ICA-RAG: Information Completeness Guided Adaptive Retrieval-Augmented Generation for Disease Diagnosis

Anonymous ACL submission

Abstract

Retrieval-Augmented Large Language Models (LLMs), which integrate external knowledge, have shown remarkable performance in medical domains, including clinical diagnosis. However, existing RAG methods often struggle to tailor retrieval strategies to diagnostic difficulty and input sample informativeness. This limitation leads to excessive and often unnecessary retrieval, impairing computational efficiency and increasing the risk of introducing noise that can degrade diagnostic accuracy. To address this, we propose ICA-RAG (**I**nformation **C**ompleteness Guided **A**daptive **R**etrieval-**A**ugmented **G**eneration), a novel framework for enhancing RAG reliability in disease diagnosis. ICA-RAG utilizes an adaptive control module to assess the necessity of retrieval based on the input's information completeness. By optimizing retrieval and incorporating knowledge filtering, ICA-RAG better aligns retrieval operations with clinical requirements. Experiments on three Chinese electronic medical record datasets demonstrate that ICA-RAG significantly outperforms baseline methods, highlighting its effectiveness in clinical diagnosis.

1 Introduction

Large Language Models (LLMs) (Achiam et al., 2023; Saab et al., 2024) have demonstrated exceptional capabilities in medical tasks, including clinical diagnosis (Zhou et al., 2024a). However, their adoption faces challenges such as hallucination—the generation of plausible but incorrect information (Maynez et al., 2020; Huang et al., 2023b)—and the resource-intensive nature of knowledge updates (Zhang et al., 2023b; Kasai et al., 2024). Retrieval-augmented generation (RAG) (Lewis et al., 2020) offers a solution by integrating trustworthy external documents to reduce hallucinations and ensure up-to-date information.

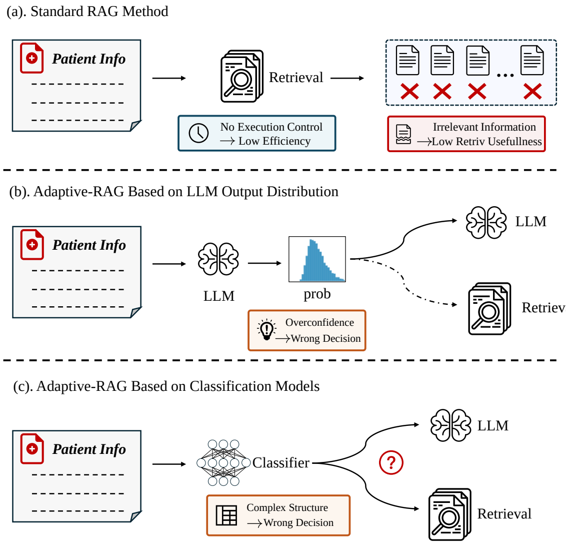


Figure 1: Illustration of three different RAG paradigms for solving clinical diagnosis task.

While researchers have extensively explored RAG to enhance LLM accuracy in high-risk domains (Zhou et al., 2024b), not all medical cases require this approach. Many common diseases or cases with mild symptoms and clear diagnoses can be accurately addressed by LLMs without retrieval (Jeong et al., 2024). However, most existing RAG methods lack selective retrieval logic, instead performing retrievals for all queries indiscriminately. This approach not only increases computational and time costs but may also introduce errors through low-quality retrievals (as shown in Figure 1.a), potentially degrading rather than improving performance.

To improve the efficiency of retrieval systems, researchers have proposed adaptive RAG paradigms (Jeong et al., 2024; Su et al., 2024; Yao et al., 2024), which establish control logic to activate the retrieval system only when certain conditions are met. There are two common approaches in these paradigms: (1) setting judgment conditions based on LLM's output text or probability distributions (Yao et al., 2024; Su et al., 2024); (2) training a relatively smaller judgment model to determine whether to perform retrieval at a lower cost (Jeong et al., 2024). Figure 1-(b) and (c) provide corresponding examples of these approaches.

However, the former approach has limitations as LLMs tend to be overconfident, generating high-confidence probability distributions even when lacking relevant knowledge (Huang et al., 2023a; Xu et al., 2024). Additionally, these methods typically require access to LLM output probability distributions (logits), limiting adaptability for API services or closed-source model applications. While the latter approach relies heavily on input content characteristics. For instance, Jeong et al. (Jeong et al., 2024) define "simple questions" as single-hop queries (e.g., "When is Michael F. Phelps's birthday?") and "difficult questions" as multi-hop queries (e.g., "What currency is used in Bill Gates's birthplace?"). Such question-answering tasks have distinct difficulty gradients, making them relatively easy for models to differentiate.

Unlike single-hop or multi-hop question answering tasks, input texts in the medical domain typically do not exhibit obvious structural patterns that can be captured, making it extremely challenging for smaller language models to understand the difficulty of answering them. Therefore, the successful experiences from this approach cannot be directly transferred to other tasks.

Based on above analysis, we proposed a disease diagnosis approach ICA-RAG, using adaptive retrieval decision optimization, specifically tailored for complex structured and long-context medical texts. The core innovation introduces a retrieval decision optimization module based on input information completeness. This module segments long inputs into text units, employs a classification model to predict each unit's importance, and calculates global information completeness to determine retrieval necessity. Since the classifier already identifies important text units, these can be prioritized during retrieval, minimizing interference from irrelevant information. Through a single prediction round, this module achieves both retrieval decision optimization and query selection, effectively addressing the limitations in existing RAG paradigms.

Our main contributions are as follows:

- We propose ICA-RAG, a framework for adaptive retrieval-augmented disease diagnosis without the need for tuning backbone LLMs.

- We desgined a novel data annotation methodology that employs masking operations to elicit varied responses from LLMs, thereby acquiring label information. Concurrently, we have optimized the retrieval process to better accommodate clinical scenarios with complex context.

- We conducted extensive experiments on three Chinese EMR datasets to demonstrate the effectiveness of our ICA-RAG framework.

2 Related Work

2.1 RAG in Clinical Disease Diagnosis

To improve diagnostic accuracy, model reliability, and reduce hallucination issues without retraining, recent studies widely adopt Retrieval-Augmented Generation (RAG) to integrate external medical knowledge (Wen et al., 2023; Wu et al., 2024; Shi et al., 2023; Thompson et al., 2023; Zhao et al., 2024b). Most research uses basic retrieval methods (Ge et al., 2024; Shi et al., 2023; Zhang et al., 2023a; Zhao et al., 2024b; Oniani et al., 2024), typically leveraging embedding models to encode external knowledge and task queries into vector representations. Relevant knowledge is retrieved via vector similarity and used in LLMs through tailored prompts for diagnosis generation. Besides, knowledge graphs are also widely employed (Wen et al., 2023; Wu et al., 2024; Gao et al., 2023).

2.2 Adaptive-RAG

Adaptive Retrieval-Augmented Generation (RAG) dynamically determines whether a large language model (LLM) requires external knowledge retrieval to mitigate inaccuracies. FLARE (Jiang et al., 2023b) and DRAGIN (Su et al., 2024) activate search engines when the LLM generates low-confidence tokens. Wang et al. (Wang et al., 2024a) use a prompting mechanism for LLMs to autonomously decide on retrieval. Self-Awareness-Guided Generation (Wang et al., 2023c) trains a classifier to assess output authenticity, while Adaptive-RAG (Jeong et al., 2024) evaluates query complexity to determine retrieval necessity. Mallen et al. (Mallen et al., 2023) propose activating retrieval based on entity frequency in queries, though this may fail for complex, multi-step reasoning tasks. Asai et al. introduce Self-RAG (Asai et al., 2023), which trains a model to dynamically retrieve, critique, and generate text.

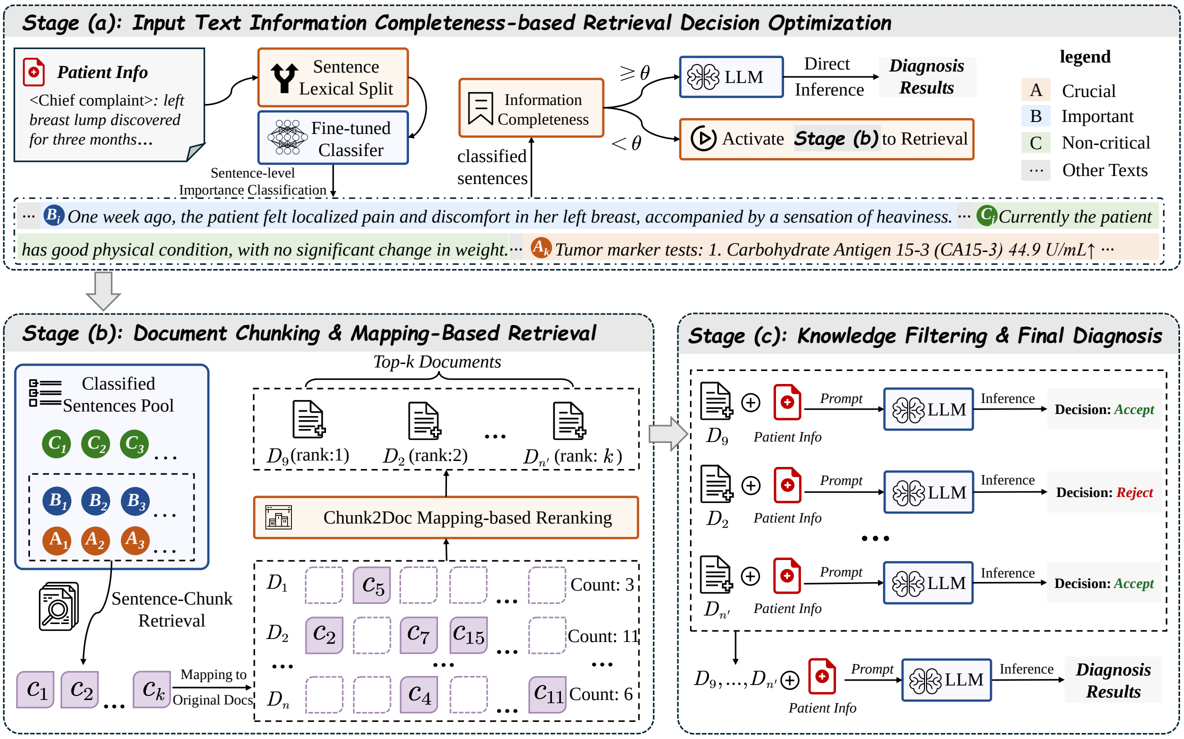

Stage (a): Input Text Information Completeness-based Retrieval Decision Optimization
Patient Info
<Chief complaint>: left breast lump discovered for three months…
Sentence Lexical Split
Fine-tuned Classifer
Sentence-level Importance Classification
Information Completeness
classified sentences
≥ θ
< θ
LLM
Direct Inference
Diagnosis Results
Activate Stage (b) to Retrieval
legend
A Crucial
B Important
C Non-critical
··· Other Texts
··· B One week ago, the patient felt localized pain and discomfort in her left breast, accompanied by a sensation of heaviness. ··· C Currently the patient has good physical condition, with no significant change in weight.··· A Tumor marker tests: 1. Carbohydrate Antigen 15-3 (CA15-3) 44.9 U/mL↑ ···
Stage (b): Document Chunking & Mapping-Based Retrieval
Classified Sentences Pool
C_1 C_2 C_3 …
B_1 B_2 B_3 …
A_1 A_2 A_3 …
Sentence-Chunk Retrieval
c_1 c_2 … c_k
Mapping to Original Docs
Top-k Documents
D_9(rank:1) D_2 (rank:2) … $D_{n'}$ (rank: k)
Chunk2Doc Mapping-based Reranking
D_1 c_5 … Count: 3
D_2 c_2 c_7 c_{15} … Count: 11
D_n c_4 … c_{11} Count: 6
Stage (c): Knowledge Filtering & Final Diagnosis
D_9 ⊕ Patient Info Prompt LLM Inference Decision: Accept
D_2 ⊕ Patient Info Prompt LLM Inference Decision: Reject
$D_{n'}$ ⊕ Patient Info Prompt LLM Inference Decision: Accept
$D_9, \ldots, D_{n'}$ ⊕ Patient Info Prompt LLM Inference Diagnosis Results

Figure 2: The overall architecture of our proposed framework ICA-RAG. It consists of three stages. Stage(a) involves inference & Retrieval Decision Making Based on Fine-Grained Information Density. Stage (b) focuses on knowledge retrieval and integration. Note that Stage (b) and (c) is activated only when the score computed in Stage (a) falls below a predefined threshold.

3 Methods

In this section, we first present the formal definition of disease diagnosis task and the task settings for adaptive-RAG-based disease diagnosis. Then we will introduce the details of each components of our proposed ICA-RAG framework.

3.1 Preliminaries

Direct Disease Diagnosis via LLM: Given a token sequence $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ representing input text, LLM-based text generation can be formalized as $\mathbf{y} = \text{LLM}(\mathbf{x}, prompt)$, where $prompt$ is a task-specific template and $\mathbf{y} = [y_1, y_2, \ldots, y_n]$ is the generated output. For disease diagnosis, the input $\mathbf{x}$ is patient information $\mathcal{Q}$, and the output $\mathbf{y}$ is the predicted diagnosis $\hat{\mathcal{D}}$, formalized as: $\hat{\mathcal{D}} = \text{LLM}(\mathcal{Q}, prompt)$.

RAG-based Disease Diagnosis: This approach retrieves relevant knowledge d from an external knowledge source $\mathcal{K}$ using a retrieval module Retriever. The diagnosis is then generated by incorporating this knowledge: $\hat{\mathcal{D}} = \text{LLM}(\mathcal{Q}, d, prompt)$, where $d = \text{Retriever}(\mathcal{K}, \mathcal{Q})$. In this paper, we use a document knowledge base KB as the external knowledge source, detailed in Appendix C.1.

Adaptive-RAG-based Disease Diagnosis: This paradigm introduces a control function F that evaluates input $\mathcal{Q}$ to determine whether retrieval is necessary:

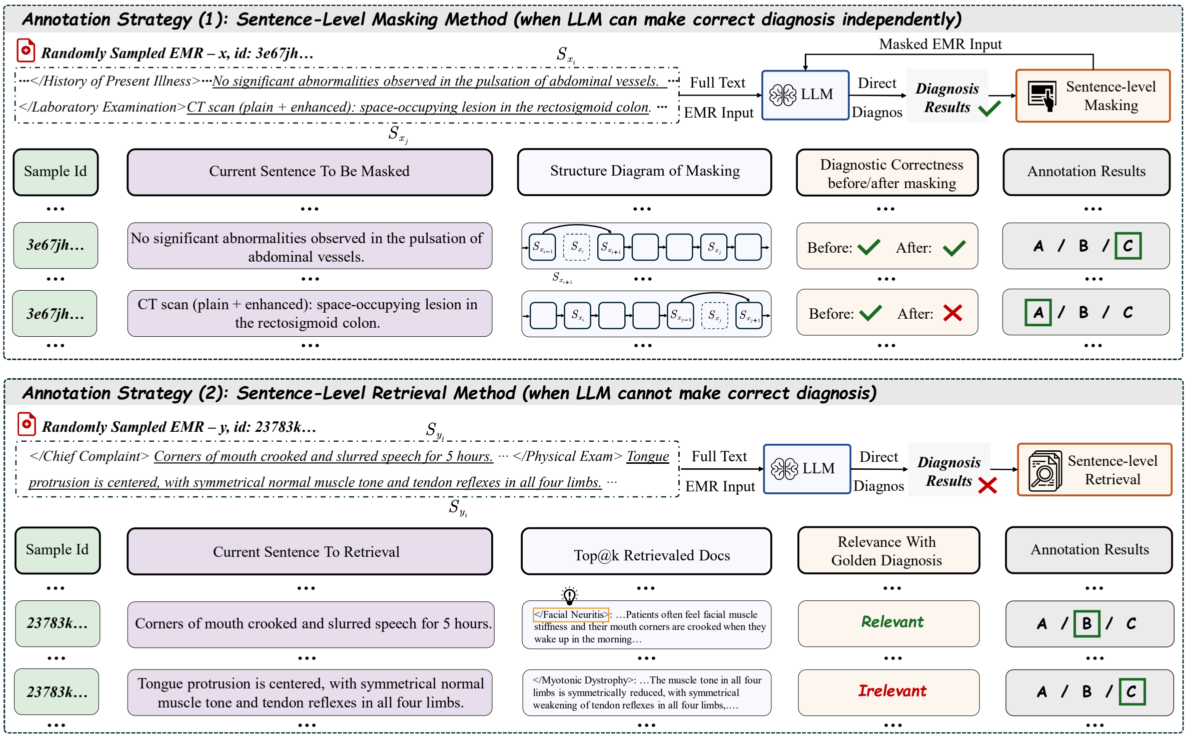
$$\hat{\mathcal{D}} = \begin{cases} \text{LLM}(\mathcal{Q}, prompt), & \text{if } F(\mathcal{Q}) = \langle Activate \rangle \\ \text{LLM}(\mathcal{Q}, d, prompt), & \text{otherwise} \end{cases} \tag{1}$$

where $d = \text{Retriever}(\mathcal{K}, \mathcal{Q})$. The control function F can be implemented through various approaches, such as LLM token probability distributions, confidence levels, or a smaller trained decision model.

3.2 Retrieval Decision Optimization Based on Input Information Completeness

3.2.1 Calculation of Input Information Completeness

Although smaller language models can evaluate the complexity of input questions and make retrieval decisions (Jeong et al., 2024), they struggle with long, complex medical diagnostic contexts. These models often rely on superficial features rather than semantic understanding when processing extensive inputs. Training larger models specifically for this


Annotation Strategy (1): Sentence-Level Masking Method (when LLM can make correct diagnosis independently)
Randomly Sampled EMR – x, id: 3e67jh…
S_{x_i}
…</History of Present Illness>…No significant abnormalities observed in the pulsation of abdominal vessels. …
</Laboratory Examination>CT scan (plain + enhanced): space-occupying lesion in the rectosigmoid colon. …
S_{x_j}
Full Text
EMR Input
LLM
Direct
Diagnos
Diagnosis Results ✓
Masked EMR Input
Sentence-level Masking
$S_{x_{i-1}}$
S_{x_i}
$S_{x_{i+1}}$
S_{x_j}
$S_{x_{j-1}}$
$S_{x_{j+1}}$
Annotation Strategy (2): Sentence-Level Retrieval Method (when LLM cannot make correct diagnosis)
Randomly Sampled EMR – y, id: 23783k…
S_{y_j}
</Chief Complaint> Corners of mouth crooked and slurred speech for 5 hours. … </Physical Exam> Tongue protrusion is centered, with symmetrical normal muscle tone and tendon reflexes in all four limbs. …
S_{y_i}
Full Text
EMR Input
LLM
Direct
Diagnos
Diagnosis Results ✗
Sentence-level Retrieval

Sample Id	Current Sentence To Be Masked	Structure Diagram of Masking	Diagnostic Correctness before/after masking	Annotation Results
…	…	…	…	…
3e67jh…	No significant abnormalities observed in the pulsation of abdominal vessels.	$S_{x_{i-1}}$ S_{x_i} $S_{x_{i+1}}$ S_{x_j}	Before: ✓ After: ✓	A / B / C
…	…	…	…	…
3e67jh…	CT scan (plain + enhanced): space-occupying lesion in the rectosigmoid colon.	S_{x_i} $S_{x_{j-1}}$ S_{x_j} $S_{x_{j+1}}$	Before: ✓ After: ✗	A / B / C
…	…	…	…	…

Sample Id	Current Sentence To Retrieval	Top@k Retrievaled Docs	Relevance With Golden Diagnosis	Annotation Results
…	…	…	…	…
23783k…	Corners of mouth crooked and slurred speech for 5 hours.	</Facial Neuritis>: …Patients often feel facial muscle stiffness and their mouth corners are crooked when they wake up in the morning…	Relevant	A / B / C
…	…	…	…	…
23783k…	Tongue protrusion is centered, with symmetrical normal muscle tone and tendon reflexes in all four limbs.	</Myotonic Dystrophy>: …The muscle tone in all four limbs is symmetrically reduced, with symmetrical weakening of tendon reflexes in all four limbs,….	Irelevant	A / B / C
…	…	…	…	…

Figure 3: Details of our proposed annotation strategy. During the annotation process, we adopt different annotation strategies based on the responses generated by the LLM.

purpose (Asai et al., 2023) is resource-intensive and contradicts RAG paradigm objectives.

To address this limitation, we segments the input $\mathcal{Q}$ into manageable text units (defaulting to sentences): $\mathcal{Q} = \{s_i\}_{i=1}^{n}$, and trains a language model Classifier to predict each unit's importance. As shown in the left half of Figure 2.Stage (a):

$$l_i = \mathsf{Classifier}(s_i) \quad \forall i \in \{1, 2, ..., n\} \tag{2}$$

This approach transforms complex document comprehension into simpler sentence-level tasks. Each text unit s_i receives one of three labels $\{\mathsf{A}, \mathsf{B}, \mathsf{C}\}$: A for information critical to diagnostic decisions, B for information that positively contributes to retrieval without directly inferring the correct result, and C for relatively unimportant information.

Based on the classification results, we calculate the global information completeness of input $\mathcal{Q}$ as follows:

$$I_{\text{norm}}(\mathcal{Q}) = \frac{1}{\alpha \cdot n} \sum_{i=1}^{n} \Big(\alpha \cdot \mathbb{I}(l_i = \mathsf{A}) \tag{3}$$

$$+ \beta \cdot \mathbb{I}(l_i = \mathsf{B}) + \gamma \cdot \mathbb{I}(l_i = \mathsf{C}) \Big) \tag{4}$$

where l_i is the classification result of text unit s_i, α, β, and γ are weights for the three category labels, and $\mathbb{I}(\cdot)$ is an indicator function that returns 1 when the condition is true and 0 otherwise. The denominator $\alpha \cdot n$ in the equation represents the maximum information completeness (when all sentences are classified as A), serving as normalization. Inputs with more critical clues increase the LLM's potential for accurate diagnosis. When I_{norm} exceeds θ_1, the input contains sufficient information for direct diagnosis:

$$\mathcal{D}_{final} = \text{LLM}(\mathcal{Q}, prompt_{diag}) \tag{5}$$

If I_{norm} falls between θ_1 and θ_2, the retrieval program activates (see Section 3.3). When I_{norm} is below θ_2, a warning signal is issued alongside normal retrieval and reasoning, indicating sparse critical information and potential misdiagnosis risk.

3.2.2 Annotation Method for Classifier Training Data Based on Masking Strategy

In the first part of this subsection, we detailed the implementation approach of the retrieval decision optimization module based on input information completeness. However, due to the lack of annotated datasets meeting our requirements for training importance classification models, we propose a simple yet effective strategy to construct and annotate training datasets. Inspired by dynamic token deletion from single-stage Weakly Supervised Rationale Extraction (Jiang et al., 2023a), we annotate the importance category of each text unit by sequentially masking them, as illustrated in Figure 3.

For a given input $\mathcal{Q}$, we set the doctor's diagnostic result $\bar{\mathcal{R}}$ as the reference answer, then segment $\mathcal{Q}$ into multiple text units $\mathcal{Q} = [s_1, s_2, \dots, s_n]$. We sequentially mask each text unit s_i to obtain the masked input $\mathcal{Q}' = [s_1, s_2, \dots, s_{i-1}, s_{i+1}, \dots, s_n]$. The LLM then performs diagnostic reasoning based on both $\mathcal{Q}$ and $\mathcal{Q}'$ to generate predicted diagnoses:

$$\hat{\mathcal{D}} = \text{LLM}(\mathcal{Q}, prompt_{diag}) \tag{6}$$

$$\hat{\mathcal{D}}' = \text{LLM}(\mathcal{Q}', prompt_{diag}) \tag{7}$$

where $\hat{\mathcal{D}}$ and $\hat{\mathcal{D}}'$ represent the diagnostic results based on the complete and masked inputs, respectively. The prompt template $prompt_{diag}$ is detailed in Table 7 in Appendix G. By comparing these diagnostic results with the standard answer $\bar{\mathcal{R}}$, we present two annotation strategies:

Annotation Strategy (1). If $\hat{\mathcal{D}} \approx \bar{\mathcal{R}}$ (the LLM makes correct predictions with complete input): If $\hat{\mathcal{D}}' \approx \bar{\mathcal{R}}$ also holds, indicating that masking s_i does not significantly impact the reasoning process, then s_i is labeled as C (non-critical information). If $\hat{\mathcal{D}}'$ differs from $\hat{\mathcal{D}}$ resulting in an incorrect diagnosis, s_i is labeled as A (critical diagnostic information). This strategy is illustrated in the upper part of Figure 3.

Annotation Strategy (2). If $\hat{\mathcal{D}} \neq \bar{\mathcal{R}}$ (the LLM cannot make correct predictions with complete input): In this case, we implement annotation by searching the knowledge base. We use s_i as the retrieval query with the BM25 method. If documents corresponding to the disease in $\bar{\mathcal{R}}$ can be retrieved, s_i is labeled as B (valuable diagnostic information). Otherwise, s_i is labeled as C (low importance). This strategy is illustrated in the lower part of Figure 3.

3.3 Knowledge Retrieval and Reranking Based on Document Segmentation and Mapping

Considering the complex structures, large context spans, and semantic discontinuities in clinical texts, we adapt the RAG process following Zhao et al. (Zhao et al., 2024a). This approach divides documents in the knowledge base KB into text chunks with length restrictions, using sentences as the minimum segmentation unit (details in Appendix C.1). Figure 2.Stage-b illustrates our retrieval and reranking workflow.

Given an input text $\mathcal{Q} = \{s_i\}_{i=1}^{n}$ with n sentences, we first perform sentence-level importance classification and calculate overall information completeness I_{norm} as described in Section 3.2.1 (1). When I_{norm} falls below a preset threshold, the retrieval module Stage-b activates. To optimize retrieval efficiency, we only retain sentences with $\mathsf{label} = \mathsf{A}$ and $\mathsf{label} = \mathsf{B}$, excluding those with $\mathsf{label} = \mathsf{C}$ (shown on the left side of Figure 2.Stage-b). This exclusion is justified as $\mathsf{label} = \mathsf{C}$ sentences typically contain non-pathological descriptions that contribute minimally to retrieval and may introduce noise.

The retrieval algorithm operates on knowledge base KB through chunk-level retrieval and document-level reranking. Each sentence $s_i \in \mathcal{Q}$ serves as a query to retrieve the top m relevant text chunks:

$$\mathcal{C}_i = \mathsf{Retriever}(s_i, m) \quad \forall i \in \{1, 2, \dots, n\} \tag{8}$$

where $\mathcal{C}_i = \{c_{i,j}\}_{j=1}^{m}$, and $c_{i,j}$ is the j-th chunk retrieved using s_i. All text chunk sets are merged into $\mathcal{C} = \bigcup_{i=1}^{n} \mathcal{C}_i$. Each chunk $c \in \mathcal{C}$ is mapped to its original document $doc \in \mathsf{KB}$. For each document doc, a score S_{doc} counts the number of retrieved chunks from that document:

$$S_{doc} = \sum_{c \in \mathcal{C}} \mathbb{I}(c \in doc) \tag{9}$$

where $\mathbb{I}(\cdot)$ is the indicator function, returning 1 if c belongs to doc and 0 otherwise. Documents are reranked based on S_{doc}, and the top k documents with highest scores are selected as the final retrieval results: $\mathcal{K}_{rerank} = \{doc_{(1)}, doc_{(2)}, \dots, doc_{(k)}\}$, where $doc_{(l)}$ represents the document with the l-th highest score.

3.4 Knowledge Filtering and Diagnosis Generation Based on Prompt Guidance

Despite optimizing the retrieval process, retrieved documents may not always be relevant, particularly in clinical diagnostic tasks requiring complex reasoning. Drawing inspiration from medical "differential diagnosis" procedures, where doctors examine potentially confusing diseases based on patient symptoms and test results, we designed a prompt template $prompt_{diff}$ to filter irrelevant information.

This template guides the LLM to identify conflicts between patient information and document descriptions, determining which documents to retain. The process is illustrated in Figure 2.Stage-c, with the complete prompt template detailed in Table 9 in Appendix G.

Given the reranked knowledge document set $\mathcal{K}_{rerank} = \{doc_{(1)}, doc_{(2)}, \ldots, doc_{(k)}\}$, we filter documents by evaluating their relevance to the diagnosis. The filtering function $V(\mathcal{Q}, doc_{(i)}, prompt_{diff})$ is defined as:

$$V(\mathcal{Q}, doc_{(i)}, prompt_{diff}) = \begin{cases} \text{True}, & \text{if } \langle\text{support}\rangle \\ \text{False}, & \text{otherwise} \end{cases} \tag{10}$$

where $\langle\text{support}\rangle$ represents the LLM output when provided with query $\mathcal{Q}$, document $doc_{(i)}$, and prompt template $prompt_{diff}$. The term $\langle\text{support}\rangle$ indicates that the LLM determines $doc_{(i)}$ is critical for diagnosis. The final reference knowledge document set $\mathcal{K}^*$ retains only documents that satisfy the filtering condition:

$$\mathcal{K}^* = \{doc_{(i)} \in \mathcal{K}_{rerank} \mid V(\cdot, doc_{(i)}, \cdot) = \text{True}\} \tag{11}$$

The final RAG-based diagnostic process is formalized as:

$$\mathcal{D}_{final} = \text{LLM}(\mathcal{Q}, \mathcal{K}^*, prompt_{rag}) \tag{12}$$

where $prompt_{rag}$ represents the RAG-based diagnostic prompt template. The complete content of this prompt template can be found in Table 8 in Appendix G.

4 Experimental Setup

4.1 Datasets

We evaluated our framework using three Chinese EMR datasets: CMEMR (Jia et al., 2025), ClinicalBench (Yan et al., 2024), and CMB-Clin (Wang et al., 2023a), to assess its ability in analyzing complex clinical information and making accurate diagnoses. For the task setup, all three datasets are configured into end-to-end diagnostic tasks, where patient information (such as chief complaints, medical history, and examination findings) serves as input, with physicians' diagnostic conclusions as ground-truth labels. Details can be found in Appendix B.

4.2 Baseline Methods

We compare our approach with three categories of methods. Details of all the baselines below are shown in Appendix A.

Non-Retrieval methods: We include Chain-of-Thought (CoT) (Wei et al., 2022a), Self-Consistent Chain of Thought(Sc-CoT) (Wang et al., 2023b) and Atypical Prompting (Qin et al., 2024).

Standard-Retrieval methods: We include two representative RAG methods: RAG^2 (Rationale-Guided RAG)(Sohn et al., 2024) and LongRAG (Zhao et al., 2024a).

Adaptive-Retrieval methods: We include Adaptive-RAG (Jeong et al., 2024), DRAGIN (Su et al., 2024), and SEAKR (Yao et al., 2024).

4.3 Evaluation Metric

Following (Fan et al., 2024), we use the International Classification of Diseases (ICD-10) (Percy et al., 1990) to standardize disease terminologies. We extract disease entities from diagnostic results and EMR labels, then perform fuzzy matching with a threshold of 0.5 to link them to ICD-10, creating normalized sets $S_{\hat{\mathcal{D}}}$ and $S_{\mathcal{R}}$. These sets are used to calculate set-level metrics Precision, Recall, and F1-score. Details are shown in Appendix D.

4.4 Implementation Details

We choose qwen2.5-7B-instruct as the backbone model for inference in our experiments by default. For the classifier we choose BERT-base-Chinese (Devlin et al., 2019). For the retriever we use BM25(Robertson et al., 2009) by default. For the external knowledge corpus we use CMKD (Clinical Medicine Knowledge Database)[1]. Detailed settings of each module and hyperparameters are provided in Appendix C.

5 Results and Analyses

5.1 Overall Performance

Our experiments evaluate the framework against baselines on three Chinese EMR datasets. Table 1 highlights key findings:

(1) ICA-RAG demonstrates consistent performance across all benchmark datasets, achieving optimal or near-optimal F1 scores compared to baseline methods.

(2) Compared to LongRAG, a superior conventional retrieval approach, ICA-RAG improves Set-level F1 values by 1.81%, 1.54%, and 1.72% respectively on the three datasets. This indicates that standard RAG methods without retrieval decision optimization rely excessively on knowledge

[1] http://cmkd.juhe.com.cn/

Method	CMEMR			ClinicalBench			CMB-Clin		
	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)	R (%)	P (%)	F1 (%)
Non Retrieval Methods									
CoT	49.09	48.56	48.82	44.12	34.09	38.46	68.03	42.27	52.14
SC-CoT	49.49	48.21	48.84	42.74	33.41	37.50	69.27	43.43	53.39
ATP	49.68	47.72	48.68	43.01	33.82	37.87	70.83	44.73	**54.83**
Standard Retrieval Methods									
RAG[2]	47.13	44.34	45.69	42.43	34.57	38.10	58.33	35.29	43.98
LongRAG	49.85	48.31	49.07	44.65	35.02	39.25	69.44	41.32	51.81
Adaptive Retrieval Methods									
DRAGIN	47.09	46.92	47.00	43.67	35.54	39.19	59.72	36.13	45.03
Adaptive-RAG	50.23	48.35	49.27	42.23	34.61	38.04	65.37	**45.20**	53.44
SEAKR	47.37	45.90	46.62	40.66	33.13	36.51	59.60	34.34	43.57
ICA-RAG (ours)	**53.42**	**48.58**	**50.88**	**46.63**	**36.24**	**40.79**	**71.62**	42.74	53.53

Table 1: Experimental results on CMEMR, ClinicalBench and CMB-Clin datasets. Bold indicates the best performances and the second-best performances are underlined.

base quality in complex disease diagnosis scenarios. They initiate retrieval even when LLMs can independently complete diagnoses, reducing efficiency and potentially introducing errors.

(3) ICA-RAG outperforms other adaptive RAG methods significantly. Compared to the best-performing Adaptive-RAG method, ICA-RAG exhibits enhanced robustness when handling structurally complex and long context inputs due to its adaptive decision-making based on local-to-global information completeness calculations. Most other baselines, on the other hand, are designed primarily for simpler question answering tasks, so their performance fluctuations when applied to disease diagnosis without appropriate adaptations.

5.2 Ablation Study

Table 2: Ablation study on CMEMR dataset. *w/o* denotes removing the corresponding module.

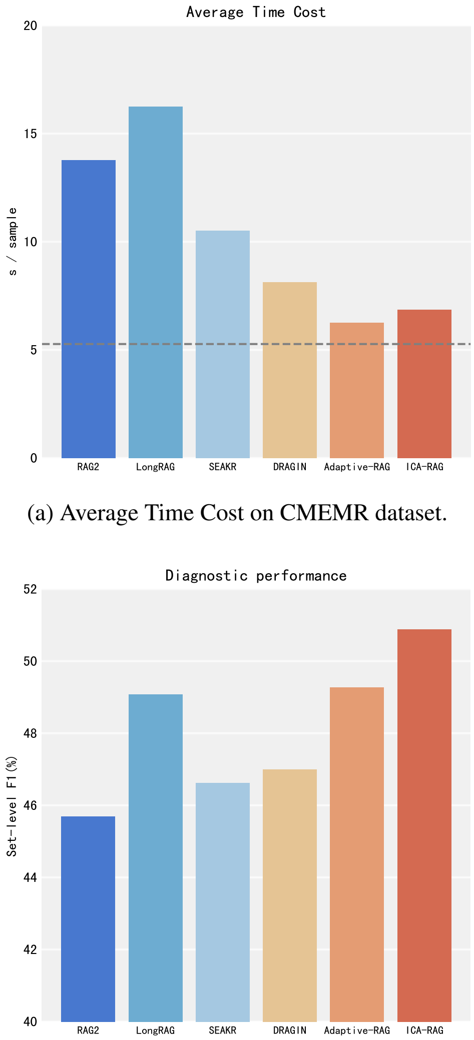
Method	R (%)	P (%)	F1 (%)
ICA-RAG	**53.42**	**48.58**	**50.88**
w/o Decision	49.74	46.52	48.07
w/o Chunk	52.26	47.53	49.78
w/o M-rerank	52.22	47.20	49.59
w/o Diff	52.70	47.29	49.85

To analyze the contribution of different modules in ICA-RAG to its performance, we conducted ablation experiments on the CMEMR dataset: (a) *w/o* Decision: removing the retrieval decision optimization module; (b) *w/o* Chunk: replacing ICA-RAG's document segmentation and mapping-based knowledge retrieval with direct retrieval of complete documents; (c) *w/o* M-rerank (Mapping-based Rerank): replacing ICA-RAG's text chunk mapping-based reranking with the bge-reranker-v2-m3 model; (d) *w/o* Diff: removing the LLM knowledge filtering module based on differential diagnosis prompting. The results are shown in Table 2, leading to the following conclusions:

(1) Without the retrieval decision optimization module, ICA-RAG's F1 value dropped by 2.81%. This occurs because all inputs undergo retrieval indiscriminately, forcing samples that LLM could diagnose independently to undergo unnecessary retrieval, reducing efficiency and increasing error risk from irrelevant information.

(2) Replacing ICA-RAG's document segmentation and mapping-based retrieval with original retrieval decreased F1 by 1.1%. This demonstrates that general RAG methods struggle with sparse information distribution and semantic incoherence in clinical texts, hampering accurate matching between inputs and knowledge base documents.

(3) Substituting ICA-RAG's reranking method with the bge-reranker-v2-m3 model reduced performance, validating ICA-RAG's reranking design. ICA-RAG's approach relies solely on numerical calculations from retrieval results without additional models, reducing memory overhead while maintaining higher compatibility with the retrieval



(a) Average Time Cost on CMEMR dataset.

(b) Diagnostic Performance on CMEMR dataset.

Figure 4: A Comparative Analysis of Computational Time Expenditure and Diagnostic Performance Between the Proposed Method and Selected Baseline Methods on the CMEMR Dataset.

workflow.

(4) Removing the differential diagnosis-based knowledge filtering mechanism meant all retrieved documents were provided to the LLM without discrimination. This increased the difficulty of LLM's reasoning and raised the probability of exceeding input length limits, negatively impacting overall performance.

5.3 Analysis of Retrieval Decision Optimization Effects

We compare our method with other retrieval-based baselines in terms of efficiency and diagnostic performance, as shown in Figure 4. Based on the experimental results, the following conclusions can be drawn:

(1) As illustrated in Figure 4.a, our method demonstrates significant time efficiency advantages compared to non-adaptive RAG methods (RAG^2 and LongRAG), reflecting the improvements from decision optimization.

(2) Compared to adaptive RAG methods (SEAKR, DRAGIN, and Adaptive-RAG), our approach shows competitive time consumption, only slightly higher than Adaptive-RAG but lower than SEAKR and DRAGIN. Unlike SEAKR and DRAGIN, which require access to LLMs' output probability distributions, our method maintains adaptability for closed-source LLMs and API-based deployments. While both Adaptive-RAG and our method employ classifiers for decision optimization, our BERT-Base classifier (110M parameters) is more lightweight than Adaptive-RAG's T5-Large (770M parameters).

(3) Figure 4.b demonstrates that our method achieves superior diagnostic performance. Overall, the proposed approach better balances efficiency and performance compared to baseline methods.

6 Conclusion

In this paper, we propose ICA-RAG, an adaptive retrieval decision optimization method for disease diagnosis that addresses the rigid retrieval strategy issue in traditional retrieval-augmented methods. ICA-RAG establishes a decision mechanism based on input information completeness to flexibly determine retrieval necessity, and introduces a retrieval and reranking strategy using document segmentation and mapping. Experimental results demonstrate ICA-RAG's strong adaptability in complex clinical scenarios. Future work may explore further optimization of the retrieval process and ICA-RAG's application to other medical tasks.

Limitations

Although our classification data annotation strategy is straightforward and effective, it still exhibits certain shortcomings in practical application. Due to the potential presence of repetitive content within the input patient information, LLMs may still arrive at a correct diagnosis even after masking a critical sentence. This can result in inaccurate annotation labels, necessitating manual inspection and revision on top of our proposed annotation strategy. Moreover, clinical medical texts, particularly EMRs, often contain abbreviations, synonyms, and aliases. And the manner in which identical patient information is recorded can vary significantly among different physicians, leading to a high degree of inconsistency. This issue to some extent hampers the search accuracy of our retrieval system. In the future, we aim to explore more effective preprocessing strategies for medical texts to enhance retrieval quality.

Ethical Consideration

In this paper, we focus on the medical domain, specifically on enhancing the reliability and efficiency of retrieval-augmented generation (RAG) systems for disease diagnosis using large language models (LLMs). Our goal is to support better-informed decision-making by adaptively determining the necessity of information retrieval based on the information completeness of the input data. While our results demonstrate significant improvements in diagnostic accuracy and efficiency with the ICA-RAG framework, we need to stress that LLMs, even when augmented with retrieval mechanisms, should not be solely relied upon without the oversight of a qualified medical expert. The involvement of a physician or an expert is essential to validate the model's recommendations and ensure a safe and effective decision-making process.

Moreover, we acknowledge the profound ethical implications of deploying AI in healthcare. It is crucial to recognize that LLMs are not infallible and can produce erroneous outputs, even with advanced retrieval mechanisms. Transparency in how these models, including the decision-making process of ICA-RAG (e.g., why retrieval was or was not triggered), reach their conclusions, and incorporating continuous feedback from healthcare professionals are vital steps in maintaining the integrity and safety of medical practice.

References

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

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

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

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Interactive evaluation and collaboration of llms as intern doctors for clinical diagnosis. *arXiv preprint arXiv:2402.09742*.

Yanjun Gao, Ruizhe Li, Emma Croxford, Samuel Tesch, Daniel To, John Caskey, Brian W Patterson, Matthew M Churpek, Timothy Miller, Dmitriy Dligach, et al. 2023. Large language models and medical knowledge grounding for diagnosis prediction. *medRxiv*, pages 2023–11.

Jin Ge, Steve Sun, Joseph Owens, Victor Galvez, Oksana Gologorskaya, Jennifer C Lai, Mark J Pletcher, and Ki Lai. 2024. Development of a liver disease-specific large language model chat interface using retrieval augmented generation. *Hepatology*, pages 10–1097.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023a. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7029–7043.

Mingyi Jia, Junwen Duan, Yan Song, and Jianxin Wang. 2025. medikal: Integrating knowledge graphs as assistants of llms for enhanced clinical diagnosis on emrs. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 9278–9298. Association for Computational Linguistics.

Han Jiang, Junwen Duan, Zhe Qu, and Jianxin Wang. 2023a. You only forward once: Prediction and rationalization in a single forward pass. *arXiv preprint arXiv:2311.02344*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie

Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. 2024. Realtime qa: what's the answer right now? *Advances in Neural Information Processing Systems*, 36.

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.

Dingkun Long, Qiong Gao, Kuan Zou, Guangwei Xu, Pengjun Xie, Rui Guo, Jianfeng Xu, Guanjun Jiang, Luxi Xing, and P. Yang. 2022. Multi-cpr: A multi domain chinese dataset for passage retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 22.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

David Oniani, Xizhi Wu, Shyam Visweswaran, Sumit Kapoor, Shravan Kooragayalu, Katelyn Polanska, and Yanshan Wang. 2024. Enhancing large language models for clinical decision support by incorporating clinical practice guidelines. *arXiv preprint arXiv:2401.11120*.

Constance Percy, Valerie van Holten, Calum S Muir, World Health Organization, et al. 1990. *International classification of diseases for oncology*. World Health Organization.

Jeremy Qin, Bang Liu, and Quoc Nguyen. 2024. Enhancing healthcare llm trust with atypical presentations recalibration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2520–2537.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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

Wenqi Shi, Yuchen Zhuang, Yuanda Zhu, Henry Iwinski, Michael Wattenbarger, and May Dongmei Wang. 2023. Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–10.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2024. Rationale-guided retrieval augmented generation for medical question answering. *arXiv preprint arXiv:2411.00300*.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. *arXiv preprint arXiv:2403.10081*.

Will E Thompson, David M Vidmar, Jessica K De Freitas, John M Pfeifer, Brandon K Fornwalt, Ruijun Chen, Gabriel Altay, Kabir Manghnani, Andrew C Nelsen, Kellie Morland, et al. 2023. Large language models with retrieval-augmented generation for zero-shot disease phenotyping. *arXiv preprint arXiv:2312.06457*.

Keheng Wang, Feiyu Duan, Peiguang Li, Sirui Wang, and Xunliang Cai. 2024a. Llms know what they need: Leveraging a missing information guided framework to empower retrieval-augmented generation. *arXiv preprint arXiv:2404.14043*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024b. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Jiangnan Xia, Zuyi Bao, Liwei Peng, and Luo Si. 2020. Structbert: Incorporating language structures into pre-training for deep language understanding. In *8th International Conference on Learning Representations*.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. 2023a. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023c. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10303–10315.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022a. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2023. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *arXiv preprint arXiv:2308.09729*.

Jiageng Wu, Xian Wu, and Jie Yang. 2024. Guiding clinical reasoning with large language models via knowledge seeds. *arXiv preprint arXiv:2403.06609*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Yang Wang. 2024. Perils of self-feedback: Self-bias amplifies in large language models. *arXiv e-prints*, pages arXiv–2402.

Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, Jiayi Wang, Weishan Zhao, Yixin Zhang, Renjun Zhang, et al. 2024. Clinicallab: Aligning agents for multi-departmental clinical diagnostics in the real world. *arXiv preprint arXiv:2406.13890*.

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Weichuan Liu, Lei Hou, and Juanzi Li. 2024. Seakr: Self-aware knowledge retrieval for adaptive retrieval augmented generation. *arXiv preprint arXiv:2406.19215*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923.

Haodi Zhang, Jiahong Li, Yichi Wang, and Yuanfeng Songi. 2023a. Integrating automated knowledge extraction with large language models for explainable medical decision-making. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1710–1717. IEEE.

Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *Preprint*, arXiv:2110.06696.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023b. How do large language models capture the ever-changing world knowledge? a review of recent advances. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311.

Qingfei Zhao, Ruobing Wang, Yukuo Cen, Daren Zha, Shicheng Tan, Yuxiao Dong, and Jie Tang. 2024a. Longrag: A dual-perspective retrieval-augmented generation paradigm for long-context question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22600–22632.

Wenting Zhao, Zhongfen Deng, Shweta Yadav, and Philip S Yu. 2024b. Heterogeneous knowledge grounding for medical question answering with retrieval augmented large language model. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1590–1594.

Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Sirui Ding, Jiashuo Wang, Kaishuai Xu, Yi Fang, et al. 2024a. Large language models for disease diagnosis: A scoping review. *arXiv preprint arXiv:2409.00097*.

Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Sirui Ding, Jiashuo Wang, Kaishuai Xu, Yi Fang, et al. 2024b. Large language models for disease diagnosis: A scoping review. *ArXiv preprint*, abs/2409.00097.

A Details of Baseline Methods

In this section, we provide a detailed introduction to the three categories of baseline methods used in this paper, namely Non-Retrieval methods, Standard-Retrieval methods and Adaptive-Retrieval methods, Their methodological descriptions and implementation details are listed below.

A.1 Non-Retrieval Methods

These methods do not rely on external knowledge, but rather leverage the internal knowledge of LLMs through prompt optimization for reasoning. This chapter selects the classic Chain of Thought (CoT) (Wei et al., 2022b) and Self-Consistent Chain of Thought (SC-CoT) (Wang et al., 2023b) as baselines. Additionally, we include a method called Atypical Prompting (ATP) (Qin et al., 2024), which is designed specifically for the medical domain and enhances reasoning capabilities by focusing on non-typical factors such as scenarios and symptoms.

A.2 Standard-Retrieval Methods

In this category, "Standard" corresponds to "Adaptive" mentioned later, referring to the RAG method that initiates retrieval for all inputs uniformly without Adaptive settings. This paper selects two representative baselines: RAG^2 (Rationale-Guided RAG) (Sohn et al., 2024) and LongRAG (Zhao et al., 2024a).

RAG^2 (Rationale-Guided RAG) (Sohn et al., 2024) enhances the original input by utilizing rationales generated by LLM based on the input question, and then performs subsequent retrieval operations with the enhanced input. This approach was tested on medical question-answering tasks.

LongRAG (Zhao et al., 2024a) designs a strategy that integrates global information perspective and factual detail perspective for long-text retrieval tasks. It improves the overall understanding and processing capability for long texts by prompting LLM to extract global information and analyze retrieved document information.

A.3 Adaptive-Retrieval Methods

This paradigm enhances retrieval flexibility and controllability by presetting conditions or introducing additional models. Retrieval is only activated when the input meets preset conditions; otherwise, results are directly inferred by the LLM. We selected three representative baselines, including Adaptive-RAG (Jeong et al., 2024), DRAGIN (Su et al., 2024), and SEAKR (Yao et al., 2024).

Adaptive-RAG (Jeong et al., 2024) labels training data based on the correctness of LLM responses to certain samples, and trains a classification model to determine the complexity of multi-hop question answering problems to decide whether to perform retrieval.

DRAGIN (Su et al., 2024) measures uncertainty by calculating the entropy of token probability distributions, utilizing the Transformer's self-attention mechanism to quantify the influence of tokens on subsequent content.

SEAKR (Yao et al., 2024) introduces self-aware uncertainty, determining whether to activate the retrieval model based on this value.

A.4 Settings of Baseline Methods

To ensure a fair comparison, we implement all baseline methods using the same backbone LLM, retriever, and external knowledge corpus by default. For baseline methods that require training a classifier (RAG^2 (Sohn et al., 2024) and Adaptive-RAG (Jeong et al., 2024)), we adopt the same language model as used in our framework, namely Mengzi-T5-base (Zhang et al., 2021).

B Details of Datasets

CMEMR (Jia et al., 2025) CMEMR is sourced from a Chinese medical website[2] and comprises 10,450 electronic medical records (EMRs) spanning 15 departments. During the collection process, records with missing critical information or other deficiencies were excluded via screening. The official repository for this dataset has not provided a formal license.

ClinicalBench (Yan et al., 2024) ClinicalBench originates from authentic EMRs from officially certified Grade A Class III hospitals in China, encompassing 1,500 records across 24 departments. The creators of this dataset have furnished a comprehensive data usage license, which explicitly stipulates that the dataset is limited to non-commercial academic research use.

CMB-Clin (Wang et al., 2023a) CMB-Clin is a constituent dataset of CMB benchmark, primarily derives its content from official medical textbooks. It comprises diagnostic procedures for various disease types, compiled into 74 complete medical records and 208 associated clinical diagnostic questions. The official repository for this dataset is licensed under the Apache-2.0 license.

For language, all these datasets are in Chinese. According to the source papers of the above three datasets, their construction processes all strictly adhered to privacy protection principles, with the personally identifiable information and sensitive data such as treatment locations being concealed or removed. During experiments, we have strictly adhered to the stipulations set forth by the creators of each dataset, employing these datasets exclusively for the purpose of experimental evaluation.

[2] https://bingli.iiyi.com/

C Implementation Details

C.1 Details of the Retrieval Module

We employ the CMKD (Clinical Medicine Knowledge Database) as the external knowledge corpus. Knowledge documents for all 5,200 diseases were obtained from the official website. An example is provided in Chinese (Figure 7) and English (Figure 8) version. Following (Zhao et al., 2024a), we preprocess the documents in the knowledge base by segmenting them into chunks prior to retrieval.

Specifically, we impose a length constraint on the chunks, using sentences as the minimum segmentation unit. A sliding window is then applied to extend the context by merging overlapping content from the end of the previous sentence, thereby preventing semantic discontinuity at truncation points. Short chunks at the end of a document are merged with preceding chunks to ensure better semantic coherence. Furthermore, since the knowledge documents for each disease are inherently semi-structured, containing fixed fields such as "Etiology," "Clinical Manifestations," "Laboratory Tests," and "Other Auxiliary Examinations," we terminate the current chunk at the end of the text corresponding to each field during the segmentation process. By default, we set the chunk size to 200 words after segmenting the documents of CMKD.

During the retrieval process, we set the default retriever as the sparse retriever bm25(Robertson et al., 2009), and all retrievers are implemented using the retriv library[3]. The number of text chunks m retrieved for each sentence s_i is set to 100, and the number of documents after mapping text chunks to documents and reranking k is set to 5. During retrieval, chunks with similarity scores below 50% are discarded.

C.2 Details of the Classifier

We adopt BERT-Base-Chinese (Devlin et al., 2019) as the foundation model for our classifier, training it for 2 epochs with a learning rate of 3e-5 and AdamW (Loshchilov and Hutter, 2019) optimizer. For training data, we extract samples from existing medical record datasets. Since the CMEMR dataset provides comprehensive departmental coverage and is significantly larger than ClinicalBench and CMB-Clin datasets, we sample 5% of CMEMR records according to departmental proportions ($CMEMR_{subset}$, 516 samples) for entity weight calculation. These samples are completely excluded from subsequent experiments, with testing conducted only on the remaining 95% of CMEMR. For ClinicalBench and CMB-Clin datasets, all samples are used for evaluation.

In the annotation process, we follow the strategy described in Section 3.2.2. For scenario (1), even when the LLM makes correct predictions after removing sentence s_i, this sentence may still contain valuable information due to content redundancy across sections (e.g., symptoms appearing in both chief complaint and present illness history). To prevent information loss and annotation errors, we implement an additional retrieval step for sentences labeled as $\mathsf{label} = \mathsf{C}$. If documents corresponding to diseases in $\bar{\mathcal{R}}$ can be retrieved using s_i, we update its label to B.

C.3 Details of LLM Inference Settings

We conducted our experiments using a single NVIDIA GeForce RTX 3090 GPU. Due to memory constraints, for inference with large-scale backbone models (such as Qwen2.5-14B), we utilized the API provided by the Siliconflow platform[4]. During inference, we set the maximum generation length of the LLM to 2048. To ensure reproducibility, we set `do_sample` to `False` by default.

C.4 HyperParameters

When calculating the information density based on the classifier's predictions, the weights α, β, and γ for labels A, B, and C are set to 1.0, 0.5, and 0.1, respectively. The thresholds θ_1 and θ_2 are set to 0.3 and 0.1, respectively.

For the retrieval process, the number of chunks m retrieved for each sentence s_i is set to 100, and the number of documents k after chunk-to-document mapping and re-ranking is set to 5. During retrieval, chunks with a similarity score below 50% to the given query s_i are discarded.

D Evaluation Metrics Calculation

To enhance evaluation rigor, we follow Fan et al. (Fan et al., 2024) by adopting the International Classification of Diseases (ICD-10) (Percy et al., 1990) to link natural language diagnoses with standardized clinical terminology. For predicted disease entities $\hat{\mathcal{D}}$ and reference diagnoses $\mathcal{R}$, we employ fuzzy matching (threshold 0.5) to map these entities to standardized disease sets $S_{\hat{\mathcal{D}}}$ and $S_{\mathcal{R}}$.

[3] https://github.com/AmenRa/retriv

[4] https://www.siliconflow.cn/

Based on the above setup, this chapter redefines the following statistical values:

True Positives (TP): The number of standard disease terms in the prediction results $S_{\hat{\mathcal{D}}}$ that correctly correspond to the reference diagnosis $S_{\mathcal{R}}$.

False Positives (FP): The number of standard disease terms that appear in the prediction results $S_{\hat{\mathcal{D}}}$ but do not correctly match with the reference diagnosis $S_{\mathcal{R}}$.

False Negatives (FN): The number of standard disease terms that appear in the reference diagnosis $S_{\mathcal{R}}$ but are omitted in the prediction results $S_{\hat{\mathcal{D}}}$.

Finally, based on the above statistical values, this chapter can calculate set-level evaluation metrics for the two sets $S_{\hat{\mathcal{D}}}$ and $S_{\mathcal{R}}$: Set-level Recall, Set-level Precision, and Set-level F1 score:

$$\text{Set-level } R = \frac{TP}{TP + FN} \tag{13}$$

$$\text{Set-level } P = \frac{TP}{TP + FP} \tag{14}$$

$$\text{Set-level } F1 = \frac{2 \times P \times R}{P + R} \tag{15}$$

In all experiments of this paper, any metrics related to "P", "R", and "F1" refer to the set-level metrics defined above.

E Detailed Experimental Results

E.1 The Effects of Different Classification Models

To verify the universality and robustness of our proposed retrieval decision optimization module based on input information completeness across various classification models, we evaluated two additional pre-trained language models: Struct-BERT (Wang et al., 2020; Yin et al., 2019) and T5 (Zhang et al., 2021). We trained these models on our annotated data and assessed their text unit importance classification accuracy and final diagnostic performance. As shown in Table 3, BERT-base achieved the highest classification accuracy (86.28%), while the generative T5-base model performed slightly lower than the self-encoding architectures of BERT and StructBERT, despite having more parameters. Nevertheless, all models maintained robust classification performance, with trends consistent with their final diagnostic performance. These results demonstrate the strong cross-model adaptability and robustness of ICA-RAG's adaptive retrieval decision optimization module.

To further analyze the rationality of ICA-RAG's text unit importance categorization, Figure 5 presents the confusion matrices for the three models. The matrices reveal that label B is frequently misclassified as label A, while label C is rarely misclassified. This pattern is intuitive—models can easily distinguish between expressions like "normal diet and sleep" and "obvious purpura on both lower limbs" based on their importance difference. However, differentiating between similarly pathological information such as "obvious purpura on both lower limbs" and "multiple thyroid nodules" proves challenging. This indicates current limitations in distinguishing between decisive and important information, highlighting directions for future improvements.

Table 3: Performances of different classification models on CMEMR dataset.

Model	Acc (%)	R (%)	P (%)	F1 (%)
BERT	**86.28**	**53.42**	48.58	**50.88**
StructBERT	85.28	53.09	**48.61**	50.75
T5	84.34	52.50	48.20	50.25

E.2 The Effects of Different Retrievers

To further investigate the retrieval module design and verify the universality of our method, we compare our approach with other RAG methods using multiple retrievers: the sparse retriever BM25 (our default choice), and three dense retrievers: E5 (Wang et al., 2024b), BGE-m3 (Xiao et al., 2024), and CoROM (Long et al., 2022). Results in Table 4 show that our method achieves optimal performance across different retrievers with minimal variation (maximum difference of only 0.96% in Set-level F1 scores). This validates the rationality of our document segmentation and mapping-based knowledge retrieval strategy, which reduces document-level search to simpler text segment matching tasks. Furthermore, it demonstrates that our adaptive control module can selectively retain or filter information without additional overhead, maintaining high stability across different retrievers and facilitating broader application scenarios.

E.3 Performance Analysis Across Different LLMs for Inference

Table 5 demonstrates the performance of different foundation models as inference models on the CMEMR dataset. Considering that the official LLaMA models from metaAI perform poorly on Chinese tasks, we use the Chinese versions of

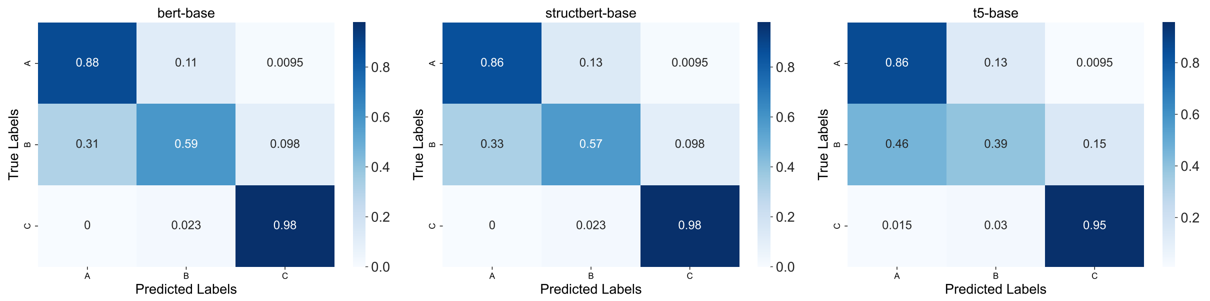


Figure 5: Classification accuracy comparison between ICA-RAG and other two baseline methods Adaptive-RAG (Jeong et al., 2024) and RAG^2 (Sohn et al., 2024). We also provide the confusion matrix across three labels (Right).

Methods	bm25	BGE	E5	CoROM
RAG^2	45.69	42.80	46.64	44.97
LongRAG	49.07	48.93	49.15	48.24
Adaptive-RAG	49.27	48.19	48.83	47.78
DRAGIN	47.00	45.51	47.62	42.25
SEAKR	46.62	43.07	44.31	45.18
ICA-RAG(ours)	**50.88**	**50.21**	**51.17**	**50.34**

Table 4: Performance comparison (in F1-score) of using different retrievers on CMEMR dataset. Bold indicates the best performances.

Backbone	R (%)	P (%)	F1 (%)
Qwen2.5-7B-Instruct	53.42	48.58	50.88
Qwen2.5-14B-Instruct	**56.68**	**51.49**	**53.96**
GLM4-9B-Chat	43.48	41.56	42.50
LLaMA3-8B-Chinese	46.66	38.87	42.41
LLaMA3.1-8B-Chinese	45.58	47.81	46.67

Table 5: Performance comparison of different inference LLMs on CMEMR dataset. Bold indicates the best performances.

LLaMA3[5] and LLaMA3.1[6] released by Wang et al. for our experiments. The experimental results indicate that the inference capabilities of these LLMs significantly influence diagnostic performance in three aspects:

(1) Under the same conditions, larger models generally yield better performance. For instance, when the parameter size of the Qwen2.5 model increases from 7B to 14B, its performance on the CMEMR dataset improves by 3.08%. (2) With the iterative upgrades of model versions, diagnostic performance also shows qualitative improvements. For example, from LLaMA3 to LLaMA3.1, the Set-level F1 increases by 4.26%. (3) When pre-training corpora, training strategies, and model architectures differ, model performance varies accordingly. For instance, although GLM4-9B-Chat has 1∼2B more parameters than Qwen2.5-7B and LLaMA3.1-8B, its actual diagnostic performance lags significantly behind the other two models.

E.4 Results on Different Clinical Departments

To investigate ICA-RAG's performance in diagnostic tasks across different medical departments, we compared it with several representative baseline methods on samples from major departments in the CMEMR dataset. The results in Figure 6 reveal that:

(1) ICA-RAG consistently outperforms other baseline methods using adaptive retrieval strategies across all departments, confirming its effectiveness in various departmental diagnostic tasks.

(2) All methods, including ICA-RAG, show relatively lower diagnostic performance in dermatology, oncology, and obstetrics and gynecology. This can be attributed to the high feature overlap in dermatological conditions, the heavy reliance on imaging information in oncology, and the unique nature of obstetric cases where routine pregnancy examinations are often analyzed through a pathological diagnostic framework. These observations provide valuable directions for future improvements.

[5] https://huggingface.co/shenzhi-wang/Llama3-8B-Chinese-Chat

[6] https://huggingface.co/shenzhi-wang/Llama3.1-8B-Chinese-Chat

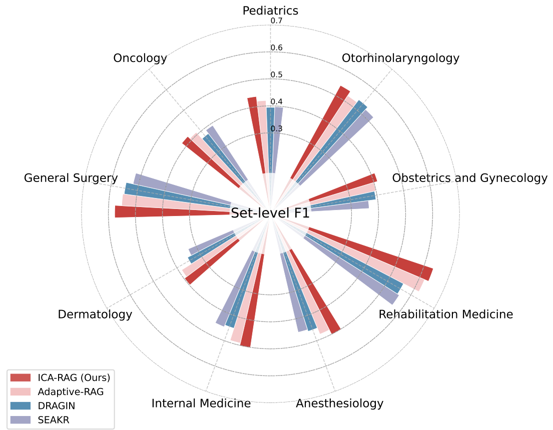


Figure 6: Diagnostic performance of ICA-RAG (ours) and selected baseline methods on samples from major clinical departments in the CMEMR dataset.

F Case Study

Table 6 presents representative case studies demonstrating the practicality of our proposed ICA-RAG. The results show that ICA-RAG conducts fine-grained importance assessment of patient information, accurately determining if the current data suffices for diagnosis and initiating retrieval when appropriate. Unlike previous Adaptive-Retrieval methods, our approach warns when I_{norm} falls below a threshold, indicating potential diagnostic failure. This meets clinical requirements for balancing accuracy and reliability, highlighting practical significance.

G Prompt Templates

Document Example (Chinese)

【疾病名】: 病毒性心肌炎
【英文名】: Viral Myocarditis
【ICD号】: I41.0*
【分类】: 心血管内科

概述: 病毒性心肌炎(viral myocarditis)是一种与病毒感染有关的局限性或弥漫性炎症性心肌疾病，是最常见的感染性心肌炎。近年来随着检测技术的提高，发现多种病毒可引起心肌炎，其发病率呈逐年增高趋势，是遍及全球的常见病和多发病。

流行病学: ... 病毒性心肌炎可发生于各个年龄段，但从临床发病情况看以儿童和40岁以下成人居多。由于许多病毒感染具有明显的季节分布特点，如流感病毒感染多发生在冬季，而肠道病毒感染则多发生于夏秋季，因此，病毒性心肌炎的发病也具有明显的季节特征，夏秋季发病率较高，冬春季较少。...

病因: ...目前已证实能引起心肌炎的病毒包括：(1)小核糖核酸病毒：肠道病毒如柯萨奇(Coxsackie)、埃可(ECHO)、脊髓灰质炎病毒、鼻病毒等；(2)虫媒病毒：如黄热病毒、登革热病毒、白蛉热病毒、流行性出血热病毒等；...

临床表现: ... 在临床就诊的患者中，90%左右以心律失常为主诉或首发症状，常诉心悸、乏力、胸闷、头晕等，严重者可出现晕厥或阿-斯综合征。部分患者可有程度不一的胸痛，其原因可能有：...

实验室检查: ...血清心肌肌钙蛋白I(cTnI)或肌钙蛋白T(cTnT)增高(以定量测定为准)有较大价值。...

其他辅助检查: ...X线检查 约1/4病人有不同程度心脏扩大，搏动减弱，其扩大程度与心肌损害程度一致，有时可见心包积液(病毒性心肌心包炎)，严重病例因左心功能不全有肺淤血或肺水肿征象。...

鉴别诊断: 1.风湿性心肌炎 有典型风湿热表现者，则两者鉴别不难，一般可从以下几点作鉴别：风湿性心肌炎常有扁桃体炎或咽峡炎等链球菌感染史，抗"O"增高，血沉降多明显增快，C反应蛋白(CRP)阳性，心电图改变以P-R间期延长较常见，咽拭物培养常有链球菌生长，且多有大关节炎，鉴于风湿性心肌炎常有心内膜炎，因此二尖瓣反流性收缩期杂音多较明显，且可因瓣膜水肿、炎症出现舒张期杂音(Carey Coombs杂音)，若心脏扩大不明显，而杂音较响亮，则风湿性可能性更大。...

治疗: (1)应用改善心肌细胞营养与代谢药物：该类药物包括维生素C、维生素B、辅酶A 50～100U或肌苷200～400mg，每天肌内注射或静脉注射1～2次；细胞色素C 15～30mg，每天静脉注射1～2次，该药应先皮试，无过敏者才能注射。...

Figure 7: A document example from CMKD (in Chinese).

Data Example (English)

[Disease Name]: Viral Myocarditis
[English Name]: –
[ICD Code]: I41.0*
[Classification]: Cardiovascular Medicine

Overview: Viral myocarditis is a localized or diffuse inflammatory myocardial disease associated with viral infections, and it is the most common infectious myocarditis. In recent years, with the improvement of detection techniques, it has been found that various viruses can cause myocarditis, and its incidence has been increasing year by year, making it a common and frequently occurring disease worldwide.

Epidemiology: ... Viral myocarditis can occur in all age groups, but clinically, it is more common in children and adults under 40 years old. Since many viral infections have distinct seasonal distribution characteristics, such as influenza virus infections occurring mostly in winter and enterovirus infections occurring mostly in summer and autumn, the incidence of viral myocarditis also has obvious seasonal characteristics, with higher incidence in summer and autumn and lower incidence in winter and spring....

Etiology: ... It has been confirmed that viruses that can cause myocarditis include: (1) Picornaviruses: enteroviruses such as Coxsackie, ECHO, poliovirus, rhinovirus, etc.; (2) Arboviruses: such as yellow fever virus, dengue virus, sandfly fever virus, epidemic hemorrhagic fever virus, etc.;...

Clinical Manifestations: ... Among patients who seek clinical consultation, about 90% report arrhythmia as their main complaint or initial symptom, often complaining of palpitations, fatigue, chest tightness, dizziness, etc. In severe cases, syncope or Adams-Stokes syndrome may occur. Some patients may experience varying degrees of chest pain, which may be due to:...

Laboratory Tests: ... Elevated serum cardiac troponin I (cTnI) or troponin T (cTnT) (based on quantitative measurements) is of significant value....

Other Auxiliary Examinations: ... X-ray examination shows that about 1/4 of patients have varying degrees of cardiac enlargement and weakened pulsations, with the degree of enlargement consistent with the degree of myocardial damage. Sometimes pericardial effusion (viral myopericarditis) can be seen, and severe cases may show signs of pulmonary congestion or pulmonary edema due to left heart dysfunction....

Differential Diagnosis: 1. Rheumatic myocarditis: For those with typical rheumatic fever manifestations, the differentiation is not difficult. Generally, the following points can be used for differentiation: rheumatic myocarditis often has a history of streptococcal infections such as tonsillitis or pharyngitis, elevated anti-streptolysin O (ASO), significantly increased erythrocyte sedimentation rate (ESR), positive C-reactive protein (CRP), and common ECG changes such as prolonged P-R interval. Throat swab cultures often grow streptococci, and there is often polyarthritis. Since rheumatic myocarditis often involves endocarditis, the systolic murmur of mitral regurgitation is usually more pronounced, and diastolic murmurs (Carey Coombs murmur) may appear due to valve edema and inflammation. If the heart is not significantly enlarged but the murmur is loud, rheumatic myocarditis is more likely....

Treatment: (1) Use of drugs that improve myocardial cell nutrition and metabolism: These drugs include vitamin C, vitamin B, coenzyme A 50~100U, or inosine 200~400mg, administered intramuscularly or intravenously once or twice daily; cytochrome C 15~30mg, administered intravenously once or twice daily. This drug should be skin-tested first, and only those without allergies can be injected....

Figure 8: A document example from CMKD (translated).

Case 1: Non-retrieval
[Patient Info]: <Chief Complaint>: Pain in the right upper abdomen for 2 days... <History of Present Illness>: ...Persistent pain with paroxysmal exacerbation, accompanied by nausea and vomiting (vomitus consisted of gastric contents), as well as abdominal distension and poor appetite... <Physical Examination>: ...No abdominal muscle tension or palpable masses... <Auxiliary Examination>: ...Color ultrasound indicates gallbladder sludge and stones... **[I_{norm}]:** 0.63 **[Activate_Retreival]:** False **[Raise_Warning]:** False **[LLM Diagnosis]:** Gallstones and acute cholecystitis (✓)
Case 2: Retrieval
[Patient Info]: ...<History of Present Illness>: ...Previously treated at a local hospital with enteric-coated aspirin tablets and isosorbide mononitrate, but no significant improvement was observed...<Physical Examination>: ...The heart rhythm is regular, and no pathological murmurs are heard in any of the valve auscultation areas... <Auxiliary Examination>: ...During the Bruce protocol exercise test, at 2 minutes and 14 seconds, tall tent-shaped T waves appeared in the precordial leads, accompanied by upsloping ST-segment elevation in the corresponding leads (Figure 2). Simultaneously, the patient experienced chest tightness... **[I_{norm}]:** 0.47 **[Activate_Retreival]:** True **[Raise_Warning]:** False **[Retrieved Documents]:** ...Transient episodes of chest pain induced by exercise or other conditions that increase myocardial oxygen demand... In some patients with spontaneous angina, transient ST-segment elevation occurs during episodes, known as variant angina. New-onset exertional angina, worsening exertional angina, and spontaneous angina are often collectively referred to as "unstable angina."... **[LLM Diagnosis]:** Coronary heart disease, unstable angina (✓)
Case 3: Warning
[Patient Info]: <Chief Complaint>: Recurrent pain in both knees for 8 years, worsening over the past month.... <Past Medical History> Previously healthy, preoperative blood tests and coagulation function tests were normal, and color Doppler ultrasound of the arteries and veins of both lower limbs showed no abnormalities. <Physical Examination>: ...No localized redness or swelling in the bilateral knee joints, with normal muscle tone.... **[I_{norm}]:** 0.27 **[Activate_Retreival]:** True **[Raise_Warning]:** True **[Retrieved Documents]:** Knee synovitis: The common sites of disease are the knee and hip joints, and the general symptoms are joint pain and significant limitation of movement... **[LLM Diagnosis]:** Knee synovitis (✗)

Table 6: Case Study. For clarity, only part of the key information of the selected samples is presented.

[**Role**]<SYS>

You are an outstanding AI medical expert. You can perform a preliminary disease diagnosis based on the patient's Information.

[**Role**]<USR>

Below is a medical record summary of a patient from the ${department}. Please act as the attending physician and provide a diagnosis based on your expertise and knowledge.

[Medical Record Summary]:

###

${summary}

###

[Requirements]:

1. You need to comprehensively analyze the patient's symptoms, medical visits, medical history, and various examination results.

2. Please provide your diagnosis using the following template.

[Output Template]:

Diagnosis: [Predicted Disease 1: [Disease Name 1]; Predicted Disease 2: [Disease Name 2]; ...; Predicted Disease n: [Disease Name n]]

Please strictly adhere to the output template and do not include any irrelevant information!

Table 7: The default prompt template for LLM direct diagnosis. The presence of a "$" symbol indicates a placeholder variable that needs to be filled with specific content.

[**Role**]<SYS>

You are an outstanding AI medical expert. You can perform a preliminary disease diagnosis based on the patient's Information.

[**Role**]<USR>

Below is a medical record summary of a patient from the ${department}. Please act as the attending physician and provide a diagnosis based on your expertise and knowledge.

[Medical Record Summary]:

###

${summary}

###

Additionally, by searching the medical knowledge base, you have identified several suspected diseases and have extracted relevant information from them as follows, for reference:

[Knowledge Document]:

${External Knowledge Documents}

[Requirements]:

1. You need to comprehensively analyze the patient's symptoms, medical history, examination results, and other relevant information.

2. You should make full use of your medical knowledge and may refer to the knowledge documents you retrieved. Please note! The knowledge in the documents may contain errors or misleading information, so you must carefully evaluate and avoid blindly following them!

3. After the above analysis and thinking process, please provide your diagnosis using the following template.

[Output Template]:

Diagnosis: [Predicted Disease 1: [Disease Name 1]; Predicted Disease 2: [Disease Name 2]; ...; Predicted Disease n: [Disease Name n]]

Please strictly adhere to the output template and do not include any irrelevant information!

Table 8: The default prompt template for RAG-based LLM diagnosis. The presence of a "$" symbol indicates a placeholder variable that needs to be filled with specific content.

[**Role**]<SYS>
You are an outstanding AI medical expert. You can perform a preliminary disease diagnosis based on the patient’s Information.

[**Role**]<USR>
Below is a medical record summary of a patient from the ${department}.
[Medical Record Summary]:
###
${summary}
###
Based on the above, you tried to search in the medical knowledge base and retrieved the following document from the knowledge base:
${External Knowledge Documents}

[The Conception of Differential Diagnosis]
When analyzing the given documents, you may refer to the method of "differential diagnosis" in clinical medicine: by analyzing the degree of concordance between the patient’s onset cause, presenting symptoms, examination indicators, and the characteristics of the diseases described in the current document, you can determine the relevance of the document for reference. Additionally, you need to compare whether there are contradictions or significant inconsistencies between the patient’s condition and the descriptions in the document. If such inconsistencies exist, you should consider that the current document may not provide accurate diagnostic guidance.

[Requirements]:
Your task is to match the patient’s condition with the description in the knowledge base document, analyzing any content that matches or conflicts. Then, use your knowledge to think critically and ultimately determine whether the knowledge base document is valuable for diagnosis. If you think it is valuable, select "True"; if you think it is misleading or irrelevant, select "False".
Please output in the following JSON format and do not output anything else:
{"status": "the value of status"}

Table 9: The default prompt template for LLM filtering the retrieved document via differential diagnosis prompt. The presence of a "$" symbol indicates a placeholder variable that needs to be filled with specific content.