# GPT-RagAD: Two-layer Retrieval-Augmented Multilingual Diagnosis System

**Xinyi Liu** LIU323@ILLINOIS.EDU
*University of Illinois Urbana-Champaign, 201 N Goodwin Ave, Urbana, IL 61801, USA*

**Dachun Sun** DSUN18@ILLINOIS.EDU
*University of Illinois Urbana-Champaign, 201 N Goodwin Ave, Urbana, IL 61801, USA*

**Yi R. Fung** YRFUNG@UST.HK
*Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong SAR*

**Dilek Hakkani-Tür** DILEK@ILLINOIS.EDU
*University of Illinois Urbana-Champaign, 201 N Goodwin Ave, Urbana, IL 61801, USA*

**Tarek Abdelzaher** ZAHER@ILLINOIS.EDU
*University of Illinois Urbana-Champaign, 201 N Goodwin Ave, Urbana, IL 61801, USA*

## Abstract

We introduce GPT-RagAD, a multilingual, zero-shot automated diagnosis system that achieves high accuracy without relying on real patient data.[1] GPT-RagAD adopts a two-layer Retrieval-Augmented Generation (RAG) architecture: a knowledge graph-based retriever selects disease candidates from 1,058 conditions, and an LLM-based re-ranker applies prompt-based reasoning to refine predictions. Unlike traditional diagnostic models that require supervised training and large clinical datasets, GPT-RagAD is privacy-preserving, scalable, and language-agnostic. Extensive evaluations on three multilingual datasets (Chinese and English) show that GPT-RagAD achieves 40.6% Hit@1 and 56.7% NDCG@10 on the Symptom2Disease benchmark—substantially outperforming embedding-based and direct LLM baselines. Ablation and sensitivity analyses further validate its robustness. GPT-RagAD presents a practical, lightweight solution for clinical triage and pre-diagnosis support.

**Keywords:** Automated Diagnosis, Large Language Models, Retrieval-Augmented Generation, Knowledge Graph, Zero-shot Learning, Medical NLP, Heterogeneous Information Network

**Data and Code Availability** We build our disease–symptom knowledge graph from the public *Mayo Clinic Symptoms and Diseases* dataset (1,058 diseases), and evaluate on three existing benchmarks: DX Xu et al. (2019), IMCS21 Chen et al. (2023), and Symptom2Disease (Kaggle). All datasets are publicly available from their original providers; we do not use or release any private clinical records. An anonymized implementation of GPT-RagAD (graph construction, inference, and evaluation scripts) is included as supplementary material for review. If the paper is accepted, we will de-anonymize and publicly release the code.

**Institutional Review Board (IRB)** Our work uses only previously published, de-identified datasets and does not involve new data collection or access to identifiable health information. Under our institution's policy, this study is not considered human subjects research and therefore did not require IRB review or approval.

## 1. Introduction

Automated diagnosis (AD) is a cornerstone of modern healthcare systems, enabling swift and accurate identification of potential diseases based on patient-reported symptoms. Traditional AD methods often rely on supervised learning models trained on vast amounts of patient data and diagnostic reports Hosseini et al. (2018); Wang et al. (2021); Shoham and Rappoport (2023); Tu et al. (2024). However, several challenges remain with this approach, including concerns about data privacy, the difficulty in collecting comprehensive and diverse patient data, the need

1. Code is available at https://github.com/tracy3057/GPT-RagAD.

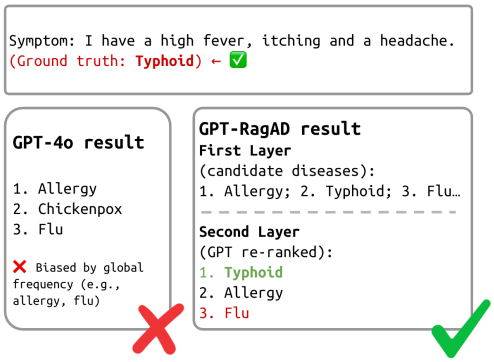


Figure 1: An example of GPT-4o making incorrect diagnosis for a patient, influenced by global priors bias.

for high-quality data labeling, and the challenge of acquiring sufficient patient cases for training. These limitations impede the effectiveness of conventional AD systems.

Recently, large language models (LLMs) have demonstrated impressive generalization abilities across a variety of tasks and have shown considerable promise in the AD domain. LLMs benefit from extensive knowledge learned from large-scale real-world data, enabling them to perform automatic diagnoses with remarkable accuracy. The use of LLMs not only mitigates the challenges of data collection and privacy but also facilitates more flexible diagnosis systems. However, general-purpose LLMs are not explicitly trained for medical diagnosis tasks. As a result, they may suffer from biases introduced by global priors, which can lead to suboptimal or incorrect diagnoses. As illustrated in Figure 1, a general-purpose LLM such as GPT-4o can be influenced by global priors, leading to incorrect diagnoses.

One potential solution is fine-tuning domain-specific LLMs for AD tasks Brown et al. (2020); Fan et al. (2024); Zhao et al. (2023). However, this approach is not cost-effective or feasible due to concerns over patient data privacy and the challenges related to the quality of available data. To address these concerns, Retrieval-Augmented Generation (RAG) methods have gained traction Brown et al. (2020); Fan et al. (2024); Zhao et al. (2023). RAG facilitates the dynamic retrieval of external knowledge sources, thereby introducing personalized information into the diagnostic process. This method helps mitigate the global prior bias. However, directly applying RAG to AD tasks presents two significant challenges. First, relying on large-scale external medical resources, such as PubMed, UMLS, and clinical guidelines, incurs high computational costs, making these sources unsuitable for real-time diagnosis. This reliance on massive datasets complicates the efficient use of essential data for medical diagnosis. Second, using patient case retrieval for augmentation raises privacy concerns, as it may inadvertently expose sensitive health data from previous patients.

To address these challenges, we propose GPT-RagAD, a novel and effective two-layer framework for knowledge-graph-based AD. Unlike traditional methods that depend on large and computationally expensive medical datasets, GPT-RagAD generates disease knowledge graphs from medical encyclopedias, which focus on describing the core features of diseases. This allows for the efficient use of essential, high-quality information without the computational burden of vast medical datasets. By combining these medical knowledge graphs with the powerful language understanding capabilities of LLMs, GPT-RagAD predicts diseases from a more comprehensive disease space while alleviating biases from global priors. In this way, GPT-RagAD provides a more accurate, privacy-preserving, and scalable approach to automated diagnosis.

Figure 2 illustrates the overall architecture of GPT-RagAD, which integrates knowledge graph generation in the first layer and leverages the strengths of LLMs in the second layer to refine the diagnostic process. This two-layer framework not only addresses computational cost and privacy concerns but also enhances the accuracy and scalability of AD systems.

We introduce GPT-RagAD, a multilingual, two-layer diagnosis framework that combines symbolic knowledge retrieval with prompt-based LLM reasoning. The first layer performs coarse-grained disease retrieval using a disease-symptom knowledge graph, while the second layer refines predictions through a retrieval-augmented generation (RAG) strategy.

To efficiently represent complex and uncertain symptom-disease associations, we build a Heterogeneous Information Network (HIN) and apply a Variational Graph Autoencoder (VGAE)Kipf and Welling (2016) that produce deterministic representations, VGAE models uncertainty explicitly, helping reduce overconfidence in noisy matches. Compared to alternatives like GAT Veličković et al. (2017) and R-GCN Schlichtkrull et al. (2018), VGAE offers a better balance of robustness, scalability, and expressiveness—especially when applied to loosely structured

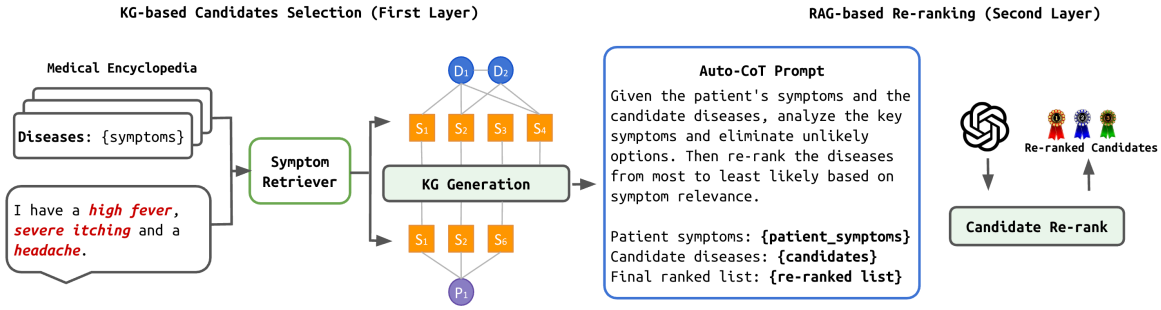


Figure 2: Overview of GPT-RagAD: Knowledge Graph-Based Disease Candidate Selection and RAG-based Re-ranking.

encyclopedia-derived graphs rather than curated clinical datasets.

Finally, the learned HIN embeddings guide the RAG module to focus on high-quality candidates and consider the relative importance of different symptoms, enhancing the precision and interpretability of downstream predictions.

Our contributions are as follows:

**Contributions.**

- A RAG-enhanced AD framework that builds a disease–symptom HIN from curated encyclopedic data, encoding disease–symptom/disease–disease/symptom–symptom relations.
- A two-stage system: fast HIN-based retrieval plus LLM re-ranking, achieving scalability and high accuracy with low inference overhead.
- Multilingual diagnosis over 1,058 diseases with >70% Hit@10 on English and Chinese datasets, outperforming state-of-the-art baselines.
- Unified symbolic retrieval and LLM reasoning that avoids annotation, preserves privacy, and supports practical deployment.

## 2. Related Work

### 2.1. Graph- and LLM-based Diagnosis

Graph-based AD encodes EHRs into heterogeneous information networks (HINs) to model disease–symptom–entity relations Hosseini et al. (2018); Wang et al. (2021). These methods work well with structured data but rely on large, clean EHRs, limiting cross-domain and multilingual generalization. LLM-based approaches Shoham and Rappoport (2023); Wang et al. (2023); Tu et al. (2024) predict diseases or synthesize symptoms directly from text; they are flexible and annotation-free but exhibit global prior bias and lack explicit reasoning over structured medical knowledge.

### 2.2. RAG for Clinical Reasoning and Our Positioning

RAG grounds LLMs with retrieved context Lewis et al. (2020); Ram et al. (2023); Shi et al. (2023), yet most pipelines retrieve free text, risking hallucination, higher latency, and privacy concerns in clinical settings. **GPT-RagAD** retrieves from a disease–symptom knowledge graph (public encyclopedias), models uncertainty via VGAE, and applies LLM re-ranking with optimized prompting—improving interpretability and scalability while preserving privacy and enabling cross-lingual generalization. See Appendix D for an extended survey.

## 3. GPT-RagAD Framework

The overview of GPT-RagAD is in Figure 2. We firstly develop the first layer: KG-based Candidates Selection, shown in Figure 3. Then, based on the selected candidate diseases, we establish a RAG-based candidate Re-ranking, introduced in Section 3.2. The detailed procedures are shown in Algorithm 1.

### 3.1. First Layer: Graph-based Candidate Disease Generation

In the first layer of GPT-RagAD, we construct an enhanced multi-relational disease-symptom knowledge graph and use a graph-based similarity framework to generate candidate diseases. The process in-

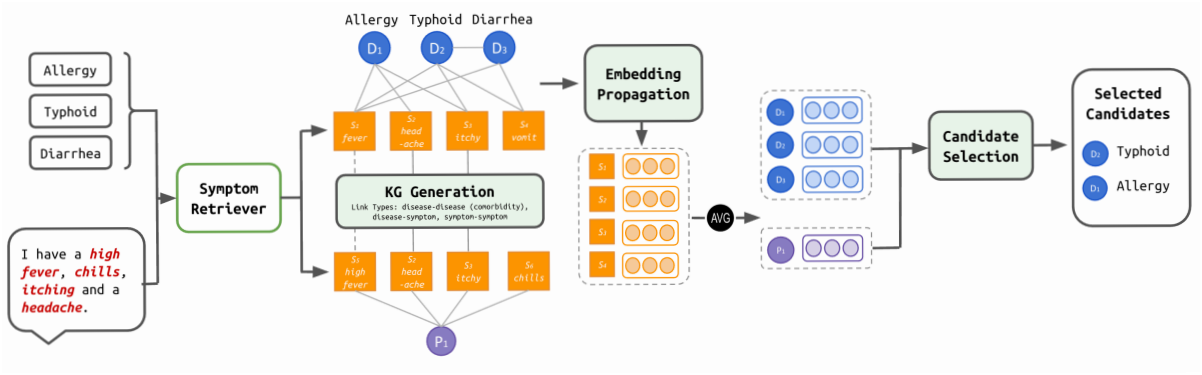


Figure 3: Detailed procedure of KG-based Candidates Selection (First Layer).

**Algorithm 1:** GPT-RagAD Disease Diagnosis Process

**Input:**
  $\mathcal{D}$: Disease descriptions from medical encyclopedia
  $\mathcal{P}$: Patient symptom inputs (free text)
**Output:** Final diagnosis $\hat{d} \in \mathcal{D}$ for each patient
**Step 1: Knowledge Graph Construction**
  Extract symptom sets $S(D_i)$ from each $D_i \in \mathcal{D}$ via LLM
  Build multi-relational graph $G = (V, E)$ with:
    $E^{\text{ds}}$: disease–symptom edges from encyclopedic links
    $E^{\text{sym}}$: symptom–symptom similarity edges (BERT similarity $\geq \tau$)
    $E^{\text{dd}}$: disease–disease comorbidity or Jaccard similarity edges
**Step 2: Patient Symptom Normalization**
  For each $P_i \in \mathcal{P}$, extract raw symptoms via LLM
  Normalize symptoms to KG vocabulary using cosine similarity $\geq \beta$
  Obtain patient symptom set $\hat{S}_i^P$
**Step 3: Graph-based Embedding Learning**
  Apply relation-aware VGAE on $G$ to obtain node embeddings $\mathbf{z}_v$
  Aggregate:
    $\mathbf{z}_{P_i} = \text{mean}(\mathbf{z}_{s_j})$ for $s_j \in \hat{S}_i^P$
    $\mathbf{z}_{D_j} = \text{mean}(\mathbf{z}_{s_k})$ for $s_k \in S(D_j)$
**Step 4: Candidate Disease Retrieval**
  Compute similarity scores:
    $\text{sim}(P_i, D_j) = \cos(\mathbf{z}_{P_i}, \mathbf{z}_{D_j})$
  Select top-$K$ diseases: $\mathcal{C}_i = \text{TopK}_{D_j}(\text{sim}(P_i, D_j))$
**Step 5: LLM-based Re-ranking**
  Construct prompt $\mathcal{T}_i = \texttt{PromptTemplate}(P_i, \mathcal{C}_i)$
  Use LLM to re-rank $\mathcal{C}_i$: $\mathcal{D}_i^{\text{final}} = \text{LLM}(\mathcal{T}_i)$
  Output top prediction: $\hat{d}_i = \mathcal{D}_i^{\text{final}}[0]$

volves three key steps: (1) heterogeneous graph construction with multiple relation types, (2) relation-aware embedding propagation using VGAE, and (3) candidate disease selection via patient-disease similarity. The overall workflow is illustrated in Figure 3.

**Multi-Relational Knowledge Graph Construction** We begin by extracting symptoms from both disease encyclopedia entries and patient inputs. Let $\mathcal{D}$ denote the set of diseases, and $\mathcal{P}$ denote patient cases. For each disease $D_i \in \mathcal{D}$, we use an LLM-based extractor to obtain its associated symptom set $S(D_i)$ from its encyclopedia description $R_i$:

$$S(D_i) = \text{LLM}(R_i). \tag{1}$$

Beyond the canonical disease–symptom edges, we introduce two additional relation types to enrich the graph semantics:

- **Symptom–Symptom Edges:** If two symptoms co-occur frequently across diseases or exhibit high textual embedding similarity, we connect them with a semantic edge. Formally, for symptoms $s_a, s_b \in S^D$,

$$(s_a, s_b) \in E^{\text{sym}} \quad \text{if} \quad \text{sim}(s_a, s_b) \geq \tau,$$

  where similarity is computed using BERT embeddings and $\tau$ is a fixed threshold.
- **Disease–Disease Edges:** If two diseases share a high number of overlapping symptoms (i.e., $\text{Jaccard}(S(D_i), S(D_j)) \geq \gamma$), or are commonly co-mentioned in clinical literature (co-morbidity), we create a disease–disease edge.

Combining all relation types, we obtain a heterogeneous graph $G = (V, E)$ with node set $V = \mathcal{D} \cup S^D$ and edge set $E = E^{\text{ds}} \cup E^{\text{sym}} \cup E^{\text{dd}}$ for disease–symptom, symptom–symptom, and disease–disease relations respectively.

Each patient input $C_i \in \mathcal{P}$ is similarly processed using an LLM extractor, followed by BERT-based symptom alignment. The normalized symptom set for patient $i$ is:

$$\hat{S}_i^P = \left\{ S_j \in S^D \mid \max_{S_k \in \text{LLM}(C_i)} \text{sim}(S_j, S_k) \geq \beta \right\}. \tag{2}$$

**Multi-Relational Embedding Propagation** To capture the semantics of multi-type interactions, we apply a relation-aware Variational Graph Autoencoder (VGAE). Let $\mathbf{X} \in \mathbb{R}^{|V| \times F}$ be the initial BERT-based node feature matrix, and $\mathbf{A}^{(r)}$ the adjacency matrix for relation $r \in \{\text{ds}, \text{sym}, \text{dd}\}$. We aggregate information across $R$ relation types using a shared GCN encoder with relation-specific propagation:

$$\mathbf{H}^{(l)} = \gamma \left( \sum_{r=1}^{R} \tilde{\mathbf{A}}^{(r)} \mathbf{H}^{(l-1)} \mathbf{W}^{(r)} \right), \tag{3}$$

where $\tilde{\mathbf{A}}^{(r)}$ is the normalized adjacency matrix of relation $r$, $\mathbf{W}^{(r)}$ is the trainable weight for relation $r$, and $\gamma$ is a non-linear activation (e.g., ReLU).

Each node's latent representation is modeled as a Gaussian distribution:

$$q(\mathbf{z}_i | \mathbf{X}, \{\mathbf{A}^{(r)}\}) \sim \mathcal{N}(\boldsymbol{\mu}_i, \text{diag}(\boldsymbol{\sigma}_i^2)), \tag{4}$$

and trained via a standard reconstruction loss using inner product decoding.

**Patient and Candidate Representation** After embedding propagation, we compute patient and disease embeddings via mean aggregation over their respective symptom representations:

$$\begin{aligned} \mathbf{z}_{D_i} &= \frac{1}{|S(D_i)|} \sum_{s_j \in S(D_i)} \mathbf{z}_{s_j}, \\ \mathbf{z}_{P_i} &= \frac{1}{|\hat{S}_i^P|} \sum_{s_j \in \hat{S}_i^P} \mathbf{z}_{s_j}. \end{aligned} \tag{5}$$

We then compute cosine similarity between the patient embedding and each disease embedding:

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

The top-$K$ most similar diseases form the candidate set:

$$\mathcal{C}_i = \text{Top-}K \left( \{\text{sim}(P_i, D_j) \mid D_j \in \mathcal{D}\} \right). \tag{7}$$

These candidates are passed to the second-layer LLM for semantic re-ranking (Section 3.2).

### 3.2. Second Layer: LLM-based Disease Re-ranking

After retrieving a top-$K$ candidate disease list from the first-layer knowledge graph, we leverage the language understanding and reasoning capabilities of Large Language Models (LLMs) to re-rank these candidates based on patient-specific information. This step introduces contextual and semantic reasoning into the diagnostic process, allowing the model to consider nuanced symptom-disease relationships beyond structural similarity.

**Prompt Construction and Optimization** To enable effective re-ranking, we design structured prompts that instruct the LLM to identify the most likely disease based on the patient's symptoms and candidate list. A base version of the prompt includes the following components:

- **Patient Symptoms:** a natural language description of the patient's reported symptoms.
- **Candidate Diseases:** the list of top-$K$ diseases retrieved from the knowledge graph.

We explore three prompt strategies for re-ranking candidate diseases:

- **Vanilla Prompting**: A basic template that directly asks the model to re-rank diseases based on the given symptoms.
- **Chain-of-Thought (CoT) Prompting**: Enhances reasoning by instructing the model to analyze symptoms step-by-step and eliminate unlikely candidates.
- **Auto-CoT Prompting**: Automatically generates and ranks multiple CoT-style prompts using GPT-4, selecting the best-performing variant on a held-out validation set Zhang et al. (2022).

These strategies are designed to improve the alignment between the LLM's reasoning process and the subtle symptom variations across languages and patient descriptions.

**Candidate Re-ranking** Given a patient $P_i$ with symptom description $\mathcal{S}_i$, and a candidate disease set $\mathcal{C}_i$ from the first layer, the re-ranking model selects a prompt $\mathcal{T}_i^*$ from the optimized prompt set and uses it to re-order the candidates. Formally:

$$\mathcal{D}_i^{\text{final}} = \text{LLM}(\mathcal{T}_i^*(\mathcal{S}_i, \mathcal{C}_i)), \tag{8}$$

where $\mathcal{D}_i^{\text{final}}$ is the re-ranked list of diseases output by the LLM. The final prediction $\hat{d}_i$ is selected from the top of this list. The right panel of Figure 2 provides an overview of this re-ranking process.

In our experiments (Section 4), we demonstrate that optimized prompts lead to consistent gains in diagnostic accuracy, particularly in multilingual and high-ambiguity cases. This confirms the importance of task-specific prompt design in complex reasoning tasks such as medical diagnosis.

## 4. Experiment

*Further dataset statistics, training details, metrics, and full ablations are in Appendix A.*

### 4.1. Datasets

**Dataset Statistics.** Table 2 summarizes sizes, languages, and diagnosis diversity across splits.

**Disease Knowledge Base** We use a structured medical encyclopedia to construct the disease-symptom knowledge graph used in the first layer of GPT-RagAD. we adopt the following dataset:

- **Mayo Clinic Symptoms and Diseases**[2]: A curated dataset consisting of textual symptom descriptions for 1058 diseases collected from Mayo Clinic's public medical knowledge base.

**Multilingual Patient Symptom Data** To evaluate GPT-RagAD's multilingual diagnosis capabilities, we construct three test sets from real-world patient-doctor dialogues, including two Chinese datasets and one English dataset. Descriptive statistics are summarized in Table 2.

- **DX (Chinese)** Xu et al. (2019): A medical dialogue dataset from dxy.com. We select 104 cases among 5 diseases.
- **IMCS21 (Chinese)** Chen et al. (2023): A Chinese clinical QA dataset from Muzhi, a Baidu health platform. We select 500 pneumonia cases among 10 diseases.
- **Symptom2Disease (English)**: An English dataset containing free-text symptom descriptions from 24 diseases, curated from Kaggle. We randomly select 500 samples.

### 4.2. Baseline Models

Recent domain-specific LLMs, such as Med-PaLM 2 Singhal et al. (2025), PMC-LLaMA Wu et al. (2024), and ClinicalCamel Toma et al. (2023), have shown strong performance in medical reasoning tasks by leveraging large-scale supervised training on curated clinical corpora. However, such models typically require access to sensitive patient records, expensive medical annotations, or proprietary biomedical datasets—conditions that limit their scalability and practical use in privacy-sensitive or low-resource environments.

2. http://huggingface.co/datasets/celikmus/mayo_clinic_symptoms_and_diseases_v1

In contrast, GPT-RagAD is designed as a lightweight, training-free framework that builds a symbolic disease-symptom knowledge graph from publicly available medical encyclopedias (e.g., Mayo Clinic) and augments reasoning through prompt-based use of general-purpose LLMs. This hybrid design eliminates the need for fine-tuning on medical data, reduces hallucination risks through constrained candidate spaces, and generalizes across languages and LLM backbones. To assess its effectiveness, we compare GPT-RagAD against two categories of baseline systems:

- **Embedding Similarity Models**: These methods compute cosine similarity between patient symptom descriptions and disease texts using pre-trained sentence encoders. We include BERT Devlin et al. (2018), RoBERTa Liu et al. (2019), and MPNet Song et al. (2020). They serve as retrieval-only baselines without reasoning or external knowledge integration.
- **Direct Generation (LLM-only)**: We evaluate GPT-4o in a direct zero-shot diagnosis setting using both vanilla and CoT-style prompts, without knowledge graph grounding or candidate filtering. This baseline reflects the upper bound of general-purpose LLM reasoning in the absence of structured symbolic support.

This comparison isolates the added value of graph-based candidate narrowing and prompt-based re-ranking in our two-layer architecture, showing how symbolic retrieval and prompt optimization complement LLM generation to improve both accuracy and reliability.

### 4.3. Metrics

We adopt two widely used ranking metrics:

- **Hit@K (H@K)**: Measures whether the ground truth disease is among the top-K predicted results.
- **NDCG@K (N@K)**: Evaluates the ranking quality of the disease list by rewarding correct predictions ranked higher.

### 4.4. Implementation Details

In the first layer, we generate 100 candidate diseases for each patient from the 1058-disease knowledge base using our graph-based retriever. In the second layer, we use three LLMs—LLaMA3, GPT-3.5, and GPT-4o—for re-ranking. Due to cost constraints, GPT-3.5 and GPT-4o are evaluated once on each dataset,

Table 1: Performance comparison across baseline models and our proposed GPT-RagAD variants. Best results are in **bold**, second-best are underlined.

| Dataset | Metric | BERT | RoBERTa | MPNet | GPT-RagAD (GPT-4o) | | GPT-RagAD (Auto-CoT) | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Vanilla | CoT | LLaMA3 | GPT-3.5 | GPT-4o |
| **DX** (Chinese) | H@1 | 0.1541 | 0.1497 | 0.1754 | 0.2663 | 0.2779 | **0.3074** | 0.2391 | 0.3057 |
| | H@10 | 0.3054 | 0.2711 | 0.3502 | 0.7016 | 0.7141 | **0.7823** | 0.7581 | 0.7357 |
| | H@20 | 0.4007 | 0.2741 | 0.3690 | 0.7624 | 0.7791 | **0.8075** | 0.7973 | 0.7901 |
| | H@50 | 0.4537 | 0.2976 | 0.4321 | 0.7845 | 0.7812 | **0.8205** | 0.8132 | 0.8093 |
| | N@1 | 0.1541 | 0.1497 | 0.1754 | 0.2663 | 0.2779 | **0.3074** | 0.2391 | 0.3057 |
| | N@10 | 0.2169 | 0.2058 | 0.2246 | 0.4827 | 0.4924 | 0.4028 | 0.4125 | **0.5014** |
| | N@20 | 0.2486 | 0.2114 | 0.2642 | 0.4865 | 0.4973 | 0.4793 | 0.4702 | **0.5104** |
| | N@50 | 0.2674 | 0.2198 | 0.2819 | 0.4911 | 0.5056 | 0.4801 | 0.4289 | **0.5120** |
| **IMCS21** (Chinese) | H@1 | 0.1579 | 0.1520 | 0.1725 | 0.2754 | 0.2908 | **0.3116** | 0.2391 | 0.2989 |
| | H@10 | 0.3158 | 0.2602 | 0.3450 | 0.7025 | 0.7189 | 0.7625 | **0.7743** | 0.7267 |
| | H@20 | 0.3889 | 0.2690 | 0.3684 | 0.7592 | 0.7726 | **0.8039** | 0.8013 | 0.7846 |
| | H@50 | 0.4678 | 0.2953 | 0.4415 | 0.7815 | 0.7663 | 0.8157 | **0.8182** | 0.8031 |
| | N@1 | 0.1579 | 0.1520 | 0.1725 | 0.2754 | 0.2908 | **0.3116** | 0.2391 | 0.2989 |
| | N@10 | 0.2366 | 0.2030 | 0.2489 | 0.4612 | 0.4827 | 0.4667 | 0.4001 | **0.4924** |
| | N@20 | 0.2552 | 0.2051 | 0.2548 | 0.4705 | 0.4823 | 0.4763 | 0.4131 | **0.5066** |
| | N@50 | 0.2706 | 0.2102 | 0.2703 | 0.4890 | 0.4967 | 0.4787 | 0.4248 | **0.5115** |
| **Symptom2Disease** (English) | H@1 | 0.0588 | 0.0783 | 0.1110 | 0.3840 | 0.3991 | 0.2971 | 0.2824 | **0.4064** |
| | H@10 | 0.2677 | 0.2481 | 0.2655 | 0.7291 | 0.7368 | **0.7566** | 0.7340 | 0.7445 |
| | H@20 | 0.3286 | 0.2884 | 0.3025 | 0.8024 | 0.8107 | 0.7901 | **0.8171** | 0.8159 |
| | H@50 | 0.3972 | 0.3493 | 0.3885 | 0.8116 | 0.8248 | 0.8193 | **0.8382** | 0.8313 |
| | N@1 | 0.0588 | 0.0783 | 0.1110 | 0.3840 | 0.3991 | 0.2971 | 0.2824 | **0.4064** |
| | N@10 | 0.1555 | 0.1617 | 0.1822 | 0.5412 | 0.5615 | 0.4650 | 0.4598 | **0.5677** |
| | N@20 | 0.1709 | 0.1720 | 0.1915 | 0.5496 | 0.5637 | 0.4724 | 0.4791 | **0.5779** |
| | N@50 | 0.1846 | 0.1838 | 0.2085 | 0.5684 | 0.5727 | 0.4790 | 0.4847 | **0.5835** |

Table 2: Dataset Statistics.

| Dataset | #Diseases | #Samples | Language |
|---|---|---|---|
| **DX** | 5 | 104 | Chinese |
| **IMCS21** | 10 | 500 | Chinese |
| **Symptom2Disease** | 24 | 500 | English |

while LLaMA3 is evaluated with three independent runs and averaged.

### 4.5. Main Results

Table 1 presents a comprehensive comparison of GPT-RagAD variants against embedding-based retrieval baselines (BERT, RoBERTa, MPNet) and GPT-4o direct generation (Vanilla/CoT prompting), across three multilingual datasets.

**Substantial Performance Gains.** GPT-RagAD consistently and significantly outperforms all baselines across datasets and evaluation metrics. On the Chinese **DX** dataset, our best variant (Auto-CoT + LLaMA3) achieves a Hit@10 of **78.2%**, more than doubling the best baseline (MPNet: 35.0%) and substantially improving over GPT-4o with vanilla prompt (70.2%). Similar patterns are observed in English: on the **Symptom2Disease** dataset, our method improves Hit@1 from 11.1% (MPNet) to **40.6%** (Auto-CoT + GPT-4o), a nearly **3.7×** gain. This validates the core design of GPT-RagAD—combining symbolic retrieval with prompt-based neural re-ranking—as a powerful framework for multilingual, zero-shot diagnosis.

**Prompting Strategy Matters.** Across all datasets, prompting strategies significantly impact final performance. CoT-style prompting consistently outperforms vanilla instructions (e.g., IMCS21, NDCG@10 improves from 46.1% to 48.3% with GPT-4o), and our **Auto-CoT prompt generation** further improves results. Notably, Auto-CoT + GPT-4o achieves the best NDCG@10 across datasets (DX: **50.1%**, IMCS21: **49.2%**, Symptom2Disease: **56.8%**). This confirms the importance of context-aware reasoning and demonstrates that prompt optimization is a crucial component of LLM-based diagnosis.

**LLM Portability and Scalability.** While GPT-4o yields the highest overall accuracy, GPT-RagAD remains effective across diverse LLMs. The LLaMA3 variant achieves best-in-class performance on several metrics (e.g., Hit@10 = **78.2%** on DX, Hit@50 = **81.8%** on IMCS21), showing strong competitiveness despite being open-source. GPT-3.5 provides a lightweight alternative with reasonable performance

Table 3: Ablation Study: Impact of Second-Layer Re-ranking. Performance across three datasets comparing candidate retrieval (1st layer) and full re-ranking (2nd layer). Best results are in **bold**.

| Dataset | Layer | Hit@K | | | | NDCG@K | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | H@1 | H@10 | H@20 | H@50 | N@1 | N@10 | N@20 | N@50 |
| **DX** (Chinese) | 1st Layer | 0.0919 | 0.3024 | 0.4135 | 0.6445 | 0.0919 | 0.1883 | 0.2160 | 0.2615 |
| | 2nd Layer | **0.3057** | **0.7357** | **0.7901** | **0.8093** | **0.3057** | **0.5014** | **0.5104** | **0.5120** |
| **IMCS21** (Chinese) | 1st Layer | 0.0893 | 0.2831 | 0.4039 | 0.6277 | 0.0893 | 0.1679 | 0.2057 | 0.2602 |
| | 2nd Layer | **0.2989** | **0.7267** | **0.7846** | **0.8031** | **0.2989** | **0.4924** | **0.5066** | **0.5115** |
| **Symptom2Disease** (English) | 1st Layer | 0.0667 | 0.3767 | 0.5301 | 0.7842 | 0.0667 | 0.1954 | 0.2339 | 0.2696 |
| | 2nd Layer | **0.4064** | **0.7445** | **0.8159** | **0.8313** | **0.4064** | **0.5677** | **0.5779** | **0.5835** |

Table 4: Bucket-based Accuracy by Disease Type (3 Lowest Hit@1)

| Disease Type | #Cases | Hit@1 | Hit@10 | NDCG@10 |
|---|---|---|---|---|
| Cough | 34 | 0.3219 | 0.6757 | 0.5039 |
| Diarrhea | 33 | 0.3342 | 0.6941 | 0.5227 |
| Indigestion | 28 | 0.3692 | 0.7046 | 0.5465 |

(e.g., NDCG@10 = 41.2% on DX), making the system scalable across computational budgets. This demonstrates that our framework is not tied to any specific proprietary model and can generalize across language model backbones.

**Cross-lingual Robustness.** GPT-RagAD performs robustly across both Chinese and English datasets without any language-specific tuning. On the Chinese **IMCS21** dataset, Auto-CoT + GPT-4o achieves 81.5% Hit@50, and on the English **Symptom2Disease** dataset, it reaches 83.1%, demonstrating high recall across languages. The model's ability to generalize from structured KG retrieval to LLM reasoning makes it well-suited for multilingual deployment in low-resource settings.

**Summary.** Together, these results validate the efficacy of our two-layer design and highlight several key takeaways: (1) symbolic KG-based retrieval drastically improves LLM inference, (2) prompt engineering—especially via Auto-CoT—is critical for accurate re-ranking, and (3) GPT-RagAD generalizes well across models, languages, and data domains. Our system achieves state-of-the-art results across all benchmarks without requiring access to private patient data or task-specific fine-tuning.

### 4.6. Ablation Summary

We ablate (i) knowledge graph retrieval depth, (ii) reranker architecture, and (iii) reasoning style (Vanilla vs. CoT). Results indicate the two-layer design contributes the largest gains, and CoT yields consistent improvements at higher K. See Appendix A for protocol details.

## 5. Error Analysis

To better understand failure modes in GPT-RagAD, we conduct qualitative and quantitative error analysis on the DX (Chinese) dataset. Table 10 summarizes representative failure cases, and Table 9 quantifies the proportion of errors attributed to retrieval-stage (top-100 candidate miss) and re-ranking-stage (top-10 miss) failures. We categorize four primary error sources:

**(1) Semantic Overlap.** Diseases with highly overlapping symptoms (e.g., *pneumonia* vs. *asthma*) often confuse both stages. Although symptom nodes are correctly matched in the knowledge graph, the LLM struggles to disambiguate without more precise contextual cues, leading to misclassification.

**(2) Global Prior Bias.** Even with candidate inclusion from the knowledge graph, LLM re-ranking remains biased toward globally frequent diseases. For instance, a case with canonical *herpes* symptoms was incorrectly ranked below *influenza*, reflecting residual training bias. Notably, 19.4% of errors occur despite the correct label being in the top-100 candidates—indicating re-ranking failures.

**(3) Ambiguous Input.** Patient descriptions with vague or nonspecific phrasing (e.g., "belly discomfort") often lead to incorrect semantic interpretation by the LLM. These input-level ambiguities account for a notable portion of re-ranking errors, where lack of discriminative features undermines accurate reasoning.

**(4) Retrieval Miss.** In 12.9% of failures, the ground-truth disease was absent from the top-100

candidates, limiting re-ranking effectiveness regardless of LLM capability. This is often due to symptom normalization mismatch or incomplete extraction during KG construction—highlighting the need for improved recall in the first layer.

bucket-level analysis (Table 4) reveals lower accuracy for generalized symptom clusters like **cough** (Hit@1: 32.2%), **diarrhea** (33.4%), and **indigestion** (36.9%). These vague disease categories introduce high lexical and semantic variability, complicating both graph matching and LLM interpretation.

**Summary.** GPT-RagAD's two-stage design mitigates some limitations of end-to-end generation, but failures persist due to residual bias, retrieval incompleteness, and language ambiguity. Future directions include contrastive training to reduce global prior reliance, symptom paraphrase expansion for KG recall, and more robust prompt formatting to reduce misalignment.

## 6. Conclusion

We presented **GPT-RagAD**, a two-layer retrieval-augmented diagnosis system: a graph-based generator maximizes recall and an LLM re-ranker improves precision, with no task-specific training. Across DX, IMCS21, and Symptom2Disease, it outperforms neural and prompting baselines on HIT@K/NDCG@K; ablations show the two-layer design yields most gains, and CoT helps at larger $K$ (e.g., $K$=100 is a good trade-off). Remaining issues are overlapping symptoms, sparse input, and occasional retrieval misses; next steps include stronger recall under noise, safer reasoning traces, and broader multilingual coverage. This system is for research, not clinical use.

## 7. Acknowledgement

This research was supported in part by NSF CNS-20-38817, DARPA HR0011-24-3-0325 (BRIES), and The Boeing Company. It also received partial support from ACE, one of the seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. The views and conclusions expressed are those of the authors and do not necessarily reflect the official policies of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notice.

## References

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

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

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

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

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

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

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

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

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and

Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*, pages 593–607. Springer, 2018.

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

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

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8, 2025.

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

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *CoRR*, 2023.

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

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

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

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

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, 2024.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 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, 2019.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

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

# Appendix A. Additional Experimental Details

## A.1. Datasets

**Disease Knowledge Base** We utilize a structured medical encyclopedia to construct the disease-symptom knowledge graph used in the first layer of GPT-RagAD. Specifically, we adopt the following dataset:

- **Mayo Clinic Symptoms and Diseases**[3]: A curated dataset consisting of textual symptom descriptions for 1058 diseases collected from Mayo Clinic's public medical knowledge base.

**Multilingual Patient Symptom Data** To evaluate GPT-RagAD's multilingual diagnosis capabilities, we construct three test sets from real-world patient-doctor dialogues, including two Chinese datasets and one English dataset. Descriptive statistics are summarized in Table 2.

- **DX (Chinese)** Xu et al. (2019): A medical dialogue dataset from dxy.com. We select 104 cases among 5 diseases.
- **IMCS21 (Chinese)** Chen et al. (2023): A Chinese clinical QA dataset from Muzhi, a Baidu health platform. We select 500 pneumonia cases among 10 diseases.
- **Symptom2Disease (English)**: An English dataset containing free-text symptom descriptions from 24 diseases, curated from Kaggle. We randomly select 500 samples.

## A.2. Baseline Models

Recent domain-specific LLMs, such as Med-PaLM 2 Singhal et al. (2025), PMC-LLaMA Wu et al. (2024), and ClinicalCamel Toma et al. (2023), have shown strong performance in medical reasoning tasks by leveraging large-scale supervised training on curated clinical corpora. However, such models typically require access to sensitive patient records, expensive medical annotations, or proprietary biomedical datasets—conditions that limit their scalability and practical use in privacy-sensitive or low-resource environments.

In contrast, GPT-RagAD is designed as a lightweight, training-free framework that builds a symbolic disease-symptom knowledge graph from publicly available medical encyclopedias (e.g., Mayo Clinic) and augments reasoning through prompt-based use of general-purpose LLMs. This hybrid design eliminates the need for fine-tuning on medical data, reduces hallucination risks through constrained candidate spaces, and generalizes across languages and LLM backbones. To assess its effectiveness, we compare GPT-RagAD against two categories of baseline systems:

- **Embedding Similarity Models**: These methods compute cosine similarity between patient symptom descriptions and disease texts using pre-trained sentence encoders. We include BERT Devlin et al. (2018), RoBERTa Liu et al. (2019), and MPNet Song et al. (2020). They serve as retrieval-only baselines without reasoning or external knowledge integration.
- **Direct Generation (LLM-only)**: We evaluate GPT-4o in a direct zero-shot diagnosis setting using both vanilla and CoT-style prompts, without knowledge graph grounding or candidate filtering. This baseline reflects the upper bound of general-purpose LLM reasoning in the absence of structured symbolic support.

This comparison isolates the added value of graph-based candidate narrowing and prompt-based re-ranking in our two-layer architecture, showing how symbolic retrieval and prompt optimization complement LLM generation to improve both accuracy and reliability.

## A.3. Metrics

We adopt two widely used ranking metrics:

- **Hit@K (H@K)**: Measures whether the ground truth disease is among the top-K predicted results.
- **NDCG@K (N@K)**: Evaluates the ranking quality of the disease list by rewarding correct predictions ranked higher.

## A.4. Implementation Details

In the first layer, we generate 100 candidate diseases for each patient from the 1058-disease knowledge base using our graph-based retriever. In the second layer, we use three LLMs—LLaMA3, GPT-3.5, and GPT-4o—for re-ranking. Due to cost constraints, GPT-3.5 and GPT-4o are evaluated once on each dataset, while LLaMA3 is evaluated with three independent runs and averaged.

3. http://huggingface.co/datasets/celikmus/mayo_clinic_symptoms_and_diseases_v1

Table 5: Ablation Study: Impact of Second-Layer Re-ranking. Performance across three datasets comparing candidate retrieval (1st layer) and full re-ranking (2nd layer). Best results are in **bold**.

| Dataset | Layer | Hit@K | | | | NDCG@K | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | H@1 | H@10 | H@20 | H@50 | N@1 | N@10 | N@20 | N@50 |
| **DX** (Chinese) | 1st Layer | 0.0919 | 0.3024 | 0.4135 | 0.6445 | 0.0919 | 0.1883 | 0.2160 | 0.2615 |
| | 2nd Layer | **0.3057** | **0.7357** | **0.7901** | **0.8093** | **0.3057** | **0.5014** | **0.5104** | **0.5120** |
| **IMCS21** (Chinese) | 1st Layer | 0.0893 | 0.2831 | 0.4039 | 0.6277 | 0.0893 | 0.1679 | 0.2057 | 0.2602 |
| | 2nd Layer | **0.2989** | **0.7267** | **0.7846** | **0.8031** | **0.2989** | **0.4924** | **0.5066** | **0.5115** |
| **Symptom2Disease** (English) | 1st Layer | 0.0667 | 0.3767 | 0.5301 | 0.7842 | 0.0667 | 0.1954 | 0.2339 | 0.2696 |
| | 2nd Layer | **0.4064** | **0.7445** | **0.8159** | **0.8313** | **0.4064** | **0.5677** | **0.5779** | **0.5835** |

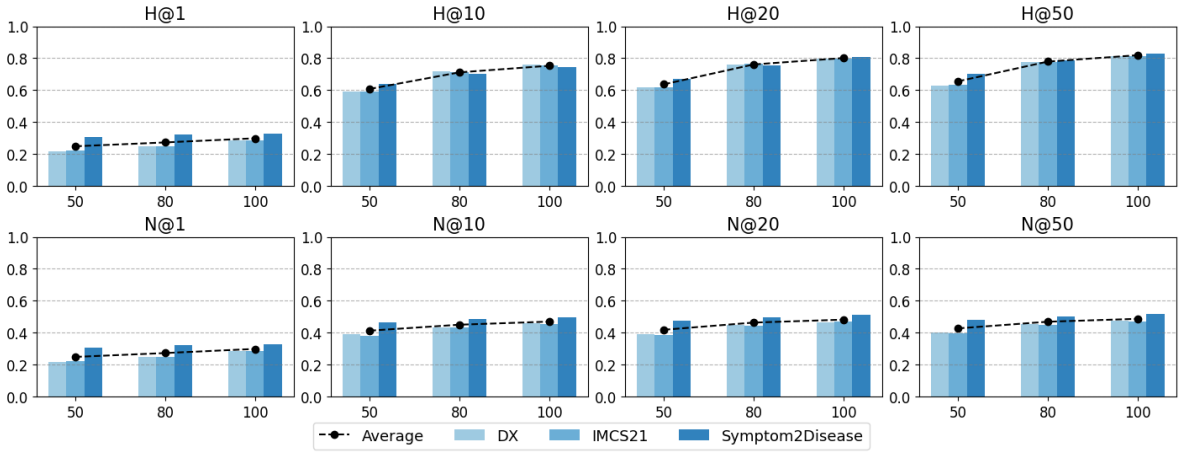


Figure 4: Impact of Candidate Set Size on Retrieval Performance Across Datasets and Metrics.

Table 6: Error Attribution Between Candidate Retrieval and Re-ranking

| Error Source | Failure Rate (%) |
|---|---|
| 1st-layer recall miss | 12.93 |
| 2nd-layer re-ranking error | 19.41 |

### A.5. Ablation and Sensitivity

To assess the impact of candidate set size on final diagnosis performance, we conduct a sensitivity analysis by varying the number of retrieved disease candidates from 50 to 80 and 100. Results across three multilingual datasets and three LLM backbones (LLaMA3, GPT-3.5, GPT-4o) are shown in Figure 4.

**Larger Candidate Sets Yield Consistent Gains.** We observe consistent performance improvements as the candidate pool expands. For instance, on the Chinese **DX** dataset using LLaMA3, **Hit@1** increases from 17.7% (50 candidates) to 30.7% (100 candidates), a relative gain of 73%. Similarly, **NDCG@50** rises from 36.9% to 48.0%, underscoring the benefit of enhanced candidate coverage.

**Improvements Generalize Across Models and Languages.** This trend holds across both English and Chinese datasets, and across all LLMs. On the English **Symptom2Disease** dataset, **GPT-4o**'s **Hit@10** improves from 65.5% to 74.4%, and **NDCG@10** increases from 52.1% to 56.8%. On the Chinese **IMCS21** dataset, **GPT-3.5**'s **Hit@50** improves from 63.5% to 81.8%, and **NDCG@50** increases from 39.8% to 42.5%. These results confirm that broader candidate pools facilitate deeper and more context-aware reasoning by the LLM.

**Top-K Precision Metrics Benefit Most.** Metrics such as **Hit@1** and **NDCG@10**, which are critical in high-stakes clinical scenarios, improve most significantly with larger candidate sets. For example, on Symptom2Disease with GPT-4o, **Hit@1** improves from 37.8% to 40.6%, and **NDCG@10** increases from 52.1% to 56.8%. Even modest improve-

Table 7: Qualitative Error Analysis: Sample Failure Cases in DX (Chinese) Dataset

| True Label | Top-1 Prediction | Error Type | Observation |
|---|---|---|---|
| Pneumonia | Asthma | Semantic Overlap | Both diseases mention "shortness of breath" and "cough", causing confusion. |
| Herpes | Influenza | Global Prior Bias | GPT favors flu due to higher frequency in corpus despite clear herpes symptoms. |
| Appendicitis | Gastritis | Ambiguous Input | LLM misinterpreted pain location due to vague phrasing: "belly discomfort". |
| HFMD | Common Cold | Insufficient KG Match | First-layer candidate set lacked the ground-truth disease. |

ments in top-ranked predictions can substantially enhance clinical reliability.

**Why 100 Candidates is the Default.** While expanding the candidate set increases computational cost, performance continues to improve up to 100 candidates with no sign of saturation. For example, on DX with GPT-4o, **Hit@1** improves from 24.5% to 30.6% as the candidate size grows from 50 to 100—a 25% relative gain. Thus, we select 100 as the default candidate size to ensure robust performance while maintaining computational feasibility.

**Summary.** This analysis underscores the importance of candidate set size in enabling accurate and robust multilingual medical diagnosis. The consistent improvements observed across datasets and models demonstrate the scalability of GPT-RagAD's two-layer architecture, and support our choice of 100 candidates as a practical and effective design default.

To isolate the contribution of each component in our two-layer GPT-RagAD framework, we conduct an ablation study comparing the performance of (1) the first-layer graph-based candidate selection alone and (2) the full two-layer system, where an LLM reranks the candidate diseases. Results are summarized in Table 5 across three datasets and multiple evaluation metrics.

**Second Layer Re-ranking Is Critical for High-Precision Diagnosis.** Across all datasets and metrics, the second-layer LLM re-ranking yields substantial performance improvements over the first-layer candidate retrieval alone. On the Chinese **DX** dataset, **Hit@1** improves from 9.2% (1st layer) to 30.6% (2nd layer), a 3.3× relative gain. Similarly, **NDCG@10** increases from 18.8% to 50.1%, highlighting that the re-ranking module not only improves recall but significantly enhances the ordering of predictions, which is crucial for clinical use cases.

**Consistent Gains Across Languages.** This pattern holds across datasets and languages. On the Chinese **imcs21** dataset, **Hit@10** increases from 28.3% to 72.7%, while **NDCG@50** grows from 26.0% to 51.1%, nearly doubling ranking quality. On the English **Symptom2Disease** dataset, the impact is even more dramatic: **Hit@1** improves from 6.7% to 40.6%, and **NDCG@10** from 19.5% to 56.8%. These consistent gains indicate that the language model is able to contextualize and reason over symptoms beyond what structural similarity alone can capture.

**LLMs Resolve Structural Limitations of the Graph Layer.** The first layer relies on structural similarity between symptom nodes and diseases in the knowledge graph, which is effective for broad filtering but limited in semantic nuance. For example, the first layer often ranks diseases with overlapping but non-definitive symptoms near the top. The second-layer LLM leverages natural language reasoning and global medical knowledge to refine these predictions, better capturing subtle linguistic cues and symptom importance. The boost from 53.0% to 81.6% in **Hit@20** on Symptom2Disease directly reflects this capability.

**Conclusion.** The ablation study clearly demonstrates that the second-layer LLM re-ranking is essential for achieving state-of-the-art performance in multilingual medical diagnosis. It dramatically enhances both recall and ranking fidelity, making GPT-RagAD more reliable in real-world applications. These findings validate our two-layer architecture and highlight the importance of combining symbolic retrieval with neural language reasoning.

Table 8: Candidate Size Sensitivity Analysis on Three Datasets with Different LLM Backbones. **Bold** indicates best and underline indicates second-best performance.

| Dataset | Metric | LLaMA3 | | | GPT-3.5 | | | GPT-4o | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 50 | 80 | 100 | 50 | 80 | 100 | 50 | 80 | 100 |
| **DX** (Chinese) | H@1 | 0.1769 | 0.2389 | **0.3074** | 0.2278 | 0.2344 | 0.2391 | 0.2454 | 0.2699 | 0.3057 |
| | H@10 | 0.5907 | 0.7052 | **0.7823** | 0.5974 | 0.7293 | 0.7581 | 0.5907 | 0.7125 | 0.7357 |
| | H@20 | 0.6057 | 0.7532 | **0.8075** | 0.6239 | 0.7754 | 0.7973 | 0.6245 | 0.7596 | 0.7901 |
| | H@50 | 0.6179 | 0.7698 | **0.8205** | 0.6421 | 0.7894 | 0.8132 | 0.6303 | 0.7698 | 0.8093 |
| | N@1 | 0.1769 | 0.2389 | **0.3074** | 0.2278 | 0.2344 | 0.2391 | 0.2454 | 0.2699 | 0.3057 |
| | N@10 | 0.3641 | 0.4007 | 0.4028 | 0.3829 | 0.4096 | 0.4702 | 0.4197 | 0.4901 | **0.5014** |
| | N@20 | 0.3659 | 0.4512 | 0.4793 | 0.3904 | 0.4047 | 0.4125 | 0.4231 | 0.4894 | **0.5104** |
| | N@50 | 0.3685 | 0.4503 | 0.4801 | 0.4073 | 0.4201 | 0.4289 | 0.4312 | 0.4927 | **0.5120** |
| **IMCS21** (Chinese) | H@1 | 0.1893 | 0.2419 | **0.3116** | 0.2239 | 0.2322 | 0.2391 | 0.2478 | 0.2760 | 0.2989 |
| | H@10 | 0.5853 | 0.7064 | 0.7625 | 0.6047 | 0.7357 | **0.7743** | 0.5874 | 0.7038 | 0.7267 |
| | H@20 | 0.6121 | 0.7701 | **0.8039** | 0.6166 | 0.7711 | 0.8013 | 0.6290 | 0.7522 | 0.7846 |
| | H@50 | 0.6345 | 0.7754 | 0.8157 | 0.6345 | 0.7807 | **0.8182** | 0.6345 | 0.7665 | 0.8031 |
| | N@1 | 0.1893 | 0.2419 | **0.3116** | 0.2239 | 0.2322 | 0.2391 | 0.2478 | 0.2760 | 0.2989 |
| | N@10 | 0.3556 | 0.4320 | 0.4667 | 0.3801 | 0.3944 | 0.4001 | 0.4128 | 0.4715 | **0.4924** |
| | N@20 | 0.3605 | 0.4436 | 0.4763 | 0.3870 | 0.4091 | 0.4131 | 0.4128 | 0.4847 | **0.5066** |
| | N@50 | 0.3666 | 0.4452 | 0.4787 | 0.3976 | 0.4131 | 0.4248 | 0.4257 | 0.4883 | **0.5115** |
| **Symptom2Disease** (English) | H@1 | 0.2735 | 0.3003 | 0.2971 | 0.2747 | 0.2759 | 0.2824 | 0.3784 | 0.3854 | **0.4064** |
| | H@10 | 0.6299 | 0.7103 | **0.7566** | 0.6264 | 0.6848 | 0.7340 | 0.6550 | 0.7141 | 0.7445 |
| | H@20 | 0.6645 | 0.7600 | 0.7901 | 0.6719 | 0.7591 | **0.8171** | 0.6795 | 0.7461 | 0.8159 |
| | H@50 | 0.7008 | 0.7804 | 0.8193 | 0.7008 | 0.7928 | **0.8382** | 0.7008 | 0.7842 | 0.8313 |
| | N@1 | 0.2735 | 0.3003 | 0.2971 | 0.2747 | 0.2759 | 0.2824 | 0.3784 | 0.3854 | **0.4064** |
| | N@10 | 0.4402 | 0.4575 | 0.4650 | 0.4385 | 0.4534 | 0.4598 | 0.5213 | 0.5422 | **0.5677** |
| | N@20 | 0.4473 | 0.4676 | 0.4724 | 0.4519 | 0.4710 | 0.4791 | 0.5272 | 0.5496 | **0.5779** |
| | N@50 | 0.4557 | 0.4728 | 0.4790 | 0.4597 | 0.4794 | 0.4847 | 0.5317 | 0.5575 | **0.5835** |

Table 9: Error Attribution Between Candidate Retrieval and Re-ranking

| Error Source | Failure Rate (%) |
|---|---|
| 1st-layer recall miss | 12.93 |
| 2nd-layer re-ranking error | 19.41 |

## Appendix B. Extended Error Analysis

## Appendix C. Error Analysis

To better understand failure modes in GPT-RagAD, we conduct qualitative and quantitative error analysis on the DX (Chinese) dataset. Table 10 summarizes representative failure cases, and Table 9 quantifies the proportion of errors attributed to retrieval-stage (top-100 candidate miss) and re-ranking-stage (top-10 miss) failures. We categorize four primary error sources:

**(1) Semantic Overlap.** Diseases with highly overlapping symptoms (e.g., *pneumonia* vs. *asthma*) often confuse both stages. Although symptom nodes are correctly matched in the knowledge graph, the LLM struggles to disambiguate without more precise contextual cues, leading to misclassification.

**(2) Global Prior Bias.** Even with candidate inclusion from the knowledge graph, LLM re-ranking remains biased toward globally frequent diseases. For instance, a case with canonical *herpes* symptoms was incorrectly ranked below *influenza*, reflecting residual training bias. Notably, 19.4% of errors occur despite the correct label being in the top-100 candidates—indicating re-ranking failures.

**(3) Ambiguous Input.** Patient descriptions with vague or nonspecific phrasing (e.g., "belly discomfort") often lead to incorrect semantic interpretation by the LLM. These input-level ambiguities account for a notable portion of re-ranking errors, where lack of discriminative features undermines accurate reasoning.

**(4) Retrieval Miss.** In 12.9% of failures, the ground-truth disease was absent from the top-100 candidates, limiting re-ranking effectiveness regardless of LLM capability. This is often due to symptom normalization mismatch or incomplete extrac-

Table 10: Qualitative Error Analysis: Sample Failure Cases in DX (Chinese) Dataset

| True Label | Top-1 Prediction | Error Type | Observation |
|---|---|---|---|
| Pneumonia | Asthma | Semantic Overlap | Both diseases mention "shortness of breath" and "cough", causing confusion. |
| Herpes | Influenza | Global Prior Bias | GPT favors flu due to higher frequency in corpus despite clear herpes symptoms. |
| Appendicitis | Gastritis | Ambiguous Input | LLM misinterpreted pain location due to vague phrasing: "belly discomfort". |
| HFMD | Common Cold | Insufficient KG Match | First-layer candidate set lacked the ground-truth disease. |

tion during KG construction—highlighting the need for improved recall in the first layer.

Additionally, bucket-level analysis (Table 4) reveals lower accuracy for generalized symptom clusters like **cough** (Hit@1: 32.2%), **diarrhea** (33.4%), and **indigestion** (36.9%). These vague disease categories introduce high lexical and semantic variability, complicating both graph matching and LLM interpretation.

**Summary.** GPT-RagAD's two-stage design mitigates some limitations of end-to-end generation, but failures persist due to residual bias, retrieval incompleteness, and language ambiguity. Future directions include contrastive training to reduce global prior reliance, symptom paraphrase expansion for KG recall, and more robust prompt formatting to reduce misalignment.

## Appendix D. Extended Related Work

### D.1. Graph-based Diagnosis

Graph-based AD systems Hosseini et al. (2018); Wang et al. (2021) typically encode medical records into Heterogeneous Information Networks (HINs) to capture structured interactions among diseases, symptoms, and other clinical entities. While effective in structured-data scenarios, these systems depend on high-quality, large-scale EHRs, which constrains generalization across domains or languages.

### D.2. LLM-based Diagnosis

LLM-based approaches Shoham and Rappoport (2023); Wang et al. (2023); Tu et al. (2024) leverage pretrained transformers for direct disease prediction or symptom synthesis. Although flexible and annotation-free, such methods are prone to global prior bias and often lack explicit reasoning over structured medical knowledge. Prior work commonly finetunes LLMs on curated datasets, which limits scalability and can reduce transparency.

### D.3. Retrieval-Augmented Generation (RAG)

RAG enhances LLMs by retrieving external knowledge to ground generation, and has proven effective in various NLP tasks Lewis et al. (2020); Ram et al. (2023); Shi et al. (2023). However, most RAG pipelines retrieve free-text passages, which may introduce hallucinations, increase latency, and raise privacy concerns—issues that are especially salient in clinical contexts.

### D.4. Positioning of Our Work

**GPT-RagAD** advances RAG-based diagnosis by introducing a symbolic retrieval layer over a disease–symptom knowledge graph constructed from public medical encyclopedias, rather than raw text or private records. The graph is embedded with a Variational Graph Autoencoder (VGAE) to model uncertainty in symptom–disease associations, and candidates are re-ranked with an LLM via optimized prompting. This design balances interpretability, scalability, and accuracy, while supporting privacy preservation and cross-lingual generalization—capabilities under-explored in prior AD or RAG literature.