DOI 10.15302/J-SSCAE-2024.07.020

医疗大模型技术及应用发展研究

陈晓红[1,2,3,4]，**刘浏**[1,2,3,4]，**袁依格**[2*]，**王俊普**[5,6]，**李大元**[1]，**邱建华**[7]

（1. 中南大学商学院，长沙 410083；2. 湘江实验室，长沙 410205；3. 湖南工商大学前沿交叉学院，长沙 410205；
4. 湖南工商大学管理科学与工程学院，长沙 410205；5. 中南大学湘雅医院，长沙 410008；
6. 中南大学基础医学院，长沙 410008；7. 智慧眼科技股份有限公司，长沙 410036）

摘要：医疗大模型基于深度神经网络架构进行复杂医疗数据的高效处理与模式识别，为智慧医疗提供新型的决策支持；需要系统分析医疗大模型技术及应用情况，以精准把握医疗大模型的发展方向、精准应对面临的发展挑战，进而基于医疗大模型提升医疗文本、医学图像、药械研发、医学教育等方面的能力。本文梳理了医疗大模型的技术范式与应用场景，剖析了由基础层、模型层、应用层、公共模块构成的医疗大模型技术体系，覆盖评价指标体系构建、数据集范围与题型、模型对齐方法、模型评测平台的医疗大模型评测体系，辨识出医疗大模型应用存在的数据安全、技术风险、落地挑战、伦理道德等方面的难点。为此建议，发挥政府引导优势、保障数据安全，加快基础理论研究、突破技术风险，强化应用场景牵引、缓解落地挑战，建立健全监管机制、规范伦理道德，完善公共服务体系、营造创新生态，以加快医疗大模型创新应用，推动我国智慧医疗的高端化、智能化、绿色化发展。

关键词：医疗大模型；多模态数据；预训练微调；提示工程；技术体系；评测体系

中图分类号：TP18；R-1　　**文献标识码：**A

Technology and Application Development of Medical Foundation Model

Chen Xiaohong[1,2,3,4], **Liu Liu**[1,2,3,4], **Yuan Yige**[2*], **Wang Junpu**[5,6], **Li Dayuan**[1], **Qiu Jianhua**[7]

(1. School of Business, Central South University, Changsha 410083, China; 2. Xiangjiang Laboratory, Changsha 410205, China;
3. School of Advanced Interdisciplinary Studies, Hunan University of Technology and Business, Changsha 410205, China;
4. School of Management Science and Engineering, Hunan University of Technology and Business, Changsha 410205, China; 5. Xiangya Hospital of Central South University, Changsha 410008, China; 6. School of Basic Medical Sciences, Central South University, Changsha 410008, China; 7. Athena Eyes Co., Ltd., Changsha 410036, China)

Abstract: The medical foundation model performs efficient processing and pattern recognition of complex medical data based on a deep neural network architecture, providing a new type of decision support for intelligent medical care. It is necessary to systematically analyze the technologies and application of the medical foundation model, thus to identify the development directions and challenges and improve the capabilities of the medical care sector in medical text writing, medical image recognition, medical equipment research and development, and medical education using the medical foundation model. This study sorts out the technology paradigm and application scenarios of the medical foundation model and proposes a technology system and an evaluation system for the model. The

收稿日期：2024-07-01；**修回日期：**2024-08-29

通讯作者：*袁依格，湘江实验室副研究员，研究方向为大模型与智慧医疗；E-mail: immyyuan23@163.com

资助项目：中国工程院咨询项目“全球未来产业发展趋势及湖南未来产业布局研究”(2024-DFZD-39)；湘江实验室项目(23XJ01008, 23XJ03001)

本刊网址：www.engineering.org.cn/ch/journal/sscae

technology system is composed of a base layer, a model layer, an application layer, and a common module. The evaluation system involves the establishment of an evaluation index system, dataset range and question types, model alignment methods, and model evaluation platforms. Moreover, application challenges of the medical foundation model are identified in terms of data security, technical risks, implementing challenges, and ethics. Furthermore, the following countermeasures are suggested: (1) ensuring data security through government guidance, (2) accelerating basic theoretical research to address technic risks, (3) focusing on application scenarios to cope with implementing challenges, (4) improving the ethics regulating mechanism, and (5) perfecting the public service system to create an innovation ecosystem, thereby accelerating the innovative development of the medical foundation model and promoting the high-end, intelligent, and green development of intelligent medical care in China.
Keywords: medical foundation model; multimodal data; pre-training and fine-tuning; prompt engineering; technology system; evaluation system

一、前言

健康是居民幸福的重要基础。深入开展健康中国行动和爱国卫生运动，加强健康、养老等民生科技研发应用等成为公共管理的重点任务之一；推动人工智能（AI）生成内容大模型绿色创新发展、以新一代信息技术为引擎加快形成新质生产力、健全算力网服务生态体系[1]等“人工智能+”发展要素成为社会关注焦点。加快推进“人工智能+”行动、布局医疗大模型创新发展，是更好服务居民健康、实现智慧医疗的必经之路[2,3]。

当前，AI领域的研究与应用热点是大模型，相关产业进入了蓬勃发展阶段[4-6]。其中，医疗大模型基于先进的数据分析和处理能力，正在显著增强医疗文本、医学图像、药械研发、医学教育等方面的发展水平[7~13]。例如，医疗大模型能够处理大量的历史病例数据，从中提取有价值的信息，为医生提供更精确的诊断与治疗建议；能够理解和生成复杂的医学术语与临床文档，自动生成病例报告、研究摘要等，减少医疗专业人员的文档编写负担；在处理文本数据之外，能够整合医学图像、实验室检测结果等数据，提供更为全面的诊断支持[2,3,14~17]。尽管如此，医疗大模型在智慧医疗领域的应用仍面临着一系列挑战[7,8,10,18]。例如，数据是制约智慧医疗发展的最大壁垒[19]，疾病诊断的智能化程度偏低[20]；幻觉问题、实时性问题构成医疗大模型普及应用的直接挑战；医疗大模型的透明度不足、可解释性差[7]，直接影响临床应用效果和接受程度；医护人员和患者对医疗大模型的基础知识及工作原理了解不足[2,3]。

发展医疗大模型是智慧医疗领域构建新质生产力[21,22]的重要举措，相关研究兼具学术探索与实践应用价值。本文围绕医疗大模型这一前沿热点与应用要点，梳理技术范式与应用场景、剖析技术体系和评测体系、探讨应用难点并提出发展建议，以期丰富医疗大模型的理论认知和实践启示，推动我国智慧医疗的高端化、智能化、绿色化发展。

二、医疗大模型的技术范式与应用场景

AI经历了漫长的孕育期，最早可追溯到演绎逻辑，随着人类对智能认识的不断深入而持续演进。1980年，机器学习成为AI发展的独立分支，以“从数据中获取经验”克服了基于规则建模的困境。2006年，深度学习概念正式形成，针对特定应用场景专门训练的深度神经网络（即小模型）开始涌现。当前，AI进入了大数据驱动的新发展阶段，生成式AI正在获得重大突破并快速演进，大模型开始涌现[2,3]。

（一）医疗大模型的技术范式

大模型是“大数据+大算力+强算法”的深度神经网络模型，在知识、数据、算法、算力等关键要素的推动下呈现爆发式发展，从自然语言处理逐步扩展和迁移到计算机视觉、多模态数据融合等信息技术方向[6,23~26]。转换器（Transformer）确定了主流的大模型架构[11,13]：首先进行模型训练，从大量数据中“学习”出一些规则并生成模型；然后进行模型推理以解决实际问题。大模型通过“预训练+微调”，增强了AI的泛化性和通用性，构建了AI发展的新范式，成为迈向通用AI的重要技术路径[13]；基于海量无标注数据进行预训练，提升模型前期学习的广度、深度、知识水平，从而低成本、高适应性地在后续（下游）任务中应用[27]。当模型参数规模足够大时，大模型会出现“智能涌现”现象，实现“少样本”“零样本”学习和推理。整体来看，大模型是实现多种AI应用的通用载体，有望成为未来

的新型基础设施并赋能诸多行业[7,23,24,28~31]，其中以落地应用、价值实现为重点[2,3]。

在上述背景下，基于大模型构建的医疗大模型为智慧医疗的发展注入新动力[12]。医疗大模型作为面向复杂、开放的智慧医疗场景的基础大模型，蕴含大数据、大算力、大参数等关键要素，具有“智能涌现”能力、良好的泛化性与通用性[26]。基础大模型适应智慧医疗领域的特定任务，在具体实施上有多种策略[9]；主要形成了4种技术范式，由易到难分别是提示工程、各种指令/任务微调、继续训练通用大模型、从头开始预训练（见表1）[32~34]。当有大量的数据、计算资源、专业知识时，可以采用从头开始或继续训练通用大模型的方式开发智慧医疗领域的特定模型，但成本较高。各种指令/任务微调、提示工程更具成本优势，对预训练后的大模型进行面向智慧医疗领域知识的微调训练（基于下游特定任务上的小规模有标注数据进行二次训练）[35]或使用提示工程[26]，即可高水平地完成多类智慧医疗应用场景中的任务，实现通用的智能能力[10]。不同技术范式可以单独或组合使用，将更好适应智慧医疗场景的需求[2,3]。

（二）医疗大模型的应用场景

医疗大模型支持提升医疗智能化水平，推动智慧医疗的创新和进步[11,18,36]。高校、科研院所较多基于开源模型进行微调，以LLaMA（美国元宇宙平台公司发布的产品）为底座的模型发展势头良好；大型科技公司积极自研通用大模型，追求平台赋能智慧医疗行业发展；药品与医疗器械类企业具有行业数据优势，多以调用接口或基于开源模型自研的方式构建医疗大模型。本研究着重分析医疗文本、医学图像、药械研发、医学教育[2,3,13]等代表性应用场景。

1. 医疗文本

医疗大模型辅助临床文档生成及医疗文本结构化，为医疗专业人员提供准确、快速、个性化的诊断和治疗建议，主要涉及：将医生口述或非结构化的记录转换为结构化的电子病历；从大量医学文献中提取关键信息，支持研究人员快速了解研究趋势和新发现；自动化生成病例报告和研究摘要，减少医疗专业人员的文档负担；为患者和医疗专业人员提供基于证据的医疗信息，辅助临床决策；支持将临床服务转化为标准化的医疗计费代码。例如，香港中文大学（深圳）、深圳市大数据研究院的研究团队开发了华佗GPT（生成式预训练）医疗大模型，结合聊天生成预训练转换器（ChatGPT）、医生回复的数据，使模型具备像医生一样的诊断能力以及提供有用信息的能力，同时保持面向用户的流畅交互与内容丰富性。

2. 医学图像

医学图像和医疗大模型相结合，优势在于提升疾病诊断的准确性及效率，为医生提供更全面的信息以辅助确定诊疗方案。医学图像成像方式独特，主要有如X线、计算机断层扫描、核磁共振、超声、病理等形式，相关数据量约占医疗数据总量的90%。医疗大模型辅助医生对发现的病变进行差异化诊断，提供额外的图像信息支持诊断，增强医生的临床决策、随访建议等能力。医疗大模型具备自动检测异常或疾病迹象等能力，有利于医生快速诊断，也可辅助医生学习和提高技能。例如，上海联影智能医疗科技有限公司、复旦大学附属中山医院联合开发了全病程智医诊疗大模型，涉及多模态、涵盖多病种，基于医学影像、临床数据等维度的信息进行疾病的早期预测、精准诊断、疗效评估。

表1　智慧医疗领域开发与应用大模型的4种范式

技术范式	主要步骤	典型案例
提示工程	设计并构建输入提示，用于控制大模型的输出，以提高生成文本的准确性和可靠性；涉及硬提示、软提示两种主要技术	GeneGPT
各种指令/任务微调	针对特定任务的微调，常用于为特定下游任务调整较小的模型；主要包括指令微调、人类反馈强化学习微调	Med-PaLM
继续训练通用大模型	从现有通用大模型的检查点开始进行参数初始化，然后在智慧医疗语料库上进一步预训练模型，实现填充或自回归大模型的训练目标	BioBERT
从头开始预训练	在大型智慧医疗语料库上采用随机初始化的参数预训练大模型，实现填充或自回归大模型的训练目标	BioGPT AlphaFold2

3. 药械研发

医疗大模型和药械研发相结合，有利于提高药械研发效率、降低综合成本、提供个性化治疗方案，成为药械研发的重要趋势。医疗大模型服务药械研制的全流程，覆盖药物发现、临床前研究、临床试验、注册申请、上市后再评价等主要环节。面向药械企业的医药信息情报、行业知识问答大模型也已出现，进一步赋能药械产品创新。在制药领域，新药研发面临成本高、周期长、成功率低的难题；医疗大模型基于大数据源，具备更高的预测能力，支持药物设计、筛选、优化、验证等关键环节的效率与效果的双提升，匹配制药企业加快研发进度、降低研发成本的目标。例如，美国英伟达公司发布的 BioNeMo AI 工具，可将预训练的大模型、预训练框架与任务微调、优化推理相结合，直接用于加快药物研发。

4. 医学教育

医疗大模型用于模拟不同类型病人与医生进行对话，创造出提高医学生知识、技能、能力的新机会。医疗大模型充当虚拟患者或虚拟测试对象，为医学生模拟临床环境、提出问题、解释响应并提供反馈，使医学生在安全和受控的环境中练习临床推理、决策、沟通等技巧；作为虚拟导师，提供即时反馈和个性化指导，辅助医学生将医学理论知识应用于接近现实世界的情境；用于课程开发、个性化教学计划及材料准备、学生测试与评估、医学写作协助等，辅助医学生精准培养。例如，美国 Hippocratic AI 公司发布的平台，可模拟各种类型的病人在差异化的性格背景下以不同的语气与医生对话，辅助医学生提升医学知识，为医学生的临床诊断技能提供反馈与评价。

三、医疗大模型技术体系

Transformer 架构的提出（2017 年），标志着多层感知机、循环神经网络、卷积神经网络等经典处理结构开始遭到摒弃；利用自注意力机制得到输入和输出之间的全局依赖关系，捕捉长距离的依赖关系和上下文信息，具有并行、灵活、可拓展的特点，成为大模型的主流算法架构基础，为发展医疗大模型提供了关键支撑。近年来，国际科技企业推出了 Med-PaLM 医疗大模型和自建的医疗模型测评数据集。从主流的研究与应用进展来看，医疗大模型赋能智慧医疗领域的生态架构主要涉及“上游基础层－中游模型层－下游应用层”（见图 1）。医疗大模型技术体系由基础层、模型层、应用层、公共模块构成：基础层提供必要的数据和算力支持，模型层通过算法和模型将海量的数据转化为有用的信息，应用层将有价值的信息转化为医疗决策和行为，公共模块确保整个技术体系的安全性、可靠性

图 1　医疗大模型+智慧医疗生态架构

和伦理性。

（一）基础层

基础层是支撑医疗大模型研发和应用落地的基石，分为算力基础设施、数据基础设施：前者包括通用计算芯片、AI计算加速芯片、计算服务器、存储服务器、通信网络、云计算服务、容器/虚拟化技术等；后者涉及数据采集和标注、生物信息学数据库、专病数据库、多模态医疗数据库资源、医疗知识图谱等。基础层的主要作用是数据收集、存储和处理，相应数据关联电子病历、医学影像、临床试验、健康记录等来源，其多样性和复杂性构成了医疗大模型构建与应用的主要挑战。此外，基础层需要确保数据的质量、完整性、安全性，通常使用云计算、大数据技术以及相关的数据治理策略以达到这些目标。

（二）模型层

模型层是医疗大模型技术体系的核心层级，用于构建研发管理和运维体系。模型研发指在一定的编程环境（语言）、算法框架、开发平台及工具的基础上，构建大语言模型、视觉大模型、图学习大模型、语言条件多智体大模型、多模态大模型、生物计算大模型等，确保医学自然语言处理、医学图像识别、医学语音语义识别、生物分子设计、预测分析等任务的可靠执行。模型管理和运维主要包括系统管理、接口管理、数据处理等，涉及机器学习和AI算法的应用。开发可理解和处理医疗数据的算法与模型是直接目标，而基于大量标注的医学数据进行训练是前提条件，如此才能实现对医疗信息的理解、推理和预测。

（三）应用层

应用层是医疗大模型技术体系的上层，实现“药械医健”多场景触达。医疗大模型直接赋能医学和药械研发，相关应用起步早、发展快、成果多；在医学影像、医疗问答与智能问诊、辅助诊疗与临床决策支持、医学信息提取及生成、行政流程优化、个人健康管理、医保与商业保险、医学教育等方面的应用价值逐步显现，相关场景应用探索加速。医疗大模型直接面向患者、医护人员等，支持医疗信息管理系统优化，辅助医生和患者作出更优的决策，提升医疗数据利用效率和医疗服务质量；将为智慧医疗领域的诸多环节带来更加精确、高效、人性化的服务，提升整个医疗系统的服务和运行质量。

（四）公共模块

公共模块是医疗大模型技术体系的保障，重在提高模型输出结果的可解释性和安全性，为医疗行业带来可靠和安全的AI解决方案；主要分为监管治理、安全能力两部分：前者包含政策与法规、验证及评测，后者包括通用安全能力、特定安全能力。依据严格设立的政策与法规，规范医疗大模型的开发与应用，保障医疗数据的隐私和安全。通过验证与评测机制，确保医疗大模型的可靠性和准确性，提高临床实践中的应用可信度。通用安全能力涵盖数据加密、身份认证等方面，用于保障医疗大模型的信息安全。特定安全能力针对医疗领域中数据的敏感性、保密性等特殊需求，提供定制化的安全解决方案。

四、医疗大模型评测体系

当前，医疗大模型的全方位评测涉及范围广、工作量大、成本高昂。例如，数据标注工作量很大，许多维度的评测基准仍属空白、有待构建；自然语言具有多样性和复杂性，无法形成标准答案或者标准答案不具唯一性，导致相应的评测指标难以量化；主要在学术研究数据集、为人类设计的医学考试上开展评估，但此类数据集具有局限性，不能确切反映大模型在真实医疗场景中的表现。临床实践、正确回答各类主/客观测试题及考试问题，二者并不等同，寻求适当的基准以衡量大模型的临床应用效能具有挑战性。需要深入探讨医疗大模型的评估方法，建立反映真实需求、具有多样化特征的基准数据集，发展优化的医疗大模型动态评测体系，由此验证医疗大模型实践应用的有效性与实用性[2,3,7]，更好适应不断变化的医疗环境和需求。

（一）评价指标体系构建

医疗大模型的现有评估没有考量人类和AI之间新型合作的价值[2]。明确并定义大模型应用于智慧医疗的价值，围绕相关价值主张设立对应的评价指标，才能更好验证医疗大模型的应用价值。构建

评价指标体系，建立明确机制以持续监控医疗大模型的性能和影响，根据新的数据和知识定期进行模型的更新与改进。参考MedBench评测体系，本研究构建了更适合中文数据及语境的医疗大模型评价指标体系（见表2）；不仅可以衡量医疗大模型的基础性能，而且能够衡量医疗大模型的高级性能、扩展性能、用户体验等。

在医疗大模型的评测过程中，基础性能指标是评价的初始与核心部分，包括准确性、特异性、灵敏度（召回率）、精确度、F1分数、受试者工作特性曲线下面积（AUC-ROC）等。医疗大模型的高级性能评估涉及更专业的指标，如医学语言理解、

表2　医疗大模型评价指标体系

指标类型	具体指标	指标说明	测试方法	评分标准
基础性能	准确性	正确预测的总数（含真阳性、真阴性）占总样本数的比例	（真阳性+真阴性）/总样本数×100%	准确性（%）
	特异性	正确识别的阴性样本占所有实际阴性样本的比例	真阴性/（真阴性+假阳性）×100%	特异性（%）
	灵敏度（召回率）	正确识别的阳性样本占所有实际阳性样本的比例	真阳性/（真阳性+假阴性）×100%	灵敏度/召回率（%）
	精确度	预测为阳性的样本中，实际为阳性的比例	真阳性/（真阳性+假阳性）×100%	精确度（%）
	F1分数	衡量模型的平衡性能	2×（精确度×召回率）/（精确度+召回率）	F1分数（0~1）
	AUC-ROC	衡量分类模型性能	计算AUC-ROC	AUC值（0~1）
高级性能	医学语言理解	涵盖医学信息抽取、医学术语标准化、医学文本分类等测试	结构化数据，常见的单项选择题和部分限定域问答，自由文本常见的开放域问答，为综合得分	评分（1~10）
	医学语言生成	包含短对话电子病历生成和长对话电子病历生成任务		评分（1~10）
	医学知识问答	包括医学考试、医学咨询、专科问答、导诊与轻问诊等任务的测试		评分（1~10）
	复杂医学推理	覆盖临床问诊、医学诊断、治疗方案等任务		评分（1~10）
	医疗安全和伦理	涵盖医学伦理考题、药物禁忌等任务		评分（1~10）
扩展性能	鲁棒性	在面对输入医疗数据中的异常、噪声或故意的对抗性攻击时，仍能保持稳定性能的能力	在输入中引入噪声或扰动（如随机变化、对抗样本），然后测量模型输出的一致性和稳定性	鲁棒性（%）
	公平性	在处理不同患者群体时，不会产生偏见或歧视	使用统计测试来比较模型输出的一致性	公平性（%）
	资源利用率	计算资源（如中央处理器、图形处理器（GPU）、内存）利用效率和能源消耗的需求	包括计算所需时间、使用内存量、电力消耗等，为综合得分	评分（1~10）
	成本效益率	评估模型的经济可行性、成本效益	计算模型部署、运行维护的总体成本以及产生的医疗效益，进行成本效益分析，为综合得分	评分（1~10）
用户体验	可解释性	评估模型决策过程的透明度、可理解性	通过专家评审、用户调查测量，为综合得分	评分（1~10）
	用户接受度	评估医疗专业人员和患者对模型的信任度与接受度	通过问卷调查、访谈、用户反馈收集测量，为综合得分	评分（1~10）

医学语言生成、医学知识问答、复杂医学推理、医疗安全和伦理等。扩展性能指标是医疗大模型评价的实用性关键指标，如鲁棒性、公平性、资源利用率、成本效益率等。用户体验指标主要用于评估医疗大模型决策过程的透明度与可理解性、医疗专业人员和患者对模型结果的信任度与接受度等。

（二）数据集范围与题型

现有的医疗大模型评测体系多采用单轮对话、选择题等形式，而对大模型的整体理解能力和生成能力考量不足；主要关注大模型生成内容与预期答案的匹配程度，忽略了对话类医疗大模型的交互特性；又以封闭式问题居多且输出简洁（如多项选择），无法反映大模型聊天助手的典型使用情况[2]。此外，对医疗大模型的伦理合规性考量不足，而相应开发与应用需符合伦理标准（如尊重患者权利、保护隐私、促进社会福利等）。为此，可在数据集的范围选定和题型设计方面进行针对性处理：综合多样化医疗机构的数据集，收集不同地区和背景的患者数据；在传统的单轮对话和选择题以外，增加开放式问题并模拟患者对话，评估大模型的交互特性与持续对话能力。例如，设计连续的病例管理任务，模拟真实的医患交互场景，测试医疗大模型在接收连续信息时的适应性和准确性，更全面地评估医疗大模型的综合能力（见表3）。

理想的医疗大模型应能连贯地跟进患者的描述和所提问题，逐步深入地收集病情信息；在每次交互后，应提供逻辑上连贯、医学上可靠的回应，利于患者的进一步交互以及医生在合适的时机介入；应能提供有价值的医疗信息、合理的初步诊断，进而给出基于当前信息的医疗建议。

表3　医疗大模型连续交互流程

环节	描述
病例背景	患者，45岁，男性，反映近1个月来反复出现头痛
初始对话	患者描述头痛的具体情况，如痛感位置、持续时间和强度
模型询问	医疗大模型基于病情信息提出相关的诊断性问题，如询问症状、生活习惯等
患者反馈	患者根据医疗大模型的询问，提供进一步的详细信息
综合判断	医疗大模型根据收集的信息，提出可能的诊断建议

（三）模型对齐方法

医疗大模型在处理复杂的医疗数据和决策时，需要确保模型输出与医疗行业规范及价值观的一致性。采用对齐方法辅助大模型更好地理解和生成与医疗相关的自然语言，能够提升跨模态交互的质量[2,3]。具有代表性的模型对齐方法有基于提示、基于监督微调（SFT）两类。

1. 基于提示的对齐方法

基于提示的对齐方法属于优化和调整医疗大模型输出的技术，通过精心设计的提示来引导大模型生成更准确、更相关的回答，支持大模型更好地理解和处理医疗相关的复杂问题，从而提高模型输出的准确性和可靠性。这种对齐方法不改变大模型的结构和权重，而是构造特定的输入提示来达到类似于微调的效果，设计合适的提示是关键。相关提示可以是一系列的问题、描述、指令，用于引导大模型在特定任务中采取正确的行动、生成恰当的输出。对医疗大模型而言，设计提示时应结合医学领域的专业知识和实际需求，以确保大模型准确理解并响应相关的医疗问题。当然，基于提示的对齐方法还需进一步的研究和验证，以不断完善和优化在医疗大模型中的应用效果；也可细分为数种不同的策略，如基于示例的提示、基于知识的提示、基于对比学习的提示、基于多任务学习的提示等。在设计提示时应避免泄露敏感信息，保障医疗数据的隐私性和安全性。

2. 基于监督微调的对齐方法

基于SFT的对齐方法属于在预训练的大模型基础上进行微调的技术，可适应特定的任务或领域（如医疗领域）。需要以预训练的医疗大模型作为基础，即基础模型具备一定的通用医学知识和理解能力；收集一系列带有正确答案或标签的医学任务数据，构建用于微调的数据集；利用微调数据集对预训练模型进行有监督的微调，使大模型逐渐适应并准确完成医学领域的各项任务。这种对齐方法关键在于利用带标签的数据指导大模型的学习过程（见图2），将改变模型结构，对参数量偏少的大模型微调效果更好；也存在一定的局限性，如大量带标注的医疗数据不易获取或者成本过高，微调过程的稳定性和可解释性依然不强。为了提高微调效果并减少对标

注数据的依赖，可以采用优化的基于SFT的对齐方法，如基于任务自适应的SFT、基于领域自适应的SFT、基于少样本学习的SFT、基于主动学习的SFT、基于知识蒸馏的SFT等。

（四）模型评测平台

大模型评测平台提供标准化、系统化的工具，用于评估和比较不同大模型的性能与可靠性，支持研究人员和开发者识别大模型的优势及不足、指导大模型的优化及改进[2,3]。目前，北京智源人工智能研究院推出的FlagEval、上海人工智能实验室牵头推出的MedBench是具有代表性的医疗大模型的开放评测平台。

FlagEval大模型评测平台具有细化的能力框架，直接提供算力并统一基于SFT的对齐方法；构建了“能力－任务－指标”三维评测框架，精细刻画基础模型的认知能力边界，以可视化方式呈现评测结果；成为科学、公正、开放的评测基准与工具集，支持研究人员全面评估基础模型和训练算法的性能。目前，该平台包含6类评测任务、约30个评测数据集（含自建的主观评测数据集、与高校共建的评测数据集），拥有评测题目约1×10^5道，可全面评估大模型的语言理解、文本生成、知识推理等能力；新增了大模型的鲁棒性评测功能，用于考察模型对输入文本的抗干扰能力。

MedBench平台面向中文医疗大模型的开放评测需求，提供科学、公平、严谨的评估服务，持续更新和维护高质量的医学数据集，成为医疗大模型评测的权威平台。基于真实的医学考试题目和临床案例，模拟我国医生的教学过程和临床经验，成为评估医疗大模型掌握知识与推理能力的可靠基准。目前，该平台包含15项评测任务、20个评测数据集，拥有评测题目约3×10^5道，涵盖医学语言理解、医学语言生成、医学知识问答、复杂医学推理、医疗安全与伦理等能力维度，能够考察大模型在处理医学相关问题时的综合性能。

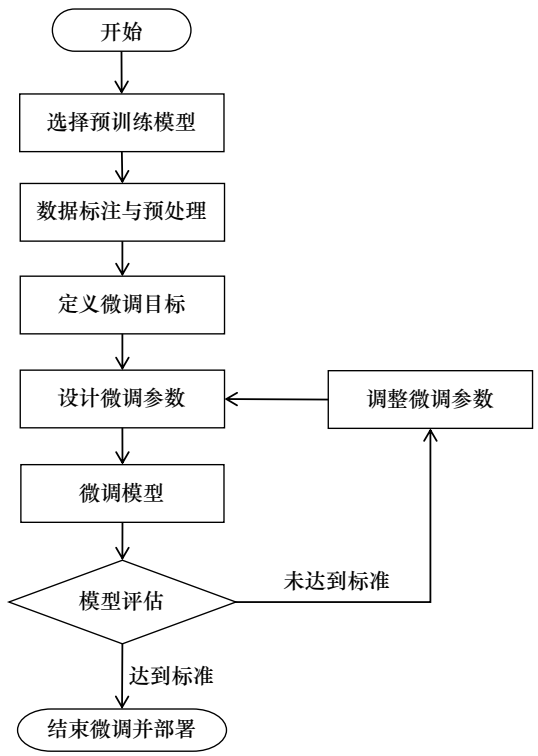

图2　基于SFT的对齐方法应用流程图

五、医疗大模型应用难点

（一）数据安全

医疗大模型涉及的数据种类多、来源广，难以共享且容易隐私泄露。不同医疗机构之间的数据“孤岛”现象也加剧了数据共享的难度。医疗大模型的训练语料库通常包含健康状况、疾病诊疗情况、临床监测数据、生物基因信息等数据，不仅涉及患者隐私，而且具有敏感性和特殊价值；一旦数据泄露，可能造成潜在的重大损失。用于大模型训练的医疗数据也可能被滥用至其他用途。医疗大模型可能尝试根据用户输入来预测患者的性别、民族、收入、宗教信仰等，造成侵犯个人隐私的潜在后果。即使在使用训练模型之后，有时只需检查生成的模型即可重建训练使用的原始数据点、从模型层参数中还原出原始的输入信息，从而泄露患者隐私。在训练之前对输入进行加密可以更好地保护患者数据，但会影响模型输出的可解释性。

医疗大模型易遭受投毒和对抗性攻击，面临安全风险。在数据收集阶段，可能存在医疗记录中插入虚假的诊断信息和治疗建议、敏感的医疗数据遭到泄露等风险。在数据预处理阶段，攻击者篡改医学图像并滥用人与机器之间的（视觉）认知差异，会导致模型对病变区域的误判。在模型训练阶段，大模型最易受到投毒攻击，如攻击者通过注入大量“错误”的医疗数据来搅乱数据的分布，调整大模型朝着期望的方向偏移，甚至利用后门进行深度隐

藏以等待合适的攻击时机。在推理阶段，大模型在受到对抗攻击（针对输入的医疗样本，故意添加一些人类无法察觉的细微干扰）后往往高置信度地给出错误的输出。在使用阶段，需要规避大模型引起的医疗数据泄漏、模型接口滥用等情况。

（二）技术风险

医疗大模型存在幻觉问题，其准确性和可靠性仍待提升。大模型理论（如上下文学习）处于黑箱阶段，暂时无法确保输出的完全可控：有时会生成看似合理但医学上不准确的响应，说明只是学习了表面上的语言模式，没有真正理解医疗数据的深层含义。一方面，训练数据的准确性、完整性有待确认及验证。大模型训练需要快速处理和分析大量的医疗数据，才能提出及时准确的诊断和治疗建议；而人工检查数据质量难以实现，也可能出现训练数据集、测试数据集重叠导致模型的过度预测现象。另一方面，大模型没有经历“理解”训练过程，不像医生一样理解输入和输出的信息。大模型尚无法全面应对医学知识和临床决策的复杂性，也无法完全复制临床医生的经验和细致判断能力；如果不满足可控与合规条件，其应用价值难以显现。

医疗大模型缺乏可解释性和透明度，依然难以取得用户的信任。一方面，医疗大模型问答过程、决策逻辑的可解释性不足，较难获取患者、医护人员、监管机构的信任。大模型通常采用深度神经网络，拥有多个隐藏层，涉及海量参数，使用多种策略进行并行加速，因而难以追踪单一医疗数据在模型中的处理过程，很难获得模型推理结果的有效解释，也不清楚训练数据集的哪些部分被用于生成结果。另一方面，医疗大模型的透明度和开源发展不足。许多用于预训练大模型的大型数据集仍然闭源，大模型代码通常不会公开发布，甚至有些大模型仅限于研究人员接触，导致业内很难独立验证和重建以前的结果。医疗数据通常受到严格审查，医疗大模型透明度和可验证性的不足将进一步引发用户的疑虑和不信任心理。

（三）落地挑战

医疗大模型的训练和推理成本偏高，制约落地应用。医疗大模型在经费、时间、算力、环境方面体现出高昂的开发和运营成本，高质量的医疗数据成本也很高；算力即使在不断增加，但无法匹配医疗大模型在网络参数、网络深度、数据量上的更快速增长。例如，GPT-4 的训练涉及 2.5×10^4 个 A100 GPU，耗时 90~100 d，训练成本为千万美元级；即使应用新一代 H100 GPU 进行预训练，时间也要 55 d 左右。医疗大模型的日常运营、模型迭代也会消耗大量算力，产生电力消耗和碳排放，给医院私有化部署带来挑战。实际医疗场景中部署医疗大模型，需要做好模型大小选择、成本与收益权衡。

医疗大模型引发新型权责问题，而相关问责制有待探讨完善。医疗大模型尽管具有较强的生成能力、支持人类医疗决策的潜力，但存在限制个人自主权、产生新责任和新纠纷的风险。患者可能过度依赖医疗大模型，失去对自身健康的自主理解或控制，意味着他们需要对自己的健康决策承担更大的个人责任。医疗大模型生成内容可能存在知识产权侵权的风险，其来源和生成内容也存在权属不清、取证与损害认定困难等问题。医疗大模型协助或参与哪些医学、医疗任务，有何种自主权（作为自主、半自主、完全从属工具）？这些问题有待深入探讨才能达成共识。医疗大模型设定为不同的角色和权限，将直接影响医疗流程、患者安全、医疗质量。此外，权责界定模糊会导致出现问题后难以明确原因并认定参与者的责任，会降低潜在患者、医护人员的使用意愿。

（四）伦理道德

医疗大模型强化偏见现象，将加剧歧视和社会不公平性。由于性别、年龄、民族、收入、教育、地理等因素的差异，世界上大多数国家在医疗方面存在不平等、不公平现象，而医疗大模型倾向于延续并放大导致医疗不公平的系统性差异和人类偏见。医疗大模型的规模和能力在不断增长，可能比相对较小的模型表现出更高的偏见和歧视水平。预训练数据规模庞大，难以人工开展收集、标注和检查。如果医疗数据集主要包含某一种疾病的病例记录而忽略其他类型的疾病，那么在处理不同类型疾病时可能表现出明显的偏好或偏见。设计算法时，如果未能适当地考虑患者的特定背景和实际情况，就可能产生不准确的预测或建议，从而影响患者接受适当医疗服务的机会，进一步加剧社会不公平性。

医疗大模型有可能生成有害内容、传播虚假错误信息。在生成有害内容方面，随着医疗大模型的兴起和发展，患者会无意中接触到可能导致严重情感伤害的话题。医疗大模型拟人化程度高，更容易获得患者的信任，但往往缺乏额外的个性化情感支持能力；如果患者通过与医疗大模型交流得知自己患有某种致命疾病，这种突然的信息冲击可能导致患者严重的心理反应和情感伤害。在虚假信息传播方面，医疗大模型生成的文本与人类书写的文本越来越难以区分，虽然医疗大模型可以减轻临床医生的文档编写负担，但也使他人恶意使用医疗大模型生成虚假的医疗记录或误导性的健康建议成为可能，对患者健康、医疗系统的信誉造成潜在威胁。

六、医疗大模型发展建议

（一）发挥政府引导优势，保障数据安全

建议适时发布国家层面的政策和法规，支持破除医疗数据流通的行政壁垒，解决医疗大模型构建方面的数据难题。建设国家级医疗大模型数据集公共服务平台，整合多源医疗数据，提供数据收集、处理、标注等工具，促进数据要素价值释放。加强训练数据的来源及质量审核，加快数据确权、共享、安全防护等机制研究，明确数据收集和使用的合规要求，防止数据滥用和隐私泄露。构建统一的安全框架和标准协议，集成对抗性攻击检测系统，实时监控和识别潜在的对抗性攻击，提高医疗大模型的鲁棒性。定期对医疗大模型进行安全审计，检查潜在的安全漏洞和不当的数据访问行为，确保模型应用的安全性与合规性。制定详细的医疗大模型应用应急预案和快速响应机制，确保发生安全事件时可以迅速有效地应对。

（二）加快基础理论研究，突破技术风险

加大基础研究投入，攻关医疗大模型算法、框架等基础性、原创性技术，提升医疗大模型在泛化性、准确性、透明性、可解释性、公平性等方面的能力，突破技术局限。实施全面的数据质量控制流程，包括数据验证、清洗、去重，减少训练数据集中的误差和偏差。制定医疗大模型的性能评估和验证标准，通过标准化测试确保模型应用的准确性和可靠性。设计高级可视化、解释性等工具，辅助用户理解医疗大模型的决策过程，提高模型输出的可解释性和透明度。鼓励“产学研医”联合开发相关技术，聚合AI科技企业、制药企业、合同研究组织企业、医疗机构、科研机构、算力资源，通过传统医学人才、大模型技术人才的联合攻关，突破新理论、新技术、新算法，变革医疗大模型科研范式。

（三）强化应用场景牵引，缓解落地挑战

加强大模型开源社区、公共算力平台建设，降低大模型的获取成本。加快研发模型计算效率提升技术，剪枝、量化、知识蒸馏等模型压缩技术，降低大模型的开发和使用成本。推广模型训练、微调、优化工具，建立指令微调数据集，降低医疗大模型开发和部署门槛。明确医疗大模型代码、微调技术及其生成内容的版权归属。探索建立医疗大模型问责制度，明确研发、部署、使用各方的权利及义务。鼓励有条件的机构开展大模型技术在智慧医疗实际场景中的应用探索，采用“揭榜挂帅”“重点工程”“试点示范”等机制，发挥应用场景的纽带作用。加速推动医疗大模型技术的产业化，促进成果落地、产业链协同，积极培育新模式、新业态。

（四）建立健全监管机制，规范伦理道德

开发和应用偏见检测工具，定期审查医疗数据和模型输出，识别并纠正医疗大模型应用潜在的偏见和不公平性。建立内容审核和过滤机制，明确有害内容的定义与监管策略，防止医疗大模型生成有害、虚假、误导性的信息。鼓励患者和医护人员提供反馈，监测医疗大模型应用中的错误和问题，及时进行模型修正和更新。通过跨部门、跨领域协同，建立全方位、多层次、立体化的监管体系，提高监管效能。采取分类管理、风险分级策略，根据医疗大模型的预期用途、风险等级进行监管，确立各类医疗大模型的许可使用及限制范围。完善大模型治理机制，建立全面的伦理审查机制，确保各类医疗大模型在研制和训练过程中遵循严格的伦理标准。推动医疗大模型的立法研究和制定，确保技术应用符合社会伦理标准，规范医疗大模型建设生态。

（五）完善公共服务体系，营造创新生态

从顶层设计出发，论证医疗大模型发展战略，制定医疗大模型中长期规划，探索医疗大模型高质

量发展的“中国路径”。推进智慧医疗领域医疗大模型“一张网”布局，有组织地开展国产医疗大模型的研制和部署，抢占科技创新制高点。合理加大公共资金投入和资源支持力度，培养掌握医学、数据分析、AI等技能的复合型人才，为医疗大模型建设筑牢智力保障。积极参与国际标准制定，提升国产医疗大模型在国际应用市场上的影响力。鼓励医疗大模型领域的龙头企业强化产业生态布局，凝练典型案例、推广成功经验，提供第三方开发能力和解决方案，带动产业链上的中小企业协同发展。探索多方参与、合作共赢的医疗大模型商业模式，形成贯通数据、算法、算力、算网、应用的医疗大模型创新生态。

利益冲突声明

本文作者在此声明彼此之间不存在任何利益冲突或财务冲突。

Received date: July 1, 2024; **Revised date:** August 29, 2024

Corresponding author: Yuan Yige is an associate research fellow from Xiangjiang Laboratory. His major research fields include large language model and intelligent medical. E-mail: immyyuan23@163.com

Funding project: Chinese Academy of Engineering project “Research on Global Future Industrial Development Trend and Hunan Future Industrial Layout” (2024-DFZD-39); Xiangjiang Laboratory Project (23XJ01008, 23XJ03001)

参考文献

[1] 陈晓红, 许冠英, 徐雪松, 等. 我国算力服务体系构建及路径研究 [J]. 中国工程科学, 2023, 25(6): 49–60.
Chen X H, Xu G Y, Xu X S, et al. Computing power service system of China and its development path [J]. Strategic Study of CAE, 2023, 25(6): 49–60.

[2] 人工智能医疗器械创新合作平台, 中国信息通信研究院. 人工智能大模型赋能医疗健康产业白皮书 (2023) [R]. 北京: 人工智能医疗器械创新合作平台, 中国信息通信研究院, 2023.
Artificial Intelligent Medical Device Innovation and Cooperation Platform, China Academy of Information and Communications Technology. White paper on empowering the healthcare industry with AI large language models (2023) [R]. Beijing: Artificial Intelligent Medical Device Innovation and Cooperation Platform, China Academy of Information and Communications Technology, 2023.

[3] 亿欧智库. 2023医疗健康AI大模型行业研究报告 [R]. 北京: 亿欧智库, 2023.
EO Intelligence. Research report healthcare AI large language model industry (2023) [R]. Beijing: EO Intelligence, 2023.

[4] 陈晓红, 陈姣龙, 胡东滨, 等. 面向环境司法智能审判场景的人工智能大模型应用探讨 [J]. 中国工程科学, 2024, 26(1): 190–201.
Chen X H, Chen J L, Hu D B, et al. Application of artificial intelligence large language model for smart environmental judicial adjudication [J]. Strategic Study of CAE, 2024, 26(1): 190–201.

[5] Wang H F, Li J W, Wu H, et al. Pre-trained language models and their applications [J]. Engineering, 2023, 25: 51–65.

[6] 车万翔, 窦志成, 冯岩松, 等. 大模型时代的自然语言处理: 挑战、机遇与发展 [J]. 中国科学: 信息科学, 2023, 53(9): 1645–1687.
Che W X, Dou Z C, Feng Y S, et al. Towards a comprehensive understanding of the impact of large language models on natural language processing: Challenges, opportunities and future directions [J]. Scientia Sinica Informationis, 2023, 53(9): 1645–1687.

[7] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge [J]. Nature, 2023, 620(7972): 172–180.

[8] Thirunavukarasu A J, Ting D S J, Elangovan K, et al. Large language models in medicine [J]. Nature Medicine, 2023, 29(8): 1930–1940.

[9] Webster P. Six ways large language models are changing healthcare [J]. Nature Medicine, 2023, 29: 2969–2971.

[10] Moor M, Banerjee O, Abad Z S H, et al. Foundation models for generalist medical artificial intelligence [J]. Nature, 2023, 616(7956): 259–265.

[11] 陈润生. 医疗大数据结合大语言模型的应用展望 [J]. 四川大学学报(医学版), 2023, 54(5): 855–856.
Chen R S. Prospects for the application of healthcare big data combined with large language models [J]. Journal of Sichuan University (Medical Sciences), 2023, 54(5): 855–856.

[12] 韩晓光, 朱小龙, 姜宇桢, 等. 人工智能与机器人辅助医学发展研究 [J]. 中国工程科学, 2023, 25(5): 43–54.
Han X G, Zhu X L, Jiang Y Z, et al. Development strategies for artificial intelligence and robotics in medicine [J]. Strategic Study of CAE, 2023, 25(5): 43–54.

[13] 颜见智, 何雨鑫, 骆子烨, 等. 生成式大语言模型在医疗领域的潜在典型应用与面临的挑战 [J]. 医学信息学杂志, 2023, 44(9): 23–31.
Yan J Z, He Y X, Luo Z Y, et al. Generative large language lodels in the medical domain: Potential and typical applications and challenges [J]. Journal of Medical Intelligence, 2023, 44(9): 23–31.

[14] Huang Z, Bianchi F, Yuksekgonul M, et al. A visual-language foundation model for pathology image analysis using medical twitter [J]. Nature Medicine, 2023, 29(9): 2307.

[15] Veen D V, Uden C V, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization [J]. Nature Medicine, 2024, 30: 1134–1142.

[16] Zhang X M, Wu C Y, Zhang Y, et al. Knowledge-enhanced visual-language pre-training on chest radiology images [J]. Nature Communications, 2023, 14: 4542.

[17] Yu G, Sun K, Xu C, et al. Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images [J]. Nature Communications, 2021, 12: 6311.

[18] Zhou H J, Liu F L, Gu B Y, et al. A survey of large language models in medicine: Principles, applications, and challenges [EB/OL]. (2024-02-02)[2024-07-15]. https://arxiv.org/html/2311.05112v3.

[19] 杨善林, 丁帅, 顾东晓, 等. 医疗健康大数据驱动的知识发现与知识服务方法 [J]. 管理世界, 2022, 38(1): 219–228.
Yang S L, Ding S, Gu D X, et al. Healthcare big data driven knowledge discovery and knowledge service approach [J]. Journal

of Management World, 2022, 38(1): 219–228.

[20] 蛋壳研究院. 2023 数字智慧病理行业研究报告 [R]. 重庆: 蛋壳研究院, 2023.
VCBeat Research. Research report on digital intelligence pathology industry (2023) [R]. Chongqing: VCBeat Research, 2023.

[21] 郑永年. 如何科学地理解“新质生产力”? [J]. 中国科学院院刊, 2024, 39(5): 797–803.
Zheng Y N. How to scientifically understand “new quality productivity” [J]. Bulletin of Chinese Academy of Sciences, 2024, 39(5): 797–803.

[22] 周文, 许凌云. 论新质生产力: 内涵特征与重要着力点 [J]. 改革, 2023 (10): 1–13.
Zhou W, Xu L Y. On new quality productivity: Connotative characteristics and important focus [J]. Reform, 2023 (10): 1–13.

[23] Bi K F, Xie L X, Zhang H H, et al. Accurate medium-range global weather forecasting with 3D neural networks [J]. Nature, 2023, 619(7970): 533–538.

[24] Romera-Paredes B, Barekatain M, Novikov A, et al. Mathematical discoveries from program search with large language models [J]. Nature, 2024, 625(7995): 468–475.

[25] Fei N Y, Lu Z W, Gao Y Z, et al. Towards artificial general intelligence via a multimodal foundation model [J]. Nature Communications, 2022, 13: 3094.

[26] 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.
Guo H Y, Liu P, Lu R G, et al. Research on a massively large artificial intelligence model and its application in medicine [J]. Scientia Sinica Vitae, 2024, 54(3): 482–506.

[27] 余艳, 张文, 熊飞宇, 等. 融合知识图谱与神经网络赋能数智化管理决策 [J]. 管理科学学报, 2023, 26(5): 231–247.
Yu Y, Zhang W, Xiong F Y, et al. Fusion of knowledge graph and neural network to empower data-intelligence for management decisions [J]. Journal of Management Sciences in China, 2023, 26(5): 231–247.

[28] Dias A L, Rodrigues T. Large language models direct automated chemistry laboratory [J]. Nature, 2023, 624(7992): 530–531.

[29] Shanahan M, McDonell K, Reynolds L. Role play with large language models [J]. Nature, 2023, 623(7987): 493–498.

[30] Boiko D A, MacKnight R, Kline B, et al. Autonomous chemical research with large language models [J]. Nature, 2023, 624(7992): 570–578.

[31] Piktus A. Online tools help large language models to solve problems through reasoning [J]. Nature, 2023, 618(7965): 465–466.

[32] Zhen C Q, Shang Y L, Liu X Y, et al. A survey on knowledge-enhanced pre-trained language models [EB/OL]. (2022-12-27)[2024-07-15]. https://arxiv.org/abs/2212.13428.

[33] Pan S R, Luo L H, Wang Y F, et al. Unifying large language models and knowledge graphs: A roadmap [J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(7): 3580–3599.

[34] Yin S K, Fu C Y, Zhao S R, et al. A survey on multimodal large language models [EB/OL]. (2023-06-23)[2024-07-15]. https://arxiv.org/abs/2306.13549.

[35] Ding N, Qin Y J, Yang G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models [J]. Nature Machine Intelligence, 2023, 5: 220–235.

[36] Arasteh S T, Han T Y, Lotfinia M, et al. Large language models streamline automated machine learning for clinical studies [J]. Nature Communications, 2024, 15: 1603.