	0.1720	0.1838
mpnet	0.1110	0.2655	0.3025	0.3885	0.1110	0.1822	0.1915	0.2085
Dx-LLM (1st layer)	0.0566	0.3667	0.5201	0.7008	0.0566	0.1854	0.2239	0.2596
Dx-LLM (Llama3)	0.3003	0.7103	**0.7600**	0.7804	0.3003	0.4575	0.4676	0.4728
Dx-LLM (GPT3.5)	0.2759	0.6848	0.7591	0.7928	0.2759	0.4534	0.4710	0.4794
Dx-LLM (GPT4)	**0.3854**	**0.7141**	0.7461	**0.7842**	**0.3854**	**0.5422**	**0.5496**	**0.5575**

Table 2: Examples of Dx-LLM's generated diagnosis results.

initial symptoms	Dx-LLM (Llama3)	Dx-LLM (GPT3.5)	Dx-LLM (GPT4)
Chinese dataset			
Hello, doctor. My daughter is 28 months old. She had repeated fever for three or four days last week, with the highest temperature reaching 39.2 degrees. There are several herpes on her left hand, but not on her right hand, feet, mouth, or around the anus. **real disease: hand foot and mouth disease**	1. Viral illnesses, such as rotavirus or flu, 2. Staph infections, 3. Kawasaki diseas, 4. Chicken pox, **5. Hand foot and mouth disease**, 6. Measles, 7. Scarlet fever, 8. Rubella, 9. Other infectious diseases like impetigo, measles, etc., 10. The remaining diseases are less likely given the provided symptoms, but I would not rule them out completely without further investigation.	**1. Hand foot and mouth disease**, 2. Measles, 3. Rubella, 4. Scarlet fever, 5. Shingles, 6. Smallpox, 7. Chicken pox, 8. Herpes, 9. Staph infections, 10. Chronic granulomatous disease	**1. Hand foot and mouth disease**, 2. Staph infections, 3. Kawasaki disease, 4. Chicken pox, 5. Scarlet fever, 6. Rheumatic fever, 7. Flu, 8. X linked agammaglobulinemia, 9. Primary immunodeficiency, 10. Chronic granulomatous disease
English dataset			
I've been drinking a lot more water and going to the bathroom a lot more often. My mouth and throat are always dry. I've been really hungry and eating a lot more lately. **real disease: diabetes**	**1. Diabetes**, 2. Diabetic ketoacidosis, 3. Hyperglycemia, 4. Chronic kidney disease, 5. Polydipsia, polyuria, and polyphagia can also be seen in other endocrine disorders	**1. Diabetes**, 2. diabetic ketoacidosis, 3. hyperglycemia, 4. dehydration, 5. hyperparathyroidism, 6. h pylori, 7. pancreatic neuroendocrine tumors, 8. gastroparesis, 9. addisons disease, 10. systemic mastocytosis	**1. Diabetes**, 2. Hyperglycemia, 3. Diabetic ketoacidosis, 4. Dehydration, 5. Chronic kidney disease, 6. Glomerulonephritis, 7. Addison's disease, 8. Hyperparathyroidism, 9. Primary immunodeficiency, 10. Sjogren's syndrome

into English. **Disease information data**

- mayo_clinic_symptoms_and_diseases[1]: A disease symptom knowledge base from mayo clinics, which contains 1058 types of diseases.

Multilingual patient symptom data. For the patient symptom data, we generate a Chinese test dataset from 2 Chinese patient-doctor conversation datasets, and an English dataset from 1 patient-

[1] http://huggingface.co/datasets/celikmus/mayo_clinic_symptoms_and_diseases_v1

Table 3: Statistics of selected multilingual patient dataset.

language	sum # disease types	sum # patient cases	sub-dataset	disease examples	# disease type	# patient case
Chinese	3	300	DX	hand foot and mouth disease, bronchial asthma	2	200
			imcs21	pneumonia	1	100
English	19	500	Symptom2Disease	malaria, psoriasis, jaundice, arthritis, gastroesophageal reflux disease, chicken pox, urinary tract infection, cervical spondylosis, typhoid, impetigo, hypertension, bronchial asthma, peptic ulcerdisease, diabetes, common cold, varicose veins, migraine, dengue, pneumonia	19	500

doctor conversation English dataset. The statistics are shown in Table 3.

- DX (Chinese) (Xu et al., 2019): A dataset collected from dxy.com where users ask doctors for medical diagnosis. We select 200 samples with the disease of "hand foot and mouth disease" or "bronchial asthma" from this dataset.
- imcs21 (Chinese) (Chen et al., 2023): A dataset collected from Muzhi[2], a Chinese online health community that provides professional medical consulting services for patients. We select 100 cases with the disease of "pneumonia" from this dataset.
- Symptom2Disease (English): A dataset containing diseases and natural language symptom descriptions from kaggle[3]. We random sampled 500 cases from 19 diseases in this dataset.

4.2 Baseline Models

We compare the Dx-LLM's performance with three other LLMs, in which we re-rank the diseases based on the cosine similarity of the embedding of the patient's symptoms and the diseases' symptom descriptions.

- BERT (Devlin et al., 2018): A bidirectional encoder representations from Transformers are designed to pre-train deep bidirectional representations from the unlabeled text.
- Roberta (Liu et al., 2019): An improved variant of BERT that enhances performance in natural language understanding tasks by optimizing the pre-training process with more data, longer training times, and larger batch sizes.
- Mpnet (Song et al., 2020): A pre-training model that enhances language understanding by combining masked and permuted language modeling techniques to effectively capture both local and global dependencies in text.

4.3 Metrics

We compare the Dx-LLM's performance with three other LLMs, in which we re-rank the diseases based on the cosine similarity of the embedding of the patient's symptoms and the diseases' symptom descriptions.

- **HIT@K (H@K)**. Whether any of the top-K recommended items were in the test set for a given user.
- **NDCG@K (N@K)**. NDCG is a widely used metric in information retrieval. It is used to calculate a cumulative score of an ordered set of items.

4.4 Setting

In our experiment, we generate 80 candidate diseases among 1058 diseases from the disease knowledge graph after the first layer. We performed the experiments on three LLMs: Llama3, GPT3.5 and GPT4. When the size of the candidate disease set is 100 (in ablation study), we perform the experiment one time for GPT3.5 and GPT4 model due to the budget limit, and perform experiment three times for Llama3. When the size of the candidate disease set is 50 (in ablation study) and 80, we performed the experiment three times and calculated the average.

[2] http://muzhi.baidu.com

[3] http://www.kaggle.com/datasets/niyarrbarman/symptom2disease

Table 4: Dx-LLM diagnose performance with the candidate disease size of 50.

Model name	H@1	H@10	H@20	H@50	N@1	N@10	N@20	N@50
Chinese dataset								
Dx-LLM (Llama3)	0.1893	0.5853	0.6121	**0.6345**	0.1893	0.3556	0.3605	0.3666
Dx-LLM (GPT3.5)	0.2239	**0.6047**	0.6166	**0.6345**	0.2239	0.3801	0.3870	0.3976
Dx-LLM (GPT4)	**0.2477**	0.5874	**0.6290**	**0.6345**	**0.2478**	**0.4128**	**0.4244**	**0.4257**
English dataset								
Dx-LLM (Llama3)	0.2735	0.6299	0.6645	**0.7008**	0.2735	0.4402	0.4473	0.4557
Dx-LLM (GPT3.5)	0.2747	0.6264	0.6719	**0.7008**	0.2747	0.4385	0.4519	0.4597
Dx-LLM (GPT4)	**0.3784**	**0.6550**	**0.6795**	**0.7008**	**0.3784**	**0.5213**	**0.5272**	**0.5317**

Table 5: Dx-LLM diagnose performance with the candidate disease size of 100.

Model name	H@1	H@10	H@20	H@50	N@1	N@10	N@20	N@50
Chinese dataset								
Dx-LLM (Llama3)	0.3116	0.7625	**0.8039**	**0.8157**	**0.3116**	0.4667	0.4763	0.4787
Dx-LLM (GPT3.5)	0.2239	**0.7743**	0.8013	0.8182	0.2239	0.4001	0.4131	0.4248
Dx-LLM (GPT4)	**0.2989**	0.7267	0.7846	0.8031	0.2989	**0.4924**	**0.5066**	**0.5115**
English dataset								
Dx-LLM (Llama3)	0.2971	0.7566	0.7901	0.8193	0.2971	0.4650	0.4724	0.4790
Dx-LLM (GPT3.5)	0.2824	0.7340	**0.8171**	**0.8382**	0.2824	0.4598	0.4791	0.4847
Dx-LLM (GPT4)	**0.4064**	**0.7445**	0.8159	0.8313	**0.4064**	**0.5677**	**0.5779**	**0.5835**

4.5 Major Results

We compare the performance of two-layer Dx-LLM with different SOTA baselines and the 1st layer Dx-LLM. Results are shown in Table 1.

From the result, we can see our proposed Dx-LLM model can outperform other models consistently. Without using patient cases as training data, Dx-LLM can make accurate diagnoses among 1058 different types of diseases. In particular, the hit@10 for both the Chinese dataset and English dataset can achieve around 70% with multiple LLMs, whereas for other baseline models, most results are below 30%. The evaluation result shows a stable performance over patient cases with different languages, which demonstrates Dx-LLM's ability to resolve multilingual diagnosis problems.

We showcase examples of top-10 re-ranked diagnosis results generated by Llama3, GPT3.5 and GPT4 when the candidate set size is 80. Examples are shown in Table 2. As can be seen from the examples, Dx-LLM can make high quality diagnosis for all three LLMs. We also notice that Llama3-based Dx-LLM output have a poor performance in following output instructions.

4.6 Ablation Study

To see the influence of candidate disease size on Dx-LLM's performance, we tested on two other cases when the candidate size is 50 or 100. The result when the candidate size is 50 is shown in Table 4, the result when the candidate size is 50 is shown in Table 5.

As can be seen from the result, when the candidate size is larger, the overall diagnosis performance is better. However, even with a candidate size of 50, the performance can still consistently outperform SOTA baselines, which shows the superiority in Dx-LLM.

5 Conclusion

We proposed Dx-LLM, a two-layer retrieval-augmented multilingual diagnosis system, that does not require abundant patient cases as training data for high performance. Instead, we applied the RA-LLMs technique to generate a disease-symptom graph for representation learning. To effectively utilize LLMs' understanding and generation ability, we proposed a two-layer diagnosis, where we selected the most possible diseases as diagnosis candidates in the first layer, and then prompted LLMs to re-rank the potential diseases. Extensive results showed the superiority of Dx-LLM and demonstrated its ability for multilingual diagnosis.

6 Limitations

In most of our experiments, GPT3.5 and GPT4 are used as the backbone model. Therefore, the result might be biased with different prompts of datasets. Besides, our proposed Dx-LLM does not perform well on distinguishing diseases with similar symptoms. Future research can work on this aspect to better improve the performance.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2023. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39(1):btac817.

Edward Choi, Zhen Xu, Yujia Li, Michael W Dusenberry, Gerardo Flores, Yuan Xue, and Andrew M Dai. 2019. Graph convolutional transformer: Learning the graphical structure of electronic health records. *arXiv preprint arXiv:1906.04716*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Wenqi Fan, Shijie Wang, Jiani Huang, Zhikai Chen, Yu Song, Wenzhuo Tang, Haitao Mao, Hui Liu, Xiaorui Liu, Dawei Yin, et al. 2024. Graph machine learning in the era of large language models (llms). *arXiv preprint arXiv:2404.14928*.

Anahita Hosseini, Ting Chen, Wenjun Wu, Yizhou Sun, and Majid Sarrafzadeh. 2018. Heteromed: Heterogeneous information network for medical diagnosis. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 763–772.

Anahita Hosseini, Tyler Davis, and Majid Sarrafzadeh. 2019. Hierarchical target-attentive diagnosis prediction in heterogeneous information networks. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 949–957. IEEE.

Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. 2016. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):17–37.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Ofir Ben Shoham and Nadav Rappoport. 2023. Cpllm: Clinical prediction with large language models. *arXiv preprint arXiv:2309.11295*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

Huimin Wang, Wai-Chung Kwan, Kam-Fai Wong, and Yefeng Zheng. 2023. Coad: Automatic diagnosis through symptom and disease collaborative generation. *arXiv preprint arXiv:2307.08290*.

Zifeng Wang, Rui Wen, Xi Chen, Shilei Cao, Shao-Lun Huang, Buyue Qian, and Yefeng Zheng. 2021. Online disease diagnosis with inductive heterogeneous graph convolutional networks. In *Proceedings of the Web Conference 2021*, pages 3349–3358.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7346–7353.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.