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

Anonymous NAACL submission

Abstract

In this paper, we release the **largest** ever medical Question Answering (QA) dataset with **26 Million** QA pairs named Huatuo-26M and its streamlined version **Huatuo-Lite** with 177K QA pairs. 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 traing 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.

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 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.

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 Table 8, this dataset is two orders of magnitude larger than the existing QA datasets. To improve the data quality, we also release a lite version called 'Huatuo-Lite' that has only 177K QA pairs. 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 conducted benchmark evaluations of text generation, covering a series of autoregressive language models from GPT2 (Brown et al., 2020) and T5 (Raffel et al., 2020b) 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-tnuening data for medical large language models (LLMs); (ii) as external knowledge for RAG; (iii) transferability to other QA datasets; and (vi) as a pre-trained corpus.

Use Case I: As Fine-tuning Data for Medical LLMs We utilize Huatuo-Lite as a corpus to enhance the capabilities of two existing medical large language models, HuatuoGPT and Disc-MedLLM. Experimental results on multiple-choice questions and complex medical record interpretation shows that both models could benefit from Huatuo-Lite in fine-tuning.

Use Case II: As an External Knowledge Source for RAG Large-scale medical QA datasets themselves explicitly contain rich medical knowledge, and we leverage it as external knowledge in the context of retrieval-augmented generation (Lewis et al., 2020). Experimental results on cMedQA2 and webMedQA datasets show that using this dataset as an external knowledge base

[1] see https://en.wikipedia.org/wiki/World_Health_Organization_ranking_of_health_systems_in_2000

can greatly improve the quality of generated texts.

Use Case III: Transferability to Other QA Datasets Since the Huatuo-26M dataset is large, we also expect that the models trained by the dataset could encapsulate general medical knowledge. Therefore, we use the trained models on two existing medical QA datasets, namely cMedQA2 (Zhang et al., 2018) and webMedQA (He et al., 2019). Experimental results in Sec. 7 show that the model can achieve competitive performance even in few or zero samples.

Use Case IV: As a Pre-training Corpus Since data scale of Huatuo-26M is large, we use the text corpus of Huatuo-26M as a pre-trained corpus that could inject implicit knowledge into the model through pre-training. We improve Bert and RoBERTa in a continuously-training manner on the dataset by using QA pairs as pre-training corpora. The experimental results show the performance of pre-trained models on biomedical tasks could be largely improved by using Huatuo-26M as an additional 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) along with its streamlined version (with 177,703 QA pairs); (**ii**) we benchmark some existing models for the proposed methods for both retrieval and generation; and (**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 Appendix E 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. Each record is a QA pair: a patient raises a question and a medical doctor answers the question. 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.

Data Processing We directly capture patient questions and doctor answers that meet the requirements as QA pairs, getting 31,677,604 pairs. Subsequently, we conduct a filtration process to eliminate QA pairs that contained special characters and expunged any redundant pairs. Finally, we get 25,341,578 QA pairs.

[2]Qianwen Health in https://51zyzy.com/, we only released the URL links instead of extracted full texts to avoid data license issue.

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 1: Basic statistics of Huatuo-26M.

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 We first structure an article. Each article is divided into title-paragraph pairs. For example, such titles in articles about medicines could be usage, contraindications, and nutrition; for articles about medicines about diseases, they could be diagnosis, clinical features, and treatment methods. We remove the titles of paragraphs that have appeared less than five times, finally resulting in 733 unique titles. Based on these titles, we artificially design templates to transform each title into a question that could be answered by the corresponding paragraph. Note that a disease name or a drug name could be a placeholder in the templates. See the Appendix F for details.

2.1.3 Online Medical Knowledge Bases

Data Sources Some existing 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** [4] 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**[5] and **Xywy-KG**[6] are two open source knowledge graphs. Basic information is shown in Tab.9.

Data Processing We clean the three knowledge graphs by removing invalid characters and then merge entities and relationships among entities among these three knowledge graphs, resulting in 43 categories. Each category is associated with either a relationship between entities or an attribute of entities. Subsequently, we

[3]zh.wikipedia.org/wiki/

[4]https://cpubmed.openi.org.cn/graph/wiki

[5]https://github.com/zhihao-chen/QASystemOnMedicalGraph

[6]https://github.com/baiyang2464/chatbot-base-on-Knowledge-Graph

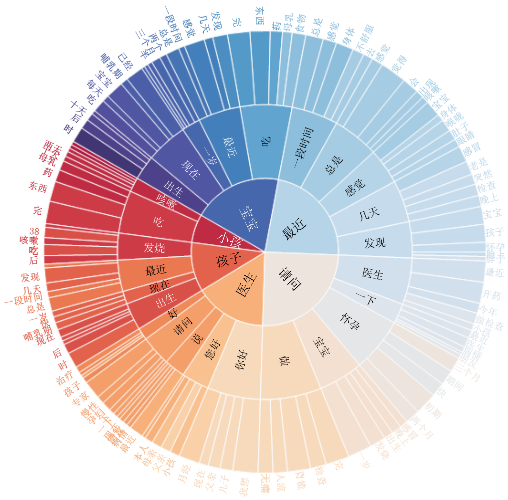

Figure 1: Distribution of questions. We present the relative distribution of these recurring problems and their subsequent distributions.

manually design templates to convert each category to a *question*. The *question* is either 1) querying the object entity based on the subject entity or 2) querying an attribute of an entity. The object entity will be the *answer* w.r.t the *question* in both cases. Finally, we obtain 798,444 QA pairs by constructing questions and answers with corresponding templates. See Appendix G for details.

2.2 Data Statistics and Analysis

The basic statistics of Huatuo-26M are shown in Table 1, 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 Huatuo-26M consists of a large number of colloquial QA pairs, which are closer to the offline medical diagnosis and contain a lot of medical knowledge. As shown in the sample from online medical consultation in Table 13, the patient's question contains patient characteristics and daily symptoms accompanied by life-like scenes, while the doctor's answers are targeted and with contextual semantic continuity.

Diverse Question Topics Our heuristic analysis of the dataset's questions, detailed in Figure 1, reveals a focus on issues concerning newborns, pregnant women, children, and the elderly, highlighting the role of online consultations in addressing the health needs of these demographics in the context of China's aging population.

Significant Topics in Huatuo-26M Word clouds in Appendix D show the dataset's coverage of health issues, from common to complex diseases. Answers provide medical prescriptions, lifestyle guidance, and hospital referrals. Compared to online consultations, Wikipedia-based QA pairs show more topics in specialized fields, while knowledge base QA pairs emphasize complex conditions, with answers suggesting advanced diagnostic and treatment procedures.

2.3 Data Licence and Privacy Issues

Data licence For question-answer pairs extracted from open-source online encyclopedias and knowledge bases, we provide full texts unrestrictedly. In contrast, for online consultation records, we release only the question and its URL, without the full texts. To access the full texts, one must visit the URL. This method is adopted to prevent data misuse, as we do not hold the license to disseminate it fully.

Privacy issues As discussed in Sec. 2.1.1, our data originates from three sources. The open-source knowledge sources, such as encyclopedias and knowledge bases, are publicly available and generally free of private information. For the online medical consultation records, all data is shared proactively by users (e.g., doctors and patients), who are aware that this information will become public. Additionally, we have released a 'lite' version in which considerable effort has been made to remove any private information, see Sec. 3.

2.4 Quality Assessment by Experts

We enlist the expertise of a licensed medical doctor to manually assess the quality of the collected data. We select 100 examples from each data source and ask the doctors to evaluate whether the answers accurately address the questions without containing any factual errors. The accuracy rates for the three sources, namely a) online medical consultation, b) medical encyclopedia, and c) medical knowledge bases, are 71%, 88%, and 79% respectively. Notably, the data obtained from online medical consultation is deemed the least reliable. Despite each answer in online consultations being provided by a medical doctor with a specific name and affiliation, the quality of the data is not guaranteed. We also introduce Huatuo-Lite in Sec. 3, a curated, high-quality subset. Huatuo-Lite underwent rigorous deduplication, refinement, and thorough privacy checking, including manual sampling and pattern searches, to meet a high standard of privacy and quality for research use.

3 A Lite Version: Huatuo-Lite

3.1 Data Creation of Huatuo-Lite

To create Huatuo-Lite, a refined medical consultation dataset, a comprehensive pipeline was employed, emphasizing both quality and privacy. See details in Appendix J and here is the sketch of pipeline.

Step I: Data depublication The dataset underwent a thorough Data depublication. Word embeddings for each question were generated using the BGE (Xiao et al., 2023), and Euclidean distance measured the semantic

similarity. Questions with high similarity were grouped into neighbor sets through the FAISS (Johnson et al., 2019), while we select the most representative items and remove redundant ones.

Step II: Data filter Low-frequency questions and those lacking sufficient semantic and linguistic resemblance were excluded. We employ the GPT-3.5-turbo model 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 assessed questions based on clarity, completeness, and relevance, retaining only those scoring 4 or above. The prompts and scoring statistics is in the Appendix J.

Step III: Data Polishing The final phase involved GPT-3.5-turbo rewriting the answers for enhanced clarity and conciseness. This meticulous process, along with rigorous efforts to remove any private information, resulted in a dataset of 177,703 high-quality question-answer pairs, underlining a strong commitment to both data quality and privacy protection in Huatuo-Lite.

3.2 Data Statistics and Analysis

The basic statistics of Huatuo-Lite are shown in Table 10, and it can be seen that compared with Huatuo-26M, the average question and answer length of the streamlined version increase, which are 80.1 and 143.9 respectively. In order to better know the distribution of knowledge in the data and ensure the representativeness and diversity of Huatuo-Lite, we analyze it from two perspectives: department and disease.

Departments Distribution of Huatuo-Lite Taking into consideration the medical institution diagnostic subject directory [7], expert recommendations, and the composition of the dataset, we firstly define 16 department categories shown in Table 11.

Subsequently, we randomly sample 30,000 records and employ the GPT-4 model for medical departments annotation. We then split the data at a ratio of 8:2 into training and testing sets, and choose BERT[8](Devlin et al., 2018) as our foundational classification model. The model achieve an accuracy rate of 93.88 % on the test set. We further utilize it to annotate the entire dataset with department labels. From the "Ratio" column of Table 11, we can see that the data set covers various departments, and focuses on obstetrics and gynecology, internal medicine, dermatology and venereology, and pediatrics. Detailed prompts, training procedures and parameter settings can be referenced in Appendix K.

Significant Diseases in Huatuo-Lite In order to ascertain the types of diseases covered by our dataset, we initially collect disease and complication names from Chinese disease knowledge website[9], obtaining a vocabulary of 15,717 disease entities. We then define the disease entities covered in the question as relevant diseases. For problems with several entities, we define priorities following expert opinions and select the entity with the highest priority. In the end, 80.80% of the samples successfully extracted relevant disease labels, including a total of 2,702 entity categories. A Word cloud based on these frequency and more detailed information for annotation are shown in Appendix K.

3.3 Privacy Issues in Huatuo-Lite

In the development of Huatuo-Lite, a rigorous layered filtering process was employed to prioritize privacy and eliminate personal data. Initially, the filter step was crucial for identifying and removing sensitive content. Subsequently, ChatGPT was utilized to further sanitize both questions and answers, a vital step in eradicating private information while preserving the integrity of the content. The final stage involved an exhaustive effort to use regular expression filters for privacy information, to ensure the absolute exclusion of private data.

3.4 Quality Assessment by Experts

To validate the effects of the refinement, we randomly sample 10 entries from each department, resulting in a total of 160 samples. These are evaluated using GPT-4 and 10 users. Evaluation options include preferences for the original answer, the refined answer, or considering both of equal quality. The scoring prompt for GPT-4 can also be found in Appendix K. GPT-4 and the users find that 48.12% and 47.81% of the answers, respectively, are superior to the original ones, while 40.00% and 40.46% of the answers are deemed to be of comparable quality. We conduct a department-wise evaluation, as detailed in Table 11, revealing that for each department, *the vast majority of answers are of higher or equivalent quality compared to before refinement*.

4 Benchmarking

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

4.1 Retrieval Based Benchmark

4.1.1 Baselines and Experimental Settings

For a given question, we rank the top 1000 relevant answers from the answer pool, which consists of answers from both training and test sets. For encyclopedias and knowledge bases, we use 90% questions for training and the rest for testing. For consultant records or all categories, we use 99% questions for training and the rest for testing, since testing with 1% questions is enough and could save more evaluation time than that with 10% questions. We use BM25, DeepCT (Dai and Callan, 2020) and DPR (Karpukhin et al., 2020) as our baselines, BM25 and DeepCT are sparse retrieval methods while DPR is a dense retrieval method. See baseline details in App. I.1.

[7] http://www.nhc.gov.cn/fzs/s3576/201808/\protect\@normalcr\relax345269bd570b47e7aef9a60f5d17db97.shtml

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

[9] https://m.120ask.com/jibing/

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

Evaluation Metrics We use Recall@k and MRR@10 as indicators. Recall@k measures the percentage of top k retrieved passages that contain the answer. MRR@10 calculates the average of the inverse of the ranks at which the first relevant document is retrieved.

4.1.2 Results

The experimental results are shown in Table 2. Both DeepCT and DPR outperform BM25, evidencing the effectiveness of neural IR models. In most cases, DPR performs better than DeepCT, this is probably because dense IR models might be generally more powerful than sparse neural IR models. Note that the recall performance is relatively low in experiments involving consultant records since the pool of retrieval candidates (i.e., 26M) is too large to recall 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 selected 100 questions from three data sources, namely, consultation records, encyclopedias, and knowledge bases, and retrieved the top five answers for each question using DPR. Subsequently, we enlisted 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 had answers among the top five retrieved responses. This suggests that the retrieval performance is actually significantly better than what is reported in Table 2. For specific sample analysis, please refer to App. H.

⚠ It is worth noting that retrieval-based solutions for medical QA assume that 1) there should be predefined answers for all medical questions; and 2) answers should be static for a given question and independent of the different backgrounds of patients. The two assumptions sometimes do not hold. First, there are always some new emergent situations in the medical domain, e.g. COVID-19, which people have little information about it when it just emerges. Second, the answers (e.g., suggestions and treatment) to a given medical question depend on the individual's situation, such as age and gender, symptoms and complications, and whether the symptoms are in an early or late stage. Therefore, a static answer might not be enough for medical consultation.

4.2 Generation Based Benchmark

4.2.1 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., 2020a), ChatGLM3 (Zeng et al., 2023), Qwen (Bai et al., 2023), Baichuan2 (Yang et al., 2023), InternLM (Team, 2023) and ChatGPT (GPT-3.5-turbo). 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. I.2.

Evaluation Metrics Evaluation Metrics include **BLEU**, **ROUGE**, **GLEU**, and **Distinct**. **BLEU** assesses generated text similarity to references via k-gram overlap. **ROUGE-N** gauges N-gram concurrence with references, while **ROUGE-L** focuses on the longest matching word sequence. **GLEU** inspects sentence fluency through parsing comparisons. **Distinct-1/2** measures response diversity by counting unique n-grams. However, these reference-dependent metrics may not fully apply to medical question answering due to the potential variability in correct responses.

4.2.2 Results

The results of the generation benchmark are summarized in Table 3. Fine-tuning significantly enhances the performance of T5 and GPT2 models, with T5 showing the best results in most evaluation metrics. Large language models like ChatGPT and ChatGLM-6B, however, underperform compared to the fine-tuned T5 due to their respective zero-shot and full-shot learning approaches. While reference-based metrics are effective for fine-

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 3: Generation based benchmark. T5 and GPT2 are fine-tuned using Huatuo-26M, while LLMs are fine-tuned using Huatuo-Lite.

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 4: Knowledge Evaluation for Medical LLMs

tuned models, large language models still provide reasonable results, though they may differ from ground truth. This necessitates further evaluation by medical experts. Moreover, large language models show improvement when fine-tuned with Huatuo-Lite, a subset comprising 0.6% of Huatuo-26M, indicating efficient fine-tuning with a smaller yet comprehensive dataset. The lower performance in generation metrics is likely due to the fact that it is challenging to exactly generate long answers as expected.

⚠ We warn that generation-based medical QA is risky. Since it is difficult to verify the correctness of generated content; misleading information in the medical domain might lead to severe ethic issues. We benchmark these generation methods because generation methods in QA are nowadays more promising than retrieval methods thanks to the success of ChatGPT. However, it is not ready to be deployed in the real world.

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

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, with its multiple-choice task, CMB-Exam, covering single and multiple selections, and CMB-Clin focused on consultation question answering using complex medical records. CMExam, derived from the Chinese National Medical Licensing Examination, includes over 60,000 multiple-choice questions. C-Eval and CMMLU, which also utilize a multiple-choice format, measure large models' knowledge capabilities.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	GLEU	ROUGE-1	ROUGE-2	ROUGE-L	Distinct-1	Distinct-2
cMedQA2 Fine-tuned										
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
webMedQA Fine-tuned										
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 5: The comparison with or without using Huatuo-26M as an external RAG corpus. The difference with Tab. 6 is that here we finally fine-tune these models in the target datasets.

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

Table 6: Performance of models trained on Huatuo-26M. † indicates fine-tuning while others are zero-shot.

For C-Eval, we concentrate on Clinical Medicine and Basic Medicine, while for CMMLU, the focus is on anatomy, clinical knowledge, college medicine, genetics, nutrition, traditional Chinese medicine, and virology. Our evaluation strategy involves directly generating answers for these multiple-choice questions to effectively gauge the models' mastery of medical knowledge.

Results As shown in Table4, 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 notic 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.

6 Application II: As an External Knowledge Source for RAG

Problem Setting RAG (Lewis et al., 2020) combines pre-trained parametric and non-parametric 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., 2017) 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 Table 5, 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. 7; 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.

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

7 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 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. 4.2.1.

Results As shown in Table 6, 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 Huatuo-26M that might have related information in cMedQA2. This shows a great potential ofHuatuo-26M for transfer learning.

8 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., 2022). CBLUE is the first Chinese medical language understanding evaluation benchmark platform, including a collection of natural language understanding tasks.

Results As shown in Table 7, 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.

9 Conclusion

In this paper, we propose the largest Chinese medical QA dataset to date, consisting of **26 Million** medical QA pairs, expanding the size of existing datasets by more than 2 orders of magnitude. At the same time, we benchmark many existing works based on the data set and demonstrate the possible uses of the dataset in practice. We also release a streamlined version of the dataset to lower the barrier for the community to utilize it. The experimental results show that the dataset contains rich medical knowledge that can be very helpful to existing datasets and tasks.

10 Limitation

Our study has two limitations. First, although our dataset collects a lot of data, multi-modal data is needed to build a complete diagnostic process. Secondly, in the process of creating Huatuo-Lite, we utilize GPT-3.5-turbo to rewrite answers which may cause information loss or errors.

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.

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*.

Zhuyun Dai and Jamie Callan. 2020. Context-aware term weighting for first stage passage retrieval. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 1533–1536.

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*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *CoRR*, abs/1702.08734.

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*.

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. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 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, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

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 Limitation

⚠ The dataset might contain some wrong medical information since its scale is large with 26M QA pairs and manual checking by experts in nearly impossible in the current stage. To better maintain the dataset, we aim to build an online website where clinical doctors or experts could modify these QA pairs. This might be done by recruiting part-time doctors to first check these data and regularly update them.

This dataset might be translated into other languages, especially low-resource languages. Note that the translation might introduce some additional errors. Moreover, one should also be noticed some basic differences between traditional Chinese medicine and western medicine.

For medical consultation, the treatment/suggestions vary from person to person. In other words, it might be highly dependent on the individual's situation, e.g., age and gender, whether the main symptoms such as pain are accompanied by other symptoms, or whether the symptoms are early or late. The information might need to be confirmed in a multi-turn dialogue instead of single-turn QA. In the future, we would explore dialogue systems for medical QA.

B 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 have removed all private information of 'patient' users in all data sources.

C 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.

D Word Clouds for Huatuo-26M Dataset

As shown in Figure 3, 4, and 5, 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 combination between formal and informal questions making this dataset diverse.

E Examples of Huatuo-26M Dataset

Table 13 shows 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.

F 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. 14.

G Questions Templates for Knowledge Bases

Tab. 14 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. 14. 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.

Domain	Dataset	Lang	Domain	Source	#Q
medical	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
	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
	Huatuo-Lite	Chinese	Medical	Consultation records, Encyclopedia, KBs	177K
	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 8: Existing QA dataset.

	# 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 9: Basic statistics of the three knowledge bases.

Step	# Pairs	Len(Q)	Len(A)
Aft. Semantic&N-gram	1,316,730	75.6	131.9
Aft. ChatGPT Score	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. Refinement	177,703	80.1	143.9

Table 10: Statistics in the streamlining process.

Department	Orig.	Aft.	Eq.	Ratio
Obstetrics & Gynecology(妇产科)	25.0	35.0	40.0	19.3
Internal Medicine(内科)	0.0	62.5	37.5	16.7
Dermato & Venereology(皮肤与性病科)	10.0	50.0	40.0	13.9
Pediatrics(儿科)	7.5	60.0	32.5	11.9
Otorhinolaryngology(耳鼻喉科)	15.0	37.5	47.5	7.8
Oncology(肿瘤)	12.5	42.5	45.0	5.7
Neuroscience(神经科)	2.5	47.5	50.0	5.6
Surgery(外科)	22.5	47.5	30.0	5.4
Andrology(男科)	17.5	32.5	50.0	4.6
Infectious & Immune(感染与免疫科)	7.5	37.5	55.0	2.8
Dentistry(牙科)	12.5	40.0	47.5	1.8
Psychology(心理科)	12.5	55.0	32.5	1.7
Trad. Chinese Medicine(中医)	10.0	72.5	17.5	1.4
Emergency Medicine(急诊科)	2.5	60.0	37.5	1.0
Reproductive Health(生殖健康科)	15.0	27.5	57.5	0.8
Others(其他)	15.0	57.5	27.5	0.6

Table 11: Evaluation Preferences by Department for Original and Refined Answers of Huatuo-Lite. We consolidate the preference data from GPT-4 with human preferences. The data in the table are percentages. "Orig." indicates a preference for the original answer, "Aft." indicates a preference for the polished answer, and "Eq." indicates consistent quality. "Ratio" represents the proportion of the data set occupied by this department.

H Examples of Retrieval Based Benchmark

We select DPR for the case study since it has the best overall performance. Table 15,16,17 shows the retrieved results using DPR. Interestingly, the top-ranked answers are relevant and generally valid, especially for the first case in online consultant records in table 17 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.

I Details about Baselines

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

DeepCT (Dai and Callan, 2020) uses BERT [10] 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 model used in our experiments was trained using the batch-negative setting with a batch size of 192 and additional BM25

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

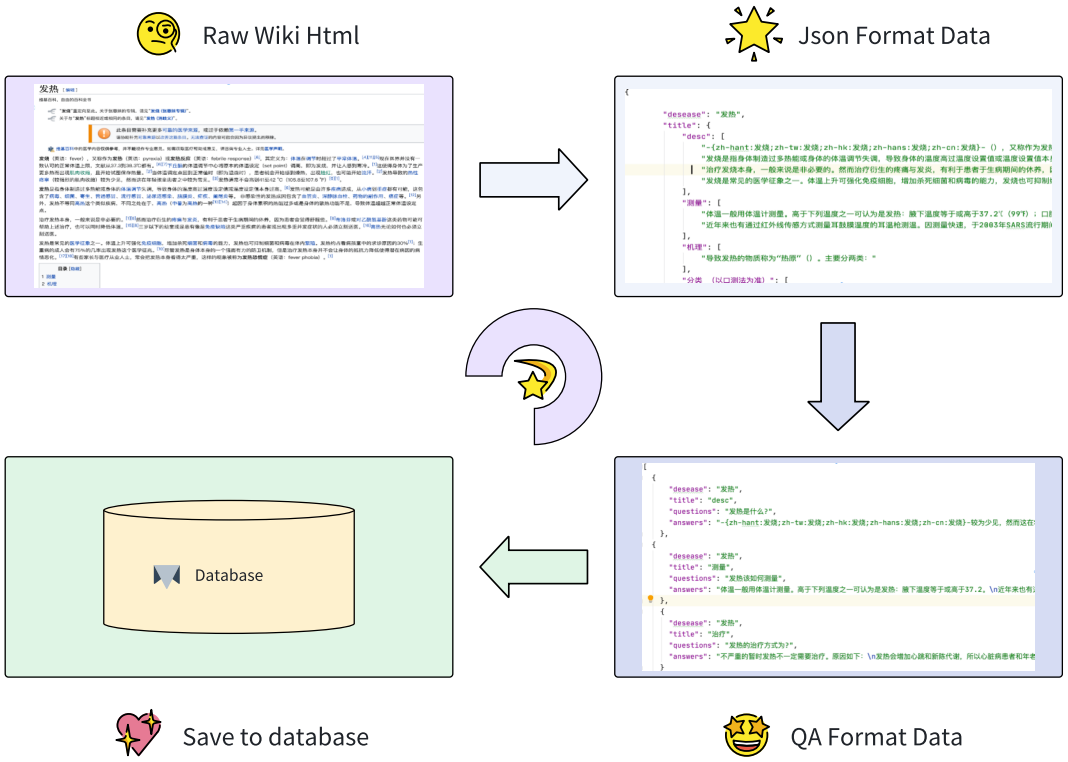

Figure 2: Workflow for extracting QA pairs from WIKI pages.

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 12: Data access

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

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.

I.2 Baseline Details for Generation

T5 (Raffel et al., 2020a) 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

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

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

of 10^{-4} using Adam, linear scheduling with a warm-up rate of 0.1. The Chinese T5 model has 12 layers T5 [11].

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 [12].

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.

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

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

Baichuan2-7B-Chat (Yang et al., 2023) is the new generation of open-source large language models launched by Baichuan Intelligent Technology. 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 ChatGPT (GPT-3.5-turbo) on 10th May 2023.

I.3 Baseline Details for CBLUE

BERT **BERT-base (Huatuo-26M)** is the model initialized by **BERT-base** [13] 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**[14] 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 to inject entity and linguistic domain knowledge into representation learning.

J Details for Creation of Huatuo-Lite

Reduction Based on Semantic&N-gram Initially, using the BGE (Xiao et al., 2023)[15], 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 Subsequently, we employ the GPT-3.5-turbo model 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. A detailed distribution of scores can be found in the table 10, while the specific scoring prompts are delineated in the Figure 8.

Refinement Using GPT-3.5-turbo We employ GPT-3.5-turbo to rewrite the answers, with the specific prompt provided in Figure 10. The original answer is also fed into the prompt as reference information for GPT-3.5-turbo. We exclude samples where the length of the answer text is less than 5 characters post-refinement, ultimately obtaining 177,703 high-quality question-answer pairs.

K Details for Statistics of Huatuo-Lite

Details of Department Annotation Based on the established department classification structure, we utilize GPT-4 to annotate a randomly sampled set of 30,000 instances for training the classification model. The specific prompt can be found in Figure 7. We train the Bert model [16] on the dataset annotated by GPT-4. We set the batch size to 64 and weight decay to 0.1, and trained for 10 epochs. We utilize the Adam optimizer for model optimization, with a learning rate of 5e-5 and a warm up proportion of 0.1.

Details of Related Disease Annotation We determine which diseases are mentioned in the question based on the constructed vocabulary. In cases where a question contained multiple disease names, we select the one with the highest importance defined as follows: Assuming disease name d with a length $L(d)$ and a priority $P(d)$, the importance $I(d)$ can be calculated as follows:

$$I(d) = L(d) \times P(d)$$

Priority $P(d)$ is based on the presence of specific characters in d. It's set 5 for containing "肿瘤"(tumor), 4 for containing "癌"(cancer), 3 for containing "病"(disease) and 2 for containing"炎"(-itis) and 1 for others. We get related diseases for 80.80% samples, encompassing a total of 2,702 categories.

Wordcloud Based on Related Diseases Based on the "related diseases" labels and their proportions, we construct a word cloud as shown in Figure 9.

GPT-4 Evaluation Prompt To assess whether the quality of the text improved after refinement, we employ GPT-4 and invite users for evaluation. The prompt used for GPT-4's assessment can be found in Figure 11.

[13] https://huggingface.co/bert-base-chinese
[14] https://huggingface.co/hfl/chinese-roberta-wwm-ext
[15] https://huggingface.co/BAAI/bge-large-zh
[16] https://huggingface.co/bert-base-chinese

Multiple choices 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 6: Prompt for Multiple choices answering

Classification Prompt

Prompt: 内科, 外科, 妇产科, 儿科, 神经科学, 眼耳鼻喉科, 皮肤性病科, 心理科学, 中医科, 感染与免疫科, 急诊科, 肿瘤科, 遗传与发育科, 保健科, 生殖健康科, 男性健康科, 口腔科, 药剂科, 检验科, 医学影像科, 麻醉科, 其他。请判断下面的这个问题属于以上哪个科室的问题，只用给出科室标签即可，不需任何额外的文本输出。{问题}

Figure 7: Department classification prompt for creation of Huatuo-Lite

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 8: Scoring Prompt for creation of Huatuo-Lite

Figure 9: Main Diseases of Huatuo-Lite.

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 10: The Prompt for ChatGPT Refinement to create Huatuo-Lite

Scoring Prompt

Prompt: Please evaluate the quality of each answer, A1 and A2, based on the following medical inquiry questions, output the final choice (options include: 'A1 is better', 'A2 is better', 'Quality is comparable').
Please note the following requirements:

Accuracy of the answers: Are they based on the latest medical research and appropriately explained medical terms?
Clarity of advice: Are clear recommendations provided, considering the patient's specific situation?
Ensure your thought process is logical and well-structured. The input data format should be in JSON, as shown: {"Question": "xxx", "A1": "xxx", "A2": "xxx"}
After completing your analysis, please output your chosen label using the following format: ###A1 is better###, ###A2 is better###, or ###Quality is comparable###

Figure 11: Department classification prompt for creation of Huatuo-Lite

	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.
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	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 cannot 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.
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	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
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 13: Examples from various sources of the dataset

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 14: Templates to transform relations in knowledge bases to questions.

Question	前列腺钙化灶是怎么治呢?
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。
Top 1	如果只是单纯的钙化灶那么不必担心多数可能是以前有过炎症治愈后留下的证据可以先查一个前列腺液常规 原问题:再去检查一次前列腺实质内钙化灶什么问题？有什么要注意？想治疗好
Top 2	你好根据你的情况看来单纯的钙化灶一般不会有什么症状也无需治疗如果合并有前列腺炎需要针对前列腺炎进行治疗可以通过一些抗菌素中药等等进行治疗以及温水坐浴勿劳累过度熬夜避免饮酒辛辣刺激性食物等等同时注意自己的饮食还有休息情况如果严重请到医院就医祝你早日康复 原问题:最近到医院检查身体发现自己的前列腺增生钙化，在很小的时候曾经换过前列腺炎不过后来发现的及时已经治愈了，现在看来是不是以前留下来的后遗症。请问我这种前列腺增生钙化怎么治疗？可以治愈吗?
Top 3	前列腺钙化灶这种情况下是因为有之前炎症导致的疤痕情况，现在要注意再看一下是否还有尿路感染等导致的情况，可以通过药物来进行治疗的。要注意多喝水，不要憋尿，而且要注意避免辛辣刺激性的食物，平时的话注意增强个人体质来改善的。 原问题:18岁前列腺钙化灶怎么办?
Question	什么是生物药剂学?
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。
Top 1	生物药理学，在生物制药和医药生物技术是跨学科领域之间的药理学和生物技术，认为是一种新兴的科学。包括获得药物的生物起源在生物反应器。一个主要的优势使用这条路线，而不是获得化学合成，避免了消旋的产品，这样就可以获得大量易纯化产品的同类产品，提高性能并降低成本。另一个优势是获得化合物，几乎无法获得任何其他方式尽可能多的重组蛋白。这有助于科学的设计和开发新疗法。 原问题:什么是生物药理学?
Top 2	生物产药，又译基因产药术或药耕，是遗传工程学的一种透过基因改造的动植物来生产药物的方法。以此方式生产的通常是重组蛋白质或者其代谢产物。重组蛋白质通常用在生物反应器皿中通过细菌和酵母生产，但是通过生物产药不需要高昂的基础设备，产能可以更加低廉的费用按需而变。 原问题:什么是生物产药?
Top 3	原答案

Table 15: Examples of retrieval results of DPR model on questions generated from Encyclopedia

Question	脓腔穿刺的辅助治疗有些什么?
Answer	消毒隔离；皮肤的护理；营养支持
Top 1	不留死腔；引流通畅；支管开窗引流；了解脓腔范围 原问题:粘膜下脓肿的辅助治疗有些什么?
Top 2	保持引流通畅；护理干预；严格拔管；严格无菌操作；保持引流瓶的合适高度 原问题:双侧脑室外引流的辅助治疗有些什么?
Top 3	破膜；程控降温法；换液；选择合适的供者和采集方法；康复训练；洗血；巴氏消毒；贴壁法；55℃水浴加热 原问题:脐血的辅助治疗有些什么?
Question	气道吸痰的辅助治疗有些什么?
Answer	足量补液
Top 1	加温湿化器联合密闭式吸痰法 原问题:NSICU人工气道的辅助治疗有些什么?
Top 2	平卧位；床头抬高30°到45°体位；按需吸痰原则 原问题:胃肠反流的辅助治疗有些什么?
Top 3	常规雾化排痰法；气管按压法；吸痰管刺激法 原问题:诱导性主动咳嗽的辅助治疗有些什么?

Table 16: Examples of retrieval results of DPR model on questions generated from Knowledge bases

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?
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。
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。 原问题: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。 原问题: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?
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。 原问题: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?
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?
Answer	只要是咽后宝宝没有憋气的现象，那就是咽下去了。
Top 1	就目前的这种情况首先要确认一下是否已经吞下，一般的情况下宝宝都会有感觉，比如腹疼了，呕吐了等。 原问题:13个月宝宝，昨天发现窗帘上的小挂钩少了一个，怀疑让宝宝误吞了，需要到医院做什么检查吗?
Top 2	既然能够咽得下去应该是没事的，你可以注意观察孩子的呼吸状况和面色情况。如有异常问题立即就诊。 原问题:一岁八个月宝宝吃果冻噎住又咽下去了，刚才又喝了点水，有事没有?
Top 3	看核的大小，一般都可以排出来，可以密切观察孩子进食情况，只要吃的好，不呕吐，就没问题，如果进食差或呕吐，就要去医院检查了。 原问题:十个月的宝宝吞下荔枝核有没有事，急求答案

Table 17: Examples of retrieval results of DPR model on questions from consultant records