基于大语言模型的医疗质量控制应用系统构建*

吴 宏[1] 胡 军[1] 陈尔真[2] 董晨杰[2] 李建华[3] 叶 琪[3]

([1] 上海市卫生健康委员会 上海 200125 [2] 上海市医疗质量控制管理事务中心 上海 200040
[3] 华东理工大学 上海 200237)

〔摘要〕 **目的/意义** 开发基于大语言模型的医疗质量控制系统，以提升医疗质量控制自动化水平与准确性。**方法/过程** 获取高质量医学数据（包括通用医疗数据和质量控制相关数据）对 PULSE 模型微调，结合自动化指标计算与检索增强生成技术，高效执行复杂质量控制任务。**结果/结论** 该系统在医疗质量控制的自动化计算与推理准确性方面性能显著提升，准确度达 93.31%，优于基座大模型和常见推理方法，具有重要的应用价值。

〔关键词〕 大语言模型；指令微调；医疗质量控制

〔中图分类号〕R-058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2025.02.002

Construction of a Medical Quality Control Application System Based on Large Language Models

WU Hong[1], *HU Jun*[1], *CHEN Erzhen*[2], *DONG Chenjie*[2], *LI Jianhua*[3], *YE Qi*[3]

[1]*Shanghai Municipal Health Commission, Shanghai 200125, China*; [2]*Shanghai Medical Quality Control Management Center, Shanghai 200040, China*; [3]*East China University of Science and Technology, Shanghai 200237, China*

〔**Abstract**〕 **Purpose/Significance** To develop a medical quality control system based on large language models (LLM), in order to enhance the automation and accuracy of quality control processes. **Method/Process** The system utilizes a diverse set of high-quality medical data, including general medical data and quality control-related data, to fine tune the PULSE model. Automated indicator calculation and retrieval-augmented generation techniques are employed to enable the system to efficiently perform complex quality control tasks. **Result/Conclusion** The system demonstrates significant improvements in automated computation and inference accuracy, achieving an accuracy of 93.31%. It outperforms baseline language models and common inference methods, offering significant implications for medical quality control.

〔**Keywords**〕 large language models (LLM); instruction fine-tuning; medical quality control

〔修回日期〕 2025-01-22

〔作者简介〕 吴宏，博士，二级巡视员，发表论文 30 余篇；通信作者：叶琪。

〔基金项目〕 国家重点研发计划（项目编号：2023YFF1204904）。

1 引言

医疗质量直接关乎人民群众的健康权益和医疗服务体验。为强化医疗质量管理，规范医疗服务行为，保障医疗安全，原国家卫生和计划生育委员会于 2016 年出台《医疗质量管理办法》。通过构建全面的医疗质量管理指标体系[1]，相关行政部门或医疗质量控制组织能够在医疗服务能力、质量与安全、运行效率及综合管理等方面实施精准的质量控制[2-3]，为医院管理和医疗质量的持续改进提供坚实的数据支撑。

既往质量控制指标主要通过各条线工作人员手动填报上传，主观性强，难以全面、准确反映医院真实医疗质量。由于多数质量控制指标针对过程性医疗质量评价，无法直接从医嘱、病案首页、收费数据等结构性数据中获取，目前仍有大量指标需人工填报。大语言模型（large language models, LLM）[4-6]在自然语言处理方面具有优势，且能减少数据标注需求，通过输入脱敏后的电子病历[7]信息及指标计算要求，可自动生成质量控制指标统计结果，大幅降低管理成本。因此，本研究提出基于大语言模型的医疗质量控制应用系统，通过微调基座大语言模型指令，提升其医疗质量管理知识理解能力，实现质量控制指标的自动化计算，并确保计算过程高置信度。

2 相关研究

2.1 医疗大语言模型

自然语言文本在日常生活中无处不在，LLM 依托大规模文本数据迅猛发展，代表性模型包括 ChatGPT[8]、Gemini[9]、LLaMA[10]，以及国内的通义千问[11]、文心一言、PULSE 等。医学领域积累了丰富的医学文献和临床数据，但其高专业性及中文语法特性增加了使用预训练语言模型进行处理的复杂性。学术界、企业界和医疗界相继提出基于中文医学文本的 LLM[12-13]，以优化医学数据处理。针对标注人员医学知识不足的问题，华佗模型[14]使用医生提供的真实数据进行微调，同时辅以 ChatGPT 生成的蒸馏数据优化。扁鹊模型[15]采用自建多轮对话数据集训练，模拟医生通过多次交流了解病情。明医模型[16]通过混合专家技术克服传统医学 LLM 局限性，适应复杂多样的医学任务，如医学自然语言处理和医师资格考试。PULSE 模型使用约 400 万个中文医学领域和通用领域指令微调数据进行调优，支持医学领域的各种自然语言处理任务，包括健康教育、医师考试问题、报告解读、医疗记录结构化以及模拟诊断和治疗等。LLM 在医学领域已应用于医疗命名实体识别、医学因果关系抽取等任务，但在医疗质量控制领域的应用仍不足。

2.2 大语言模型在医疗质量控制中的应用

多数医院仍采用传统质量控制方式，主要依赖人工抽查。面对庞大的电子病历数据[17-18]，传统质量控制方式难以确保样本代表性，易导致数据偏差。电子病历以文本形式存储，LLM 依赖大规模自然语言文本（如互联网数据）训练，在解决数据偏差问题方面展现出应用前景。吉林大学第二医院构建智能医疗质量控制系统，通过预设质量控制规则及知识图谱模型匹配，实现质量控制的自动识别[19]。南京鼓楼医院开发单病种质量控制管理平台[20]，运用自然语言处理技术将非结构化数据转化为结构化数据，结合数据仓库技术抽取与清洗数据，实现单病种质量控制自动化。上述方法主要利用 LLM 处理数据，LLM 在医疗质量控制指标计算中的应用尚处起步阶段。

3 基于大语言模型的医疗质量控制应用系统框架（图 1）

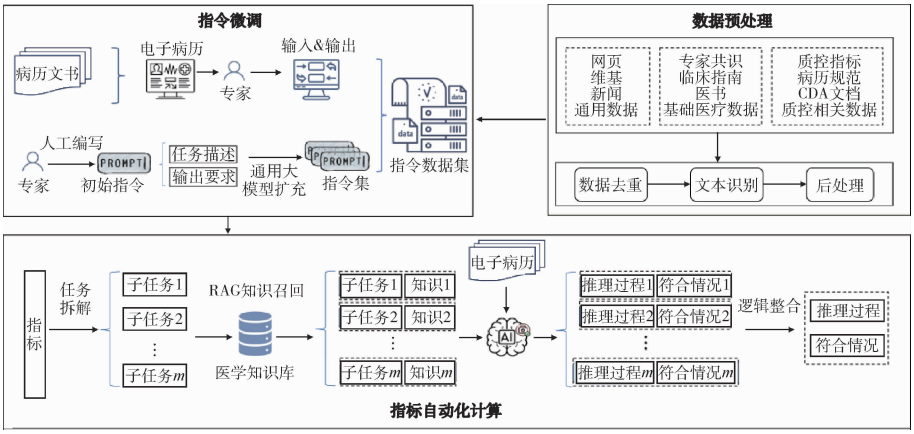

图 1　基于大语言模型的医疗质量控制应用系统

3.1　数据获取与处理

数据获取与处理模块负责采集并处理高质量医学数据，为指令微调模块提供基础。通用 LLM 缺乏专业医疗数据训练，直接用于医疗领域表现欠佳。获取高质量医学数据是提升模型性能的关键。数据主要源自临床指南、专家共识及专业质量控制指标等文本，涵盖中/西医术语、通用领域语料及各临床专业指南。为提升数据质量和有效性，采用多步处理：去重减少冗余；分割文本防止遗忘；删除无关或噪声信息。

3.2　大语言模型指令微调

指令微调[21-23]指通过多样化数据和精心设计的指令优化 LLM 性能。第 1 步：设计任务模板，明确指令内容。第 2 步：使用脱敏电子病历数据构造任务模板中的输入、输出数据。第 3 步：通过人工编写与通用 LLM 协作，构建多样化任务指令，包括通用指令、基础医疗指令、质量控制指令等不同类型。为防止模型微调过程中产生遗忘，采用“任务渐进式微调”方法，以获得性能稳定的模型。

3.3　指标自动化计算

目前各级卫生行政部门、质量控制中心通过单病种管理模式监控医疗质量。单病种对应多个质量控制指标，每个指标含多个任务。人工编写提示词耗时费力，自动化提示词生成方法可提高效率。使用 LLM 分解单病种指标为细粒度任务；使用知识召回技术从医学知识库提取相关信息；利用 LLM 生成准确描述指标的提示词；将提示词与电子病历输入模型，系统即自动判断病历是否符合指标要求。

4　基于大语言模型的医疗质量控制应用系统关键技术与结果分析

4.1　面向医疗质量控制的大语言模型指令微调

采用 PULSE 基座模型进行指令微调，兼顾成本效益与模型优化，为医疗质量控制提供可操作性和精准度更高的解决方案。

4.1.1　多样化指令构建　为确保模型在各种医疗场景下准确理解并执行质量控制指标计算任务，提升泛化能力及应用有效性，引入多样化任务指令提示。每条任务指令提示由任务描述和占位符组成。任务描述以清晰的文本形式明确传达具体任务，指导模型执行；占位符表示需动态填入的数据项，以“{}”形式呈现。为高效构建多样化任务指令集，采取人工编写与通用 LLM 协作策略。基于对不同医疗质量控制任务的深入理解，手动编写一系列初始任务指令。例如，肺癌切缘阴性率计算任务指令：“{ </start > } instruction：根据提供电子病历判断肿瘤支气管切缘是否为阴性。描述为‘切缘无肿瘤组织累及’或相似表述判为阴性；‘切缘可见肿瘤细胞浸润’或相似表述判为阳性。如为阴性，请判

断是否存在转移淋巴结。input：支气管切缘无肿瘤组织累及｛</end>｝”。在该示例中，任务描述为 instruction 以及 input，而占位符为 </start>、</end>对。结合开源数据集和人工设计模板，从真实世界医疗数据中提取关键信息，构建与实际医疗实践密切相关的多样化质量控制指标任务指令集，提升模型在医疗质控领域的应用效果。

4.1.2 指令微调 采用“任务渐进式微调”训练策略，综合分析任务难度和数据特性，优化训练顺序和训练轮次，逐步增强模型在医疗领域的专业处理能力，适应多样化医疗质量控制任务。以电子病历规范检查任务为初始训练任务，进行少轮次训练，掌握医疗文档基本处理要求和格式检查能力。当模型逐步适应初始任务，引入医疗质量控制指标计算任务，进行较多轮次训练，提高文本处理和数据分析能力，灵活适应医疗质控复杂任务。相比传统单一任务或联合任务训练，通过阶段性调整训练任务的难度和频次，确保模型在每个阶段都有针对性的能力提升，避免因任务复杂而产生不稳定现象。有序的任务安排则有助于模型建立不同任务间的联系，提升多任务学习的整体表现。各项单独任务训练完成后，整合所有任务数据，进入联合训练阶段。该阶段旨在通过综合性数据输入，使模型整合各任务学习成果，提升其在多样化医疗任务中的表现，强化模型泛化能力。模型不仅能在单项任务中保持较高的准确度，还能在面对多任务组合时展现更为一致的稳定性和适应性。

4.2 医疗质量控制指标自动化计算

不同质量控制指标计算复杂性存在显著差异，LLM 直接从电子病历抽取文本进行判断难以满足所有需求。例如，“非瓣膜性心房颤动（房颤）患者血栓栓塞风险评估率”和“冠脉介入治疗术后即刻冠状动脉造影成功率”，前者只需判断电子病历是否记录 CHA_2DS_2 – VASc 评分或栓塞风险评分，后者需综合残余狭窄程度、冠状动脉血流心肌梗死溶栓情况以及手术类型等因素进行准确计算，见表 1。为提高计算准确性和效率，将复杂指标拆分为若干独立子任务。通过逻辑规则整合子任务符合情况，判断原始质量控制指标符合情况。以“冠脉介入治疗术后即刻冠状动脉造影成功率”指标为例，其定义要求：支架术后病变残余狭窄 <20% 或单纯经皮冠状动脉腔内血管成形术后病变残余狭窄 <50%，且冠状动脉血流心肌梗死溶栓分级达到 3 级。基于此定义，可将该指标拆分为 3 个子任务。任务 1：残余狭窄 <20% 时，判断患者是否接受过支架术。任务 2：残余狭窄 <50% 时，判断患者是否接受过单纯经皮冠状动脉腔内血管成形术。任务 3：判断患者冠状动脉血流心肌梗死溶栓分级是否为 3 级。计算逻辑表达式（Task1 ∨ Task2）∧ Task3，得出该患者冠脉介入治疗术后即刻冠状动脉造影是否成功的结论。

表 1 质量控制指标要求及说明

指标名称	指标定义	指标说明
非瓣膜性心房颤动（房颤）患者血栓栓塞风险评估率	单位时间内，行血栓栓塞风险评估的非瓣膜性房颤患者数占同期非瓣膜性房颤患者总数的比例	血栓栓塞风险评估推荐采用 CHA_2DS_2 – VASc 评分
冠脉介入治疗术后即刻冠状动脉造影成功率	单位时间内，冠脉介入治疗术后即刻冠状动脉造影成功的例数占同期接受冠脉介入治疗的总例数的比例	冠状动脉造影成功指支架术后病变残余狭窄 <20% 或单纯经皮冠状动脉腔内血管成形术后病变残余狭窄 <50%，且冠状动脉血流心肌梗死溶栓分级 3 级

尽管各子任务准确描述了所需关注的细节，但其表述过于简洁，存在潜在歧义或遗漏具体应用场景的情况，可能导致 LLM 误解子任务目标，产生偏离预期的回复。采用检索增强生成（retrieval – augmented generation，RAG）技术，从事先收集的质量控制指标说明和权威医学文献中提取与子任务最相关的段落，作为额外的知识补充。

按原定义，LLM 接收的输入为 < 指标定义，指

标计算公式，电子病历 >，对自然语言中蕴含的谓词逻辑可能存在理解偏差。通过任务拆分，以 < 指标定义，指标计算公式 > 为输入，生成 < ｛子任务｝ m，电子病历 > 。再利用 RAG 技术，为每个子任务匹配相关医学知识，输出 < ｛子任务｝ m，知识，电子病历 > 。此转化减轻了 LLM 对复杂谓词逻辑的理解偏差，发挥了其在处理弱逻辑化子任务时的潜力，提高了可解释性。任务拆分与 RAG 结合增强了 LLM 的分析能力，为准确判断奠定了基础。

4.3 结果对比分析

选取 ChatGLM3 - 6B、框架基座模型 PULSE - 20B 作为对比模型，选取 Zero - shot[24]（仅提供指令）和 ICL[25]（提供完整指令、输入、输出的示例）作为对比推理方法。计算结果对比，见表 2。与 ChatGLM3 - 6B 相比，医疗质量控制应用系统参数量更大，具备更强的学习与计算能力。因此，即便采用 Zero - shot 方法，准确度也领先 ChatGLM3 - 6B 约 14.11 个百分点；结合自动化指标计算方法，准确度达 93.31%。与基座模型 PULSE - 20B 相比，医疗质量控制应用系统经医学知识微调，理解更深入和透彻。使用相同 ICL 方法，准确度提升 4.15 个百分点。综上，本研究提出的医疗质量控制应用系统在推理方法与模型性能方面均展现出潜力和优势。

表 2 指标计算结果对比

模型	方法	准确度（%）
ChatGLM3 - 6B	ICL	71.47
PULSE - 20B	ICL	84.38
医疗质量控制应用系统	Zero - shot	85.58
医疗质量控制应用系统	ICL	88.53
医疗质量控制应用系统	Ours	93.31

5 结语

本研究提出一种基于 LLM 的医疗质控应用系统，旨在提高医疗质量控制自动化水平和准确性。采用标准化数据处理和指令微调，解决医疗 LLM 中数据质量和专业性问题；通过预处理过程构建的高质量医学数据和“任务渐进式微调”策略，确保模型能应对复杂医学任务；提出自动化指标计算方法，利用任务拆分与检索增强生成技术，解决医疗指标计算中的复杂逻辑问题，提升计算效率与准确性。该系统为医疗质量控制的自动化与智能化发展，以及 LLM 在医疗领域的深度应用提供了可行的技术路线，将在更多医疗机构广泛应用，推动医疗质量管理现代化。

作者贡献：吴宏、胡军负责论文撰写；陈尔真、董晨杰负责数据采集与整理、文献调研；李建华、叶琪负责研究设计、论文审核。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 王飞，张春燕，张付静，等．“大质控”理念下医疗质量管理体系的构建与实践探索［J］．中国卫生标准管理，2023，14（13）：50 - 54.

2 LEE R I. The fundamentals of good medical care：an outline of the fundamentals of good medical care and an estimate of the service required to supply the medical needs of the United States［M］. Chicago：University of Chicago Press，1933.

3 黄素华．论医疗机构单病种质量控制的研究［J］．广西医学，2004，26（3）：443 - 444.

4 刘少堃，何仲廉，李彬，等．基于大模型的电子病历自动生成系统的设计与应用探讨［J］．中国数字医学，2024，19（8）：8 - 13.

5 陈依萍，周露，李亦学．基于大模型的电子病历与 CDA 标准模式映射方法研究［J］．中国卫生信息管理杂志，2023，20（6）：867 - 874.

6 浙江大学医学院附属第二医院：在电子病历系统中嵌入人工智能大模型［J］．中华医学信息导报，2024，39（15）：9.

7 郝宁．信息化管理在医疗质量控制中的应用［J］．邯郸职业技术学院学报，2023，36（4）：34 - 37.

8 BROWN T B，MANN B，RYDER N，et al. Language models are few - shot learners［EB/OL］.［2024 - 12 - 25］. https：//arxiv.org/abs/2005.14165.

9 GEMINI TEAM，ANIL R，BORGEAUD S，et al. Gemini：a family of highly capable multimodal models［EB/OL］.［2024 - 12 - 25］. https：//arxiv.org/abs/2312.11805.

10 TOUVRON H，LAVRIL T，IZACARD G，et al. Llama：open and efficient foundation language models［EB/OL］.

[2024 - 12 - 25]. https://arxiv.org/abs/2302.13971.
11 BAI J, BAI S, CHU Y, et al. Qwen technical report [EB/OL]. [2024 - 12 - 25]. https://arxiv.org/abs/2309.16609.
12 阮彤, 卞俣昂, 余广涯, 等. 医学大语言模型研究与应用综述 [J]. 中国卫生信息管理杂志, 2023, 20 (6): 853 - 861.
13 刘泓泽, 王耀国, 唐圣晟, 等. 医学大语言模型的应用现状与发展趋势研究 [J]. 中国数字医学, 2024, 19 (8): 1 - 7, 13.
14 ZHANG H, CHEN J, JIANG F, et al. HuatuoGPT, towards taming language model to be a doctor [C]. Singapore: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023.
15 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 [EB/OL]. [2024 - 12 - 25]. https://arxiv.org/abs/2310.15896.
16 LIAO Y, JIANG S, WANG Y, et al. MING - MOE: enhancing medical multi - task learning in large language models with sparse mixture of low - rank adapter experts [EB/OL]. [2024 - 12 - 25]. https://arxiv.org/abs/2404.09027.
17 宋雪君. 电子病历在医疗质量控制管理中的应用与研究 [J]. 现代计算机, 2023, 29 (6): 88 - 90.
18 冯园园, 苏源, 来庆玲, 等. 基于 3 级质控管理的电子病历系统流程优化研究 [J]. 医学信息学杂志, 2023, 44 (4): 73 - 77.
19 周芮, 陈彦东, 宿明. 基于大数据和 AI 的电子病历自动质控系统构建 [J]. 医学信息, 2022, 35 (4): 13 - 16.
20 刘晓娇, 黄春芳, 朱玉婷, 等. 基于临床数据中心的单病种质控管理平台的实践 [J]. 中国医疗设备, 2023, 38 (2): 90 - 94, 130.
21 蔡子杰, 方荟, 刘建华, 等. 基于大型语言模型指令微调的心理健康领域联合信息抽取 [J]. 中文信息学报, 2024, 38 (8): 112 - 127.
22 DING N, QIN Y, YANG G, et al. Parameter - efficient fine - tuning of large - scale pre - trained language models [J]. Nature machine intelligence, 2023, 5 (3): 220 - 235.
23 LONGPRE S, HOU L, VU T, et al. The flan collection: designing data and methods for effective instruction tuning [C]. Honolulu: The 40th International Conference on Machine Learning (ICML23), 2023.
24 WEI J, BOSMA M, ZHAO V Y, et al. Finetuned language models are zero - shot learners [C]. Virtual: The Tenth International Conference on Learning Representations, 2022.
25 WIES N, LEVINE Y, SHASHUA A. The learnability of in - context learning [C]. New Orleans: Advances in Neural Information Processing Systems, 2023.

(上接第 7 页)

究阶段向临床应用迈进。通过多模态数据融合、因果推理的引入、模型可解释性的提升，大模型不仅能提升诊断效率和准确性，还将为数智医院的建设和医疗服务模式的重构提供强有力的技术支持。

作者贡献：闾海荣负责论文构思与撰写；江瑞、张学工参与医学大模型技术创新讨论并提出建议；王存库参与医疗服务模式重构讨论并提出建议。
利益声明：所有作者均声明不存在利益冲突。

参考文献

1 刘哲, 石钰, 林延带, 等. 智能医学的现状与未来 [J]. 科学通报, 2023, 68 (10): 1165 - 1181.
2 DeepSeek - V3 technical report [EB/OL]. [2025 - 01 - 10]. https://github.com/deepseek - ai/DeepSeek - V3/blob/main/DeepSeek_V3.pdf.
3 DeepSeek - R1: incentivizing reasoning capability in LLMs via reinforcement learning [EB/OL]. [2025 - 01 - 10]. https://github.com/deepseek - ai/DeepSeek - R1/blob/main/DeepSeek_R1.pdf.
4 郭华源, 刘盼, 卢若谷, 等. 人工智能大模型医学应用研究 [J]. 中国科学: 生命科学, 2024, 54 (3): 482 - 506.
5 肖仰华, 徐一丹. 大规模生成式语言模型在医疗领域的应用: 机遇与挑战 [J]. 医学信息学杂志, 2023, 44 (9): 1 - 11.
6 相东升, 赖宇, 陈浩, 等. 智能时代下的空间数据科学: 基础模型研究与行业应用 [J]. 无线电工程, 2025, 55 (1): 184 - 195.