• 医学信息研究 •

中文医学知识大模型问答语料数据集构建研究*

吕婷钰[1] 李晓瑛[1] 张 颖[1] 刘宇炀[1] 杜晋华[2] 李心怡[2] 罗 妍[1] 唐小利[1]
任慧玲[1] 刘 辉[1] 尹 浩[2]

([1] 中国医学科学院/北京协和医学院医学信息研究所/图书馆 北京 100005
[2] 清华大学网络大数据研究中心 北京 100084)

〔摘要〕 **目的/意义** 构建中文医学知识问答语料数据集，为医学垂域大模型提供标准化的评测基准，进而提升大模型处理中文医学问答任务的准确率和效率。**方法/过程** 构建中文医学论文知识问答数据集、医学名词解释问答数据集和以中国执业医师资格考试真题为基础的问答数据集，整理相关开源数据集。**结果/结论** 自主构建的中文医学知识问答语料数据集丰富了中文医学问答语料来源，能够作为一项标准化的评测基准，推动医学领域大模型实现客观全面的定量评估，今后将利用电子病历、在线健康社区等数据，为健康中国战略的实施提供更坚实的人工智能支持。

〔关键词〕 大语言模型；语料数据集；模型评测；医学
〔中图分类号〕R-058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2024.05.004

Study on the Construction of a Question-Answer Corpus Dataset for Chinese Medical Knowledge Large Language Models

LYU Tingyu[1], LI Xiaoying[1], ZHANG Ying[1], LIU Yuyang[1], DU Jinhua[2], LI Xinyi[2], LUO Yan[1], TANG Xiaoli[1], REN Huiling[1], LIU Hui[1], YIN Hao[2]

[1] Institute of Medical Information & Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, China; [2] Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China

〔**Abstract**〕 **Purpose/Significance** To construct a Chinese medical knowledge Q&A corpus dataset as a standardized evaluation benchmark for large language models (LLMs) in the medical domain, so as to improve the accuracy and efficiency of LLMs in handling Chinese medical questions. **Method/Process** Chinese medical paper knowledge, medical terminology explanations and supplementary questions are acquired from the Chinese medical licensing examination, and open-source Chinese medical Q&A datasets are encompassed in the developed Q&A datasets. **Result/Conclusion** The Chinese medical knowledge Q&A corpus datasets enrich the sources of existing datasets and promote the objective and comprehensive quantitative evaluation of large models in the medical field. In the near future, additional data such as electronic medical records and those from online health communities will be used to strengthen the support of artificial intelligence for the Healthy China strategy.

〔**Keywords**〕 large language models; corpus dataset; model evaluation; medicine

〔修回日期〕 2024-04-04
〔作者简介〕 吕婷钰，硕士研究生；通信作者：李晓瑛，刘辉。
〔基金项目〕 国家社会科学基金项目（项目编号：20BTQ062）；中央高校基本科研业务费资助项目（项目编号：3332023163）。

1 引言

大语言模型（large language models，LLMs）采用深度学习技术，通过预训练学习大量自然语言文本，发挥自注意力机制和 Transformer 结构优势，可用于自然语言处理任务[1]。生成式预训练转换器等 LLMs 以生成式自监督学习为基础，从 TB 级训练数据中学习隐含的语言规律和模式，训练出千亿级别参数量的大语言模型，使机器能够理解、回答和生成人类语言，已在自然语言处理领域取得巨大成功[2]。医学领域数据庞大且复杂，基于医学语料微调的 LLMs 可以更好地满足医学场景下的语境和需求。

医学领域 LLMs 能在一定程度上改变人们与在线医学信息系统的交互方式，如利用 BioGPT[3] 可以丰富和改进医学信息搜索结果、辅助临床决策、总结医学文献等。虽然 LLMs 能够分析海量数据，但医学领域专业术语表达多样化、语义关系高度复杂，因此其无法完全模拟医学专家长期累积的知识经验和判断决策能力[4]。利用大规模医学数据进行预训练则可以增强 LLMs 专业知识覆盖度和上下文理解能力，提升 LLMs 在多种医学自然语言处理任务中的性能[5]。定量评估中文医学领域 LLMs 是当前的关键任务之一，本研究尝试构建中文语境下的医学知识问答语料数据集，用于模型的训练、调优和评测，以促进中文医学知识问答任务的不断发展和应用创新。

2 国内外相关研究进展

目前，国内外正处于医学语料数据集建设和医学知识 LLMs 研发训练、评测应用的关键阶段。医学数据资源类型丰富、规模体量庞大，且应用范围不断扩大，为构建医学领域语料数据集奠定了基础[6]。

2.1 国际研究进展

通过文献调研和网络调研发现，国际医学问答语料数据集工作起步早，从基于生物医学论文的语料数据延伸到医学考试真题、在线医学问答数据和专家注释的问答实例等；数据来源广泛，从以单一数据源为主题发展到多源融合的数据集库。国际上广泛使用且具有较大影响力的医学语料数据集包括 MedQA、MedMCQA、MultiMedQA、PubMedQA 等，见表 1。其中，MedQA[7] 是首个公开可用的针对医疗问题的大规模开放领域多项选择数据集，主要收集医师执照考试真题，支持 LLMs 利用广泛、先进的医学领域知识回答问题；该数据集研发团队还收集并发布了一个大规模医学教科书语料库。MedMCQA[8] 包含两个印度医学院入学试题的模拟题和历年发布的考试中的多项选择题，主题多样丰富，涵盖 2 000 多个医疗保健主题和 21 个医学科目，不仅是可靠和多样化的医学问答评估基准，还可以用于改进医学领域的多项选择题答案自动生成模型，提高医学教育领域的自动化水平。MultiMedQA 已包含 6 个开放式问答数据集[7-11]，新增由谷歌和 DeepMind 开发的在线医学问答数据集 HealthSearchQA，能够用于回答医学考试、医学研究等相关问题，作为评估医学领域 LLMs 临床知识和问答能力的多样化基准[12]。PubMedQA 来源于 PubMed 论文题目和摘要结论句[9]，也是首个针对生物医学科研文本推理回答医学问题的问答数据集；除基于启发式规则实现自动收集的实例外，还拥有最大规模的基于专家注释的生物医学问答实例，能够帮助提升问答系统的稳健性和泛化能力。这些医学问答语料数据集都能被研究人员和开发者用于调优和评估 LLMs，使其更好地理解和回答与医学相关的自然语言问题。

表 1 国际医学语料数据集建设

序号	数据集名称	数据概况
1	MedQA	收集美国和中国的医学委员会考试题，涵盖英语、简体中文和繁体中文的 12 723、34 251 和 14 123 个问题
2	MedMCQA	整合超过 19.4 万条高质量的医学院入学考试多项选择题，涵盖 2 000 多个医疗保健主题和 21 个医学科目
3	MultiMedQA	包含 HealthSearchQA 在内的 3 173 个在线搜索医学问题与 6 个现有的开源数据集
4	PubMedQA	拥有 1 000 个专家注释、60 000 余个未标记和 21 000 余个人工构建的医学论文问答实例

2.2 国内研究进展

在国际医学 LLMs 和语料数据集的启发下，中文医学语料数据集建设取得显著成果，典型代表如 MLEC - QA、CMExam、CMB（Comprehensive Medical Benchmark in Chinese）等，见表 2。其中，MLEC - QA 是目前规模最大的中国执业医师资格考试题数据集，涵盖临床、口腔医学、公共卫生、中医、中西医结合等多个学科[13]。CMExam 作为首个提供医学注释的中国临床医学检查数据集，包含医学考试的多项选择题和医学专家的注释[14]。中文医疗模型评估基准 CMB 包括不同临床职业、不同职业阶段考试中的多项选择题 CMB - Exam 和基于真实病例的复杂临床诊断问题 CMB - Clin[15]。CMB 数据集能更好地基于中国本土文化环境提供情境化的基准，支持评估医学领域 LLMs 的综合知识问答和推理能力。但是，目前用作 LLMs 评测的中文医学语料数据集的数据来源单一，主要聚焦于中国执业医师资格考试题、临床检查数据。

表 2 国内语料数据集建设

序号	数据集名称	数据概况
1	MLEC - QA	由 5 个子集构成，包含 136 236 个生物医学领域多项选择题，配有医学专家注释的图像或表格
2	CMExam	包含来自中国执业医师资格考试的 60 000 多个多项选择题，医学专家标注包括疾病群组、临床科室、医学学科、能力领域和问题难度级别 5 个问题的注释
3	CMB	包括中国执业医师资格考试中的多项选择题和基于实际病例研究的复杂临床诊断问题。临床诊断问题基于实际的病例设置，由医学专家达成共识后确定正确答案

3 问答语料数据集构建研究

自然语言理解的长期目标是构建自然语言推理的智能系统，而以标准化和规范化医学问答语料数据集为基础的自动问答任务是评估智能系统推理能力的通用方法之一。目前国际相关机构建设的数据集大多基于美国执业医师考试题、英文医学论文摘要和临床医学检查等数据，针对中文医学论文的语料数据集构建实践较少。本研究以中文医学论文题目和摘要特征为主体构建中文医学论文知识问答数据集，以中文医学专业名词、中国执业医师资格考试真题、医学领域开源问答数据集作为补充，整合并转化为 LLMs 评测可直接使用的规范格式。基于该语料数据集调优的医学垂域 LLMs 已在同步建设中，也将在调优完成后采用此语料数据集评测模型。

3.1 中文医学论文知识问答数据集

以中文医学论文摘要为基础的数据涵盖广泛的生物医学主题，如药理学、医学影像学、遗传学、分子生物学等众多与医学相关的学科和领域，是训练和评估医学领域 LLMs 的重要资源。因此，本研究基于自主加工的中文医学论文题目、摘要和摘要结论句，构建中文医学论文问答数据集，形成由“问题”“上下文”和“长答案”组成的结构化问答对。其中，“问题”衍生自现有的医学论文标题，“上下文”是论文的摘要部分，“长答案”为摘要的结论部分。鉴于中文医学论文很少直接以疑问句作为论文标题，改造论文标题作为问题，将论文摘要结论句作为问题的答案（即长答案）。已初步整理出 10 类摘要处理模式，见表 3，经过人工审核后转化 312 243 条问答对，后续将进一步对其补充完善。

表 3 中文医学论文知识问答数据集构建模式

序号	模式	问题	回答
1	A 对 B 的敏感性	miR－503 下调 Bcl－2 增强 BEL－7402/5－FU 细胞对 5－氟尿嘧啶的敏感性有影响吗？	miR－503 可能通过下调 Bcl－2 的表达，进而增强 BEL－7402/5－FU 细胞对 5－氟尿嘧啶的药物敏感性，抑制细胞增殖
2	A 对 B 的干预作用	维康颗粒对亚健康疲劳小鼠不同组织氧化应激损伤的干预作用是什么？	健康疲劳状态下小鼠存在显著的氧化应激损伤，维康颗粒可通过降低模型小鼠的氧化应激损伤，促进亚健康疲劳状态向正常状态转化
3	A 的药理作用	气道黏液调节剂的药理作用是什么？	在慢性呼吸道炎症疾病过程中，气道分泌细胞增生肥大、黏液分泌亢进，形成气道黏液高分泌。气道黏液调节剂能够抑制黏蛋白的产生及分泌，对气道黏液高分泌有治疗作用
4	A 对 B 的影响	Ad－HIF－1α 双突变体对大鼠缺血心肌及 Na～＋－K～＋ATP 酶的影响是什么？	Ad－HIF－1α－564/402 能减轻心肌细胞梗死程度，保护心脏功能，对 Na～＋－K～＋ATP 酶水平无明显影响
5	A 与 B 的关系探讨	HCG 日孕酮水平与体外受精－胚胎移植的治疗结局是什么关系？	HCG 日血清孕酮浓度为 3.0～4.77 nmol/L 能获得更成熟的卵子，同时不增加因多卵泡发育引起卵巢过度刺激综合征导致的周期取消率
6	A 与 B 的比较	MRA 与颈部血管彩超诊断颈动脉狭窄的比较结果是什么？	MRA 对脑梗死患者颈部血管检查准确率优于颈部血管彩超，当患者颈部血管病变越严重时二者差异越大，且两种检查方法均与 DSA 有一定的差距，应仔细斟酌二者的结果
7	A 与 B 的关联性	血脂相关基因 DNA 甲基化水平与冠心病有什么关联性？	两组间脂代谢相关基因的 DNA 甲基化水平差异有统计学意义，提示脂代谢相关基因的 DNA 甲基化在冠心病中具有一定的调控作用
8	A 的临床疗效	阿替普酶静脉溶栓对穿支梗死患者的临床疗效如何？	采用 rt－PA 静脉溶栓治疗穿支梗死患者，有利于降低炎症因子水平及神经功能缺损程度，能提升临床疗效，临床应用安全可靠
9	A 的治疗策略	颅颈交界区髓周动静脉瘘患者血管构筑特点及治疗策略是什么？	颅颈交界区 PAVF 患者多伴脊髓动脉参与供血和动脉瘤样结构，多以出血起病，更多使用介入栓塞联合显微手术治疗。PAVF 患者预后与非 PAVF 组类似，大多患者预后良好
10	A 对 B 的治疗效果	氯吡格雷联合丁苯酞注射液对急性脑梗死的治疗效果如何？	氯吡格雷联合丁苯酞注射液对急性脑梗死患者的治疗效果较好，能促使患者的神经功能、抗氧化功能得到改善，值得在临床上推广

3.2 医学名词解释问答数据集

医学领域专业词汇对提升 LLMs 的学科知识理解能力非常重要，可以弥补医学论文知识问答数据集构建中对于医学术语解释深度和专业度的不足[16]。本研究构建的中文医学语料数据集涵盖 2.7 万个来自《中文医学主题词表》（Chinese Medical Subject Headings，CMeSH）的医学名词及其定义注释信息。从词表中获取“名词”和“定义注释”部分，以“什么是｛名词｝?”的格式作为问题提出，并以“｛定义注释｝”作为问题的答案。最终，基于医学名词构建的结构化问答数据集收录了经人工审核的疾病、解剖学、有机体、化学物质和药物、诊疗技术与设备等多种类别的基础医学知识，具有较高的专业性和通用性。其中肿瘤、病毒、身体部位 3 类的问答数据示例，见表 4。

表 4 医学名词解释问答模式构建示例

序号	名词	问题	回答
1	炎性乳腺肿瘤	什么是炎性乳腺肿瘤？	转移性乳腺癌，其特点为由淋巴转移和最终癌细胞阻塞淋巴管所致的受累乳腺的水肿和红斑
2	肝炎病毒	什么是肝炎病毒？	任何引起肝脏炎症的病毒。包括来源于人类和动物的 DNA 病毒和 RNA 病毒
3	消化系统	什么是消化系统？	一组从口腔延伸至肛门的器官，用于分解食物、吸收营养和排泄废物。人类消化系统包括胃肠道和附腺（肝、胆道、胰腺）

3.3 中国执业医师资格考试真题

语料数据集能否及时更新是衡量其内容质量和生命力的重要指标。本研究筛选了 2 960 道近年来中国执业医师资格考试真题作为 CMB – Exam 的补充，涵盖临床、中医、口腔、公共卫生 4 个领域。执业医师资格考试真题设计严谨、考察范围全面，通常涉及真实复杂的医学场景和案例，具有较高的专业性和权威性，有助于提高语料数据集的适用性。以 2022 年某单选题为例，经格式转化后的数据示例如下。

问题：急性肾损伤后，患者出现肾小管原尿回漏的原因是?

选项：A 肾小球滤过率降低

B 肾间质水肿，挤压肾小管

C 输尿管梗阻

D 肾小管上皮细胞坏死、脱落

E 肾小管的水钠重吸收减弱

回答：D

3.4 开源问答数据集

开源数据集含有大量真实和丰富的医学知识，为构建和扩展中文医学语料数据集提供了重要资源。具体而言，Chinese – medical – dialogue – data 数据集[17]涵盖内科、妇产科、肿瘤科等 6 个医学领域，包含 792 099 个问答对。MedDialog 中文数据集[18]收录医生和患者的 110 万条中文对话，数据总量达 400 万条。Huatuo Encyclopedia QA 数据集[19]不仅收集中文维基百科中的疾病和药物百科条目，而且覆盖来自健康网站的医学文章，共计 364 420 条医学问答数据。以上开源问答数据集在经过专家审核后纳入问答语料数据集的构建，因其包含现实的医疗场景问答数据和丰富的网络医学资源，能够在提高语料数据集灵活性方面有所助益。

4 结语

本研究通过分析中文医学论文题目和摘要特征，总结归纳了 10 种问答模式，形成中文医学论文知识问答数据集；基于中文医学专业名词建设成果，构建医学名词解释问答数据集；整理近年来中国执业医师资格考试真题，作为 CMExam、CMB – Exam 的补充数据；收集 Chinese – medical – dialogue – data、MedDialog 等开源问答数据集，提升医学知识 LLMs 的多轮问答能力。这些数据集经整合处理，形成中文医学知识问答语料数据集库，并转化为 LLMs 调优和评测可直接使用的规范格式，以期作为中文医学领域 LLMs 专业知识理解能力和生成能力的评估基准。

自主构建中文医学问答语料数据集，既能为研发医学知识 LLMs 提供坚实的基础数据支持，实现标准化和客观全面的定量评估，也能从语料数据层面提升医学 LLMs 的知识覆盖度、问答能力和行业认可度，促进医学科技自立自强。今后，将进一步扩充和更新中文医学语料数据集的内容，利用电子病历、在线健康社区等真实世界数据持续开展相关研究，探索医学知识 LLMs 在临床智能诊疗、病历自动书写等任务的应用效果，拓展新一代人工智能技术的应用范围，为健康中国战略实施提供更广泛、更坚实的人工智能支持。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 岳增营，叶霞，刘睿珩. 基于语言模型的预训练技术研究综述 [J]. 中文信息学报，2021，35（9）：15 – 29.

2 RADFORD A，NARASIMHAN K，SALIMANS T，et al. Improving language understanding by generative pre – training [EB/OL]. [2023 – 09 – 10]. https：//cdn. openai. com/research – covers/language – unsupervised/language_ understanding_ paper. pdf

3 LUO R，SUN L，XIA Y，et al. BioGPT：generative pre – trained transformer for biomedical text generation and mining [J]. Briefings in bioinformatics，2022，23（6）：409.

4 马武仁，弓孟春，戴辉，等. 以 ChatGPT 为代表的大语言模型在临床医学中的应用综述 [J]. 医学信息学杂志，2023，44（7）：9 – 17.

5 BOLTON E，HALL D，YASUNAGA M，et al. BioMedLM [EB/OL]. [2023 – 12 – 22]. https：//crfm. stanford. edu/2022/12/15/biomedlm. html.

6 吕文蓉. 医学域机器阅读理解研究及系统实现 [D]. 兰州：西北民族大学，2023.

7 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.

8 PAL A, UMAPATHI L K, SANKARASUBBU M. MedmcQA: a large – scale multi – subject multi – choice dataset for medical domain question answering [EB/OL]. [2023 – 12 – 10]. https://proceedings.mlr.press/v174/pal22a.html.

9 JIN Q, DHINGRA B, LIU Z, et al. PubMedQA: a dataset for biomedical research question answering [C]. Hong Kong: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019.

10 ABACHA A B, AGICHTEIN E, PINTER Y, et al. Overview of the medical question answering task at TREC 2017 LiveQA [EB/OL]. [2023 – 12 – 10]. https://trec.nist.gov/pubs/trec26/papers/Overview – QA.pdf.

11 HENDRYCKS D, BURNS C, BASART S, et al. Measuring massive multitask language understanding [EB/OL]. [2023 – 12 – 10]. https://github.com/hendrycks/test.

12 SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge [J]. Nature, 2023, 620 (7972): 172 – 80.

13 LI J, ZHONG S, CHEN K. MLEC – QA: a Chinese multi – choice biomedical question answering dataset [C]. Punta Cana: The 2021 Conference on Empirical Methods in Natural Language Processing, 2021.

14 LIU J, ZHOU P, HUA Y, et al. Benchmarking large language models on CMExam – – a comprehensive Chinese medical exam dataset [EB/OL]. [2023 – 09 – 10]. https://arxiv.org/abs/2306.03030.

15 WANG X, CHEN G H, SONG D, et al. CMB: a comprehensive medical benchmark in Chinese [EB/OL]. [2023 – 09 – 10]. https://arxiv.org/abs/2308.08833.

16 GU Y, TINN R, CHENG H, et al. Domain – specific language model pretraining for biomedical natural language processing [J]. ACM transactions on computing for healthcare (HEALTH), 2021, 3 (1): 1 – 23.

17 Toyhom. Chinese – medical – dialogue – data [EB/OL]. [2023 – 09 – 10]. https://github.com/Toyhom/Chinese – medical – dialogue – data.

18 ZENG G, YANG W, JU Z, et al. MedDialog: a large – scale medical dialogue dataset [C]. online: The 2020 Conference on Empirical Methods in Natural Language Processing, 2020.

19 LI J, WANG X, WU X, et al. Huatuo – 26M, a large – scale Chinese medical QA dataset [EB/OL]. [2023 – 09 – 10]. https://arxiv.org/abs/2305.01526. publications/i/item/depression – global – health – estimates.

(上接第 19 页)

27 RAMEZANI M, TAKIAN A, BAKHTIARI A, et al. The application of artificial intelligence in health financing: a scoping review [J]. Cost effectiveness and resource allocation, 2023, 21 (1): 83.

28 XIANG Y, DU J, FUJIMOTO K, et al. Application of artificial intelligence and machine learning for HIV prevention interventions [J]. Lancet HIV, 2022, 9 (1): e54 – e62.

29 于帆, 何海洪, 周义文. 人工智能在检验医学领域的应用进展 [J]. 国际检验医学杂志, 2023, 44 (18): 2267 – 2273.

30 Intel. 智慧医院解析: 智能、互联、安全 [EB/OL]. [2024 – 01 – 22]. https://www.intel.cn/content/www/cn/zh/internet – of – things/smart – hospital – intelligent – interconnect – safe.html.

31 吴英华, 罗家锋, 林成创. 迈入人工智能新时代: ChatGPT 在智慧医疗应用场景研究与思考 [J]. 数据通信, 2023, (4): 33 – 38, 54.

32 彭亮, 孙磊. 人工智能医疗器械监管研究进展 [J]. 中国食品药品监管, 2022 (2): 30 – 35.

33 徐冬. 智慧医院评价指标研究 [J]. 医学信息, 2024, 37 (6): 42 – 46, 51.

34 曾杏珍, 陈芸, 卢红, 等. 我国智慧医院技术应用现状及问题对策研究 [J]. 中国卫生信息管理杂志, 2021, 18 (4): 514 – 520.

35 RODRIGUES L, GONÇALVES I, FÉ I, et al. Performance and availability evaluation of an smart hospital architecture [J]. Computing, 2021, 103 (10): 2401 – 2435.

36 但汉亮, 黎宗毅, 曾凝, 等. 基于全新诊疗模式的智慧医院实践与探索 [J]. 现代医院, 2024, 24 (1): 84 – 87.

37 秦帆, 吴永仁, 邵军, 等. 公立医院智慧后勤管理体系建设研究与实践 [J]. 中国卫生信息管理杂志, 2024, 21 (2): 192 – 197.

38 KONG X, AI B, KONG Y, et al. Artificial intelligence: a key to relieve China's insufficient and unequally – distributed medical resources [J]. American journal of translational research, 2019, 11 (5): 2632 – 2640.