AutoMIR: Effective Zero-Shot Medical Information Retrieval without Relevance Labels

Lei Li[1], **Xiangxu Zhang**[1], **Xiao Zhou**[123*], **Zheng Liu**[4*]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Beijing Key Laboratory of Research on Large Models and Intelligent Governance
[3]Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE
[4]Beijing Academy of Artificial Intelligence
{leil, xansar, xiaozhou}@ruc.edu.cn, zhengliu1026@gmail.com

Abstract

Medical information retrieval (MIR) is vital for accessing knowledge from electronic health records, scientific literature, and medical databases, supporting applications such as medical education, patient queries, and clinical diagnosis. However, effective zero-shot dense retrieval in the medical domain remains difficult due to the scarcity of relevance-labeled data. To address this challenge, we propose **S**elf-**L**earning **Hy**pothetical **D**ocument **E**mbeddings (**SL-HyDE**), a framework that leverages large language models (LLMs) to generate hypothetical documents conditioned on a query. These documents encapsulate essential medical context, guiding dense retrievers toward the most relevant results. SL-HyDE further employs a self-learning mechanism that iteratively improves pseudo-document generation and retrieval using unlabeled corpora, eliminating the need for labeled data. In addition, we introduce the Chinese Medical Information Retrieval Benchmark (CMIRB), a comprehensive evaluation suite reflecting real-world medical scenarios, comprising five tasks and ten datasets. By benchmarking ten models on CMIRB, we provide a rigorous standard for evaluating MIR systems. Experimental results demonstrate that SL-HyDE significantly outperforms HyDE in retrieval accuracy, while exhibiting strong generalization and scalability across diverse LLM and retriever configurations. Our code and data are publicly available at: https://github.com/ll0ruc/AutoMIR.

1 Introduction

Medical information retrieval (MIR) (Luo et al., 2008; Goeuriot et al., 2016) focuses on retrieving relevant medical information from sources like electronic health records, scientific papers, and medical knowledge databases, based on specific medical queries. Its applications are wide-ranging, supporting doctors in clinical decision-making (Sivarajkumar et al., 2024), assisting patients in seeking health-related information (McGowan et al., 2009), and aiding researchers in accessing pertinent studies (Zheng and Yu, 2015).

Dense retrievers (Karpukhin et al., 2020; Xu et al., 2024) have shown strong performance when trained on large labeled datasets in information retrieval (IR). Several studies (Xiong et al., 2020; Li et al., 2023; Xiao et al., 2024) have successfully employed contrastive learning to develop general-purpose text embedding models, achieving promising results in zero-resource retrieval scenarios. They leverage large-scale weakly supervised data through web crawling as well as high-quality text pairs derived from data mining or manual annotation. However, the availability of such large-scale datasets cannot always be guaranteed, particularly in non-English languages or specialized domains.

Recently, large language models (LLMs) have demonstrated exceptional performance in zero-resource retrieval scenarios (Wang et al., 2023a; Shen et al., 2023), primarily due to their extensive knowledge and powerful text generation capabilities. This makes them particularly effective in situations where labeled data are scarce or unavailable. One such approach, HyDE (Gao et al., 2023), employs zero-shot prompts to guide an instruction-following language model to generate hypothetical documents, effectively narrowing the semantic gap between the query and the target document. Similarly, Query2doc(Wang et al., 2023a) uses few-shot prompting of LLMs to generate pseudo-documents, which are then used to augment the original query.

However, applying these HyDE-style methods to medical information retrieval presents three critical challenges: (1) **LLMs lack the specialized medical knowledge necessary to generate highly relevant hypothetical documents**. HyDE employs general-purpose LLMs for pseudo-document generation, which are often insufficiently equipped

*Xiao Zhou and Zheng Liu are corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24028–24047
November 4-9, 2025

with domain-specific knowledge, particularly in fields like medicine. Community efforts such as HuatuoGPT (Zhang et al., 2023) and PMC-LLaMA (Wu et al., 2024) highlight the necessity of fine-tuning on medical data to enhance domain-specific capabilities. Although these medical LLMs have richer medical knowledge, their outputs are suboptimal in aligning with retrieval optimization needs. (2) **General text embedding models are inadequate for representing medical queries and documents effectively**. These versatile retrievers (Xiao et al., 2024; Wu et al., 2023) are typically designed for multi-domain and multi-task settings, but fail to capture the nuanced and knowledge-intensive nature of the medical domain. (3) **The medical domain suffers from a scarcity of high-quality, relevance-labeled datasets**. Such resources are particularly limited in non-English languages, where annotation demands substantial domain expertise and is further constrained by strict privacy regulations. This shortage substantially raises the cost of training and fine-tuning retrieval models to achieve strong performance.

To address these issues, we propose **S**elf-**L**earning **Hy**pothetical **D**ocument **E**mbedding (**SL-HyDE**), an effective fully zero-shot dense retrieval system that requires no relevance-labeled data for medical information retrieval. During the inference phase, SL-HyDE first employs an LLM as the generator to produce a relevant hypothetical document in response to a medical query. A retrieval model is then used to identify the most relevant target document from the candidates based on the generated hypothetical document. In the training phase, we design a self-learning mechanism that enhances the retrieval performance of SL-HyDE without the need for labeled data. Specifically, this mechanism leverages the retrieval model's ranking capabilities to select high-relevance hypothetical documents that align with the outputs of the generator (LLM), simultaneously injecting medical knowledge into the LLM. In turn, the generator's ability to produce high-quality hypothetical documents provides pseudo-labeled data for the training of the retrieval model, enabling it to efficiently encode medical texts. This interactive and complementary approach generates supervisory signals that enhance both the generation and retrieval capabilities of the system. Notably, SL-HyDE begins with unlabeled medical corpora and completes the training process through a self-learning mechanism, thereby circumventing the heavy reliance on labeled data typically required for training both large language models and text embedding models.

To evaluate SL-HyDE's performance in Chinese medical information retrieval, we develop a valuable **C**hinese **M**edical **I**nformation **R**etrieval **B**enchmark (**CMIRB**). CMIRB is constructed from real-world medical scenarios, including online consultations, medical examinations, and literature retrieval. It comprises five tasks and ten datasets, representing the first comprehensive and authentic evaluation benchmark for Chinese medical information retrieval. This benchmark is expected to accelerate advancements toward building more robust and generalizable MIR systems in the future.

Through extensive experimentation on CMIRB, we find that our proposed method significantly enhances retrieval performance. We validate SL-HyDE across various configurations involving three large language models as generators and three embedding models as retrievers. Notably, SL-HyDE surpasses the HyDE (Qwen2 as generator + BGE as retriever) combination by an average of 4.9% in NDCG@10 across ten datasets, and achieves a 7.2% improvement compared to using BGE alone for retrieval. These outcomes underscore the effectiveness and versatility of SL-HyDE. In summary, our contributions are as follows:

- We propose Self-Learning Hypothetical Document Embeddings for zero-shot medical information retrieval, eliminating the need for relevance-labeled data.
- We introduce the first comprehensive Chinese Medical Information Retrieval Benchmark and evaluate the performance of various text embedding models on it.
- SL-HyDE enhances retrieval accuracy across five tasks and demonstrates strong generalizability and scalability with different combinations of generators and retrievers.

2 Related Work

2.1 Dense Retrieval

Recent advancements in deep learning and natural language processing have significantly advanced information retrieval and recommendation systems (Xiong et al., 2020; Xiao et al., 2024; Ma et al., 2024; Li et al., 2024; Li and Zhou, 2025). Contriever (Izacard et al., 2021) leverages unsupervised contrastive learning for dense retrieval.

PEG (Wu et al., 2023) and BGE (Xiao et al., 2024) enhance Chinese general embeddings by training on large-scale text pairs. These works illustrate the impact of well-structured training strategies on effective retrieval across multiple domains. Beyond embedding-based techniques, large language models have demonstrated exceptional performance in zero-resource retrieval scenarios. GAR (Mao et al., 2021) enriches query semantics with generated content. HyDE (Gao et al., 2023) generates hypothetical documents for the retriever, effectively narrowing the semantic gap between the query and the target document. Query2doc (Wang et al., 2023a) utilizes few-shot prompts to expand queries, boosting both sparse and dense retrieval.

However, retrieval using LLM-generated documents often yields suboptimal results when domain-specific knowledge is limited. To address this limitation, we propose a self-learning framework that jointly optimizes the generator and retriever without the need for relevance labels, thereby improving retrieval performance.

2.2 Information Retrieval Benchmark

To better guide the development of retrieval models, researchers have developed various datasets and benchmarks. For instance, DuReader (He et al., 2018), a large-scale Chinese reading comprehension dataset, has substantially advanced text understanding and information retrieval research. BEIR(Thakur et al., 2021), a zero-shot retrieval evaluation benchmark, covers diverse retrieval tasks and offers a unified evaluation platform. MTEB (Muennighoff et al., 2023) establishes a framework for evaluating multilingual text embeddings. More recently, C-MTEB (Xiao et al., 2024) specifically targets Chinese text embedding evaluations. However, these benchmarks are designed for general domains, limiting their applicability to specialized fields such as medical retrieval. Existing medical benchmarks like TREC Collections (Voorhees et al., 2021) and NFCorpus (Boteva et al., 2016) are highly valuable for MIR evaluation, but they are limited in scale and cover few medical scenarios. To bridge this gap, we develop the first comprehensive and realistic evaluation benchmark based on real-world medical scenarios for Chinese medical information retrieval tasks.

2.3 Large Language Models in Medicine

Large language models (Team, 2024; Guo et al., 2025; Yong et al., 2025; Zhou et al., 2025) have shown strong potential in general domains. HuatuoGPT (Zhang et al., 2023) distills clinician-supervised consultation data, and PMC-LLaMA (Wu et al., 2024) leverages large-scale biomedical literature for instruction-tuning, highlighting the necessity of medical data fine-tuning to improve clinical reasoning and QA. Similarly, biomedical retrievers like BMRetriever (Xu et al., 2024) demonstrate that domain-adaptive fine-tuning on biomedical corpora is crucial for accurate evidence retrieval. However, the HyDE paradigm requires a retrieval model that is robust to hypothetical documents and a generator that produces retrieval-preferred documents. In this work, we jointly optimize the generator and retriever toward a shared objective of enhancing retrieval under the HyDE-style pipeline.

3 Methodology

3.1 Preliminary

Zero-shot document retrieval is a fundamental component of search systems. Given a user query q and a document set $D = \{d_1, ..., d_n\}$, where n represents the number of candidate documents, the goal of a retrieval model ($\mathcal{M}_r$) is to identify documents that align with the user's genuine search intent for the given query q. These models map an input query q and a document d into a pair of vectors $<v_q, v_d>$, using their inner product as a similarity function $s(q, d)$:

$$s(q, d) =< \mathcal{M}_r(q), \mathcal{M}_r(d) > . \quad (1)$$

The retrieval model then selects the top-k documents, denoted as D_{topk}, which achieve the highest similarity scores when compared to the query q.

Large language models have achieved remarkable success across various natural language processing tasks, including question answering (Liu et al., 2022) and text generation (Dathathri et al., 2019). Recently, there has been a growing interest in leveraging these models to generate query-relevant documents, thereby improving retrieval accuracy. Hypothetical Document Embeddings (HyDE) (Gao et al., 2023) decompose dense retrieval into two components: a generative task performed by an instruction-following language model and a document-document similarity task executed by a retrieval model.

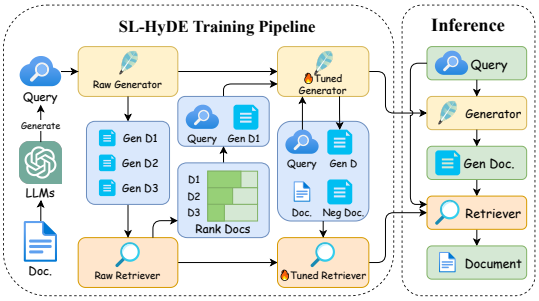

Figure 1: Training and inference pipeline of SL-HyDE.

3.2 Overview

Applying HyDE to the medical domain presents two primary challenges: (1) LLMs often lack specialized medical domain knowledge, and (2) retrievers may struggle to accurately encode medical texts due to inadequate training on medical corpora. These challenges hinder the successful application of HyDE in the medical field, making it difficult to achieve substantial performance improvements in retrieval tasks. A common strategy to enhance medical domain knowledge involves fine-tuning with labeled medical data (Zhang et al., 2023; Wang et al., 2024b; Xu et al., 2024). However, these approaches depend on high-quality, manually constructed data to adapt general models to the medical domain. Unfortunately, obtaining such high-quality labeled data in practice is particularly challenging, rendering the training of a medical LLM both difficult and costly.

In this paper, we introduce a self-learning hypothetical document embedding mechanism designed to exploit the potential of unlabeled medical corpora. The labels are entirely generated by the generator and retriever within SL-HyDE, eliminating the need for external labeled data collection. Figure 1 presents the overall framework.

3.3 SL-HyDE Training

Self-Learning Generator. An unlabeled medical corpus, such as Huatuo26M (Wang et al., 2025), serves as the primary resource for domain-specific content. To construct queries, we employ a robust offline LLM, Qwen2.5-32B-Instruct (Team, 2024), leveraging in-context learning (Brown, 2020). With a well-designed prompt, the model effectively generates medically grounded, context-aware queries:

$$q = \mathrm{LLM}(d, \mathrm{prompt}). \tag{2}$$

To facilitate retrieval, the raw generator $\mathcal{M}_g$ produces a hypothetical document that encapsulates the relevant information from the true target document. Concretely, we provide both the query and the corresponding target document as input to the generator, along with a carefully designed prompt to guide the pseudo-document generation:

$$d' = \mathcal{M}_g(q, d, \mathrm{prompt}). \tag{3}$$

Notably, we avoid using the true target document as the output label, as the generator's primary role is to craft a hypothetical document that assists the retriever in locating it. Expecting the generator to replicate the exact target document itself would be impractical and overly demanding.

Given that not all hypothetical documents generated by the generator are equally effective for retrieval, we leverage the retriever $\mathcal{M}_r$ to select the most optimal one. Specifically, the generator $\mathcal{M}_g$ creates L hypothetical documents for a given query. Each hypothetical document d'_i is used to retrieve documents from the corpus, and we record the rank position r_i of the true target document d. The pseudo-document with the highest retrieval quality (the lowest r_i) is selected.

This process yields a collection of question-answer pairs in the form of (q, d^*), where the query q functions as the question and the optimal hypothetical document d^* as the corresponding answer. The generator is subsequently trained via supervised fine-tuning on the resulting dataset $D_{llm} = \{(q, d^*) | q \in Q\}$. The standard supervised fine-tuning (SFT) loss is computed as:

$$\mathcal{L}_{\mathrm{slg}} = -\sum\nolimits_{q \in Q} \sum\nolimits_{t} \log \mathcal{M}_g(d'_t | d'_{<t}, q). \tag{4}$$

Importantly, the self-learning generator is trained entirely without relying on supervision signals from labeled medical data. Instead, it leverages unlabeled corpora and combines the generator's text generation with the retriever's ranking function to construct high-quality, domain-tailored question-answer pairs for hypothetical document generation.

Self-Learning Retriever. Given a passage d from the corpus D and its corresponding query q, the pair (q, d) naturally constitutes the labeled query-document data required for retriever fine-tuning. However, since SL-HyDE retrieves the target document by encoding both the query and a generated hypothetical document during inference, we adopt a triplet $(q, d'; d)$ as the labeled data for retriever training. This approach effectively aligns the training data format with the inference stage, thereby

enhancing consistency and bridging the gap between training and deployment.

To achieve this, we utilize the fine-tuned generator $\mathcal{M}_g^t$ from the previous stage to generate hypothetical documents for all queries, constructing a labeled fine-tuning dataset $D_{emb} = \{(q, d'; d)|q \in Q\}$. Following previous research (Li et al., 2023; Xiao et al., 2024), we further increase the complexity of the training data through hard negative mining. Specifically, a retriever is used to identify challenging negative samples from the original corpus D via an ANN-based sampling strategy (Xiong et al., 2020), resulting in a hard negative dataset:

$$D^- = \text{ANN}(\mathcal{M}_r(q, d'), \mathcal{M}_r(D)). \quad (5)$$

In addition to the negatives mined from the corpus, we also incorporate in-batch negatives. Contrastive learning loss is then applied for the supervised fine-tuning of the retriever, with the objective function formulated as:

$$\mathcal{L}_{\text{slr}} = \min. \sum_{(q,d)} -\log \frac{e^{s(q,d)/\tau}}{e^{s(q,d)/\tau} + \sum_{B \cup D^-} e^{s(q,d^-)/\tau}}, \quad (6)$$

where τ is the temperature coefficient, and B represents the negative samples in a batch. The score $s(q, d)$ incorporates the generated document, as described in Equation 1.

At this stage, we can obtain a retriever endowed with medical domain knowledge, coherently adapted to the characteristics of retrieval queries by leveraging hypothetical documents. In SL-HyDE, the generator and retriever are trained separately in a sequential manner, allowing each component to be optimized with the most appropriate supervision signals available at its respective training phase.

3.4 SL-HyDE Inference

As illustrated in Figure 1, the inference stage of SL-HyDE introduces a hypothesis generation step prior to standard retrieval. Specifically, the input query q is first rewritten by a fine-tuned generator $\mathcal{M}_g^t$ to produce a pseudo-document d', as defined by the following equation:

$$d' = \mathcal{M}_g^t(q, \text{prompt}). \quad (7)$$

The prompt is a carefully designed instruction tailored to the requirements of each task. Detailed formulations of the prompts used in our experiments are provided in Appendix A.2.

To better integrate the hypothetical documents, we sample N documents from the hypothetical documents. Following (Gao et al., 2023), a fine-tuned retriever $\mathcal{M}_r^t$ encodes these documents into an embedding vector v_q:

$$v_q = \frac{1}{N+1}[\sum_{k=1}^{N} \mathcal{M}_r^t(d'_k) + \mathcal{M}_r^t(q)]. \quad (8)$$

Subsequently, the inner product is computed between v_q and all document vectors:

$$s(q, d) =< v_q, \mathcal{M}_r^t(d) >, \forall d \in D. \quad (9)$$

This aggregated vector representation identifies a neighborhood in the corpus embedding space, from which semantically similar real documents are retrieved based on vector similarity.

3.5 SL-HyDE vs. HyDE

Our approach, SL-HyDE, builds upon HyDE (Gao et al., 2023) with several key enhancements while retaining some similarities. First, both approaches follow the same inference pipeline: a large model generates a hypothetical document based on the query, which the retriever then uses to identify relevant documents. Second, neither SL-HyDE nor HyDE requires labeled data, enabling rapid deployment. This makes HyDE particularly useful in real-world settings, where effective retrieval can be achieved simply by choosing a generator and a retriever. However, for domain-specific tasks such as medical information retrieval, directly deploying HyDE often yields suboptimal results. One option is to fine-tune the generator and retriever separately with labeled medical data, but this approach faces the dual challenges of scarce labeled data and the risk of suboptimal adaptation when models are trained independently.

SL-HyDE addresses these limitations by introducing a self-learning mechanism that transforms HyDE into a trainable end-to-end framework. This mechanism enables the generator and retriever to adapt jointly to the medical domain. Supervision signals for the generator are derived from the retriever, and vice versa, enabling mutual reinforcement. This integrated training strategy substantially improves retrieval performance. In summary, SL-HyDE provides an efficient and practical solution for enhancing HyDE in medical domains, particularly when working with unlabeled corpora.

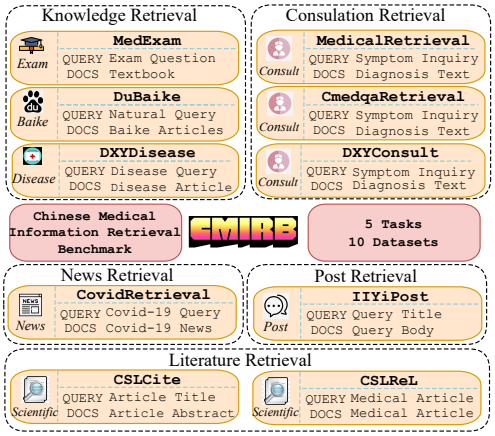

Figure 2: An overview of CMIRB.

Task	Dataset	#Samples		Avg. Word Lengths	
		#Query	#Document	Query	Document
Medical Knowledge Retrieval	MedExam	697	27,871	96.9	493.7
	DuBaike	318	56,441	7.6	403.3
	DXYDisease	1,255	54,021	24.3	191.1
Medical Consultation Retrieval	MedicalRet.	1,000	100,999	17.9	122.0
	CmedqaRet.	3,999	100,001	48.4	307.7
	DXYConsult	943	12,577	170.4	370.1
News Ret.	CovidRet.	949	100,001	25.9	332.4
Post Ret.	IIYiPost	789	27,570	15.9	150.1
Literature Retrieval	CSLCite	573	36,703	21.9	269.6
	CSLRel	439	36,758	281.8	292.2

Table 1: Statistics of datasets in CMIRB.

4 CMIRB Benchmark

4.1 Overview

The CMIRB benchmark is a specialized, multi-task dataset designed for Chinese medical information retrieval. As shown in Figure 2, it comprises five different tasks. *Medical knowledge retrieval*: Retrieve relevant medical knowledge snippets from textbooks or encyclopedias based on a given medical query. *Medical consultation retrieval*: Extract relevant doctor responses to online medical consultation questions posed by patients. *Medical news retrieval*: Focus on retrieving news articles that address queries related to COVID-19. *Medical post retrieval*: Retrieve the content of a forum post corresponding to its title. *Medical literature retrieval*: Retrieve abstracts of cited references based on a medical title or identify a similar paper based on the given medical paper.

4.2 Data Construction

The CMIRB benchmark integrates 10 datasets, including several existing resources: **MedicalRetrieval** (Long et al., 2022), **CmedqaRetrieval** (Qiu et al., 2022), and **CovidRetrieval** (Qiu et al., 2022), covering patient-doctor consultations and COVID-19-related news retrieval.

In addition, we construct several datasets by combining existing query resources with curated medical corpora. **MedExam** pairs questions with textbook passages from MedQA (Jin et al., 2021). **DuBaike** uses queries from DuReader (He et al., 2017) and documents collected from Baidu Baike pages[1]. We also curate two datasets from the medical website DingXiangYuan[2]. **DXYDisease** focuses on structured disease-related Q&A, while **DXYConsult** captures richer patient-doctor dialogues that include symptom descriptions, medication history, and diagnostic queries. **IIYiPost** is curated by crawling posts from the IIYi forum[3].

Finally, **CSLCite** and **CSLRel** are constructed based on the CSL dataset (Li et al., 2022), targeting different literature retrieval scenarios. **CSLCite** uses journal titles as queries and their cited references from WanFangMedical[4] as documents, while **CSLRel** pairs each paper with the most relevant similar paper recommended by the platform.

To ensure dataset quality, we apply ChatGPT to filter out non-medical content and low-quality query-document pairs. Additional query-document matching is performed for MedExam and DuBaike to ensure content relevance. Full details are provided in the Appendix B.1. Table 1 summarizes dataset statistics, revealing broad variability in query and document length, ranging from short titles to long passages, thereby ensuring the benchmark's diversity and practical relevance.

5 Experiments

5.1 Experimental Setup

Implementation Details. We sample 10,000 documents from the Huatuo26M_encyclopedia dataset as the unlabeled corpus. In our training framework, we utilize Qwen2-7B-Instruct (Yang et al., 2024) as the generator and BGE-Large-zh-v1.5 (Xiao et al., 2024) as the retriever. Unless otherwise specified, all experiments are conducted under this Qwen+BGE configuration. Model training and evaluation are conducted on up to 5 NVIDIA A100 GPUs, each equipped with 40GB of memory. For fine-tuning the LLM, we employ the AdamW op-

[1] https://baike.baidu.com/

[2] https://dxy.com/

[3] https://bbs.iiyi.com/

[4] https://med.wanfangdata.com.cn/

Task	Knowledge Retrieval			Consulation Retrieval			News	Post	Literature Retrieval		Average
Dataset	MedExam	DuBaike	DXYDis.	Medical	Cmedqa	DXYCon.	Covid	IIYiPost	CSLCite	CSLRel	
Text2Vec(large)	41.39	21.13	41.52	30.93	15.53	21.92	60.48	29.47	20.21	23.01	30.56
mContriever	51.50	22.25	44.34	38.50	22.71	20.04	56.01	28.11	34.59	33.95	35.20
BM25	31.95	17.89	40.12	29.33	6.83	17.78	78.90	66.95	33.74	29.97	35.35
OpenAI-Ada-002	53.48	43.12	58.72	37.92	22.36	27.69	57.21	48.60	32.97	43.40	42.55
M3E(large)	33.29	46.48	62.57	48.66	30.73	41.05	61.33	45.03	35.79	47.54	45.25
mE5(large)	53.96	53.27	72.10	51.47	28.67	41.35	75.54	63.86	42.65	37.94	52.08
piccolo(large)	43.11	45.91	70.69	59.04	41.99	47.35	85.04	65.89	44.31	44.21	54.75
GTE(large)	41.22	42.66	70.59	62.88	43.15	46.30	88.41	63.02	46.40	49.32	55.40
BGE(large)	58.61	44.26	71.71	59.60	42.57	47.73	73.33	67.13	43.27	45.79	55.40
PEG(large)	52.78	51.68	77.38	60.96	44.42	49.30	82.56	70.38	44.74	40.38	57.46
BGE(large)	58.61	44.26	71.71	59.60	42.57	47.73	73.33	67.13	43.27	45.79	55.40
HyDE	64.39	52.73	73.98	57.27	38.52	47.11	74.32	73.07	46.16	38.68	56.62
SL-HyDE	71.49*	60.96*	75.34*	58.58*	39.07*	50.13*	76.95*	73.81*	46.78*	40.71*	59.38*
Improve.	↑ **11.03%**	↑ **15.61%**	↑ **1.84%**	↑ **2.29%**	↑ **1.43%**	↑ **6.41%**	↑ **3.54%**	↑ **1.01%**	↑ **1.34%**	↑ **5.25%**	↑ **4.87%**

Table 2: Performance of various Retrieval models on nDCG@10. The first part shows ten base retrieval models, and the second shows retrieval models enhanced by hypothetical documents. * denotes the result outperforms baseline model (HyDE) in t-test at $p < 0.05$ level.

timizer (Loshchilov, 2017) in conjunction with a cosine learning-rate scheduler. Training is conducted for 1 epoch with a learning rate of 1e-5 and a batch size of 2. We set 200 warmup steps and configure the LoRA rank to 8. Retriever fine-tuning also uses the AdamW optimizer with a linear decay schedule and an initial learning rate of 1e-5. The batch size per GPU is set at 4, and the maximum input sequence length is limited to 512. We apply a temperature of 0.02 and mine 7 hard negatives for each query to increase training difficulty.

Evaluation Settings. For simplicity, we employ the LLM to generate a single hypothetical document for each query. The retrieval model embeds all queries, hypothetical documents, and corpus documents, with similarity scores calculated using cosine similarity. Documents in the corpus are ranked for each query based on these scores, and nDCG@10 is adopted as the primary evaluation metric to assess retrieval effectiveness. We set the LLM temperature to 0.7 and repeat each experiment five times with different random seeds.

Baseline Models. To comprehensively evaluate CMIRB, we select several widely used retrieval models. These include lexical retriever BM25 (Robertson et al., 2009); dense retrieval models such as Text2Vec-Large-Chinese (Xu, 2023), PEG (Wu et al., 2023), BGE-Large-zh-v1.5 (Xiao et al., 2024), GTE-Large-zh (Li et al., 2023), and Piccolo-Large-zh (SenseTime, 2023); multilingual retrievers like mContriever (mas-marco) (Izacard et al., 2021), M3E-Large (Wang et al., 2023b), mE5 (multilingual-e5-large) (Wang et al., 2024a); and text-embedding-ada-002 (OpenAI). For more details about baselines, please refer to Appendix A.1.

5.2 Main Results

The experimental results of various retrieval models, including SL-HyDE, on the CMIRB benchmark are presented in Table 2. We highlight the following key observations.

(1) BM25 remains highly competitive in specific medical tasks. As a lexical retriever, it ranks documents based on TF-IDF matching scores between queries and documents. Although it underperforms on the overall CMIRB benchmark, it achieves strong results in tasks like medical news retrieval (78.9 vs. 73.33 for BGE) and medical post retrieval (66.95 vs. 67.13 for BGE). This advantage can be attributed to the higher keyword overlap in these datasets.

(2) No single retrieval model achieves optimal performance across all ten tasks. PEG and GTE each deliver the best performance on four datasets, while BGE and mE5 lead on one dataset each. Dense models with stronger performance typically employ contrastive learning, leveraging large-scale pretraining on unlabeled data followed by fine-tuning on labeled datasets. Differences in training data distribution influence model effectiveness across different datasets, underscoring the need for specialized approaches.

(3) SL-HyDE consistently outperformed HyDE across all ten datasets. While HyDE provides modest overall improvements over BGE, it excels in medical knowledge retrieval but lags in medical consultation tasks. This gap may stem from LLMs being more adept at handling encyclopedia-type knowledge than the nuanced reasoning required for patient-doctor dialogues. In contrast, SL-HyDE achieves consistent improvements over HyDE due

Task	Know.	Consu.	News	Post	Literature	Avg.(All)
ChatGLM3 as Generator + BGE as Retriever						
HyDE	62.43	46.43	73.89	70.88	44.46	56.02
SL-HyDE	66.26	48.55	76.78	72.29	46.40	58.63
Improve.	↑ **6.14%**	↑ **4.57%**	↑ **3.91%**	↑ **1.99%**	↑ **4.36%**	↑ **4.65%**
Llama2 as Generator + BGE as Retriever						
HyDE	55.74	40.62	72.90	72.22	45.30	52.48
SL-HyDE	63.66	45.44	77.17	71.99	45.75	56.80
Improve.	↑ **14.21%**	↑ **11.87%**	↑ **5.86%**	↓ **0.32%**	↑ **0.99%**	↑ **8.23%**

Table 3: Performance of different generators.

Task	Know.	Consu.	News	Post	Literature	Avg.(All)
Qwen2 as Generator + mE5 as Retriever						
HyDE	65.77	43.15	75.92	68.15	38.58	54.80
SL-HyDE	68.60	44.83	77.59	66.81	42.33	56.94
Improve.	↑ **4.31%**	↑ **3.90%**	↑ **2.20%**	↓ **1.97%**	↑ **9.72%**	↑ **3.90%**
Qwen2 as Generator + PEG as Retriever						
HyDE	66.03	49.73	80.49	72.51	38.87	57.80
SL-HyDE	69.96	50.97	80.89	75.93	45.03	60.97
Improve.	↑ **5.96%**	↑ **2.50%**	↑ **0.50%**	↑ **4.72%**	↑ **15.86%**	↑ **5.48%**

Table 4: Performance of different retrievers.

Task	**Know.**	**Consu.**	**News**	**Post**	**Literature**	**Avg.(All)**
SL-HyDE	69.26	49.26	76.95	73.81	43.75	59.38
w/ D.	68.00	41.86	71.94	68.02	37.36	54.43
w/ con.	69.04	45.51	73.38	69.53	44.81	57.62
w/ K-D.	69.30	50.17	77.38	74.55	45.42	60.12

Table 5: Performance of different fusing strategies.

to its self-learning mechanism, which not only enhances medical knowledge integration within both the generator and the retriever but also better aligns the outputs of the two components.

5.3 Performance Analysis

Effect of Different Generators. In Table 3, we present SL-HyDE's performance when using alternative fine-tuned LLMs as the generator, including ChatGLM3-6B (Team et al., 2024) and Llama2-7b-Chat (Touvron et al., 2023).

Both models yield improvements under SL-HyDE compared to HyDE. For instance, we observe a 4.65% improvement with ChatGLM3 and an 8.23% improvement with Llama2. However, for Llama2, HyDE performs slightly worse than BGE. This issue likely stems from the pseudo-documents generated by the English-based Llama2 containing English text, which the downstream Chinese BGE retriever struggled to encode effectively. After fine-tuning, SL-HyDE improves by approximately 8%, benefiting from both the reduction of English content and the retriever's enhanced ability to encode medical knowledge, demonstrating SL-HyDE's adaptability across different generator architectures.

Effect of Different Retrievers. We further investigate SL-HyDE's generalizability by fine-tuning two additional retrievers: PEG, the strongest baseline on CMIRB, and a multilingual retriever mE5.

As shown in Table 4, HyDE provides moderate gains compared to using the retriever alone. However, the application of SL-HyDE yields substantially larger improvements across both models. For instance, PEG, which achieves the best baseline performance on CMIRB, improves from 57.46% to 60.97%, marking a notable increase in retrieval effectiveness. These results highlight SL-HyDE's robustness in enhancing retrieval performance across various retriever models.

Effect of Different Fusing Strategies. We also evaluate multiple strategies for incorporating hypothetical documents into retrieval. SL-HyDE encodes the original query and the hypothetical documents separately, then aggregates them via mean pooling to obtain the final query vector. SL-HyDE w/ D uses only the hypothetical document as the query. SL-HyDE w/ con concatenates the original query and the hypothetical document into a single string before encoding. SL-HyDE w/ K-D generates multiple (five) hypothetical documents for each query before retrieval.

Table 5 demonstrates that combining the original query and hypothetical documents is the most effective approach. Relying solely on hypothetical documents significantly degrades performance, particularly in medical consultation tasks where the original query contains indispensable clinical details. Concatenation introduces additional noise at the string level, leading to performance drops, whereas mean pooling mitigates this issue by integrating information at the representation level. Generating multiple hypothetical documents further improves coverage and boosts performance across tasks, but incurs a proportional increase in inference cost. Thus, an efficiency–accuracy trade-off must be considered when selecting the number of generated documents.

5.4 Ablation Study

To further analyze the gains brought by the internal architecture of SL-HyDE, we conduct two sets of ablation experiments: (1) SL-HyDE w/o BGE-FT, which employs the fine-tuned LLM as the generator while keeping the retriever unfine-tuned; (2) SL-HyDE w/o Qwen-FT, which uses the raw LLM as the generator alongside the fine-tuned retriever.

As shown in Table 6, fine-tuning both components further improves performance across most datasets, validating the efficacy of the self-learning

Task	Know.	Consu.	News	Post	Literature	Avg.(All)
HyDE	63.70	47.63	74.32	73.07	42.42	56.62
SL-HyDE	**69.26**	**49.26**	**76.95**	73.81	**43.75**	**59.38**
w/o BGE-FT	64.32	47.95	74.87	72.91	43.24	57.11
w/o Qwen-FT	68.75	48.85	76.63	**74.52**	43.11	58.77

Table 6: Performance of different variants.

Query: How to treat a hernia?
Target Doc: Inguinal Hernia Treatment Plan. For conventional treatment, a 1-year-old infant can use a hernia belt for compression. As the muscles gradually strengthen, there may be a possibility of spontaneous recovery. For elderly and frail individuals a hernia belt can be worn, but for other patients, surgery is generally recommended...
HyDE: Hernia is a common disease caused by a weak area in the abdominal wall, Treatment usually includes conservative and surgical methods. For most patients, especially young and healthy individuals, surgery is the preferred option... (Rank: 10)
SL-HyDE: Hernia is a common condition that typically occurs For infants,... the use of a hernia belt to apply localized pressure can help alleviate symptoms and promote the development of the abdominal muscles,... For elderly or frail patients, or those with severe underlying conditions,... wearing a hernia belt can help manage symptoms and reduce the risk of the hernia progressing further... (Rank: 2)

Table 7: The case study comparing with HyDE.

framework. Notably, retriever fine-tuning provides larger gains, indicating that BGE particularly benefits from domain-specific adaptation. Nevertheless, jointly fine-tuning both the retriever and the generator leads to the most robust improvements, demonstrating the synergistic effect of SL-HyDE's co-adaptation mechanism.

5.5 Case Study

To provide an intuitive illustration of SL-HyDE's impact, Table 7 compares hypothetical documents generated by HyDE and SL-HyDE for the query *How to treat a hernia?*. HyDE produces a general document discussing *conservative and surgical treatments*, but it lacks specificity for different patient groups. In contrast, SL-HyDE generates a more tailored document mentioning *hernia belts for infants and elderly patients*, aligning closely with the target document's content. This refinement results in a significantly higher retrieval ranking ($2nd$ vs. $10th$), clearly demonstrating how more precise hypothetical documents can enhance retrieval effectiveness.

5.6 Cross-Domain Generalization

In this subsection, we further evaluate SL-HyDE in the legal domain to demonstrate its adaptability beyond medicine. In the legal field, labeled datasets are also scarce due to the high cost of annotation and the complexity of domain expertise. To showcase the generality of our approach, we sample 10k unlabeled law texts from pile-of-law (Henderson et al., 2022) to construct a domain-specific corpus, and build the SL-HyDE system using Llama-2-7b-chat-hf as the generator and BGE-Large-en-V1.5 as the retriever. We then evaluate the system on three information retrieval datasets from MTEB in the law domain. As shown in Table 8, while vanilla HyDE brings only a slight improvement over BGE, the fine-tuned SL-HyDE (77.25%) significantly outperforms HyDE (75.52%). These results highlight the strong cross-domain generalization ability of SL-HyDE and its potential to serve as a versatile solution for low-resource domains.

Dataset	legal_ summar.	legalbench_ contracts_qa	legalbench_ lobbying	Average
BGE	59.99	73.52	91.51	75.01
HyDE	58.95	74.82	92.78	75.52
SL-HyDE	**63.50**	**75.10**	**93.15**	**77.25**

Table 8: Performance of SL-HyDE in legal domain.

6 Conclusions

In this paper, we propose SL-HyDE, an automated framework for zero-shot medical information retrieval that operates without reliance on labeled relevance data. Leveraging an unlabeled medical corpus, SL-HyDE employs a self-learning training paradigm where the retriever guides the generator's training, and the generator in turn produces pseudo-documents that enhance retriever training. This process injects domain-specific medical knowledge into both components, yielding hypothetical documents that are highly effective in guiding retrieval. Furthermore, we introduce CMIRB, a comprehensive benchmark for Chinese medical information retrieval, encompassing five tasks and ten datasets. Extensive experiments demonstrate that SL-HyDE consistently outperforms HyDE across all datasets. Additionally, SL-HyDE shows strong adaptability and scalability, effectively enhancing retrieval performance across various combinations of generators and retrievers. In future work, we plan to extend SL-HyDE to other data-scarce domains to further evaluate its generalizability across different settings. In addition, we will explore reinforcement learning techniques to further enhance retriever capabilities and improve reasoning in complex medical retrieval scenarios.

Limitations

While our work effectively addresses the adaptation challenges of HyDE in low-resource scenarios, several limitations remain. First, our study primarily focuses on the medical domain and only provides an initial exploration of the legal domain (see Appendix A.3), without extending the evaluation to other vertical fields such as economics or education. Second, although we experiment with three open-source LLMs, Qwen2, LlaMA2, and ChatGLM3, as generators, we do not include more recent or diverse model families such as Qwen3 or Gemini, which may exhibit different generative behaviors and impact performance differently. Third, our data construction pipeline relies on LLMs for query-document matching and pseudo-relevant pair filtering. The effectiveness of these components depends on the model's instruction-following ability and sensitivity to domain-specific nuances, which may introduce hallucinations or spurious correlations and potentially affect reliability.

Ethical Considerations

We clarify that CMIRB is constructed from publicly accessible medical platforms and sources. For datasets such as MedExam, we utilize existing IR community benchmarks released under open licenses (e.g., MIT, CC-BY license). Data from Wanfang explicitly states that it may be used for learning and scientific research. Data from IIYi and DXY, consistent with prior research (Yim et al., 2024; Jia et al., 2025), can be used for research and educational purposes. We adhere strictly to data collection protocols to ensure compliance with copyright and privacy requirements, including the removal of all personal identifiers. Nevertheless, we acknowledge that the dataset may contain medically sensitive or potentially distressing content. We emphasize that CMIRB is released solely for academic research and evaluation purposes, and it is not intended for clinical practice or real-world medical decision-making under any circumstances.

Acknowledgements

This work was supported by the Public Computing Cloud at Renmin University of China and the Fund for Building World-Class Universities (Disciplines) at Renmin University of China.

References

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *European Conference on Information Retrieval*, pages 716–722. Springer.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Lorraine Goeuriot, Gareth JF Jones, Liadh Kelly, Henning Müller, and Justin Zobel. 2016. Medical information retrieval: introduction to the special issue. *Information Retrieval Journal*, 19:1–5.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, and 1 others. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073*.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

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*, pages 9278–9298.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Jinlin Li and Xiao Zhou. 2025. Curegraph: Contrastive multi-modal graph representation learning for urban living circle health profiling and prediction. *Artificial Intelligence*, 340:104278.

Lei Li, Jianxun Lian, Xiao Zhou, and Xing Xie. 2024. Ada-retrieval: An adaptive multi-round retrieval paradigm for sequential recommendations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8670–8678.

Yudong Li, Yuqing Zhang, Zhe Zhao, Linlin Shen, Weijie Liu, Weiquan Mao, and Hui Zhang. 2022. Csl: A large-scale chinese scientific literature dataset. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3917–3923.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2356–2362.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114.

Dingkun Long, Qiong Gao, Kuan Zou, Guangwei Xu, Pengjun Xie, Ruijie Guo, Jian Xu, Guanjun Jiang, Luxi Xing, and Ping 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*, pages 3046–3056.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Gang Luo, Chunqiang Tang, Hao Yang, and Xing Wei. 2008. Medsearch: a specialized search engine for medical information retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 143–152.

Yijun Ma, Chaozhuo Li, and Xiao Zhou. 2024. Tailsteak: Improve friend recommendation for tail users via self-training enhanced knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 8895–8903.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100, Online. Association for Computational Linguistics.

Jessie McGowan, Roland Grad, Pierre Pluye, Karin Hannes, Katherine Deane, Michel Labrecque, Vivian Welch, and Peter Tugwell. 2009. Electronic retrieval of health information by healthcare providers to improve practice and patient care. *Cochrane Database of Systematic Reviews*, (3).

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

OpenAI. 2022. New and improved embedding model.

Yifu Qiu, Hongyu Li, Yingqi Qu, Ying Chen, Qiaoqiao She, Jing Liu, Hua Wu, and Haifeng Wang. 2022. Dureader-retrieval: A large-scale chinese benchmark for passage retrieval from web search engine. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5326–5338.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

SenseTime. 2023. Text2vec: Text to vector toolkit. `https://github.com/timczm/piccolo-large-zh`.

Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Tianyi Zhou, and Daxin Jiang. 2023. Large language models are strong zero-shot retriever. *arXiv preprint arXiv:2304.14233*.

Sonish Sivarajkumar, Haneef Ahamed Mohammad, David Oniani, Kirk Roberts, William Hersh, Hongfang Liu, Daqing He, Shyam Visweswaran, and Yanshan Wang. 2024. Clinical information retrieval: A literature review. *Journal of Healthcare Informatics Research*, pages 1–40.

GLM Team, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv e-prints*, pages arXiv–2406.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423, Singapore. Association for Computational Linguistics.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024b. Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*.

Xidong Wang, Jianquan Li, Shunian Chen, Yuxuan Zhu, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Junying Chen, Jie Fu, Xiang Wan, and 1 others. 2025. Huatuo-26m, a large-scale chinese medical qa dataset. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3828–3848.

Yuxin Wang, Qingxuan Sun, and sicheng He. 2023b. M3e: Moka massive mixed embedding model.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045.

Tong Wu, Yulei Qin, Enwei Zhang, Zihan Xu, Yuting Gao, Ke Li, and Xing Sun. 2023. Towards robust text retrieval with progressive learning. *arXiv preprint arXiv:2311.11691*.

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.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Ming Xu. 2023. Text2vec: Text to vector toolkit. https://github.com/shibing624/text2vec.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May Dongmei Wang, Joyce Ho, Chao Zhang, and Carl Yang. 2024. Bmretriever: Tuning large language models as better biomedical text retrievers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22234–22254.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Fei Xia, Meliha Yetisgen-Yildiz, and Martin Krallinger. 2024. Overview of the mediqa-m3g 2024 shared task on multilingual multimodal medical answer generation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 581–589.

Xixian Yong, Jianxun Lian, Xiaoyuan Yi, Xiao Zhou, and Xing Xie. 2025. Motivebench: How far are we from human-like motivational reasoning in large language models? *arXiv preprint arXiv:2506.13065*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, and 1 others. 2023. Huatuogpt, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885.

Jiaping Zheng and Hong Yu. 2015. Key concept identification for medical information retrieval. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 579–584.

Xiao Zhou, Zhongxiang Zhao, and Hanze Guo. 2025. Tricolore: Multi-behavior user profiling for enhanced candidate generation in recommender systems. *IEEE Transactions on Knowledge & Data Engineering*, (01):1–12.

A Models

A.1 Baselines

To comprehensively evaluate the performance of existing retrievers on CMIRB, we selected ten representative models that have demonstrated strong results on the MTEB leaderboard[5]. For details regarding the retrievers and large reasoning models evaluated in this paper, please refer to Table 9.

BM25 (Robertson et al., 2009). BM25 is a widely used baseline retriever that relies on bag-of-words and TF-IDF to perform lexical retrieval. In this paper, BM25 is implemented with Pyserini (Lin et al., 2021) with default hyperparameters to index snippets from all corpora.

Text2Vec (Xu, 2023). It is a cosine sentence model based on a linguistically-motivated pre-trained language model (LERT).

PEG (Wu et al., 2023). Proposed by Wu et al., (Wu et al., 2023), PEG is trained over 100 million data points, spanning a broad range of domains and covering multiple downstream tasks.

BGE (Xiao et al., 2024). BGE adopts a compound training recipe that integrates pre-training, contrastive learning with advanced negative sampling, and instruction-based fine-tuning to build general-purpose text embeddings.

GTE (Li et al., 2023). GTE introduces a multi-stage contrastive learning framework for training text embedding models that can be applied to various retrieval and similarity tasks.

Piccolo (SenseTime, 2023). Piccolo is a general-purpose Chinese embedding model trained via a two-stage process that combines weakly supervised learning with manually labeled text pairs.

Contriever (Izacard et al., 2021). It is a multilingual dense retriever with contrastive learning, which fine-tunes the pre-trained mContriever model on MS MARCO dataset.

M3E (Wang et al., 2023b). M3E (Moka Massive Mixed Embedding) is a bilingual text embedding model trained on over 22 million Chinese sentence pairs, supporting tasks like cross-lingual text similarity and retrieval.

mE5 (Wang et al., 2024a). mE5 is a multilingual E5 text embedding model trained with a multi-stage pipeline, including contrastive pre-training on one billion multilingual text pairs and fine-tuning on labeled datasets.

OpenAI-Ada-002 (OpenAI). A highly efficient text embedding model that converts natural language into dense vectors for a wide range of applications, including semantic search, clustering, and similarity tasks.

For the generator, we selected three highly powerful large language models.

Qwen2 (Yang et al., 2024). Qwen2 is a comprehensive suite of foundational and instruction-tuned language models, encompassing a parameter range from 0.5 to 72 billion, featuring dense models and a Mixture-of-Experts model.

ChatGLM3 (Team et al., 2024). ChatGLM3-6B is a next-generation conversational pre-trained model that demonstrates strong performance in semantics, reasoning, and code execution.

Llama2 (Touvron et al., 2023). Llama2 is an auto-regressive language model with an optimized transformer architecture. The fine-tuned versions leverage supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to better align with human preferences for helpfulness and safety.

Q2P Prompt
Please generate a medical content paragraph to answer this question. Question: QUESTION Paragraph:
T2P Prompt
Please generate a medical content paragraph based on this title. Title: TITLE Paragraph:
P2P Prompt
Please generate a similar medical paragraph for the following text. Text: TEXT Similar Paragraph:

Table 10: Evaluation prompts for generators.

A.2 Evaluation Settings

We use the C-MTEB[6] framework to evaluate the performance of various retrieval models on CMIRB. To ensure stability, we set the temperature of LLM to 0.7 and repeat each experiment five times with different random seeds. For each dataset, the prompts used to generate pseudo-documents are shown in Figure 10. The IIYIPost and CSLCite datasets utilize the Title-to-Paragraph (T2P) template to prompt LLMs to generate documents from titles. The CSLRel dataset adopts the Paragraph-

[5] https://huggingface.co/spaces/mteb/leaderboard

[6] C-MTEB

Model	Size	Model Link
Retrieval Models		
BM25 (Robertson et al., 2009)	N/A	https://github.com/castorini/pyserini
Text2Vec (Xu, 2023)	325M	https://huggingface.co/GanymedeNil/text2vec-large-chinese
PEG (Wu et al., 2023)	335M	https://huggingface.co/TownsWu/PEG
BGE (Xiao et al., 2024)	335M	https://huggingface.co/BAAI/bge-large-zh-v1.5
GTE (Li et al., 2023)	335M	https://huggingface.co/thenlper/gte-large-zh
Piccolo (SenseTime, 2023)	335M	https://huggingface.co/sensenova/piccolo-large-zh
Contriever (Izacard et al., 2021)	109M	https://huggingface.co/facebook/mcontriever-msmarco
M3E (Wang et al., 2023b)	340M	https://huggingface.co/moka-ai/m3e-large
mE5 (Wang et al., 2024a)	560M	https://huggingface.co/intfloat/multilingual-e5-large
OpenAI-Ada-002 (OpenAI)	N/A	https://openai.com/index/new-and-improved-embedding-model/
Large Language Models		
Qwen2 (Yang et al., 2024)	7B	https://huggingface.co/Qwen/Qwen2-7B-Instruct
Llama2 (Touvron et al., 2023)	7B	https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
ChatGLM3 (Team et al., 2024)	7B	https://huggingface.co/THUDM/chatglm3-6b

Table 9: Detailed information on all of the retrieval models and large language models in our paper.

Similar Example
Query: What causes seborrheic alopecia?
Target Doc: Seborrheic alopecia (androgenetic alopecia) is linked to genetics, androgens, excess scalp oil...
HyDE: Seborrheic alopecia is a common hair loss type linked to genetics, DHT, scalp oil, immune factors, causing hair thinning and loss... (Rank: 6)
SL-HyDE: Seborrheic alopecia involves excess oil, follicle blockage, genetics, hormones, lifestyle, affecting hair growth... (Rank: 6)
Degraded Example
Query: What causes snoring?
Target Doc: Snoring occurs when airflow is blocked by relaxed throat muscles, obesity, or nasal issues. Factors include age, weight, alcohol, posture....
HyDE: Snoring is caused by airway structure issues, obesity, alcohol, posture, genetics, and may require CPAP or lifestyle changes... (Rank: 2)
SL-HyDE: Snoring can result from fatigue, stress, poor sleep habits, with suggestions on healthy routines, lacking key medical causes... (Rank: 8)

Table 11: More illustrative examples of hypothetical documents.

to-Paragraph (P2P) template to generate semantically similar text. For the remaining datasets, the Question-to-Paragraph (Q2P) template is employed to generate answers to medical questions.

A.3 More Experiment Results

Table 12 presents the Recall@100 performance of 10 retrieval models on CMIRB. In Table 13, we provide a detailed breakdown of performance for various combinations of generators and retrievers across the 10 datasets.

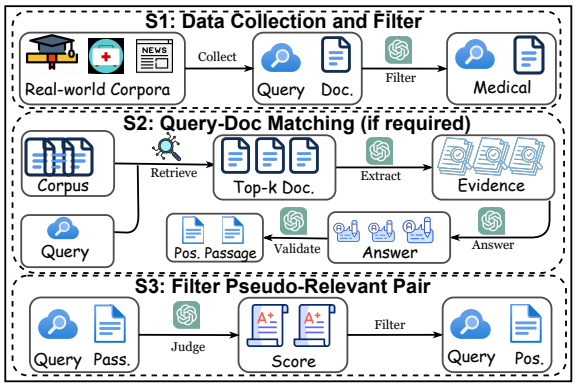

Figure 3: CMIRB benchmark construction pipeline.

Additionally, in Table 11, we present two illustrative examples of generated hypothetical documents. In one case, the HyDE-generated pseudo-document retrieved the target document without change in rank, while in the other, the retrieval performance slightly degraded.

B CMIRB Datasets

B.1 Data Process

We curated a substantial dataset from multiple medical resources, summarized in Table 14, which details both source distribution and data volume. Our data preprocessing pipeline, illustrated in Figuer 3 and Algorithm 1, leverages prompt templates shown in Figure 4 and Figure 5.

Initially, ChatGPT[7] is used to filter out non-

[7]https://openai.com/chatgpt

Task	**Knowledge Retrieval**			**Consulation Retrieval**			**News**	**Post**	**Literature Retrieval**		**Average**
Dataset	MedExam	DuBaike	DXYDis.	Medical	Cmedqa	DXYCon.	Covid	IIYiPost	CSLCite	CSLRel	
BM25	75.61	56.92	72.91	44.20	17.26	37.33	96.47	89.98	67.19	72.66	63.05
Text2Vec(large)	89.81	79.25	78.01	52.80	42.99	64.58	88.83	74.78	61.96	70.39	70.34
mContriever	93.40	86.48	84.06	61.50	53.40	62.67	84.93	70.72	72.25	84.97	75.44
mE5(large)	93.83	98.43	96.02	70.90	57.95	80.38	97.05	91.64	77.31	91.12	85.46
M3E(large)	86.08	98.43	93.55	74.00	70.61	86.96	93.26	88.97	76.09	**96.58**	86.45
GTE(large)	87.52	96.54	95.86	**87.00**	**84.95**	89.50	**99.47**	93.41	**83.25**	**96.58**	91.41
piccolo(large)	89.67	99.06	96.81	82.80	84.81	91.09	**99.47**	95.69	83.07	92.25	91.47
PEG(large)	95.41	**98.74**	**98.01**	83.70	84.64	89.50	98.74	**96.83**	81.15	92.25	**91.90**
BGE(large)	**97.42**	**98.74**	96.81	81.20	82.57	**91.30**	98.10	95.69	80.80	96.36	**91.90**

Table 12: Performance of various Retrieval models on CMIRB benchmark. All scores denote Recall@100. The best score on a given dataset is marked in bold.

Task	Knowledge Retrieval			Consulation Retrieval			News	Post	Literature Retrieval		Average
Dataset	MedExam	DuBaike	DXYDis.	Medical	Cmedqa	DXYCon.	Covid	IIYiPost	CSLCite	CSLRel	
ChatGLM3 as Generator + BGE as Retriever											
HyDE	61.96	54.25	71.07	56.32	37.73	45.23	73.89	70.88	45.11	43.80	56.02
SL-HyDE	67.12	59.40	72.25	57.16	38.77	49.71	76.78	72.29	45.81	46.98	58.63
Improve.	↑ **8.33%**	↑ **9.49%**	↑ **1.66%**	↑ **1.49%**	↑ **2.76%**	↑ **9.90%**	↑ **3.91%**	↑ **1.99%**	↑ **1.55%**	↑ **7.26%**	↑ **4.65%**
Llama2 as Generator + BGE as Retriever											
HyDE	53.10	45.78	68.34	53.51	31.29	37.07	72.90	72.22	44.19	46.41	52.48
SL-HyDE	64.88	56.30	69.81	54.68	36.93	44.72	77.17	71.99	44.62	46.88	56.80
Improve.	↑ **22.18%**	↑ **22.98%**	↑ **2.15%**	↑ **2.19%**	↑ **18.02%**	↑ **20.64%**	↑ **5.86%**	↓ **0.32%**	↑ **0.97%**	↑ **1.01%**	↑ **8.23%**
Qwen2 as Generator + mE5 as Retriever											
HyDE	65.18	56.35	75.77	54.31	32.02	43.12	75.92	68.15	45.66	31.50	54.80
SL-HyDE	71.36	59.50	74.95	54.68	33.95	45.87	77.59	66.81	45.65	39.01	56.94
Improve.	↑ **9.48%**	↑ **5.59%**	↓ **1.08%**	↑ **0.68%**	↑ **6.03%**	↑ **6.38%**	↑ **2.20%**	↓ **1.97%**	↓ **0.02%**	↑ **23.84%**	↑ **3.90%**
Qwen2 as Generator + PEG as Retriever											
HyDE	64.87	55.04	78.18	58.47	41.47	49.25	80.49	72.51	43.56	34.17	57.80
SL-HyDE	72.04	60.26	77.59	59.81	40.43	52.68	80.89	75.93	47.53	42.53	60.97
Improve.	↑ **11.05%**	↑ **9.48%**	↓ **0.75%**	↑ **2.29%**	↓ **2.51%**	↑ **6.96%**	↑ **0.50%**	↑ **4.72%**	↑ **9.11%**	↑ **24.47%**	↑ **5.48%**

Table 13: Performance of different combinations of generators and retrievers on CMIRB benchmark.

medical content (lines 3-8). Subsequently, ChatGPT assesses query-document relevance, removing low-relevance pairs (lines 27-33). The relevance evaluation considers both semantic alignment and practical utility for the target tasks, as detailed in Figure 5.

For the MedExam and DuBaike datasets, the direct query-document signal isn't initially provided. Both queries and documents in the MedExam dataset originate from Work (Jin et al., 2021), where 100 randomly selected questions have corpus documents containing evidence sufficient to answer them, verified manually by the authors. For DuBaike, queries from Baidu Search and Baidu Zhidao generally align with the content of Baidu Baike, enabling a query-matching algorithm to identify relevant documents.

To pinpoint the most relevant documents, we first retrieve the top 20 candidates for a given query using BM25. ChatGPT then ranks these candidates and selects the top 3 most relevant documents. These documents are expected to be semantically aligned with the query and provide adequate answers or supporting evidence. ChatGPT further extracts evidence segments from these documents to form the basis for query answering.

To verify the sufficiency of this evidence, GPT generates an answer to the query based on the extracted evidence fragment. A self-verification step follows: if the GPT-generated answer aligns with the document, the document is deemed a positive match for the query. For MedExam, multiple-choice questions are validated against ground-truth answers. For DuBaike, the generated answers are compared with encyclopedic references for consistency in conveying the same medical knowledge. This detailed process is outlined in lines 10-26.

By leveraging ChatGPT's reasoning and domain knowledge capabilities throughout this iterative loop, we ensure the creation of high-quality, semantically relevant query-document pairs suitable for downstream retrieval tasks.

B.2 Data Example

The constructed datasets cover a wide range of real-world medical scenarios. Representative examples from the ten constituent datasets are shown in Table 15 and Table 16. Queries vary in type,

Dataset	Query URL	#Samples	Document URL	#Samples
MedExam	https://github.com/jind11/MedQA	3,426	https://github.com/jind11/MedQA	27,871
DuBaike	https://github.com/baidu/DuReader	2,000	https://baike.baidu.com/	56,441
DXYDisease	https://dxy.com/diseases	61,840	https://dxy.com/diseases	61,840
MedicalRetrieval	https://huggingface.co/datasets/C-MTEB/MedicalRetrieval	1,000	https://huggingface.co/datasets/C-MTEB/MedicalRetrieval	100,999
CmedqaRetrieval	https://huggingface.co/datasets/C-MTEB/CmedqaRetrieval	3,999	https://huggingface.co/datasets/C-MTEB/CmedqaRetrieval	100,001
DXYConsult	https://dxy.com/questions/	13,057	https://dxy.com/questions/	13,057
CovidRetrieval	https://huggingface.co/datasets/C-MTEB/CovidRetrieval	949	https://huggingface.co/datasets/C-MTEB/CovidRetrieval	100,001
IIYiPost	https://bbs.iiyi.com/	37,065	https://bbs.iiyi.com/	37,065
CSLCite	https://github.com/ydli-ai/CSL	934	https://med.wanfangdata.com.cn/	36,783
CSLRel	https://github.com/ydli-ai/CSL	934	https://med.wanfangdata.com.cn/	36,783

Table 14: Dataset collection sources and quantity statistics.

including medical paper titles, patient symptom descriptions, and examination questions. The corresponding documents consist of medical paper abstracts, doctor-patient diagnostic dialogues, and reference materials for exam questions, illustrating the diversity and practical relevance of CMIRB.

Algorithm 1 Data Preprocessing Pipeline

Input: Query set Q, Document set D, A large language model LLM (e.g., ChatGPT)
Output: High-quality, highly relevant query-document pair collection
// Step 1: Filter out medically irrelevant
for each query $q \in Q$, $d \in D$ **do**
    $med_{score} \leftarrow$ LLM.med_score(q/d)
    if $med_{score} < threshold$ **then**
        Remove q/d
    end if
end for
// Step 2: Matching positive pairs
if query-document matching **then**
    for each query $q \in Q$ **do**
        // Retrieve top-k documents
        $D_k \leftarrow$ BM25(q, D)
        $D_k \leftarrow$ LLM.reranking$(q; D_k)$
        // Extract evidence snippets
        $E_k \leftarrow$ LLM.extract_evidence(q,D_k)
        // Generate answers
        $A_k \leftarrow$ LLM.answer(q, E_k)
        for each document d_i **do**
            if LLM.validate(a_i, d_i) **then**
                Store (q, d_i)
            end if
        end for
    end for
end if
// Step 3: Filter out pseudo-relevant pairs
for each matched pair (q, d) **do**
    $rel_{score} \leftarrow$ LLM.filter_score(q, d)
    if $rel_{score} < threshold$ **then**
        Remove (q, d)
    end if
end for

Medical Relevance Prompt

You will receive a question-answer pair from Baidu Search. Your task is to evaluate whether the Q&A is related to the medical field and output the result in JSON format.
The JSON object must include the following keys:
- "reason": a string explaining the reason for your judgment.
- "label": an int, 0/1.
Please adhere to the following steps:
- If the content mentioned in the question and answer includes medical information and is related to the medical field, the label should be 1.
- If most of the content in the question and answer is unrelated to the medical field, the label should be 0.
You need to make a judgment and provide a reason. Please output the result as required, and do not output any other content.
Here is the text:
Question: [QUESTION]
Answer: [ANSWER]

Passage Reranking Prompt

You will be given a medical question, a reference (standard) answer, and a model-generated answer. Your task is to evaluate the content similarity between the reference answer and the model-generated answer to determine whether they are conveying the same meaning. Your output is a JSON object, which must contain the following keys:
- "similarity_score": a number between 0 and 1 indicating the content similarity between the two answers.
- "explanation": a detailed explanation of the similarities or differences that justify your similarity score.
Please adhere to the following steps:
- 1. Carefully read the medical question.
- 2. Review the reference answer and the model-generated answer.
- 3. Compare the two answers, focusing on content similarity—whether they convey the same meaning, and lead to the same conclusion.
- 4. Provide a similarity score between 0 and 1, where 1 indicates that the answers are identical in meaning, and 0 indicates different.
- 5. Justify your score by explaining the similarities or differences between the two answers.
The "explanation" should be in Chinese. and your output must always be a JSON object, do not output anything else.
Now here are the question, standard answer, and generated answer.
Question: [QUESTION]
Reference Answer: [REFERENCE ANSWER]
Model-generated Answer: [MODEL-GENERATED ANSWER]

Evidence Extracting Prompt

You will be given a medical question, its answer and a related document. Your task is to extract evidence spans from the document that directly or indirectly support the answer to the medical question. Your output is a JSON object, which must contain the following keys:
- "evidence_spans": a list, a list of passages. Please adhere to the following steps:
- 1. Carefully read the medical question and its answer.
- 2. Review the content of the provided document.
- 3. Identify and extract the passage from the document that directly supports the correct answer to the question.
- 4. If no passage in the document can directly support the correct answer or answer the question, return an empty list.
The "explanation" should be in Chinese. and your output must always be a JSON object, do not output anything else.
Here is the medical question, its answer, and the related document
Question: [QUESTION]
Answer: [ANSWER]
Document: [DOCUMENT]

Figure 4: Prompt for data processing (I).

Answer by Evidence Prompt

You will be given a medical exam question and one or more evidence spans that were extracted from related documents. Your task is to provide a detailed and comprehensive answer to the question based solely on the provided evidence spans. Your output is a JSON object, which must contain the following keys:
- "answer": a string, the answer you derive from the reference documents.
- "reason": a detailed explanation of your reasoning process leading to the answer.
Please adhere to the following steps:
- 1. Review the exam question.
- 2. Review the provided evidence spans.
- 3. Based solely on the information contained in the evidence spans, provide a detailed and comprehensive answer to the question.
- 4. If the evidence spans do not provide sufficient information to answer the question, state "The evidence passage can not answer the question." in "answer" and explain why. If you don't know the answer, don't guess.
You must not use any common knowledge, personal knowledge, or external information beyond the provided evidence spans. The "answer" and "reason" should be in Chinese. and your output must always be a JSON object, do not output anything else.
Now here are the exam question and reference documents.
Question: [QUESTION]
Evidence Spans: [EVIDENCE SPANS]

Validate Answer Prompt

You will be given a medical question, a reference (standard) answer, and a model-generated answer. Your task is to evaluate the content similarity between the reference answer and the model-generated answer to determine whether they are conveying the same meaning. Your output is a JSON object, which must contain the following keys:
- "similarity_score": a number between 0 and 1 indicating the content similarity between the two answers.
- "explanation": a detailed explanation of the similarities or differences that justify your similarity score.
Please adhere to the following steps:
- 1. Carefully read the medical question.
- 2. Review the reference answer and the model-generated answer.
- 3. Compare the two answers, focusing on content similarity—whether they convey the same meaning, and lead to the same conclusion.
- 4. Provide a similarity score between 0 and 1, where 1 indicates that the answers are identical in meaning, and 0 indicates different.
- 5. Justify your score by explaining the similarities or differences between the two answers.
The "explanation" should be in Chinese. and your output must always be a JSON object, do not output anything else.
Now here are the question, standard answer, and generated answer.
Question: [QUESTION]
Reference Answer: [REFERENCE ANSWER]
Model-generated Answer: [MODEL-GENERATED ANSWER]

Query-Document Relevance Prompt

You will be given a medical search query and its associated passage. Your task is to evaluate the quality of query-passage pairs intended for use in a medical encyclopedia knowledge retrieval evaluation dataset. Your output is a JSON object, which must contain the following keys:
- "quality_score": an integer, a score from 1 to 5.
- "explanation": a string, providing a brief rationale for the given score.
Please adhere to the following steps:
- 1. Carefully read the query to understand the user's information need.
- 2. Review the passage to assess its relevance and targeted content in relation to the query.
- 3. Assign a quality score from 1 to 5 and explain your reasoning.
The "explanation" should be in Chinese. and your output must always be a JSON object, do not output anything else.
Now here are the query and passage. Query: [QUERY]
Passage: [PASSAGE]

Figure 5: Prompt for data processing (II).

MedExam

Query: 问题: 胃癌最常发生的转移途径是（）。选项: A:直接蔓延, B:血性转移, C:种植转移, D:淋巴转移, E:沿肠管转移.
(EN) *Question: The most common metastasis route for gastric cancer is (). Options: A: Direct spread, B: Hematogenous metastasis, C: Seeding metastasis, D: Lymphatic metastasis, E: Along the intestinal tract.*
Document: 外科学 3.胃癌的扩散与转移 (2)淋巴转移：是胃癌的主要转移途径，进展期胃癌的淋巴转移率高达70%左右，侵及黏膜下层的早期胃癌淋巴转移率近20%。通常将引流胃的淋巴结分为16组，有的组还可以进一步分为若干亚组...
(EN) *Surgery 3. Gastric cancer dissemination and metastasis (2) Lymphatic metastasis: It is the primary route of metastasis for gastric cancer, with a lymphatic metastasis rate of about 70% in advanced gastric cancer and approximately 20% in early gastric cancer invading the submucosa. Lymph nodes draining the stomach are usually classified into 16 groups, with some groups further divided into several subgroups...*

DuBaike

Query: 强迫症的表现是什么?
(EN) *What are the manifestations of obsessive-compulsive disorder (OCD)?*
Document: 强迫症 临床表现 多发人群焦虑症与遗传因素、个性特点、不良事件、应激因素等均有关系，尤其与患者的个性特点紧密相关，比如：过分追求完美、犹豫不决、谨小慎微、固执等，具备这些不良个性特征容易患强迫症...
(EN) *Obsessive-Compulsive Disorder Clinical Manifestations Prevalent Population Anxiety disorders are related to genetic factors, personality traits, adverse events, and stress factors, particularly closely linked to the patient's personality traits. For instance, excessive perfectionism, indecisiveness, meticulousness, and stubbornness are traits that increase the risk of developing OCD...*

DXYDisease

Query: 维生素 A 缺乏症者需要做哪些检查来诊断?
(EN) *What tests are needed to diagnose vitamin A deficiency?*
Document: 最准确的就是血液学检查。抽血检查血清维生素 A 的水平，对于成人来说，如果在 1.05～3.15 μmol/L，那么就表明不存在维生素 A 缺乏。 如果低于参考范围下限，那就是维生素 A 缺乏了。 ...
(EN) *The most accurate test is a hematological examination. A blood test to check the serum vitamin A levels is conducted. For adults, if the levels are between 1.05 and 3.15 mol/L, it indicates that there is no vitamin A deficiency. If the levels are below the lower limit of the reference range, it indicates vitamin A deficiency....*

MedicalRetrieval

Query: 一般宝宝的肚脐眼要多久愈合?
(EN) *How long does it take for a baby's belly button to heal?*
Document: 你好，宝宝的肚脐一般是1-2周左右会好的，时间长的也有一个月的，不过这个时候可能会有脐茸了。
(EN) *Hello, a baby's belly button generally heals in about 1 to 2 weeks, although it may take up to a month in some cases. During this time, there might also be umbilical granuloma.*

CmedqaRetrieval

Query: 甲状腺手术后多久可以干活?
(EN) *How long after thyroid surgery can one return to work?*
Document: 皮肤的修复一般由两周左右就会不影响你的颈部活动了，至于皮下软组织以及肌肉组织的修复可能时间长一下，一般一个月后就不会有明显影响了，你就可以工作了。工作中注意不要劳累，调整好自己的情绪。
(EN) *The skin usually heals in about two weeks, and you should no longer have restrictions on neck movement. However, the repair of subcutaneous soft tissue and muscle tissue may take longer. Generally, after about a month, there should be no significant impact, and you can return to work. During work, be sure to avoid overexertion and manage your emotions well.*

Table 15: Data example in CMIRB (I).

DXYConsult

Query: 症状及患病时长：感冒，鼻炎，失去嗅觉一周。就医及用药情况：未就医，自行服用泰诺。需要解答的问题：鼻炎，失去嗅觉怎么办

(EN) *Symptoms and Duration of Illness: Cold, rhinitis, loss of smell for one week. Medical Consultation and Medication: No medical consultation, self-medicated with Tylenol. Questions Needing Answers: What to do about rhinitis and loss of smell?*

Document: 你好，如果近期有这种感冒的病史的话，就会导致出现嗅觉功能下降，建议在口服感冒药的技术上的话，用海盐水冲洗鼻腔，一天两次，鼻喷辅舒良或者内舒拿看看效果，如果分泌过多的话，可以口服桉柠蒎胶囊，每天三次每次一粒。

(EN) *Hello, if there has been a recent history of cold symptoms, this can lead to decreased olfactory function. It is recommended to use saline nasal irrigation twice a day while taking cold medicine. You may also try nasal sprays like Budesonide or Fluticasone to see if they help. If there is excessive secretion, you can take Eucalyptus and Menthol capsules, three times a day, one capsule each time.*

CovidRetrieval

Query: 如何对待因履行工作职责感染新冠肺炎的医务人员?

(EN) *How should healthcare workers who contract COVID-19 while fulfilling their duties be treated?*

Document: ...为进一步加强疫情防控期间医务人员防护工作，切实保障医务人员身心健康，现将有关要求通知如下：一、高度重视医务人员防护工作做好医务人员防护工作，是预防和减少医务人员感染的关键举措，...

(EN) *...To further enhance the protection of healthcare workers during the pandemic and ensure their physical and mental well-being, the following requirements are hereby notified: Pay great attention to the protection of healthcare workers Ensuring proper protection for healthcare workers is a key measure in preventing and reducing infections among them, ...*

IIYiPost

Query: 病例讨论：静脉输入阿昔洛韦2天，出现腰痛、尿少

(EN) *Case Discussion: Two days of intravenous acyclovir, followed by lower back pain and reduced urine output*

Document: 1.病例资料,患者，男，31岁。因静脉输入阿昔洛韦2天，出现腰痛、尿少伴恶心、呕吐6天入院。患者8天前因受凉感冒，出现咳嗽、发热（最高体温38.6°C），无明显咳痰，院外静脉给予NS500ml+青霉素钠盐800万U，vd，1次/日,...

(EN) *Case Data, Patient: Male, 31 years old. The patient was admitted after experiencing lower back pain and reduced urine output, accompanied by nausea and vomiting for six days following two days of intravenous acyclovir administration. Eight days prior, the patient had caught a cold due to exposure, presenting with a cough and fever (highest temperature of 38.6°C), without significant sputum production. He received intravenous administration of ...*

CSLCite

Query: 微球在组织工程中的应用

(EN) *Application of Microspheres in Tissue Engineering*

Document: 背景:骨组织工程骨构建中如何使生长因子持续高效发挥作用是影响成骨速度和质量的关键,现多以各种材料的微球或支架作为缓释载体,但缓释作用有待提高.目的:实验拟制备壳聚糖微球,然后复合到纳米羟基磷灰石/聚乳酸羟基乙酸支架上...

(EN) *Background: In bone tissue engineering, maintaining the sustained and efficient activity of growth factors is key to influencing the speed and quality of bone formation. Currently, microspheres or scaffolds made from various materials are commonly used as sustained-release carriers, but the release efficiency needs improvement. Objective: This experiment aims to prepare chitosan microspheres and incorporate them into a nano-hydroxyapatite/polylactic-co-glycolic acid (nHA/PLGA) scaffold, ...*

CSLRel

Query: 高血压病的辨治及预防 高血压病可归属中医学"眩晕"、"头痛"等范畴,其起病隐匿,不易引起患者的充分重视,中后期可致心脑血管疾病、肾损害...

(EN) *Differentiation and Treatment of Hypertension and Its Prevention Hypertension can be categorized under the terms "dizziness" and "headache" in traditional Chinese medicine (TCM). Its onset is insidious, often not receiving enough attention from patients, ...*

Document: 辨证施治高血压 高血压病是现代医学病名,在中医归属眩晕病范畴,中医认为高血压与风、火、痰、虚有关.高血压的界定根据世界卫生组织(WHO)的标准,成人在休息状态下,收缩压持续高于140毫米汞柱...

(EN) *TCM Syndrome Differentiation and Treatment of Hypertension Hypertension is a modern medical term, categorized under dizziness in TCM. TCM holds that hypertension is related to wind, fire, phlegm, and deficiency ...*

Table 16: Data example in CMIRB (II).