PediaBench: A Comprehensive Chinese Pediatric Dataset for Benchmarking Large Language Models

Anonymous ACL submission

Abstract

The emergence of Large Language Models (LLMs) in the medical domain has stressed a compelling need for standard datasets to evaluate their performance. Although there have been several benchmark datasets for medical problems, they either cover common knowledge across different departments or are specific to another department rather than pediatrics. Moreover, some of them are limited to objective questions and do not measure the generation capacity of LLMs. Therefore, they cannot comprehensively assess the ability of LLMs in pediatrics. To fill this gap, we construct PediaBench, the first Chinese pediatric dataset for LLM evaluation. Specifically, it contains 4,565 objective questions and 1,632 subjective questions spanning 12 pediatric disease groups. It adopts an integrated scoring criterion based on five types of questions to thoroughly assess the proficiency of an LLM in instruction following, knowledge understanding, clinical case analysis, etc. Finally, we validate the effectiveness of PediaBench with extensive experiments on 21 open-source and commercial LLMs. Through an in-depth analysis of experimental results, we offer insights into the ability of LLMs to handle pediatric questions in the Chinese context, highlighting their limitations for further improvements. Our code and data are published anonymously at https://anonymous.4open.science/r/PediaBench-E8E2.

1 Introduction

Question answering (QA) is an important task in natural language processing (NLP) that has received considerable attention over several decades (Simmons, 1965; Hirschman and Gaizauskas, 2001; Lan and Jiang, 2020; Li et al., 2024). Recently, Large Language Models (LLMs) (OpenAI et al., 2024), with their remarkable language understanding, reasoning, and generation capabilities, have shown exceptional performance in QA tasks compared to traditional methods, especially for subjective questions. Naturally, there is also an increasing interest in applying LLMs to medical QA (Chen et al., 2023a; Singhal et al., 2023; Zhang et al., 2023a; Liévin et al., 2024). In this paper, we focus on the QA tasks in *pediatrics*, a medical department that involves the care of infants, children, adolescents, and young adults. Since pediatrics often involves the manifestations and treatments of diseases that differ from those of adults, LLMs with common medical knowledge might not perform equally well on pediatric QA. Therefore, evaluating the proficiency of LLMs in pediatric QA is an urgent need for their application in this domain.

Several medical QA benchmarks have been proposed in the literature (Jin et al., 2019, 2021; Pal et al., 2022; Fansi Tchango et al., 2022; Zhang et al., 2022; Li et al., 2023b; Liu et al., 2023; Wang et al., 2023a,b; Zhu et al., 2023; Cai et al., 2024a; Yue et al., 2024). However, they still have some limitations that hinder their effectiveness in the pediatric context. First, they are mostly general medical benchmarks across multiple departments and are not tailored for pediatrics. As such, their coverage of knowledge in pediatrics is often very limited. Second, most of them contain only objective questions, e.g., true/false and multiple-choice questions. Although they can serve as an indicator of the capacity of LLMs to comprehend medical knowledge, they cannot assess the capacity of LLMs to generate medical texts. In addition, some QA datasets originating from doctor-patient interactions in real-world scenarios (Li et al., 2023b) include some subjective questions to evaluate the conversation ability of LLMs. However, they focus only on a limited number of common diseases and often lack depth of knowledge in medicine. Therefore, existing benchmarks are insufficient to provide comprehensive evaluations of LLMs in terms of pediatric capability.

To address the above challenges, in this paper, we introduce PediaBench, the first comprehensive

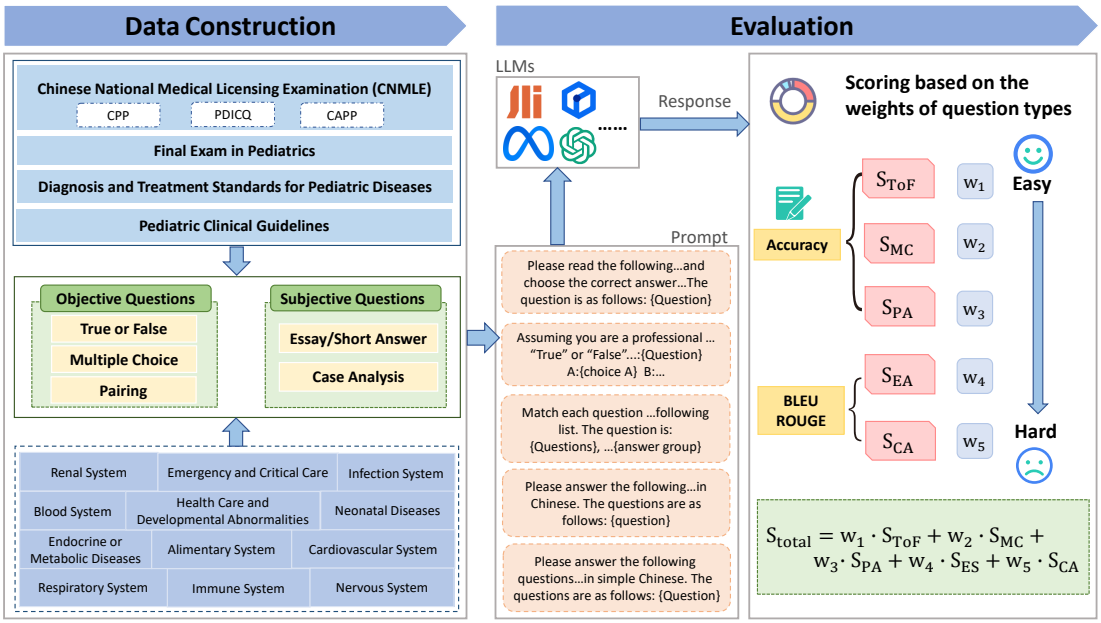

Figure 1: The Framework of PediaBench.

Chinese benchmark dataset in pediatrics. As shown in Figure 1, PediaBench contains 4,565 objective questions and 1,632 subjective questions collected from various sources, such as the Chinese National Medical Licensing Examination (CNMLE), final exams in medicine from renowned universities, pediatric disease diagnosis and treatment standards, and clinical guidelines. PediaBench encompasses five distinct types of questions, namely true or false, multiple choice, pairing, essay/short answer, and case analysis, across 12 typical pediatric disease groups. Furthermore, to provide an accurate evaluation of the performance of each LLM for QA in pediatrics, we adopt a scoring criterion that integrates distinct measures for different types of questions (*accuracy* for objective questions and *BLEU* (Papineni et al., 2002) & *ROUGE* (Lin, 2004) for subjective questions) and assigns a weight to each type according to difficulty. Finally, we conduct extensive experiments with 21 LLMs on the PediaBench dataset and provide an in-depth analysis of the performance of LLMs for QA in pediatrics. Our main contributions are summarized as follows:

- We introduce PediaBench, a high-quality QA dataset specific to pediatrics in the Chinese context. We also devise an integrated scoring scheme to measure the QA performance of each LLM on PediaBench.
- We evaluate PediaBench with 21 LLMs, including open-source and commercial general-purpose models of different scales and specialized models in the medical domain. The results indicate that PediaBench is a challenging dataset that can gauge the capacities of LLMs in terms of pediatric QA.
- We provide a detailed analysis of the results, highlighting the limitations of current LLMs and suggesting their potential future improvements in this emerging domain.

2 Related Work

General LLM Benchmarks The rapid advances in LLMs (Touvron et al., 2023; Yang et al., 2023; Du et al., 2022; Bai et al., 2023; OpenAI et al., 2024) have underscored the need for benchmarks to evaluate their performance in a variety of NLP tasks. To this end, a large number of benchmarks specifically designed for LLMs, e.g., (Wang et al., 2019; Xu et al., 2020, 2023; Zhang et al., 2023b; Zhong et al., 2023; Huang et al., 2023; Li et al., 2023a; Suzgun et al., 2023; Gu et al., 2024), were proposed. We refer interested readers to (Chang et al., 2024) for an extensive survey. These general benchmarks often require LLMs to answer questions from standard examinations in a broad spectrum of domains to evaluate their capacity in text understanding, logical reasoning, calculation, generation, etc. However, although medical questions are included in some of these benchmarks, they are about common knowledge that can be answered without specialization, which cannot accurately reflect the proficiency of LLMs in medical tasks.

Medical LLMs and Benchmarks More recently, many efforts have been made to build specialized LLMs in the medical domain, e.g., DoctorGLM (Xiong et al., 2023), ChatDoctor (Li et al., 2023c), BianQue (Chen et al., 2023a), PMC-LLaMA (Wu et al., 2024), BioMistral (Labrak et al., 2024), MEDITRON (Chen et al., 2023b), ZhongJing (Yang et al., 2024), and QiLin-Med (Ye et al., 2023). Generally, they use general-purpose LLMs as foundation models, construct training corpora with medical articles, textbooks, guidelines, etc., and fine-tune foundation models on the training corpora to inject medical knowledge.

With the development of medical LLMs, there has also been an increasing interest in benchmarking the medical knowledge of LLMs. Jin et al. (2019) constructed PubMedQA, a biomedical QA dataset from PubMed abstracts. MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and CMExam (Liu et al., 2023) consisted of questions from standardized medical examinations. The datasets above contain only multiple-choice questions for evaluation and cannot fully capture the generation capability of LLMs. CMB (Wang et al., 2023a) and MedBench (Cai et al., 2024a) incorporated clinical case analysis into the evaluation. CBLUE (Zhang et al., 2022) and PromptCBLUE (Zhu et al., 2023) included eight NLU tasks (e.g., named entity recognition, information extraction, and sentence classification) to evaluate the capabilities of models for language understanding in medicine. Fansi Tchango et al. (2022) provided DDXPlus, a large-scale medical diagnosis dataset. Huatuo-26M (Li et al., 2023b) contained a large number of real-world medical dialogues. Chen et al. (2024) built a benchmark dataset with an emphasis on rare diseases. Yue et al. (2024) collected a benchmark dataset that focuses on traditional Chinese medicine. These datasets have contained subjective questions to evaluate the generation and conversation ability of LLMs. However, to the best of our knowledge, none of the existing datasets is specific to pediatrics and all have a poor coverage of knowledge in pediatrics.

3 The PediaBench Dataset

To evaluate the question-answering ability of LLMs in pediatric problems, we have constructed PediaBench. This section provides a detailed description of the construction process of PediaBench, as illustrated in Figure 1.

3.1 Question Types

The motivation behind PediaBench is to establish a benchmark to assess how well LLMs can serve as AI assistants for pediatricians. To this end, we should provide a comprehensive and multifaceted assessment of LLMs in real-world scenarios. As such, PediaBench incorporates the following five typical types of medical questions:

- **True or False (ToF):** This type of question asks whether a statement is factual or not. It requires an LLM to match the statement with its corresponding concepts and facts in the corpus, to understand their semantic meanings, and to reason about them so as to detect possible errors and contradictions.
- **Multiple Choice (MC):** This type of question asks for the selection of one (or more) appropriate choices from multiple candidates to complete a sentence or answer a question. It requires an LLM to distinguish among similar or related concepts. Some also evaluate the mathematical and logical skills of an LLM, as basic calculations are essential to obtain the correct answer.
- **Pairing (PA):** This type of question requires one to exactly match all sentences with their corresponding missing words from the candidate list. Distinguishing among similar concepts is also essential for PA. But since any single mismatch leads to an entire erroneous answer, PA is even more challenging than MC.
- **Essay/Short Answer (ES):** This type of question asks one to elaborate on a specific concept. It requires an LLM to generate coherent and accurate text relevant to the concept.
- **Case Analysis (CA):** This type of question presents an LLM with a description of a particular instance and asks the LLM to make a medical diagnosis and provide treatment measures. It can comprehensively evaluate the medical capacity of an LLM in terms of comprehension, reasoning, and problem solving.

Figure 2 presents examples of different types of questions, as well as their sample answers, in the PediaBench dataset.

3.2 Data Collection and Processing

The questions in PediaBench are collected from diverse yet reliable sources, including the Chinese National Medical Licensing Examination, final exams of universities in medicine, and pediatric dis-

Figure 2: Examples for different types of questions and their answers in PediaBench.

ease diagnosis and treatment standards and clinical guidelines. Next, we will describe each of them.

Chinese National Medical Licensing Examination (CNMLE) We have gathered 3,576 multiple-choice questions from CNMLE, including the Clinical Practicing Physician Examination (CPP), the Clinical Assistant Practicing Physician Examination (CAPP), and the Pediatric Doctor In-Charge Qualification Examination (PDICQ). 3,383 of them were from the question bank of PDICQ, and the rest were from the question banks of CPP and CAPP and relevant to pediatrics. There are 50 questions with more than one correct answer. For these questions, an answer is considered correct if it includes and only includes all correct options.

Final Exams in Medicine We have collected the final exams of renowned universities in medicine from their official websites. We manually extracted 258 true-or-false questions, 167 essay/short-answer questions, and 67 case analysis questions in pediatrics from the original PDF files.

Pediatric Disease Diagnosis and Treatment Standards & Clinical Guidelines We also curated 1,398 essay/short-answer questions from the *Pediatric Disease Diagnosis and Treatment Standards* series and *Clinical Guidelines*, covering critical aspects such as etiology, diagnostic criteria, treatment plans, and preventive measures for diseases across diverse pediatric specialties.

Manual Construction We manually constructed a set of 283 pairing questions based on the multiple-choice questions with shared stems or answer options. Since pairing questions require an exact matching between all missing parts in the stems and options, they are more challenging than multiple-choice questions.

Classification of Disease Groups The questions curated from the *Pediatric Disease Diagnosis and Treatment Standards* series and *Clinical Guidelines* have been annotated with disease groups to which they belong referring to the International Classification of Diseases (ICD-11) standard issued by the

Question Type	#Questions	Data Source	Disease Groups?	Measure
ToF	258	FE, PCG	Yes	Accuracy
MC	3,576	CPP, CAPP, PDICQ	Yes	Accuracy
PA	283	CPP, CAPP, PDICQ	Yes	Accuracy
ES	1,565	FE, DTSPD, PCG	Yes	BLEU, ROUGE
CA	67	FE	No	BLEU, ROUGE

Table 1: Statistics of the five question types in PediaBench. For *data source*, "FE" stands for final exams of universities, "DTSPD" stands for diagnosis and treatment standards for pediatric diseases, "PCG" stands for pediatric clinical guidelines, "CPP" stands for Clinical Practicing Physician Examination, "CAPP" stands for Clinical Assistant Practicing Physician Examination , and "PDICQ" stands for Pediatric Doctor In-Charge Qualification Examination.

WHO. We further annotated the remaining unclassified questions. Specifically, we used GLM-4 to guide the task of disease group classification. We first wrote cue phrases for GLM-4 to suggest a disease group for each question. Then, we manually eliminated duplicate questions, double-checked the classification results of GLM-4, and corrected the misclassified ones.

3.3 Dataset Statistics

As shown in Table 1, the PediaBench dataset consists of 5,749 questions, including 258 true-or-false questions, 3,576 multiple-choice questions, 283 pairing questions, 1,565 essay/short-answer questions, and 67 case analysis questions. Except for case analysis questions, the remaining 5,682 questions are organized into 12 distinct disease groups, namely, diseases of the renal system, emergency and critical care, diseases of the infection system, diseases of the blood system, diseases of the cardiovascular system, diseases of the immune system, diseases of the respiratory system, endocrine or metabolic diseases, health care and developmental abnormalities, neonatal diseases, diseases of the alimentary system, and diseases of the nervous system. Note that case analysis questions do not need to be classified because (1) they require an LLM to diagnose the disease groups to which the instance belongs according to symptoms and (2) the instance might belong to more than one disease group. The number of questions for different disease groups in PediaBench is shown in Table 2. We can see that PediaBench features a balanced distribution among all disease groups, which comprehensively covers a wide spectrum of both prevalent and rare pediatric diseases and establishes an unbiased benchmark.

3.4 Evaluation Criteria

For objective questions (i.e., ToF, MC, and PA), we use *accuracy* as the performance measure. We also note that for MC with more than one correct option and PA, we only consider *fully* correct answers for scoring. For each LLM, we use the percentage of questions of each type that it correctly answers as its score for this type, denoted as S_{ToF}, S_{MC}, and S_{PA}, respectively.

For subjective questions (i.e., ES and CA), we adopt the BLEU score (Papineni et al., 2002) and three ROUGE scores (Lin, 2004) (ROUGE-1, 2, l), which gauge the semantic similarity of the answer returned by an LLM and the ground truth in different aspects, as performance measures. Each score is in the range $[0, 1]$, and a higher value implies that the answer has a semantic meaning closer to the ground truth. For each LLM, we take the average of the four scores multiplied by 100 for all the questions of each type as its score for this type, denoted as S_{ES} and S_{CA}, respectively.

Furthermore, we realize that the difficulty of each type of question differs, and thus it is unreasonable to treat them all equally. As such, we follow the design principle of standardized examinations and devise a weighted scoring scheme based on their relative difficulty. Generally, ToF, for which random answers are expected to achieve a score of 50, is considered the easiest among the five types and is assigned with the lowest weight. MC and PA questions, which are provided with candidate answers to select, present a moderate level of difficulty. PA is harder than MC because it requires exact matching. ES and CA are open questions without unique answers. Answering these questions necessitates the generation capacity of LLMs. Therefore, they are much more challenging than objective questions. Finally, CA is even harder than ES because it may require the integration of multiple pieces of knowledge to answer. Therefore, we assign different weights to each type of question to reflect their respective difficulty levels. Taking into account all the above issues, the total score for

Disease Group	Abbreviation	ToF	MC	PA	ES	CA
Renal system	DRnS	293	21	24	139	Unclassified (67)
Emergency and critical care	ECC	270	20	21	104	
Infection system	DInS	297	20	20	150	
Blood system	DBS	294	22	24	122	
Cardiovascular system	DCS	295	21	20	138	
Immune system	DImS	294	21	20	128	
Respiratory system	DRpS	299	21	26	132	
Endocrine or metabolic diseases	EMD	316	22	21	145	
Health care and developmental Abnormalities	HCDA	318	26	40	149	
Neonatal diseases	ND	293	22	21	144	
Alimentary system	DAS	314	22	21	104	
Nervous system	DNS	293	20	25	110	

Table 2: Statistics of the number of questions of each type from the 12 disease groups in PediaBench.

each LLM is calculated as Equation 1.

$$S_{\text{total}} = w_1 \cdot S_{\text{ToF}} + w_2 \cdot S_{\text{MC}} + w_3 \cdot S_{\text{PA}} + w_4 \cdot S_{\text{ES}} + w_5 \cdot S_{\text{CA}}. \quad (1)$$

In practice, we set $w_i = 0.05 \cdot (1 + i)$ for $i = 1, 2, \ldots, 5$ as the weight for each question type. Other weighting schemes that reflect the difficulty of each question type are also allowed.

4 Experiments

4.1 Experimental Setup

We validate PediaBench through experiments with 21 general-purpose and medical LLMs, including open-source models of various scales and commercial models. Specifically, the models we use include: (1) medical LLMs such as BianQue (Chen et al., 2023a), QiZhenGPT (CMKRG et al., 2024), PULSE-7B, and PULSE-20B (Zhang et al., 2023a); (2) the GLM series (Zeng et al., 2023), including ChatGLM3-6B and GLM4; (3) the Baichuan series (Yang et al., 2023), comprising Baichuan2-7B and Baichuan2-13B; (4) the Qwen series (Bai et al., 2023), including Qwen1.5-7B, Qwen1.5-14B, Qwen1.5-72B, and Qwen-MAX; (5) the InternLM series (Cai et al., 2024b), with InternLM2-7B and InternLM2-20B; (6) the LLaMA3 series (Touvron et al., 2023), featuring LLaMA3-8B and LLaMA3-70B; (7) the Sparse Mixture of Experts language models, namely Mixtral 8x7B and Mixtral 8x22B (Jiang et al., 2024); (8) the GPT series (OpenAI et al., 2024), including GPT3.5-turbo and GPT-4o; and (9) ERNIE-3.5-8K-0329 (Sun et al., 2021). For open-source LLMs, we deploy and evaluate them locally using pretrained weights on a server with eight NVIDIA RTX A6000 GPUs. For commercial LLMs, we access and evaluate them through their official APIs. More detailed information on these LLMs is given in Appendix B.

We developed a standardized set of prompts for all LLMs. This ensures that each LLM can generate the desired responses to each type of question based on pre-specified requirements, thereby achieving a standardized and fair evaluation process. Appendix C details the prompts for different types of questions. We adopted the zero-shot prompt setting in all experiments.

4.2 Evaluation Results

Overall Performance The results for the overall performance of LLMs are shown in Table 3. We present the measures of each LLM for the five types of questions and sort them in ascending order of total score S_{total} of Equation 1. Generally, the performance on the objective questions is related to the scale of the models. BianQue (7B) and QiZhenGPT (13B) cannot correctly understand and follow the instructions for the objective questions, and thus their scores are 0 for these types. In terms of ToF, Mixtral-8x7B (47B) shows the lowest score of 38.76 among all models, which is even worse than random answers. GPT-4o generally has the best performance for objective questions. We note that all models with at least 60B parameters have achieved passing scores (i.e., ≥ 60) on ToF and MC. PA is the most discriminative question type, where most models exhibit poor performance. However, GPT-4o achieves a score of more than 10 points higher than the second-ranked model. This confirms the effectiveness of introducing PA into the PediaBench dataset, as PA is not included in existing benchmarks. In terms of subjective questions, larger models also generally achieve higher scores, demonstrating their strong generation capacity. However, their advantages in subjective

Model	ToF	MC	PA	ES				CA				S_{total}
	Accuracy	Accuracy	Accuracy	BLEU	RG-1	RG-2	RG-l	BLEU	RG-1	RG-2	RG-l	
BianQue (7B)	0	0	0	14.21	29.81	4.77	17.46	16.54	29.65	7.58	17.99	9.52
QiZhenGPT (13B)	0	0	0	16.05	**32.05**	**6.83**	19.37	15.8	30.94	8.98	21.49	10.43
LLaMa3-8B	52.71	36.38	9.54	15.27	17.25	1.92	10.47	16.16	19.78	7.49	14.68	19.8
Mixtral-8x7B (47B)	38.76	41.34	11.66	16.25	19.15	2.42	11.52	25.4	29.71	11.81	23.14	22.25
PULSE (7B)	51.94	30.62	6.71	15.12	27.33	4.59	15.99	22.61	39.24	17.53	28.05	23.13
ChatGLM3-6B	53.49	38.48	7.77	18.16	23.5	3.27	13.06	29.24	36.93	15.98	27.51	24.52
InternLM2-20B	61.24	53.02	15.9	15.59	18.65	2.64	8.34	23.75	27.49	9.6	16.84	26.32
InternLM2-7B	56.2	51.76	13.07	17.19	21.03	2.92	9.5	26.54	31.22	12.21	20.05	25.91
Qwen1.5-7B	56.2	41.69	10.95	13.15	31.37	5.57	**22.05**	27.73	37.54	16.48	30.12	26.96
Baichuan2-7B	55.42	48.01	7.42	19.42	24.05	3.46	13.81	30.21	40.8	17.47	31.47	27.02
Mixtral-8x22B (141B)	62.02	47.46	26.15	18.8	25.67	3.73	14.9	25.96	30.96	12.41	21.78	29.33
Baichuan2-13B	58.91	51.54	11.66	20.11	22.54	3.31	12.27	26.94	46.41	19.89	36.86	29.35
PULSE (20B)	56.2	49.52	12.01	15.67	26.26	3.63	17.01	24.03	49.96	23.62	40.1	29.69
GPT3.5-turbo	61.24	48.46	27.91	16.56	30.54	4.9	18.64	24.05	**51.61**	**24.81**	**43.43**	34.18
LLaMa3-70B	72.87	63.84	39.22	18.46	22.46	3.09	12.62	24.53	31.63	11.1	21.38	34.89
Qwen1.5-14B	64.34	61.16	33.57	15.49	30.45	5.27	19.93	28.31	38.42	16.63	30.36	35.3
GLM-4	75.97	78.15	43.81	21.29	24.35	4.76	11.96	31.05	38.03	14.98	26.14	40.24
Qwen-72B	74.42	**83.39**	48.06	18.37	31.18	6.22	19.54	31.29	43.91	19.53	35.67	44.05
Qwen-MAX	77.43	82.33	63.6	20.62	22.35	3.58	11.26	32.21	35.87	14.82	26.6	44.64
ERNIE-3.5-8K-0329	81.01	75.42	68.55	**21.84**	25.13	5.61	13.89	**37.73**	47.71	22.43	35.43	48.03
GPT-4o	**84.88**	80.39	**79.86**	18.97	23.48	3.69	10.14	32.42	44.61	18.87	30.74	**49.53**

Table 3: Results of different LLMs for the scores for five question types and the total scores on PediaBench, where RG-1, RG-2, and RG-l are short for ROUGE-1, 2, l. Here, all LLMs are sorted in ascending order of total score, and the highest score of each question type for each performance measure is highlighted in bold.

Model	HCDA	DRpS	DRnS	DAS	ND	EMD	ECC	DNS	DInS	DCS	DBS	DImS
PULSE(7B)	11.75	11.07	10.53	10.97	12.51	12.67	10.78	11.07	15.45	11.95	10.68	11.25
PULSE(20B)	18.05	15.85	17.54	12.79	16.26	14.31	14.31	13.45	15.29	15.32	15.79	16.56
ChatGLM3-6B	15.33	14.90	13.49	12.84	12.65	11.96	12.91	13.64	11.56	11.45	11.43	12.63
InternLM2-7B	17.75	17.49	16.14	14.76	15.88	13.97	13.00	13.70	18.51	15.05	17.36	16.96
LLaMa3-8B	16.08	13.24	13.61	13.65	12.36	10.15	8.60	13.33	14.87	10.29	14.54	15.12
Qwen1.5-7B	14.92	15.50	16.31	13.24	15.12	12.83	13.46	12.47	14.92	11.78	16.41	14.76
Baichuan2-7B	17.15	14.74	13.24	13.30	14.79	13.99	11.78	13.09	13.84	13.01	15.74	14.35
Baichuan2-13B	16.64	18.40	18.42	14.56	15.12	14.51	12.38	15.98	16.06	14.34	17.73	15.97
Qwen1.5-14B	23.33	25.94	27.34	18.74	22.87	21.82	18.15	24.19	22.68	18.05	23.32	22.19
InternLM2-20B	19.56	19.42	16.19	15.37	16.88	17.56	15.93	15.08	17.10	14.50	16.78	21.18
Mixtral-8x7B (47B)	13.96	14.10	12.48	13.62	11.17	12.66	9.08	14.51	14.63	11.13	15.53	14.56
LLaMa3-70B	23.27	28.48	25.55	21.68	27.11	24.39	20.82	24.92	28.00	20.16	26.68	26.73
Qwen-72B	28.31	30.79	33.83	26.92	30.63	30.82	23.02	34.25	30.24	26.87	30.64	28.13
Mixtral-8x22B (141B)	19.19	19.68	19.67	22.36	19.40	16.30	17.48	17.50	20.90	16.04	19.04	16.05
GPT3.5-turbo	20.69	19.50	21.53	16.43	18.36	19.07	13.83	17.94	20.15	17.32	21.39	20.46
GLM-4	25.80	29.21	32.03	26.42	29.06	24.49	24.03	31.28	33.53	25.33	30.02	27.35
Qwen-MAX	30.52	35.70	32.26	34.95	**36.01**	34.99	30.71	33.68	34.00	27.66	31.45	33.83
ERNIE-3.5-8K-0329	34.41	34.85	34.65	32.45	34.25	33.90	29.64	36.23	34.52	27.08	33.66	29.18
GPT-4o	**37.73**	**38.20**	**37.37**	**35.06**	35.34	**35.11**	**32.96**	**38.75**	**37.45**	**33.63**	**37.69**	**37.26**

Table 4: Scores of different LLMs for three types of objective questions in different disease groups on PediaBench. Here, the highest score in each disease group is highlighted in bold.

questions are less obvious than in objective questions. We also find that CA generally yields higher scores than ES. This is a bit surprising because CA requires the model to perform diagnoses and necessities more precise and nuanced responses.

Performance on Different Disease Groups Table 4 shows the scores of each LLM calculated using Equation 1 specifically for objective questions across 12 disease groups. Here, BianQue (7B) and QiZhenGPT (13B) are omitted. We find that GPT-4o achieves the highest scores in 11 of 12 disease groups. These results are consistent with those of Table 3. Table 5 shows the scores of each LLM calculated using Equation 1 specifically for ES across 12 disease groups. We observe that most models achieve their highest scores in the two disease groups of HCDA of DImS. As for questions related to HCDA, they often involve basic medical knowledge, which does not require in-depth specialization. Moreover, many diseases in the DImS group share similar clinical manifestations and treatment strategies, providing models with clearer clues for answers. There are no models that can perform well for subjective questions in all performance measures across different groups of diseases.

Summary Our main experimental findings are summarized as follows:

1. Until now, none of the LLMs have been able to pass the standardized exam in pediatrics.

Model	HCDA	DRpS	DRnS	DAS	ND	EMD	ECC	DNS	DInS	DCS	DBS	DImS
BianQue (7B)	15.92	17.90	**18.40**	14.36	17.73	16.38	17.18	15.51	16.57	16.49	16.58	15.96
QiZhenGPT (13B)	20.06	20.30	17.50	18.82	17.43	**19.11**	17.13	17.77	18.63	18.33	**19.68**	**20.80**
PULSE (7B)	18.18	17.16	13.93	15.47	15.49	16.44	14.52	15.58	14.14	14.82	16.80	18.18
PULSE (20B)	17.93	17.05	15.05	15.28	15.33	16.38	13.78	14.95	14.82	14.89	15.45	17.44
ChatGLM3-6B	16.89	15.42	13.42	13.86	13.99	15.02	12.84	14.10	14.48	14.46	14.50	15.58
InternLM2-7B	14.69	13.70	11.28	11.00	11.84	13.71	11.16	11.90	12.72	13.47	13.06	15.48
LLaMa3-8B	13.32	12.14	10.57	10.26	11.18	11.32	9.90	10.12	11.10	11.64	11.28	12.88
Qwen1.5-7B	19.62	**20.65**	16.65	17.36	17.08	18.19	16.63	16.89	18.56	17.85	18.19	19.70
Baichuan2-7B	16.80	16.84	14.03	14.66	14.37	15.52	14.21	14.90	15.30	15.29	14.93	16.19
Baichuan2-13B	16.79	15.86	13.09	13.50	13.69	15.21	13.30	13.90	14.88	15.00	14.64	15.80
Qwen1.5-14B	19.49	18.70	16.75	17.90	17.12	17.89	**17.40**	17.12	18.62	17.02	17.87	20.32
InternLM2-20B	9.06	8.38	7.47	7.05	7.34	8.83	6.66	7.41	7.68	8.57	8.26	9.43
Mixtral-8x7B (47B)	14.56	12.94	11.91	11.76	12.16	12.22	11.42	11.51	10.88	12.31	13.10	12.13
LLaMa3-70B	16.21	14.81	13.36	12.98	13.44	15.32	12.81	13.06	14.32	14.72	14.20	14.88
Qwen-72B	**20.57**	20.04	17.55	**18.95**	**18.71**	18.75	17.03	**18.47**	**20.19**	**18.39**	18.81	19.96
Mixtral-8x22B(141B)	17.70	16.65	14.77	15.01	14.67	16.45	14.70	15.54	14.61	15.71	17.10	16.37
GPT3.5-turbo	19.37	18.13	16.45	17.06	17.07	18.24	16.18	17.80	17.11	17.67	18.00	19.55
GLM-4	17.97	16.56	13.68	14.03	15.21	15.96	13.26	14.64	16.00	17.04	16.63	17.33
Qwen-MAX	16.66	15.25	13.13	12.73	14.36	15.42	13.26	13.47	14.56	14.91	14.67	16.49
ERNIE-3.5-8K-0329	19.35	18.08	14.62	15.39	16.06	17.38	14.33	16.37	16.61	17.81	16.67	18.82
GPT-4o	16.20	14.64	12.44	11.85	14.39	15.65	11.69	12.83	14.53	15.11	14.48	16.11

Table 5: Scores of different LLMs for essay/short-answer questions in different disease groups on PediaBench. Here, the highest score in each disease group is highlighted in bold.

GPT-4o achieved the highest total score of 49.53, firmly securing its top position but still far from passing the exam. On the one hand, this confirms that PediaBench is challenging and discriminative enough to assess the applicability of LLMs in pediatric scenarios. On the other hand, this also indicates that, despite the commendable efforts, there remains a huge gap to meet the requirement for deploying LLMs as AI assistants for pediatricians.

2. Injecting medical knowledge can somehow improve the capability of an LLM in generative QA tasks, whereas foundation models also play a more critical role. For objective questions, two medical LLMs, BianQue (7B) and QiZhenGPT (13B), cannot generate answers following the instructions provided. This can be attributed to the limitations of their foundation models. Nevertheless, QiZhenGPT (13B) achieves high ROUGE scores for essay/short-answer questions. Commercial models without medical specialization, however, fall short in essay/short-answer questions. In summary, medical knowledge injection is an effective way to improve the capability of an LLM in generative QA.
3. Small-scale models sometimes show good performance on PediaBench. For example, Qwen1.5-14B, with only 14B parameters, outperforms GPT3.5-turbo and LLaMA3-70B with much more parameters. This could potentially be explained by the fact that GPT3.5-turbo and LLaMA3-70B may lack sufficient exposure to Chinese medical corpora during their training. Nevertheless, large-scale and commercial models still achieve better performance in most cases.

5 Conclusion

In this paper, we introduce PediaBench, a comprehensive Chinese benchmark dataset encompassing 12 pediatric disease groups for LLMs. Specifically, PediaBench consists of 4,565 objective questions and 1,632 subjective questions. It adopts an integrated scoring criterion based on five types of questions to thoroughly assess the proficiency of an LLM in pediatric problems. Finally, the effectiveness of PediaBench is confirmed through extensive experiments on 21 open-source and commercial LLMs. Through an in-depth analysis of experimental results, we offer insights into the ability of LLMs to handle pediatric QA in the Chinese context, highlighting their limitations for further improvements. PediaBench aims to calibrate the current progress of LLMs in pediatrics and facilitates further exploration of medical LLMs.

Limitations

Despite the abundance of pediatric questions in the PediaBench dataset, it still cannot encompass many pediatric diseases and their corresponding treatments in the real world. Therefore, the PediaBench dataset should be maintained with a continual effort for better coverage. Currently, PediaBench focuses mostly on pediatrics. In future work, we plan to extend it to more complex medical departments. Furthermore, we also notice the BLEU

and ROUGE scores face difficulties in accurately gauging the quality of responses to subjective questions. To address this issue, we will explore using more comprehensive evaluation measures, possibly based on GPT-4, or consider incorporating a human evaluation for subjective questions.

Ethics Statement

In this work, all data sources we use to construct the PediaBench dataset are publicly available and free to use. All questions in the PediaBench dataset have been appropriately anonymized so that they do not contain sensitive private information. We do not foresee any other possible negative societal impacts of this work.

References

Jinze Bai, Shuai Bai, Yunfei Chu, et al. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024a. MedBench: A large-scale Chinese benchmark for evaluating medical large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17709–17717.

Zheng Cai, Maosong Cao, Haojiong Chen, et al. 2024b. InternLM2 technical report. *Preprint*, arXiv:2403.17297.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):39:1–39:45.

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. 2024. RareBench: Can LLMs serve as rare diseases specialists? *Preprint*, arXiv:2402.06341.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Huimin Zheng, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, and Xiangmin Xu. 2023a. BianQue: Balancing the questioning and suggestion ability of health LLMs with multi-turn health conversations polished by ChatGPT. *Preprint*, arXiv:2310.15896.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023b. MEDITRON-70B: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

CMKRG et al. 2024. QiZhenGPT: An open source Chinese medical large language model. https://github.com/CMKRG/QiZhenGPT.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. DDXPlus: A new dataset for automatic medical diagnosis. *Advances in Neural Information Processing Systems*, 35:31306–31318.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, et al. 2024. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18099–18107.

Lynette Hirschman and Robert J. Gaizauskas. 2001. Natural language question answering: the view from here. *Natural Language Engineering*, 7(4):275–300.

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. *Advances in Neural Information Processing Systems*, 36:62991–63010.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, et al. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

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

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.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Yunshi Lan and Jing Jiang. 2020. Query graph generation for answering multi-hop complex questions from knowledge bases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 969–974.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. CMMLU: Measuring massive multitask language understanding in Chinese. *Preprint*, arXiv:2306.09212.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023b. Huatuo-26M, a large-scale Chinese medical QA dataset. *Preprint*, arXiv:2305.01526.

Xiaoxi Li, Yujia Zhou, and Zhicheng Dou. 2024. UniGen: A unified generative framework for retrieval and question answering with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8):8688–8696.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023c. ChatDoctor: A medical chat model fine-tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge. *Cureus*, 15(6):e40895.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3):100943.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, LEI ZHU, and Michael Lingzhi Li. 2023. Benchmarking large language models on CMExam - A comprehensive Chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36:52430–52452.

OpenAI et al. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

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*, pages 248–260.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Robert F. Simmons. 1965. Answering English questions by computer: a survey. *Communications of the ACM*, 8(1):53–70.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Yu Sun, Shuohuan Wang, Shikun Feng, et al. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *Preprint*, arXiv:2107.02137.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *The Seventh International Conference on Learning Representations*.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023a. CMB: A comprehensive medical benchmark in Chinese. *Preprint*, arXiv:2308.08833.

Yuqing Wang, Yun Zhao, and Linda Petzold. 2023b. Are large language models ready for healthcare? a comparative study on clinical language understanding. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, pages 804–823.

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

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. DoctorGLM: Fine-tuning your Chinese doctor is not a herculean task. *Preprint*, arXiv:2304.01097.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020.

CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. SuperCLUE: A comprehensive Chinese large language model benchmark. *Preprint*, arXiv:2307.15020.

Aiyuan Yang, Bin Xiao, Bingning Wang, et al. 2023. Baichuan 2: Open large-scale language models. *Preprint*, arXiv:2309.10305.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. Zhongjing: Enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19368–19376.

Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, Fenglin Liu, Meng Cao, Ziming Wang, Xuxin Cheng, Zhu Lei, and Zhenhua Guo. 2023. Qilin-Med: Multi-stage knowledge injection advanced medical large language model. *Preprint*, arXiv:2310.09089.

Wenjing Yue, Xiaoling Wang, Wei Zhu, Ming Guan, Huanran Zheng, Pengfei Wang, Changzhi Sun, and Xin Ma. 2024. TCMBench: A comprehensive benchmark for evaluating large language models in traditional Chinese medicine. *Preprint*, arXiv:2406.01126.

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*.

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.

Xiaofan Zhang, Kui Xue, and Shaoting Zhang. 2023a. PULSE: Pretrained and unified language service engine. Accessed on 2024-06-09.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023b. Evaluating the performance of large language models on gaokao benchmark. *Preprint*, arXiv:2305.12474.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

Wei Zhu, Xiaoling Wang, Huanran Zheng, Mosha Chen, and Buzhou Tang. 2023. PromptCBLUE: A Chinese prompt tuning benchmark for the medical domain. *Preprint*, arXiv:2310.14151.

A Question Types in Medical Exams

Table 6 summarizes the different types of questions in standardized medical examinations. PediaBench covers all these types of questions while also involving pairing questions that are not included in existing medical examinations and benchmarks.

B Comparison of LLMs

Table 7 presents detailed information on the LLMs we evaluate, where "Domain" indicates whether an LLM is of general purpose or specialized in the medical domain, "#Parameters" presents the number of parameters of an LLM ("n/a" for commercial models with disclosed parameter numbers), "Context Window" reveals the size of the context window of an LLM, "How Accessed" indicates how we accessed an LLM for experimentation (open-source models are obtained through their weights and deployed locally on our own servers and commercial models are accessed via their official APIs).

C Prompts

C.1 Prompt for Disease Group Classification

In this section, we present the prompt to classify the disease group of questions in PediaBench using GLM-4. Specifically, we list the 12 disease groups and extract questions that have not yet been categorized. Next, we set restrictive conditions to ensure that GLM-4 only returns the disease group that matches a question without irrelevant information. Finally, we conduct a thorough review of the results labeled with GLM-4 and manually correct the questions that are misclassified. Figure 3 is an example of the prompt for disease group classification and the response of GLM-4.

C.2 Prompt Templates for LLM Evaluation

Figure 4 illustrates the templates of prompts for different question types when evaluating the LLMs in the experiments.

Question Type	Description
A1 (Single-Select Multiple-Choice)	The question consists of a single-sentence stem and five alternative answers, among which only one is the best answer.
A2 (Individual Case Analysis)	The question features a brief, small instance as the stem and a question with five alternative answers, among which the examinee should select the only best answer.
A3/A4 (Multiple Case Analysis)	The question features a brief, small instance as the stem and multiple questions each with five alternative answers, among which the examinee should select the only best answer for each question.
B1 (Shared-Answer)	Given the same five alternative answers and a set of questions, the examinee should select the only best answer for each question.
True or False	Given a sentence or a paragraph, determine whether the statement is accurate or not.
Essay/Short-Answer	Provide a concise answer to the question, with an elaboration on the issue involved.
Case Analysis	Present one or more questions based on a case study of a simulated clinical scenario, requiring the examinee to provide the corresponding answers according to the requirements.

Table 6: Types of questions in standardized medical examinations.

Name	#Parameters	Context Window	Domain	Open Source	How Accessed
BianQue	7B	/	Medical	Yes	Weights
QiZhenGPT	13B	/	Medical	Yes	Weights
PULSE-7B	7B	/	Medical	Yes	Weights
PULSE-20B	20B	/	Medical	Yes	Weights
ChatGLM3-6B	6B	8K	General	Yes	Weights
Baichuan2-7B-Chat	7B	4K	General	Yes	Weights
Qwen1.5-7B	7B	32K	General	Yes	Weights
InternLM2-7B	7B	200K	General	Yes	Weights
Baichuan2-13B-Chat	13B	4K	General	Yes	Weights
Qwen-14B-Chat	14B	32K	General	Yes	Weights
Qwen-72B-Chat	72B	32K	General	Yes	Weights
Mixtral 8x7B	47B	32K	General	Yes	API
InternLM2-20B	20B	200K	General	Yes	API
LLaMa3-70B	70B	8K	General	Yes	API
LLaMa3-8B	8B	8K	General	Yes	API
Mixtral 8x22B	141B	32K	General	Yes	API
Qwen-MAX	n/a	8K	General	No	API
GLM-4	n/a	128K	General	No	API
ERNIE-3.5-8K-0329	n/a	8K	General	No	API
ChatGPT3.5-turbo	n/a	16K	General	No	API
GPT-4o	n/a	128K	General	No	API

Table 7: List of LLMs evaluated in the experiments.

D Additional Results for Performance on Different Disease Groups

Figure 5 shows the performance of LLMs for objective questions in different disease groups. In terms of ToF, the models score higher in addressing neonatal diseases and pediatric infectious diseases, whereas their performance drops slightly in pediatric basic health care and developmental behavioral diseases. In terms of MC, the models show robust capabilities in responding to inquiries related to neonatal diseases and pediatric hematological diseases. However, their skill in handling questions related to pediatric immunodeficiency is comparatively limited. In terms of PA, the models demonstrate exceptional performance in pediatric urinary system diseases and pediatric infectious system diseases. It should be noted that some models, hampered by smaller sizes, display relatively weaker instruction-following abilities, resulting in exceptionally low scores in PA and, in specific cases, achieving zero points for certain disease groups. The detailed results further confirm those presented in Section 4.

E Examples of Model Responses

In this section, we select some representative questions and answers provided by different LLMs. These cases clearly illustrate that LLMs still suffer from obvious shortcomings in pediatric QA.

E.1 True-or-False Question

For true-or-false questions, we compare the answers of PULSE-7B, LLaMA3-8B, Qwen-7B-Chat, and ERNIE-3.5-8K-0329, as shown in Figure 6. After in-depth analysis, we found an interesting phenomenon: models with smaller parameters, such as Qwen-7B-Chat, could judge statements correctly, even though there are errors in their explanations. Furthermore, despite our restrictions on the language generated by the models, the output of

LLaMA3-8B always contains a mixture of Chinese and English.

E.2 Multi-Choice Question

We conducted an in-depth analysis of a multiple-choice question on the disease of infectious system. The answers provided by several LLMs are shown in Figure 7. In this case, LLaMA3-8B and PULSE-20B provided incorrect options and explanations, failing to accurately capture the key points of the question. Baichuan2-7B was even unable to make a reasonable choice. However, ERNIE-3.5-8K-0329 not only selected the correct answer but also provided a detailed and reasonable explanation.

E.3 Pairing Question

The answers to an exemplar pairing question are shown in Figure 8. In this pairing question, only a few LLMs are able to provide complete and accurate results. ChatGLM3-6B's answer contained duplicate answers, while InternLM-20B and Mixtral-8x22B provided incorrect matchings. Unfortunately, PULSE-20B directly copied the order of the given answer sets in the question without performing any effective matching. However, it should be mentioned that GPT-4o provided a completely correct answer to this question, demonstrating its superior performance.

E.4 Essay/Short-Answer Question

Figure 9 presents several responses of LLMs to a essay/short-answer question. Medical LLMs, such as BianQue and QiZhenGPT, tend to provide concise answers, while general-purpose LLMs prefer to elaborate and explain in more detail, resulting in richer content in their answers. In particular, the response from PULSE-20B is not satisfactory, containing a significant amount of repetitive and logically unsound content, which to some extent affects the quality and precision of its response.

E.5 Case Analysis Question

In case analysis questions, the answer to each sub-question serves as the foundation for the solution to the subsequent sub-questions. Therefore, once the answer to the previous question is incorrect, it will further mislead the conclusions of subsequent questions like dominoes. In the example shown in Figure 10, due to BianQue and Baichuan2-13B-Chat's misinterpretation of the first sub-question, their judgments and suggestions for subsequent questions become totally erroneous.

Prompt

ZH: 给定以下儿科疾病组：[肾脏系统疾病，急诊与危重症，感染性疾病，血液系统疾病，心血管系统疾病，免疫系统疾病，呼吸系统疾病，内分泌与代谢系统疾病，儿童保健与发育行为疾病，新生儿疾病，消化系统疾病，神经系统疾病]，请判断下列问题应该归类为哪个疾病组。请注意，你应该选择一个与问题最相关的疾病组作为回答，不能生成其他多余的内容。问题如下：营养性巨幼细胞性贫血是什么病？

EN: Given the following pediatric disease groups: [Diseases of the Renal System, Emergency and Critical Care, Diseases of the Infection System, Diseases of the Blood System, Diseases of the Cardiovascular System, Diseases of the Immune System, Diseases of the Respiratory System, Endocrine or metabolic diseases, Pediatric Health Care, Neonatal Diseases, Diseases of the Alimentary System, Diseases of the Nervous System], please determine which disease group the following question should be classified into. Please note that you should choose the most relevant disease group for the question as your answer, and cannot generate other unnecessary content. The question is as follows: What is nutritional megaloblastic anemia?

Possible Responses From GLM-4

ZH:这个问题既可以归类为[儿童保健与发育行为疾病]，也可以归类为[血液系统疾病]，如果非要选择一个最合适的类别，那可以将该问题归类为[儿童保健与发育行为疾病]，因为营养性巨幼细胞性贫血是由于叶酸和或维生素 B_12 等营养元素缺乏所导致的一种疾病。

EN: This issue can be classified as either [Pediatric Health Care] or [Diseases of the Blood System]. If one must choose the most appropriate category, it can be classified as [Pediatric Health Care], as nutritional megaloblastic anemia is a disease caused by a deficiency of nutrients such as folic acid and/or vitamin B12.

Prompt

ZH: 给定以下儿科疾病组：[肾脏系统疾病，急诊与危重症，感染性疾病，血液系统疾病，心血管系统疾病，免疫系统疾病，呼吸系统疾病，内分泌与代谢系统疾病，儿童保健与发育行为疾病，新生儿疾病，消化系统疾病，神经系统疾病]，请判断下列问题应该归类为哪个疾病组。请注意，你应该选择一个与问题最相关的疾病组作为回答，不能生成其他多余的内容。问题如下：如何诊断原发性肉碱缺乏症？

EN: Given the following pediatric disease groups: [Diseases of the Renal System, Emergency and Critical Care, Diseases of the Infection System, Diseases of the Blood System, Diseases of the Cardiovascular System, Diseases of the Immune System, Diseases of the Respiratory System, Endocrine or metabolic diseases, Pediatric Health Care, Neonatal Diseases, Diseases of the Alimentary System, Diseases of the Nervous System], please determine which disease group the following question should be classified into. Please note that you should choose the most relevant disease group for the question as your answer, and cannot generate other unnecessary content. The question is as follows: How is primary carnitine deficiency diagnosed?

Possible Responses From GLM-4

ZH:内分泌与代谢系统疾病

EN: Endocrine or metabolic diseases

Figure 3: Prompt for disease group classification and an exemplar response of GLM-4.

Types of Questions	Prompt
True or False	ZH: 假如你是一位专业的儿科医生，现在，请用中文判断以下题目表述或做法是否正确。你的回答只能为对或者错，随后简要给出这样判断的理由。题目如下：{题目+选项} EN: Assuming you are a professional pediatrician, please use Chinese to judge whether the following statement or practice is correct or incorrect. Your answer should only be “True" or “False", followed by a brief explanation for your judgment. Here is the question: {question}
Multiple Choice	ZH: 请阅读以下选择题，根据题意选择正确的答案选项。请确保你给出的答案是正确的，并给出简要理由。请注意，你只需要直接给出你认为正确的选项和相应的解释。不能输出其他内容。题目如下：{题目} EN: Please read the following multiple choice questions and choose the correct answer option based on the meaning of the question. Please ensure that the answer you provide is correct and provide a brief explanation. Please note that you only need to directly provide the correct option and corresponding explanation. No other content can be output. The question is as follows: {Question+options}
Pairing	ZH: 请将每一个问题和正确的答案进行匹配，只需要输出最终的匹配结果，严格按照以下python列表的格式输出：[['此处输出提供的问题','此处输出正确的答案'],['此处输出提供的问题','此处输出正确的答案']]。 EN: (Match each question to the correct answer, just output the final match, in the exact format of the following python list: [[' Output question provided here ',' output correct answer here '],[' Output question provided here ',' output correct answer here ']]].)
Essay/Short Answer	ZH: 请用中文回答下列关于儿科疾病的问题。题目如下：{题目} EN: Please answer the following questions about pediatric diseases in Chinese. The questions are as follows: {question}
Case Analysis	ZH: 请用简洁的中文回答下列关于儿科案例分析题。题目如下：{题目} EN: Please answer the following questions about pediatric cases analysis in simple Chinese. The questions are as follows: {Question}

Figure 4: Prompt templates for different question types.

	ND	DAS	DNS	DBS	DCS	DRnS	EMD	DImS	ECC	PHC	DRpS	DInS
Bianque(7B)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
QiZhenGPT(13B)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PULSE(7B)	40.91	27.27	40.00	50.00	61.90	61.90	59.09	52.38	75.00	46.15	52.38	60.00
PULSE(20B)	68.18	50.00	55.00	45.45	71.43	52.38	50.00	52.38	65.00	44.83	61.90	75.00
ChatGLM3-6B	63.64	54.55	55.00	50.00	61.90	47.62	63.64	52.38	40.00	46.15	47.62	60.00
InternLM2-7B	63.64	54.55	60.00	50.00	61.90	42.86	45.45	52.38	65.00	53.85	57.14	70.00
LLaMa3-8B	72.73	45.45	55.00	54.55	52.38	52.38	36.36	52.38	60.00	42.31	71.43	70.00
Qwen1.5-7B	45.45	45.45	60.00	45.45	76.19	52.38	68.18	61.90	60.00	46.15	57.14	60.00
Baichuan2-7B	72.73	54.55	45.00	50.00	66.67	57.14	40.91	52.38	60.00	42.31	66.67	60.00
Baichuan2-13B	54.55	59.09	80.00	50.00	66.67	52.38	45.45	66.67	50.00	46.15	76.19	65.00
Qwen1.5-14B	77.27	59.09	80.00	50.00	66.67	61.90	54.55	80.95	55.00	53.85	66.67	70.00
InternLM2-20B	68.18	54.55	60.00	50.00	66.67	76.19	50.00	47.62	65.00	53.85	61.90	85.00
mixtral-8x7B(47B)	40.91	50.00	55.00	31.82	38.10	47.62	27.27	47.62	40.00	34.62	52.38	65.00
LLaMa3-70B	68.18	81.82	75.00	72.73	76.19	80.95	63.64	66.67	70.00	57.69	85.71	80.00
Qwen-72B	81.82	77.27	85.00	68.18	90.48	85.71	63.64	61.90	60.00	65.38	76.19	80.00
mixtral-8x22B(141B)	63.64	72.73	55.00	72.73	61.90	71.43	54.55	47.62	65.00	61.54	66.67	65.00
GPT3.5-turbo	72.73	54.55	50.00	63.64	57.14	71.43	50.00	61.90	65.00	57.69	71.43	65.00
GLM-4	81.82	72.73	90.00	72.73	76.19	85.71	72.73	80.95	80.00	53.85	80.95	70.00
Qwen-MAX	90.91	77.27	80.00	85.71	85.71	85.71	68.18	66.67	65.00	57.69	80.95	90.00
ERNIE-3.5-8K-0329	86.36	77.27	80.00	86.36	76.19	100.00	72.73	71.43	75.00	76.92	85.71	85.00
GPT-4o	100.00	77.27	85.00	77.27	90.48	90.48	86.36	80.95	70.00	80.77	90.48	90.00

(a) Accuracy on ToF

	ND	DAS	DNS	DBS	DCS	DRnS	EMD	DImS	ECC	PHC	DRpS	DInS
Bianque(7B)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
QiZhenGPT(13B)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PULSE(7B)	37.71	35.12	32.39	39.8	35.76	36.86	32.48	33.56	32.99	42.22	30.72	28.33
PULSE(20B)	51.52	51.84	52.52	48.64	48.1	47.78	55.73	44.07	45.24	45.56	52.9	47.1
ChatGLM3-6B	43.1	42.47	42.14	45.92	36.71	41.64	43.63	40	37.07	38.89	38.91	37.54
InternLM2-7B	55.89	49.5	50.94	52.38	51.9	51.88	56.37	45.76	53.4	51.11	49.83	46.42
LLaMa3-8B	38.72	37.46	37.42	48.3	41.14	32.76	33.12	37.97	39.12	33.7	38.23	34.13
Qwen1.5-7B	52.53	52.51	46.54	45.24	43.67	37.88	44.27	31.19	39.46	27.78	54.61	45.05
Baichuan2-7B	49.16	46.49	47.17	48.98	47.78	48.81	51.27	41.69	45.58	45.19	54.95	42.32
Baichuan2-13B	54.55	52.51	52.83	51.02	50	49.15	52.23	46.1	53.74	44.81	56.31	49.83
Qwen1.5-14B	67.34	66.89	62.26	59.86	63.61	59.73	59.24	59.32	61.22	57.78	66.55	54.61
InternLM2-20B	58.25	57.19	51.26	56.46	49.05	53.58	53.82	47.46	50.68	47.41	53.92	51.19
mixtral-8x7B(47B)	45.79	40.13	46.54	44.22	42.72	39.93	42.36	38.31	44.22	37.78	46.42	47.1
LLaMa3-70B	66.33	63.55	64.78	64.29	66.46	64.16	64.65	63.05	66.67	55.93	65.19	64.85
Qwen-72B	87.54	81.94	85.53	76.87	86.71	78.5	85.67	85.76	81.63	82.22	81.23	80.89
mixtral-8x22B(141B)	52.19	46.82	50	56.12	50	41.98	48.41	36.95	49.32	45.93	49.15	43.69
GPT3.5-turbo	46.13	47.49	49.06	54.42	52.53	47.78	46.18	41.02	51.02	50.37	50.51	46.42
GLM-4	80.81	79.6	75.79	76.87	79.43	74.4	79.94	74.58	76.87	79.63	79.52	75.68
Qwen-MAX	82.83	83.95	83.96	80.61	87.66	80.89	83.12	84.07	83.33	79.26	72.35	78.84
ERNIE-3.5-8K-0329	81.82	78.26	77.67	76.19	82.28	76.79	79.3	76.61	73.47	62.59	67.24	64.51
GPT-4o	84.85	74.92	75.79	80.61	86.39	78.5	79.62	81.69	82.99	77.04	79.86	75.09

(b) Accuracy on MC

	ND	DAS	DNS	DBS	DCS	DRnS	EMD	DImS	ECC	PHC	DRpS	DInS
Bianque(7B)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
QiZhenGPT(13B)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PULSE(7B)	4.76	0.00	4.00	4.17	5.00	8.33	4.76	5.00	0.00	10.00	15.38	15.00
PULSE(20B)	9.52	4.76	8.00	8.33	20.00	20.83	9.52	10.00	4.76	17.50	15.38	10.00
ChatGLM3-6B	4.76	4.76	12.00	4.17	5.00	8.33	4.76	5.00	0.00	12.50	15.38	10.00
InternLM2-7B	9.52	9.52	8.00	20.83	10.00	12.50	9.52	15.00	0.00	15.00	23.08	20.00
LLaMa3-8B	4.76	4.76	12.00	8.33	5.00	12.50	0.00	15.00	0.00	15.00	15.38	15.00
Qwen1.5-7B	4.76	9.52	8.00	12.50	15.00	16.67	9.52	10.00	0.00	12.50	15.38	15.00
Baichuan2-7B	4.76	4.76	8.00	4.17	10.00	8.33	4.76	10.00	0.00	12.50	11.54	5.00
Baichuan2-13B	4.76	9.52	12.00	8.33	15.00	12.50	9.52	10.00	0.00	15.00	23.08	15.00
Qwen1.5-14B	33.33	23.81	36.00	33.33	20.00	50.00	33.33	35.00	19.05	27.50	50.00	40.00
InternLM2-20B	14.29	9.52	16.00	12.50	10.00	12.50	9.52	25.00	14.29	20.00	26.92	15.00
mixtral-8x7B(47B)	4.76	19.05	20.00	16.67	10.00	0.00	9.52	5.00	0.00	15.00	15.38	20.00
LLaMa3-70B	47.62	23.81	44.00	41.67	30.00	41.67	33.33	45.00	23.81	32.50	53.85	55.00
Qwen-72B	42.86	42.86	76.00	54.17	40.00	62.50	52.38	40.00	19.05	35.00	53.85	60.00
mixtral-8x22B(141B)	28.57	33.33	36.00	25.00	15.00	33.33	14.29	15.00	23.81	25.00	26.92	35.00
GPT3.5-turbo	23.81	9.52	28.00	33.33	20.00	45.83	23.81	35.00	9.52	32.50	34.62	30.00
GLM-4	47.62	38.10	60.00	50.00	40.00	58.33	23.81	45.00	23.81	27.50	50.00	70.00
Qwen-MAX	71.43	71.43	72.00	62.50	50.00	58.33	71.43	65.00	57.14	45.00	76.92	75.00
ERNIE-3.5-8K-0329	71.43	61.90	88.00	75.00	50.00	75.00	61.90	55.00	52.38	67.50	76.92	80.00
GPT-4o	66.67	76.19	92.00	83.33	70.00	87.50	71.43	85.00	61.90	75.00	96.15	90.00

(c) Accuracy on PA

Figure 5: Performance on each type of objective questions across disease groups.

Question	假性性早熟除有第二性征发育外，还有性腺的发育。 （The pseudo-precocious puberty is characterized by the development of secondary sexual characteristics as well as the development of the gonads.）
Answer	错（false）
Explanation	假性性早熟的症状是有第二性征的发育、乳房增大、阴毛腋毛出现、无性腺的发育。 （The symptoms of pseudo-precocious puberty include the development of secondary sexual characteristics, breast enlargement, the appearance of pubic and axillary hair, and the absence of gonadal development.）
LLaMA3-8B	错 理由：假性性早熟（Pseudo-Precocious Puberty）是指在正常发育期之前出现第二性征,but it does not necessarily imply development of the sex glands (such as ovaries or testes). The term "假性性早熟" typically refers to the development of secondary sexual characteristics before the expected age of puberty (usually around 10-14 years old), **but it does not necessarily imply the development of the sex organs themselves.**
Qwen-7B-chat	对。假性性早熟是儿童内分泌系统疾病的一种，除了出现第二性征发育外，还会伴随着性腺的提前发育，如乳房增大、阴毛生长等。 （True. Pseudo-precocious puberty is a type of childhood endocrine system disorder that, in addition to the development of secondary sexual characteristics, is also accompanied by premature development of the gonads, such as breast enlargement and the growth of pubic hair.）
PULSE-7B	Helper: 假性性早熟除有第二性征发育外，还有性腺的发育。 这个表述是正确的。假性性早熟是指在8岁前女孩出现乳房发育、月经初潮、阴毛生长等第二性征，但骨龄提前，而骨龄提前是由于性激素分泌过多引起的。性腺的发育是指卵巢和睾丸的发育，这是假性性早熟的另一个特征。因此，假性性早熟不仅有第二性征发育，还有性腺的发育。 （Helper: The statement that pseudo-precocious puberty involves not only the development of secondary sexual characteristics but also the development of the gonads is correct. Pseudo-precocious puberty refers to the appearance of secondary sexual characteristics such as breast development, menarche, and pubic hair growth in girls before the age of 8, accompanied by advanced bone age due to excessive secretion of sex hormones. The development of the gonads, referring to the development of the ovaries and testes, is another characteristic of pseudo-precocious puberty. Therefore, pseudo-precocious puberty indeed involves both the development of secondary sexual characteristics and the development of the gonads.）
ERNIE-3.5-8K-0329	错 解释：假性性早熟的特点是第二性征发育与性腺发育步调不一致，即仅有第二性征发育，但性腺（如卵巢或睾丸）并未达到与年龄相符的成熟状态。因此，**假性性早熟中并不伴有性腺的发育。** Wrong. Explanation: The characteristic of pseudo-precocious puberty is that the development of secondary sexual characteristics and the development of the gonads are inconsistent, meaning that there is only the development of secondary sexual characteristics, but the gonads (such as the ovaries or testes) have not reached a maturity level commensurate with the child's age. Therefore, pseudo-precocious puberty does not involve the development of the gonads.

***注：绿色字体代表回答的关键正确部分，红色字体代表回答明显出错的部分。**
***Note: The green font represents the key correct part of response, while the red font indicates the part of response that is obviously incorrect.**

Figure 6: Answers and explanation notes of an exemplar true-or-false question.

Question	患儿男，1岁8个月。咳嗽15天伴加剧6天，呈阵发性痉挛性咳嗽伴鸡鸣样回声，以夜间为甚，病初低热3天。体检:体温正常，咽红，舌系带溃疡，呼吸音粗。X线胸片显示间质改变，外周血WBC13.8×10^9/L，中性粒细胞0.18，淋巴细胞0.82，未接种过疫苗。首选的抗感染药物是 A.红霉素 B.氯霉素 C.阿莫西林 D.复方磺胺甲噁唑 E.头孢菌素 (The patient is a 1-year-old and 8-month-old male child. He has been coughing for 15 days, with symptoms worsening in the past 6 days, presenting as paroxysmal spasmodic cough with crowing-like echo, which is more severe at night. There was low fever for 3 days at the beginning of the illness. Physical examination: normal body temperature, red pharynx, ulcerated frenulum of tongue, and coarse breath sounds. X-ray chest film shows interstitial changes. Peripheral blood WBC 13.8×10^9/L, neutrophils 0.18, lymphocytes 0.82. The child has not been vaccinated. The preferred anti-infective drug is A. Erythromycin B. Chloramphenicol C. Amoxicillin D. Compound sulfamethoxazole E. Cephalosporin)
Answer	A.红霉素 (A. Erythromycin)
Explanation	患儿为幼儿，病程较长，以咳嗽为主要症状，特征为阵发性痉挛性咳嗽伴鸡鸣样回声，有舌系带溃疡，为痉咳时舌外伸，舌系带反复摩擦门齿所致。外周血WBC增高且以淋巴细胞分类为主，而胸片仅为间质改变，无百白破疫苗接种史，故百日咳可能性大。百日咳可并发吸入性肺炎、支气管肺炎、肺气肿、百日咳脑病、营养不良及结核病恶化等，但以并发支气管肺炎最常见。虽红霉素、氯霉素及阿莫西林均可选用，但以红霉素为首选抗感染药物。 (The patient is a young child with a long duration of illness, and the main symptom is cough. The characteristics are paroxysmal spasmodic cough with crowing-like echo, and frenulum ulceration due to repeated friction between the protruding tongue and the incisor teeth during coughing spasms. The peripheral blood WBC is elevated, mainly lymphocytes, and the chest X-ray shows only interstitial changes. Since there is no history of vaccination for diphtheria, pertussis, and tetanus, pertussis is highly likely. Pertussis can be complicated by aspiration pneumonia, bronchopneumonia, pulmonary emphysema, pertussis encephalopathy, malnutrition, and exacerbation of tuberculosis, but bronchopneumonia is the most common complication. Although erythromycin, chloramphenicol, and amoxicillin can all be used, erythromycin is the preferred anti-infective drug.)
LLaMA3-8B	患儿为幼儿，病程较长，以咳嗽为主要症状，特征为阵发性痉挛性咳嗽伴鸡鸣样回声，有舌系带溃疡，为痉咳时舌外伸，舌系带反复摩擦门齿所致。外周血WBC增高且以淋巴细胞分类为主，而胸片仅为间质改变，无百白破疫苗接种史，故百日咳可能性大。百日咳可并发吸入性肺炎、支气管肺炎、肺气肿、百日咳脑病、营养不良及结核病恶化等，但以并发支气管肺炎最常见。虽红霉素、氯霉素及阿莫西林均可选用，但以红霉素为首选抗感染药物。 (The patient is a young child with a long duration of illness, and the main symptom is cough. The characteristics are paroxysmal spasmodic cough with crowing-like echo, and frenulum ulceration due to repeated friction between the protruding tongue and the incisor teeth during coughing spasms. The peripheral blood WBC is elevated, mainly lymphocytes, and the chest X-ray shows only interstitial changes. Since there is no history of vaccination for diphtheria, pertussis, and tetanus, pertussis is highly likely. Pertussis can be complicated by aspiration pneumonia, bronchopneumonia, pulmonary emphysema, pertussis encephalopathy, malnutrition, and exacerbation of tuberculosis, but bronchopneumonia is the most common complication. Although erythromycin, chloramphenicol, and amoxicillin can all be used, erythromycin is the preferred anti-infective drug.)
LLaMA3-8B	D. 复方磺胺甲噁唑 理由：患儿的症状表明其可能患有 streptococcus pneumoniae 导致的社区获得性肺炎，具有患者的呼吸音粗、咽红、舌系带溃疡等特征，且X线胸片显示的间质改变进一步支持这种判断。红霉素无法治疗社区获得性肺炎，氯霉素可以治疗，但在儿童中可能会导致听力损害，阿莫西林对社区获得性肺炎效果不清，头孢菌素对社区获得性肺炎效果更差。复方磺胺甲噁唑是常用的社区获得性肺炎首选抗感染药物。 (D. Compound sulfamethoxazole Reason: The child's symptoms indicate that he may suffer from community-acquired pneumonia caused by streptococcus pneumoniae, characterized by coarse breath sounds, red pharynx, frenulum ulceration, etc. The interstitial changes shown on the X-ray chest film further support this judgment. Erythromycin cannot treat community-acquired pneumonia, chloramphenicol can treat it but may cause hearing impairment in children, the effect of amoxicillin on community-acquired pneumonia is unclear, and the effect of cephalosporins on community-acquired pneumonia is even worse. Compound sulfamethoxazole is a commonly used preferred anti-infective drug for community-acquired pneumonia.)
Baichuan2-7B-Chat	首先，我们需要了解患儿的症状和体征以及可能的病因。患儿有阵发性痉挛性咳嗽、鸡鸣样回声、咽红、舌系带溃疡等表现，这可能与病毒性感染或过敏性咳嗽有关。体检发现体温正常、呼吸音粗等，提示可能有呼吸道感染。X线胸片显示间质改变，外周血WBC升高，中性粒细胞降低，淋巴细胞升高，这些指标支持细菌感染的可能。由于患儿未接种过疫苗，可能存在免疫力低下的情况。 根据以上症状和体征，我们可以初步判断患儿可能患有急性支气管炎，一种常见的呼吸道感染。在这种情况下，首选的抗感染药物应该是能够覆盖细菌感染的抗生素，如头孢菌素类。然而，题目中没有提供足够的信息来确定具体的抗生素选择。因此，我们无法确定首选的抗感染药物是A、B、C还是E。 (First and foremost, we need to understand the symptoms, physical signs, and possible causes of the illness in the child. The child presents with paroxysmal spasmodic cough, crowing-like echo, red pharynx, and frenulum ulceration, which may be related to viral infection or allergic cough. The physical examination reveals a normal body temperature and coarse breath sounds, indicating a possible respiratory tract infection. The X-ray chest film shows interstitial changes, and the peripheral blood WBC is elevated with a decrease in neutrophils and an increase in lymphocytes, which support the possibility of bacterial infection. As the child has not been vaccinated, there may be a weakened immune system. Based on the above symptoms and physical signs, we can initially diagnose the child with acute bronchitis, a common respiratory tract infection. In this case, the preferred anti-infective drug should be an antibiotic that can cover bacterial infections, such as cephalosporins. However, the question does not provide sufficient information to determine the specific antibiotic choice. Therefore, we cannot determine whether the preferred anti-infective drug is A, B, C, or E.)
PULSE-7B	Helper: E.头孢菌素 根据患儿症状和体检结果，可能存在呼吸道感染。首选的抗感染药物应该是针对细菌感染的抗生素，而头孢菌素是一种广谱抗生素，对呼吸道感染的病原体有较好的抗菌活性，因此是首选药物。其他选项如红霉素、氯霉素、阿莫西林、复方磺胺甲噁唑等抗生素可能对病原体无效，因此不是首选药物。 (Helper: E. Cephalosporin Based on the symptoms and physical examination results of the child, there may be a respiratory tract infection. The preferred anti-infective drug should be an antibiotic targeting bacterial infections. Cephalosporin is a broad-spectrum antibiotic with good antibacterial activity against pathogens causing respiratory tract infections, thus making it the first choice. Other options such as erythromycin, chloramphenicol, amoxicillin, and compound sulfamethoxazole may be ineffective against the pathogens, therefore they are not the preferred drugs.)
ERNIE-3.5-8K-0329	A.红霉素。 解释：患儿男，1岁8个月。出现咳嗽、阵发性痉挛性咳嗽伴鸡鸣样回声，为百日咳典型症状，病初低热，X线胸片显示间质改变，外周血白细胞升高，淋巴细胞比例升高，根据患儿临床表现和实验室检查，该患儿应诊断为百日咳。百日咳的病原菌为百日咳鲍特菌，为细菌的一种，属于鲍特菌属，侵入呼吸道黏膜在纤毛上皮进行繁殖并产生毒素，引起纤毛坏死、黏膜炎症等病变，并引起全身反应。百日咳首选的抗感染药物是**红霉素**，可渗透呼吸道黏膜，达到有效浓度，控制感染。 选项B氯霉素主要不良反应是抑制骨髓造血机能，包括可逆的各类血细胞减少和不可逆的再生障碍性贫血，其次为消化道反应和二重感染等，临床上较少使用，故排除。 选项C阿莫西林为青霉素类抗生素，对百日咳杆菌效果不确切，故不选。 选项D复方磺胺甲噁唑为磺胺类抗生素，主要用于敏感菌引起的尿路感染、肠道感染、成人慢性支气管炎急性发作等，对百日咳治疗效果不佳，故不考虑。 选项E头孢菌素属于β-内酰胺类抗生素，对百日咳杆菌效果不确切，可排除。

***注：绿色字体代表回答的关键正确部分，红色字体代表回答明显出错的部分。**

***Note: The green font represents the key correct part of response, while the red font indicates the part of response that is obviously incorrect.**

Figure 7: Answers and explanation notes of an exemplar multi-choice question.

Question	请将每一个问题和正确的答案进行匹配，只需要输出最终的匹配结果，严格按照以下python列表的格式输出：[['此处输出提供的问题','此处输出正确的答案'],['此处输出提供的问题','此处输出正确的答案']]。题目为：["新生儿常见的特殊生理状态之一是（）。","新生儿生理性黄疸的主要原因是（）。","早产儿的呼吸特点是（）。","早产儿的外观特点之一是（）。","足月儿的外观特点之一是（）。"]，对应的答案组为["皮肤色红润，皮下脂肪丰满，囊毛少","足月新生儿生后3～5天出现乳腺肿大","皮肤发亮，水肿囊毛多","呼吸常不规则，甚至呼吸暂停","肝葡萄糖醛酸基转移酶活性低"] (Match each question to the correct answer, just output the final match, in the exact format of the following python list: [[' Output question provided here ',' output correct answer here '],[' Output question provided here ',' output correct answer here ']]. The question is: ["One of the common special physiological states of newborns is.","The main cause of physiological jaundice in newborns is.","The respiratory characteristics of premature infants are.","One of the appearance characteristics of premature infants is.","One of the appearance characteristics of full-term infants is. "], and the corresponding answer group is: ["The skin is ruddy, the subcutaneous fat is plump, and the hair is less.","The breast enlargement occurs in full-term newborns 3-5 days after birth.","The skin is shiny, the edema is plump, and the hair is more.","The breathing is often irregular, even apnea.","The activity of hepatic glucuronyl transferase is low"])
Answer	[['新生儿常见的特殊生理状态之一是（）。','足月新生儿生后3～5天出现乳腺肿大'], ['新生儿生理性黄疸的主要原因是（）。','肝葡萄糖醛酸基转移酶活性低'], ['早产儿的呼吸特点是（）。','呼吸常不规则，甚至呼吸暂停'], ['早产儿的外观特点之一是（）。','皮肤发亮，水肿囊毛多'], ['足月儿的外观特点之一是（）。','皮肤色红润，皮下脂肪丰满，囊毛少']] ([['One of the common special physiological states of newborns is', 'the enlargement of the breast glands in full-term newborns 3-5 days after birth'], ['The main cause of physiological jaundice in newborns is', 'low activity of hepatic glucuronyltransferase'], ['The respiratory characteristics of premature infants are', 'often irregular breathing, even apnea'], ['One of the appearance characteristics of premature infants is', 'bright skin, edema, and many fine hairs'], ['One of the appearance characteristics of full-term infants is', 'ruddy skin, plump subcutaneous fat, and few fine hairs']])
ChatGLM3-6B	[['新生儿常见的特殊生理状态之一是（皮肤色红润，皮下脂肪丰满，囊毛少）', '新生儿生理性黄疸的主要原因是（肝葡萄糖醛酸基转移酶活性低）'], ['早产儿的呼吸特点是（呼吸常不规则，甚至呼吸暂停）', '早产儿的外观特点之一是（肝葡萄糖醛酸基转移酶活性低）'], ['足月儿的外观特点之一是（皮肤发亮，水肿囊毛多）', '早产儿的呼吸特点是（呼吸常不规则，甚至呼吸暂停）'], ['足月儿的外观特点之一是（肝葡萄糖醛酸基转移酶活性低）']] [['One of the common special physiological states of newborns is a ruddy skin, plump subcutaneous fat, and few fine hairs', 'The main cause of physiological jaundice in newborns is low activity of hepatic glucuronyl transferase'], ['The respiratory characteristics of premature infants are often irregular breathing, even apnea', 'One of the appearance characteristics of premature infants is low activity of hepatic glucuronyl transferase'], ['One of the appearance characteristics of full-term infants is bright skin, edema, and many fine hairs', 'The respiratory characteristics of premature infants are often irregular breathing, even apnea', 'One of the appearance characteristics of full-term infants is low activity of hepatic glucuronyl transferase']]

InternLM-20B	以下是您提供的问题和答案的匹配结果：
	``python
	[
	['新生儿常见的特殊生理状态之一是（）。','皮肤色红润，皮下脂肪丰满，囊毛少'],
	['新生儿生理性黄疸的主要原因是（）。','肝葡萄糖醛酸基转移酶活性低'],
	['早产儿的呼吸特点是（）。','呼吸常不规则，甚至呼吸暂停'],
	['早产儿的外观特点之一是（）。','皮肤发亮，水肿囊毛多'],
	['足月儿的外观特点之一是（）。','足月新生儿生后3～5天出现乳腺肿大']
	]
	``
	注意，这里的匹配结果基于您提供的问题和答案，如果问题和答案之间存在任何错误或不一致，匹配结果可能会受到影响。
		Here are the matching results of your questions and answers:
		``python
		[One of the common special physiological states of newborns is that their skin is ruddy, their subcutaneous fat is plump, and their hair is sparse,
		The main cause of physiological jaundice in newborns is low activity of hepatic glucuronyltransferase,
		The respiratory characteristics of premature infants are often irregular, even with episodes of apnea,
		One of the appearance characteristics of premature infants is that their skin is shiny, swollen, and covered with fine hair,
		One of the external features of full-term infants is that they will experience breast enlargement 3 to 5 days after birth]
		``
		Please note that the matching results here are based on the questions and answers you provide. If there are any errors or inconsistencies between the questions and answers, the matching results may be affected.

PULSE-20B	[['新生儿常见的特殊生理状态之一是（）。','皮肤色红润，皮下脂肪丰满，囊毛少'], ['新生儿生理性黄疸的主要原因是（）。','足月新生儿生后3～5天出现乳腺肿大'], ['早产儿的呼吸特点是（）。','皮肤发亮，水肿囊毛多'], ['早产儿的外观特点之一是（）。','呼吸常不规则，甚至呼吸暂停'], ['足月儿的外观特点之一是（）。','肝葡萄糖醛酸基转移酶活性低']] [['One of the common special physiological states of newborns is', 'ruddy skin, plump subcutaneous fat, and few vellus hair'], ['The main cause of physiological jaundice in newborns is', 'the appearance of breast enlargement in full-term newborns 3-5 days after birth'], ['The respiratory characteristics of premature infants are', 'bright skin, edema, and many vellus hair'], ['One of the appearance characteristics of premature infants is', 'often irregular breathing, even apnea'], ['One of the appearance characteristics of full-term infants is', 'low activity of hepatic glucuronyltransferase']]
Mixtral-8x22B	[['新生儿常见的特殊生理状态之一是（）。','皮肤色红润，皮下脂肪丰满，囊毛少'], ['新生儿生理性黄疸的主要原因是（）。','肝葡萄糖醛酸基转移酶活性低'], ['早产儿的呼吸特点是（）。','呼吸常不规则，甚至呼吸暂停'], ['早产儿的外观特点之一是（）。','皮肤发亮，水肿囊毛多'], ['足月儿的外观特点之一是（）。','足月新生儿生后3～5天出现乳腺肿大']] [['One of the common special physiological states of newborns is', 'ruddy skin, plump subcutaneous fat, and few vellus hair'], ['The main cause of physiological jaundice in newborns is', 'low activity of hepatic glucuronyl transferase'], ['The respiratory characteristics of premature infants are', 'often irregular breathing, even apnea'], ['One of the appearance characteristics of premature infants is', 'bright skin, edema, and many vellus hair'], ['One of the appearance characteristics of full-term infants is', 'breast enlargement appears 3-5 days after birth in full-term newborns']]
GPT-4o	[['新生儿常见的特殊生理状态之一是（）。','足月新生儿生后3～5天出现乳腺肿大'], ['新生儿生理性黄疸的主要原因是（）。','肝葡萄糖醛酸基转移酶活性低'], ['早产儿的呼吸特点是（）。','呼吸常不规则，甚至呼吸暂停'], ['早产儿的外观特点之一是（）。','皮肤发亮，水肿囊毛多'], ['足月儿的外观特点之一是（）。','皮肤色红润，皮下脂肪丰满，囊毛少']] [['One of the common special physiological states of newborns is', 'mammary gland enlargement in full-term newborns 3-5 days after birth'], ['The main cause of physiological jaundice in newborns is', 'low activity of hepatic glucuronyltransferase'], ['The respiratory characteristics of premature infants are', 'often irregular breathing, even apnea'], ['One of the appearance characteristics of premature infants is', 'bright skin, edema, and many fine hairs'], ['One of the appearance characteristics of full-term infants is', 'ruddy skin, plump subcutaneous fat, and few fine hairs']]

***注：绿色字体代表回答的关键正确部分，红色字体代表回答明显出错的部分。**

***Note: The green font represents the key correct part of response, while the red font indicates the part of response that is obviously incorrect.**

Figure 8: Answers and explanation notes of an exemplar pairing question.

Question	儿童风湿性疾病的自身抗体检查有哪些？ (What are the autoantibody tests for pediatric rheumatic diseases?)	
Answer	儿童风湿性疾病的自身抗体检查有1. 抗核抗体(ANA)检测；2. 抗双链 DNA(dsDNA)抗体检测；3. 抗着丝点抗体(ACA)检测；4. 抗 ENA 抗体检测；5.抗组蛋白抗体(AHA)检测；6.抗核小抗体检测；7. 抗磷脂抗体谱；8. 抗中性粒细胞胞浆抗体谱；9. 类风湿因子(RF)检测；10. 抗角蛋白抗体(AKA)检测；11.抗核周因子(APF)抗体检测；12.抗环瓜氨酸肽(CCP)抗体检测；13.葡萄糖 -6- 磷酸异构酶(GPI)检测；14.抗 C1q 抗体检测。 (Autoantibody tests for pediatric rheumatic diseases include: 1. Antinuclear antibody (ANA) test; 2. Anti-double-stranded DNA (dsDNA) antibody test; 3. Anti-centromere antibody (ACA) test; 4. Anti-ENA antibody test; 5. Anti-histone antibody (AHA) test; 6. Anti-nucleosome antibody test; 7. Antiphospholipid antibody spectrum; 8. Anti-neutrophil cytoplasmic antibody spectrum; 9. Rheumatoid factor (RF) test; 10. Anti-keratin antibody (AKA) test; 11. Anti-perinuclear factor (APF) antibody test; 12. Anti-cyclic citrullinated peptide (CCP) antibody test; 13. Glucose-6-phosphate isomerase (GPI) test; 14. Anti-C1q antibody test.)	
Bianque-7B	儿童风湿性疾病的自身抗体检查包括： 1. 抗核抗体(ANA)和抗平滑肌抗体(RFA) 2. 抗中性粒细胞胞浆抗体(NSA) 3. 抗磷脂抗体(anti-cardiolipin) 4. 抗β2-微球蛋白抗体(β2-MG) 5. 抗β1-微球蛋白抗体(β1-MG) 6. 抗组蛋白抗体(anti-follicular unit extraction) 7. 抗单链肌酸激酶(CK) 8. 抗单链肌酸激酶同工酶(CK-MB) 9. 抗β2-微球蛋白抗体(β2-MG) 10. 抗单链肌酸激酶同工酶(CK-MB) 以上是常见的儿童风湿性疾病的自身抗体检查项目，具体需要根据医生的建议进行选择。	The autoantibody tests for pediatric rheumatic diseases include: Antinuclear antibody (ANA) and anti-smooth muscle antibody (RFA) Anti-neutrophil cytoplasmic antibody (NSA) Anti-phospholipid antibody (anti-cardiolipin) Anti-β2-microglobulin antibody (β2-MG) Anti-β1-microglobulin antibody (β1-MG) (Note: "Anti-histone antibody" is typically referred to as "AHA" rather than "anti-follicular unit extraction") Anti-single-chain creatine kinase (CK) Anti-single-chain creatine kinase isoenzyme (CK-MB) (Note: This is a repeat of Anti-β2-microglobulin antibody (β2-MG) mentioned in item 4) (Note: This is a repeat of Anti-single-chain creatine kinase isoenzyme (CK-MB) mentioned in item 8) The above are common autoantibody tests for pediatric rheumatic diseases, and specific tests should be selected based on the doctor's recommendations.
Qizhen-13B	儿童风湿性疾病的自身抗体检查主要包括抗核抗体、类风湿因子、抗中性粒细胞胞浆抗体、抗心磷脂抗体、抗核周因子、抗组蛋白抗体、抗核抗原抗体等。 (The autoantibody tests for pediatric rheumatic diseases mainly include antinuclear antibodies (ANA), rheumatoid factor (RF), anti-neutrophil cytoplasmic antibodies (ANCA), anti-cardiolipin antibodies, anti-perinuclear factor (APF), anti-histone antibodies, and anti-nuclear antigen antibodies, etc.)	
Mixtral-8x7B	儿童风湿性疾病的自身抗体检查有： 1. 反 Rheumatoid Factor（RF）：风湿因子是一种抗体，通常在多种自身免疫系统疾病中可检测到，包括 Still 综合征、系统性红斑狼疮和干燥综合征等。 2. 抗核抗体（ANA）：ANA 是一种常见的自身抗体，可在 verschiedene Autoimmunerkrankungen wie systemischer Lupus erythematodes (SLE) und Sjögren-Syndrom gefunden werden. 3. 抗双珠蛋白抗体（dsDNA-AK）：dsDNA-AK ist ein spezifischer ANA-Subtyp und tritt hauptsächlich bei SLE auf. 4. 抗extractable nuclear antigen antibodies (ENA)：ENA-Antikörper sind eine Gruppe von Autoantikörpern, die gegen verschiedene Kernantigene gerichtet sind, wie Sm, RNP, SSA und SSB. Sie werden bei verschiedenen Autoimmunerkrankungen wie SLE, Sjögren-Syndrom und Mixed Connective Tissue Disease (MCTD) gefunden. 5. 抗цикличеskih citrullinated peptide antibodies (ACPA)：ACPA wird bei rheumatoider Arthritis (RA) gefunden und ist spezifisch für diese Erkrankung. 6. 抗Jo-1抗体：Jo-1-Antikörper werden bei einer Untergruppe von Autoimmunerkrankungen gefunden, die als polymyositis und dermatomyositis bekannt sind. 7. 抗Scl-70抗体：Scl-70-Antikörper werden hauptsächlich bei Sklerodermie gefunden. Diese Selbstantikörper können bei der Diagnose von Autoimmunerkrankungen bei Kindern hilfreich sein, obwohl sie nicht immer spezifisch für eine bestimmte Erkrankung sind und andere diagnostische Tests und klinische Bewertungen erforderlich sein können.	
PULSE-20B	儿童风湿性疾病的自身抗体检查包括： 1. 抗核抗体（ANA）：是一种自身抗体，可以检测到多种自身免疫性疾病，包括风湿性关节炎、系统性红斑狼疮等。 2. 抗双链DNA抗体（ds-DNA）：是一种自身抗体，可以检测到系统性红斑狼疮等自身免疫性疾病。 3. 抗Sm抗体：是一种自身抗体，可以检测到系统性红斑狼疮等自身免疫性疾病。 4. 抗SSA抗体：是一种自身抗体，可以检测到干燥综合征等自身免疫性疾病。 5. 抗SSB抗体：是一种自身抗体，可以检测到干燥综合征等自身免疫性疾病。 这些自身抗体检查可以帮助医生诊断儿童风湿性疾病的类型和严重程度，并制定相应的治疗方案。 (The autoantibody tests for pediatric rheumatic diseases include: Antinuclear antibody (ANA): It is an autoantibody that can detect multiple autoimmune diseases, including rheumatoid arthritis, systemic lupus erythematosus, etc. Anti-double-stranded DNA antibody (ds-DNA): It is an autoantibody that can detect autoimmune diseases such as systemic lupus erythematosus. Anti-Sm antibody: It is an autoantibody that can detect autoimmune diseases such as systemic lupus erythematosus. Anti-SSA antibody: It is an autoantibody that can detect autoimmune diseases such as Sjögren's syndrome. Anti-SSB antibody: It is an autoantibody that can detect autoimmune diseases such as Sjögren's syndrome. These autoantibody tests can help doctors diagnose the type and severity of pediatric rheumatic diseases and formulate corresponding treatment plans.)	

Figure 9: Answers and explanation notes of an exemplar essay/short-answer question.

Question	男婴，第一胎第一产，孕 40 周，3 天龄，生后 18 小时发现黄疸，并进行性加重。反应，吃奶好，生后 12 小时解胎便，目前大便已转黄色。查体：反应好，皮肤、巩膜重度黄疸，心肺体检未见异常，腹平软，肝肋下 3cm，脐带无渗液。查血胆红素 342μmol/L，直接胆红素 18μmol/L，血红蛋白 135g/L。血型为"A"，其母为"O"，抗体释放试验阳性。问题： 一、该患儿最可能的诊断是什么。 二、该病的病因是什么？ 三、简述该病的治疗方法。 (Male infant, first pregnancy and first delivery, 40 weeks gestation, 3 days old. Jaundice was found 18 hours after birth and progressively worsened. The infant has a good response and appetite for breastfeeding. Meconium passed 12 hours after birth, and the stool has turned yellow. Physical examination: good response, severe jaundice of skin and sclera, no abnormality in heart and lung examination, abdomen is flat and soft, liver palpable 3cm below the rib cage, no exudate from the umbilical cord. Blood test results: bilirubin 342μmol/L, direct bilirubin 18μmol/L, hemoglobin 135g/L. The blood type of the infant is "A" and the mother's is "O". Antibody release test is positive. Questions: What is the most likely diagnosis for this infant? What is the cause of this disease? Briefly describe the treatment methods for this disease.)
Answer	一、该患儿诊断为新生儿 ABO 溶血症。 二、病因为由父亲遗传而母亲所不具有的显性胎儿红细胞血型抗原通过胎盘进入母体，刺激母体产生相应的血型抗体并通过胎盘进入胎儿血循环，与致敏红细胞结合发生溶血。 三、治疗方法包括：①光照疗法，降低血清未结合胆红素；②药物治疗如供给白蛋白、纠正酸中毒、肝酶诱导剂；③换血疗法；④其他治疗：防止低血糖、低体温，纠正贫血，缺氧等. (1.The diagnosis for this infant is ABO hemolytic disease of the newborn. 2.The cause of this disease is that the dominant fetal red blood cell blood group antigen, which is inherited from the father but not possessed by the mother, enters the mother's body through the placenta, stimulating the mother to produce corresponding blood group antibodies. These antibodies then enter the fetal blood circulation through the placenta, binding with sensitized red blood cells and causing hemolysis. 3.The treatment methods include: ① Phototherapy to reduce unbound bilirubin in the serum; ② Medication such as albumin supplementation, acidosis correction, and liver enzyme inducers; ③ Blood exchange therapy; ④ Other treatments: prevention of hypoglycemia, hypothermia, correction of anemia, hypoxia, etc.)
Qizhen-13B	该患儿最可能的诊断是新生儿溶血病，病因是母婴血型不合引起的同族免疫性溶血，治疗方法是光照疗法、药物治疗、换血疗法等。 新生儿溶血病是由于母婴血型不合引起的同族免疫性溶血，临床上以ABO血型不合最常见，Rh血型不合较少见。患儿生后可出现黄疸、贫血、肝脾肿大等症状。本病的治疗方法有光照疗法、药物治疗、换血疗法等。 1.光照疗法：当血清胆红素水平为151 ~ 342μmol/L时，可选择光疗。当血清胆红素水平超过342μmol/L时，禁忌光疗。 2.药物治疗：静脉注射促红细胞生成素可促进红细胞增殖，纠正贫血。肝酶诱导剂可减轻肝脏负担。 3.换血疗法：当血清胆红素水平超过342μmol/L时，可进行换血疗法。 The most likely diagnosis for this infant is hemolytic disease of the newborn, which is caused by the incompatible blood types between the mother and the infant leading to isoimmune hemolysis. The treatment methods include phototherapy, medication, and blood exchange therapy. Neonatal hemolytic disease is caused by incompatible blood types between the mother and the infant, leading to isoimmune hemolysis. ABO blood type incompatibility is the most common clinically, while Rh blood type incompatibility is less common. Infants may experience symptoms such as jaundice, anemia, hepatomegaly, and splenomegaly after birth. The treatment methods for this disease include phototherapy, medication, and blood exchange therapy. Phototherapy: Phototherapy can be selected when the serum bilirubin level is between 151 and 342 μmol/L. However, phototherapy is contraindicated when the serum bilirubin level exceeds 342 μmol/L. Medication: Intravenous injection of erythropoietin can promote red blood cell proliferation and correct anemia. Liver enzyme inducers can reduce the burden on the liver. Blood exchange therapy: Blood exchange therapy can be performed when the serum bilirubin level exceeds 342 μmol/L.
Bianque(7B)	1. 最可能的诊断是新生儿肝炎综合症。该病是由于新生儿肝脏功能未完全发育，导致胆汁不能正常排出，从而引起黄疸。病因可能是病毒感染、细菌感染、遗传代谢性疾病等。 2. 病因可能是病毒感染、细菌感染、遗传代谢性疾病等。 3. 治疗方法包括药物治疗、营养支持等。具体治疗方案需要根据患儿的具体情况而定。 以上是一些可能的诊断和治疗方法，具体治疗方案需要根据患儿的具体情况而定。建议您尽快就医，由专业医生进行诊断和治疗。 The most probable diagnosis is neonatal hepatitis syndrome. This condition is caused by the incomplete development of the liver function in newborns, resulting in the inability of bile to be discharged normally, thus leading to jaundice. The causes may include viral infection, bacterial infection, inherited metabolic diseases, etc. The potential causes may include viral infection, bacterial infection, inherited metabolic diseases, etc. The treatment methods include medication and nutritional support. The specific treatment plan needs to be determined based on the specific condition of the infant. The above are some possible diagnoses and treatment methods, but the specific treatment plan needs to be tailored to the infant's specific condition. It is recommended that you seek medical attention as soon as possible and allow a professional doctor to conduct the diagnosis and treatment.
Baichuan2-13B-chat	一、最可能的诊断是新生儿病理性黄疸（母乳性黄疸）。 二、病因是胎儿期母亲与胎儿血液通过胎盘接触，使胎儿产生对抗母亲抗体的免疫应答，导致出生后新生儿与母乳中抗体结合，阻塞胆红素代谢，引起黄疸。 三、治疗方法是暂停母乳喂养，改用其他配方奶粉喂养，以降低母乳中抗体浓度，促进黄疸消退；同时观察患儿黄疸变化和肝功能恢复情况。 The most likely diagnosis is neonatal pathological jaundice (breast milk jaundice). The cause is the contact between the mother's and fetus's blood through the placenta during the fetal period, which prompts the fetus to produce an immune response against the mother's antibodies. This results in the newborn's antibodies binding with those in breast milk after birth, blocking bilirubin metabolism and causing jaundice. The treatment method involves suspending breastfeeding and switching to other formula milk to reduce the concentration of antibodies in breast milk and promote the regression of jaundice. At the same time, the changes in jaundice and the recovery of liver function in the infant should be observed.
GPT-4o	一、该患儿最可能的诊断是新生儿溶血病。 二、该病的病因是ABO血型不合引起的母婴血型不合溶血。 三、该病的治疗方法包括： 1. 光疗：通过蓝光照射降低血清胆红素水平。 2. 免疫球蛋白治疗：减少溶血反应。 3. 换血疗法：在严重情况下，通过换血降低胆红素水平和溶血程度。 4. 支持治疗：包括维持水电解质平衡和营养支持。 I. The most likely diagnosis for this child is hemolytic disease of the newborn. II. The etiology of this disease is hemolysis caused by ABO blood group incompatibility between the mother and the child. III. The treatment methods for this disease include: (1)Phototherapy: Lower serum bilirubin levels through blue light irradiation. (2)Immunoglobulin therapy: Reduce hemolytic reactions. (3)Exchange transfusion: In severe cases, lower bilirubin levels and reduce the degree of hemolysis through blood exchange. (4)Supportive treatment: Includes maintaining water-electrolyte balance and nutritional support.

***注：绿色字体代表回答的关键正确部分，红色字体代表回答明显出错的部分。**
***Note: The green font represents the key correct part of response, while the red font indicates the part of response that is obviously incorrect.**

Figure 10: Answers and explanation notes of an exemplar case analysis question.