Benchmarking and Enabling Efficient Chinese Medical Retrieval via Asymmetric Encoders

Angqing Jiang[1,2,3] **Jianlyu Chen**[1,2] **Zhe Fang**[2,3] **Yongcan Wang**[2,3] **Xinpeng Li**[2,3]
Keyu Ding[4*] **Defu Lian**[1,2]
[1]University of Science and Technology of China
[2]State Key Laboratory of Cognitive Intelligence [3]iFlytek Research
[4]HeFei Institute of Technology
{philipgaq, chenjianlv}@mail.ustc.edu.cn kyding@hfit.edu.cn
liandefu@ustc.edu.cn {zhefang, ycwang12, xpli}@iflytek.com

Abstract

Effective medical text retrieval requires both high accuracy and low latency. While LLM-based embedding models possess powerful retrieval capabilities, their prohibitive latency and high computational cost limit their application in real-time scenarios. Furthermore, the lack of comprehensive and high-fidelity benchmarks hinders progress in Chinese medical text retrieval. In this work, we introduce the **C**hinese **Med**ical **T**ext **E**mbedding **B**enchmark (**CMedTEB**), a benchmark spanning three kinds of practical embedding tasks: retrieval, reranking, and semantic textual similarity (STS). Distinct from purely automated datasets, CMedTEB is curated via a rigorous multi-LLM voting pipeline validated by clinical experts, ensuring gold-standard label quality while effectively mitigating annotation noise. On this foundation, we propose the **C**hinese Medical **A**symmetric **RE**triever (**CARE**), an asymmetric architecture that pairs a lightweight BERT-style encoder for online query encoding with a powerful LLM-based encoder for offline document encoding. However, optimizing such an asymmetric retriever with two structurally different encoders presents distinctive challenges. To address this, we introduce a novel two-stage training strategy that progressively bridges the query and document representations. Extensive experiments demonstrate that CARE surpasses state-of-the-art symmetric models on CMedTEB, achieving superior retrieval performance without increasing inference latency.

1 Introduction

Text embedding models are essential for a wide range of natural language processing (NLP) tasks, including retrieval, reranking, and classification (Reimers and Gurevych, 2019). Their role is crucial in retrieval-augmented generation (RAG) systems (Lewis et al., 2020; Zhang et al., 2026), which leverage external knowledge to enhance large language models (LLMs). In specialized domains such as healthcare, where LLMs often lack deep expert knowledge, accurate and low-latency access to medical knowledge can enhance clinical decision support and mitigate hallucinations in RAG, making domain-specific, low-latency embeddings indispensable.

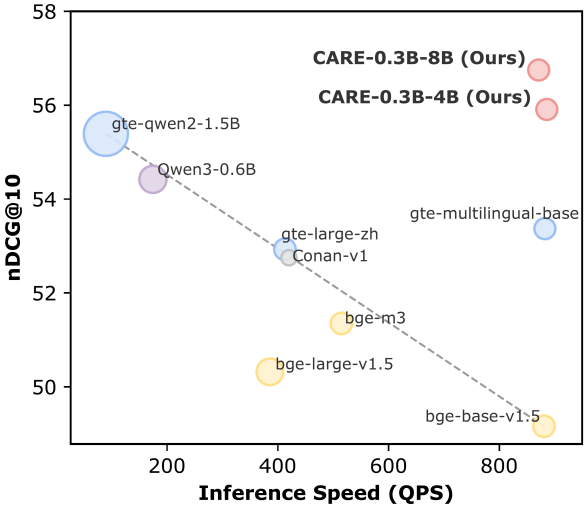

Figure 1: **Efficiency-performance trade-off on CMedTEB Retrieval.** The x-axis shows queries per second (QPS) on a single A100 80GB GPU, while the y-axis reports nDCG@10. Notably, **CARE breaks the conventional trade-off**: it matches the high retrieval quality of heavy LLM-based models while sustaining the high throughput of lightweight BERT-style models.

Despite recent rapid progress in general-domain embedding models (e.g., BGE (Chen et al., 2024a), GTE (Li et al., 2023), Qwen3-Embedding (Zhang et al., 2025)), Chinese medical text embedding has received limited attention. Existing benchmarks like C-MTEB (Xiao et al., 2024) include only two Chinese medical retrieval datasets, both suffer from annotation sparsity and substantial false negatives (see analysis in Section 3). While recent benchmark CMIRB (Li et al., 2025) scale up data via single-LLM generation (e.g., ChatGPT), it lacks

*Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20000–20020

rigorous multi-model consensus and expert verification. Besides, it focuses exclusively on retrieval tasks while still incorporating legacy datasets with false negative issues, leaving a gap for a comprehensive, high-fidelity benchmark. Moreover, while current state-of-the-art embedding models are mostly LLM-based (Lee et al., 2024; Lin et al., 2025), they incur prohibitive latency and high computational costs, which limit their applications in latency-sensitive scenarios such as real-time medical QA. This highlights the challenge of accuracy-latency trade-off.

To address the challenges of false negative noise and ensure robust evaluation beyond single LLM generated labels, we introduce the **C**hinese **Med**ical **T**ext **E**mbedding **B**enchmark (**CMedTEB**). CMedTEB consists of three newly curated tasks: retrieval, reranking, and medical synonym STS. To ensure high data quality, we employ a **multi-LLM consensus pipeline** for annotation. Our experiments indicate that even advanced general-purpose embedders underperform on CMedTEB, highlighting the benchmark's difficulty and value for domain-specific optimization.

To address the trade-off between retrieval accuracy and online inference latency in medical retrieval, we propose **C**hinese Medical **A**symmetric **RE**triever (**CARE**), an asymmetric architecture that pairs a lightweight BERT-style encoder for online query encoding with a powerful LLM-based encoder for offline document encoding. By employing a novel two-stage alignment strategy, CARE effectively bridges the semantic gap, achieving LLM-level accuracy with BERT-level online latency. As shown in Figure 1, while most embedding models exhibit a clear accuracy-latency trade-off, CARE breaks this trend. It matches the retrieval accuracy of large-scale LLM-based embedding models while sustaining Queries Per Second (QPS) levels comparable to small-size BERT-style embedding models. We further observe that as the document encoder scales up, the asymmetric model progressively closes the gap with LLM-based embedding models, offering a practical path to scale retrieval performance without sacrificing online latency.

The primary contributions of our work are summarized as follows:

- We introduce **CMedTEB**, a comprehensive, high-fidelity benchmark for Chinese medical text embedding, establishing a reliable standard for medical domain-specific evaluation.
- We propose the **CARE framework**, which integrates an asymmetric inference architecture with a progressive two-stage training strategy. Our design reconciles the efficiency-accuracy conflict, enabling LLM-level retrieval quality with BERT-style inference latency.
- To facilitate future research in medical text retrieval, we will open-source our benchmark, models, and code at our repository[1].

2 Related Work

Medical Retrieval Benchmarks. MTEB (Muennighoff et al., 2022) provides a comprehensive benchmark across languages and tasks, and its Chinese extension C-MTEB (Xiao et al., 2024) includes several Chinese embedding model datasets. However, domain-specific evaluation of Chinese medical remains scarce. Existing Chinese medical benchmarks in C-MTEB, such as CmedqaRetrieval (Zhang et al., 2017) and MedicalRetrieval (Long et al., 2022), both exhibit annotation noise and false negatives (see analysis in Appendix E). Although recent works like CMIRB (Li et al., 2025) attempt to address data scarcity via automated generation, they rely on unverified single-source LLM judgments and incorporate noisy legacy datasets. Furthermore, it focuses exclusively to retrieval tasks. As a result, the field still lacks a comprehensive, high-quality, and human-verified benchmark for Chinese medical text embedding.

Embedding Models. Text embedding models have advanced rapidly alongside pretrained language models. Early works such as Contriever (Izacard et al., 2021) explored unsupervised contrastive pretraining, while more recent models like E5 (Wang et al., 2022), GTE (Li et al., 2023), and the BGE series (Chen et al., 2024a) leveraged large-scale contrastive pretraining to obtain strong general-purpose embeddings. In the biomedical domain, specialized models such as MedCPT (Jin et al., 2023) and BMRetriever (Xu et al., 2024) leverage large-scale medical corpus and tuning language models for enhanced retrieval. Recently, decoder-only embedding models such as Qwen3-Embedding (Zhang et al., 2025), bge-en-icl (Li et al., 2024a), and NV-Embed (Lee et al., 2024) have achieved state-of-the-art performance on MTEB (Muennighoff et al., 2022).

Despite these advances, most LLM-based models contain billions of parameters. While they

[1] https://github.com/PhilipGAQ/CARE

deliver strong accuracy, their high latency and computational overhead make them impractical for latency-sensitive applications such as real-time medical retrieval. This gap highlights the urgent need for lightweight yet effective embedding models in specialized domains.

Asymmetric Retrieval Architecture. While existing works of asymmetric architectures offer a promising path for retrieval efficiency, achieving effective alignment between disparate encoders remains a significant challenge. Existing approaches primarily follow two paradigms: (1) Homogeneous Distillation: Methods like KALE (Wang and Lyu, 2023; Campos et al., 2023) prune layers from a teacher model to initialize a student. However, this creates a rigid architectural dependency, constraining the student model to the specific design of the teacher. (2) Heterogeneous Alignment: Recent works such as ScalingNote (Huang et al., 2024) or HotelMatch (Askari et al., 2025) aligns different architectures. However, aligning heterogeneous encoders presents a significant semantic gap. Without tailored training, lightweight models fail to adapt to the complex embedding space of large document encoders, resulting in suboptimal retrieval performance.

In contrast, we propose a novel **two-stage training strategy** that bridges the representational gap between the lightweight query encoder and the LLM-based document encoder. This progressive approach ensures stable convergence, enabling the smaller model to accurately map inputs into the rich semantic space of the larger model without architectural constraints.

3 CMedTEB

Chinese medical text embedding benchmarks remain scarce. Among the few available benchmarks, CmedqaRetrieval (Zhang et al., 2017) and MedicalRetrieval (Long et al., 2022) are well known and widely used. These datasets are constructed primarily from human-labeled query-answer pairs sourced from online medical Q&A platforms, such as patient inquiries and physician responses. However, this methodology inherently ignores potentially relevant yet unlabeled candidate answers associated with other pairs. The medical domain further exhibits *topic intensity*: common diseases or medications often generate a large volume of semantically similar queries and answers, increasing the risk of false negatives (See Table 22 for examples).

To quantify this issue, we performed an analysis of current benchmarks using LLM-assisted annotation followed by human verification (details in Appendix E). The results suggested a significant presence of **potential false negatives**: on average, each query is associated with approximately **9 False Negatives** in MedicalRetrieval, and **19 False Negatives** in CmedqaRetrieval. To validate the reliability of these findings, we employed human assessors to re-judge a random sample of 500 pairs which LLM labeled as false negatives. yielding a **92% consistency rate** with model's suggestions, providing strong evidence that existing retrieval benchmarks suffer from notable annotation noises.

To address the challenges of false negatives and ensure robust evaluation beyond single LLM generated labels, we construct CMedTEB (Figure 2) via a rigorous multi-LLM consensus pipeline. CMedTEB comprises three new tasks: Retrieval, Reranking, and Synonym STS, along with two high-quality, human-verified existing public datasets CMedQAv1-reranking (Zhang et al., 2017) and CMedQAv2-reranking (Zhang et al., 2018). We construct the document corpus using data from **XunYiWenYao**[2], and sampling queries from logs of an online medical service (construction details and anonymization steps on queries and documents are available in Appendix A).

3.1 Construction Method

Retrieval Prior studies like AIR-Bench (Chen et al., 2024b) and Thomas et al. (2024) demonstrate the reliability of LLM-generated relevance labels in information retrieval benchmark. Building on these findings, we adopt a multi-LLM labeling pipeline. Given a query $q_i \in \mathcal{Q}$, we used gte-multilingual-base, bge-m3, Conan-embedding-v1 to retrieve and gather a candidate pool of top-500 documents $\mathcal{D}_i = \{d_1, \ldots, d_{500}\}$. Three strong LLMs, DeepSeek-V3 (Liu et al., 2024), Doubao-1.5-Pro (Guo et al., 2025) and GPT-4o (Hurst et al., 2024) then rated each (q_i, d_j) pair on a 5-point relevance scale. To ensure label quality, a document was retained as positive only when all three LLMs agreed, while pairs with partial agreement (only 1 or 2 agreements) were discarded. The final retrieval dataset comprise a query set $\mathcal{Q}$, a refined corpus $\mathcal{D}' \subseteq \mathcal{D}$, and relevance labels $\mathcal{R} = \{(q_i, d_j, y_{ij}) \mid y_{ij} \in \{0, 1\}\}$.

[2] https://www.xywy.com/

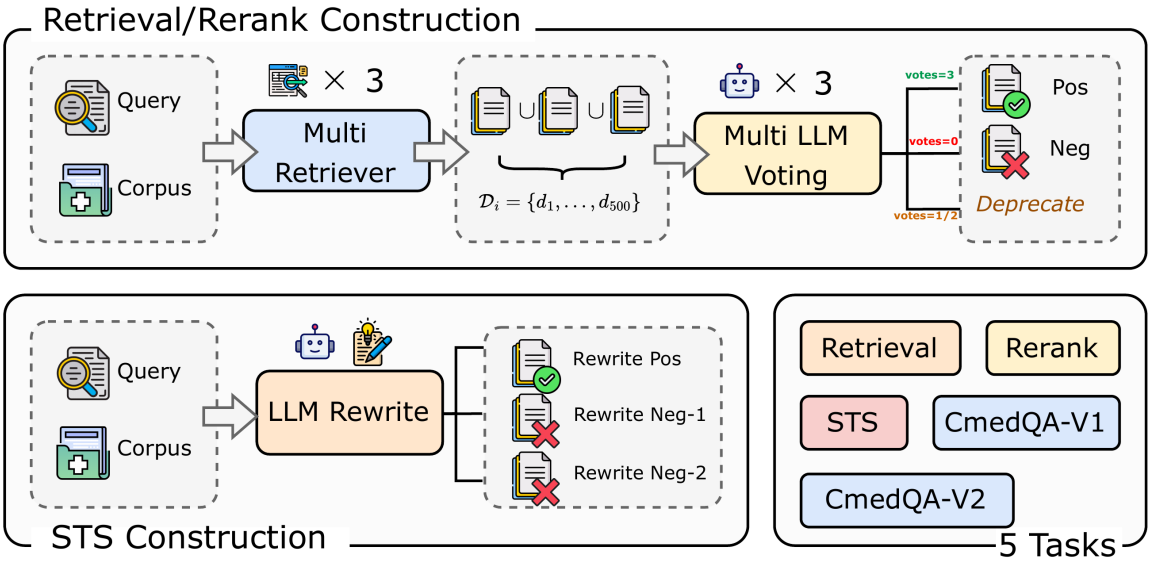

Figure 2: **Workflow for constructing the CMedTEB benchmark.** The Figure shows the distinct curation strategies for discriminative tasks (Retrieval/Rerank) and semantic similarity tasks (STS).

Task	Test	Train	Main Metric
New tasks			
Retrieval	734	20,000	nDCG@10
Rerank	1,128		MAP@10
Synonym STS	5,000	10,000	Pearson
Public datasets			
CMedQA-v1-rk.	1,000	50,000	MAP@10
CMedQA-v2-rk.	1,000		MAP@10

Table 1: CMedTEB statistics

Rerank For each query $q_i \in \mathcal{Q}$, we derive positives $P_i = \{d_j \in \mathcal{D}' : y_{ij} = 1\}$ and negatives $N_i = \{d_j \in \mathcal{D}' : y_{ij} = 0\}$ labeled from the same multi-LLM labeling pipeline as in Retrieval. The reranking dataset is a collection of triplets $\mathcal{T}_{\textbf{Rerank}} = \{(q_i, \mathcal{P}_i, \mathcal{N}_i)\}$, where $\mathcal{P}_i$ is a list sampled from P_i and $\mathcal{N}_i$ is a list sampled from N_i.

STS We first build a medical synonym dictionary with domain experts. For each $q_i \in \mathcal{Q}$, GPT-4o generates three sentences: a positive s_i^+ (synonym substitution with semantics preserved), a hard negative $s_{i,1}^-$ (synonym substitution with semantics changed), and an easy negative $s_{i,2}^-$ (no synonym substitution with semantics changed). We then sample $s_i \in \{s_i^+, s_{i,1}^-, s_{i,2}^-\}$ and pair it with q_i to form (q_i, s_i, y_i), where $y_i = \mathbf{1}[\, s_i = s_i^+ \,] \in \{0, 1\}$. The dataset is $\mathcal{T}_{\text{STS}} = \{(q_i, s_i, y_i)\}$, evaluating fine-grained synonym understanding.

3.2 Quality Analysis

The statistics of CMedTEB are summarized in Table 1, with detailed breakdowns provided in Appendix A.2.

Annotation Accuracy Assessment. To ensure annotation reliability, we computed Fleiss' Kappa (Fleiss, 1971) among the results of three LLMs in our pipeline, yielding a score of **0.731**, indicating substantial agreement. Furthermore, a clinical expert independently re-annotated a large-scale sample of 5,000 query-document pairs, achieving a **93.3%** agreement rate with our final labels. These metrics confirm that our automated pipeline aligns closely with clinical expertise.

Task Difficulty and Distinctness. To demonstrate the necessity of this benchmark, we evaluated existing general-domain embedding models (full zero-shot results are detailed in Appendix Table 11). We observe a sharp performance contrast: while models achieve high accuracy on legacy CMedQA tasks (Avg. 85.15), their performance drops drastically on CMedTEB new tasks (Avg. 57.85). Moreover, the Spearman rank correlation between model rankings on CMedQA versus CMedTEB new tasks is notably weak ($\rho = 0.354, p = 0.215 \gg 0.05$). This statistical lack of significant correlation indicates that CMedTEB is **non-redundant**, aiming to evaluate medical semantic capabilities overlooked by prior datasets. Finally, while massive decoder-only models like Qwen3-Embedding-8B

achieve stronger results (Avg. 64.52), their prohibitive latency underscores the urgent need for efficient, domain-specialized solutions.

4 CARE

To bridge the gap between retrieval accuracy and inference latency observed in CMedTEB, Section 3.2, we propose the **Chinese Medical Asymmetric Retriever (CARE)**. As illustrated in Figure 3, CARE adopts a decoupled architecture where a powerful LLM serves as the offline document encoder to capture rich semantics, while a lightweight BERT model acts as the online query encoder to ensure low-latency inference.

In this section, we describe our high-quality training data construction for medical domain, and a two-stage training strategy designed for our asymmetric embedding architecture.

4.1 Training Data Construction

Standard hard negative mining fails in the medical domain due to *topic intensity*, where abundant latent positives lead to severe false negatives. We address this via a **diversity-aware curation pipeline** applied independently to query and document sets. The process initializes a vector index with 5,000 seed samples. For each new candidate x, we retrieve its top-k neighbors from the evolving index; if the count of neighbors exceeding similarity threshold t surpasses n, x is discarded as redundant. Otherwise, it is added to update the distribution. Finally, we employ GPT-4o to verify top-50 retrieved candidates, distinguishing hard negatives from false positives. This yields **500K high-fidelity triples** (q, d^+, d^-). Details are presented in Appendix B.

4.2 Asymmetric Embedding Architecture

While LLM-based embedders achieve state-of-the-art retrieval accuracy, their high computational cost and latency are prohibitive for real-time applications. To resolve this trade-off, we propose an **Asymmetric embedding architecture**, which pairs a lightweight query encoder E_Q for fast online inference with a powerful document encoder E_D whose embeddings are pre-computed offline.

To establish strong foundation before alignment, we first initalize E_Q and E_D independently with contrastive learning. Crucially, E_D employs Matryoshka Representation Learning (MRL) (Kusupati et al., 2022) to truncate its native embedding dimension to match the smaller E_Q.

Despite dimensional compatibility, a significant semantic gap persists between the embedding spaces of these heterogeneous models. We bridge this gap via a progressive two-stage strategy: (1) **Query Encoder Alignment**, which maps the query encoder's space to the frozen document encoder; (2) **Joint Fine-Tuning**, which optimizes both encoders for end-to-end retrieval performance.

4.2.1 Asymmetric Stage I: Query Encoder Alignment

In this stage, we freeze the document encoder (the *teacher*) and update only the query encoder (the *student*) to align the student's space to the teacher's.

To bridge the representational gap without consuming an amount of labeled data, we propose a **Self-Contrastive** strategy over abundant unlabeled corpora. Specifically, we treat each input text $\mathbf{x}$ as a positive anchor for itself. This unsupervised paradigm effectively leverages millions of raw texts, allowing the alignment process to scale up free from the constraints of annotation.

Objective Function We employ a hybrid objective to enforce alignment from two perspectives:

Asymmetric Contrastive Loss. We use Asym-InfoNCE with the frozen document encoder as the teacher:

$$\mathcal{L}_{\text{InfoNCE}}^{Asym} = -\log \frac{\exp(s^+/\tau)}{\exp(s^+/\tau) + \sum_{i=1}^{N} \exp(s_i^-/\tau)}, \tag{1}$$

where $s^+ = \text{sim}(E_Q(\mathbf{x}), E_D(\mathbf{x}))$ represents the similarity of the same text's embedding across two encoders, and $s_i^- = \text{sim}(E_Q(\mathbf{x}), E_D(\mathbf{x}_i^-))$ represents similarity with in-batch negatives. This explicitly trains the student E_Q to identify the teacher E_D's representation of the same input against negative distractors.

MSE Loss. For stricter alignment, we minimize the L2 distance:

$$\mathcal{L}_{\text{MSE}} = \|E_Q(\mathbf{x}) - E_D(\mathbf{x})\|_2^2, \tag{2}$$

where embeddings are normalized. This penalizes absolute deviations in the embedding space.

The final objective is

$$\mathcal{L}_{\text{Stage 1}} = \lambda_1 \mathcal{L}_{\text{InfoNCE}}^{Asym} + \lambda_2 \mathcal{L}_{\text{MSE}} \tag{3}$$

We empirically set $\lambda_1 = \lambda_2 = 1$ to balance soft ranking alignment (InfoNCE) with hard structural alignment (MSE).

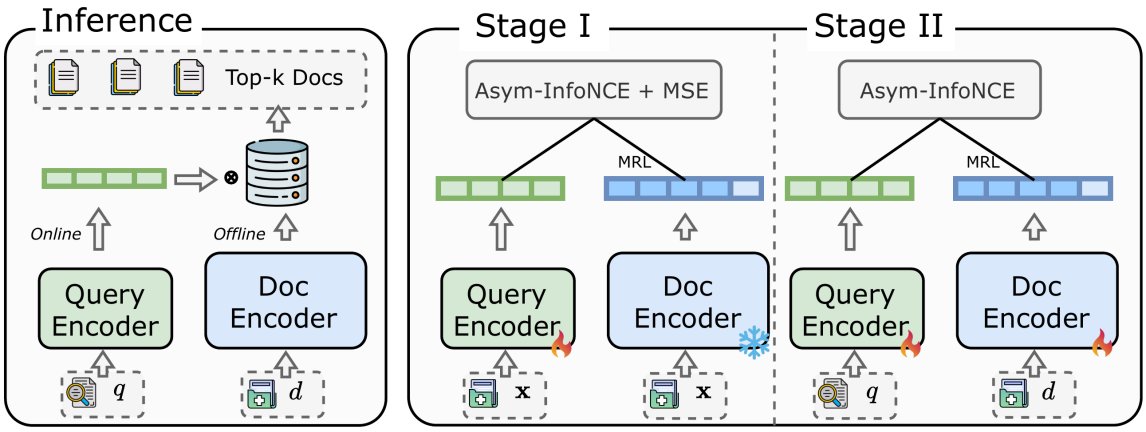

Figure 3: **Inference and Training pipeline for asymmetric embedding model.** Stage I: Query encoder is trained to align with the frozen document encoder using Asym-InfoNCE and MSE losses. Stage II: Both encoders are jointly fine-tuned with Asym-InfoNCE loss on retrieval data.

4.2.2 Asymmetric Stage II: Joint Fine-tuning

After alignment, we unfreeze both encoders and perform end-to-end joint fine-tuning. The goal of this stage is to further enhance retrieval performance by jointly optimizing the two encoders to better discriminate between positive and negative documents. We adopt the **Asym-InfoNCE loss** as the sole objective:

$$\mathcal{L}_{\text{Stage 2}} = -\log \frac{e^{s(q,d^+)/\tau}}{e^{s(q,d^+)/\tau} + \sum_{d^- \in \mathcal{N}} e^{s(q,d^-)/\tau}}, \tag{4}$$

where $s(q,d) = \text{sim}(E_Q(q), E_D(d))$, and $\mathcal{N}$ includes both in-batch negatives and the hard negatives. Unlike Stage 1, q and d^+ are semantically relevant pairs. This end-to-end optimization directly refines the decision boundary for retrieval.

5 Experiments

In our experiments, we aim to answer the following research questions:

- **RQ1:** How does CARE compare with state-of-the-art baseline embedding models on CMedTEB?
- **RQ2:** Can CARE match the performance of large symmetric models with **lightweight** inference?
- **RQ3:** Is the proposed CARE framework superior to other efficient retrieval methods?
- **RQ4:** What are the contributions of different components in our training strategy?

5.1 Setup

Training Data. We use the curated dataset (Section 4.1) and CMedTEB training splits. Although some baselines have previously seen CMedQA during pre-training, we explicitly include it to prevent potential performance degradation on this task.

Baselines. As we target efficient online deployment, we prioritize comparisons with *lightweight* yet strong open-source models (e.g., BGE (Xiao et al., 2024), GTE (Li et al., 2023), Conan (Li et al., 2024b), Stella (Zhang et al., 2024)) and moderate-sized LLM embedders (Qwen3-Embedding (Zhang et al., 2025), gte-Qwen2 (Li et al., 2023)), rather than prohibitive large size models.

Implementation Details. The query encoder (**Med-Emb-base**) is initialized from `gte-multilingual-base`. Document encoders (**Med-Emb-4B/8B**) are fine-tuned from Qwen3 using LoRA. Implementation details are in Appendix C.1. We apply Matryoshka Representation Learning (MRL) to project their native dimensions (2560/4096) to the query encoder's 768-dim space. We evaluate two asymmetric variants: **CARE-0.3B-4B** and **CARE-0.3B-8B**. Experiments run on 32×A100 GPUs. Details are in Appendix C.3.

5.2 Main results on CMedTEB (RQ1)

Table 2 presents the evaluation of CARE series on the CMedTEB benchmark, alongside strong open-source baselines. We observe two key findings: (1) CARE establishes a new state of the art: the 0.3B-4B variant achieves an average score of 78.13, and the 0.3B-8B variant reaches 78.94, surpass-

Model	Params (Q/D)	CMed v1 MAP@10	CMed v2 MAP@10	Retrieval nDCG@10	Rerank MAP@10	STS Pearson	Avg
Baselines							
bge-small-zh-v1.5	24M / 24M	80.21	81.69	44.33	62.30	70.50	67.81
bge-base-zh-v1.5	102M / 102M	83.37	83.31	49.16	66.73	76.24	71.76
bge-large-zh-v1.5	326M / 326M	83.23	85.15	50.32	67.55	78.95	73.04
bge-m3	568M / 568M	82.98	83.32	51.35	66.90	78.34	72.58
Conan-embedding-v1	326M / 326M	**89.89**	88.77	52.75	69.31	81.49	76.44
stella-base-zh-v3-1792d	102M / 102M	87.16	88.28	53.31	69.56	80.52	75.77
gte-multilingual-base	305M / 305M	86.21	86.37	53.37	69.38	82.36	75.54
gte-base-zh	102M / 102M	85.31	86.44	52.62	69.35	79.73	74.69
gte-large-zh	326M / 326M	85.44	86.97	52.93	69.97	81.48	75.36
gte-Qwen2-1.5B-instruct	1.78B / 1.78B	87.68	87.15	55.39	72.35	85.50	77.61
Qwen3-Embedding-0.6B	596M / 596M	85.58	86.09	54.42	70.94	80.42	75.49
Ours							
CARE-0.3B-4B†	305M / 4.02B	86.04	87.31	55.91	72.84	**88.53**	78.13
CARE-0.3B-8B†	305M / 8.19B	88.34	**88.86**	**56.75**	**73.67**	87.07	**78.94**

Table 2: **Results of our models compare to the baselines on CMedTEB.** Best results in **bold**, second-best in underline. Asymmetric models are marked with †.

ing the strongest baseline gte-Qwen2-1.5B-instruct (77.61, a decoder-only model), despite using a much smaller query encoder. (2) Both baseline models and our asymmetric variants exhibit consistent performance scaling with model size: enlarging the document encoder from 4B to 8B improves the average score by 0.81 with *zero* increase in online query cost, verifying the superior accuracy-latency trade-off.

5.3 Asymmetric vs. Symmetric Architectures (RQ2)

Figure 4 visualizes the performance trade-off between symmetric (Med-Emb) and asymmetric (CARE) architectures. While symmetric models (Med-Emb-4B/8B) set a high performance ceiling, they incur prohibitive computational costs, indicated by the sharp spike in the blue line. In contrast, CARE strikes a balance: CARE-0.3B-8B achieves an average score of 65.21, trailing the fully symmetric 8B giant (65.63) by only 0.6%, yet requires 27× fewer online inference parameters (0.3B vs 8.2B). This confirms that CARE effectively scales performance via the offline document encoder without increasing online latency.

5.4 Comparison with Efficient Retrieval Baselines (RQ3)

We further compare our two-stage asymmetric training framework against several alternative symmetric or asymmetric approaches to efficient retrieval (Table 3, implementation details are in C.2). CARE consistently outperforms all baselines. Asymmetric approaches like KALE and Wang and Lyu (2023) underperform (55.05 and 53.13), likely because their distillation designs originally tailored for encoder-only models transfer poorly to decoder-based architectures.

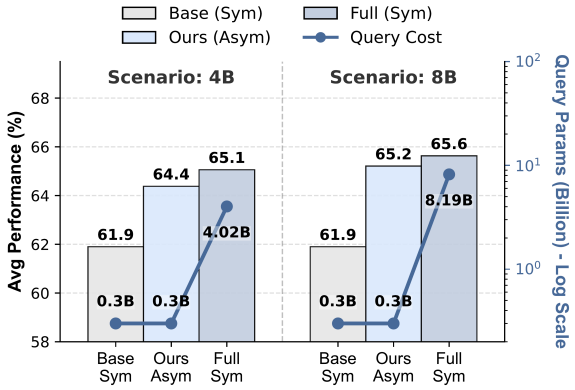

Figure 4: **Asymmetric vs. Symmetric Architectures on performance and parameters.** We evaluate models in two scenarios: leveraging power from 4B (left) and 8B (right) document encoders. The bars denote average performance on CMedTEB Retrieval and Rerank (higher is better), while the blue line represents the inference cost in terms of Query Encoder parameters (log scale, lower is better). CARE (Ours) bridge the performance gap to symmetric giants (Full) while maintaining same online inference cost to baselines.

Moreover, CARE surpasses the distillation baseline (Ren et al., 2021) (distillation 62.72 vs. ours 64.38). While distillation attempts to transfer capabilities via soft labels (KL-divergence), it implies an inherent information loss. In contrast, our framework allows the query encoder to directly interact with teacher document encoder during inference,

Model	Asym	Params	Retrieval nDCG@10	Rerank MAP@10	Avg
KALE	✓	0.3B / 4B	42.67	67.42	55.05
Wang and Lyu (2023)	✓	0.3B / 4B	39.99	66.26	53.13
ScalingNote	✓	0.3B / 4B	34.81	64.17	49.49
Distill-from-4B	×	0.3B	54.68	70.76	62.72
CARE-0.3B-4B	✓	0.3B / 4B	**55.91**	**72.84**	**64.38**

Table 3: **Ablation study on alternative approaches to efficient retrieval.** Params indicate parameter counts for Query / Document encoders.

Setting	Retrieval nDCG@10	Rerank MAP@10	Avg
Independent Initialization			
w/o query init	50.46	68.85	59.66
w/o doc init	37.30	63.21	50.26
Asymmetric Stage			
w/o query align	35.34	66.79	51.07
w/o joint fine-tuning	42.69	68.28	55.49
Loss Design (Asymmetric Stage I)			
w/o MSE	55.19	71.94	63.57
w/o Contrastive	55.48	72.58	64.03
Full Model	**55.91**	**72.84**	**64.38**

Table 4: Ablation study on training stages and loss functions.

effectively bridging the information gap.

5.5 Ablation Study (RQ4)

We conduct ablation study on CARE-0.3B-4B except specially mentioned. To better reflect downstream applications, we report performance on both Retrieval and Reranking tasks.

5.5.1 Training Design

We conduct experiments on the contribution of different components in our asymmetric training framework. Results are in Table 4.

For Independent Initialization, removing either query or document encoder initialization leads to severe performance degradation (*w/o query init* scores: 59.66 and *w/o doc init* scores: 50.26 vs. 64.38 for full model). This underscores the necessity of independent pre-training, where robust symmetric backbones serve as superior initializations for asymmetric alignment.

For Asymmetric Stage, skipping the query alignment stage (*w/o query align* scores: 51.07) or the joint fine-tuning stage (*w/o joint fine-tuning* scores: 55.49) results in clear performance drops (full model 64.38). This validates that the query alignment stage effectively bridges the semantic gap between student and teacher, while joint fine-tuning optimizes both encoders to downstream retrieval. Note that our *w/o query align* configuration is close to HotelMatch (Askari et al., 2025), though not identical: HotelMatch applies a linear projection to up-project the small-LM query embeddings to the document encoder dimension and uses separate learning rates for the two encoders, whereas we remove the projection layer and use a single learning rate since we use LoRA to fine-tune our document encoder.

Finally, we study the loss design in the query align stage. Removing either MSE or Asym-InfoNCE contrastive loss weakens performance. The full model, combining both, consistently achieves the best results. This indicates that both objectives are complementary for effective embedding space alignment.

Training Phase	Stage-I Strategy	Retrieval nDCG@10	Rerank MAP@10	Avg
Stage-I Only	Supervised (Labeled)	44.83	69.69	57.26
	Self-Contrastive (Ours)	42.69	68.28	55.49
Stage-I → Stage-II (Full Pipeline)	Supervised (Labeled)	49.95	71.02	60.49
	Self-Contrastive (Ours)	**55.91**	**72.84**	**64.38**

Table 5: **Impact of Stage-I Strategy.** We compare using Supervised data vs. our Self-Contrastive approach during Stage-I alignment, followed by identical Stage-II fine-tuning.

5.5.2 Impact of Self-Contrastive Alignment

We investigate the necessity of the unsupervised Self-Contrastive alignment task in Stage-I (Section 4.2.1) by comparing it against using supervised contrastive fine-tuning data. As detailed in Table 5, using supervised data in Stage-I yields higher immediate metrics compared to our self-contrastive approach (57.26 vs. 55.49). However, the trend reverses significantly after the Stage-II end-to-end fine-tuning: models initialized via our self-contrastive alignment achieve a decisive performance gain (64.38), surpassing those pre-trained with labeled data (60.49).

Early supervision in Stage-I likely causes *premature convergence* to local minima. Conversely, the self-contrastive strategy provides a robust unsupervised alignment, establishing a superior foundation for subsequent fine-tuning.

6 Conclusion

In this work, we introduce CMedTEB, a new benchmark for Chinese medical text embedding, and propose CARE, an asymmetric model designed for

efficient, low-latency medical retrieval. Our architecture, which pairs a lightweight query encoder with a powerful document encoder via a two-stage training strategy, achieves state-of-the-art performance on CMedTEB. By releasing the benchmark, models, and training pipeline, we provide both a practical solution for real-world medical retrieval systems and a foundation for future research in domain-specific embedding learning. Our future work will include exploring more effective strategies for asymmetric alignment.

Limitations

Our work presents a novel asymmetric alignment framework, yet it comes with certain limitations. First, regarding the architecture, the performance of our query encoder is inherently upper-bounded by the representational quality of the document encoder. Second, our work is primarily concentrated on the Chinese medical domain. While this setting presents significant challenges due to specialized terminology, the generalizability of our strategy to other languages or broader open-domain retrieval tasks remains to be verified. Finally, concerning the CMedTEB benchmark, although it utilizes a sophisticated multi-LLM pipeline for data construction, the annotation accuracy remains dependent on the capabilities of LLMs. However, this limitation may be mitigated by the advancement of LLMs.

Ethical considerations

This research has been approved by the National Technology Ethics (Review) Committee, which is an IRB-equivalent ethics committee. We strictly adhered to ethical guidelines regarding data collection and privacy. User queries were sourced from participants who explicitly consented to a user experience improvement program for non-commercial research. Medical documents were crawled from publicly accessible, non-paywalled websites (e.g., XunYiWenYao[3]) in compliance with robots.txt protocols and applicable copyright regulations. Both queries and documents were strictly anonymized, and details on the anonymization process are provided in Appendix A.3. We emphasize that these resources are for research purposes only and require rigorous validation before clinical deployment. Expert re-annotation was performed by clinicians from a partner tertiary hospital, compensated at institution-approved rates via official project budgets. CMedTEB is released under a CC BY-NC-SA 4.0 license, with model cards explicitly disclaiming diagnostic utility to ensure strict non-commercial, research-only usage.

[3] https://www.xywy.com/

References

Arian Askari, Emmanouil Stergiadis, Ilya Gusev, and Moran Beladev. 2025. Hotelmatch-llm: Joint multi-task training of small and large language models for efficient multimodal hotel retrieval. *arXiv preprint arXiv:2506.07296*.

Daniel Campos, Alessandro Magnani, and ChengXiang Zhai. 2023. Quick dense retrievers consume kale: Post training kullback leibler alignment of embeddings for asymmetrical dual encoders. *arXiv preprint arXiv:2304.01016*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Jianlyu Chen, Nan Wang, Chaofan Li, Bo Wang, Shitao Xiao, Han Xiao, Hao Liao, Defu Lian, and Zheng Liu. 2024b. Air-bench: Automated heterogeneous information retrieval benchmark. *arXiv preprint arXiv:2412.13102*.

Yongqi Fan, Nan Wang, Kui Xue, Jingping Liu, and Tong Ruan. 2025. Medeureka: A medical domain benchmark for multi-granularity and multi-data-type embedding-based retrieval. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2825–2851.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, and 1 others. 2025. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Suyuan Huang, Chao Zhang, Yuanyuan Wu, Haoxin Zhang, Yuan Wang, Maolin Wang, Shaosheng Cao, Tong Xu, Xiangyu Zhao, Zengchang Qin, and 1 others. 2024. Scalingnote: Scaling up retrievers with large language models for real-world dense retrieval. *arXiv preprint arXiv:2411.15766*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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*.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and 1 others. 2022. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024a. Making text embedders few-shot learners. *arXiv preprint arXiv:2409.15700*.

Lei Li, Xiangxu Zhang, Xiao Zhou, and Zheng Liu. 2025. AutoMIR: Effective zero-shot medical information retrieval without relevance labels. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24028–24047, Suzhou, China. Association for Computational Linguistics.

Shiyu Li, Yang Tang, Shizhe Chen, and Xi Chen. 2024b. Conan-embedding: General text embedding with more and better negative samples. *arXiv preprint arXiv:2408.15710*.

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*.

Ailiang Lin, Zhuoyun Li, Kotaro Funakoshi, and Manabu Okumura. 2025. Causal2vec: Improving decoder-only llms as versatile embedding models. *arXiv preprint arXiv:2507.23386*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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 *SIGIR*, pages 3046–3056. ACM.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking. *arXiv preprint arXiv:2110.07367*.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1930–1940.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Yuxuan Wang and Hong Lyu. 2023. Query encoder distillation via embedding alignment is a strong baseline method to boost dense retriever online efficiency. *arXiv preprint arXiv:2306.11550*.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. Retromae: Pre-training retrieval-oriented language models via masked auto-encoder. *arXiv preprint arXiv:2205.12035*.

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.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May Dongmei Wang, Joyce C. 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, Miami, Florida, USA. Association for Computational Linguistics.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. 2026. Stable-rag: Mitigating retrieval-permutation-induced hallucinations in retrieval-augmented generation. *arXiv preprint arXiv:2601.02993*.

S. Zhang, X. Zhang, H. Wang, L. Guo, and S. Liu. 2018. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071.

Sheng Zhang, Xin Zhang, Hui Wang, Jiajun Cheng, Pei Li, and Zhaoyun Ding. 2017. Chinese medical question answer matching using end-to-end character-level multi-scale cnns. *Applied Sciences*, 7(8):767.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Details of CMedTEB Benchmark

A.1 Instruction

Table 6 presents the instruction used on CMedTEB benchmarks.

A.2 Statistics

For detailed statistics of the CMedTEB datasets, please refer to Tables 7. To measure sequence lengths, we utilize the `tiktoken` tokenizer with the `cl100k_base` encoding scheme to count tokens for queries and corpus documents.

A.3 Anonymization steps of CMedTEB.

All user queries and web documents were processed as follows. 1) Automated PII (personally identifiable information) Detection: We deployed an offline, locally hosted large language model to detect and mask potential PII, including names, locations, phone numbers, and ID numbers. 2) Rule-based Validation: After initial masking, we applied a rule-based validation module to scan residual digits, and keywords. 3) Human Checks: 1% of anonymized data were checked by human, and no re-identifiable content found.

A.4 CMedTEB Case Examples

We present representative cases from CMedTEB tasks in Table 8, Table 9 and Table 10.

A.5 Zero-shot results on CMedTEB.

Table 11 presents the full zero-shot performance of all evaluated models across individual CMedTEB tasks. Results show significant performance gaps between general-domain embedders and the medical-specific retrieval challenge. Note that most of baselines have already trained on CMedQA train dataset before.

A.6 Dataset Comparative Analysis

We compare CMedTEB with CMIRB and MedEureka (Fan et al., 2025) across four key dimensions. **False Negative Handling:** CMIRB aggregates legacy datasets with annotation sparsity; MedEureka verifies original pairs but does not systematically retrieve other candidates. CMedTEB mitigates this by retrieving top candidates using multiple retrievers and re-annotating the entire pool. **Label Construction:** CMIRB uses ChatGPT for filtering; MedEureka employs GPT-4o with $\sim$10% human correction. CMedTEB implements Multi-LLM Voting followed by Clinical Ex-

Task Name	Instruction Template
CMedQAv1-reranking	Based on a Chinese medical question, evaluate and rank the medical information that provide answers to the question.
CMedQAv2-reranking	Based on a Chinese medical question, evaluate and rank the medical information that provide answers to the question.
CMedTEB-Retrieval	Given a Chinese medical question, retrieve medical documents that answer the question.
CMedTEB-Rerank	Based on a Chinese medical question, evaluate and rank the medical information that provide answers to the question.
CMedTEB-STS	Retrieve semantically similar text.

Table 6: Instruction used on CMedTEB benchmarks

Split	Dataset	Type	# Queries	Avg. $\|Q\|$	# Corpus / Cand.	Avg. $\|D\|$	Avg. Pos / Pos. Ratio
Train	Retrieval / Rerank	Mixed	20,000	21.24	229,457	470.90	-
	CMedQA	Rerank	50,000	66.76	196,902	134.22	-
	STS	STS	10,000	23.52	24,906	29.95	-
Test	CMedTEB-Retrieval	Retr.	734	20.68	229,457	470.90	8.43
	CMedTEB-Rerank	Rerank	1,128	18.52	27.83	502.75	7.83
	CMedQA-v1	Rerank	1,000	75.58	100.00	143.03	1.93
	CMedQA-v2	Rerank	1,000	66.98	100.00	135.99	1.91
	CMedTEB-STS	STS	5,000	35.45	-	-	47.92%

Table 7: **Detailed Statistics of CMedTEB Datasets.** We aggregate statistics for both Training and Test splits across Retrieval, Reranking, and STS tasks. *Avg.* $|Q|$ and *Avg.* $|D|$ denote the average lengths of queries and documents, respectively. For STS, the last column indicates the ratio of positive pairs.

pert Verification, achieving Fleiss' Kappa of 0.731 and 93.3% expert agreement. **Query Authenticity:** MedEureka relies partially on LLM-generated queries, while CMedTEB derives from real-world user logs preserving authentic distributions. **Task Scope:** CMIRB and MedEureka focus on Retrieval only; CMedTEB extends to Reranking and STS.

A.7 Validation of Consensus Annotation Pipeline

Unlike general domain retrieval, medical tasks demand high precision due to complex knowledge structures. To validate our strict consensus pipeline, we analyzed 100 random samples where three LLMs (GPT-4o, Kimi, Claude-3.5-Sonnet) disagreed. As shown in Table 12, 50% involve boundary relevance with multiple ambiguous intents, 38% suffer from LLM hallucinations on complex terminology, and 12% contain low-quality queries/documents. Including these would compromise reliability by increasing false negatives rather than contributing valid difficulty.

Consensus does not imply triviality. To verify CMedTEB retains inherent complexity, we sampled 100 queries and employed GPT-4o to annotate difficulty levels. As shown in Table 13, only 26% can be solved by lexical matching, while 53% require semantic understanding and 21% need deep inference involving medical knowledge or multi-hop logic. Furthermore, strong baselines exhibit a sharp performance drop from 85.15% on CMedQA to 57.85% on CMedTEB's new tasks (Section 3.2), confirming our pipeline filters ambiguity while preserving rigorous evaluation difficulty.

B Training Data Construction Details

B.1 Data Construction

Data Diversification. We apply diversification for query and corpus independently. We first initialized a vector index seeded with 5,000 documents encoded by `gte-multilingual-base`. For each new candidate x (query or document), we retrieve top-k neighbors and discard x if more than n neighbors exceed similarity threshold t; otherwise we insert x. This is applied separately to queries and corpus, preserving diversity while removing near-duplicates. We summarize the key parameters used during Data Diversification in Table 14, where k represents for top-k retrieved relevant candidates from vector index, t for similarity score threshold, and n for maximum number of related documents.

LLM annotation. For each diversified query q, we retrieve top-50 candidates from the diversified corpus and have GPT-4o assign a 5-point relevance score. From scored pools, we select positives and

Query 肾结石如何判断是酸性还是碱性结石？ *How to ascertain whether a renal calculus is acidic or alkaline in composition?*
Positive example 咋知道肾结石是有酸性碱性引起 病情分析：一般通过尿检判断肾结石是酸性的还是碱性的，可以到本地正规医院做尿液，酸碱度检查也可以观察一下pH值的变化，然后再明确一下尿液的酸碱度。如果怀疑身体有肾结石的症状，可以到正规医院做影像学检查检查一下大小。如果结石比较大的话，一定要及时到医院做激光碎石治疗。 *How can I tell whether a kidney stone is related to acidic or alkaline urine? Clinical assessment: In general, urinalysis is used to determine whether a renal calculus is associated with acidic or alkaline urine. You can have a urine pH test at a qualified local hospital and monitor the pH value to establish urinary acidity or alkalinity. If kidney stone symptoms are suspected, undergo imaging studies to assess the stone size. If the calculus is relatively large, timely laser lithotripsy is recommended.*

Table 8: CMedTEB-Retrieval example.

Query 耳鸣需要吃什么药？ *Which medications are indicated for tinnitus?*
Positive example 耳鸣的药有哪些 病情分析：耳鸣常用的药物有，1.盐酸氟桂利嗪胶囊、尼莫地平等，用于改善耳蜗的供血，扩张耳蜗血管。2.三磷酸干、辅酶A、甲钴胺等，用于改善耳道的代谢功能，可以促进耳部的新陈代谢，清理耳道杂质。3.卡马西平、路硝西泮等，用于抗惊厥，能够缓解耳朵受到刺激造成的耳鸣。4. 抗生素、红霉素、万古霉素等，这些药物含有非类固醇消炎药物，可以给耳道涂抹起到消炎的作用，以此来缓解耳鸣。 *What medications are available for tinnitus? Clinical assessment: Commonly used drugs include flunarizine hydrochloride capsules and nimodipine to improve cochlear perfusion by dilating cochlear vessels; adenosine triphosphate (ATP), coenzyme A, and mecobalamin (methylcobalamin) to enhance metabolic function of the auditory pathway, promote aural metabolism, and help clear debris from the ear canal; carbamazepine and clonazepam as anticonvulsants to relieve tinnitus triggered by neural irritation; and antibiotics such as erythromycin and vancomycin, as well as nonsteroidal anti-inflammatory agents, which can be applied to the ear canal for anti-inflammatory effects to help alleviate tinnitus.*
Negative example 吃补肾的药怎么耳鸣呢 病情分析：患者是由于肾阴亏虚而引起的上火症状，进而导致患者出现耳鸣。首先，患者应该服用一些滋阴补肾的药物来进行补肾，比如六味地黄丸或者知柏地黄丸。等到患者的肾虚得到一定的恢复之后，耳鸣的症状也会逐渐的消失。另外，患者可以搭配服用一些清热泻火的药物来进行治疗。 *Why would taking kidney-tonifying medicine lead to tinnitus? Clinical assessment: From a traditional Chinese medicine perspective, the patient's tinnitus is due to kidney-yin deficiency with endogenous heat, which precipitates tinnitus. It is advisable to use yin-nourishing, kidney-tonifying formulas such as Liuwei Dihuang Wan or Zhibai Dihuang Wan. As the kidney deficiency improves, the tinnitus should gradually resolve. In addition, heat-clearing and fire-purging agents can be used concomitantly.*

Table 9: CMedTEB-Rerank example.

negatives to form triples, yielding 500K fine-tuning instances of triplets $\mathcal{T} = \{(q_i, \mathcal{P}_i, \mathcal{N}_i)\}$, where $\mathcal{P}_i$ is a list sampled from positives P_i and $\mathcal{N}_i$ is a list sampled from negatives N_i.

B.2 Ablation studies on Data Diversification

To evaluate the effectiveness of our diversity-aware data curation pipeline, we conduct an ablation study on the role of query and document-side diversification on Medical-Embedder-base. All configurations use the same amount of training data. As shown in Table 15, the full setting achieves the best performance, demonstrating that both query and document diversification are essential: the former ensures broad topic coverage, while the latter improves the reliability and difficulty of negative samples. This validates the importance of our diversity-aware curation strategy in building high-quality medical retrieval datasets.

C Implementation Details

C.1 Independent Initialization

C.1.1 Query Encoder Training

RetroMAE Pretrain. We first adopt RetroMAE (Xiao et al., 2022) pretrain, which mask inputs differently in the encoder and a lightweight decoder; the encoder outputs sentence embeddings and the decoder reconstructs the original text via masked language modeling. This stage leverages a 60M unsupervised Medical Q&A corpus.

Sentence1 碳酸氢钠片是否会引起头皮痒 *Do sodium bicarbonate tablets cause scalp itching?*
Sentence2 服用小苏打片是否可能导致头皮发痒? *Could taking baking soda tablets lead to an itchy scalp?*

Table 10: CMedTEB-STS example.

Model	Param.	CMedv1 MAP@10	CMedv2 MAP@10	Avg CMed	Retr. nDCG@10	Rerank MAP@10	STS Pearson	Avg. New
gte-multilingual-base	305M	86.11	87.40	86.76	47.80	61.51	72.39	60.57
gte-base-zh	102M	86.79	87.20	86.99	44.18	58.40	75.07	59.22
gte-large-zh	326M	86.09	86.46	86.28	29.75	53.70	68.02	50.49
gte-Qwen2-1.5B-instruct	1.78B	88.16	88.12	88.14	45.14	58.99	76.81	60.31
gte-Qwen2-7B-instruct	7.61B	88.20	89.31	88.76	40.94	61.07	72.67	58.23
bge-small-zh-v1.5	24M	77.40	79.86	78.63	35.22	55.39	57.87	49.49
bge-base-zh-v1.5	102M	80.47	84.88	82.68	33.11	53.56	67.45	51.37
bge-large-zh-v1.5	326M	83.45	85.44	84.45	43.05	58.31	71.90	57.75
bge-m3	568M	77.71	79.19	78.45	41.14	57.68	63.67	54.16
Conan-embedding-v1	326M	**91.39**	**89.72**	**90.56**	41.60	61.89	72.86	58.78
stella-base-zh-v3-1792d	102M	88.35	89.06	88.71	45.77	60.43	74.96	60.39
Qwen3-Embedding-0.6B	596M	80.06	81.35	80.71	47.54	64.51	68.31	60.12
Qwen3-Embedding-4B	4.02B	84.43	85.06	84.75	50.14	**66.67**	**76.49**	64.43
Qwen3-Embedding-8B	7.57B	86.13	86.39	86.26	**51.15**	66.31	76.09	**64.52**
Average performance		84.62	85.67	85.15	42.61	59.89	71.04	57.85
Spearman Rank Correlation Coefficient (P-value)								0.354 (0.215)

Table 11: Zero-shot results on CMedTEB (%). Best results in **bold**.

Unsupervised Pretrain. We perform contrastive unsupervised pretrain using InfoNCE loss (Oord et al., 2018):

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\mathbf{q}^\top \mathbf{d}^+/\tau)}{\sum_{\mathbf{d}\in\mathcal{D}} \exp(\mathbf{q}^\top \mathbf{d}/\tau)},$$

where $\mathbf{q}, \mathbf{d}^+$ are embeddings of a matched (query, document) pair, $\mathcal{D}$ contains one positive and $|\mathcal{D}|-1$ negatives, and τ is a learnable temperature. We use the same unsupervised medical Q&A corpus for RetroMAE pretraining, treating title–content pairs as positives and other documents within the same batch as in-batch negatives.

Supervised Finetuning. The final stage finetunes the encoder on high quality fine-tuning datasets described in Section 4.1 together with the training splits of **CMedTEB** (retrieval, reranking, CMedQA, and Synonym STS) using the InfoNCE loss.

C.1.2 Document Encoder Training

We fine-tune Qwen3-4B and Qwen3-8B and apply LoRA with rank=32, $\alpha = 64$. We adopt Matryoshka Representation Learning (MRL) (Kusupati et al., 2022), whose training objective aggregates the contrastive loss across this predefined set of dimensions. Specifically, the final loss is the average of the InfoNCE losses computed at each target dimension:

$$\mathcal{L}_{\text{MRL}} = \frac{1}{|M|} \sum_{m\in M} \mathcal{L}^{(m)}_{\text{InfoNCE}}, \tag{5}$$

where M is the set of nested dimensions and $\mathcal{L}^{(m)}_{\text{InfoNCE}}$ is the standard InfoNCE loss calculated using embeddings truncated to the first m dimensions.

C.1.3 Impact of Pretraining on Query Encoder

We evaluate the impact of pretraining on the query encoder. As shown in Table 16, combining RetroMAE and unsupervised domain pretraining achieves the best performance (54.16), outperforming ablated variants. This confirms that multi-stage pretraining enhances the encoder's performance in medical retrieval.

C.2 Implementation of Efficient Baselines.

To ensure a fair comparison, we re-implemented representative efficient retrieval methods within our decoder-only framework. For KALE(Campos et al., 2023) and Wang and Lyu (2023), we initialized the

Reason	Description	Ratio (%)
Boundary relevance	Multiple or ambiguous intents make relevance labeling inherently difficult.	50
LLM hallucination	Limited medical knowledge causes misjudgment of complex terms or relationships.	38
Noise	Low-quality or incomplete queries/documents mislead into incorrect labels.	12

Table 12: Analysis of LLM disagreement reasons (N=100).

Level	Description	Ratio (%)
L1: Lexical	High keyword overlap between query and document.	26
L2: Semantic	Conceptual match without keyword overlap; requires dense retrieval.	53
L3: Hard	Deep inference requiring medical knowledge, cause-effect, or multi-hop logic.	21

Table 13: Difficulty distribution of CMedTEB samples (N=100).

Parameter	Query	Document
k (retrieved candidates)	5	5
t (score threshold)	0.85	0.78
n (maximum number)	1	1

Table 14: Key parameters used during data generation.

Diversification Setting	Retrieval nDCG@10	Rerank MAP@10	Avg
w/o query, w/o doc	51.17	68.74	59.96
w/ query, w/o doc	<u>52.23</u>	<u>68.98</u>	<u>60.61</u>
w/ query, w/ doc	**54.16**	**69.63**	**61.90**

Table 15: Impact of query and document diversification on retrieval performance.

Training Strategy	Retrieval nDCG@10
Finetune only	52.88
RetroMAE + Finetune	53.21
RetroMAE + Unsup + Finetune	**54.16**

Table 16: Ablation study on pretraining strategies for Medical-Embedder-Base. Combining RetroMAE and unsupervised domain pretraining leads to the best retrieval performance.

query encoder using the first 3 layers of Qwen3-4B (approximately 303M parameters excluding the LM head) with an embedding dimension of 2560. For ScalingNote(Huang et al., 2024), we employed Medical-Embedder-base as the query encoder. For the Distill baseline (Ren et al., 2021), following standard protocols, we utilized scores from Medical-Embedder-4B as soft labels and optimized the student model (Medical-Embedder-base) using a combination of KL-divergence and InfoNCE loss.

C.3 Training Details

For the query encoder, we use the final hidden state of the [CLS] token as the sentence embedding. For the document encoder, we append an [EOS] token to the input sequence and use its output hidden state as the document embedding. The maximum input length for both queries and documents is set to 512 tokens.

We summarize the training configurations in Table 17. For memory efficiency, we enable gradient checkpointing and use DeepSpeed Stage 0. For models up to 4B parameters, we train in `fp16`, while for the 8B model we switch to `bf16` to ensure stability. All document encoders are fine-tuned with LoRA (rank 32, $\alpha = 64$). For fair comparison, all symmetric baseline models are fine-tuned for 2 epochs, matching the total exposure of our asymmetric models, which observe the fine-tuning data once during independent initialization and once again during joint fine-tuning.

In our asymmetric architecture, both query and document encoders are first initialized by one epoch of fine-tuning. For Stage I, we align query and document embeddings using 8.4M pairs of query alignment data for one epoch. For Stage II, we further fine-tune for one epoch to ensure comparability with other baselines. We apply the same learning rate (1×10^{-4}) to both query and document encoders, as we observed that asymmetric learning rates led to performance degradation.

Hyperparameter	General Training			Asymmetric Training	
	RetroMAE	Unsup.	Fine-tuning	Stage I	Stage II
Peak Learning Rate	2e-4	1e-4	1e-4	1e-4	1e-4
Warmup Ratio	0.0	0.05	0.05	0.05	0.05
Global Batch Size	384	19,200	640	2,560	640
Epochs	3	3	2	1	1

Table 17: **Detailed Training Hyperparameters.** We list the configurations for both the general pre-training stages (RetroMAE, Unsupervised) and our proposed asymmetric alignment stages (Stage I & II). All stages utilize the AdamW optimizer with a linear decay scheduler.

D Cost-Benefit Analysis

We address the cost-benefit concern by providing detailed offline metrics. In industrial retrieval systems, document indices are updated much less frequently than queries are processed. Therefore, our architecture strategically leverages powerful offline models to boost representation quality without penalizing the critical online user experience.

Offline Efficiency. As shown in Table 18, processing 1 million documents with our largest 8B model takes only 3.77 hours on a single A100 GPU. Given the low update frequency (monthly or weekly) typical in production medical retrieval systems, this one-time offline cost is negligible compared to the continuous benefits in online retrieval accuracy. Although `gte-multilingual-base` is faster (0.21h), the 8B model's overhead is entirely acceptable for periodic updates. Note that index memory is constant as MRL aligns the document embedding dimension with the query encoder regardless of model size.

Model	Offline Params	Time (h)
gte-multilingual-base	305M	0.21
CARE-0.3B-4B	4B	2.95
CARE-0.3B-8B	8B	3.77

Table 18: Offline Efficiency Analysis. Encoding time for 1M documents on A100 80GB. Index memory constant (2,929.69 MB) due to MRL dimension alignment.

Online Advantage. Crucially, the primary bottleneck in real-world applications lies in online inference latency rather than offline costs. While offline costs increase marginally, CARE-0.3B-8B achieves a superior average score of 78.94 with only 0.3B online parameters. It surpasses the fine-tuned `gte-Qwen2-1.5B` by +1.33 with 9× higher QPS. This decoupled design allows independent scaling: document encoders can be upgraded for accuracy gains without any impact on online serving latency.

Model	Stage-1 (h)	Stage-2 (h)	Finetune Baseline (h)
CARE-0.3B-4B	316	260	280
CARE-0.3B-8B	608	529	544

Table 19: Training Cost Analysis (GPU Hours on A100-40G).

Training Cost. We present a direct comparison of GPU hours in Table 19. Our two-stage training strategy is highly efficient. For example, Stage 2 for the 4B model requires 260 hours, comparable to standard fine-tuning of the document encoder alone (280 hours). While Stage 1 represents an initial alignment cost, it establishes a reusable foundation. This confirms that our approach improves performance without introducing prohibitive computational costs compared to conventional fine-tuning paradigms.

E Quality Analysis of Open-Source Benchmarks

We investigate the false negative issue in CmedqaRetrieval and MedicalRetrieval. For each query, we use gte-multilingual-base to retrieve the top-50 candidate documents and re-annotate them using GPT-4o under a 5-point relevance scale with prompt in Table 23.

Results in Table 20 suggest that a large number of retrieved documents, though unlabeled in the original datasets, are judged as relevant by the LLM. Table 21 and Table 22 shows several examples of false negatives and false positives, together with the topic intensity phenomenon in medical domain that certain diseases or drugs generate a large volume of semantically similar queries and answers. This indicates annotation incompleteness in existing benchmarks. These findings raise concerns about the validity of current benchmarks for

a reliable evaluation of medical retrieval capability.

F Annotation Prompts

Table 23 and Table 24 present the prompts templates for CMedTEB construction. Table 25 shows prompt for our training data annotation.

Benchmark	Orig. Pos.	LLM-Labeled Pos.	False Positive	False Negative
MedicalRetrieval	0.81	9.11	0.26	**8.56**
CmedqaRetrieval	1.42	19.94	0.46	**18.98**

Table 20: LLM re-annotation on open-source medical retrieval benchmarks. To aid interpretation, we assume the LLM labels are pseudo–ground truth. We measure the average number of positive documents per query in the original dataset vs. LLM-labeled data, and identify false positives and false negative.

Query 查出说是贫血孩子老烧还有咳嗽 *The child was diagnosed with anemia and has been running a fever with coughing.*
False Positive (Labeled as positive, but not mention fever and coughing.) 如果是检查有贫血，可以结合贫血的类型和严重的程度，根据检查结果进行治疗的考虑即可。 *If anemia is detected, treatment can be determined based on the type and severity of anemia, as indicated by the test results.*
False Negative (Labeled as negative, but annotated as positive by LLM) 如果只是简单地烧咳嗽，等相应症状，是无需，担心的，但是贫血的原因必须地须要查清楚，一般情况下评选分为营养不良性贫血，还有其他病理性贫血。所以建议到医院进行系统检查，看到底是？出现的什么方面的贫血。然后对症治疗。 *If it is just a simple cough or other corresponding symptoms, there is no need to worry. However, the cause of anemia must be clarified. Generally, anemia is categorized into nutritional deficiency anemia and other pathological anemias. Therefore, it is recommended to go to the hospital for a comprehensive examination to determine the specific type of anemia and then treat it accordingly.*

Table 21: An example of false negatives in CmedqaRetrieval.

Query
感冒发烧一起来怎么办
What should I do if I have a cold and fever at the same time?

Positive
你好，应该是流行感冒吧，典型流感，急起高热，全身疼痛，显著乏力，呼吸道症状较轻。颜面潮红，眼结膜外眦充血，咽充血，软腭上有滤泡。具体吃药建议咨询医生。 期间注意：多饮开水，多食新鲜的蔬菜、瓜果。饮食宜清淡，多食易消化，且富于营养及富含维生素的食物，如稀饭、豆浆、面条、水果等。窗子经常开下来通通风。 一些简单食疗： 1.生姜20克、大蒜头5～6瓣、红糖适量。 用法：水煎服。 主治：流行性感冒初起，头痛，怕冷发热，无汗，伴有恶心者。 说明：流行性感冒是一种急性呼吸道传染病，表现为急起高热，全身疼痛，乏力，呼吸道症状如咽干喉痒，干咳等，胃肠道症状如恶心呕吐、腹泻水样便等。 2.冬瓜粥 粳米50克。将冬瓜适量切成小块，与米同煮，粥熟即可食用。此粥对病毒型流行性感冒病人有效。 3.葱白500克、大蒜250克。 用法：上药切碎加水2000毫升煎煮。日服3次，每次250毫升，连服2～3天。 愿早日康复！
Hello, it sounds like you have the flu. Typical symptoms include sudden high fever, body aches, significant fatigue, and mild respiratory symptoms. You may also have facial flushing, conjunctival injection, pharyngeal congestion, and follicles on the soft palate. Please consult a doctor for specific medication advice. During this time, drink plenty of water and eat more fresh vegetables and fruits. Keep your diet light and easy to digest, focusing on nutritious and vitamin-rich foods like porridge, soy milk, noodles, and fruits. Ventilate your room regularly by opening windows. Here are some simple home remedies: 1. 20 grams of fresh ginger, 5–6 cloves of garlic, and an appropriate amount of brown sugar. Decoct in water and take orally. This is for the early stages of influenza with headache, chills, fever, no sweating, and nausea. Influenza is an acute respiratory infectious disease characterized by sudden high fever, body aches, fatigue, and respiratory symptoms like sore throat and dry cough. It may also cause gastrointestinal symptoms like nausea, vomiting, and watery diarrhea. 2. Winter melon porridge: 50 grams of japonica rice. Cut an appropriate amount of winter melon into small pieces and cook with rice. This porridge is effective for patients with viral influenza. 3. 500 grams of green onion whites and 250 grams of garlic. Chop the ingredients and decoct in 2000 milliliters of water. Take three times a day, 250 milliliters each time, for 2–3 days. Hope you recover soon!

False Negative 1 (Labeled as negative, but annotated as positive by LLM)
建议口服抗病毒药物和感冒冲剂试试.有炎症还是应该加上抗生素。口服药物不见效的，建议输液治疗为好。在当地医生指导下使用。发烧用退热贴
It is suggested to try oral antiviral medications and cold granules. If there is an infection, antibiotics should be added. If oral medications are not effective, it is recommended to consider intravenous therapy. This should be done under the guidance of a local doctor. For fever, you can use fever patches.

False Negative 2
感冒发烧是临床上最常见的疾病和症状，具体吃药要根据具体的症表现以及病人身体状况而定。如果是儿童出现感冒发烧的情况一般选择以单药为主，出现发烧时主要可选择对乙酰氨基酚或者布洛芬口服液来进行治疗；如果还有其他的症状，比如出现鼻塞流涕，可以使用氨咖黄敏颗粒。如果是成人感冒发烧，一般多选择复合剂型，比如酚麻美敏片或者复方氨酚烷胺等。如果持续发烧不退，要及时完善血液分析和胸片检查排除并发肺炎的可能。
A cold with fever is one of the most common illnesses and symptoms clinically. The specific medication should be determined based on the specific symptoms and the patient's physical condition. For children with a cold and fever, monotherapy is usually chosen. For fever, acetaminophen or ibuprofen oral suspension can be used for treatment. If there are other symptoms, such as nasal congestion and runny nose, pheniramine and caffeine granules can be used. For adults with a cold and fever, compound formulations are generally preferred, such as phenylephrine, dextromethorphan, and acetaminophen tablets, or compound paracetamol and amantadine. If the fever persists, it is important to promptly complete blood tests and chest X-rays to rule out the possibility of pneumonia.

False Negative 3
你好，建议口服抗病毒药物和感冒冲剂试试.即使是病毒性感冒也容易继发细菌感染,所以最好还是应该加上抗生素口服.建议口服药物不见效的,建议输液抗炎治疗为好.因为还是输液血药浓度更高见效更快更好啊.有痰的加上鲜sd竹沥口服试试.发烧还需要适当加上额外的退烧药物.一般需要7-10天才能治愈的.最好还是看医生啊
Hello, it is suggested to try oral antiviral medications and cold granules. Even viral colds can easily lead to secondary bacterial infections, so it is better to add oral antibiotics. If oral medications are not effective, it is recommended to consider intravenous anti-inflammatory treatment, as it provides higher blood drug concentration and faster results. For those with phlegm, you can try adding fresh bamboo extract orally. Fever also requires the addition of extra antipyretic drugs. It usually takes 7-10 days to recover. It is best to see a doctor.

Table 22: An example of false negatives in MedicalRetrieval. This example shows the *topic intensity* phenomenon in medical domain: certain diseases or drugs generate a large volume of semantically similar queries and answers.

Prompt:
This is a medical information retrieval task: given a medical query (Query), retrieve documents (Passages)that can answer the question.

Given a medical query (Query) and {len(docs)} passages, your task is to rate the relevance between the Query and each Passage.

Relevance scoring criteria:
S: The subject (e.g., disease name, drug name, inquiry target) and inten of Query and Passage are fully consistent. The Passage can directly, completely, and correctly answer the Query.
A: The subject and intent of Query and Passage are consistent. The Passage contains content that candirectly and correctly answer the Query.
B: The subject of Query and Passage is consistent, but the intent differs.
The Passage cannot directly answer the Query, but it is useful for inference.
C: The subject of Query and Passage is related, but the intent is inconsistent.
It can only partially match the Query from the text, but cannot answer the Query.
D: The subject and intent of Query and Passage are unrelated. Cannot answer the Query.

Notes:
1. Query and Passage are independent; there is no contextual relationship.
Do not infer or supplement the subject/intent of Query based on Passage.
2. If the Query is low-quality (e.g., missing subject, like "How to treat this disease?"), the maximum relevance score for all Passages should not exceed B.
3. All Passages are independent; they are randomly ordered and have no contextual relationship.

Output format:
Your output must be a JSON object, containing only the required fields. The format is as follows:
{ "Passage-0": "A", "Passage-1": "C", ... }
Query and Passages are as follows:
- Query: {query}
{passages}
...
Remember: do not output any other content or explanation.
Your output must be only a JSON object with the required fields. Output:

Table 23: Prompt template for CMedTEB Retrieval and Rerank tasks

Medical Query Rewriting Sample Generation (Positive and Negative Examples)

Task Objective: Your task is to generate one positive example and two negative examples based on a given original medical query and a set of synonyms.
You will receive a JSON object containing the following fields:
"origin": "Original medical term",
"replace": "Synonym medical term for replacement",
"query_pairs": { "origin": "Query sentence using the original term", "replace": "Query sentence using the replaced term" }

Generation Rules:
1. General Quality Standards (applicable to all outputs):
- Professional Expression: Use professional, fluent, and natural medical language.
- Medical Accuracy: Content must conform to medical knowledge and avoid ambiguity.
- Format Requirement: All outputs must be complete, fluent interrogative sentences.

2. Specific Sample Requirements:
- positive (Positive Example):
- Task: Optimize and rewrite the second query in query_pairs (the one containing the "replace" term).
- Intent: Must preserve the exact same intent as the original query.
- Terminology: Must use the term specified in the "replace" field.
- Constraint: Rewritten query length must be within $\pm 30\%$ of the original query's length.
- negative-1 (Negative Example 1):
- Task: Create a new query based on the topic of the original query, similar but distinctly different.
- Terminology: Must use the term specified in the "replace" field.
- Intent: Significantly alter the intent of the original query.
- negative-2 (Negative Example 2):
- Task: Create a new query based on the topic of the original query, similar but distinctly different.
- Terminology: Must use the term specified in the "origin" field.
- Intent: Significantly alter the intent of the original query (same rule as negative-1).

Output Format: Must be a JSON object containing only the following three fields. Do not add any extra explanations or comments.

Input: {input}
Output:

Table 24: Prompt template for CMedTEB STS tasks

This is a retrieval task in the Chinese medical domain, requiring classification of positive and negative documents based on the user's medical query and search engine returned documents.
You will receive data containing the following fields:
"query": User input in the medical domain.
"documents": A candidate document set containing multiple documents, some relevant and some irrelevant — capable or incapable of answering the query.
Your task is to identify "positive_document" and "negative_document" from the provided documents.
"positive_document": Relevant to the query; the document contains sentences that can answer the query.
"negative_document": Either relevant or irrelevant to the query, but the document content does NOT contain any sentence that can answer the query.
Please follow these guidelines:
- Both "positive_document" and "negative_document" must come from the candidate document set.
- "positive_document" and "negative_document" are mutually exclusive — no document overlap is allowed.
Output Requirements:
Example: {out_exam}.
Your output must always be ONLY a JSON object, containing ONLY document indices (e.g., "doc-1"). Do NOT include document content, explanations, or any additional text.
Input Data Format:
{"positive_document":["doc-1","doc-2"], "negative_document":["doc-3","doc-4"]}
Classify the documents in the input data according to the above rules, ensuring the output strictly follows the required format.
Output:

Table 25: Prompt template for training data annotation