·中医药大模型研究·

基于检索增强生成技术的中医药问答大语言模型的构建

张玉铭[1,2],李红岩[1,2],郎许锋[1,2],周作建[1,2],凌云[3],王子琰[4]
(1. 南京中医药大学人工智能与信息技术学院,江苏 南京 210023;2. 江苏省智慧中医药健康服务工程研究中心,江苏 南京 210023;3. 南京中医药大学中医学院,江苏 南京 210023;4. 南京中医药大学中医药文献研究院,江苏 南京 210023)

摘要:目的 构建基于检索增强生成技术的中医药问答大语言模型。**方法** 收集中医古籍《伤寒论》、中医教材、名老中医经方及其他人工标注的中医数据集组建中医药语料库,构建中医药知识向量库;将检索增强生成(RAG)技术结合 P-Tuning v2 微调方法与大语言模型(ChatGLM2-6B)进行融合构建中医药问答大语言模型。**结果** 以精确率、召回率与 *F*1 值为知识问答任务的评价指标进行验证,在简单类中医问答可以达到 90%以上的准确率,其中成分类问题的回答准确性最高,*F*1 值达到 0.928,中高难度问答准确率在 75.8%~87.7%之间,*F*1 值均达到 0.766 以上;以多样性和准确性为中医问题生成任务的评价指标进行专家打分,研究模型相较于基座模型高出了 9.5 分。**结论** 研究模型具备良好的语义理解能力和较高的可靠性,有效缓解了模型幻觉并帮助患者明确问题意图,对推进中医药知识的研究以及人性化的交互式回答具有重要意义,为促进中医经验的传承与普及、中医诊疗智能化建设提供了创新方式。

关键词: 中医药知识库; 大语言模型; 问答系统;检索增强生成技术
中图分类号: R2-03 **文献标志码:** A **文章编号:** 1672-0482(2024)12-1375-08
DOI: 10.14148/j.issn.1672-0482.2024.1375
引文格式: 张玉铭,李红岩,郎许锋,等. 基于检索增强生成技术的中医药问答大语言模型的构建[J]. 南京中医药大学学报,2024,40(12):1375-1382.

Construction of Traditional Chinese Medicine Question-Answering Large Language Model Based on Retrieval-Augmented Generation Technology

***ZHANG Yuming*[1,2], *LI Hongyan*[1,2], *LANG Xufeng*[1,2], *ZHOU Zuojian*[1,2], *LING Yun*[3], *WANG Ziyan*[4]**
(1. School of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, Nanjing 210023, China; 2. Jiangsu Province Engineering Research Center of TCM Intelligence Health Service, Nanjing 210023, China; 3. School of Chinese Medicine, Nanjing University of Chinese Medicine, Nanjing 210023, China; 4. Institute of Literature in Chinese Medicine, Nanjing University of Chinese Medicine, Nanjing 210023, China)

ABSTRACT: OBJECTIVE To construct a large language model for TCM question-answering. **METHODS** TCM corpora were built by collecting TCM classics such as *Treatise on Cold Damage*, TCM textbooks, prescriptions from famous TCM doctors, and other manually annotated TCM datasets. A TCM knowledge vector library was constructed. The RAG technology was fused with the P-Tuning v2 fine-tuning method and the large language model (ChatGLM2-6B) to build the TCM question-answering large language model. **RESULTS** Recision, Recall, and *F*1 score were used as evaluation metrics for knowledge question-answering tasks. The model achieved over 90% accuracy in simple TCM question-answering, with the highest accuracy in component-type questions, reaching an *F*1 score of 0.928. The accuracy of medium to high difficulty questions ranged from 75.8% to 87.7%, with *F*1 scores all exceeding 0.766. Expert ratings based on diversity and accuracy were used as evaluation metrics for TCM question generation tasks, and the model in this paper scored 9.5 points higher than the baseline model. **CONCLUSION** The model in this paper demonstrates good semantic understanding and high reliability, effectively alleviating model hallucinations and helping patients clarify their question intentions. It is of great significance for advancing research on TCM knowledge and providing personalized interactive answers. It also provides an innovative approach to promoting the inheritance and popularization of TCM experience and the intelligent construction of TCM diagnosis and treatment.

KEYWORDS: TCM knowledge base; large language model; question-answering system; retrieval-augmented generation technology

收稿日期: 2024-05-15
基金项目: 国家中医药管理局高水平中医药重点学科建设项目(国中医药人教函〔2023〕85 号);江苏省中医药科技发展计划项目(MS2023010);2023 年江苏省研究生科研创新计划(KYCX23_2084)
第一作者: 张玉铭,女,硕士研究生,E-mail:zym@njucm.edu.cn
通信作者: 周作建,男,研究员,主要从事中医药人工智能与大数据分析研究,E-mail:zhouzj@njucm.edu.cn

中医药涉及复杂的理论体系，如阴阳五行、经络脏腑等，有效理解中医药文本中的专业术语和概念至关重要。加之医学信息的歧义和语境丰富性对自然语言处理提出的一系列挑战，主要体现在术语在不同上下文具有不同的含义、问题及文本规范化等方面，因此研究者将焦点转向了智能问答系统，以期更好提高医疗服务效率、辅助医疗决策。

目前智能问答系统主要分为基于文档集的问答系统、基于知识图谱的问答系统和基于生成式模型的问答系统 3 类。但基于文档集的问答系统存在问答的准确率、系统性能较低等问题[1]。基于知识图谱的问答系统具有较强的可解释性[2-3]，但面临知识图谱的构建和更新维护成本高、不可扩展等问题。生成式问答系统更符合人类交互习惯[4]，多用于文本生成和对话系统等任务。其中，神农大语言模型[5]、华佗大语言模型[6]、ChatDoctor[7]等均是基于医学专业领域知识训练的大语言模型，通过结合中医临床辨证思维策略，模拟现有病案生成患者主诉进行情景互动式交流，可辅助提高中医诊疗质量。但当前生成式语言模型多基于大量参数存储知识，在未经训练的数据上存在过度泛化的风险，易生成一些虚构的“事实”，从而引发“幻觉”现象[8]，误导用户。近年来，Meta AI 提出了检索增强生成(RAG)技术[9]，该方法利用文本嵌入模型和生成模型的协同作用，将带有参数记忆的隐式知识模型与非参数记忆的外部知识库结合，有利于提升知识回答的可信度和准确度。同时 ChatGLM2-6B 推出了基于 P-Tuning v2[10] 的高效参数微调，允许创建和优化特定任务的提示来指导预训练模型生成所需要的输出内容，小样本甚至是零样本的微调性能也有较好的效果，可以更灵活地适应各种下游任务。

因此，本研究在预训练的 ChatGLM2-6B 基础之上，利用 RAG 技术进行中医知识库的检索以构造 prompt 提示输入基座模型增强生成响应的能力，使用 P-Tuning v2 方法对模型进行问题生成微调，构建中医药问答大语言模型。完成微调后通过双向评估模型由问题生成答案及由答案反推问题的准确性，验证本研究模型在理解和生成中医领域答案与问题的一致性和准确性，从而依据自身反馈缓解模型“幻觉”，引导用户明确问题意图，以期实现中医药问答大语言模型在实际场景中的高效应用与自动问答。

1　中医药问答大语言模型

本研究模型的构造和训练共需要 4 个步骤：①收集中医典籍数据并使用 m3e-base 进行语义向量化，构建中医药知识向量库；②利用 RAG 技术检索知识库获得问题对应的答案作为 prompt 优化基座模型；③通过 P-Tuning v2 方法使用中医文献问题进行微调；④构造标准评估集并使用 BertScore 指标及专家评估对优化后的模型进行验证。具体方案设计如图 1 所示。

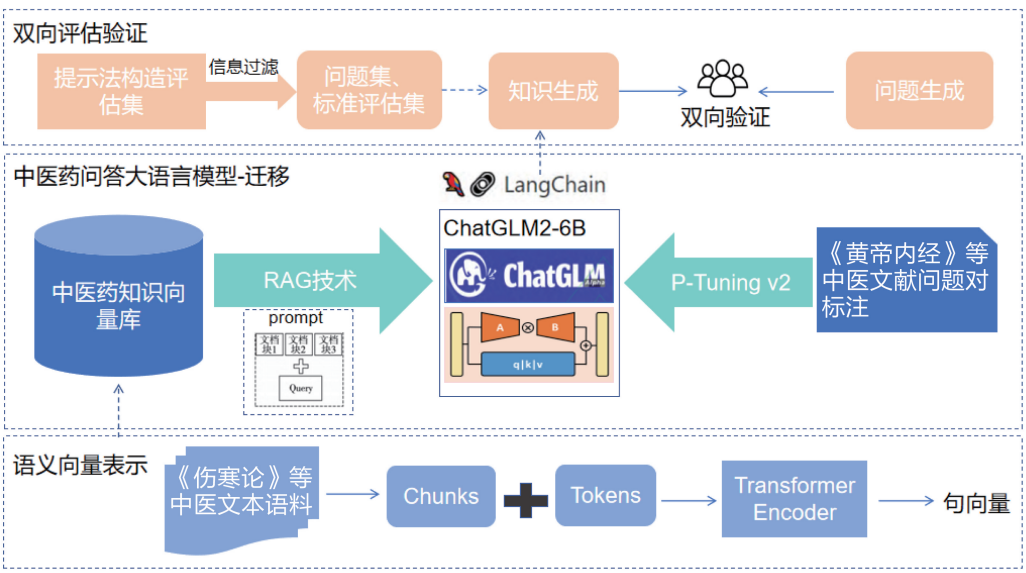

图 1　中医药问答大语言模型构建流程图

Fig. 1　Flowchart of the construction of a TCM question-answering large language model

1.1　基座模型选择

中医药理论体系复杂、知识丰富,因此该领域的智能问答模型需具备长距离记忆和推理能力,本研究选择了多头自注意力机制和序列到序列架构等特性的模型。大语言模型(GLM)[11]是一种专门针对中文问答进行优化的千亿参数规模的中英文语言模型,在 NLU、条件和无条件生成的任务上表现显著优于 BERT、T5 和 GPT,证明了其对不同下游任务的普适性。GLM 通过优化自回归填空目标进行训练,令 Z_m 为长度为 m 的索引序列[1,2,……,m] 的所有可能排列的集合,从输入文本 x 采样出多个文本片段[S_1,……,S_m],$S_{z<i}$ 表示[S_{z1},……,S_{zi-1}],输入 x 被分成 2 个部分:一部分是损坏的文本 $x_{corrupt}$,另一部分包含了被屏蔽的跨度。定义了预训练目标为公式(1)。

$$\max_{\theta} E_{z\sim Z_m}\left[\sum_{i=1}^{m}\log p_\theta(s_{zi}\mid x_{corrupt},s_{z<i})\right] \tag{1}$$

ChatGLM2-6B 作为 ChatGLM-6B 的升级版,采用了 GLM 的混合目标函数,展现出了更强大的性能。与前代模型相比,它在 MMLU、CEval、GSM8K、BBH 等多个数据集上实现了显著的性能提升,分别达到了 23%、33%、571%、60% 的增长[12]。此外,ChatGLM2-6B 在推理速度上也有了 42%的提高,特别是在 INT4 量化技术的支持下,该模型在 6 GB 显存的条件下支持的对话长度从 1 K 提升到了 8 K,便于在资源受限的设备上进行部署,故本研究采用 ChatGLM2-6B 作为基座模型。

1.2　中医药知识库的构建与响应

本研究收集多个权威可靠的中医药数据用于中医药知识库构建,包括《伤寒论》、名老中医经验方、“十三五”规划中医教材,并使用《黄帝内经》翻译版等开源数据人工标注的问答对微调问题生成模型。通过手工筛除近似内容、处理特殊字符,并将所有内容进行结构化,例如中药方剂注明方名、成分、功用、方解、主治疾病、证候、证等。在以上收集到的语料数据预处理后对其进行语义向量表示。首先使用 TextSplitter 对文本进行分块,设该文本数据中的第 i 个文件经过分块处理后得到的文本块表示为 Dij,其中 i=[1,2,3,……,n] 表示文件序号,j=[1,2,3,……,m]表示文本块序号;再利用 Sentence Transformes 库[13]中的嵌入模型 m3e-base 将每个文本块转换为句向量表示。对于每个文本块 Dij,其对应的向量表示为 Vij,E 表示嵌入模型,由此构建中医药知识向量,见公式(2)。

$$Vij=E(Dij) \tag{2}$$

本研究所使用的 RAG 技术包括 2 部分,首先从中医药知识向量库检索相关答案组成问答对,再将其输入大语言模型进行生成增强以验证本研究模型回答的准确性。为加速在大规模向量数据中的搜索,采用了近似最近邻搜索算法,利用 FAISS 库[14]构建向量索引。FAISS 库基于倒排索引的方法,通过将相似的向量聚类,并将它们分配到同一个组中。在检索时,先在组间搜索,粗略定位答案所在组,以减小搜索空间,提升检索效率[15];确定答案所在组之后,再对组内向量进行逐一匹配,以实现更精确的搜索结果。通过这种方法,可以快速检索出与给定问题最相似的文本块。最终,采用向量相似度计算的方法从知识库中搜索出与查询条件最相关的前 k 个条目,并返回其对应的文本块作为参考依据,见公式(3)(4)。

$$Top\ k(candidates)=argmax_k similarity(query,candidate) \tag{3}$$

$$similarity(ab)=\frac{a\cdot b}{\|a\|\,b\|} \tag{4}$$

其中,$candidate$ 候选集,$query$ 是查询向量,$similarity(query,candidate)$是查询向量和候选向量之间的相似度。基于以上步骤构建的中医药知识库能够接受文本查询并返回前 k 条参考答案。

通过将检索到的前 k 个条目与问题一起添加到 prompt 提示模板,并输入到预训练的 ChatGLM2- 6B 大语言模型中生成最终的答案,确保了答案是根据上下文语义信息得出并提供了解释或来源,如图 2 所示。

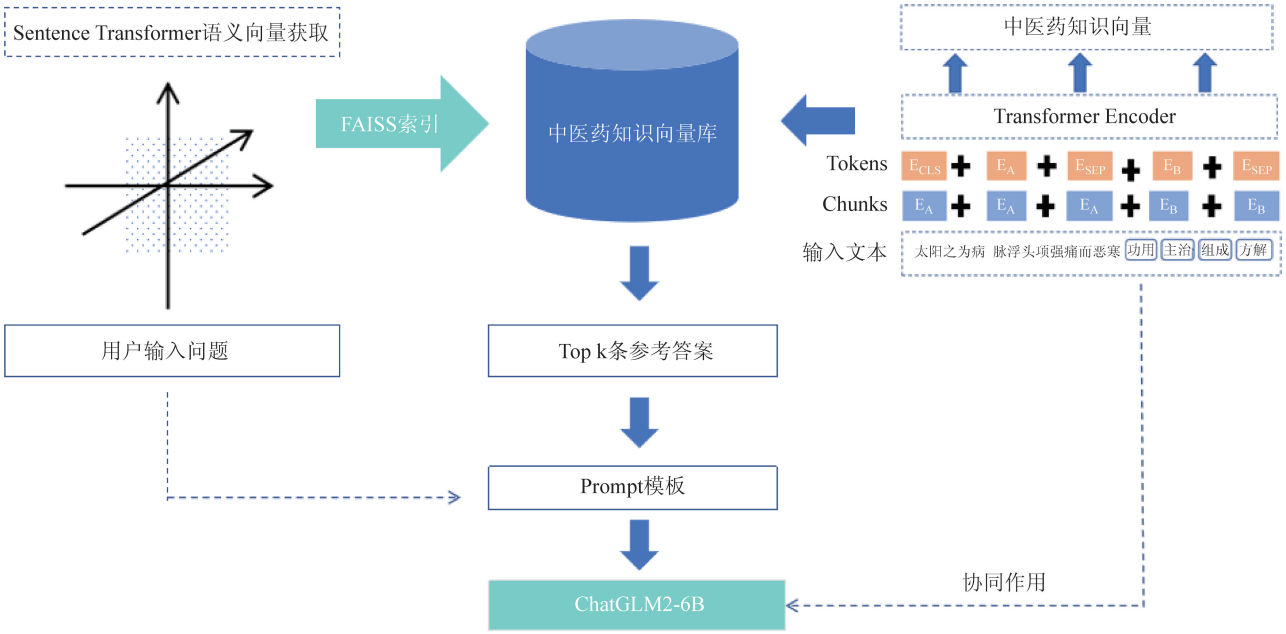

图 2　中医药知识库构建与响应

Fig. 2　Construction and response of TCM knowledge base

1.3　中医药问题生成微调

为了进一步明确用户意图、引导其规范表达，在 RAG 技术优化模型的基础上采用 P-Tuning v2 方法对 ChatGLM2-6B 进行微调，使其根据输入答案生成问题，评估大语言模型生成问题的准确性。通过对开源数据《黄帝内经》翻译版等文档进行整理，每篇由人工标注产生 1～4 对问题、文档、答案用作训练数据，随机抽取 20 条作为测试集。在指令“instruction”中，提供了模型当前的任务身份和任务描述；“input”字段提供了文档的内容，模型需要根据这个文档来生成问题；而“output”字段则表示人工标注的问题。以清晰地指示模型在训练过程中的角色和任务目标。指令训练格式如下。

{“instruction”：“你现在作为一个问题生成模型，请根据下面文档生成一个问题文档”

“input”：(文档)

“output”：(人工标注的问题)}

1.4　标准评估集构造

为客观地评估本研究模型回答结果的准确性，本研究构建了中医药大语言模型在线版本(http://10.120.53.205:8501)，通过分析用户所提问题的类型了解人们对该领域的认知程度及重点关注的问题。结果显示人们对于中药的分类、成分、属性、归经、治疗作用、适用症状以及药物配伍等方面较感兴趣。根据用户提问的需求点，可以将其归纳为成分类、判断类、功效类、主治类以及属性类。

依据以上分类结果，本研究将专家经方输入 ChatGPT，并提供 prompt 模板生成规范化问答以表示其中的方剂、成分、症状、功效等内容，汇入问题集 Q 和评估集 E 中作为标准数据集，如图 3 所示。除此之外，本研究还采用 TF-ID 算法结合余弦距离方法对同一类问题的相似度进行两两计算，之后选择相似度高于 0.8 的数据对进行合并删除，构成一个经过滤的问题集[16]。

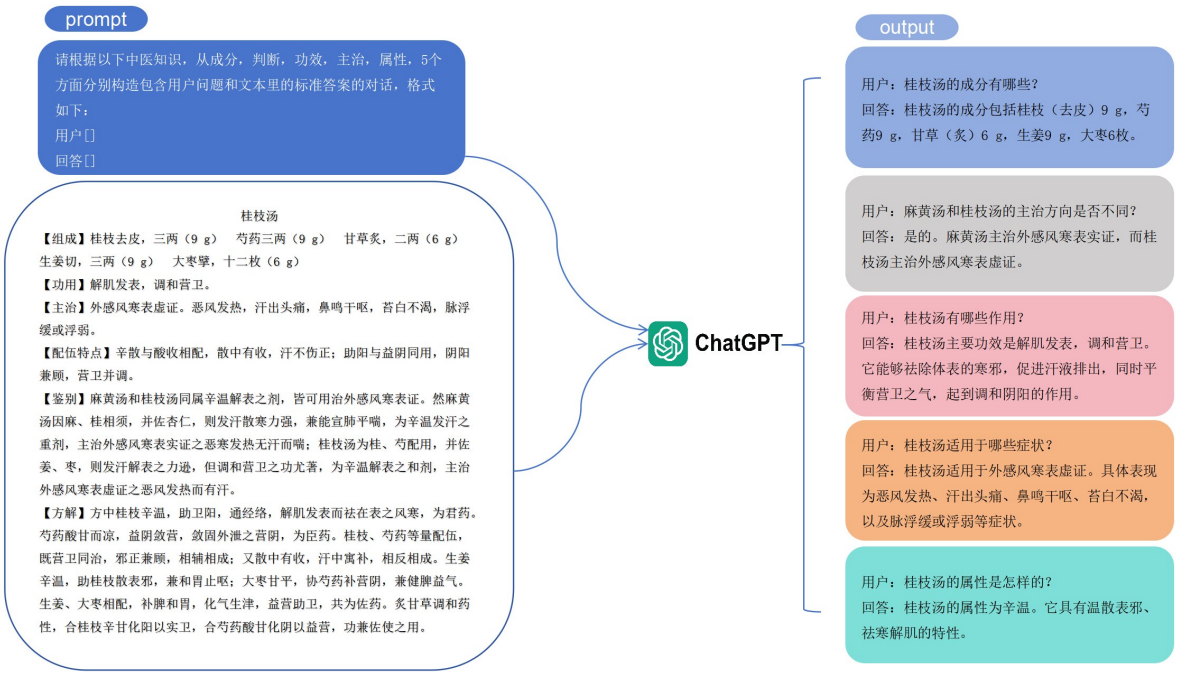

图 3　标准评估集构造

Fig. 3　Construction of standard evaluation set

2　方法与结果

2.1　实验环境及参数配置

本研究实验环境配置为:4 块 NVIDIA T4 GPU;python 版本 3.10;cuda 版本 12.0。实验超参数设置如表 1 所示。

表 1　实验超参数设置表

Table 1　Experimental hyperparameter settings

参数名称	参数值	说明
per_device_train_batch_size	1	每个设备的训练批次大小
max_len	1 560	输入序列的最大长度
max_src_len	1 024	源序列的最大长度
learning_rate	$1e^{-4}$	学习率
weight_decay	0.1	权重衰减

2.2　实验结果与讨论

BERTScore[17]通过 BERT 提取生成文本和参照文本的特征,计算 2 个句子中每个词的内积生成相似性矩阵,再基于该矩阵使用最大似然性得分的累加进行归一化计算相似度,得到最终的相似度评分。BERTScore 在系统级和语段级的相关性方面与人类评判更加接近,相较于 BLEU 具有更强的模型选择性能。以下为对川芎功效的生成文本与参考文本的相似性矩阵计算示意图(图 4)。

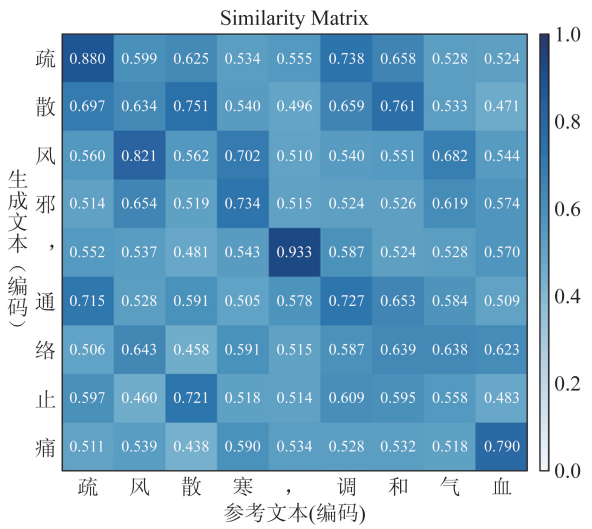

图 4　川芎功效相似性矩阵示意图

Fig. 4　Schematic diagram of the similarity matrix of Ligusticum wallichii efficacy

2.2.1　模型生成答案效果分析　依据模型回答结果的好坏,简单类问题包括方剂的组成和成分问题、简单的方剂鉴别以及是非判断问题;中高难度问题则涉及探究经验方的益处和功能的功效类问题,对疾病的治疗效果的主治类问题以及关注方剂配伍特点的属性类问题。通过对中医领域的提问进行分

类,可以更全面地评估本研究模型的性能。在本研究中,基于相似度矩阵计算最终得到准确率(Precision)、召回率(Recall)和 $F1$ 值,随机选用标准评估集的 50 条问题数据进行测评,如表 2 所示。

表 2　答案生成评估

Table 2　Answer generation evaluation

问题类别	本研究模型			ChatGLM2-6B		
	准确率	召回率	$F1$	准确率	召回率	$F1$
成分类	0.927	0.931	0.928	0.832	0.790	0.810
判断类	0.916	0.895	0.905	0.804	0.790	0.794
功效类	0.877	0.868	0.872	0.777	0.816	0.795
主治类	0.837	0.822	0.829	0.799	0.766	0.782
属性类	0.758	0.775	0.766	0.579	0.570	0.573

由表 2 可知,通用模型在成分类问题表现较好,$F1$ 值达到 0.810,但中高难度问题上表现较差,表明原有模型在处理需要具体理解细节的任务时,其生成的回答可能不够精确。引入本地知识库后,模型能够获得更可靠的信息来源,显著提高了回答的准确性和整体性能,在处理答案确定的成分类等简单问题时,本研究模型往往能够给出较为准确的响应,$F1$ 值达到 0.928。由于实际问题的广泛性和复杂性可能导致模型理解上有偏差,还需综合人工评估并结合任务需求和关注点进行综合分析。

2.2.2　模型生成问题效果评价　本研究模型还需通过专家从多样性和准确性两个维度判断中医问题生成能力的好坏,因此制定了评分规则。每个样本有 5 个评价维度 $score_x$,$x=1,2,\cdots\cdots,5$,每个 $score_x$ 为 1 分,每个样本总计 5 分,总共 20 个样本,评分满分 100 分;$score_{ix}$ 为第 i 个样本的第 x 个评价维度得分,$i=1,2,\cdots\cdots,20$。每个样本生成 4 个问题,每个问题占 $score_x$ 的 0.25 分,评分结果如表 3 所示。

$$score_{overall}=\sum_{i=1}^{20}(score_{i1}+score_{i2}+score_{i3}+score_{i4}+score_{i5}) \quad (5)$$

多样性:$score_1$ 生成的问题是否高度相似,每重复一个问题扣 0.25 分;$score_2$ 生成的问题对应答案是否相同,每有一个重复答案或找不到答案,扣 0.25 分;

准确性:$score_3$ 问题能否从文档中找到答案,每有一个找不到答案,扣 0.25 分;$score_4$ 问题内容是否通顺,每有一个问题不通顺,扣 0.25 分;$score_5$ 问题内容是否有误导倾向,每有一个有误导倾向,扣 0.25 分。

表 3　问题生成评估

Table 3　Question generation evaluation

性能对比	本研究模型	ChatGLM2-6B
多样性($score_1$、$score_2$)	29.5	24.5
准确性($score_3$、$score_4$、$score_5$)	53.75	49.25
总计($score_{overall}$)	83.25	73.75

由以上结果可知,本研究模型在中医问题生成的多样性和准确性上均具有良好表现。通过两组实验结果可知,引入中医药知识库和经过 P-Tuning v2 微调的模型在该领域有着较好的表现,既能够从问题生成答案又能够从答案生成问题,体现出本研究模型可以在问题和答案之间进行双向转换的可逆能力,模型表现出的语义理解、信息抽取和逻辑推理能力具有一致性。在中医领域的应用相比于通用模型不仅可以根据用户提问生成可靠准确的回答,也能生成多样性和针对性的提问,对不明确自己提问意图的用户提供更多的问题选择,以更好地满足用户需求。

2.3　模型问答示例

本研究进一步对模型生成内容进行定性讨论,挑选最具代表性的问题以及特殊示例进行实例分析。明显错误以及重复部分标记为点线,与标准答案相符部分为黑色加粗,下划线为模型推理部分,如表 4 所示。

由表 4 可见:本研究模型生成的结果更符合人类语言表达习惯,内容更严谨准确,能根据其理解自行增添或总结内容,由下划线部分可知本研究模型推理修改内容合理且具有条理。从问题生成的角度来看,本研究模型生成的问题更具有多样性,符合用户表达意图,引导用户高效提问。综上,本研究模型能够根据用户的提问对回答内容酌情修改并生成多样化的中医问题进行引导,与传统直接返回设定答案的方式相比,有效提升了问诊服务的智能性、可交互性。

表 4　实例分析表

Table 4　Example analysis

问题示例	本研究模型	ChatGLM2-6B	参考答案
当归补血汤组成	根据已知信息,当归补血汤的组成为**黄芪一两(30 g)和当归酒洗,二钱(6 g)**。黄芪味甘补气,当归味厚,为阴中之阴,故能养血。两药相合,补血气自生,故名为补血汤。	当归补血汤是一种传统的中药方剂,主要用于治疗血虚血弱、面色萎黄、眩晕耳鸣、心悸失眠等症状。它的组成包括以下药材:当归 2 钱(6 g),熟地黄 2 钱(6 g),白芍 2 钱(6 g),川芎 1 钱(3 g),桃仁 1 钱(3 g),红花 1 钱(3 g),当归 2 片(4 g)。将以上药材煎煮后,加入适量的水,煮沸后改用文火煎煮 30 min,即可得到当归补血汤。	黄芪一两(30 g) 当归酒洗,二钱(6 g)
伤寒中风,有柴胡证,但见一证便是,不必悉具	根据已知信息,可以判断出该问题涉及中医诊断和治疗。柴胡证是指**口苦、咽干、目眩、往来寒热、胸胁苦满、默默不欲饮食、心烦喜呕等症状**,而小柴胡汤是一种常用的解表发汗方剂,可以用于治疗伤寒、中风等疾病。但是,该问题中提到了“伤寒中风,有柴胡证,但见一证便是,不必悉具”,说明在某些情况下,只需要根据一种症状就可以应用小柴胡汤进行治疗,而不必追求全部症状的齐全。	根据《伤寒论》和《金匮要略》的相关内容,伤寒中风有柴胡证、浮紧脉、不汗出、身疼痛、脉弦等症状时,不用悉具,只需要证候符合其中一项即可。因此,伤寒中风,有柴胡证,但见一证便是,不必悉具的答案是正确的。	“伤寒中风”,即不论伤寒还是中风。“有柴胡证”,是指口苦、咽干、目眩、往来寒热、胸胁苦满、默默不欲饮食、心烦喜呕诸症。“但见一证便是,不必悉具”,是言临床凡见到柴胡证的一部分主症,只要能反映少阳病枢机不利,胆火上炎的病机特点,确认为少阳病,即可应用和解之法,投以小柴胡汤,而不必待其主症全部具备再行其方。
“有一分恶寒,便有一分表证”,便可以生成一条逻辑规则,有“恶寒”情况下,是否有“表证”	根据已知信息,可以得出结论:若出现发热恶寒、头身疼痛、无汗、脉浮紧等症状,则必为太阳表证。但是,如果同时出现下利症状,则说明病涉阳明胃肠,因此太阳与阳明合病。	根据所提供的问题,我们可以生成一条逻辑规则:“有一分恶寒,便有一分表证”。这条规则的意思是,当出现恶寒的情况时,就会有一分表证。	恶寒为贯穿太阳病始终的一个主症,前人有“有一分恶寒,即有一分表证”之说,虽非绝对,但道出了恶寒在太阳病中的重要地位。
容颜早衰、腰膝酸软、失眠健忘,但有些食物对女人的肾伤害很大的。请生成多种针对性问题。	哪些食物对女人的肾伤害很大? 喝饮料对人体有什么影响? 长期食用咸菜会对人体有什么影响? 过量使用喹诺酮类药物和氨基糖苷类抗生素会对肾脏造成哪些伤害?	哪些食物对女人的肾伤害很大? 哪些食物容易导致容颜早衰、腰膝酸软和失眠健忘? 哪些食物对女性的肾脏健康不利? 哪些食物容易引起肾脏疾病?	女人肾虚会引起什么? 长时间饮料对人的坏处是什么? 长时间吃咸菜对人的坏处是什么? 饮用过量的浓茶对人的坏处是什么?

3　讨论

本研究针对中医药问答场景,提出中医药知识库融合中医问题生成的方法,将 RAG 技术与微调后的 ChatGLM2-6B 模型相结合进行智能问答,通过精准率、召回率、*F*1 值以及问题生成多样性的解析等定量和定性的共同对比分析,结果显示:①针对知识问答和问题生成的下游任务,本研究模型在中医自动问诊领域的适用性更优;②基于 RAG 技术大模型“幻觉”优化策略可以显著提升模型的可靠性与准确性;③通过从问题到答案和从答案到问题的双向评估验证,表明本研究模型在理解语义和逻辑推理能力的一致性。本研究模型能够理解问题的语义并提供有依据的答案,也能够根据给定的答案理解其语境生成相关问题,引导用户提问,挖掘用户个性化需求,优化模型的问答效果。针对大语言模型普遍存在“幻觉”现象和专业领域精度有限等问题,今后团队将更进一步探索模型的反馈优化方法并实现用户的友好交互和知识库的扩充,提升系统人性化和鲁棒性,为中医临床辅助诊疗、中医药传承与教育研究领域提供新思路和新方法。

参考文献：

[1] 文森，钱力，胡懋地，等. 基于大语言模型的问答技术研究进展综述[J]. 数据分析与知识发现，2024，8(6)：16-29.
WEN S，QIAN L，HU M D，et al. Review of research progress on question-answering techniques based on large language models [J]. Data Anal Knowl Discov，2024，8(6)：16-29.

[2] 王润周，张新生. 基于混合动态掩码与多策略融合的医疗知识图谱问答[J]. 计算机科学与探索，2024，18(10)：2770-2786.
WANG R Z，ZHANG X S. Medical knowledge graph question answering based on hybrid dynamic masking and multi-strategy fusion [J]. Comput Sci Exp，2024，18(10)：2770-2786.

[3] 田迎，单娅辉，王时绘. 基于知识图谱的抑郁症自动问答系统研究[J]. 湖北大学学报(自然科学版)，2020，42(5)：587-591，596.
TIAN Y，SHAN Y H，WANG S H. The research of depression automatic question answering system based on knowledge graph[J]. J Hubei Univ Nat Sci，2020，42(5)：587-591，596.

[4] 李启渊，张静，徐权光，等. ChatGPT 在中医医院智慧化建设中的应用、挑战及对策[J]. 卫生软科学，2024，38(4)：78-81.
LI Q Y，ZHANG J，XU Q G，et al. Application，challenges and countermeasures of ChatGPT in the construction of smart TCM hospitals[J]. Soft Sci Health，2024，38(4)：78-81.

[5] WEI S B，PENG X P，WANG Y F，et al. BianCang：A traditional Chinese medicine large language model[J]. arXiv preprint arXiv：2411.11027，2024.

[6] WANG H C，LIU C，XI N W，et al. HuaTuo：Tuning LLaMA model with Chinese medical knowledge[J]. arXiv preprint arXiv：2304.06975，2023.

[7] 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 [J]. Cureus，2023，15(6)：e40895.

[8] GILBERT S，KATHER J N，HOGAN A. Augmented non-hallucinating large language models as medical information curators [J]. NPJ Digit Med，2024，7(1)：100.

[9] LEWIS P，PEREZ E，PIKTUS A，et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. Adv Neural Inf Proc Syst，2020，33：9459-9474.

[10] LIU X，JI K X，FU Y C，et al. P-tuning v2：Prompt tuning can be comparable to fine-tuning universally across scales and tasks [J]. arXiv preprint arXiv：2110.07602，2021.

[11] DU Z X，QIAN Y J，LIU X，et al. GLM：General language model pretraining with autoregressive blank infilling[J]. arXiv preprint arXiv，2021，2103.10360.

[12] ZENG A H，LIU X，DU Z X，et al. GLM-130B：An open bilingual pre-trained model [J]. arXiv preprint arXiv，2022，2210.02414.

[13] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[J]. Adv Neural Inf Proc Syst，2017，30.

[14] JOHNSON J，DOUZE M，JEGOU H. Billion-scale similarity search with GPUs[J]. IEEE Trans Big Data，2021，7(3)：535-547.

[15] 王翼虎，白海燕，孟旭阳. 大语言模型在图书馆参考咨询服务中的智能化实践探索[J]. 情报理论与实践，2023，46(8)：96-103.
WANG Y H，BAI H Y，MENG X Y. Exploration of intelligent practice of large language models in library reference and consultation services [J]. Inf Sci Theor Appl，2023，46(8)：96-103.

[16] 张君冬，杨松桦，刘江峰，等. AIGC 赋能中医古籍活化：Huang-Di 大模型的构建[J]. 图书馆论坛，2024，44(10)：103-112.
ZHANG J D，YANG S H，LIU J F，et al. AIGC empowering the revitalization of ancient books on traditional Chinese medicine：Building the Huang-Di large language model[J]. Libr Tribune，2024，44(10)：103-112.

[17] ZHANG T Y，KISHORE V，WU F，et al. BERTScore：Evaluating text generation with BERT[J]. arXiv preprint arXiv：1904.09675，2019.

(编辑：董盈妹)