# PeMeBench: 中文儿科医疗问答基准测试方法

张芊[1,2], 陈攀峰[1,2], 冯林坤[1,2], 刘淑钰[1,2], 马丹[1,2], 陈梅[1,2], 李晖[1,2]
1. 公共大数据国家重点实验室, 贵州 贵阳 550000;
2. 贵州大学计算机科学与技术学院, 贵州 贵阳 550000

**摘要**
大语言模型在医疗领域显现出巨大的应用潜力, 如何评估其在医疗领域中的性能成为挑战。现有医疗评测基准测试多为选择题形式, 难以全面和精准地评估模型在儿科医疗场景中的性能。为此, 提出首个中文儿科医疗问答基准测试方法——PeMeBench。该方法基于双视角评估维度, 参考来自10个儿科疾病系统的诊疗规范类书籍, 将儿科医疗问答任务细分为疾病知识、治疗方案、用药剂量、疾病预防和药理作用5个儿科医疗问答子任务, 构建超1万个开放式的问答题目, 引入一种融合实体召回和检测语句幻觉的多粒度自动化评估方案, 旨在对大语言模型在儿科基础医疗领域中的性能进行全面、准确的评估, 深入剖析其潜在局限性, 为提升医疗服务的智能化水平奠定坚实的基础。

**关键词**
儿科医疗; 基准测试; 大语言模型; 问答

中图分类号: TP399 文献标志码: A doi: 10.11959/j.issn.2096-0271.2024058

# *PeMeBench: Chinese pediatric medical Q&A benchmark testing method*

**ZHANG Qian[1,2], CHEN Panfeng[1,2], FENG Linkun[1,2], LIU Shuyu[1,2], MA Dan[1,2], CHEN Mei[1,2], LI Hui[1,2]**
1. State Key Laboratory of Public Big Data, Guiyang 550000, China
2. College of Computer Science and Technology, Guizhou University, Guiyang 550000, China

***Abstract***
Large language model (LLM) has demonstrated significant application potential in the medical field. However, evaluating the performance of LLM in medical scenarios poses a challenge. Existing medical benchmarks, predominantly in the form of multiple-choice questions, struggle to comprehensively and accurately assess LLM's performance in pediatric domains. To address this issue, PeMeBench, the first Chinese pediatric question-answering benchmark, was proposed. Leveraging a dual-perspective evaluation dimensions and referencing diagnostic and treatment guidelines from 10 pediatric disease systems, PeMeBench meticulously categorized pediatric medical question-answering tasks into five subdomains: disease knowledge, treatment plans, medication dosages, disease prevention, and pharmacological effects. It comprised over 10 000 open-ended question-answering items and introduced a multi-grained automated

evaluation scheme that integrated entity retrieval with the detection of hallucinated sentences. This approach aimed to provide a comprehensive and precise assessment of LLM's performance in pediatric healthcare, delving into their potential limitations and laying a solid foundation for enhancing the intelligence level of medical services.

***Key words***
pediatric medicine, benchmark testing, large language model, Q&A

## 0 引言

大语言模型(large language model, LLM)的蓬勃兴起推动了自然语言处理范式的变革。其在文学创作、机器翻译、问答等多个任务上的优异表现,促使研究者将LLM应用到医疗领域,以期解决医疗实践中遇到的各种复杂问题。

随着医疗领域大模型的不断发展[1-3],如何客观、准确地评估模型在医疗任务上的性能具有一定的挑战性。由于医疗领域的特殊性,通用LLM基准测试方法的数据集中与医疗相关的评测数据往往非常少,因此使用通用的LLM基准测试方法无法较准确地衡量LLM解决医疗问题的能力。

一些针对医疗领域的基准测试方法通常以选择题的方式来评估LLM解决医疗领域各类问题的能力[4-6]。但这种方式往往存在一定的局限性,假设LLM无法理解某道题目的真正含义,它仍然可以随机选择一个选项作为答案,但通常只有20%～25%的正确率,人们无法判断出LLM是否真正掌握了某个知识,因此使用选择题的方式并不能精准地评估LLM在医疗领域的能力。部分开放式医疗问答数据集源于真实的医患对话[7],但这些数据集中疾病类型的数据分布呈现出高度的不均衡性,某些疾病类型的样本数量远多于其他类型。另外医患对话中的口语化表达会影响数据集的医学专业性,无法深层次评估LLM的诊疗能力。现有的医疗基准测试方法并未根据儿科医疗场景进行明确的区分,无法详尽地评估LLM对不同儿科疾病的了解和处理能力。对于开放式问答来说,大模型生成的幻觉问题往往是阻碍医疗大模型落地的原因之一。然而,这些基准测试方式并未关注大模型生成的幻觉问题。

为了解决上述问题,本文提出了一个中文儿科医疗问答基准测试方法——PeMeBench,以弥补儿科医疗领域LLM基准测试方法的不足。该基准测试方法的数据来自《儿科疾病诊疗规范》丛书,涵盖了10个儿科科室的589种疾病。PeMeBench基于此构建了专业且全面的开放式问答数据集,从科室导向和任务驱动两个维度进行评估,引入融合实体召回和检测语句幻觉的多粒度自动化评估方案,进一步精细地评估医疗LLM在各维度的诊疗分析能力,为模型后续的优化指明方向。

本文的主要贡献如下。

- 基于儿科疾病诊疗规范书籍构建儿科医疗问答数据集,从科室导向和任务驱动两个维度全面评估医疗LLM各方面的能力。
- 引入实体召回率和语句幻觉率两个新指标,同时结合传统的语义相似度指标,设计了一种新的自动化评估策略。
- 在PeMeBench上对多个LLM进行评测,结果表明,现有LLM在儿科医疗问答领域的能力有很大的提升空间,PeMeBench为评估LLM对医疗基础知识的理解和应用能力提供了新的测试方法。

# 1 相关工作

## 1.1 医疗LLM

为了更好地解决医疗领域的问题，当前众多研究通过微调改进LLM，以应对医疗领域的特定挑战。很多研究倾向于在规模较小的基座LLM（参数数量为6 GB、7 GB或13 GB）上进行微调。Doctor-GLM[1]、ChatGLM-Med、BianQue[3]基于ChatGLM-6B进行微调。ChatDoctor[8]、ZhongJing[2]基于LLaMA-7B进行微调，PMC_LLaMA（13 GB）通过整合大量医学学术论文和教科书的内容构建医学知识数据集，并在LLaMA模型上进行微调[9]。基于Baichuan-7B的孙思邈医学LLM致力于提供安全、可靠、普惠的中文医疗LLM。基于BLOOMZ-7B的明医（MING）能够执行医疗问答和智能问诊任务。QiZhenGPT、BenTsao[10]是基于多个基座LLM微调得到的医疗大模型集。这些医疗LLM通常基于LoRA[11]这一高效参数微调方法（parameter-efficient fine-tuning，PEFT）进行微调，能够在资源消耗较少的情况下使LLM学习到更多的领域知识。

## 1.2 LLM的通用基准测试方法

近年来，随着自然语言理解领域的迅速发展，一系列通用的中文问答基准测试方法相继出现。为了评估模型的医疗能力，早期的模型通常将这类通用基准测试方法作为评估工具，以考察LLM在医疗相关任务上的性能。CMMLU（Chinese multi-modal learning for understanding）是一个包括社会科学、人文科学等方面的综合性中文基准测试方法，用于评估CMMLU在中文语言和文化背景下的高级知识推理能力[12]。C-EVAL涵盖了4个难度水平的多项选择题，旨在分析基础模型的重要优缺点，从而促进LLM的发展[13]。这些通用基准测试方法覆盖的学科范围较广，但在医疗方面的内容却相对缺乏，因此这些通用基准测试方法并不能全面地评估模型在医疗任务上的表现。

## 1.3 医疗LLM的基准测试方法

为了应对这一挑战，研究人员提出了一些医疗领域的LLM基准测试方法。MedMCQA是一个英文医疗基准测试方法，其收集美国医学考试的题目来评估模型的医疗能力[14]。为了填补中文医疗基准测试方法的空白，MedBench整理了大量中文医学考试的题目和真实临床案例来构建医疗评测基准测试方法[15]。CBLUE进一步将医疗任务细分为命名实体识别、信息提取、临床诊断标准化、短文分类、问题回答、意图分类、查询词-页面标题相关性、查询词-查询词相关性8个子任务，从而实现对LLM在医疗领域中各方面能力的评估[16]。然而，对于实际临床诊疗中最重要的疾病知识、用药剂量、药理作用、治疗方案和疾病预防等能力，以上基准测试方法均未进行更细致的区分。考虑到这一问题，PeMeBench采用任务驱动的评估策略评估LLM在不同医疗子任务上的能力。

# 2 PeMeBench

为了精准地评估通用LLM在儿科医疗任务上的问答能力，笔者基于权威的儿科疾病诊疗规范，构建了PeMeBench，该方法主要包括数据集构造方法、根据数据集构造方法获得的数据集、评估维度和评估方法4个部分。

### 2.1 数据集构造方法

本文设计了一套精细化的数据处理流程和问答对生成方案，旨在获得高质量的儿科医疗问答基准测试方法数据集。

首先，参考儿科诊疗规范专业书籍，按照儿科疾病种类、诊疗技术手段、常用量表类型、常用药物等分类，对相应的文本内容和关键要素进行提取。其次，构造专业的提示模板，引导LLM基于儿科医疗文本内容生成一系列与儿科医疗紧密相关的问题和答案，提示模板中要求LLM生成的内容必须源于给定文本。为了防止LLM生成的问题过于单一，PeMeBench的数据集构造方法分为两类：零样本提示方法仅调用一次LLM，使其同时生成问题和答案；少样本提示方法则构造出多样化的提问形式。利用LLM构造问答对使用的零样本、少样本提示模板如**图1**、**图2**所示。

少样本提示方法的具体步骤如下。

（1）构造种子数据

经过初步的数据处理之后，从每个科室中挑选出一份包含疾病概述、疾病诊断、疾病治疗与疾病预防等内容的层次结构清晰的文本。手动构造相应的问题，依靠LLM接口得到该问题的回答并进行核验。这些提问将作为自动化构建数据集的种子数据。

（2）设计提问逻辑

收集到种子数据后，利用少样本提示方法，随机从种子数据中挑出3个问题作为样例让LLM构造问答对。规定：对于疾病概述和病因的内容，使用考察定义类提示模板；对于疾病诊断、疾病治疗和疾病预防的内容，使用询问解决方案类和询问原因类提示模板。

（3）生成问题回答

获取到问题之后，将问题与相关的儿科医疗文本再次送入LLM，从而获取问题相应的回答。

结合两种提示模板得到的问答对对知识的覆盖更加全面。最后对用于LLM能力评测的问题和答案进行严格的人工审查，删除低质量问答对。主要的筛选原则如下。

- 删掉与儿科不相关的问题。
- 剔除指代不明确的问题。
- 核对问题对应的回答是否与文本内容相符。

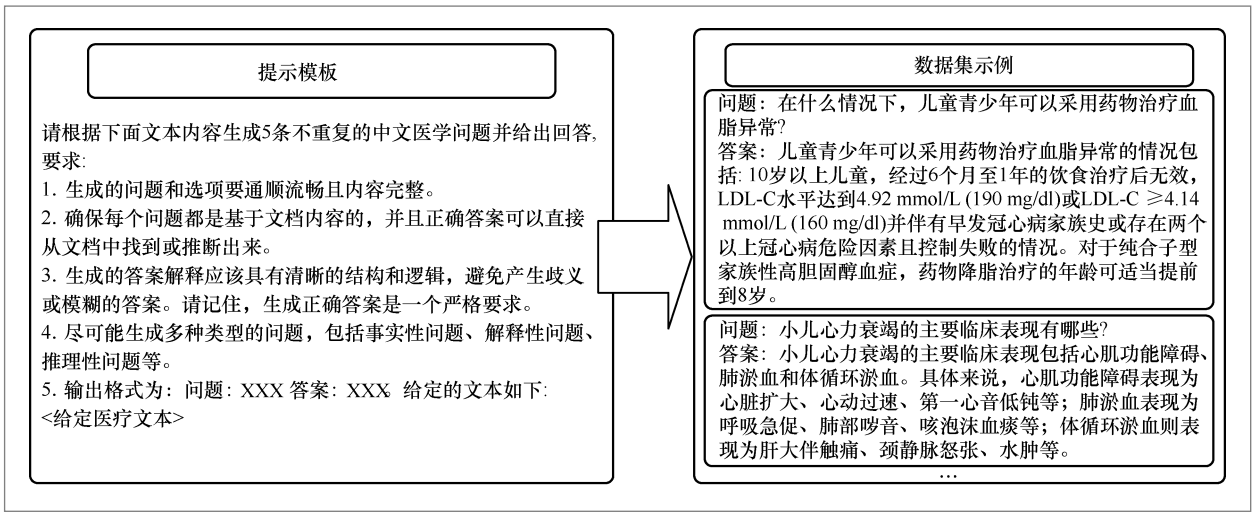



图1 利用 LLM 构造问答对使用的零样本提示模板

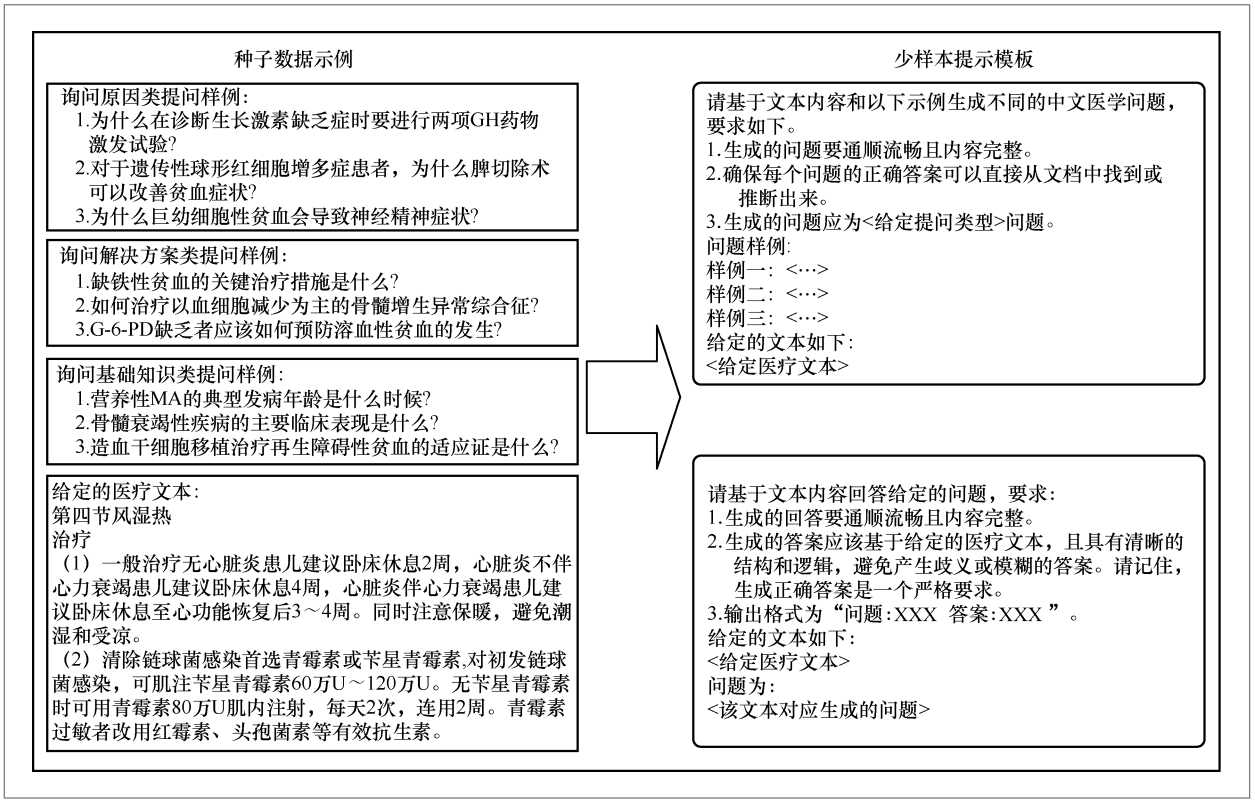



图 2 利用 LLM 构造问答对使用的少样本提示模板

## 2.2 数据集

利用GLM-4大模型的API，依靠上述数据集构造方法得到13 113条问答对，经过严格的人工审查之后保留了12 068条数据，从而形成了PeMeBench数据集。

### 2.2.1 数据分布

本基准测试方法的数据集涵盖了儿科肾脏系统疾病、儿科急诊与危重症、儿科感染性疾病、儿科血液系统疾病、儿科心血管系统疾病、儿科免疫系统疾病、儿科呼吸系统疾病、儿科内分泌与代谢性疾病、儿童保健与发育行为、新生儿疾病共10类儿科疾病以及这些疾病涉及的儿科诊疗技术和临床常用量表的内容。该数据集包含589种儿科疾病、236种儿科诊疗技术、56种儿科常见药物和70种儿科常用量表。**图3**详细展示了PeMeBench数据集中不同细分类别的问答对数量及占比。

PeMeBench数据集在保证数据集质量的同时囊括了大部分的儿科常见疾病与罕见病知识，**表1**为PeMeBench数据集的统计信息，其中Q表示数据集中的问题，A表示对应的问答。各科室的样本长度基本一致，回答可以是精练的一句话，也可以是复杂的解释和分析。

PeMeBench与其他医疗基准测试方法的数据集的对比见**表2**。由**表2**可以看出，选择题类型的数据集中儿科相关的题目较少，Huatuo-26M数据集虽然有足够数量的儿科相关问答对，但其数据来源于网络在线医患对话，很少涉及儿科疑难杂症。而PeMeBench中的数据集聚焦于儿科

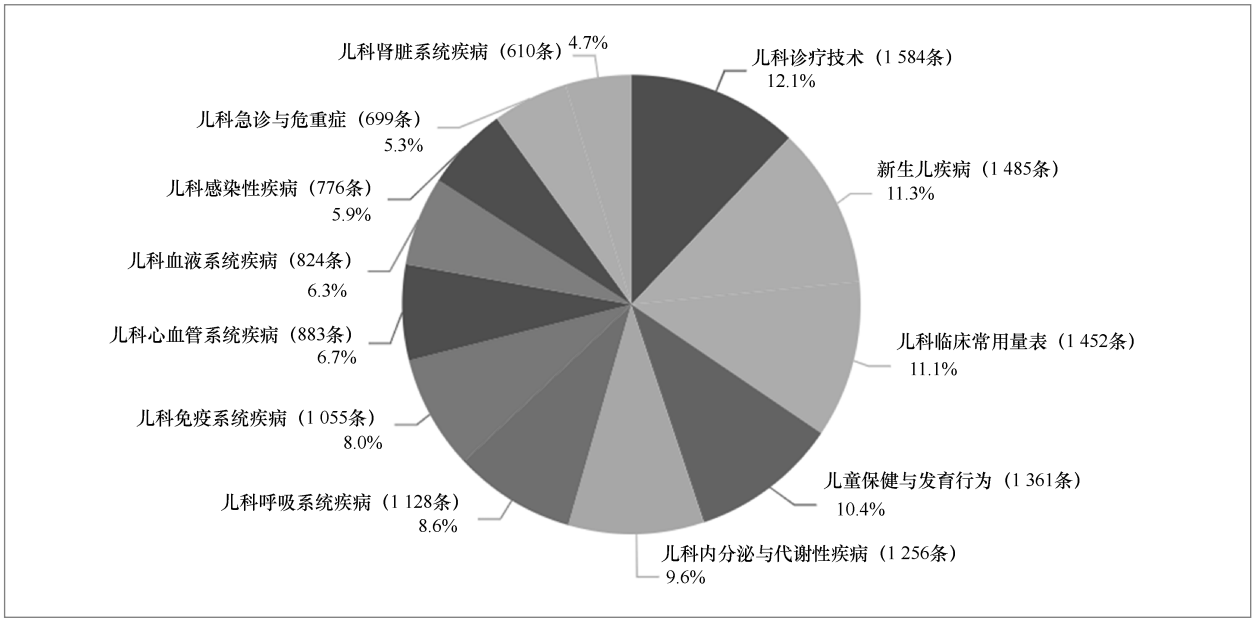



图 3 PeMeBench 数据集中不同细分类别的问答对数量及占比

表 1 PeMeBench 数据集的统计信息

| 类别 | 最小样本长度/字符 | | 最大样本长度/字符 | | 平均样本长度/字符 | |
|---|---|---|---|---|---|---|
| | Q | A | Q | A | Q | A |
| 新生儿疾病 | 6 | 14 | 52 | 294 | 21.7 | 92.7 |
| 儿科肾脏系统疾病 | 9 | 19 | 46 | 219 | 22.9 | 92.5 |
| 儿科急诊与危重症疾病 | 6 | 17 | 61 | 323 | 21.1 | 135.6 |
| 儿童保健与发育行为疾病 | 6 | 19 | 46 | 350 | 20.6 | 131.1 |
| 儿科免疫系统疾病 | 7 | 19 | 67 | 307 | 22.2 | 125.6 |
| 儿科感染性疾病 | 7 | 11 | 48 | 359 | 20.4 | 118.8 |
| 儿科呼吸系统疾病 | 7 | 13 | 60 | 338 | 22.4 | 125.9 |
| 儿科心血管系统疾病 | 8 | 15 | 49 | 259 | 21.8 | 95.3 |
| 儿科内分泌与代谢性疾病 | 8 | 21 | 72 | 418 | 22.7 | 130.4 |
| 儿科血液系统疾病 | 6 | 19 | 59 | 393 | 23.5 | 128.3 |
| 儿科诊疗技术 | 7 | 13 | 53 | 366 | 22.6 | 87.5 |
| 儿科临床常用量表 | 6 | 15 | 90 | 276 | 24.3 | 117.5 |
| PeMeBench数据集（不分类别） | 6 | 11 | 90 | 418 | 22.0 | 112.6 |

表 2 PeMeBench 与其他医疗基准测试方法数据集的对比

| 数据集 | 语言 | 类型 | 是否划分科室 | 数据量（儿科题目数量）/条 | 数据来源 |
|---|---|---|---|---|---|
| CMExam | 简体中文 | 多选项选择题 | 否 | 60 000（2 936） | MCMLE |
| MedQA[17] | 简体中文、繁体中文、英文 | 多选项选择题 | 是 | 61 097（3 844） | MCMLE |
| MedMCQA | 英文 | 多选项选择题 | 是 | >194 000（190） | AIIMS & NEET PG 入学考试题目 |
| Huatuo-26M | 简体中文 | 开放式问答 | 否 | >26 000 000+（>30 000） | 在线医疗网站 |
| PeMeBench | 简体中文 | 开放式问答 | 是 | 12 068（12 068） | 《儿科疾病诊疗规范》丛书 |

医疗问答领域，在规模与品质上均展现出卓越性，为评估LLM在儿科医疗相关问答任务中的效能提供了坚实的保障。

#### 2.2.2 数据特点

（1）语言专业化

儿科医疗领域存在大量生活中不常见的医疗术语和专业表述，日常的口语化表达很难准确无误地传达医疗信息。一些医疗问答基准测试方法侧重于通过医患沟通中较口语化的问题和答案来评估LLM的性能，容易忽视医疗服务沟通表达过程中的专业性。这类基准测试方法的数据大多源于在线问诊平台，由于医患之间口语化表达存在不精准的因素，以及提供在线问诊服务的医生的真实医疗水平存在不确定性，基准测试方法中医疗问答内容的专业度和可信度可能会受到一定程度的影响。构造PeMeBench的数据集时，笔者严格参照书中表述及标准，这种做法很大程度上保证了数据集的准确性和专业性。

（2）问题多样化

PeMeBench的数据集包含一系列多样化的提问方式，这些提问方式既考虑了儿科疾病的多样性，还兼顾了语言理解的复杂性，使数据集的表述能够符合儿科领域的语言风格。其中，询问原因类问题用于探究LLM对儿科疾病发生机制的理解程度，询问解决方案类问题旨在评估LLM治疗儿科疾病的水平，考察定义类问题则是为了检验LLM对儿科基础知识和专业术语的掌握程度。这些提问方式能够全面考察LLM在儿科医疗领域的综合性能。3种类型的问答示例如**图4**所示。

### 2.3 评估维度

PeMeBench旨在全面评估不同维度下LLM对儿科医疗知识的掌握能力，因此笔者提出基于科室导向和任务驱动的双视角评估维度，从专业细分与实践需求两个层面，对LLM的医疗问答能力进行全方位、深层次的评估。科室导向的评估，能够精准捕捉模型在不同儿科领域（如新生儿科、小儿呼吸科等）的专业知识掌握情况。任务驱动的方法则进一步模拟真实医疗场景中的具体任务（如治疗方案、用药剂量等），评估模型在解决实际问题时的表现与效能。

#### 2.3.1 基于科室导向的评估维度

医学的目的是预防和治疗疾病，对于儿科医疗领域而言，每一类儿科疾病都

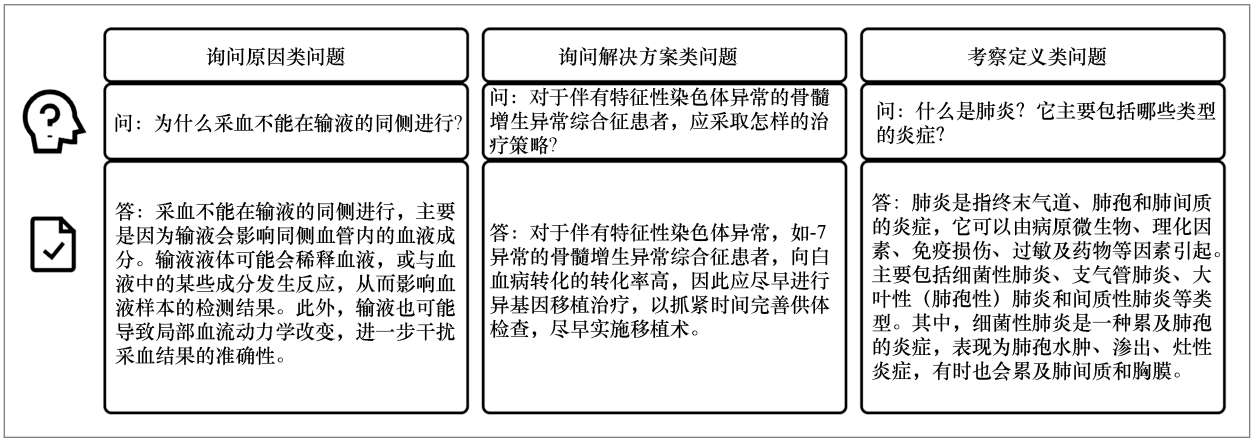



图 4 3 种类型的问答示例

有相应的疾病分析、诊断与治疗流程，将各种儿科疾病归类至各个科室，能够显著提高疾病管理与治愈的效率。受此启发，基于科室导向的评估维度的核心目标是深入挖掘并评估LLM在儿科各子领域知识掌握方面的深度与广度，不仅仅是衡量简单的儿科医疗问诊能力。对不同科室疾病进行详细考察，能够精准识别LLM在特定疾病领域内的专业知识短板，为儿科医疗大模型指明了后续的优化方向。

具体将儿科疾病归为10个科室类别，每个类别的数据包含该科室下某种疾病的相关问题。在对LLM进行评估时，为LLM精心设定角色（如告诉模型它是一个××领域的专家），引导LLM基于某科室领域专家的视角回答问题。以下是一个聚焦于儿科肾脏系统疾病的评估样例。

提示词：假设你是一位能力出众的儿科医生，你擅长诊断与治疗各种肾脏系统疾病，请你基于你的专业能力回答下面关于肾脏系统疾病的问题。

问题：如何治疗肾性贫血。

#### 2.3.2 基于任务驱动的评估维度

现有医疗基准测试方法往往只从宏观角度划分医疗场景，如医疗知识问答、复杂医疗推理等，这极大地限制了人们从更加精细的维度评估LLM应对不同医疗任务的能力。因此，笔者通过分析儿科医疗领域的特点，创新性地对任务进行细分，将儿科医疗问答任务细化为疾病知识、治疗方案、用药剂量、疾病预防和药理作用5类医疗任务。其中，疾病知识类任务旨在评估LLM对基础疾病知识和概念的了解，治疗方案类任务用于评估LLM在不同问诊场景下是否能给出合理的治疗方案，疾病预防类任务用于评估LLM是否了解疾病风险的防范。由于诊疗过程中的用药剂量需要考虑病情、治疗对象等方面因素综合确定，因此特别设计了用药剂量类任务和药理作用类任务，用于精准评估医疗LLM作为一名“医生”的专业度和可靠度。

通过各个细分任务可以更精准地定位模型在诊治过程中的哪些环节表现出色，以及在哪些环节存在不足，从而为后续针对性地优化模型结构和提升模型性能提供精细化的数据洞察。

### 2.4 评估方法

现有的基准测试方法通常使用自然语言处理中的BLEU（bilingual evaluation understudy）和ROUGE（recall-oriented understudy for gisting evaluation）来评估大语言模型的回答与真实回答之间的相似度。ROUGE和BLEU是生成式问答中被广泛使用的两类指标，通过比较生成文本与参考文本之间n-gram的重叠度来捕捉文本之间的相似度，从而支持不同粒度下的文本评估。然而，这些基于文本相似性的指标忽略了一个重要的问题，即LLM生成的幻觉问题[18]。在某些情况下，模型的回答与真实回答之间有较高的语义相似性，但是模型本身的能力有限，导致它的回答不准确或者不可靠。因此，笔者引入了一种自动化的多粒度评估策略，将实体级评估策略与句子级评估策略相结合，能够更加全面地评估大模型在儿科医疗领域的能力。笔者加入了两类指标，分别是实体召回率和语句幻觉率。整个流程使用GLM4大模型作为一个中立的评估者，通过调用API来获得评估结果。

对于实体召回率，基于提示词获取答案中的关键实体，然后通过代码计算出模型回答的实体召回率。对于语句幻觉率，直接让GLM4判断对应的模型回答和真实

答案之间是否存在对立的内容，以此判定该回答是否存在幻觉。在获取到GLM4给出的判定后，人工核实修改了实体抽取的结果，以确保评估的可靠性。利用GLM4大模型提取回答的实体召回率并判断是否为幻觉答案的提示词，如**图5**所示。

（1）BLEU

BLEU是一种评估自动文摘或机器翻译质量的指标，用于衡量模型回答的准确性以及句子的流畅性。BLEU的计算式如下。

$$\mathrm{BLEU}-n=\frac{\sum_{c\in\{\mathrm{Candidates}\}}\sum_{n-\mathrm{gram}\in c}\mathrm{Count}_{\mathrm{clip}}(n-\mathrm{gram})}{\sum_{c'\in\{\mathrm{Candidates}\}}\sum_{n-\mathrm{gram}'\in c'}\mathrm{Count}_{\mathrm{clip}}(n-\mathrm{gram}')} \quad (1)$$

其中，Candidates表示模型生成答案，$c$表示给定的Candidates中出现在reference（标准回答）中的$n$-gramg词语的个数，$c'$表示给定的Candidates中所有的$n$-gramg词语的个数，$\mathrm{Count}_{\mathrm{clip}}$表示在reference中某一个$n$-gramg词语的个数，分子表示标准答案中Candidates的$n$-gram个数，分母为Candidates中的所有$n$-gram个数[19]。

（2）ROUGE

ROUGE同样是用于评估自动文摘或机器翻译质量的评估指标，通常用来衡量LLM在自然语言生成任务上的性能。ROUGE-1、ROUGE-2、ROUGE-$L$分别基于1-gram、2-gram和最长公共子序列$L$计算真实值与预测值之间的相似度。由于生成式问答的开放性，笔者着重关注ROUGE指标中的召回率（用$R$表示），并将其简写为$R$-1、$R$-2、$R$-$L$[20]。

（3）实体召回率（$\mathrm{Recall}_{\mathrm{entity}}$）

实体召回率指的是模型回答中包括了多少个真实答案中的实体。具体来说，首先利用GLM4抽取出真实回答中的关键实体，然后统计模型的回答中出现了多少个关键实体，从而计算出句子中实体的召回率。

（4）语句幻觉率（$\mathrm{Hall}_{\mathrm{sentence}}$）

笔者将评测任务看作一个分类任务。当模型回答与真实回答的内容表述一致，则认为回答是准确的；当表述不一致时，则认为该回答是存在幻觉的。它的计算式如下。

$$\mathrm{Hall}_{\mathrm{sentence}}=\frac{|P_{\mathrm{hallucination}}|}{|S_{\mathrm{total}}|}\times 100\% \quad (2)$$

其中，$S_{\mathrm{total}}$表示所有的样本，$P_{\mathrm{hallucination}}$表示GLM4判定为存在幻觉的样本。$100-\mathrm{Hall}_{\mathrm{sentence}}$表示回答未出现幻觉部分的分值。

最后，计算4个语义相似度指标、$\mathrm{Recall}_{\mathrm{entity}}$、$100-\mathrm{Hall}_{\mathrm{sentence}}$的平均值作为模型回答的最终分数。

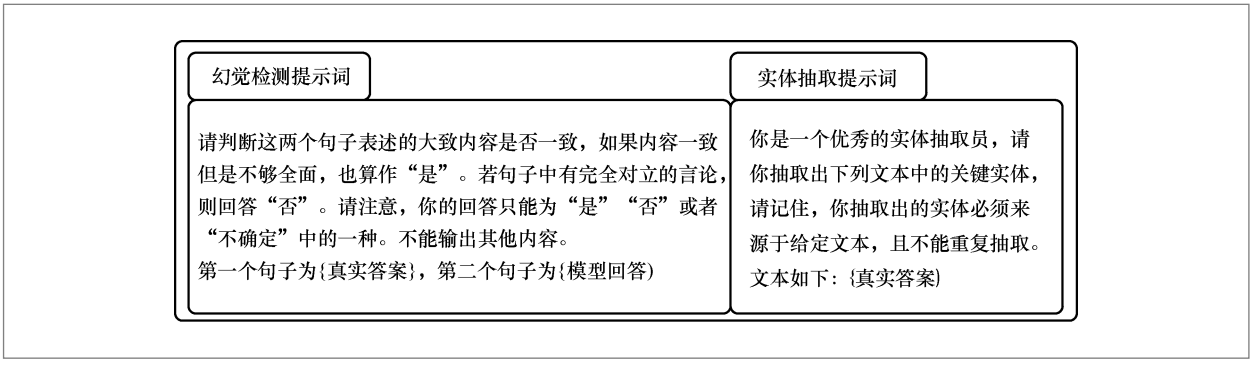



图 5 利用 GLM4 模型提取实体并判断是否为幻觉答案的提示词

$$S_{\text{total}} = \left( \frac{\text{BLEU} + R\text{-}1 + R\text{-}2 + R\text{-}L}{4} + \text{Recall}_{\text{entity}} + (100 - \text{Hall}_{\text{sentence}}) \right) / 3 \quad (3)$$

# 3 PeMeBench效能评估

在多个开源LLM和闭源商用LLM上测试PeMeBench，以评估现有LLM在儿科医疗任务上的能力。

## 3.1 实验设置

针对不同维度的实验，采用不同的提示词引导待测模型进行回答。对于基于科室导向的评估维度，使用角色扮演的提示词，引导待测模型作为一位专业的儿科医生进行回答。对于基于任务驱动的评估维度，则为模型设定不同的医疗场景提示词，引导模型基于该场景给出更合理的回答。

### 3.1.1 模型选择

选取了一系列公开的具有代表性的中文通用LLM和医疗LLM进行评测，具体的LLM如下。

- 开源通用LLM：Baichuan2-7B-Chat、Baichuan2-13B-Chat[21]、Qwen-7B-Chat、Qwen-14B-Chat[22]、InternLM2-7B-Chat、InternLM2-20B-Chat。
- 开源医疗LLM：BianQue2（7 GB）[3]。
- 闭源LLM：文心一言ERNIE-Bot4.0[23]、星火认知Spark3.5 Max、ChatGPT-3.5-turbo。

对于开源通用LLM和开源医疗LLM，利用本地一台包含8张NVIDIA A6000 GPU的服务器进行测试，对于闭源LLM，调用API进行访问。

### 3.1.2 数据集选择

选取CMexam、MedQA以及Huatuo-26M医疗问答数据集和PeMeBench进行比较。其中CMexam和MedQA数据形式为选择题，Huatuo-26M以问答题的形式呈现。利用正则表达式从CMexam、MedQA以及Huatuo-26M的测试集中挑选出关于儿科医疗的问题，以此构建一个专门的评估子集，从而确保评估结果的准确性和公平性。

### 3.1.3 评估指标选择

（1）准确率（accuracy）

选择题存在唯一确定的正确答案，因此笔者采用准确率（用ACC表示）作为选择题的评估指标，其计算方法为ACC=模型回答正确的题目数量/所有题目数量。

（2）语义相似度指标BLEU和ROUGE

BLEU和ROUGE是两类常用的基于语义相似度的生成式问答评估指标，但是仅靠这两类指标无法衡量模型是否存在幻觉。

（3）实体召回率 $\text{Recall}_{\text{entity}}$ 和幻觉率 $\text{Hall}_{\text{sentence}}$

本文提出了两种新的指标，分别从实体和语句层面来衡量模型回答的性能。第2.4节详细介绍了这些评估指标。

## 3.2 结果及分析

### 3.2.1 引入语句幻觉率指标的必要性

各模型在PeMeBench上的结果见**表3**。PeMeBench在数据形式的丰富性、知识覆盖的深度以及实际应用的有效性方面，均展现出显著的优势。

表 3 各模型在 PeMeBench 上的结果

| 模型名称 | PeMeBench | | | | | | |
|---|---|---|---|---|---|---|---|
| | BLEU | *R*-1 | *R*-2 | *R-L* | $\text{Hall}_{\text{sentence}}$ | $\text{Recall}_{\text{entity}}$ | $S_{\text{total}}$ |
| Baichuan2-7B-Chat | 8.17% | 20.58% | 5.94% | 12.07% | 22.08% | 32.04% | 40.55% |
| Baichuan2-13B-Chat | 7.75% | 19.61% | 5.50% | 11.23% | 20.83% | 33.05% | 41.08% |
| Qwen-7B-Chat | 12.83% | 22.74% | 6.38% | 16.27% | 36.25% | 25.01% | 34.44% |
| Qwen-14B-Chat | 14.23% | 25.73% | 7.67% | 17.85% | 31.25% | 25.52% | 36.88% |
| InternLM2-7B-Chat | 4.95% | 15.65% | 3.81% | 7.29% | 23.33% | 35.22% | 39.94% |
| InternLM2-20B-Chat | 5.13% | 15.04% | 3.75% | 7.49% | 22.91% | 36.68% | 40.54% |
| BianQue2(7 GB) | 11.08% | 27.60% | 8.07% | 17.04% | 41.66% | 23.67% | 32.65% |
| Spark3.5 Max | 16.77% | 20.94% | 6.63% | 12.55% | 17.08% | 34.75% | 43.96% |
| ERNIE-Bot4.0 | 16.67% | 19.95% | 6.56% | 12.31% | 15% | 36.92% | 45.26% |
| ChatGPT-3.5-turbo | 25.06% | 29.17% | 9.87% | 18.60% | 28.33% | 30.39% | 40.91% |

由**表3**可知，规模较大的通用模型在回答儿科问题时展现出较高的准确性，这一发现与先前的研究结论相契合[24]。整体上来看，越高的实体召回率对应的语句幻觉率相对越低。在诸多测试模型中，ChatGPT-3.5-turbo在语义相似度相关的指标BLEU、*R*-1、*R*-2和*R-L*上均高于其他模型，但是其表现出来的幻觉问题非常严重，因此ChatGPT-3.5的最终得分并不突出。而闭源模型Spark3.5 Max、ERNIE-Bot4.0虽然在语义相似度指标上的分数不太理想，但是它们有较高的实体召回率和较低的语句幻觉率，因此它们的总体得分位于前二。

BianQue2(7 GB)模型的综合得分是最后一名，笔者发现该模型的幻觉问题非常严重。尽管通过医疗领域知识微调后的LLM能够学习到更多医学方面的表达，但是带来了更多的幻觉问题。模型容易生成很多看似合理但实际上并不正确的回答。由此可见引入语句幻觉率指标是有效的。

### 3.2.2 模型在其他数据集上的性能

各模型在CMexam、MedQA以及Huatuo-26M上的得分见**表4**。由于选择题和开放式问答的评分标准不同，因此CMexam、MedQA与Huatuo-26M的得分相差较大。

通过**表4**的数据可知，各模型在选择题类型的基准测试方法上普遍表现得较理想，多数得分已接近及格线(60分)。然而医疗模型BianQue却不能很好地遵循指令，总是生成与答案选项无关的内容，因此在CMexam和MedQA上的表现很差。

尽管ChatGPT-3.5-turbo在CMexam、MedQA上有较好的指令遵循能力，但它依然反馈给用户一些看似合理但并不正确的答案选项。Spark3.5 Max并不能很好地理解和遵循指令，在测试其在MedQA上的性能时，尽管利用提示模板规定回复格式，Spark3.5 Max依旧会生成许多与要求不符的内容。使用简单的正则表达式计算Spark3.5 Max原始回答的得分，仅有49.19。为了挖掘Spark3.5 Max在MedQA上的真实水平，人工对它生成的回答进行了筛选与核实，发现得分达到了77.42。这一结果也和上述结论相印证，Spark3.5 Max的指令遵循能力较差，但是依然拥有

表 4 各模型在 CMexam、MedQA 以及 Huatuo-26M 上的得分

| 模型名称 | Huatuo-26M | | | | CMexam | MedQA |
|---|---|---|---|---|---|---|
| | BLEU | *R*-1 | *R*-2 | *R*-*L* | ACC | ACC |
| Baichuan2-7B-Chat | 8.53 | 14.46 | 1.83 | 10.74 | 43.77 | 38.51 |
| Baichuan2-13B-Chat | 8.53 | 14.14 | 1.76 | 10.43 | 53.2 | 47.99 |
| Qwen-7B-Chat | 8.94 | 14.62 | 1.95 | 10.98 | 55.45 | 42.91 |
| Qwen-14B-Chat | 8.82 | 14.60 | 2.09 | 10.92 | 63.4 | 55.24 |
| InternLM2-7B-Chat | 5.84 | 11.87 | 1.39 | 7.56 | 41.89 | 40.12 |
| InternLM2-20B-Chat | 5.08 | 10.29 | 1.18 | 6.42 | 21.89 | 21.37 |
| BianQue2 (7 GB) | 12.01 | 19.86 | 2.96 | 15.59 | 未能准确给出回答，记为0分 | |
| ERNIE-Bot4.0 | 11.17 | 12.01 | 2.10 | 7.55 | 77.36 | 80.65 |
| Spark3.5 Max | 10.35 | 9.85 | 0.78 | 7.67 | 75.47 | 77.42 (49.19) |
| ChatGPT-3.5-turbo | 18.28 | 17.84 | 2.47 | 13.89 | 38.49 | 41.53 |

大量的医疗知识储备。这种差距体现了利用选择题评估LLM性能的局限性。利用选择题考察LLM能力往往具有随机性，并不能有力地说明LLM在儿科医疗方面的能力。

在开放性医疗问答中，LLM的表现不尽如人意，所有模型在Huatuo-26M数据集上的得分均低于PeMeBench，且Baichuan系列和Qwen系列LLM的得分不具有区分性。Huatuo-26M的ovovr数据过于口语化，模型无法精准捕捉提问中所需要的儿科医疗知识，从而影响了回答的准确性。因此在构造数据集时除了需要确保答案的准确性，还应该关注问题的有效性和精准性。

### 3.2.3 PeMeBench在基于科室导向的评估维度中不同模型的性能

分别抽取儿科各科室的问题对LLM进行测试，将总体得分绘制成雷达图，以分析LLM对不同儿科科室知识的掌握能力，模型对不同儿科类别知识的掌握情况如**图6**所示。模型在对各类儿科医疗知识的掌握程度上表现出了显著的差异。具体来说，大部分模型在儿科内分泌与代谢性疾病上的表现较好，而在儿科肾脏系统疾病上的表现普遍较差。

不同模型在不同领域的知识掌握程度存在差异，细分疾病类别能更准确地评估模型的能力，并为模型训练提供更有针对性的指导。在模型的后续训练过程中，需要充分考虑各类儿科医疗知识的特点和分布，确保模型能够全面而深入地掌握相关知识。

### 3.2.4 PeMeBench 任务驱动评估维度中不同模型的性能

笔者从疾病知识、治疗方案、用药剂量、疾病预防和药理作用5个医疗任务中随机抽取了部分问题对多个LLM进行测试，根据式(3)计算$S_{\text{total}}$，以分析LLM在各类儿科医疗场景中的性能，测试结果见**表5**。从**表5**可以看出，LLM对不同细分任务的掌握程度存在差异，无论是开源LLM还是闭源LLM，这些模型更擅长回答用药剂量相关的知识，在这类问题上得到了较高的分数，

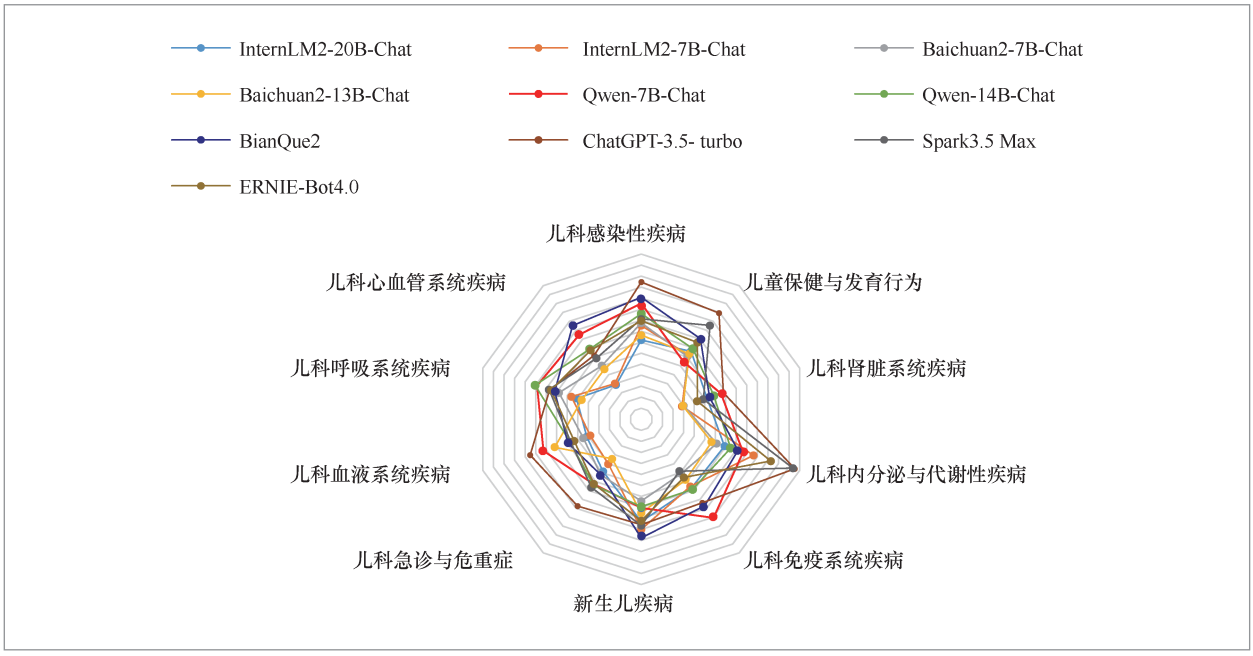



图 6 模型对不同儿科类别知识的掌握情况

表 5 不同 LLM 在 PeMeBench 的 5 个医疗子任务上的测试结果

| 模型 | BaiChuan2-7B-Chat | Qwen-7B-Chat | BianQue2 (7B) | Spark3.5 Max | ERNIE-Bot4.0 | ChatGPT-3.5-turbo |
|---|---|---|---|---|---|---|
| 治疗方案 | 21.84 | 24.37 | 25.82 | 22.02 | 22.23 | 21.49 |
| 用药剂量 | 32.93 | 28.55 | 25.06 | 33.25 | 29.48 | 33.84 |
| 疾病预防 | 19.91 | 21.87 | 23.29 | 23.62 | 19.44 | 19.55 |
| 药理作用 | 20.83 | 19.59 | 20.56 | 23.82 | 21.37 | 26.26 |
| 基础知识 | 19.84 | 24.48 | 22.05 | 21.42 | 19.32 | 21.75 |

尤其是ChatGPT-3.5-turbo在用药剂量子任务上的得分远超过其他子任务。而所有模型在治疗方案和疾病预防任务上的表现较均衡，这可能是因为同种疾病存在多种治疗和预防对策等。

### 3.2.5 案例分析

选取一个具有代表性的问题，收集几个模型的回答并给出针对性的分析，不同模型生成的答案对比如**图7**所示。其中划线部分文字说明模型生成的内容中存在明显错误的信息，加粗部分文字为与正确答案相关的内容。显然，这些模型的回答在内容准确性和完整性方面存在显著差异。每个模型均存在一定的幻觉问题，容易生成一些与正确答案不符的内容，其中InternLM2系列模型的幻觉问题尤其严重，相对而言，Baichuan2系列模型的幻觉问题较轻。尽管这些模型生成的内容都较丰富，但它们无法涵盖参考答案提及的症状等信息。这也是这些模型在PeMeBench上的分数偏低的重要原因之一。


注：加粗字体为重点正确部分，划线字体为明显错误部分

图 7 不同模型生成的答案对比

## 4 结束语

本文提出了中文儿科医疗问答基准测试方法——PeMeBench。本文基于科室导向和任务驱动双视角的评估维度，设计了一种引入融合实体召回和检测语句幻觉的多粒度自动化评估方案，并在多个LLM上进行评测。笔者发现，现有的LLM能够回答儿科医疗方面的常见基本问题，但回答的准确性仍然有待提高，且很多LLM存在一定程度的幻觉，尤其是被微调的医疗模型。然而，使用大语言模型完成自动化评估的质量取决于大语言模型自身的能力与知识储备，笔者认为，未来仍需进一步构建完善的医疗LLM性能评价体系，可通过引入更多元的评价指标、构建更全面的评测数据集、研究更先进的评估方法等工作，推动LLM在儿科医疗问答领域的发展，提升医疗服务的质量和效率，从而更好地为儿童的健康成长保驾护航。

## 参考文献:


[1] XIONG H L, WANG S, ZHU Y T, et al. DoctorGLM: fine-tuning your Chinese doctor is not a Herculean task[EB]. arXiv preprint, 2023, arXiv: 2304.01097.

[2] YANG S H, ZHAO H J, ZHU S B, et al. Zhongjing: enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue[EB]. arXiv preprint, 2023, arXiv: 2308.03549.

[3] CHEN Y R, WANG Z Y, XING X F, et al. BianQue: balancing the questioning and suggestion ability of health LLMs with multi-turn health conversations polished by ChatGPT[EB]. arXiv preprint, 2023, arXiv: 2310.15896.

[4] LIU J, ZHOU P, HUA Y, et al. Benchmarking large language models on CMExam: a comprehensive chinese medical exam dataset[C]// Proceedings of the 37th International onference on Neural Information Processings System. New York: Curran Associates Inc, 2024: 52430-52452.

[5] CAI Y, WANG L L, WANG Y, et al. MedBench: a large-scale chinese benchmark for evaluating medical large language models[EB]. arXiv preprint, 2023, arXiv: 2312.12806.

[6] ZHANG N Y, CHEN M S, BI Z, et al. CBLUE: a Chinese biomedical language understanding evaluation benchmark[EB]. arXiv preprint, 2021, arXiv: 2106.08087.

[7] LI J Q, WANG X D, WU X B, et al. Huatuo-26M, a large-scale Chinese medical QA dataset[EB]. arXiv preprint, 2023, arXiv: 2305.01526.

[8] LI Y X, LI Z H, ZHANG K, et al. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI(LLaMA) using medical domain knowledge[EB]. arXiv preprint, 2023, arXiv: 2303.14070.

[9] WU C, ZHANG X, ZHANG Y, et al. PMCLLAMA: further finetuning LLAMA on medical papers[J]. arXiv preprint, 2023, arXiv: 2304.14454.

[10] WANG H C, LIU C, XI N W, et al. HuaTuo: tuning LLaMA model with chinese medical knowledge[EB]. arXiv preprint, 2023, arXiv: 2304.06975.

[11] HU E J, SHEN Y L, WALLIS P, et al. LoRA: low-rank adaptation of large language models[EB]. arXiv preprint, 2021, arXiv: 2106.09685.

[12] LI H N, ZHANG Y X, KOTO F, et al. CMMLU: measuring massive multitask language understanding in Chinese[EB]. arXiv preprint, 2023, arXiv: 2306.09212.

[13] HUANG Y Z, BAI Y Z, ZHU Z H, et al. C-eval: a multi-level multi-discipline Chinese evaluation suite for foundation models[EB]. arxiv preprint, 2023, arxiv: 2305.08322.

[14] PAL A, UMAPATHI L K, SANKARASUBBU M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering[EB]. arXiv preprint, 2022, arXiv: 2203.14371.

[15] CAI Y, WANG L L, WANG Y, et al. MedBench: a large-scale Chinese benchmark for evaluating medical large language models[EB]. arXiv preprint, 2023, arXiv: 2312.12806.

[16] ZHANG N Y, CHEN M S, BI Z, et al. CBLUE: a Chinese biomedical language understanding evaluation benchmark[EB]. arXiv preprint, 2021, arXiv: 2106.08087.

[17] JIN D, PAN E, OUFATTOLE N, et al. What disease does this patient have? a large-scale open domain question answering dataset from medical exams[J]. Applied Sciences, 2021, 11(14): 6421.

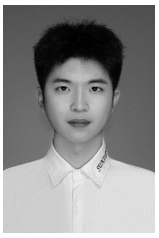
[18] ZHENG S, HUANG J, CHANG K C C. Why does ChatGPT fall short in providing truthful answers?[EB]. arXiv preprint,

2023, arXiv: 2304.10513.
[19] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2001: 311-318.
[20] LIN C Y. Rouge: a package for automatic evaluation of summaries[C]// Proceedings of the Workshop on Text Summarization Branches Out(WAS 2004). Barcelona: Association for Computational Linguistics, 2004: 74-81.
[21] YANG A Y, XIAO B, WANG, B N, et al. Baichuan 2: open large-scale language models[EB], arXiv preprint, 2023, arXiv: 2309.10305.
[22] BAI J Z, BAI S, CHU Y F, et al. Qwen technical report[EB]. arXiv preprint, 2023, arXiv: 2309.16609.
[23] SUN Y, WANG S, FENG S, et al. Ernie 3.0: large-scale knowledge enhanced pre-training for language understanding and generation[J]. arXiv preprint, 2021, arXiv: 2107.02137.
[24] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models[EB]. arXiv preprint, 2022, arXiv: 2206.07682.

## 作者简介

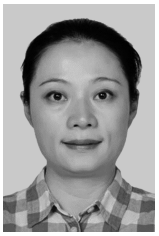
**张芊**(2000- ),女,贵州大学计算机科学与技术学院硕士生,主要研究方向为医疗大模型、智能问答。

**陈攀峰**(1982- ),男,贵州大学校聘副教授,主要研究方向为大数据的融合与集成。



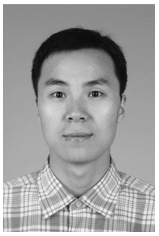
**冯林坤**(2001- ),男,贵州大学计算机科学与技术学院硕士生,主要研究方向为LLM在医学领域的探索。

**刘淑钰**(2000- ),女,贵州大学计算机科学与技术学院硕士生,主要研究方向为医疗人工智能。



**马丹**（1977- ），女，贵州大学计算机科学与技术学院副教授，主要研究方向为数据分析、大数据处理。

**陈梅**（1964- ），女，贵州大学计算机科学与技术学院教授，主要研究方向为大数据管理与分析、人工智能系统、数据库技术。



**李晖**（1982- ），男，博士，贵州大学计算机科学与技术学院教授、博士生导师，主要研究方向为大数据管理与分析、人工智能系统、数据库技术。

**收稿日期：** 2024-04-15

**通信作者：** 李晖，cse.HuiLi@gzu.edu.cn

**基金项目：** 国家自然科学基金项目（No.61462012）；2023年贵州省科技计划项目（黔科合支撑[2023]一般276）；2023年贵州省科技成果应用及产业化计划项目（黔科合成果[2023]一般010）

**Foundation Items:** The National Natural Science Foundation of China(No.61462012), Research Projects of the Science and Technology Plan of Guizhou Province(No.[2023]276), Research Projects of the Science and Technology Achievement Application and Industrialization Plan of Guizhou Province(No.[2023]010)