融合监督微调和检索增强的中医知识问答模型研究*

王欣宇[1]，杨 涛[1,2,3**]，王 松[4]，徐忆初[1]，胡孔法[1,5,6**]

（1. 南京中医药大学人工智能与信息技术学院 南京 210023；2. 南京大学信息管理学院 南京 210023；3. 江苏省中医流派研究院 南京 210023；4. 南京中医药大学卫生经济管理学院 南京 210023；5. 江苏省智慧中医药健康服务工程研究中心 南京 210023；6. 南京中医药大学唐仲英中医疫病研究中心 南京 210023）

摘 要：目的 充分利用中医问答和文献典籍等中医数据，构建推理能力强、结果可靠的中医知识问答任务模型。方法 收集和整理大规模中医语料问答数据，以ChatGLM3为基座，利用PissA方法进行监督微调，结合检索增强生成（Retrieval-augmented generation，RAG）方法，建立融合监督微调和检索增强的中医知识问答模型。将其与ChatGLM3、SFT、RAG等进行比较，从BLEU、ROUGE1、*F*值等经典指标角度评价模型效果。结果 本文建立的模型在BLEU值和ROUGE1值上分别为14.5830和34.6730，结合RAG后在中医题库推理结果中*F*值达到0.6398，优于ChatGLM3基线模型0.2654。结论 融合监督微调和检索增强的中医垂直领域大模型构建方法可有效提升模型的中医推理性能和可靠性。

关键词：监督微调 检索增强生成 大语言模型 中医知识问答

DOI: 10.11842/wst.20241121002 CSTR: 32150.14.wst.20241121002 中图分类号：R2-03 文献标识码：A

中医药凝聚了中华民族大医先贤的智慧结晶和经验知识，在各类疾病的预防和治疗过程中发挥着重要作用。高效的中医知识获取是中医学习和知识传播的重要需求。随着信息技术的发展，中医数字化和智能化程度不断提升，中医知识智能问答是帮助人们便捷获取中医药知识的有效方式。在中医知识的传承和发展中，中医文献典籍作为不可或缺的核心资源，为中医临床实践提供了重要支撑[1]。然而，中医文献典籍数量众多且内涵丰富，语句表述形式复杂多样，难以直接统一使用和整理。同时，传统的人工查阅和整理效率低下，对从业者的中医素养有较高的要求，难以满足现代中医临床、教学和科研需求，这给中医知识问答系统的构建带来了严峻挑战。

当前，大语言模型在通用领域已取得重要进展，并成功应用于法律[2]、教育[3]和医疗[4]等垂直领域中。通过针对性的领域数据训练，垂直模型在较小的模型参数体量下可以取得与较大规模模型相媲美的结果。虽然通用的大语言模型在自然语言处理上具有优势，但在中医知识问答这一垂直领域应用中仍然存在可解释性低、生成结果不稳定、缺乏确切的推理依据等问题[5]。针对上述问题，本文提出将大模型监督微调和外部知识检索增强[6]相结合，构建全新的中医知识问答模型，以期提高在中医垂直领域知识问答的可解释性和可信度。将大语言模型技术应用于中医知识问答任务，有益于中医文献典籍的传承和应用，对中医知识的广泛传播和中医药产业的现代化发展具有一定意义。

收稿日期：2024-11-21

修回日期：2025-02-16

* 国家自然科学基金委员会面上项目（82174276）：知识和数据协同驱动的中医藏象智能辨证方法研究——以心系疾病为例，负责人：杨涛；江苏省中医流派研究院开放课题（JSZYLP2024060）：基于大模型技术的江苏中医流派知识挖掘和服务创新方法学研究，负责人：杨涛。

** 通讯作者：杨涛（ORCID：0000-0002-9537-0500），博士，副教授，主要研究方向：中医药信息；胡孔法（ORCID：0000-0001-7670-3501），教授，博士研究生导师，主要研究方向：中医药人工智能与大数据分析。

1 研究现状

通用领域的开源大语言模型的训练语料中包含着一定的医学领域数据，在一定程度上具备了医学知识和理解能力，例如GPT-4模型[7]在医学领域文本生成任务上已取得较优表现。然而，由于大模型在处理中医领域相关任务时往往缺乏专业知识，无法胜任于复杂且个性化的中医问答任务，故需要充分利用中医领域数据进行模型的构建。

随着大语言模型技术的发展，医疗领域的大模型构建已取得了一定的进展。例如，基于开源的LLaMa架构[8]，研究人员通过10万条来自在线医疗网站的真实医患对话数据训练了ChatDoctor模型。该模型采用独特的信息检索机制，能够访问维基百科和专业医疗数据库，从而提升其在理解患者需求和提供诊疗建议时的准确性[9]。在中医领域，HuatuoGPT模型[10]通过结合来自ChatGPT的蒸馏数据和真实临床数据进行微调，并构建奖励模型，采用基于人工智能反馈的强化学习（Reinforcement learning from AI feedback，RLAIF）方法，使模型能综合学习多种来源的信息。结果表明，在医疗基准数据集上，HuatuoGPT在GPT-4评估和人工评估中均表现优异。在此基础上，HuatuoGPT-2模型[11]将医疗领域的预训练数据与微调整合到统一的训练阶段，并采用专门的采样技术实现领域适应，成为国内首个通过多项医疗资格考试并公开的大模型。ShenNong-TCM模型[12]采用基于实体的自指令方法，基于开源中医药知识图谱，利用ChatGPT生成11万多条中医药指令数据进行训练，从而具备了一定的中医问答能力。然而，由于中医领域数据规模庞大且复杂的特性，相关模型在中医领域仍可能存在着可信度和可解释性低的问题，缺乏推理依据。此外，监督训练数据的来源和知识注入方式相对单一，难以满足高智能和高可信度的中医基础知识问答任务。因此，本文通过建立融合监督微调与检索增强的中医知识问答模型，以提升模型对中医基础知识的理解和识别能力，同时增强回答的依据性和可信度。

2 中医知识问答模型构建方法

本文模型构建方法主要包括两个部分：其一，收集和整理大规模中医问答语料，利用PissA方法[13]进行监督微调；其二，通过检索增强生成（Retrieval-Augmented Generation，RAG）方法，以确定性知识库为基础，保证模型生成可信回复。模型结构如图1所示。

2.1 数据收集

高质量的领域监督数据是构建中医问答大语言模型的重要基础。考虑到中医语料数据的复杂性和应用场景的广泛性，本文通过融合多源中医领域数据，充分发挥各类数据的优势，帮助模型在具有充分推理依据和高度专业知识的同时也能拥有符合用户对话习惯的友好特征。本文模型构建所采用数据包括监督微调和检索增强生成数据。

2.1.1 监督微调数据

2.1.1.1 现实医患对话

数据来源于真实医疗网站的中西医医患对话记录，保留了患者问询以及医生专业回复的原始特征，

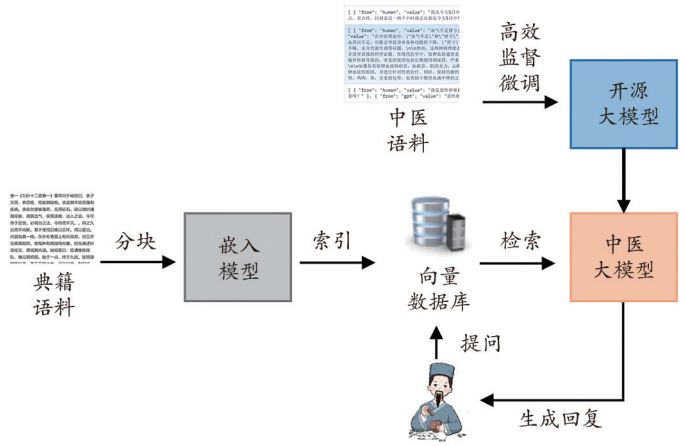

图1 模型结构图

Fig. 1 Model structure diagram

反映医疗咨询的实际情况，贴合模型的实际使用场景。为避免过度口语化所造成的影响，使用ChatGPT对原始内容进行增强和细化，构建了一个高质量的真实会话数据集，共计177703条问答对。

2.1.1.2 ChatGPT蒸馏对话

蒸馏对话来源于两个分别设定了角色指令的ChatGPT，以医疗诊断数据库作为医疗知识来源，根据患者的情况以及最终的诊断生成符合真实诊疗场景的对话数据，构建以数据库中的诊断知识为基础的专业化数据集，共计112565条问答对。

2.1.1.3 中医药知识图谱

以完备的中医药知识图谱为依托，利用ChatGPT生成以实体和关系为中心的问答数据集，涵盖知识图谱中丰富的症状、方剂和诊断信息，使得模型具备一定的中医诊疗能力并理解相应方剂、药物等细节信息，共计22215条问答对。

2.1.1.4 中医执业医师资格考试题库

数据集包含中医执业医师资格考试题库中的题干、选项、答案和解析，包含单项和多项选择题。构建问答对，问句中包含统一的提示指令、题干、选项；答句中包含问题答案以及相关的原因解释，共计54,498条问答对。

2.1.2 检索增强生成数据

检索增强采用了《伤寒杂病论》《本草纲目》《慈幼新书》等在内涵盖古代不同时期的中医文献典籍373本，通过光学字符识别（Optical Character Recognition，OCR）文字识别技术进行字符的识别并保存。为避免识别过程中产生的错误以及古汉语中晦涩的表述对模型的检索和推理产生不良影响，本文将文献典籍进行识别后，调用ChatGPT模型对其进行纠错和现代汉语的翻译，生成易于大模型理解的文献典籍语料，以此提升模型检索增强生成的成效。

2.2 监督微调

在第一阶段的监督微调过程中，对监督微调数据集进行预处理，保证每个数据单元皆以问答对的方式进行保存，包括指令、输入和相应输出。在此基础上，采取PissA技术对ChatGLM3模型进行监督训练。PissA微调的结构如图2所示。

PissA是针对大语言模型的高效微调技术，采用适配器（Adapter）的概念，在冻结模型原始参数的基础上，向模型的特定部分插入可供训练的低秩矩阵，以减少模型微调时需要训练的参数数量。同时在模型训练的过程中更完整保证模型的文本理解和分析性能，以此来保证模型后续对文献典籍进行检索增强生成时的文本分析能力。传统的LoRA[14]采用高斯噪声来初始化矩阵A，0来初始化矩阵B，没有利用到原模型的参数，而PissA对原模型参数矩阵$W \in R^{m\times n}$采用奇异值分解（Singular value decomposition，SVD），将其中前r个奇异值和奇异向量用以初始化Adapter的两个矩阵$A \in R^{m\times r}$和$B \in R^{r\times n}$，剩余的奇异值和奇异向量用以构造残差矩阵$W^{res} \in R^{m\times n}$，使$W = AB + W^{res}$。这种初始化方式使得适配器中的参数包含了原模型的核心参数，可以更好地捕捉模型的关键特征，在微调开始阶段就能更完整地复制原始模型微调的效果，并且通过微调参数量较小的核心适配器A、B，冻结参数量较大的残差矩阵W^{res}，实现了用很少的参数获取近似全参数微调的学习效果。

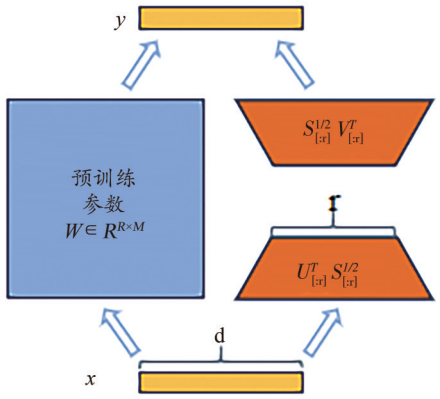

图2 PissA微调结构图

Fig. 2 PissA fine-tuning structure diagram

2.3 检索增强生成

经过领域数据微调后的大模型，仍会产生幻觉、知识过时、推理过程和依据不可追溯、可信度低以及中医术语不足等问题，需要将大模型内部的知识与外部的知识库进行整合。对于中医领域的大语言模型，文献典籍是中医知识的重要来源之一，可以通过检索增强生成的方式为大模型提供稳定、可信、准确的知识资源。检索增强生成的主要流程如图3所示。

2.3.1 文档索引

将翻译后的中医文献典籍使用TextSplitter根据符号分割为约500字符的文本块，为保持文本块的连贯性，文本块之间保留前后约50字符的重复。在此基础上构建文本块的索引，采用bge-large-zh向量化模型[15]

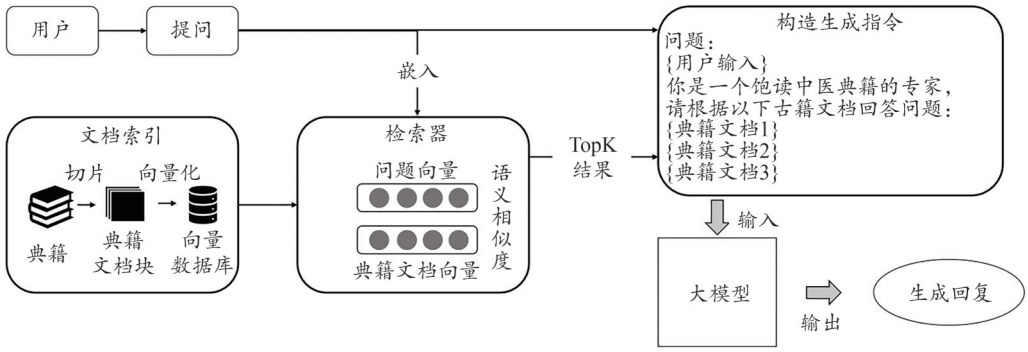

图3　检索增强生成流程

Fig. 3　Search enhanced generation process

捕捉语义信息，为每个文本块生成高维的向量表示（embedding），并将生成的向量表示存储到Faiss数据库中，同时存储相关的索引信息。其中，bge-large-zh模型特别针对中文文本进行了优化，能够捕捉文本中的语义信息，使得语义上相似的文本在向量空间中的距离更近，为后续的相似性搜索提供便利，而Faiss是一个高效的相似性搜索和稠密向量聚类的库，其核心原理包括倒排索引和乘积量化，能够大幅度减少存储和计算成本，同时保持向量间距离近似不变。

2.3.2　检索和重排序

当用户进行提问时，根据用户输入的问句，调用Faiss数据库采用向量相似度计算的方式进行相关性搜索，检索出与查询语句最相关的K个文本块。

$$Top\,k(candidates) = argmax_k similarity(query, candidate) \quad 式(1)$$

$$similarity(ab) = \frac{a \cdot b}{\|a\| \|b\|} \quad 式(2)$$

其中，candidate为候选向量，query为查询向量，similarity指其向量余弦相似度的计算方式。

为提升对于复杂查询的响应质量，使用bge-reranker-large模型对检索到的语句进行重排序，即根据用户提出的问句和检索到的文本得到的相似性评分（similarity scores）选择最相关的文本块，以供后续的增强生成阶段使用。

2.3.3　增强生成

在检索和重排序阶段得到的最相关文本块的基础上，设计提示指令，调用经过微调的大模型回答内容，使得模型生成不仅依赖于监督微调注入的知识，也结合了中医文献典籍中的信息，以生成更为准确、丰富和可信的输出结果。

3　实验

3.1　实验设置

本文中医知识问答模型建立在新一代对话预训练模型ChatGLM3-6b[16]之上，ChatGLM3-6b采用了更丰富多样的训练数据、更充分的训练步数和更合理的训练策略，在多类领域的数据上取得了优越的评测效果。本实验的硬件配置为：i7 13700KF和GPU RTX4090（24G）。软件配置为：python3.11.0、pytorch1.13.1和transformers4.23.0。采用PissA方法进行微调，学习率为0.00005，并采取Cosine策略进行梯度优化，为了节约训练成本，本文采用混合的fp16-fp32精度和梯度累积策略来提升训练效率。训练的epoch为1，按照9∶1的比例划分训练集和测试集。同时，为综合评估模型生成的效率，本文选取BLEU[17]、ROUGE[18]、F、P、R等指标进行文本生成评估。

3.1.1　BLEU

BLEU是目前业界公认的文本生成模型评价指标，偏向于评估文本生成的精确率，其实质是计算模型生成句与原句的相似度。首先，统计两者同时出现n-gram的次数，并取其中较小值作为最终匹配个数，再除以文本的总n-gram数，从而得到其n-gram的精度得分P_n。

$$Count_{\text{clip}} = \min\left(Count, Max_{Ref_{Count}}\right) \quad 式(3)$$

$$p_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{n-\text{gram} \in C} \text{Count}_{\text{clip}}(n-\text{gram})}{\sum_{C'} \in \{\text{Candidates}\} \sum_{n-\text{gram}' \in C'} \text{Count}_{\text{clip}}(\text{n}-\text{gram}')} \quad 式(4)$$

其中，Count为n元词在生成结果中出现的次数，$Max_{Ref_{Count}}$为参考文档中n元词的最大出现个数。在此基础上，对P_n求对数的算术平均数并加入长度惩罚因

子BP，得到其评价BLEU值。

$$BP = \begin{cases} 1, & \text{if} \quad c > r \\ e^{\left(1-\frac{r}{c}\right)} & \text{if} \quad c \leqslant r \end{cases} \qquad 式(5)$$

$$BLEU = BP*exp\left[\left(\sum_{n=1}^{N} w_n logp_n\right)\right] \qquad 式(6)$$

$$W_n = 1/n \qquad 式(7)$$

其中，c为机器生成文档的实际长度，r为参考文档的长度。

3.1.2 ROUGE

ROUGE是计算模型生产句与原句的相似度，但更偏向于评估模型的召回率，本章使用ROUGE1和ROUGE2指标对模型的生成性能进行评估。

$$ROUGE - N = \frac{\sum_{S \in \{ReferencesSentences\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferencesSentences\}} \sum_{gram_n \in S} Count(gram_n)} \qquad 式(8)$$

3.1.3 精确率P(precision)、召回率R(recall)和F值

$$P = \frac{TP}{TP + FP} \qquad 式(9)$$

$$R = \frac{TP}{TP + FN} \qquad 式(10)$$

$$F = \frac{2*P*R}{P + R} \qquad 式(11)$$

其中，TP(True Positives)指被正确分类为正样本的正样本数量；FP (False Positives)指被错误分类为正样本的负样本数量；FN(False Negatives)指被错误分类为负样本的正样本数量。

3.2 监督微调方法比较

为充分分析微调方式对模型监督学习性能造成的影响，本文针对整理的监督微调数据集采取LoRA和PissA两种方式进行训练，其训练结果对比见表1。

由表1可知，PissA因其使用主奇异值和奇异向量来初始化适配器，在训练时有更快的收敛速度，且在训练末期损失更低，最终减少了0.0264的损失。在测试集上，PissA微调的模型在BLEU-4、ROUGE-1、ROUGE-2和ROUGE-L指标上分别有0.0885、0.0593、0.0438和0.0885的提升，说明模型可以更好地学习到训练数据中蕴含的知识，取得更优的微调效果。

3.3 模型推理结果比较

为充分比较知识注入方式对模型推理性能的影响，本文随机选取了未在训练数据集中出现的100道中医药有关选择题，分别对微调后的模型(Supervised fine-tuning，SFT)、RAG、SFT+RAG以及Huatuo-2、GPT-4o和ChatGLM3的推理结果从P、R和F三个指标进行对比，其结果如表2所示。

表1 监督微调结果对比

Table 1 Comparison of supervision and fine adjustment results

微调方法	train_loss	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
LoRA	1.4120	14.4945	34.6137	13.7369	27.5037
PissA	1.3856	14.5830	34.6730	13.7807	27.5922

表2 推理结果对比

Table 2 Comparison of reasoning results

模型	P/(%)	R/(%)	F/(%)
Huatuo-2	47.00	49.00	47.98
GPT-4o	58.00	58.00	58.00
ChatGLM3	36.00	39.00	37.44
SFT	50.00	51.00	50.50
RAG	36.00	36.00	36.00
SFT+RAG	63.00	65.00	63.98

注：表中SFT指在ChatGLM3模型基础上的监督微调模型；RAG指在ChatGLM3基础上的检索增强生成模型；SFT+RAG指在ChatGLM3模型基础上经过监督微调后再进行检索增强生成的模型。

由表2可知，在通用领域大模型中，GPT-4o具备较强的中医推理性能，超越了经过中医领域数据微调的Huatuo-2和本文SFT得到的模型，而在经过SFT及RAG后，模型推理效果得到了大幅提升，不但超越了SFT模型的推理结果，更超越了GPT-4o的中医知识推理能力，充分证明SFT方法提升了模型对中医领域文献数据的理解和分析能力，体现了SFT和RAG技术在中医领域协同增强的有效性。

3.4 个案分析

在该部分的研究中，本文尝试用一些有关中医知识常用的提问，分别调用微调后的模型(SFT)以及在此基础上采取对中医文献典籍进行RAG的模型(SFT+RAG)进行案例分析，其结果如表3所示。

从表3可以看出，由于有稳定来源的文献典籍作为知识依据，SFT+RAG模型对于方剂的组成、功效、用法等问题的回复具有更高的准确性和可信度；对于患者症状的建议，SFT+RAG模型可以提供具体的中药治疗方案(例如六味地黄汤和安润汤)，以及详细的组成和功效，还提供了饮食建议，更具有针对性。而仅SFT模型更强调就医和专业指导，更具谨慎性但不完全符合中医知识的问答场景；由于微调的数据以及大模型的预训练数据中参杂的西医语料，仅SFT模型有时也会回复倾向于西医的答案(例如西药和手术治疗)，而

表3 案例分析表

Table 3 Table of case analysis

Query	SFT+RAG	SFT
介绍一下三生饮的组成、功效和用法	三生饮的组成、功效和用法如下：三生饮的组成包括南星、川乌、附子、木香。三生饮的功效主要是治疗中风突然昏厥、神志不清、痰涎堵塞、言语不畅等病症。三生饮的用法是每次服用五钱，用姜和水煎煮，加入一两的人参。	三生饮是一种传统中药方剂，其组成主要包括黄芪、白术、茯苓、甘草等中药材。 三生饮的功效主要包括益气健脾、补中益肾、养血生发。它可以改善脾胃虚弱、食欲不振、腹胀便溏等症状，还可以增强体质，改善头发质量。 三生饮的用法为口服，每次服用9 g，每日2次，可加入适量的清水或米汤一起煮服。建议在医生的指导下使用。祝您健康！
口干咯血怎么办？应该吃什么中药？	根据你提供的信息，口干咯血可能是由于肾阴虚导致的。建议服用六味地黄汤，同时可以配合安润汤。六味地黄汤中的熟地、山药、丹皮、泽泻、茯苓、麦冬等药材可以滋补肾阴，安润汤中的当归、白芍、熟地、川芎、麦冬、牛膝等药材可以滋阴润燥，两者配合使用可以缓解口干咯血的症状。同时，在饮食上应该多食用滋阴润燥的食物，如莲子、百合、银耳、枸杞子等，避免辛辣刺激性食物，保持心情愉悦，避免过度劳累等。如果症状严重，建议及时就医。	如果您出现口干咯血的症状，建议您立即就医，以确定病因并接受专业治疗。口干咯血可能是由于多种疾病引起的，例如肺部疾病、心脏疾病、血液病等，需要进行相关检查和治疗。如果您已经在医院接受治疗，请按照医生的建议进行治疗。中药治疗口干咯血需要根据个体情况和病因进行选择，建议您咨询专业医生或中医师，以获取更具体的指导和治疗方案。祝您早日康复！
中风偏瘫如何治疗？	对于中风偏瘫的治疗，需要针对病情进行个体化的治疗。一般来说，中风偏瘫的治疗可以分为以下几个方面： ①补充气血：中风偏瘫的治疗中，常需要补充气血，以促进肌肉和神经的恢复。可以口服一些补气血的药物，如地黄饮子、三生饮等。 ②舒筋化痰：中风偏瘫的治疗中，还需要注意舒筋化痰的方法，以促进肌肉的恢复和神经的康复。可以口服一些舒筋化痰的药物，如江子仁、独活汤等。 ③针灸治疗：针灸是一种传统的中医治疗方法，可以通过刺激穴位来促进血液循环和神经康复。针灸治疗中风偏瘫的效果较好，但需要在专业医师的指导下进行。 ④物理治疗：物理治疗是一种通过运动、热疗等方式来促进肌肉和神经的恢复的治疗方法。物理治疗可以帮助中风偏瘫患者恢复肌肉力量和协调性，提高生活质量。 总之，中风偏瘫的治疗需要综合考虑患者的具体情况，采用个体化的治疗方法。同时，患者还需要注意饮食和生活习惯的调整，保持良好的心态和积极的生活态度。	对于中风偏瘫的治疗，您可以考虑以下几种方法： ①药物治疗：中风偏瘫的治疗通常需要使用抗血小板药物、抗凝药物等药物来控制病情、稳定血压和预防血栓形成。 ②物理治疗：中风偏瘫的治疗通常需要进行物理治疗，包括康复训练和康复运动，以帮助恢复肌肉功能和活动能力。 ③针灸治疗：针灸治疗可以缓解疼痛和改善血液循环，从而促进中风偏瘫的康复。 ④手术治疗：对于严重的中风偏瘫，可能需要进行手术治疗，如脊椎固定或神经重建手术。 请注意，中风偏瘫的治疗因个体差异而异，建议在医生的指导下进行治疗。祝您早日康复！

注：表中SFT指在ChatGLM3模型基础上的监督微调模型；SFT+RAG指在ChatGLM3模型基础上经过监督微调后再进行检索增强生成的模型。

在此基础上的SFT+RAG模型则完全偏向于中医治疗方法，适合于对于中医知识的获取。

4 总结

本研究提出了结合监督微调和检索增强生成的两阶段大模型知识注入方法，充分利用中医领域知识问答语料和文献典籍数据，大幅提升了模型在中医问题上的推理能力和可信度，使模型具有中医问答的专业知识和能力。在未来的研究中，将进一步引入包含临床医案的多模态数据[19]，从数据和知识双轮驱动[20]的角度进行中医辨证论治大模型的构建和分析，以实现向用户提供更精准、更可信的中医诊断和治疗建议，确保模型能够应对更复杂的中医临床场景。

[利益冲突]本文不存在任何利益冲突。

参考文献

1 李轲, 李建生, 张瑞, 等. 基于数据挖掘中医古籍治疗肺热病遣方用药分析[J]. 中华中医药学刊, 2020, 38(8):82-85.
Li K, Li J S, Zhang R, et al. Analysis of prescriptions and medicines in treating lung heat disease in ancient books of Chinese medicine based on data mining[J]. Chinese Archives of Traditional Chinese Medicine, 2020, 38(8):82-85.

2 沈晨晨, 岳圣斌, 刘书隽, 等. 面向法律领域的大模型微调与应用[J]. 大数据, 2024, 10(5):12-27.
Shen C C, Yue S B, Liu S J, et al. Fine-tuning and application of large language model in law domain[J]. Big Data Research, 2024, 10(5): 12-27.

3 刘三女牙, 郝晓晗. 生成式人工智能助力教育创新的挑战与进路[J]. 清华大学教育研究, 2024, 45(3):1-12.
Liu S N Y, Hao X H. The challenges and approaches of AIGC in facilitating educational innovation[J]. Ysinghua Journal of Education, 2024, 45(3):1-12.

4 喻金龙, 张磊, 许宁, 等. ChatGPT在风湿科中医电子病历症状信息预处理中的应用[J]. 中华中医药学刊, 2025, 43(3):24-29,301.
Yu J L, Zhang L, Xu N, et al. Application of ChatGPT in symptom information preprocessing of electronic medical record of rheumatology [J]. Chinese Archives Traditional Chinese Medcine, 2025, 43(3): 24-29,301.

5 Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.

6 Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. Advances in Neural Information Processing Systems, 2020, 33:9459-9474.

7 Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.

8 Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language[J]. arXiv preprint arXiv:2302.13971, 2023.

9 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.

10 Zhang H B, Chen J Y, Jiang F, et al. HuatuoGPT, towards taming language model to be a doctor[J]. arXiv preprint arXiv:2305.15075, 2023.

11 Chen J Y, Wang X D, Ji K, et al. HuatuoGPT-II, one-stage training for medical adaption of LLMs[J]. arXiv preprint arXiv:2311.09774, 2023.

12 Dou Y T, Zhao X J, Zou H T, et al. ShennongGPT: A tuning Chinese LLM for medication guidance[C]. 2023 IEEE International Conference on Medical Artificial Intelligence (MedAI), Beijing, China, 2023: 67-72.

13 Meng F X, Wang Z H, Zhang M H. PiSSA: Principal singular values and singular vectors adaptation of large language models[J]. arXiv preprint arXiv:2404.02948, 2024.

14 Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.

15 Xiao S T, Liu Z, Zhang P T, et al. C-Pack: Packed resources for general Chinese Embeddings[C]. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington DC USA, 2024:641-649.

16 Team GLM. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools[J]. arXiv preprint arXiv:2406.12793, 2024.

17 Papineni K, Roukos S, Ward T, et al. Bleu: A method for automatic evaluation of machine translation[C]. Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, 2002:311-318.

18 Lin C Y. ROUGE: A package for automatic evaluation of summaries[C]. Proceedings of the Workshop on Text Summarization Branches Out, 2024.

19 和芳娟, 杨文媛, 于小刚, 等. 基于望目辨证理论和白睛成像AI技术探讨高胆固醇血症的目络特征及发病机制相关性研究[J]. 中华中医药学刊, 2024, 42(4):19-22.
He F J, Yang W Y, Yu X G, et al. Analysis on eye collaterals of hypercholesterolemia and the correlation of pathogenesis based on theory of syndrome differentiation according to inspection of eyes and white of the eye imaging AI technology[J]. Chinese Archives Traditional Chinese Medcine, 2024, 42(4):19-22.

20 杨涛, 漆隽之, 胡孔法, 等. 知识驱动的中医智能诊疗研究思路与方法[J]. 中华中医药学刊, 2024, 42(10):13-16.
Yang T, Qi J Z, Hu K F, et al. Reaeatch ideas and methods for knowledge-based intelligent diagnosis and treatment of Chinese-medicine[J]. Chinese Archives Traditional Chinese Medcine, 2024, 42 (10):13-16.

Research on a Traditional Chinese Medicine Knowledge Q&A Model Integrating Supervised Fine-Tuning and Retrieval-Augmented Generation

WANG Xinyu[1], YANG Tao[1,2,3], WANG Song[4], XU Yichu[1], HU Kongfa[1,5,6]

(1. School of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, Nanjing 210023, China; 2. School of Information Management, Nanjing University, Nanjing 210023, China;

3. Jiangsu Provincial Research Institute of Chinese Medicine Schools, Nanjing 210023, China; 4. School of Health Economics and Management, Nanjing University of Chinese Medicine, Nanjing 210023, China; 5. Jiangsu Provincial Engineering Research Centre for Intelligent Traditional Chinese Medicine Health Services, Nanjing 210023, China; 6. Tang Zhongying Traditional Chinese Medicine Epidemic Disease Research Centre at Nanjing University of Chinese Medicine, Nanjing 210023, China)

Abstract: Objective To construct a traditional Chinese medicine (TCM) knowledge question-answering model with strong reasoning capabilities and reliable results, TCM Q&A datasets and TCM literature were fully utilized. Methods Large-scale TCM corpus and Q&A data were collected and organized, with ChatGLM3 serving as the base model. The PissA method was used for supervised fine-tuning, combined with retrieval-augmented generation (RAG) techniques, to build a TCM knowledge Q&A model that integrates supervised fine-tuning and retrieval-augmented generation. The model was compared with ChatGLM3, SFT, and RAG, with evaluations based on classic metrics such as BLEU, ROUGE1, and *F*-scores. Results The model in this paper achieved BLEU and ROUGE1 scores of 14.5830 and 34.6730, respectively. After incorporating retrieval-augmented generation, the model attained an *F* score of 0.6398 in the inference results on a TCM dataset, outperforming the ChatGLM3 baseline model's 0.2654. Conclusion The construction method of a large model in the TCM domain that integrates supervised fine-tuning and retrieval augmentation can effectively enhance the model's reasoning performance and reliability in TCM.

Keywords: Supervised fine-tuning, Retrieval-augmented generation, Large language model, TCM knowledge question answering

（责任编辑：李青）