E-Commerce Letters 电子商务评论, 2024, 13(4), 1611-1620
Published Online November 2024 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2024.1341314

Hans汉斯

医疗电商平台中大语言模型驱动的中文医学对话系统研究

滚流海[1]，吴　娜[2]，曾以春[1]

[1]贵州大学大数据与信息工程学院，贵州　贵阳
[2]贵州中医药大学第一临床医学院，贵州　贵阳

收稿日期：2024年8月8日；录用日期：2024年10月30日；发布日期：2024年11月6日

摘　要

随着互联网技术和人工智能的迅猛发展，医疗电商平台在现代医药服务中扮演着越来越重要的角色。本研究提出了一种基于大语言模型(LLM)的中文医学对话系统模型MedAsst，并探讨其在医疗电商平台中的应用。该模型以Qwen2-7B为基础，通过LoRA方法在147万条医学问答数据上进行监督微调。本文在医学多项选择题测试和自定义医学问答数据集上对MedAsst的有效性进行了全面评估。实验结果显示，MedAsst在BLEU-4、ROUGE-1、ROUGE-2和ROUGE-L等评价指标上均优于其他基线模型，特别是在医学问答能力上展现出显著优势。与LlaMa-3-8B、Gemma-7B、Mistral-7B和未经微调的Qwen2-7B模型相比，MedAsst通过合理的微调策略在特定领域的任务中表现出色，证明了监督微调的必要性和有效性。本文的研究不仅提升了模型在中文医学问答任务中的表现，也展示了大语言模型在医疗电商平台中的应用潜力，为未来在更复杂场景中的优化和实际应用提供了有力支持。

关键词

大语言模型，监督微调，医疗电商，医学对话系统

Research on Chinese Medical Dialogue System Driven by Large Language Models in Medical E-Commerce Platforms

Liuhai Gun[1], Na Wu[2], Yichun Zeng[1]

[1]College of Big Data and Information Engineering, Guizhou University, Guiyang Guizhou
[2]The First College for Clinical Medicine, Guizhou University of Traditional Chinese Medicine, Guiyang Guizhou

Received: Aug. 8th, 2024; accepted: Oct. 30th, 2024; published: Nov. 6th, 2024

文章引用: 滚流海, 吴娜, 曾以春. 医疗电商平台中大语言模型驱动的中文医学对话系统研究[J]. 电子商务评论, 2024, 13(4): 1611-1620. DOI: 10.12677/ecl.2024.1341314

Abstract

With the rapid development of Internet technology and artificial intelligence, medical e-commerce platforms play an increasingly important role in modern pharmaceutical services. This study proposes a Chinese medical dialogue system model MedAsst based on Large Language Model (LLM) and explores its application in medical e-commerce platform. The model is based on Qwen2-7B, and supervised fine-tuning is performed on 1.47 million medical question and answer data by LoRA method. In this paper, the effectiveness of MedAsst is thoroughly evaluated on a medical multiple-choice test and a customised medical quiz dataset. The experimental results show that MedAsst outperforms other baseline models on the evaluation metrics of BLEU-4, ROUGE-1, ROUGE-2, and ROUGE-L, and in particular demonstrates a significant advantage in medical quizzing ability. Compared with LlaMa-3-8B, Gemma-7B, Mistral-7B, and the unfine-tuned Qwen2-7B model, MedAsst performs well in domain-specific tasks through reasonable fine-tuning strategies, demonstrating the necessity and effectiveness of supervised fine-tuning. The research in this paper not only improves the performance of the model in the Chinese medical Q&A task, but also demonstrates the potential application of large language models in medical e-commerce platforms, which provides strong support for future optimisation and practical application in more complex scenarios.

Keywords

Large Language Models, Supervised Fine-Tuning, Medical E-Commerce, Medical Dialogue Systems

1. 引言

随着互联网技术的快速发展，电子商务已经深刻地改变了各行各业的运作模式，其中医疗电商成为近年来的一个重要增长点。医疗电商平台不仅提供了药品、医疗器械的便捷购买渠道，还逐渐向在线医疗咨询、智能问诊等领域扩展[1]，使医药服务转变为健康管理服务，从而提高了医药电商的使用率。与此同时，人工智能(AI)技术，为医疗电商平台注入了新的活力和可能性，受益于大规模医疗数据和先进的神经网络模型，人工智能在医学对话领域的应用取得了显著进展[2]。

AI 在医学对话系统中的应用正在得到越来越多的关注和研究投入，传统的医学对话系统发展历程分为：专家系统和规则引擎阶段、传统机器学习算法阶段及深度学习阶段。最近，大语言模型(Large Language Model, LLM)取得了重大进展，例如，GPT3 [3]和 ChatGPT (OpenAI)，它们的对话能力十分出色，能够理解和回答各种问题，甚至在某些领域超越人类。然而，很多 LLM 主要是用英文语料库训练的，这限制了它们对中文的理解和生成能力，使其在中文环境中直接应用不太理想[4]。阿里云推出的通义千问大模型：Qwen 系列[5]和 Qen2 系列[6]在中文理解和生成等任务上表现优异，超越了大多数开源模型。

尽管 LLM 有许多优点，但它们并不是专门为医学领域设计的，缺乏医学领域专业知识，因此，使用专业的在线医学问诊数据对 LLM 进行微调显得尤为重要，微调模型可以使其学习到用于医疗电商的医学知识。

本研究提出了一个医学对话助手模型 MedAsst，这是一个专注于中文、为医疗电商平台定制的生成式对话 LLM，图 1 展示了一个基于 MedAsst 的 AI 医学对话的例子。我们的模型建立在开源的通义千问

大模型 Qwen2-7B (76 亿个参数)基础上，使用 LoRA (LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS)方法[7]在超过 147 万条的微调数据上进行监督微调(Supervised Fine-Tuning, SFT)。通过与参数相当的开源模型进行比较，评估我们模型在中文医学对话任务上的表现。通过对 Qwen2 大模型在中文医学对话的研究，本文希望为医疗电商、在线医疗的参与者和研究者提供有益的见解和指导，并促进大语言模型及其微调技术在医疗电商的应用和发展。生成式大模型的研究处于蓬勃发展阶段，通过构建 AI agent [8]，在预防、诊疗的全流程中给予患者提供建议性的帮助。

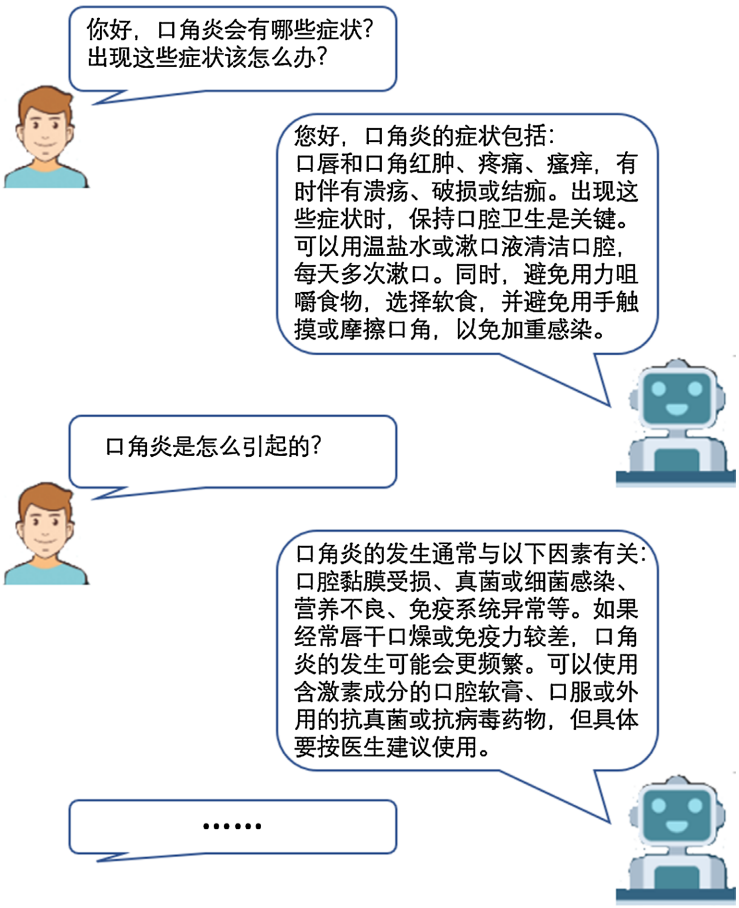

Figure 1. A sample MedAsst-based AI medical dialogue
图 1. 基于 MedAsst 的 AI 医学对话的一个样例

2. 相关工作

传统的医学对话系统研究中使用的方法包括基于规则的方法、传统机器学习方法和专家系统等。例如，Hayashi [9]试图从医疗数据和人类知识中提取基于规则的表示来进行诊断。Wilson [10]等人提出了一种基于上下文匹配的健康对话系统。Li [11]等人利用机器学习框架，在定制的数据库上训练对话系统来教授医学生解剖学。传统的处理医学对话系统的方法，只能在少数的场景中使用，局限性很大，这些方法甚至只能用在医学对话任务中的小部分模块。

随着深度学习的发展，神经网络模型在对话领域得到广泛应用。例如，颜永[12]等人提出了一种融合双向长短期记忆神经网络和强化学习模型的生成式对话方法，以尝试解决对话系统在回复过程中回答单调的问题。Wang [13]等人提出了一个用于构建智能医疗对话系统的 Bert 模型，在一些在线电

子商务数据集上实验表明，他们的模型在智慧城市中建立智能医疗对话系统很有潜力。马德草[14]等人提出了一种基于实体知识推理的端到端任务型对话系统，他们的方法提高了模型的响应生成和实体预测能力。然而这些方法的应用场景仍旧有较大限制，对话的流畅度也不够高，处理复杂任务的能力较弱。

GPT3 [3]和 ChatGPT (OpenAI 2022)等大语言模型(LLM)的显著成就受到了广泛关注，引发了人工智能的新浪潮，尽管 OpenAI 没有开放其训练策略和权重数据。随着 LLaMa [15]等开源 LLM 的出现，更多的研究机构发布了开源 LLM。Google 的 Gemma 团队[16]发布了他们的轻量级 Gemma 模型，其在语言理解、推理和安全性的基准测试中表现出强大性能。Mistral AI 发布了他们的轻量级模型 Mistral 7B [17]，其在人工和自动化基准测试上都超过了 Llama 2 13B。国内阿里巴巴也发布了开源通义千问 LLM——Qwen2 [6]系列，与先进的开源语言模型相比，Qwen2 总体上超越了大多数开源模型，并在语言理解、语言生成、推理等一系列基准测试中显示出强大的竞争力。因此本研究选择 Qwen2 作为微调的基础模型。

在需要复杂知识和高精度的医疗问诊领域，大型模型通常表现不佳。不过研究人员使用监督微调等方法取得了重大进展。例如 Wang [4]等人基于 LlaMa 模型进行了监督微调，提出了 HuaTuo 模型，他们的实验结果表明 HuaTuo 生成的响应具有更可靠的医学知识。Yang 等人基于 Llama 模型在中文多轮医疗对话数据集 CmtMedQA 上进行监督微调等训练，提出了 Zhongjing 模型[18]，该模型在复杂对话和主动问询能力上表现很好。然而，大语言模型微调的成本非常高，例如对一个 7B 的 LLM 进行全参微调，大约需要 120 GB 的显存资源。好在 Hu 等人提出了一种低秩自适应的微调方法——LoRA [7]，利用该方法对一个 7 B 的 LLM 进行微调，仅需要约 20 GB 的显存，这种小量级的显存资源即便是医疗电商平台的小型商户也容易获取到。因此本文采用 LoRA 方法对 Qwen2-7B 模型进行微调。

3. 医学对话助手模型(MedAsst)的构建

在这一章中我们将介绍我们的中文医学对话助手模型——MedAsst 的构建过程。

3.1. 基础模型

Qwen2 从根本上是基于 Transformer [19]架构构成的大型语言模型，其特点是具有因果掩码的自注意力机制。Qwen2 模型的架构包括多个 Transformer 层，每个层都配备有因果注意力机制和前馈神经网络(FFNs)。Qwen2 采用分组查询注意力(Grouped Query Attention, GQA) [20]而不是传统的多头注意力机制，GQA 在推理过程中优化了缓存的使用，显著提高了吞吐量。Qwen2 包含一系列的大语言模型，参数范围从 5 亿到 720 亿，这些模型对研究界开放源代码。本研究中，我们采用了 Qwen2-7B 来进行构建，其更适用于少量设备资源下的训练。

3.2. 医学问答数据集构建

为了让模型学习到更多医学问诊方面的知识，我们对 4 个医学数据集进行筛选和处理，来构建高质量的医疗问答数据集。这 4 个医学数据集分别源自：MedChatZH [21]、Huatuo-26M [22]、CmtMedQA [18]和 webMedQA [23]。我们只选取这些数据集中的训练集作为来构建我们的微调数据集，对于每条微调数据集，我们统一处理转换为 Alpaca [24]的微调数据样式。其中，多轮医疗问答对话的微调数据示例如图 2 所示，普通医疗对话的微调数据样例如图 3 所示。

经过筛选和处理这 4 个开源数据集后，我们构建了用于微调的医疗问答数据集，其含有超过 147 万条的中文医学问答数据，该数据集的构成信息如表 1 所示。

```
{
    "instruction": "好的，我会去正规医院检查并按医生的治疗指导进行处理。谢谢你的建议！",
    "input": "",
    "output": "不客气，希望你能早日康复。如果还有其他问题，随时向我咨询。祝你身体健康！",
    "history": [
        [
            "我最近得了前列腺炎，但是听说轻微的前列腺炎不需要治疗，是不是真的呢？",
            "轻微的前列腺炎虽然症状可能并不明显，但是依然需要积极进行治疗。如果不及时治疗，不仅会导致病情加重，还会增加治疗的难度和效果。"
        ],
        [
            "那如果我不治疗，会有什么后果呢？",
            "如果不及时治疗，前列腺炎的病情可能会进一步加重，严重的话可能导致尿频、尿急、尿痛等不适症状加重，甚至出现性功能障碍、包皮龟头炎等并发症。"
        ],
        [
            "我该怎么治疗前列腺炎呢？",
            "对于前列腺炎的治疗，首先你需要去正规的医院进行检查，以便明确导致炎症发生的原因。在医生的指导下，采取规范用药的方式进行治疗。一般使用抗炎类药物进行治疗可能会比较有效。"
        ],
        [
            "除了药物治疗，我可以采取哪些措施改善前列腺炎？",
            "除了药物治疗，你还可以在日常生活中采取一些措施来改善前列腺炎。首先，尽量少吃或者避免辛辣刺激性的食物，戒烟酒，避免频繁性生活和手淫，这些都是有助于前列腺炎康复的措施。此外，多喝水，尽量多排尿，也可以缓解炎症症状。"
        ]
    ]
}
```


Figure 2. Sample fine-tuned data from a multi-round medical Q&A dialogue
图 2. 多轮医疗问答对话的微调数据样例


```
{
    "instruction": "在我量血压时高压总是不知道怎么计算，开始听到的声音很微细等下降水银柱后到130或140时就听的很清的心跳声音，如何才是高压。",
    "input": "",
    "output": "您好，测量血压需要使用正确的方法。首先，将气袖缚于上臂，其下缘要距肘窝2-3厘米，不可过紧或过松。然后，将听诊器胸件放在肘部肱动脉搏动处。接着，向气袖内充气，待肱动脉搏动消失，再将汞柱升高2.6-4.0千帕(20-30毫米汞柱)。此时，气袖内的压力经软组织作用于肱动脉，因其压力超过动脉的收缩期压力，故血管被压扁，已无血液通过此处的血管，所以用听诊器听不到任何声音。接着由气球阀门处向外缓慢放气，使气袖内压力降低，汞柱缓慢下降。当气袖内的压力等于或稍低于收缩压时，随着心脏收缩射血，血液即可冲开被压扁的动脉而形成涡流，并发现响声。当听到第一个声音时，血压计所批示的压力数值即为收缩压。继续放气，在气袖内压力低于收缩压而高于舒张压的这段时间内，心脏每收缩一音乐会匀可以听到一次声音。当气袖内压力等于或稍低于舒张压时，血流又复通畅，涡流消失，则声音突然减弱，很快消失。此时血压计所提示的数值即为舒张压。希望这些信息能对您有所帮助。如果您还有其他问题，请随时向我提出。"
}
```


Figure 3. Sample fine-tuned data for medical dialogue
图 3. 医疗对话的微调数据样例

Table 1. Information on the composition of the medical Q&A dataset used for fine-tuning
表 1. 用于微调的医疗问答数据集的构成信息

数据来源	数据描述	数据条数
MedChatZH	医学问答对话，只抽取其中的训练数据集	763,629 条
Huatuo-26M	医学问答对话，只抽取其中的医学百科问答和微调专用数据	588,462 条
CmtMedQA	多轮医学问答对话，只抽取其中的训练数据集	68,023 条
webMedQA	医学问答对话，只抽取其中的训练数据集	50,570 条
构建的医疗问答数据集	我们构建的格式统一的医疗问答微调数据集	1,470,684 条

3.3. 监督微调(SFT)

SFT 是赋予模型在专业领域对话能力的关键阶段。通过高质量的医患对话数据，该模型可以有效地调用在训练过程中积累的医学知识，从而理解和响应用户的问询。在本研究中，我们使用了 LoRA 方法对 Qwen2-7B 模型进行微调。LoRA 是一种高效的微调方法，它通过低秩分解来减少参数量，从而降低了

训练的计算成本和存储需求。具体实现如下：

在 LoRA 方法中，预训练模型的权重矩阵 W_0 被分解为两个较小的矩阵 A 和 B，从而生成一个新的权重矩阵 W：

$$W = W_0 + \Delta W, \Delta W = AB^T \tag{1}$$

其中，W_0 是预训练权重矩阵，A 和 B 是低秩矩阵，且 $A \in \mathbb{R}^{d \times r}$，$B \in \mathbb{R}^{r \times k}$，其中 d 是输入的维度，$k$ 是输出的维度，r 是低秩矩阵的秩，通常 $r \ll d, k$。

LoRA 的核心在于低秩分解，将权重矩阵分解为两个更小的矩阵，从而降低参数量并提升计算效率。在实际操作中，权重更新部分被限制为低秩矩阵的乘积形式：

$$\Delta W = AB^T \tag{2}$$

这意味着在训练过程中，模型的参数更新只涉及 A 和 B 矩阵，而不直接修改原始的预训练权重矩阵 W_0。

我们使用构建的包含 147 万条问答对的医疗问答数据集，按照 Alpaca 的微调数据格式进行处理，确保数据的一致性和标准化。微调时使用的是 LLaMA-Factory 框架[25]，从魔塔社区[26]中加载预训练的 Qwen2-7B 含 76 亿参数的模型，并进行模型参数的初始化。微调参数中，输入模型的最大训练数据长度 cutoff_len 设置为 512，批量大小 batch_size 设置为 4，训练轮数 epochs 设置为 3.0。经过多次微调实验后，选择最大学习率为 0.0001，学习率从训练开始时为 0 到 100 步后线性提升到最大值，到达最大值后，使用余弦衰减策略将学习率平滑下降。

4. 实验结果与分析

4.1. 实验设置与基线模型

我们的实验是在 Pytorch 2.1.2 上进行的，微调过程中仅使用了一个显存为 24 GB 的 NVIDIA A10 GPU，微调的参数设置如 3.3.节中所述。对于实验过程中的所有模型，都是直接从魔塔社区种加载预训练模型，然后在相同的医疗问答数据集上进行微调。微调过程中的参数设置都是一样。

为了评估 MedAsst 的中文医学对话的有效性，我们选择了 4 个与 MedAsst 参数相当的先进大语言模型来进行比较，基线模型详细如下：

- LlaMa-3-8B：由 Meta 发布的 80 亿参数的指令调优生成式大语言模型，该模型针对对话进行了优化，在常见的行业基准测试中优于许多可用的开源聊天模型。在我们的医疗问答数据集上进行过微调。
- Gemma-7B：是 Google 推出的轻量级、先进的 70 亿参数开放式模型，非常适合各种文本生成任务，包括问答、总结和推理，仅提供英文版的指令调优变体。在我们的医疗问答数据集上进行过微调。
- Mistral-7B：由 Mistral AI 团队发布的一个具有 70 亿参数的预训练大语言模型，没有加入任何的调优机制。在我们的医疗问答数据集上进行过微调。
- Qwen2-7B：直接从魔塔社区上加载的预训练 Qwen2-7B 大模型，未经任何微调优化。

4.2. 评测数据与评价指标

我们的测试分为两个部分，一部分是中文多项选择测试，另一部分是在整合的测试数据集上做自动评估实验。

4.2.1. 多项选择题数据与指标

这部分我们的测试中，我们从 CMMLU [27]和 C-Eval [28]两个中文测试基准中，选取其中医学相关板块的多项选择题来作为测试数据集，整理后共计 1415 道题，每道题目由一个问题和四个选项组成，其

中只有一个选项是正确的，在 A’、B’、C’和 D’中选择概率最高的选项作为模型的选择。

使用分数指标来衡量模型的知识和推理能力：

$$得分数=\frac{正确个数}{总个数}\times 100 \tag{3}$$

4.2.2. 自动评估实验数据与指标

这部分的测试中，我们处理并整合了 MedChatZH、Huatuo-26M、CmtMedQA 和 webMedQA 这四个数据集中的测试数据，得到了一个包含 8975 条医学对话记录的测试数据集，其中包含了多轮和单轮的问答对话。

为了量化生成的句子和参考句子之间的相似性，使用了 BLEU-4 度量指标，其评估了 n-gram 重叠以评估句子级别的流畅性。使用 ROUGE-1，ROUGE-2 和 ROUGE-L 这几个指标基于最长公共子序列中的单词匹配来评估模型生成的质量。

4.3. 结果与分析

4.3.1. 多项选择题测试结果与分析

表 2 显示了所有模型在多项选择题测试中的得分结果。从实验结果可以看出，MedAsst 模型在 CMMLU 和 C-Eval 多项选择题测试中均取得了最高的平均得分，分别为 88.32 和 91.11，总平均得分达到 89.72。与其他基线模型相比，MedAsst 在多项选择题测试中表现出显著优势。

Table 2. Results of each model on multiple-choice tests
表 2. 各个模型在多项选择题测试上的结果

模型	CMMLU 平均得分	C-Eval 平均得分	总平均得分
LlaMa-3-8B	42.04	54.44	48.24
Gemma-7B	24.23	27.78	26.01
Mistral-7B	35.85	53.33	44.59
Qwen2-7B	87.02	87.78	87.40
我们的 MedAsst	88.32	91.11	89.72

这主要得益于 MedAsst 模型在大量医疗问答数据集上的监督微调，通过 LoRA 方法对 Qwen2-7B 模型进行优化，使其不仅继承了 Qwen2-7B 的强大基础能力，还在特定的医疗问答任务上进一步提升了模型的准确性和有效性。Qwen2-7B 的取得了 87.40 的高分，而我们的模型在其基础上得到了进一步的提升。而其他几个模型，在中文医疗领域选择题中的表现则比较差，总平均得分甚至没有达到 50 分，说明这些模型在应对中文医学任务时表现不理想。MedAsst 模型展示了在医疗电商平台中应用大语言模型驱动 AI 医学对话系统的可行性和有效性。

4.3.2. 自动评估实验结果与分析

这部分的测试主要关注模型的医学问答能力，我们在医学问答测试数据集上对所有模型进行了评估，使用了 BLEU-4、ROUGE-1、ROUGE-2 和 ROUGE-L 作为评价指标。表 3 展示了各个模型在医学问答测试中的表现。

从表 3 可以看出，我们的 MedAsst 模型在所有评价指标上都优于其他基线模型，展现出了超强的医学问答能力。MedAsst 模型在 BLEU-4、ROUGE-1、ROUGE-2 和 ROUGE-L 上的得分分别为 6.70、22.80、6.66 和 18.17，均高于其他基线模型。MedAsst 模型的出色表现主要归功于在大量医学问答数据集上的监督微调，使其在医学领域的回答更为准确和流畅。

LlaMa-3-8B 模型虽然在通用对话任务上表现优异，但在医学问答测试中得分较低，特别是 BLEU-4 仅为 1.83，ROUGE-L 为 8.16。这表明 LlaMa-3-8B 模型在中文医学领域的细粒度理解和生成能力有限。Gemma-7B 模型的表现稍好于 LlaMa-3-8B，但其在中文医学问答中的表现仍不尽如人意。尽管在我们的医疗问答数据集上进行了微调，Mistral-7B 模型在所有指标上的得分也较低。Qwen2-7B 模型在各项指标上表现良好，我们的 MedAsst 是在其基础上微调的，这也表明了他医学问答任务中仍有提升空间。

总体而言，实验结果表明，MedAsst 模型通过合理的微调策略，在医学问答任务中展现了最强的性能，显著优于其他参数相当的基线模型。这证明了大语言模型在特定领域进行监督微调的必要性和有效性，为未来在医疗电商平台中的应用提供了有力支持。

为了更直观地展示各模型在中文医学问答测试中的表现，我们绘制了如图 4 所示的柱状图，该图显示了每个模型在 BLEU-4、ROUGE-1、ROUGE-2 和 ROUGE-L 四个评价指标上的得分情况。从图 4 可以

Table 3. Results of each model on the medical quiz test
表 3. 各个模型在医学问答测试上的结果

模型	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
LlaMa-3-8B	1.83	20.39	5.20	8.16
Gemma-7B	6.01	22.20	5.86	17.56
Mistral-7B	5.36	21.53	5.46	16.36
Qwen2-7B	6.03	21.82	5.83	16.17
我们的 MedAsst	6.70	22.80	6.66	18.17

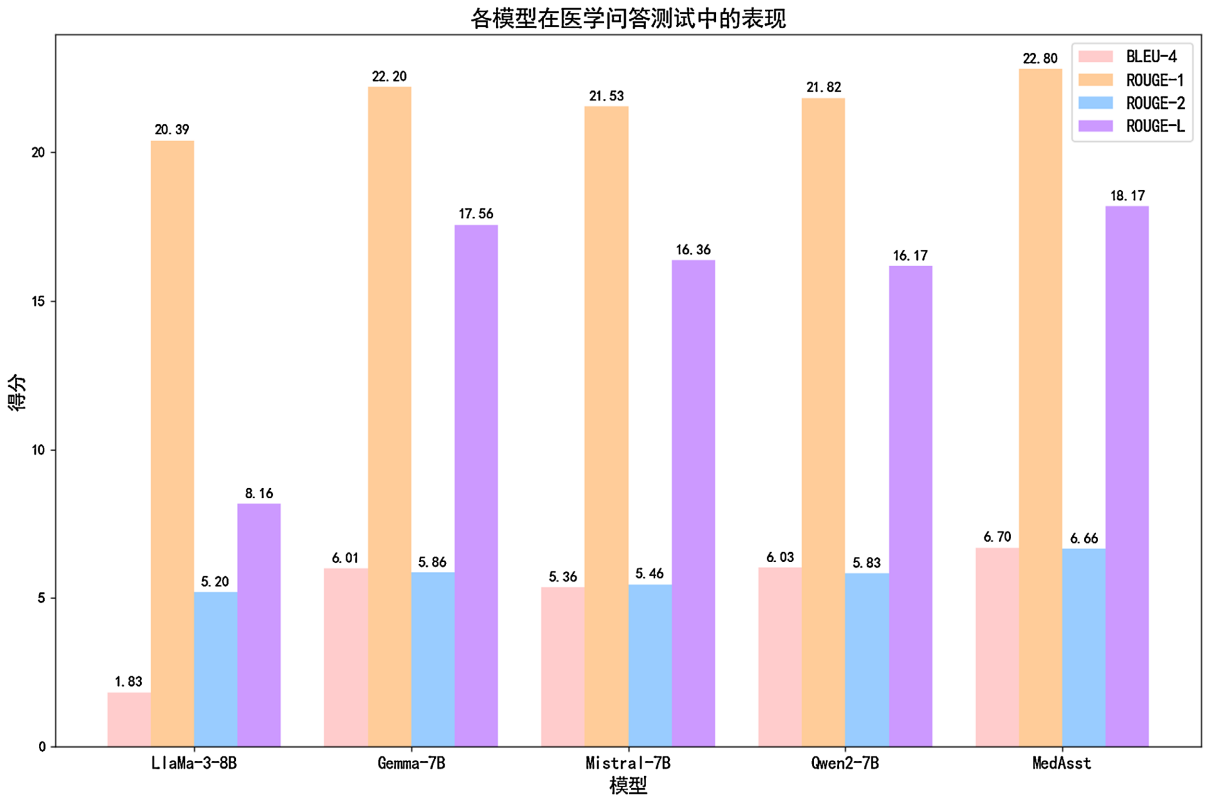

Figure 4. Scores of the models on the evaluation indicators
图 4. 各个模型在评价指标上的得分情况

清楚地看到，MedAsst 模型在所有评价指标上均优于其他基线模型，显示出其在医学问答任务中的卓越表现。相比之下，其他模型的得分较低一些。

5. 总结

本文提出了一种基于大语言模型(LLM)的中文医学对话系统模型 MedAsst，并通过对比评估了其在中文医学对话任务中的表现。我们以 Qwen2-7B 模型为基础，通过 LoRA 方法对其进行监督微调，从而构建了 MedAsst 模型。为了验证 MedAsst 的有效性，我们在医学多项选择题和自定义的医学问答数据集上对其进行了全面评测。实验结果表明，MedAsst 在 BLEU-4、ROUGE-1、ROUGE-2 和 ROUGE-L 四个指标上均优于其他基线模型，尤其在医学问答能力上展现出显著优势。通过与 LlaMa-3-8B、Gemma-7B、Mistral-7B 和未经微调的 Qwen2-7B 模型的对比，MedAsst 的优异表现证明了在特定领域进行监督微调的必要性和有效性。我们的方法不仅提升了模型在中文医学问答任务中的表现，同时也展示了大语言模型在医疗电商平台中的应用潜力。未来工作将进一步优化模型，提高其在更复杂场景中的应用效果，推动大语言模型在电商医疗领域的实际应用和发展。

参考文献

[1] 赵敏, 原超, 李朝霞. 大数据背景下医药电子商务服务模式的提升与探究[J]. 山西经济管理干部学院学报, 2018, 26(1): 45-48.

[2] 苏尤丽, 胡宣宇, 马世杰, 等. 人工智能在中医诊疗领域的研究综述[J]. 计算机工程与应用, 2024, 60(16): 1-18.

[3] Brown, T.B., Mann, B., Ryder, N., *et al*. (2020) Language Models Are Few-Shot Learners.

[4] Wang, H.C., Liu, C., Xi, N.W., Qiang, Z.W., Zhao, S.D., Qin, B. and Liu, T. (2023) HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge. https://arxiv.org/abs/2304.06975

[5] Bai, J., Bai, S., Chu, Y., *et al*. (2023) Qwen Technical Report. https://arxiv.org/abs/2309.16609

[6] Yang, A., Yang, B., Hui, B., *et al*. (2024) Qwen2 Technical Report. https://arxiv.org/abs/2407.10671

[7] Hu, E.J., Shen, Y., Wallis, P., *et al*. (2021) LoRA: Low-Rank Adaptation of Large Language Models. https://arxiv.org/abs/2106.09685

[8] 任芳慧, 郭熙铜, 彭昕, 等. 医疗领域对话系统口语理解综述[J]. 中文信息学报, 2024, 38(1): 24-35.

[9] Hayashi, Y. (1990) A Neural Expert System with Automated Extraction of Fuzzy If-Then Rules and Its Application to Medical Diagnosis. In: *Proceedings of the* 3*rd International Conference on Neural Information Processing Systems*, Morgan Kaufmann Publishers Inc., 578-584.

[10] Wong, W., Thangarajah, J. and Padgham, L. (2011). Health Conversational System Based on Contextual Matching of Community-Driven Question-Answer Pairs. *Proceedings of the* 20*th ACM international conference on Information and knowledge management*, 19-23 October 2020, 2577-2580. https://doi.org/10.1145/2063576.2064024

[11] Li, Y.S., Lam, C.S.N. and See, C. (2021) Using a Machine Learning Architecture to Create an Ai-Powered Chatbot for Anatomy Education. *Medical Science Educator*, **31**, 1729-1730. https://doi.org/10.1007/s40670-021-01405-9

[12] 颜永, 白宗文. 基于强化学习的生成式对话系统研究[J]. 数据挖掘, 2023, 13(2): 185-193.

[13] Wang, S., Wang, S., Liu, Z. and Zhang, Q. (2023) A Role Distinguishing Bert Model for Medical Dialogue System in Sustainable Smart City. *Sustainable Energy Technologies and Assessments*, **55**, Article ID: 102896. https://doi.org/10.1016/j.seta.2022.102896

[14] 马德草, 杨桂松. 基于实体知识推理的端到端任务型对话[J]. 建模与仿真, 2024, 13(3): 3212-3221.

[15] Touvron, H., Lavril, T., Izacard, G., *et al*. (2023) LLaMA: Open and Efficient Foundation Language Models. https://arxiv.org/abs/2302.13971

[16] Gemma Team, Mesnard, T., Hardin, C., *et al*. (2024) Gemma: Open Models Based on Gemini Research and Technology. https://arxiv.org/abs/2403.08295

[17] Jiang, A.Q., Sablayrolles, A., Mensch, A., *et al*. (2023) Mistral 7B. https://arxiv.org/abs/2310.06825

[18] Yang, S., Zhao, H., Zhu, S., *et al*. (2023) Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-World Multi-Turn Dialogue. https://arxiv.org/abs/2308.03549

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., *et al*. (2017) Attention is All You Need. *Annual Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, Long Beach, 5998-6008.

[20] Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebron, F. and Sanghai, S. (2023). GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 4895-4901. https://doi.org/10.18653/v1/2023.emnlp-main.298

[21] Yang, T., *et al*. (2024) Generative Large Language Models (LLMs), Question-Answering (QA), Dialogue Model, Traditional Chinese Medical QA, Fine-Tuning. https://github.com/tyang816/MedChatZH

[22] Li, J.Q., *et al*. (2023) Huatuo-26M, a Large-Scale Chinese Medical QA Dataset. https://arxiv.org/abs/2305.01526

[23] He, J., Fu, M. and Tu, M. (2019) Applying Deep Matching Networks to Chinese Medical Question Answering: A Study and a Dataset. *BMC Medical Informatics and Decision Making*, **19**, Article No. 52. https://doi.org/10.1186/s12911-019-0761-8

[24] Taori, R., *et al*. (2023) Stanford Alpaca: An Instruction-Following LLaMA Model. GitHub Repository. https://github.com/tatsu-lab/stanford_alpaca

[25] Zheng, Y., Zhang, R., Zhang, J., YeYanhan, Y. and Luo, Z. (2024) LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. *Proceedings of the* 62*nd Annual Meeting of the Association for Computational Linguistics* (*Volume* 3: *System Demonstrations*), Bangkok, August 2024, 400-410. https://doi.org/10.18653/v1/2024.acl-demos.38

[26] Model Scope. https://www.modelscope.cn/my/overview

[27] Li, H.N., *et al*. (2023) CMMLU: Measuring Massive Multitask Language Understanding in Chinese.

[28] Huang, Y., *et al*. (2023) C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. 37*th Conference on Neural Information Processing Systems* (*NeurIPS* 2023), New Orleans, 10-16 December 2023, 2-6.