Huatuo-26M, a Large-scale Chinese Medical QA Dataset

Xidong Wang$^{\alpha,\gamma*}$**, Shunian Chen**$^{\alpha,\gamma*}$**, Jianquan Li**$^{\gamma*}$ **, Yuxuan Zhu**$^{\gamma}$
Xiangbo Wu$^{\gamma}$**, Zhiyi Zhang**$^{\gamma}$**, Xiaolong Xu**$^{\gamma}$**, Junying Chen**$^{\alpha,\gamma}$**,**
Jie Fu$^{\beta}$**,Xiang Wan**$^{\alpha}$**, Anningzhe Gao**$^{\alpha}$**, Benyou Wang**$^{\alpha,\gamma}$ 🖂
$^{\alpha}$ Shenzhen Research Institute of Big Data
$^{\beta}$ Shanghai AI Lab
$^{\gamma}$ The Chinese University of Hong Kong, Shenzhen
🖂 wangbenyou@cuhk.edu.cn

Abstract

Large Language Models infuse newfound vigor into the advancement of the medical domain, yet the scarcity of data poses a significant bottleneck hindering community progress. In this paper, we release the **largest** ever medical Question Answering (QA) dataset with **26 Million** QA pairs named Huatuo-26M. We benchmark many existing approaches in our dataset in terms of both retrieval and generation. We also experimentally show the benefit of the proposed dataset in many aspects: (i) it serves as a fine-tuning data for training medical Large Language Models (LLMs); (ii) it works as an external knowledge source for retrieval-augmented generation (RAG); (iii) it demonstrates transferability by enhancing zero-shot performance on other QA datasets; and (iv) it aids in training biomedical model as a pre-training corpus. Our empirical findings substantiate the dataset's utility in these domains, thereby confirming its significance as a resource in the medical QA landscape. Our code and data are publicly released at https://github.com/FreedomIntelligence/Huatuo-26M.

1 Introduction

Pre-trained language models have made great progress in Natural Language Processing (NLP) and largely improve natural language understanding and natural language generation. This inspires researchers to apply Pre-trained Languge Models (PLMs) for fields that are not considered the core playground of NLP, for example, medicine. However, the first *bottleneck* for medicine using PLMs is the *data*, like most other breakthroughs in artificial intelligence that starts with data collection.

As shown in Tab. 1, a publicly available large-scale medical question and answer dataset has yet to be established. To break the bottleneck, this work collects the largest medical Chinese QA dataset that also might enhance medical research. Note that there are 1.4B population speaking Chinese as their native language, and more importantly, the medical care for them (particularly the mainland of China) is generally far below the western counterpart (e.g., English-speaking and developed countries) [1].

Dataset We collect the largest medical QA dataset from various sources as below: (i) collect from an online medical consultation website; (ii) automatically extract from medical encyclopedias, and (iii) automatically extract from medical knowledge bases. After screening privacy-irrelevant information, text cleaning and deduplication, we obtain the largest Chinese medical question and answer dataset, containing **26 Million** QA pairs. As seen from Tab. 1, this dataset is two orders of magnitude larger than the existing QA datasets. We call this dataset 'Huatuo-26M' to commemorate the great Chinese physician named Hua Tuo, who lived around 200 AC.

Benchmark We benchmark classical methods in the field of retrieval: for sparse retrieval, we test the performance of BM25 (Robertson et al., 2009) and DeepCT (Dai and Callan, 2019), and for dense retrieval, we test the performance of DPR (Karpukhin et al., 2020). Meanwhile, we conduct benchmark evaluations of text generation, covering a series of autoregressive language models from GPT2 (Brown et al., 2020) and T5 (Raffel et al., 2020) to Baichuan2 (Yang et al., 2023) and ChatGLM3 (Zeng et al., 2023). The results suggest the task is still challenging, probably because the medical domain involves more expert knowledge than the general domain.

Applications To further show the usefulness of the collected dataset, we leverage it in four use cases: (i) As Fine-tuning Data for Medical LLMs;

[*] The first three authors contributed to this paper equally

[1] https://en.wikipedia.org/wiki/List_of_countries_by_quality_of_healthcare

Findings of the Association for Computational Linguistics: NAACL 2025, pages 3828–3848
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

Domain	Dataset	Lang	Domain	Source	#Q
Medical	LiveQA (Zhang et al., 2018)	English	Medical	U.S. National Institutes of Health (NIH)	0.7K
	MedHop (Welbl et al., 2018)	English	Medical	MEDLINE	2.5K
	BiQA (Lamurias et al., 2020)	English	Medical	Online Medical forum	7.4K
	HealthQA (Zhu et al., 2019)	English	Medical	Medical-services website	7.5K
	MASH-QA (Zhu et al., 2020)	English	Medical	Medical article website	35K
	MedQuAD (Ben Abacha and Demner-Fushman, 2019)	English	Medical	U.S. National Institutes of Health (NIH)	47K
	ChiMed (Tian et al., 2019)	Chinese	Medical	Online Medical forum	47K
	MedRedQA (Nguyen et al., 2023)	English	Medical	Health subreddit (AskDocs)	51K
	MedQA (Jin et al., 2020)	EN&CH	Medical	Medical Exam	60K
	webMedQA (He et al., 2019)	Chinese	Medical	Medical consultancy websites	63K
	CliCR (Šuster and Daelemans, 2018)	English	Medical	Clinical case reports	100K
	cMedQA2 (Zhang et al., 2018)	Chinese	Medical	Online Medical forum	108K
	MedMCQA (Pal et al., 2022)	English	Medical	Medical Exam	192K
	PubMedQA (Jin et al., 2019)	English	Medical	Online Medical forum	272K
	Huatuo-26M	**Chinese**	**Medical**	**Consultation records, Encyclopedia, KBs**	**26M**
General	TriviaQA (Joshi et al., 2017)	English	General	Trivia	96K
	HotpotQA (Yang et al., 2018)	English	General	Wikipedia	113K
	SQuAD (Rajpurkar et al., 2016)	English	General	Wikipedia	158K
	DuReader (He et al., 2017)	Chinese	General	Web search	200K
	Natural Questions (Kwiatkowski et al., 2019)	English	General	Wikipedia	323K
	MS MARCO (Nguyen et al., 2016)	English	General	Web search	1.0M
	CNN/Daily Mail (See et al., 2017)	English	General	News	1.3M
	PAQ (Lewis et al., 2021)	English	General	Wikipedia	65M

Table 1: Existing QA datasets. Huatuo-26M is currently the largest medical QA dataset.

(ii) As an External Knowledge Source for RAG; (iii) Transferability to other QA Datasets and (iv) As a Pre-training Corpus.

Contributions of this work are as follows: (i) We release the largest Chinese Medical QA dataset with **26,504,088** QA pairs; (ii) We benchmark some existing models for the proposed methods for both retrieval and generation; (iii) We explore some additional usage of our dataset, for example, fine-tuning medical LLMs, train as external knowledge for RAG, transfer for other QA datasets, and train as a pre-trained corpus.

2 Huatuo-26M

We collect a variety of medical knowledge texts from various sources and unify them in the form of medical question-and-answer pairs. The main resources include an online medical QA website, medical encyclopedias, and medical knowledge bases. See App. D for specific examples from different sources. Here we will introduce the details of data collection.

2.1 Dataset Creation

2.1.1 Online Medical Consultation Records

Data Sources We collect data from a website for medical consultation [2], consisting of many online consultation records by medical experts. We collect data entries that record basic information about doctors, including name, hospital and department, while personal information about patients is anonymous to ensure the traceability of answers and prevent leakage of patient information, getting 31,677,604 pairs.

Data Processing We then follow standard data processing procedures, using regular expressions to remove QA pairs containing special characters and employing hashing to eliminate duplicates. Finally, we get 25,341,578 QA pairs. We will make the extraction and processing scripts public to provide the community with a clear data processing and reproduction process.

2.1.2 Online Medical Encyclopedia

Data Sources We extract medical QA pairs from plain texts (e.g., medical encyclopedias and articles), including 8,699 encyclopedia entries for diseases and 2,736 encyclopedia entries for medicines on Chinese Wikipedia [3], as well as 226,432 high-quality medical articles.

Data Processing Each article consists of title-paragraph pairs. Titles in medicine articles may include usage, contraindications, and nutrition, while those about diseases may cover diagnosis, clinical features, and treatment methods. We remove titles appearing fewer than five times, resulting in 733 unique titles. Using these titles, we create templates to convert each title into a question answerable by the corresponding paragraph, with disease or drug

[2] Qianwen Health in https://51zyzy.com/

[3] zh.wikipedia.org/wiki/

	# Entity type	# Relation	# Entity	# Triplets
CPubMed-KG	-	40	1.7M	4.4M
39Health-KG	7	6	36.8K	210.0K
Xywy-KG	7	10	44.1K	294.1K

Table 2: Basic statistics of the three knowledge bases.

Composition	# Pairs	Len(Q)	Len(A)
Huatuo-26M Train	26,239,047	44.6	120.7
Huatuo-26M Test	265,041	44.6	120.6
Data source:			
Consultant records	25,341,578	46.0	117.3
Encyclopedias	364,066	11.5	540.4
Knowledge bases	798,444	15.8	35.9
All	26,504,088	44.6	120.7

Table 3: Basic statistics of Huatuo-26M.

(%)	Fluency	Relevance	Completeness	Professionalism
Poor	0.15	0.55	0.59	0.49
Medium	21.80	23.45	40.99	67.98
Good	78.05	76.01	58.42	31.53

Table 4: Data quality statistics from four aspects.

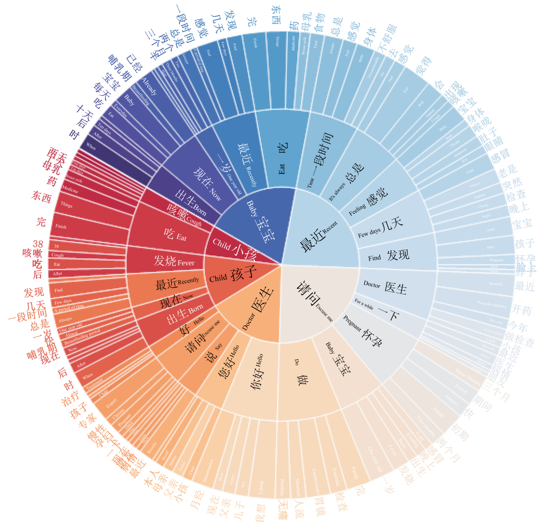

Figure 1: Distribution of questions. We present the relative distribution of these recurring problems and their subsequent distributions.

names as placeholders. See theApp. E for details.

2.1.3 Online Medical Knowledge Bases

Data Sources Some knowledge bases explicitly store well-organized knowledge, from which we extract medical QA pairs. We collect data from the following three medical knowledge bases: **CPubMed-KG** (Qingcai Chen, 2022) is a knowledge graph for Chinese medical literature, which is based on the large-scale medical literature data from the Chinese Medical Association; **39Health-KG** (Chen, 2018) and **Xywy-KG** (Chen, 2018) are two open source knowledge graphs. Basic information is shown in Tab. 2.

Data Processing We clean the three knowledge graphs by removing invalid characters and merged entities and relationships, resulting in 43 categories, each linked to a relationship or attribute. We then design templates to convert each category into a question, either querying an object entity based on a subject entity or an entity's attribute, with the object entity as the answer. This process yielded 798,444 QA pairs using the corresponding templates. See App. F for details.

2.2 Data Statistics and Analysis

The basic statistics of Huatuo-26M are shown in Tab. 3 and the examples can be found in Tab. 6. Most of the QA pairs are from online consultation records. The average length of the dataset questions is 44.6 and the average length of the answers is 120.7. Questions could be long (e.g. in consultant records) or short (in encyclopedias and knowledge bases). There exists both long answers (e.g., Encyclopedia) and short answers (e.g. consultant records and knowledge bases). We randomly take 1% QA pairs as the test set while others form the training set.

Colloquial Questions with Professional Answers As shown in Fig. 1 and examples in App. D, patient questions include characteristics and daily symptoms in realistic scenarios, while doctor answers are targeted and contextually coherent. We select 100 examples from each data source and asked three licensed physicians to evaluate the accuracy of the answers. The accuracy rates for the sources—online medical consultation, medical encyclopedia, and medical knowledge bases—are 71%, 88%, and 79%, respectively.

Quality Labeling We use data annotated by physicians and ChatGPT[4] to train classification models that generate labels based on Fluency, Relevance, Completeness, and Proficiency in medicine. The labels are categorized into three levels: Good, Medium, and Poor. Statistics are presented in Tab. 4. For more details, see App. I.

2.3 Data Licence, Privacy and Biases Issues

Data licence For pairs extracted from open-source online encyclopedias and knowledge bases,

[4] gpt-3.5-turbo-0125

Data source	Model	Recall @5	Recall @20	Recall @100	Recall @1000	MRR @10
Medical consultant records	BM25	4.91	6.99	10.37	17.97	3.82
	DeepCT	**7.60**	10.28	14.28	22.85	**6.06**
	DPR	6.79	**11.91**	**20.96**	**42.32**	4.52
Encyclopedias	BM25	4.58	8.71	17.82	39.91	3.10
	DeepCT	**20.33**	26.92	36.61	53.41	**16.25**
	DPR	16.01	**27.25**	**45.33**	**78.30**	11.20
Knowledge bases	BM25	0.52	1.02	1.82	3.51	0.38
	DeepCT	1.05	1.46	2.10	3.29	0.71
	DPR	**2.66**	**5.25**	**11.84**	**33.68**	**1.83**
ALL	BM25	4.77	6.83	10.21	17.84	3.71
	DeepCT	**7.58**	10.24	14.22	22.68	**6.04**
	DPR	6.79	**11.92**	**21.02**	**42.55**	4.53

Table 5: *Retrieval*-based benchmark for Huatuo-26M. Results are separated for different data sources.

we provide full texts. In contrast, for online consultation records, we release only the question and its URL, without the full texts. We are open to sharing the full dataset with researchers under the condition that they sign an agreement stating the data will be used for research purposes only.

Privacy issues As discussed in Sec. 2.1.1, our data is sourced from three origins. Open source knowledge, like encyclopedias and knowledge bases, is publicly available and free of private information. For online consultation records, we carefully select websites that use anonymous patient data and provide clear doctor information to ensure traceability and protect patient privacy.

Biases issues Data source bias is addressed by ensuring that encyclopedias and knowledge bases reflect authentic medical knowledge objectively. The collection of online consultation records was conducted without bias, accurately representing the QA distribution in Chinese internet consultations. App. C presents relevant distributions as word clouds. The prominence of obstetrics and pediatrics in the consultation data underscores new parents' knowledge gaps and the frequency of childhood illnesses. Additionally, the high incidence of reproductive health inquiries may reflect decreased stigma around seeking online care for such issues.

3 Benchmarking

In this section, we benchmark mainstream answer *retrieval* and *generation* methods respectively.

3.1 Retrieval Based Benchmark

3.1.1 Baselines and Experimental Settings

We rank the top 1000 relevant answers from a pool that includes both training and test set answers. For encyclopedias and knowledge bases, 90% of questions are used for training, while 10% are for testing. For consultant records and other categories, 99% are for training, as testing with 1% is sufficient and saves evaluation time. We employ BM25, DeepCT (Dai and Callan, 2019), and DPR (Karpukhin et al., 2020) as baselines, with BM25 and DeepCT as sparse retrieval methods and DPR as a dense method. See baseline details in App. H.1. Performance is measured using Recall@k and MRR@10, where Recall@k indicates the percentage of top k passages containing the answer, and MRR@10 computes the average inverse rank of the first relevant document.

3.1.2 Results

The experimental results in Tab. 5 show that both DeepCT and DPR outperform BM25, highlighting the effectiveness of neural IR models. Generally, DPR surpasses DeepCT, likely due to dense IR models being more powerful than sparse ones. However, recall performance is low in consultant records experiments, as the large candidate pool (26M) makes it difficult to retrieve the desired documents.

Interestingly, we observe that even when the desired answer is not specifically recalled, the top-ranked responses are still informative. To conduct a quantitative assessment, we randomly select 100 questions from three data sources, namely, consultation records, encyclopedias, and knowledge bases, and retrieve the top five answers for each question using DPR. Subsequently, we enlist the expertise of three general practitioners to determine if any of these answers could directly address the given questions. The research findings indicate that within these three data sources, 52%, 54%, and 42% of the questions respectively have answers among the top five retrieved responses. This sug-

	From Online Medical Consultant records:
Question	3岁宝宝把整个水果糖咽了，怎么才能知道是咽下去了呢? The 3-year-old baby swallowed the whole fruit candy, how can I know that the baby has swallowed it and not stuck it in the throat?
Answer	只要是咽后宝宝没有憋气的现象，那就是咽下去了。 As long as the baby does not hold his breath after swallowing, the baby has swallowed the thing.
	From Online Medical Encyclopedia:
Question	什么是生物药剂学? The 3-year-old baby swallowed the whole fruit candy, how can I know that the baby has swallowed it and not stuck it in the throat?
Answer	生物药剂学是研究给药后药物的吸收的整个体内过程，包含各种制剂因素和生物因素对这一过程与药效的影响。此外，生物药剂学通过药物对生物细胞产生的反应过程来达到施药者想要达到的目的。1950年代初，人们普遍认为“化学结构决定药效”，药剂学只是为改善外观、掩盖不良嗅味而便于服用。随着大量的临床实践证明，人们逐渐开始认识到剂型和生物因素对药效的影响。因此研究药物在代谢过程的各种机理和理论及各种剂型和生物因素对药效的影响，对控制药物之际的内在品质，确保最终药品的安全有效，提供新药开发和用药的严格评价，都具有重要的意义。 Biopharmaceutics is the study of the entire process of drug absorption after administration, including the effects of various preparation factors and biological factors on this process and drug efficacy. Biopharmaceutics uses the process of drug response to biological cells to achieve the expected purpose. In the early 1950s, it was generally believed that "the chemical structure determines the efficacy of the drug", and pharmacy was only for improving the appearance and masking the bad smell to make it easier to take. With a large number of clinical practices, people gradually began to realize the influence of dosage forms and biological factors on drug efficacy. It's important to study various mechanisms and theories of drugs in the metabolic process and the influence of various dosage forms and biological factors on drug efficacy, control the internal quality of drugs, ensure the safety and effectiveness of final drugs, and provide strict evaluation for new drug development.
	From Online Medical Knowledge bases:
Question	气道吸痰的辅助治疗有些什么? What are the adjunctive treatments for airway suctioning?
Answer	足量补液 Adequate rehydration

Table 6: Examples from various sources of the dataset

gests that the retrieval performance is actually significantly better than what is reported in Tab. 5. For specific sample analysis, please refer to App. G.

3.2 Generation Based Benchmark

3.2.1 Baselines and Experimental Settings

We benchmark various classic and latest general generative language models, namely GPT-2 (Radford et al., 2019), T5 (Raffel et al., 2020), ChatGLM3 (Zeng et al., 2023), Qwen (Bai et al., 2023), Baichuan2 (Yang et al., 2023), InternLM (Team, 2023) and ChatGPT. At the same time, we also select two representative medical models, namely HuatuoGPT (Zhang et al., 2023) and DISC-MedLLM (Bao et al., 2023). We use Huatuo-26M to fine-tune T5 and GPT-2, and Huatuo-Lite to fine-tune large language models. See baseline and fine-tuning details in App. H.2. Evaluation Metrics include **BLEU**, **ROUGE**, **GLEU**, and **Distinct**.

3.2.2 Results

The generation benchmark results are summarized in Tab. 7. Fine-tuning significantly enhances T5 and GPT2 performance, with T5 leading in most metrics. In contrast, large language models like ChatGPT and ChatGLM-6B underperform due to their zero-shot and full-shot learning approaches. While reference-based metrics work well for fine-tuned models, large language models yield reasonable but often differing results from the ground truth, necessitating further evaluation by medical experts. Notably, fine-tuning large models with Huatuo-Lite, a 0.6% subset of Huatuo-26M, shows promise, but generating long answers remains challenging, contributing to lower performance in generation metrics.

4 Application I: As Fine-tuning Data for Medical LLMs

4.1 A Lite Version of Huatuo-26M

In order to improve the medical capabilities of LLMs within affordable computing costs, we build a sampling version of Huatuo-26M. To create Huatuo-Lite, a comprehensive pipeline is employed, emphasizing both quality and coverage.

Step I: Data deduplication

The dataset undergoes a thorough Data dedupli-

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	GLEU	ROUGE-1	ROUGE-2	ROUGE-L	Distinct-1	Distinct-2
Language Models *without fine-tuning*										
T5	0.33	0.18	0.12	0.07	0.10	0.67	0.19	0.63	0.01	0.02
GPT2	10.04	4.60	2.67	1.62	3.34	14.26	3.42	12.07	0.17	0.22
Large Language Models *without fine-tuning*										
Baichuan2-7B-Chat	20.73	11.06	6.05	3.38	5.95	26.75	6.83	17.45	0.73	0.92
InternLM-7B-Chat	18.26	10.00	5.92	3.50	5.49	27.74	8.02	18.12	0.64	0.84
Qwen-7B-Chat	18.94	10.04	5.58	3.11	6.30	29.03	7.36	18.13	0.58	0.87
ChatGLM3-6B	14.18	7.50	4.16	2.31	4.72	26.44	6.23	16.98	0.54	0.82
HuatuoGPT	20.59	11.00	6.16	3.44	6.83	28.36	7.72	16.15	0.67	0.93
DISC-MedLLM	18.37	8.94	4.48	2.27	5.67	26.92	5.98	14.96	0.70	0.96
ChatGPT (API)	18.44	6.95	2.87	1.13	4.87	19.60	2.82	12.46	0.69	0.89
Language Models *with fine-tuning*										
T5	26.63	**16.74**	**11.77**	**8.46**	**11.38**	**33.21**	**13.26**	**24.85**	0.51	0.68
GPT2	23.42	14.00	9.35	6.33	9.47	30.48	11.36	23.15	0.43	0.58
Large Language Models *with fine-tuning*										
Baichuan2-7B-Chat	22.52	12.43	7.04	4.06	6.99	28.80	8.13	18.53	0.78	0.94
InternLM-7B-Chat	23.36	12.99	7.71	4.60	7.53	30.32	8.79	18.95	0.62	0.86
Qwen-7B-Chat	**27.30**	15.08	8.85	5.24	7.82	29.82	8.66	18.63	0.71	0.92
ChatGLM3-6B	25.65	14.24	8.38	4.97	7.69	29.37	8.67	18.92	0.75	0.93
HuatuoGPT	25.39	13.53	7.63	4.35	7.20	28.75	7.87	18.00	0.76	0.95
DISC-MedLLM	21.52	11.52	6.37	3.60	6.67	27.99	7.60	17.62	**0.82**	**0.97**

Table 7: *Generation* based benchmark. T5 and GPT2 are fine-tuned using Huatuo-26M, while LLMs are fine-tuned using Sampled version of Huatuo-26M.

Step	# Pairs	Len(Q)	Len(A)
Huatuo-26M Train	26,239,047	44.6	120.7
Aft. Deduplication	1,316,730	75.6	131.9
Aft. Filtering	237,127	81.3	141.7
Score 0	3,076	71.5	127.1
Score 1	248,256	60.8	131.6
Score 2	466,459	73.7	127.3
Score 3	361,383	84.7	131.5
Score 4	212,827	81.6	141.4
Score 5	24,300	77.7	144.1
Aft. Polishing	177,703	80.1	143.9

Table 8: Statistics before and after each step of data processing during the creation of Huatuo-Lite. *Score* refers to the quality indicator that ChatGPT assigns to the dataset questions in the Data filtering step.

cation Initially, using the BGE (Xiao et al., 2023)[5], we compute the word embeddings for each question. Euclidean distance is adopted as the metric for gauging semantic similarity between embeddings, and questions with a semantic distance less than 12 from a given question are designated as its neighbors. The neighbor count for any question is capped at 512. For the creation of neighbor sets, we employ the vector retrieval library FAISS.

During the processing phase, if the neighbor count for a question falls below 30, it is deemed a low-frequency question and removed. We also define a term frequency distance based on 2-gram overlap. Within the neighbor set. Questions with a term frequency distance exceeding 0.2 are eliminated, ensuring that questions within the set share significant semantic and linguistic resemblance. We then navigate through the entire dataset in a random manner; any new question already appearing in the neighbor set of previously included questions is excluded from consideration.

Step II: Data filtering We employ the ChatGPT to assign a score (ranging from 0 to 5) to the filtered questions. Only those questions with a score of 4 or above are retained. It assesses questions based on clarity, completeness, and relevance, retaining only those scoring 4 or above. Scoring statistics are shown in Tab. 8 and prompts are in the App. K.

Step III: Data polishing The final stage involves ChatGPT rewriting the answer to improve clarity and conciseness. Although the diversity of forum questions can improve the generalization of the model, the answers need to be consistent in style and free of grammatical errors to prevent additional negative effects on the model. This meticulous process results in a dataset of 177,703 high-quality question-answer pairs. The prompt used for polishing can be found in App. K.

Human Evaluation To ensure clinical reliability, 100 randomly sampled responses underwent rigorous assessment by licensed physicians using four criteria: Fluency, Relevance, Completeness, and Professionalism. As detailed in Tab. 10, 81% of responses achieved *Good* fluency scores (4-5)

[5] https://huggingface.co/BAAI/bge-large-zh

Models	CMB-Exam	CMExam	CMMLU (Med)	C-Eval (Med)	CMB-Clin
ChatGPT(API)	43.26	46.51	50.37	48.80	4.53
HuatuoGPT-7B	28.81	31.08	33.23	36.53	3.97
HuatuoGPT-7B (Huatuo-Lite)	32.09 +3.28	31.08 +0.00	36.04 +2.81	36.74 +0.21	3.97 +0.00
DISC-MedLLM-13B	37.51	37.98	38.73	40.07	3.58
DISC-MedLLM-13B (Huatuo-Lite)	41.56 +5.05	42.48 +4.50	44.02 +5.29	46.67 +6.60	3.67 +0.09

Table 9: Assessment of Medical for LLMs Knowledge in *Application I*.

Quality Level	Fluency	Relevance	Completeness	Professionalism
Poor	0	1	5	3
Medium	19	20	56	63
Good	81	79	39	34

Table 10: Human evaluation results (% distribution) of 100 samples assessed by medical experts. Detailed criteria in App. I.

with zero poor ratings, while 79% demonstrated strong relevance to medical contexts. Notably, 95% of answers attained at least *Medium* completeness (56% medium + 39% good), and 97% met medium-to-high professionalism standards. This expert validation confirms our dataset's capacity to deliver accurate, safe medical knowledge while maintaining natural language quality. Full evaluation protocols are documented in App. I.

4.2 Experiments

Problem Setting We use Huatuo-Lite as a fine-tuning corpus for training two representative existing medical large language models, namely HuatuoGPT and Disc-MedLLM. This process is designed to deepen the models' understanding of medical concepts and improve their diagnostic reasoning. The effectiveness of this fine-tuning is evaluated through a series of tests, including multiple-choice questions and the interpretation of complex medical records.

Experimental Settings Models are fine-tuned for 2 epoch with a batch size of 32, with a learning rate of 10^{-5} using Adam. The warm-up rate of cosine scheduling is set to 0.03. For consultation based on complex medical records, the models are set to have a maximum length of 1024, a temperature of 0.5, a top p of 0.7, and a repetition penalty of 1.2 to generate 3 returns. For multiple choice questions, we use greedy strategy to generate 3 returns with a maximum length of 10.

For evaluating our medical language models, we use CMB (Wang et al., 2023), CMExam (Liu et al., 2023), CMMLU (Li et al., 2023), and C-Eval (Huang et al., 2023). CMB offers a comprehensive assessment of clinical medical knowledge through multiple-choice tasks, including CMB-Exam for single and multiple selections and CMB-Clin for consultation question answering with complex medical records. CMExam, derived from the Chinese National Medical Licensing Examination, includes over 60,000 questions. C-Eval and CMMLU also utilize multiple-choice formats to measure large models' knowledge capabilities. For C-Eval, we focus on Clinical and Basic Medicine, while CMMLU emphasizes anatomy, clinical knowledge, college medicine, genetics, nutrition, traditional Chinese medicine, and virology. Our evaluation strategy involves generating answers for these questions, effectively gauging the models' mastery of medical knowledge. The multiple-choice prompt is detailed in App. J.

Results As shown in Tab. 9, the accuracy of multiple-choice questions of HuatuoGPT and DISC-MedLLM are improved aftering fine-tuning on Huatuo-Lite. In particular, DISC-MedLLM has improved by about 5 percentage points in different data sets. However, compared with ChatGPT, the models still have a gap after fine-tuning. At the same time, we also noptic that HuatuoGPT increase limited in CMExam and C eval. This may be because its system prompts require model answers to be as rich and friendly as possible, resulting in part of the answers being analyzed in detail before arriving at the choice. For knowledge-intensive multiple-choice questions, this is likely to exacerbate the model's hallucination, thereby affecting the model's performance (Huang et al., 2023; Wang et al., 2023). Although its performance is worse than DISC-MedLLM on multiple-choice questions, HuatuoGPT is still significantly ahead in complex medical record consultation tasks that simulate real scenarios.

5 Application II: As an External Knowledge Source for RAG

Problem Setting RAG (Lewis et al., 2020) combines pre-trained parametric and non-parametric

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	GLEU	ROUGE-1	ROUGE-2	ROUGE-L	Distinct-1	Distinct-2
On cMedQA2										
T5	20.88	11.87	7.69	5.09	7.62	27.16	9.30	20.11	0.418	0.526
T5-RAG	25.86	18.48	15.26	13.02	14.27	34.24	17.69	27.54	0.395	0.516
T5 (Huatuo-26M)	28.76	17.08	11.67	8.41	10.45	29.79	10.23	20.68	**0.647**	**0.831**
T5 (Huatuo-26M)-RAG	**31.85**	**22.77**	**18.70**	**15.96**	**17.08**	**37.01**	**19.23**	**28.72**	0.573	0.760
On webMedQA										
T5	21.42	13.79	10.06	7.38	8.94	31.00	13.85	25.78	0.377	0.469
T5-RAG	20.30	13.29	9.97	7.61	9.40	32.40	14.88	27.25	0.285	0.377
T5 (Huatuo-26M)	**31.47**	**20.74**	**15.35**	**11.60**	**12.96**	34.38	15.18	26.72	**0.651**	**0.832**
T5 (Huatuo-26M)-RAG	25.56	16.81	12.54	9.58	11.80	**34.88**	**15.59**	**27.43**	0.447	0.611

Table 11: Perfomance of T5 with or without using Huatuo-26M as an external RAG corpus in *Application II*. The difference with Tab. 12 is that here we finally **fine-tune** these models in the target datasets. T5 (Huatuo-26M) was first trained in Huatuo-26M dataset before training in the target dataset.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	GLEU	ROUGE-1	ROUGE-2	ROUGE-L	Distinct-1	Distinct-2
On cMedQA2										
T5	0.23	0.12	0.07	0.04	0.07	0.53	0.13	0.50	0.01	0.01
GPT2	9.96	4.30	2.33	1.33	3.18	13.85	3.07	11.60	0.17	0.21
T5 (cMedQA2) †	20.88	11.87	7.69	5.09	7.62	27.16	9.30	20.11	0.41	0.52
T5 (Huatuo-26M)	25.65	**14.94**	**9.79**	**6.64**	**10.03**	**30.64**	**10.49**	**21.48**	0.54	0.72
GPT2 (Huatuo-26M)	23.34	13.27	8.49	5.55	8.97	29.10	9.81	21.27	0.46	0.61
On webMedQA										
T5	0.47	0.21	0.13	0.08	0.13	1.04	0.20	0.97	0.01	0.01
GPT2	7.84	3.51	1.99	1.16	2.56	12.00	2.70	10.07	0.12	0.15
T5 (webMedQA) †	21.42	13.79	**10.06**	**7.38**	8.94	**31.00**	**13.85**	**25.78**	0.37	0.46
T5 (Huatuo-26M)	23.20	**13.80**	9.21	6.29	**9.22**	30.68	10.90	22.26	0.46	0.63
GPT2 (Huatuo-26M)	19.99	11.54	7.51	4.97	7.80	28.19	9.69	21.30	0.36	0.49

Table 12: Performance of models trained on Huatuo-26M in *Application III*. † indicates fine-tuning while others are zero-shot.

memory (i.e., external knowledge) for generation, by doing which memorization can be decoupled from generalization. Here we use the Huatuo-26M as the external knowledge resource in RAG. For a given question q, we use trained DPR as a retrieval model to get the top-ranked QA pair $(q_{\text{aug}}, a_{\text{aug}})$ from the QA dataset as an additional input.

Experimental Setting Considering that T5 performs better in zero-shot scenarios than GPT2, we use T5 instead of GPT2 to generate the answer conditioning on a concatenated text $(q_{\text{aug}}, a_{\text{aug}}, q)$. Since RAG models rely a retrieval model, we first train a Chinese DPR model using our dataset. Then we use the document encoder to compute an embedding for each document, and build a single MIPS index using FAISS (Johnson et al., 2019) for fast retrieval. In RAG training, we retrieve the closest QA pair for each question and split it into $(q_{\text{aug}}, a_{\text{aug}}, q)$ format. We define the maximum text length after splicing as 400, train for 10 epochs with batch size 24 and learning rate 3e-05. The difference between **T5** and **T5 (Huatuo-26M)** is that the latter was first trained in Huatuo-26M dataset before training in the target dataset (i.e., cMedQA2 or webMedQA).

Results As shown in Tab. 11, we find that the RAG strategy improves the quality of text generation to a certain extent. Particularly, on cMedQA2, the model can consistently benefit from the RAG strategy with and without pre-training on the Huatuo-26M dataset. For RAG, we could additionally train backbone models in Huatuo-26M before fine-tuning, as introduced in Sec. 6; the improvement of the dditional pre-training could be found in cMedQA2 (3 absolute point improvement over purely RAG) but not in webMedQA (nearly 6 absolute point decrease); this might depend on the characteristics of target datasets.

6 Application III: Transferability to Other QA Datasets

Problem Setting We directly apply the model pre-trained on the Huatuo-26M dataset and evaluate it on other answer generation datasets. A similar configuration could be found in T5-CBQA (Roberts et al., 2020).

Experimental Setting We select two

Model	CMedEE	CMedIE	CDN	CTC	STS	QIC	QTR	QQR	**Avg-ALL**
BERT-base	**62.1**	**54.0**	55.4	69.2	83.0	84.3	60.0	**84.7**	69.1
BERT-base (Huatuo-26M)	61.8	53.7	**56.5**	**69.7**	**84.6**	**86.2**	**62.2**	**84.7**	**69.9**
RoBERTa-base	62.4	53.7	56.4	69.4	83.7	85.5	60.3	82.7	69.3
RoBERTa-large	61.8	**55.9**	55.7	69.0	**85.2**	85.3	**62.8**	84.4	70.0
RoBERTa-base (Huatuo-26M)	**62.8**	53.5	**57.3**	**69.8**	84.9	**86.1**	62.0	**84.7**	**70.1**
ZEN (Diao et al., 2019)	61.0	50.1	57.8	68.6	83.5	83.2	60.3	83.0	68.4
MacBERT (Cui et al., 2020)	60.7	53.2	57.7	67.7	84.4	84.9	59.7	84.0	69.0
MC-BERT (Zhang et al., 2020)	61.9	54.6	57.8	68.4	83.8	85.3	61.8	83.5	69.6

Table 13: The performance on the test set of CBLUE evaluation for *Application IV*. We use Huatuo-26M as a pre-trained corpus. The results including Zen, MacBERT, and MC-BERT are from the official website.

existing Chinese medical QA datasets, namely cMedQA2 (Zhang et al., 2018) and webMedQA (He et al., 2019). **cMedQA2** is a publicly available dataset based on Chinese medical questions and answers consisting of 108,000 questions and 203,569 answers. **webMedQA** is a real-world Chinese medical QA dataset collected from online health consultancy websites consisting of 63,284 questions. The settings of T5 and GPT 2 follow Sec. 3.2.1.

Results As shown in Tab. 12, the performance of the model pre-trained on the Huatuo-26M dataset is much higher than the raw models. Especially, additionally training on Huatuo-26M improves the raw T5 models with 25.42 absolute points in cMedQA2 and22.73 absolute points in webMedQA. Moreover, in cMedQA2 dataset, T5 trained in Huatuo-26M which never sees neither the training set nor test of cMedQA2, outperforms T5 trained by cMedQA2 in terms of BLEU-1. This evidences that Huatuo-26M includes a wide range of medical knowledge, which is beneficial for downstream medical tasks. Moreover, using Huatuo-26M as a training set achieves better performance on cMedQA2 than using its own training set, this is probably due to the large scale of

7 Application IV: As a Pre-training Corpus

Problem Setting We use Huatuo-26M as a pre-trained corpus to continue training existing pre-trained language models like BERT and RoBERTa.

Experimental Settings **BERT** (Devlin et al., 2018) and **RoBERTa** (Liu et al., 2019) are typical pre-trained language models for natural language understanding. The base setting is with 12 layers with the large setting is with 24 layers. **BERT-base (Huatuo-26M)** and **RoBERTa-base (Huatuo-26M)** is the model initialized by **BERT-base** and **RoBERTa-base**. They are further continuously trained by the Huatuo-26M dataset using masked language model. To better contextualize the results, we also report the results of ZEN (Diao et al., 2019), MacBERT (Cui et al., 2020), and MC-BERT (Zhang et al., 2020). We evaluate BERT and RoBERTa trained on the Huatuo-26M dataset on the CBLUE (Zhang et al., 2020). CBLUE is the first Chinese medical language understanding evaluation benchmark platform, including a collection of natural language understanding tasks.

Results As shown in Tab. 13, BERT and RoBERTa trained on the Huatuo-26M dataset have improved the performance of CBLUE. The trained 12-layer RoBERTa(Huatuo-26M) model outperforms the 24-layer Roberta model in terms of average scores, demonstrating that the Huatuo-26M dataset is rich in medical information. The average score of the RoBERTa-base (Huatuo-26M) model is 0.8 percentage points higher than that of the RoBERTa-base model and 0.5 percentage points higher than that of the MC-BERT-base.

8 Conclusion

In this paper, we propose the largest Chinese medical QA dataset to date, consisting of **26 Million** QA pairs, expanding existing resources by two orders of magnitude. We benchmark existing works and demonstrate versatile applications including LLM fine-tuning, RAG knowledge bases, and transfer learning. Additionally, we introduce Huatuo-Lite - a refined subset developed through rigorous deduplication, quality filtering, and physician-validated polishing to enable efficient training of clinically reliable models. Experimental results show our datasets effectively enhance medical LLMs' capabilities while maintaining natural language quality. Both resources significantly advance Chinese medical AI development, offering comprehensive solutions for diverse real-world clinical NLP applications.

Acknowledgements

This work was supported by the Shenzhen Science and Technology Program (JCYJ20220818103001002), Shenzhen Doctoral Startup Funding (RCBS20221008093330065), Tianyuan Fund for Mathematics of National Natural Science Foundation of China (NSFC) (12326608), Shenzhen Key Laboratory of Cross-Modal Cognitive Computing (grant number ZDSYS20230626091302006), and Shenzhen Stability Science Program 2023.

Limitations

This dataset may contain some erroneous medical information because the 26M QA pairs are difficult to manually check by experts at this stage. To better maintain the dataset, we aim to build an online website where clinicians or experts can modify these QA pairs.

The dataset may be translated into other languages, especially those with low resources. And translation may introduce some additional errors. Additionally, as with medical consultations, treatment/recommendations vary from person to person. In other words, it may depend a lot on the individual's circumstances, such as age and gender, whether the main symptom like pain is accompanied by other symptoms, or whether the symptoms are early or late. This information may need to be confirmed through multiple rounds of conversations rather than a single round of QA. In the future, we will explore dialogue systems for medical quality assurance.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1):1–23.

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

Zhihao Chen. 2018. 39health-kg. `https://github.com/zhihao-chen/QASystemOnMedicalGraph`.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Zhuyun Dai and Jamie Callan. 2019. Context-aware sentence/passage term importance estimation for first stage retrieval. *arXiv preprint arXiv:1910.10687*.

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

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2019. Zen: Pre-training chinese text encoder enhanced by n-gram representations. *ArXiv*, abs/1911.00720.

Junqing He, Mingming Fu, and Manshu Tu. 2019. Applying deep matching networks to chinese medical question answering: a study and a dataset. *BMC medical informatics and decision making*, 19(2):91–100.

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

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Andre Lamurias, Diana Sousa, and Francisco M Couto. 2020. Generating biomedical question answering corpora from q&a forums. *IEEE Access*, 8:161042–161051.

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

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2023. Benchmarking large language models on cmexam–a comprehensive chinese medical exam dataset. *arXiv preprint arXiv:2306.03030*.

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

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPs*.

Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. 2023. MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–648, Nusa Dua, Bali. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Yang Xiang Qingcai Chen, Ting Ma. 2022. Cpubmed-kg. `https://cpubmed.openi.org.cn/graph/wiki`.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Empirical Methods in Natural Language Processing (EMNLP)*.

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

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Simon Šuster and Walter Daelemans. 2018. Clicr: A dataset of clinical case reports for machine reading comprehension. *arXiv preprint arXiv:1803.09720*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Yuanhe Tian, Weicheng Ma, Fei Xia, and Yan Song. 2019. Chimed: A chinese medical corpus for question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 250–260.

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

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. Huatuogpt, towards taming language model to be a doctor.

Ningyu Zhang, Qianghuai Jia, Kangping Yin, Liang Dong, Feng Gao, and Nengwei Hua. 2020. Conceptualized representation learning for chinese biomedical text mining. *arXiv preprint arXiv:2008.10813*.

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.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849.

Ming Zhu, Aman Ahuja, Wei Wei, and Chandan K Reddy. 2019. A hierarchical attention retrieval model for healthcare question answering. In *The World Wide Web Conference*, pages 2472–2482.

A Ethics Statement

As we mentioned in the limitation, the collected data might still have wrong medical information, which comes from two aspects: 1) doctors might make mistakes in online medical consultation, especially given the fact patience might expose incomplete information; and 2) the automatic extraction of QA pairs might also introduce some inaccurate information. Although the data scale is too large to manually check by medical experts, we have made some efforts to reduce its negative effects. We have highlighted these concerns in many parts of this paper and warned readers. We also remind that the model trained with the article data is limited to scientific research purposes and patient information supplementation, and cannot be abused to directly provide prescription advice to patients in real scenarios.

B Dataset Download

All data are crawled from open-source resources. For these data resources where we extract question-answering pairs, namely online encyclopedias and knowledge bases, we directly provide full-text question-answering pairs. For the raw data we crawled as question-answering pairs, like online consultation records, we provide two versions: a **raw version** that provides a URL website associated with a question-answering pair; and a **full-text version** that directly provides full texts for question-answering pairs. Huatuo-26M providing URL links for online consultation records is fully open-sourced. QA pairs from encyclopedias and knowledge bases are full-text and complete, but one has to crawl QA pairs from online medical consultation records by itself. This is to avoid data misuse from some companies or individuals. While Huatuo-26M provides full texts for all QA pairs is only open-sourced to research institutes or universities if they agree on a license to promise for the purpose of research only.

C Word Clouds for Huatuo-26M Dataset

As shown in Fig. 3, Fig. 4, and Fig. 6, we extracted the top 1000 keywords based on TF-IDF and drew word clouds for different sources of Huatuo-26M. It shows QA pairs from online consultation records are more informal since they use more daily words like ‘宝宝’ (namely ‘a lovely nickname for babies’); while they are more formal in other resources with more professional medical words, the

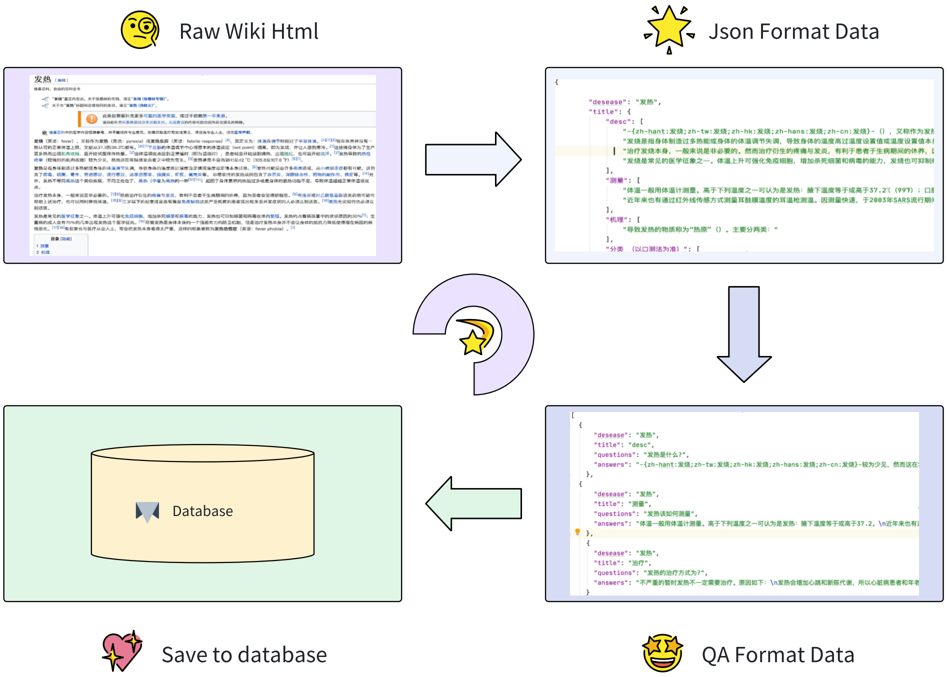

Figure 2: Workflow for extracting QA pairs from WIKI pages.

Version	Data sources			Access
	consultant records	encyclopedias	knowledge bases	
raw version	URL	full-text	full-text	public-available
full-text version	full-text	full-text	full-text	available upon application

Table 14: Data access

combination between formal and informal questions making this dataset diverse.

D Examples of Huatuo-26M Dataset

Tab. 15 shows more examples from various sources of the dataset, and the data characteristics of each data source can be roughly seen through the examples. For Q&A pairs derived from online medical consultation records, the questions are more colloquial and the answers are more targeted. For Q&A pairs sourced from online medical wikis and expert articles, the questions are more concise, rarely involving specific patient information, and the answers are more detailed and professional. For Q&A pairs from online medical knowledge bases, the questions are concise, the answers are accurate, and there are fewer identical texts between answers and questions.

E Extracting QA Pairs from Encyclopedia Pages

As shown Fig. 2 , For a given Wikipedia page, we use an HTML parsing tool to extract its structured information based on the contents of the page. Therefore, we get a title based on the contents which are associated with one or many paragraphs. Next, we transform each title to a question that could be answered by its associated paragraphs, according to a manually-designed template like Tab. 16.

F Questions Templates for Knowledge Bases

Tab. 16 shows the generated templates for all knowledge graph questions. Each question template is associated with either a relation between entities or an attribute of an entity. Each question template is conditioned on the subject entity, see the placeholder of entities like `disease` and `drug` in Tab. 16. Note that the answer to the question should be the object entity or the attribute of the subject entity. There are 43 question types in total. We manually checked 500 random examples where the 'answer' could match the question; the results show nearly every QA pair is correct.

Figure 3: Word clouds drawn from Q&A pairs from online consultation records.

Figure 4: Word clouds drawn from Q&A pairs from Encyclopedia.

Figure 5: Word clouds drawn from Q&A pairs from Knowledge bases.

G Examples of Retrieval Based Benchmark

We select DPR for the case study since it has the best overall performance. Fig. 17 shows the retrieved results using DPR. Interestingly, the top-ranked answers are relevant and generally valid since the number of QA is large and many of them might be redundant and it might lead to *false negatives*. Therefore, although the retrieval metrics (e.g. recall 5) are relatively low, its retrieval quality is moderately satisfied.

H Details about Baselines

H.1 Baseline Details for Retrieval

BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document. We use single characters as units to build indexes instead of words. We

	From Online Medical Consultant records:
Question	我可能得了戊肝，因为我饮食越来越不好，而且吃饭的时候总是想吐。问一下戊肝早期的症状是什么呢? I may have hepatitis E because my diet is getting worse and I always feel sick when I eat. What are the early symptoms of hepatitis E?
Answer	肝早期的症状会出现脂肪无法消化，造成大量脂肪存积于体内，同时胃功能也会紊乱，就会出现厌食，看见油腻的食物就恶心的症状，肤色素沉着，指甲颜色改变，毛发改变等。 The early symptoms of the liver will be that the fat cannot be digested, causing a large amount of fat to accumulate in the body. At the same time, the stomach function will also be disordered, there will be symptoms of anorexia and nausea when seeing greasy food. Symptoms such as hyperpigmentation and changes in nail color may also occur.
	From Online Medical Encyclopedia:
Question	前列腺钙化灶是怎么治呢? How is prostate calcification treated?
Answer	钙化灶是X线检测到的前列腺内的钙质沉积。前列腺内的钙化灶有大小之分：粗大的钙化灶常常为列腺内的良性病变，如前列腺内动脉的老化、陈旧性的损伤以及炎症等，不需要进一步活检。细小的钙化灶通常位于细胞生长分裂较快的部分。出现前列腺钙化或结石必须治疗，钙化会发展成结石，引发出各种症状，有的症状长期消除不了，要做全面检查，看是否有结石钙化，不治疗结石钙化难以彻底治愈前列腺病。保养治疗需要劳逸结合，防止过度疲劳进行适当的体育运动，尤其是加强盆腔肌肉的运动，忌长久静坐，忌长久骑车，忌久蹲，排便时间控制在3到5分钟，忌坐潮湿之地。便后清洁肛门。注意饮食，多饮水，忌酒及辛辣食物。多食蔬菜、水果及坚果类食物。因坚果类食物中富含铜和锌，对前列腺有益。 Calcifications are calcium deposits in the prostate that are detected on x-rays. The calcifications in the prostate can be divided into different sizes: Coarse calcifications are often benign lesions in the prostate, such as aging of the internal prostatic artery, old injury, and inflammation, and no further biopsy are required. Fine calcifications are usually located in the part where the cells are growing and dividing more rapidly. Prostate calcification or stones must be treated. Calcification will develop into stones and cause various symptoms. Some symptoms cannot be eliminated for a long time. A comprehensive examination should be done to see if there are stone calcifications. Prostate disease can not be completely cured without treatment for calcification. Maintenance treatment requires a combination of work and rest to prevent excessive fatigue and carry out appropriate physical exercises, especially exercises to strengthen pelvic muscles. Avoid sitting for a long time, riding a bicycle for a long time, and squatting for a long time. The defecation time is controlled within 3 to 5 minutes. Avoid sitting in wet places. Clean the anus after defecation. Pay attention to diet, drink plenty of water, avoid alcohol and spicy food. Eat more vegetables, fruits and nuts. Nuts are rich in copper and zinc, it is good for the prostate.
	From Online Medical Knowledge bases:
Question	脓腔穿刺的辅助治疗有些什么? What are the adjuvant treatments for abscess puncture?
Answer	消毒隔离；皮肤的护理；营养支持 Disinfection and isolation; skin care; nutritional support

Table 15: Examples from various sources of the dataset

Figure 6: Word clouds drawn from Q&A pairs from Knowledge bases.

utilize the Lucene code base and set k1 to 1.2 and b to 0.9.

DeepCT (Dai and Callan, 2019) uses BERT [6] to determine context-aware term weights. We trained the model for 3 epochs, with a learning rate of 2×10^{-5} using Adam. The batch size is set to 72 and the max sequence length is set to 256.

DPR (Karpukhin et al., 2020) learns embeddings by a simple dual encoder framework. The DPR

[6] https://huggingface.co/bert-base-chinese

疾病（disease）	症状 (symptom)	[disease]的症状是什么？（What are the symptoms of [disease]?）
疾病（disease）	并发症 (complication)	[disease]的并发症是什么？（What are the complications of [disease]?）
疾病（disease）	简介 (Introduction)	[disease]的简介是？（What is the profile of [disease]?）
疾病（disease）	预防 (prevention)	[disease]的预防措施有哪些？（What are the preventive measures of [disease]?）
疾病（disease）	病因 (Etiology)	[disease]的发病原因？（What is the cause of [disease]?）
疾病（disease）	发病率 (Morbidity)	[disease]的患病比例是多少？（What is the prevalence rate of [disease]?）
疾病（disease）	就诊科室 (Medical department)	[disease]的就诊科室是什么？（What is the clinic of [disease]?）
疾病（disease）	治疗方式 (treatment)	[disease]的治疗方式是什么？（What is the treatment of [disease]?）
疾病（disease）	治疗周期 (treatment cycle)	[disease]的治疗周期多长？（How long is the treatment cycle of [disease]?）
疾病（disease）	治愈率 (cure rate)	[disease]的治愈率是多少？（What is the cure rate in of [disease]?）
疾病（disease）	检查 (an examination)	[disease]的检查有些什么？（Which check are there for [disease]?）
疾病（disease）	多发群体 (Frequent group)	[disease]的多发群体是？（Which group of people is more likely to get [disease]?）
疾病（disease）	药物治疗 (medical treatement)	[disease]的推荐药有哪些？（What are the recommended drugs for [disease]?）
疾病（disease）	忌食 (Do not eat)	[disease]忌食什么？（What shouldn't one eat for [disease]?）
疾病（disease）	宜食 (Edible)	[disease]宜食什么？（What should one eat for [disease]?）
疾病（disease）	死亡率 (death rate)	[disease]的死亡率是多少？（What is the death rate for [disease] ?）
疾病（disease）	辅助检查 (Auxiliary inspection)	[disease]的辅助检查有些什么？（What are the auxiliary inspections of [disease]?）
疾病（disease）	多发季节 (High season)	[disease]的多发季节是什么时候？（Which season do people most likely get [disease]?）
疾病（disease）	相关（症状） (related (symptoms))	[disease]的相关症状有些什么？（What are the side symptoms of [disease]?）
疾病（disease）	发病机制 (pathogenesis)	[disease]的发病机制是什么？（What is the pathogenesis of [disease]）
疾病（disease）	手术治疗 (operation treatment)	[disease]的手术治疗有些什么？（What is the surgical treatment of [disease]?）
疾病（disease）	转移部位 (metastatic site)	[disease]的转移部位是什么？（What is the site of transfer for [disease]?）
疾病（disease）	风险评估因素 (risk assessment factors)	[disease]的风险评估因素有些什么（What are the risk assessment factors for [disease]）？
疾病（disease）	筛查 (screening)	[disease]的筛查有些什么？（What are the screenings for [disease]?）
疾病（disease）	传播途径 (way for spreading)	[disease]的传播途径有些什么？（What are the channels of transmission of [disease]?）
疾病（disease）	发病部位 (Diseased site)	[disease]的发病部位是什么？（What is the site of [disease]?）
疾病（disease）	高危因素 (high risk factors)	[disease]的高危因素有些什么？（What are the high-risk factors for [disease]?）
疾病（disease）	发病年龄 (Age of onset)	[disease]的发病年龄是多少？（What is the age of onset for [disease]?）
疾病（disease）	预后生存率 (prognostic survival rate)	[disease]的预后生存率是多少？（What is the prognosis for survival for [disease]?）
疾病（disease）	组织学检查 (Histological examination)	[disease]的组织学检查有些什么？（What are the histological examinations for [disease]?）
疾病（disease）	辅助治疗 (adjuvant therapy)	[disease]的辅助治疗有些什么？（What are adjuvant treatments of [disease]?）
疾病（disease）	多发地区 (High-risk areas)	[disease]的多发地区是哪里？（Where are the frequent occurrence areas of [disease]?）
疾病（disease）	遗传因素 (genetic factors)	[disease]的遗传因素是什么？（What is the genetic factor of [disease]?）
疾病（disease）	发病性别倾向 (Onset sex tendency)	[disease]的发病性别倾向是啥？（What is the sex tendency of onset of [disease]?）
疾病（disease）	放射治疗 (Radiation Therapy)	[disease]的放射治疗有些什么？（What is radiation therapy of [disease]?）
疾病（disease）	化疗 (chemotherapy)	[disease]的化疗有些什么？（What is the chemotherapy of [disease]?）
疾病（disease）	临床表现 (clinical manifestations)	[disease]的临床表现有些什么？（What are the clinical manifestations of [disease]?）
疾病（disease）	内窥镜检查 (endoscopy)	[disease]的内窥镜检查有些什么？（What is the endoscopy examination of [disease]?）
疾病（disease）	影像学检查 (Film degree exam)	[disease]的影像学检查有些什么？（What are the imaging tests of [disease]?）
疾病（disease）	相关（导致） (related (resulting in))	[disease]会导致什么样的结果？（What consequence does [disease] lead to?）
疾病（disease）	治疗后症状 (Symptoms after treatment)	[disease]的治疗后症状是什么？（What are the symptoms after treatment for [disease]?）
疾病（disease）	相关（转化） (relevant (conversion))	[disease]会转化成什么？（What will [disease] translate into?）
药品（drug）	推荐药（diseases cured with this drug）	[drug]能治理什么疾病?（What diseases can [drug] treat?）

Table 16: Templates to transform relations in knowledge bases to questions.

model used in our experiments was trained using the batch-negative setting with a batch size of 192 and additional BM25 negatives. We trained the question and passage encoders for 2 epochs, with a learning rate of 10^{-5} using Adam, linear scheduling with warm-up and dropout rate 0.1.

H.2 Baseline Details for Generation

T5 (Raffel et al., 2020) trains many text-based language tasks in a unified text-to-text framework. We continuously train T5 for 1 epoch on the full training set of Huatuo-26M using batch-size 8, with a learning rate of 10^{-4} using Adam, linear scheduling with a warm-up rate of 0.1. The Chinese T5 model has 12 layers T5 [7].

GPT2 (Radford et al., 2019) is a decoder-only generative language model. We fine-tune GPT2 for 1 epoch on the full training set with a batch size of 12, with a learning rate of 10^{-4} using Adam, linear scheduling with a warm-up rate of 0.1. In both T5 and GPT2, the maximum lengths of questions and answers are set to 256 and 512. The Chinese GPT is the original 12-layer GPT2 [8].

ChatGLM3-6B (Zeng et al., 2023) is an open bilingual language model based on General Language Model (GLM) framework, with 6.2 billion parameters.

Qwen-7B (Bai et al., 2023) is a strong base language model, which have been stably pre-trained for up to 3 trillion tokens of multilingual data with a wide coverage of domains, languages (with a focus on Chinese and English), etc.

Baichuan2-7B-Chat (Yang et al., 2023) is the new generation of open-source large language models launched by Baichuan Intelligent Technology.

[7] https://huggingface.co/imxly/t5-pegasus

[8] downloaded from https://huggingface.co/uer/gpt2-chinese-cluecorpussmall

Scoring Prompt

Prompt: You are an excellent rating robot. You will be given a question related to medical or health topic. You task is to provide a score to the given question in the scale of 1-5 using the judge criteria below:
1: The given text is incomplete, ambiguous. It lacks enough information for a doctor to make a judgment. It may also contain irrelevant or repetitive information, hyperlinks, or promotional content related to specific doctors.
2: The text is mostly complete and clear, with minimal repetition. But it does not provide enough information for a doctor to make a judgment, and it may not be perfectly concise or well-organized. It might contain minor grammatical errors, but they do not significantly affect its fluency.
3: The text is complete, clear, and concise, with no repetitive or irrelevant information. It provides enough information for a doctor to make a judgment, and it is well-organized and grammatically correct. However, it may still lack a specific question or contain minor ambiguities. There are no hyperlinks or promotional content.
4: The text is very complete, clear, and concise. It provides sufficient information for a doctor to make a judgment and includes a specific question. It is well-organized, grammatically correct, and free of repetition, ambiguities, hyperlinks, and promotional content. However, there may still be minor room for improvement in terms of clarity or richness of information.
5: The text is perfectly complete and concise. It provides all the necessary information for a doctor to make a judgment and includes a specific, clear question. The text is well-organized, grammatically correct, and free of repetition, ambiguities, hyperlinks, and promotional content. It could not be improved in any obvious way.
Please first provide a brief reasoning you used to derive the rating score, and then write **"Score: <score>" in the last line.**

Figure 7: Scoring Prompt for creation of Huatuo-Lite

It was trained on a high-quality corpus with 2.6 trillion tokens.

InternLM-7B-Chat (Team, 2023) a 7 billion parameter base model and a chat model tailored for practical scenarios. It leverages trillions of high-quality tokens for training to establish a powerful knowledge base.

DISC-MedLLM (Bao et al., 2023) is a large-scale domain-specific model designed for conversational healthcare scenarios. It can address a variety of your needs, including medical consultations and treatment inquiries, offering you high-quality health support services.

HuatuoGPT (Zhang et al., 2023) is a large language model (LLM) trained on a vast Chinese medical corpus to construct a more professional LLM for medical consultation scenarios.

ALL of above large language models are fine-tuneed for 2 epoch on the full training set with a batch size of 32, with a learning rate of 10^{-5} using Adam. The warm-up rate of cosine scheduling is set to 0.03. For text generation, the models are set to have a maximum length of 1024, a temperature of 0.5, a top p of 0.7, and a repetition penalty of 1.2 to generate 3 returns. The metric is the average of the three returns.

ChatGPT We use gpt-3.5-turbo-0125.

H.3 Baseline Details for CBLUE

BERT **BERT-base (Huatuo-26M)** is the model initialized by **BERT-base** [9] and continuously trained by the Huatuo-26M dataset using masked language model. We trained the model for 10 epochs with a learning rate 5^{-5} with batch size 64. Questions and answers are spliced together, and the maximum length is 256.

RoBERTa **RoBERTa-base (Huatuo-26M)** is the model initialized by **RoBERTa-base**[10] and continuously trained by the Huatuo-26M dataset using masked language model.We trained the model for 10 epochs with a learning rate 5^{-5} with a batch size 64. Questions and answers are spliced together, and the maximum length is 256.

ZEN (Diao et al., 2019) a BERT-based Chinese text encoder augmented by N-gram representations that take different character combinations into account during training.

MacBERT (Cui et al., 2020) reduces the gap between the pre-training and fine-tuning stages by covering words with a similar vocabulary to it, which is effective for downstream tasks.

MC-BERT (Zhang et al., 2020) study how the pre-trained language model BERT adapts to the Chinese biomedical corpus, and propose a new conceptual representation learning method that a coarse-to-fine cryptographic strategy is proposed

[9] https://huggingface.co/bert-base-chinese

[10] https://huggingface.co/hfl/chinese-roberta-wwm-ext

to inject entity and linguistic domain knowledge into representation learning.

I Details on Quality Labeling of Dataset

In order to facilitate users to use the dataset for different scenarios, we add quality labels to each data item in the dataset from Fluency, Relevance, Completeness and Proficiency in medicine. The scoring rules are shown in the list below.

First, we invite 3 licensed physicians to annotate 200 data items, and retained 146 data items with consistent scores on each aspect. Then, we use ChatGPT[11] to expand the dataset. Besides the same scoring rules, we also conduct 5-shot prompt chosen from the former 146 data items to further unify the score distribution of ChatGPT and physicians, and finally obtained 10K data items.

We use 70% of the data to train 4 Classification Models on each scoring aspect and the remaining data for testing, obtaining accuracies of 76.86%, 72.70%, 75.60% and 72.70% in Fluency, Relevance, Completeness and Proficiency in medicine, respectively.

```
Fluency:
1: Completely broken and unreadable
    sentence pieces
2: Mostly broken with few readable
    tokens. Or Moderately fluent but
    with limited vocabulary.
3: Mostly coherent in expressing complex
     subjects. Or Human-level fluency

Relevance:
1: Completely unrelated to the question
2: Some relation to the question, but
    mostly off-topic. Or Relevant, but
    lacking focus or key details
3: Highly relevant, addressing the main
    aspects of the question. Or Directly
     relevant and precisely targeted to
    the question

Completeness:
1: Extremely incomplete
2: Almost incomplete with limited
    information. Or Moderate
    completeness with some information
3: Mostly complete with most of the
    information displayed. Or Fully
    complete with all information
    presented

Proficiency in medicine:
1: Using plain languages with no medical
     terminology.
2: Equipped with some medical knowledge
    but lacking in-depth details. Or
    Conveying moderately complex medical
     information with clarity
3: Showing solid grasp of medical
    terminology but having some minor
    mistakes in detail. Or Fully correct
     in all presented medical knowledge
```

For the Model training, we choose to continue training **RoBERTa-base**[12], using Adam as the optimizer, with the learning rate of 1e-5, and train for 6 epochs. The generation argument of ChatGPT is set to *random_sample=False* to ensure its consistency. In addition, we also verify the reliability of ChatGPT scoring. We extract 73 of the 146 valid doctor data items as a test set, and use the remaining data items as the random sampling pool for 5-shot sample prompts. There are 71 data items that are consistent with the hysicians in every aspect, and the remaining two data items only differ by 1 in one aspect. The above experimental results fully verify the effectiveness of ChatGPT scoring.

J Multiple-choice Prompt

As shown in the Fig. 8, we design a multiple-choice prompt following related work (Zhang et al., 2023).

K Details for Creation of Huatuo-Lite

Reduction Based on Semantic&N-gram Initially, using the BGE (Xiao et al., 2023)[13], we compute the word embeddings for each question. Euclidean distance is adopted as the metric for gauging semantic similarity between embeddings, and questions with a semantic distance less than 12 from a given question are designated as its neighbors. The neighbor count for any question is capped at 512. For the creation of neighbor sets, we employ the vector retrieval library FAISS.

During the processing phase, if the neighbor count for a question falls below 30, it is deemed a low-frequency question and removed. We also define a term frequency distance based on 2-gram overlap. Within the neighbor set. Questions with a term frequency distance exceeding 0.2 are eliminated, ensuring that questions within the set share significant semantic and linguistic resemblance. We then navigate through the entire dataset in a random manner; any new question already appearing in the neighbor set of previously included questions is excluded from consideration.

Reduction Based on ChatGPT Scoring The specific scoring prompts are delineated in the

[11] gpt-3.5-turbo

[12] https://huggingface.co/hfl/chinese-roberta-wwm-ext-large

[13] https://huggingface.co/BAAI/bge-large-zh

Multiple-choice Prompt

下面是一道关于医学知识的选择题，请直接回答正确选项，不需要任何分析.
{问题}
{选项}
正确答案是:
The following is a multiple-choice question about medical knowledge. Please answer the correct option directly without any analysis.
{Question}
{Options}
The correct answer is:

Figure 8: Prompt for Multiple-choice answering

Fig. 7.

Refinement Using GPT-3.5-turbo The specific prompt for rewrite is provided in Fig. 9.

The Prompt for ChatGPT Refinement

Prompt:

```
**system**:

### You are Huatuo GPT, an AI assistant for medical questions.

### You are an AI assistant. Provide a detailed answer so user don’t need to search outside to understand the answer.

### You are an AI assistant that follows instructions extremely well. Help as much as you can.

### You should be as specific as possible, address the questioner’s concerns.

### You should answer the question in a gentle and friendly way.

### You should not answer questions that are not related to medical.

### You should not answer questions that are related to specific location, hospital, doctor, brand.

### You should not answer questions that are related to advertisement.

### You will ask for clarification if the question is not clear.

### You will ask for more information if the question is not complete.

### You should not answer questions that are beyond your ability.

### You will be given a question and a reference answer.

### You can refer to the answer given to you for your response, but this is not mandatory. If you think the given answer is not reasonable, please provide the answer you think is reasonable.

### You should give your answer in the following format


“‘Thought about reference answer
Thought
“‘

“‘Answer
Answer
“‘

**user**:
“‘Question
{question}
“‘

“‘Reference Answer
{ref_answer}
“‘
```

Figure 9: The Prompt for ChatGPT Refinement to create Huatuo-Lite

	Ground Truth:
Question	我可能得了戊肝，因为我饮食越来越不好，而且吃饭的时候总是想吐。问一下戊肝早期的症状是什么呢? I may have hepatitis E because my diet is getting worse and I always feel sick when I eat. What are the early symptoms of hepatitis E?
Answer	肝早期的症状会出现脂肪无法消化，造成大量脂肪存积于体内，同时胃功能也会紊乱，就会出现厌食，看见油腻的食物就恶心的症状，肤色素沉着，指甲颜色改变，毛发改变等。 The early symptoms of the liver will be that the fat cannot be digested, causing a large amount of fat to accumulate in the body. At the same time, the stomach function will also be disordered, there will be symptoms of anorexia and nausea when seeing greasy food. Symptoms such as hyperpigmentation and changes in nail color may also occur.
	Retrival Answers and their original questions:
Top 1	戊型肝炎通常发病比较急，并且在发病期初可能会有恶心，呕吐以及稍稍有一些黄疸的症状。这个疾病主要是通过粪口途径传播的，并且常常在老人以及孕妇或者是有乙肝基础的病人发病率比较高。这个疾病通常早期应该严格卧床休息，直到症状消失，才可以逐渐正常活动 The onset of hepatitis E is usually acute, and symptoms of nausea, vomiting, and slight jaundice may occur early in the illness. This disease is mainly transmitted through the fecal-oral route, and the incidence is often higher in the elderly, pregnant women, or patients with hepatitis B background. This disease usually requires strict bed rest in the early stages. It is not until the symptoms disappear that normal activities can gradually be resumed
Question	我最近听说我朋友得了戊肝，我不太了解这个疾病，这个是不是病毒性肝炎?
Top 2	戊型肝炎主要经粪一口途径传播，大多数报道的暴发性流行都与饮用了被粪便污染的水有关，大暴发常常是在暴雨与洪水发生之后，水源被污染时出现，多见于秋冬季。也可散发，在环境与水源卫生状况差的地区，全年都有散发病例。此外，还可通过日常生活接触和输入性传播。症状可能会出现发热、头痛、咽痛、鼻塞、呕吐、上腹不适、肝区痛、腹胀、腹泻等。每个人体质和病情不同，症状就不同。 Hepatitis E is mainly transmitted through the fecal-oral route. Most reported outbreaks are related to drinking water contaminated by feces. Major outbreaks often occur after heavy rains and floods when water sources are contami-nated, and are more common in autumn and winter. . It can also be sporadic. In areas with poor environmental and water sanitation conditions, sporadic cases occur throughout the year. In addition, it can also be transmitted through daily source contact and imported sex. Symptoms may include fever, headache, sore throat, nasal congestion, vomiting, upper abdominal discomfort, liver pain, abdominal distension, diarrhea, etc. Everyone’s body constitution and condition are different, and the symptoms will be different
Question	我最近听说很多人得了戊型肝炎，我也想预防一下，想知道一下戊肝的症状原因? I recently heard that many people have hepatitis E. I also want to prevent it. I want to know the symptoms and causes of hepatitis E?
Top 3	戊型其实是由是由肝炎病毒所致的全身性传染病，主要累及肝脏。其临床表现为食欲减退、恶心、乏力、上腹部饱胀不适、肝区疼痛，肝肿大、压痛及肝功能损害等，部分病例出现黄疸 Type E is actually a systemic infectious disease caused by the hepatitis virus, which mainly affects the liver. Its clinical manifestations include loss of appetite, nausea, fatigue, upper abdominal fullness and discomfort, pain in the liver area, hepatomegaly, tenderness, and damage to liver function. In some cases, jaundice occurs.
Question	我体检时检查出戊肝，但是我平时生活挺规律的，想要知道戊肝出现的原因有哪些呢? I was diagnosed with hepatitis E during my physical examination, but my daily life is quite regular. I want to know what are the causes of hepatitis E?

Table 17: Example of retrieval results of DPR model on question from consultant records