基于Lora微调的轻量化中医药古籍大语言模型研究*

柴景贤[1,2]，郎许锋[1,2]，李红岩[1,2]，周作建[1,2**]，凌　云[3]，战丽彬[4]，胡孔法[1,2]，乔学斌[2,5]

（1. 南京中医药大学人工智能与信息技术学院　南京　210023；2. 江苏省智慧中医药健康服务工程研究中心　南京　210023；3. 南京中医药大学中医学院　南京　210023；4. 辽宁中医药大学中医药创新工程技术中心　沈阳　110847；5. 南京中医药大学养老服务与管理学院　南京　210023）

摘　要：目的　针对中医古籍大语言模型构建难度大、微调成本高的问题，研究轻量化中医药古籍大语言模型微调方法，实现以历代《伤寒论》为核心的中医古籍知识问答模型。方法　数据集构造，设计提示词引导GPT-4生成《伤寒论》知识问答对，并融合ShenNong_TCM_Dataset与cMedQA2数据集；模型选择，选用5个通用大模型进行Lora微调，经评估选取最佳模型并验证多版本量化效果。结果　微调后的Qwen-7B-Chat的BLEU、ROUGE-1、ROUGE-2与ROUGE-L指标相较于基座模型分别提升了17.61、19.63、14.3与21.4。结论　本文所选模型能够有效理解和使用《伤寒论》等中医古籍专业术语和概念，针对用户问题给出准确答案，且相较于同类模型微调成本与算力要求更低，有助于中医药知识传播与智能化发展。

关键词：大语言模型　中医药古籍知识　《伤寒论》　Lora微调　模型轻量化

DOI: 10.11842/wst.20241014001　CSTR: 32150.14.wst.20241014001　中图分类号：R-058　文献标识码：A

中医学有独特的理论体系，其知识与实践经验的总结记载在中医经典古籍中。以历代《伤寒论》为例，其理法方药融为一体，确立了六经辨证体系，奠定了中医辨证论治的基础[1]。然而，此类古籍文本术语较多且表述不一致，因此理解难度较大。除此之外，非结构化语言使其难以自动分析，提高了中医药知识的理解和深入挖掘难度。

近些年来，大语言模型领域经历了迅猛的发展，被认为是未来通用人工智能的关键技术之一[2]。大规模参数模型如GPT-4[3]，凭借其强大的语言理解和生成能力，在自然语言处理、自动编程、智能问答等多个应用场景中表现出色。随着研究的深入，对这些大参数量模型进行垂直领域微调已成为主流趋势，例如Adapter Tuning[4]、P-Tuning[5]、Lora[6]等等。恰当的微调方法能够显著降低算力需求和微调成本，使得模型不仅能够保留强大的语言能力，还能在特定领域中展现更高的专业性和适应性。这一技术加速了人工智能在各个垂直行业的应用。医疗大模型微调最先以英文数据为主，Li等[7]收集医疗领域数据集进行全量微调构建ChatDoctor，经过海量医学事实信息微调过后的专业模型能够更好地理解和回答医学问题，提供准确的医疗信息。Wu等[8]通过整合480万篇生物医学论文和3万本医学教科书，并结合领域特定的指令微调构建了一个面向医学领域的开源大型语言模型PMC-LLaMA。谷歌和DeepMind合作在PaLM[9]的基础之上，使用指令微调方法训练获得医疗领域中表达近乎人

收稿日期：2024-10-14

修回日期：2024-11-18

*　国家自然科学基金委员会面上项目（82074580）：基于知识图谱的现代名老中医诊治肺癌用药规律及其机制研究，负责人：胡孔法；江苏省中医药科技发展计划项目合同（MS2023010）：面向《伤寒论》的中医药大语言模型构建研究及应用，负责人：周作建。

**　通讯作者：周作建（ORCID：0009-0005-4310-5683），研究员，硕士研究生导师，主要研究方向：智能医学、中医药人工智能与大数据分析研究。

类专家的模型Med-PaLM[10]，该模型能够理解临床语言、医学影像以及基因组学等等，知识面极其广泛。随着技术不断发展，中文医疗大模型也不断涌现，Xiong等[11]基于P-tuning V2[12]对ChatGLM-6B[13]进行中文医疗领域微调，初步实现了中文模拟问诊，随后又有学者将LLaMa[14]迁移至中文医疗知识领域，构建了BenTsao[15]大模型，也有研究人员结合多个开源中文医疗问答数据集根据就诊多轮问诊特征构建了BianQue[16]大模型。张君冬等[17]构造了知识引导的中医古籍对话数据集，训练获得优秀的Huang-Di大模型。然而由于上述模型庞大的参数量导致微调成本高昂，对知识领域迁移和应用落地造成一定困难，并且通用模型对文言文、半文言文理解能力较差，处理难度较大，为了解决此类问题，本文以历代《伤寒论》原文或译文以及其中所有方剂数据为例，构建一个基于Lora微调的轻量化中医药古籍大语言模型，该模型仅需较低的训练和部署成本，在对《伤寒论》等中医古籍知识的理解能力上，多项评估指标显示了一定的优势。本文的主要贡献包括：①根据中医古籍方剂、病种数据特征设计提示词调用生成式大模型GPT-4构建一个以历代《伤寒论》为核心的中医药古籍知识问答数据集；②使用中医药古籍知识数据集微调6-7B参数量的通用模型，在《伤寒论》等中医药古籍领域达到了高度的专业性。模型能够深入理解并准确运用中医经典中的复杂概念、术语和辨证体系，拥有优异的中医药专业问题解答能力；③使用低秩自适应微调方法Lora优化模型训练，在资源受限的研究环境中有效降低训练成本，提升模型的训练效率和表现，使其更好地理解和应用《伤寒论》等中医药古籍中的专业知识。不仅保证了模型的专业性，还大幅度减少了对计算资源的依赖。

1 基于Lora微调的轻量化中医药古籍大语言模型

基于历代《伤寒论》等中医古籍数据，设计提示词引导GPT-4生成有监督的高质量问答数据，并结合开源中医知识问答数据进行预处理构建中医古籍知识问答数据集。随后基于Lora在该数据集上微调训练多个通用大语言模型，并通过模型评估选出最适用于中医古籍领域的模型。最后针对该模型比对全量Lora、qLora-8bit、qLora-4bit的性能表现，综合考虑资源配置，构建轻量化中医药古籍大语言模型，如图1所示。

1.1 数据集构建

针对历代《伤寒论》等中医经典古籍以及其中方剂数据根据特征设计提示词引导生成式大模型GPT-4生成有监督的高质量问答数据，融合收集来的开源中医药知识问答数据，再对它们进行预处理，包含格式转化、数据清洗等等。最终构建一个以历代《伤寒论》为核心的中医药古籍知识问答数据集。数据集构建流程如图2所示。

1.2 模型选择

基于全量Lora微调方法在中医药古籍知识问答数据集上训练Baichuan-7B、Baichuan2-7B-Chat[18]、ChatGLM3-6B、Xverse-7B-Chat以及Qwen-7B-Chat[19]共5个初始模型，在通用模型评估基准上测试模型语言泛化能力，从中挑选出适用于古汉语领域任务的3个基座模型，对其进行《伤寒论》等中医药古籍专业知识预测并测试模型问答示例交由《伤寒论》专家评估。随后综合比对3个模型的通用评估指标、专业知识预测指标以及专家意见，选取一个最适用于中医药古籍领域的模型。

2.3 低秩自适应微调方法Lora及量化qLora

Lora微调方法通过低秩分解模拟参数的改变量，以极小的参数量来实现大模型的间接训练，实现原理如图3所示。

为了提高模型专业性能、节省算力资源，使用Lora来优化预训练模型参数，假设微调一个预训练模型，需要更新预训练参数为 $\mathcal{W}_0 + \Delta\mathcal{W}$，其中，$\mathcal{W}_0$ 是预训练模型的初始参数，$\Delta\mathcal{W}$ 是需要更新的参数，假设要全参数微调，即 $\Delta\mathcal{W} = \Delta\mathcal{W}$ 时，这样便会投入巨额的成本。而在Lora轻量级微调中只需要调整 $\Delta\mathcal{W}$，假设 $\Delta\mathcal{W} \in \mathbb{R}^{d\times k}$，参数的更新如公式1所示：

$$\mathcal{W}_0' + \Delta\mathcal{W} = \mathcal{W}_0' + BA, B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times k} \quad (1)$$

因此首先使用随机高斯分布初始化降维矩阵 A，用零矩阵初始化升维矩阵 B，确保在训练开始时微调的影响最小，从而避免破坏预训练模型的原有知识结构，接着冻结初始预训练模型的权重，只更新降维矩阵 A 与升维矩阵 B，公式(1)中 $r \ll \min(d,k)$，在Lora训练过程中，$\mathcal{W}_0$ 是固定不变的，只会更新 A 和 B 的参数，在前向过程中，$\mathcal{W}_0$ 与 $\Delta\mathcal{W}$ 都会乘以相同的输入x并相加，如公式2所示：

$$h = \mathcal{W}_0 x + \Delta W x = \mathcal{W}_0 x + BAx \quad (2)$$

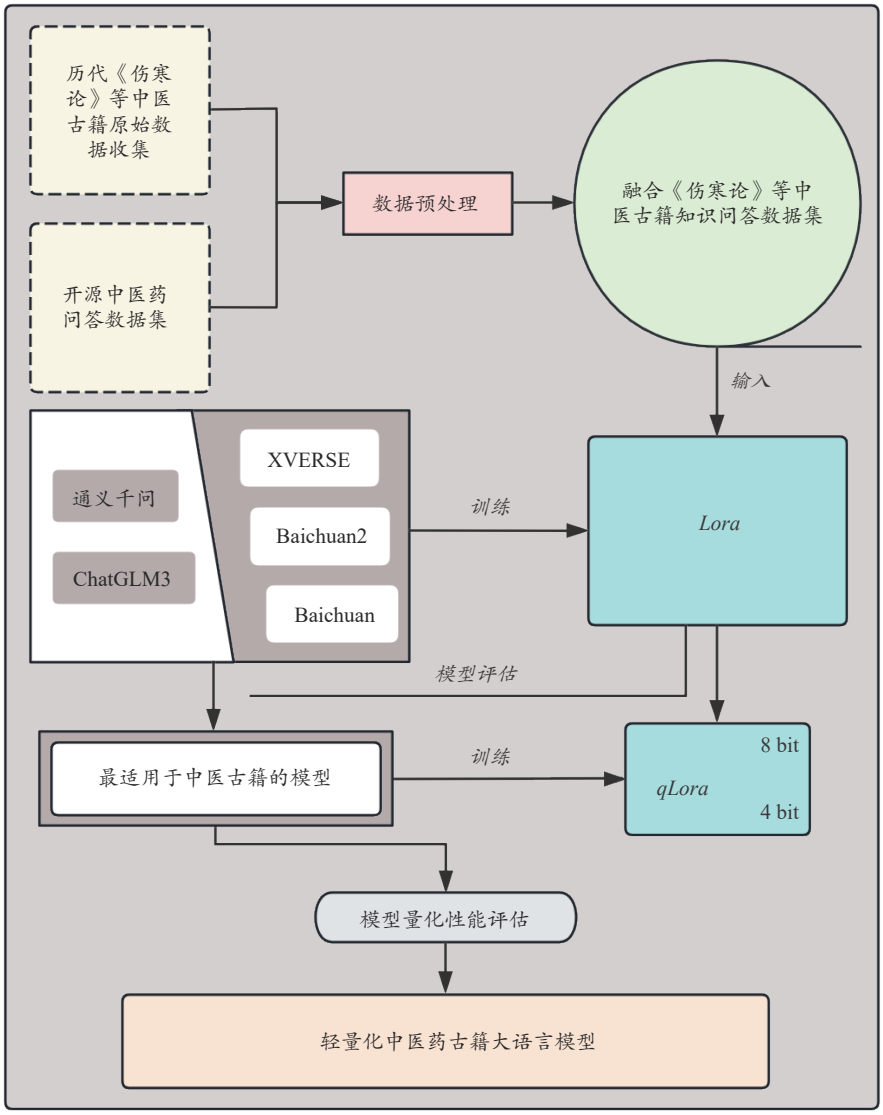

图1　轻量化中医药古籍大语言模型构建

最后将降维矩阵A和升维矩阵B注入Transformer模型的每一层，以达到以最少成本更新模型参数，构建全新的专业领域模型。在增量训练后，模型可能会遭遇基础对话能力的严重退化。因此引入了一部分通用对话数据进行训练，以确保模型在学习新知识的同时，保持其现有的通用对话能力。此外，这种方法还增强了模型的适应性和泛化能力，使其能够应对各种类型的对话需求。

Lora的一个潜在缺点是可能在微调过程中降低模型的准确性，丢失部分原始模型的高阶特征信息，从而影响模型性能。qLora则在此基础上进行了改进，通过引入高精度权重和可学习的低秩适配器，显著提升模型的准确性。其创新点在于，qLora首先将预训练模型进行量化，从而大幅减少内存占用，接着添加一组可学习的低秩适配器权重，这些适配器能够捕捉和保留原始权重的高阶特征信息。这样不仅降低了微调成本，还提高了模型在特定任务中的准确性，使其在历代《伤寒论》等中医古籍领域构建轻量化大语言模型时表现优异，兼顾性能与资源效率。

2　实验与分析

2.1　数据收集及预处理

数据内容包含：①名家医案数据：收集中医专家依据历代《伤寒论》诊断和治疗疾病，共有1000余条记录，包含症状、病机、方证、处方等。②古籍方剂、病种数据：纳入南京中医药大学《伤寒论》课题组提供的方

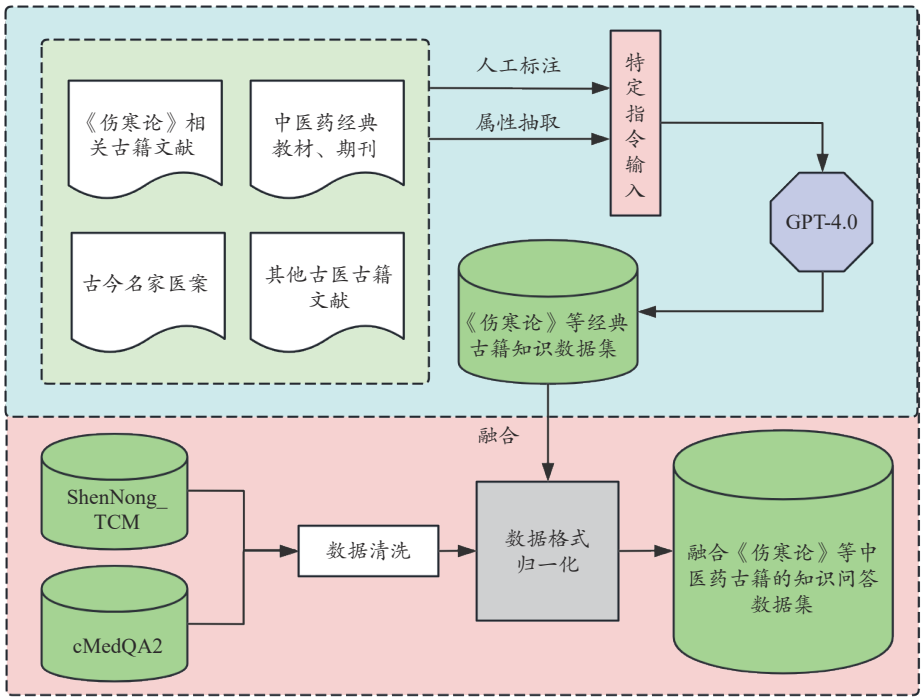

图2 中医药古籍知识问答数据集构建流程

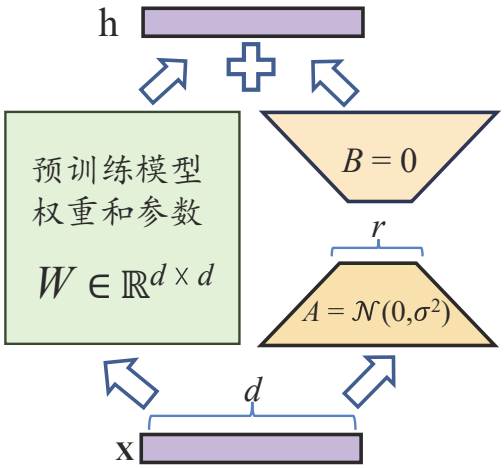

图3 Lora原理

剂、病种数据，包含疾病、证候、脉象、方剂和药物以及原文解读、方剂成分与配比以及与特定病种相关的治疗方法。③开源中医知识问答数据集：汇集开源中医知识问答数据，包括ShenNong_TCM_Dataset[20]以及cMedQA2[21]，共计10万余条问答知识对，包含患者就诊真实问答记录，涵盖了广泛的中医知识点和实际应用场景。

数据清洗：首先对收集到的数据去除重复、错误和不相关的信息；然后进行规范化处理，将其转化成知识问答对数据，最终形成JSON格式的中医药古籍知识问答数据集（见表1）。

数据标注：标注出古籍文本数据中的方剂、成分、寒热证属及其理论依据等信息，见表2与表3，表2中“+”表示出汗量级，即较多的意思。

数据增强：通过《伤寒论》中方剂、组成成分、寒热证属等特征设计如图4顶部绿色矩形框所示提示词引导GPT-4生成用于模型训练的问答对，如改写、同义词替换等手段，数据增强全流程如图4所示。

2.2 实验评估指标

选取2个中文能力评估测试集C-Eval[22]、CMMLU[23]以及1个英文测试集MMLU[24]验证各个模型的语言泛化能力，选取《伤寒学习题集》中50题作为专业知识测试集对模型进行预测分析并通过BLEU-4、ROUGE-1、ROUGE-2等指标验证《伤寒论》专业知识领域的模型预测效果。相关原理如下所示：

（1）BLEU（Bilingual Evaluation Understudy）[25]指标是一种广泛用于机器翻译质量评估的自动化工具，它比较模型输出文本与参考文本的相似度来度量模型。该方法同时考虑了短语的准确匹配（通过n-gram匹配）和整体翻译长度的适当性。如公式3所示：

$$BLEU - N = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3}$$

其中p_n是n-gram精确度，w_n是对应n-gram精确度的

表1　数据格式规范化

原格式	知识问答数据集格式
{"id": 828,"text": "疼痛有虚实之分，实性疼痛多因感受外邪、气滞血瘀、痰浊凝滞，或食积、虫积、结石等阻滞脏腑经脉，气血运行不畅所致，即所谓“不通则痛”" "annotations": [{ "Q": "不通则痛是什么意思？ ", "A": "实性疼痛多因感受外邪、气滞血瘀、痰浊凝滞，或食积、虫积、结石等阻滞脏腑经脉，气血运行不畅所致，即所谓“不通则痛”" }]}	{"instruction": "您作为一个饱含医学知识，中医知识，临床经验丰富的医生，以下是您具备的医学知识：疼痛有虚实之分，实性疼痛多因感受外邪、气滞血瘀、痰浊凝滞，或食积、虫积、结石等阻滞脏腑经脉，气血运行不畅所致，即所谓“不通则痛”；", "input": "不通则痛是什么意思？ ", "output": "实性疼痛多因感受外邪、气滞血瘀、痰浊凝滞，或食积、虫积、结石等阻滞脏腑经脉，气血运行不畅所致，即所谓“不通则痛”"}

表2　《伤寒论》方剂数据标注

方剂	组成成分	寒热	出汗	痛烦气	脉象	方证基础
桂枝汤	桂枝9 g 芍药9 g 甘草（炙）6 g 生姜（切）9 g 大枣（擘）4 g	发热、恶风	汗+	头痛，颈痛	浮弱	桂枝汤证

权重，N是考虑的最长n-gram长度（常用的是4），BP（Brevity Penalty）的计算公式如4所示：

$$\begin{cases} 1, & if\ a > b \\ e^{\left(1-\frac{b}{a}\right)}, & if\ a \leqslant b \end{cases} \tag{4}$$

其中a是机器生成文本的长度，b是参考翻译的最佳匹配长度。

（2）ROUGE（Recall-Oriented Understudy for Gisting Evaluation）[26]是一套用于自动评估文本摘要或机器翻译质量的指标，该方法计算生成文本和参考文本之间的重叠度来评估生成文本的质量，包括多种不同的评估指标，如ROUGE-N、ROUGE-L等，从不同的角度衡量文本相似度。本文采用ROUGE-N和ROUGE-L指标来评估模型的文本生成性能。ROUGE-N和ROUGE-L的公式如5、6所示：

$$ROUGE-N = \frac{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} Count(gram_n)} \tag{5}$$

其中$Count_{match}(gram_n)$是指在参考文本中和机器生成文本都出现的n-gram的数量，$Count(gram_n)$是指在参考文本中出现的n-gram的总数。

$$ROUGE-L = \frac{\sum_{S \in \{Reference\ Summaries\}} LCS(Reference, Candidate)}{\sum_{S \in \{Reference\ Summaries\}} length(Reference)} \tag{6}$$

其中$LCS(Reference, Candidate)$是参考文本和机器生成文本之间的最长公共子序列的长度，$length(Reference)$是参考摘要的长度。

（3）通用模型语言泛化能力测试基准（C-Eval、CMMLU、MMLU）：通用模型测试基准帮助研究者全面了解模型在不同任务、不同语言和不同领域中的能力。通过这些基准测试，可以衡量模型的知识广度、推理能力、语言理解能力以及特定领域的适应性，模型得分越高代表综合能力越强。

2.3　实验设置

硬件配置为Intel(R) Xeon(R) Platinum 8352V @ 2.10 GHz的CPU和24 GB显存的4090 GPU，软件环境包括Python 3.10、PyTorch 2.0、Transformers 4.33.0，以及操作系统Ubuntu 22.04。

2.4　结果与分析

模型泛化能力分析：模型微调后，通过C-Eval、CMMLU以及MMLU对模型进行泛化能力测试，各模型在3个测试集基准上的性能如表4所示。实验表明，虽然Qwen-7B-Chat在MMLU上的表现略低，但在C-Eval和CMMLU这两个更具中文语境和知识表达的基准上表现卓越。对于中医药古籍的解读和专业术语的处理，C-Eval和CMMLU的测试结果更具参考价值，在实际应用场景中，中医药古籍知识问答模型的核心任务是准确理解和传递《伤寒论》等中医古籍中的专业知识。Qwen-7B-Chat在中文基准测试中的出色表现表明其在中文任务上具备明显优势。

模型专业性能分析：为了进一步探究微调前后的模型对中医药古籍知识掌握程度，对模型进行《伤寒论》专业知识预测分析，结果如表5所示。由实验可见在中医药经典古籍领域，虽然Qwen-7B-Chat基座模型理解能力偏低，但是在经过Lora微调后，该模型性能显著提升，在所有测试集上均实现了大幅超越，尤其是在BLEU-4、ROUGE-1、ROUGE-2和ROUGE-L等评测指标上，分别达到了22.62、44.19、26.19和37.83

表3 《伤寒论》病种数据标注

原文	六经病	症状组	病机	治法	方证
太阳病，头痛，发热，汗出，恶风，桂枝汤主之。	太阳病	头痛，发热，汗出，恶风，脉浮缓	风寒袭表，卫强营弱	解肌祛风，调和营卫	桂枝汤证

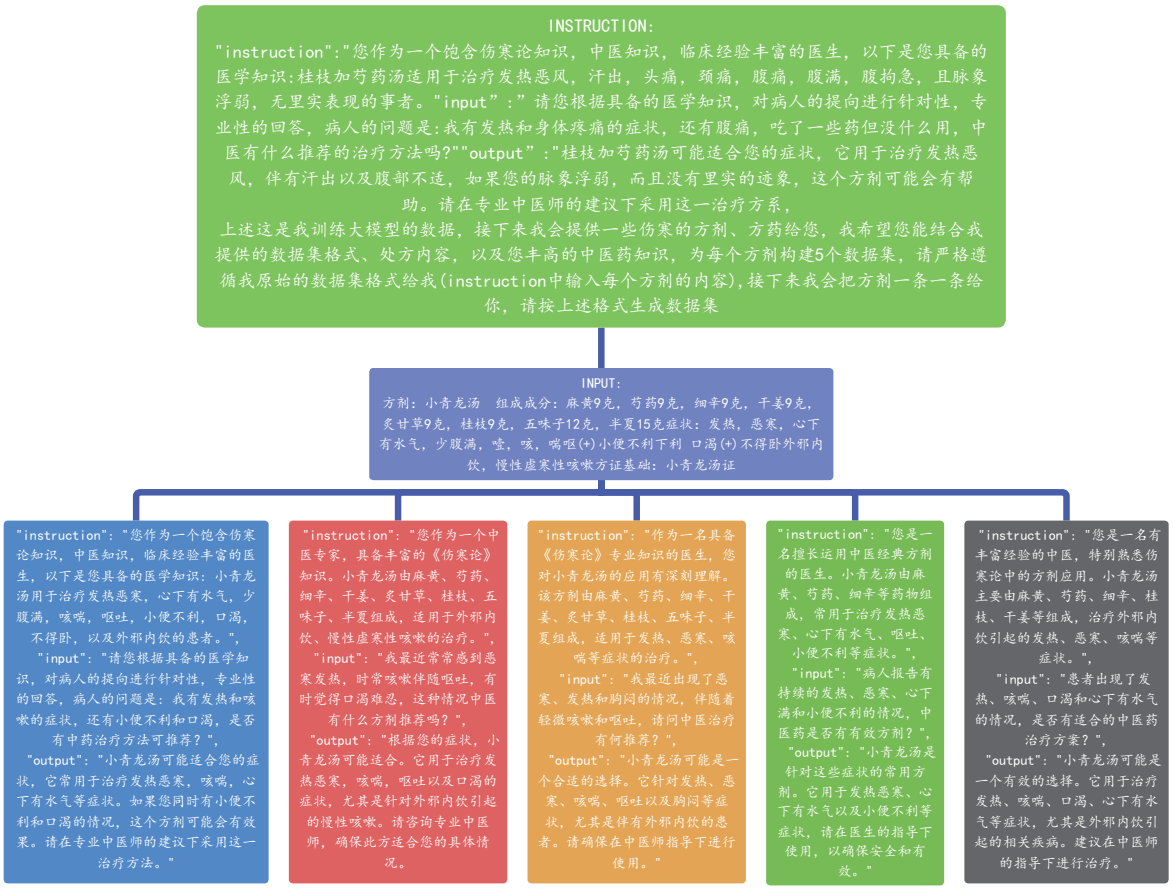

图4 数据增强流程

表4 不同测试集基准上的模型表现

模型名称	C-Eval	CMMLU	MMLU
Baichuan-7B	37.15	35.71	34.19
Baichuan2-7B-Chat	50.74	51.52	51.02
Chatglm3-6B	53.94	52.74	46.74
Xverse-7B-Chat	53.19	53.88	57.73
Qwen-7B-Chat	58.25	58.59	56.04

表5 模型微调前后预测评分

模型名称	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
Baichuan2-7B-Chat	19.01	39.54	22.46	32.74
Baichuan2-7B-Chat(Lora)	22.31	41.04	23.55	36.47
Chatglm3-6B	7.16	22.19	10.82	15.93
Chatglm3-6B(Lora)	8.77	26.28	13.31	18.99
Qwen-7B-Chat	5.01	24.56	11.89	16.43
Qwen-7B-Chat(Lora)	22.62	44.19	26.19	37.83

表6 针对Qwen-7B-Chat量化Lora测试评分

量化Lora	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
Lora	22.62	44.19	26.19	37.83
qLora-8bit	20.00	40.92	23.40	34.53
qLora-4bit	16.14	32.90	16.39	28.11

的高分，分别提升了17.61、19.63、14.3与21.4。与原始模型相比，Lora方法带来了大幅度的改进，使模型在处理《伤寒论》等中医古籍任务时表现更加优越，具备更高的准确性和专业性。

量化Lora微调性能分析：为了进一步验证不同量化版本Lora对模型性能的影响，拿Qwen-7B-Chat为例，设计了全量Lora、qLora-8bit、qLora-4bit的微调效果对比，结果如表6所示。如表6所示，虽然qLora在

表7　模型知识问答示例

问题一：桂枝加附子汤可以详细讲讲吗？		问题二：每当遇到冷风，我就发热出汗，胸部感觉闷，有时甚至会出现惊狂不安的症状，我该怎么办？	
Baichuan2-7B-Chat	桂枝加附子汤是一种中药方剂，由桂枝、芍药、生姜、大枣、甘草和附子组成。它的主要功效是温阳散寒，调和营卫，适用于风寒感冒、头痛、身痛、恶寒发热、汗出不畅、肢冷等症状。桂枝加附子汤的组成和用法如下： ·桂枝 9 g·芍药 9 g·生姜 9 g·大枣 4 枚·甘草 6 g·附子 3 g 将以上药材加水煎煮，去渣取汁，分两次服用，每日 1 剂。建议在医生指导下使用。	Baichuan2-7B-Chat	您可能正在经历焦虑或焦虑症的症状，建议您寻求专业帮助。焦虑症是一种常见的心理疾病，可以通过心理治疗、药物治疗等方式进行治疗。建议您咨询心理医生或心理咨询师，了解焦虑症的症状和治疗方法，同时保持积极的心态和良好的生活习惯，例如规律作息、适当锻炼、保持良好的饮食习惯等。
Chatglm3-6B	桂枝加附子汤是一种中药方剂，主要用于治疗感冒、发热、头痛、寒热不均等症状。它的主要成分包括桂枝、附子、生姜大枣等中药。其中，桂枝具有发汗解表、温经散寒的作用，附子则能够温阳散寒、止痛定惊，生姜和大枣则可以调和气血温中散寒。桂枝加附子汤的具体用法和剂量需要根据患者的具体情况进行调整，建议在专业医生的指导下使用。同时，在使用中药方剂时，也应注意药物的相互作用和不良反应，避免出现不良反应或药物相互作用的情况。	Chatglm3-6B	您可能正在经历焦虑或焦虑症的症状，建议您寻求专业帮助。焦虑症是一种常见的心理疾病，可以通过心理治疗、药物治疗等方式进行治疗。建议您同时保持积极的心态和良好的生活习惯，例如规律作息、适当锻炼、保持良好的饮食习惯等。
Qwen-7B-Chat	桂枝加附子汤是一种温阳祛寒的中药方剂，由桂枝、白芍、生姜、大枣、甘草、附子组成。它的主要功效是温阳祛寒、解表散寒、调和营卫、行气止痛。桂枝具有温阳散寒、调和营卫的作用；白芍具有活血止痛、调和营卫的作用；生姜具有温阳散寒、行气止痛的作用；大枣具有补中益气、调和营卫的作用；甘草具有调和诸药、缓急止痛的作用；附子具有温阳散寒、行气止痛的作用。桂枝加附子汤适用于体寒、手足冷、恶寒发热、胸闷、头痛、颈项拘挛、四肢无力等症状。但是，由于附子的毒性较大，应在中医医生的指导下使用。	Qwen-7B-Chat	您的症状可能是由体内的阳气不足、寒邪侵袭引起的，建议您采用温阳祛寒的中药方剂，如桂枝加附子汤、理中丸等。同时，注意保暖，避免接触寒冷的空气和物质，多喝温水，加强锻炼，保持良好的生活习惯。但是，如果症状严重或持续，请及时就医。

BLEU-4 和 ROUGE 指标上的表现较全量 Lora 有所下降，但这一性能牺牲是为了实现大幅度降低参数量和算力需求所做的权衡。特别是 qLora-8bit 和 qLora-4bit 在减少模型存储占用和推理计算成本上展现了显著优势。qLora-8bit 的 BLEU-4 为 20.00，ROUGE-L 为 34.53，虽然略低于 Lora 模型，但其生成质量依然能够满足《伤寒论》等中医药古籍问答等需求。尤其是在资源受限的应用场景中，qLora 凭借轻量化设计，显著降低了硬件和计算资源的压力，为大规模部署提供了可能性。在某些特定场景中，如简化版的中医药知识查询任务，性能的下降对实际效果影响较小，而 qLora 带来的效率提升则极具实际价值。qLora-4bit 虽然 BLEU-4 和 ROUGE 得分略低，但其极大的资源节省使其成为低成本环境下的理想选择。

3　模型问答示例及专家评估

本文设计模型中医药古籍知识问答示例如表 7 所示。

《伤寒论》课题组专家对不同模型回答作出的专业评估如下：

（1）Baichuan2-7B-Chat 的回答简洁明了，可以抓住问题的关键直接给出答案。尤其针对问题一的回答甚至给出了参考药量及煎煮方法，这是其它模型所没有的。但该模型逻辑欠通顺，给出的答案错误较多，专业度略显不足。如问题一附子的煎煮方法上既未明确炮附子，又未指出附子需先煎，可能会给检索者带来负面影响。

（2）Chatglm3-6B 的回答会对问题进行拓展，给用

户带来更为丰富的检索答案。不足之处在于仍然存在部分错误答案，如问题一对桂枝加附子汤组成的回答漏了甘草一药。对于问题二的回答，该模型和Baichuan2-7B-Chat完全相同，未给用户带来专业性的指导意见。

（3）Qwen-7B-Chat的回答相对丰富和准确。无论是对《伤寒论》的介绍，还是对桂枝加附子汤中附子毒性的提醒，以及问题二具体症状的用药建议和温馨提示，都标志着该模型的成熟度更高。然而，从中医药学的专业角度来说，对于“伤寒”的描述不够准确，问题二的用药也并非十分合理，有待进一步优化。

综合3个模型的表现，Baichuan2-7B-Chat回答简洁但逻辑不通，专业度不足，存在误导风险；Chatglm3-6B回答较丰富，但也有中医药方错误，专业性欠缺；Qwen-7B-Chat回答准确全面，涵盖医学理论和方剂建议，表现最佳，但细节和专业度仍可进一步提升。

4 小结

本文自主构建了一个以历代《伤寒论》为核心的中医药古籍知识问答数据集，对多个预训练模型使用Lora微调，并对性能最佳的Qwen-7B-Chat进行全量Lora、qLora-8bit、qLora-4bit微调性能比对，最终构建轻量化中医药古籍大语言模型，通过模型语言泛化能力及专业性能评估证明模型在《伤寒论》等中医药古籍领域回答丰富、精确、专业性强。qLora轻量化技术在保持可接受性能的前提下，大幅降低模型微调参数量和算力需求，展现其在资源受限场景中的强大潜力，对中医临床辅助决策以及中医药知识传播与发展有一定的意义，也为中医药古籍信息化提供了新思路。在未来的工作中，将在此基础上不断提升数据集的数量与质量，探索更高效率的模型量化构建方法并引入多模态、反馈优化等功能进一步提升模型学习能力。

参考文献

1 林睿凡, 周洪伟, 刘亮亮, 等. 基于本体方法构建《伤寒论》版本知识图谱. 中国中医基础医学杂志, 2023, 29(10):1630-1633.

2 徐月梅, 胡玲, 赵佳艺, 等. 大语言模型的技术应用前景与风险挑战. 计算机应用, 2024, 44(6):1655-1662.

3 Achiam J, Adler S, Agarwal S, *et al*. Gpt-4 technical report. *arXiv*. 2023. https://doi.org/10.48550/arXiv.2303.08774

4 Houlsby N, Giurgiu A, Jastrzebski S, *et al*. Parameter-efficient transfer learning for NLP. *International conference on machine learning. PMLR*, 2019:2790-2799.

5 Ding N, Qin Y, Yang G, *et al*. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat Mach Intell*, 2023, 5:220-235.

6 Hu E J, Shen Y, Wallis P, *et al*. Lora: Low-rank adaptation of large language models. *arXiv*. 2021. https://doi. org/10.48550/arXiv: 2106.09685

7 Li Y, Li Z, Zhang K, *et al*. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus*, 2023, 15(6):e40895.

8 Wu C, Lin W, Zhang X, *et al*. PMC-LLaMA: toward building open-source language models for medicine. *J Am Med Inform Assoc*, 2024, 31(9):1833-1843.

9 Chowdhery A, Narang S, Devlin J, *et al*. PaLM: scaling language modeling with pathways. *J Mach Learn Res*, 2023, 24(1):11324-11436.

10 Singhal K, Azizi S, Tu T, *et al*. Large language models encode clinical knowledge. *Nature*, 2023, 620(7972):172-180.

11 Xiong H, Wang S, Zhu Y, *et al*. DoctorGLM: Fine-tuning your Chinese doctor is not a herculean task. *arXiv*. 2023. https://doi. org/10.48550/arXiv.2304.01097

12 Liu X, Ji K, Fu Y, *et al*. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv*. 2021. https://doi.org/10.48550/arXiv.2110.07602

13 GLM T, Zeng A, Xu B, *et al*. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools. *arXiv*. 2024. https://doi.org/10.48550/arXiv.2406.12793

14 Touvron H, Lavril T, Izacard G, *et al*. Llama: Open and efficient foundation language models. *arXiv*. 2024. https://doi. org/10.48550/arXiv.2302.13971

15 Wang H, Liu C, Xi N, *et al*. Huatuo: Tuning llama model with Chinese medical knowledge. *arXiv* . 2023. https://doi.org/10.48550/arXiv.2304.06975

16 Chen Y, Wang Z, Xing X, *et al*. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by ChatGPT. *arXiv*. 2023. https://doi. org/10.48550/arXiv.2310.15896

17 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. 图书馆论坛, 2024, 44(10):103-112.

18 Yang A, Xiao B, Wang B, *et al*. Baichuan 2: Open large-scale language models. *arXiv*. 2023. https://doi.org/10.48550/arXiv.2309.10305

19 Bai J, Bai S, Chu Y, *et al*. Qwen technical report. *arXiv*. 2023. https://doi.org/10.48550/arXiv.2309.16609

20 Wei Zhu. Datasets:michaelwzhu/ShenNong_TCM_Dataset. 2023. https://huggingface.co/datasets/michaelwzhu/ShenNong_TCM_Dataset.

21 Zhang S, Zhang X, Wang H, *et al*. Multi-scale attentive interaction

networks for Chinese medical question answer selection. *IEEE Access*, 2018, 6:74061-74071.

22 Huang Y, Bai Y, Zhu Z, *et al*. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 2023: 62991-63010.

23 Li H, Zhang Y, Koto F, *et al*. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv*. 2023. https://doi. org/10.48550/arXiv.2306.09212

24 Hendrycks D, Burns C, Basart S, *et al*. Measuring massive multitask language understanding. *arXiv*. 2020. https://doi. org/10.48550/arXiv.2009.03300

25 Papineni K, Roukos S, Ward T, *et al*. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002: 311-318.

26 Lin C Y. Rouge: A package for automatic evaluation of summaries. *Association for Computational Linguistics*. 2004:74-81.

Research on Lightweight Large Language Models for Ancient Traditional Chinese Medicine Texts Based on Lora Fine-Tuning

CHAI Jingxian[1,2], LANG Xufeng[1,2], LI Hongyan[1,2], ZHOU Zuojian[1,2], LING Yun[3], ZHAN Libin[4], HU Kongfa[1,2], QIAO Xuebin[2,5]

(1. School of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, Nanjing 210023, China; 2. Jiangsu Province Smart Traditional Chinese Medicine Health Service Engineering Research Center, Nanjing 210023, China; 3. School of Chinese Medicine, Nanjing University of Chinese Medicine, Nanjing 210023, China; 4. Center for Innovative Engineering Technology in Traditional Chinese Medicine, Liaoning University of Traditional Chinese Medicine, Shenyang 110847, China; 5. School of Elderly Care and Management, Nanjing University of Chinese Medicine, Nanjing 210023, China)

Abstract: Objective To address the challenges of constructing large language models for traditional Chinese medicine (TCM) classics, which are complex and expensive to fine-tune, this study explores a lightweight fine-tuning method for such models, aiming to develop a question-answering model centered on TCM classics, particularly various editions of Shang Han Lun through the ages. Methods Dataset construction involved designing prompts to guide GPT-4 in generating Q&A pairs based on Shang Han Lun and integrating them with the ShenNong_TCM_Dataset and cMedQA2 datasets. Five general-purpose large models were selected for Lora fine-tuning. The best model was chosen through evaluation, and the performance of multiple quantized versions was validated. Results After fine-tuning, the BLEU, ROUGE-1, ROUGE-2, and ROUGE-L metrics for the Qwen-7B-Chat model improved by 17.61, 19.63, 14.3, and 21.4, respectively, compared to the base model. Conclusion The selected model in this study is capable of effectively understanding and utilizing professional terms and concepts from TCM classics, such as Shang Han Lun, to provide accurate answers to user queries. Compared to similar models, it requires lower fine-tuning costs and computational power, contributing to the dissemination of TCM knowledge and the development of intelligent systems.

Keywords: Large language models, Traditional Chinese Medicine ancient texts, Shang Han Lun, Low-rank adaptation, Model optimization

（责任编辑：李青）