• 专论:医学大语言模型 •

编者按: 大规模生成式语言模型是自然语言处理领域的核心技术之一,为医疗智能化带来强大的通识知识基础和高级认知能力,具有巨大的医学领域应用潜力。目前已经出现了一些面向医疗健康的大语言模型和应用,如 Med-PaLM、MedGPT、山海大模型等,将其应用于医疗信息抽取、医疗实体标准化、病历生成、药物研发、医学影像分析、医师考试、医患问答等方面结果令人惊喜。然而大模型仍然存在幻觉等问题,难以直接胜任医学的复杂决策应用,其泛化能力、准确性、解释性等方面有待提升。本期专论着眼于医学大语言模型,所载论文详细阐述了大规模生成式语言模型在医疗领域的潜在典型应用、为医疗行业发展带来的机遇、在医疗应用中面临的挑战和风险,以及基于知识增强的医学语言模型的相关研究应用,为快速适应新型人工智能技术发展、进一步促进医疗智能化提供借鉴和参考。

大规模生成式语言模型在医疗领域的应用:机遇与挑战

肖仰华[1] 徐一丹[2]

([1] 上海市数据科学重点实验室 复旦大学计算机科学技术学院 上海 200438
[2] 复旦大学生物医学研究院 上海 200032)

〔**摘要**〕 **目的/意义** 以 ChatGPT 为代表的大规模生成式语言模型带动了通用人工智能技术快速发展。大规模生成式语言模型能否在医疗领域应用取得成功是学术界和工业界密切关心的问题。本文旨在深入研究大规模生成式语言模型在我国医疗领域应用中的机遇与挑战。**方法/过程** 从知识容器、能力引擎和自治智能体 3 方面出发,分析大模型在医疗提质增效、解决我国医学发展不平衡问题、慢性病智能管理与决策、人口老龄化应对以及医学科研加速等方面的新机遇,同时指出大模型在医疗领域应用所存在的局限。**结果/结论** 大模型驱动有望成为智能医疗的新范式,针对大模型在医疗应用中的不足,提出具体发展建议。

〔**关键词**〕 大规模生成式语言模型;医疗;人工智能

〔**中图分类号**〕R-058 〔**文献标识码**〕A 〔**DOI**〕10.3969/j.issn.1673-6036.2023.09.001

The Application of Large Generative Language Models in the Medical Field: Opportunities and Challenges

XIAO Yanghua[1], XU Yidan[2]

[1] Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai 200438, China; [2] Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China

〔**Abstract**〕 **Purpose/Significance** The large generative language model (GLM) represented by ChatGPT has driven the rapid development of artificial general intelligence technology. Weather GLM can be successfully used in medical applications is a closely medical treatment issue in both academia and industry. The paper aims to delve into the opportunities and challenges of GLM in the application of

〔**修回日期**〕 2023-09-20

〔**作者简介**〕 肖仰华,博士,教授,上海市数据科学重点实验室主任,复旦大学附属眼耳鼻喉科医院等多家机构特聘教授。

medical treatment in China. **Method/Process** Starting from three aspects: knowledge container, capability engine, and autonomous agent, the paper analyzes the new opportunities of GLM in improving medical quality and efficiency, solving the imbalance problem of medical development in China, intelligent management and decision - making of chronic diseases, coping with population aging, and accelerating medical research. At the same time, it points out the limitations of the application of large models in the medical field. **Result/Conclusion** The big model driven approach is expected to become a new paradigm for intelligent healthcare. In response to the shortcomings of large models in medical applications, specific suggestions are proposed.

〔**Keywords**〕 large generative language model (GLM); medical treatment; artificial intelligence (AI)

1 引言

自 2022 年 12 月 ChatGPT 发布以来，大规模生成式语言模型（generative language model，GLM）（以下简称大语言模型或大模型）在学术界与工业界引起强烈关注，带动了一系列通用人工智能技术（artificial general intelligence，AGI）的快速发展，包括图文生成模型、具身多模态语言模型等。作为对话式大型语言模型，ChatGPT 提供开放式人机对话能力，具备文本理解能力，能理解用户的问题，提供比较精准的问答服务，同时还具有较强的文本生成能力，短期内吸引了海量用户。

ChatGPT 在开放问答场景的优异能力能否迁移到垂直领域，特别是医疗领域，是当下医疗行业十分关心的问题。一些研究开始尝试将其应用于美国执业医师资格考试（United States Medical Licensing Examination，USMLE）[1]，以及心血管疾病[2]、阿尔茨海默病等疾病诊疗[3]，近期更有报道称 ChatGPT 确诊了多名医生无法诊断的罕见病[4]。本文针对以 ChatGPT 为代表的生成式语言模型在医疗应用中的机遇与挑战、智能医疗新范式和医学大模型如何进一步发展展开分析，并给出相应观点与判断。

大模型在很多场合被视作一类基础模型，即在大规模数据上训练且可适配（如微调）广泛下游任务的模型。泛化能力比较好的基础模型通常需要海量参数，比如百亿甚至是千亿的参数规模，因而常被称作大模型。大模型通常使用大型神经网络（如 Transformer），通过自监督学习机制[5]，针对海量文本、图像等数据进行训练与建模。针对语言数据训练而得的模型就是语言模型。当参数量超过一定规模时，语言模型的性能得到显著提升，并且具备小规模语言模型所没有的能力[6]，这一现象被称作“涌现”。在通用领域中出现的大语言模型在遵循指令和产生类人回复方面表现优异，但在特定的专业领域（如医疗领域），其效果往往差强人意。将大模型应用于医学领域时，往往需要微调（如 ChatDoctor），才能具备专业领域的知识，胜任诸如临床、科研、教育等应用[7-8]。

2 大模型为我国医疗行业发展带来的新机遇

2.1 补充人类认知能力，应对复杂医学体系

近 10 年来受益于信息技术的发展与应用，医学技术迅猛发展并带来新一轮医疗变革，为疾病诊断及治疗、健康预防和管理带来全新机遇[9]。随着海量医学文献的指数级增长，医学从业人员已经难以阅读、学习、理解快速积累的医学知识。医疗专业细分造成不同医学学科之间的认知壁垒。面对日益复杂的医学体系，人类的认知能力已经难以应对不断出现的新病种（如环境病、劳动病）与新病毒。以大模型为代表的机器认知成为人类认知能力的补充，发展基于大模型的医学认知智能，实现人机认知协作，才能有效掌握日益复杂的医学体系与医疗系统。

2.2 为医疗提质增效，缓解专家资源稀缺问题

在医疗领域，以 ChatGPT 为代表的生成式语言模型可以较好地完成医疗助理类工作，如预约挂号、信息管理、健康咨询等，提高医疗人员的工作效率与质量，降低患者信息获取门槛。大模型还能就病情给出诊断建议、治疗方案、类案推荐，从而开拓医生的诊治思路、提示用药风险与临床指南，实现临床辅

助决策；同时帮助医疗专业人员阅读大量文献、提炼核心观点与结论，便于其了解行业动态，加速实验进程等[10]。随着知识体系的完善和认知能力的增强，大模型可以借助问答交互形式完成常规疾病的诊断或医疗建议，对相似病例进行搜索与推荐，从而胜任部分普通医生的工作。随着大模型在医疗行业应用的深入，医生的时间和精力有望得到进一步解放，从而缓解医学专家资源稀缺的问题。

2.3 缓解医学发展不平衡问题

大模型在医学教育中应用潜力较大。例如能生成课堂上使用的练习、测验和场景，帮助医学生进行练习和评估。基于大模型的医学智能教育使欠发达地区的医生和患者能够及时获取前沿、权威的医学知识，有助于缩小城乡医疗水平差距，提高医疗服务均等性[11]。

2.4 筑牢医疗安全底线

安全是医疗活动的底线。临床医生在执业过程中存在主观上忽视的可能。医疗大模型可提供实时提醒和辅助，例如通过提醒、核对清单等方式帮助医生遵循正确的用药流程，减少疏忽和错误；通过挖掘医疗记录和数据，识别患者的用药历史、潜在药物相互作用风险，提醒医生用药注意事项等。

2.5 加速医学科研进程

大模型在加快医学研究、推动新药研制等方面具有巨大潜力[8]。大规模参数化的深度神经网络模型已经成为蛋白质、小分子化合物表示学习的重要手段，对药物预测与发现具有积极意义，未来有望显著加快医学发现进程。基于大模型的“预训练 + 微调”范式对于充分提炼海量医学数据中的有效特征，并将其迁移应用至新疾病、新药物，已发挥巨大作用[12-14]。通过分析大量医学数据，大规模预训练语言模型可以习得不同医学学科专业知识，例如放射学、病理学和肿瘤学，为跨学科综合性疾病、疑难杂症的诊治带来机遇。

2.6 实现慢性病智能化管理与决策，应对老龄化挑战

随着我国老龄化进程的加速，慢性病（糖尿病、心脏病、癌症等）健康咨询与管理给社会化医疗带来巨大成本。慢性病管理、健康咨询、用药咨询等通常可以通过个人自助、聊天问答等形式实现。经过慢性病与健康知识增强的 ChatGPT 类语言模型能够较好地胜任此类场景。相对于医学诊治而言，慢性病与健康管理多属于健康建议，对精确性要求相对较低，很多时候给出原则正确但相对模糊的答案是可以接受的（如患者咨询如何降血脂，ChatGPT 通常会回答“饮食调节、运动锻炼、戒烟限酒、控制体重”等）。如果将用户个人信息与 ChatGPT 能力结合，有可能进一步实现个性化健康管理，引导大模型产生与患者信息高度契合的健康建议。ChatGPT 类大模型还可以与各类健康设备（如具有心率和血压检测功能的手表）协同，实现更加智能的实时健康监测与提醒。

3 大模型驱动成为智能医疗新范式

3.1 人工智能应用的两种典型技术范式

在人工智能与各行业应用深度融合过程中，形成了数据驱动的机器学习和知识驱动的符号计算两种典型的技术范式。数据驱动范式主要通过样本拟合习得统计模型以求解实际问题，是践行人工智能联结主义思想的主要形式之一。知识驱动范式主要通过构建符号表达的知识库形成推理能力来解决问题，是践行人工智能符号主义思想的主要形式之一。两种范式各有优缺点，数据驱动范式能够充分捕捉海量数据中所蕴含的隐性特征，基于深度学习的端到端学习与自监督学习进一步降低专家特征工程与样本标注成本。知识驱动的符号计算擅长符号推理，过程可解释、可干预，在环境封闭、规则明确的应用场景取得较好效果。二者均对机器认知智能的发展具有积极作用。

3.2 大模型驱动——智能医疗新范式

两种典型技术范式在实际应用中仍存在问题。以知识图谱为代表的知识驱动范式诉诸显式的符号表示表达知识，难以表达专家经验以及常识知识。传统统计模型以小规模参数化模型为主，难以对复

杂世界、复杂业务进行完整、准确建模。而大模型在一定程度上能够缓解上述问题。大模型从海量语料中习得蕴含于语料中的知识，包括常识知识、世界知识、专业知识等，从而完成对于复杂世界的完整建模。大模型本质上是基于大规模深度神经网络（如 Transformer 架构）训练出的统计模型，能够准确反映建模符号之间的统计关联，完整地表达蕴含于文本中的隐性知识，对于再现领域专家的直觉推理能力具有重要意义。大模型有望成为传统数据驱动范式与知识驱动范式发展的最后归宿。具体到医疗领域，大模型将成为医疗认知智能的重要基座。

4　大模型为医疗智能化带来的新机遇

4.1　知识容器

4.1.1　大规模语言模型是巨大的参数化知识容器

语言模型从海量语料自监督训练而来。从知识角度可以视作一个巨大的参数化知识容器，即以参数化形式编码数据中蕴含的知识容器。之所以将之视作知识容器而不是知识库，是因为其在知识存取方式上不如后者直观。语言模型中的知识获取必须经历特定的知识诱导，即构造合理的问题作为提示，才能获得所需结果。而传统知识库中的知识通常是符号表达，有着明确的语法与语义结构，可以通过明确的查询语言或者应用程序接口（application programming interface，API）获取，见图 1。

图 1　知识图谱与语言模型知识获取方式的差异

语言模型的参数化知识表达是造成上述差异的根本原因。当前语言模型多是基于 Transformer 模型架构进行训练的，不同大模型架构略有差异，其本质均是多层深度神经网络，具有数亿至千亿规模参数。虽然研究者努力尝试解释其内在结构与运行机理，但深度神经网络总体上仍是“黑盒”。这从根本上决定与传统知识库相比，作为知识容器的大模型在可理解、可解释、可编辑等方面存在不足。此外，语言模型和传统的知识容器（如数据库、规则库、知识库）在表达形式、知识内容、知识特性等方面都存在本质区别，见表 1。例如，知识图谱中的三元组使用符号表达，易于理解和编辑（增、删、改、查），因此具有较高可控性；而大模型中的参数化知识难于编辑，很难胜任可控性要求较高的应用场合。又如知识图谱推理主要使用符号推理，而大模型本质上是一个统计生成模型，其推理本质上是概率推断。

表 1　知识图谱与语言模型对比

对比维度	语言模型	知识图谱	说明
表达形式	参数知识	符号知识	语言模型的知识以参数形式存在于模型的权重中，包含在网络中的连接权重中，无法直接以符号方式表示。知识图谱是一种基于符号的结构化知识表示方法，通过节点和关系来描述实体之间的关联
知识内容	通用知识	通用、专业知识	语言模型的训练数据通常来自于互联网或大规模文本语料库，因此其所包含知识广泛而通用，涵盖语言知识、常识知识以及各领域相对宽泛的知识。知识图谱的知识内容通常针对特定领域或专业，如医学、法律、金融等，包含相应领域的概念、属性、关系和规则
知识特性	隐性知识	显性知识	语言模型本质上习得了语言符号的概率分布，以参数化形式表达了语料中蕴含的知识，可以视作一类隐性知识表达。知识图谱是一种基于符号的结构化知识表示方法，通过概念、实体、属性和关系等形式来表示知识。知识图谱中的知识是以显性的符号化知识表示的，遵行一定结构和规范
知识获取	统计生成	直接读取	语言模型通过构造问题或者提示获取相应知识，需要知识诱导过程才能获取知识。知识图谱通过图查询语言（如 Gremlin）或者资源描述框架（resource description framework，RDF）查询语言（如 SparQL）得到相应知识
知识推理	统计推断	符号推理	语言模型是一种统计模型，其推理过程本质上是统计推断，如根据给定的上下文预测某个词在该上下文中出现的概率。知识图谱的推理是一种典型的符号语义推理，如根据概念图谱进行上下位关系推理，根据规则、图谱中的路径进行推理

续表 1

对比维度	语言模型	知识图谱	说明
能否编辑	难编辑	可编辑	大规模语言模型一旦习得特定知识就往往难以编辑。其习得的知识是以一种分布式方式隐藏于相应的深度神经网络中的，难以定位，编辑相应知识有巨大挑战。知识图谱作为一种显性的符号知识可以直接存取与编辑，可以通过手动或自动的方式向知识图谱中添加、删除和修改实体、属性和关系，实现知识的编辑和更新
能否理解	难理解	易理解	大规模语言模型多是通过 Transformer 等深度神经网络进行训练的，其内在机制难以理解、不透明。总体上大模型的“黑盒”特性仍待更多研究探索。知识图谱中的知识是以直观的实体、属性和关系符号形式存在的，可以通过可视化和结构化的方式来清晰展示和解释
能否控制	不可控	可控	作为统计模型，大模型的行为呈现出一定的不确定性，如生成式大模型的生成结果往往难以控制，容易产生与给定约束相违背的内容，还容易产生所谓幻觉问题。相比较而言，知识图谱对于用户是完全透明可控的，用户可以按照需要对其内容进行相应调整
能否保护	难防护	易防护	大模型一旦习得涉及隐私、安全的数据或事实，就将其以一种分布式的方式隐藏在整个深度神经网络中，很难彻底清洗，能通过特定方式诱导出敏感信息，存在语言模型滥用或误用的风险。知识图谱中的知识则很容易实现授权授信与分级管控，更易于防护

4.1.2 医学知识表达需要知识图谱与大模型相互补充 医学领域知识密集且不断发展，包括诊断、治疗、药物、疾病等方面的信息。大模型可以通过海量医学文献、专业书籍和其他来源的内容，学习医学领域内各种概念和知识，从而具备回答医学问题的能力。很显然医学领域应用对于知识的可控与可编辑、可理解等具有较高要求，因此知识图谱等传统知识库是不可或缺的。医学大模型所编码的隐性知识可以作为医学常识的容器，对医学知识图谱具有显著补充作用。总体而言，医学领域知识仍要以医学知识图谱为主，以医学大模型为辅，才能胜任严肃应用的较高要求。

4.1.3 通用大模型的通识能力是实现医疗专业认知能力的前提 通用大模型基于来源多样的语料进行训练，具备较强通识能力，可习得不同学科的知识。这种通识能力对于大模型理解开放世界的文本与数据至关重要，是各类垂直领域专业认知能力的前提。人类理解某个概念时，包括理解该概念范畴之外的内涵与外延。例如在日常诊疗过程中，很多情况下医生要排除疾病可能性。换句话说，要想理解疾病首先必须理解健康。因此这种“先通识、再专识”的智能实现路径与人类的教育过程极为相似。医疗的垂域认知必须建立在具备通识能力的大模型基础之上。

4.2 能力引擎

大模型具备人类水平的认知与思维能力。人类认知是极为复杂的现象。从这个角度试图对 ChatGPT 的认知能力做详尽分类与分析是极具挑战性的。但就目前大量存在的评测而言，ChatGPT 类大模型已经具备了语言理解、逻辑推理、常识理解、概念理解、运筹规划、评估评测、组合泛化、价值判断、自识反思、问题求解等一系列普通人的核心思维与认知能力，在实际应用中呈现出较为出色的开放世界理解能力、组合创新能力和评估能力。大模型具有出色的开放世界理解能力，如对于任一病种，其均能通过大量文献习得相关基础知识。大模型的组合创新能力即在经过足量常见任务的指令学习后，能够胜任新的组合任务。例如，大模型可同时学习放射报告和对应的 X－Ray 图片，从而根据医学影像图片自动生成对应的放射学报告，以减轻放射科医生的工作压力[15]。目前，这种组合创新能力已经远超人类认知水平，迫使人类重新思考创新的本质。

4.3 自治智能体

语言模型将显著提升自治智能体的认知水平。真实医疗场景任务复杂，智能体需要多次交互才能够给出医学建议。基于大语言模型的自治智能体拥有海量参数化知识与认知能力，与传统智能体相比具有显著优势，主要体现在以下几方面。一是更强大的世界建模和感知能力。由于大模型拥有大量参数化知识，为基于大模型的智能体在理解世界及其复杂性方面提供强大基础。二是处理复杂任务的能力。利用大模型的复杂规划和决策能力，基于大模

型的自治智能体能够处理各种复杂任务，在各领域都表现出显著的通用性。三是高级认知和思考能力。大语言模型能展示出类人的理解能力，在处理抽象概念、进行推理、理解复杂上下文等方面能力更强。四是更好的交互性。大模型赋予智能体更好的语言处理能力和更丰富的交互方式，从而实现更为流畅与自然的人机交流。从大模型向自治智能体发展有待进一步增强规划与决策、自我监督学习、理解人类社会的复杂目标与价值、安全与伦理管理、多模态人机交互与协作能力，见图 2。

图 2　语言模型驱动的自治智能体

5　大模型在医疗应用中的局限

5.1　难以仅从文字记载中习得专家经验

医疗是一类典型的严肃、复杂应用场景，对大模型的准确性、精确性、安全性、可靠性、认知能力均提出更高要求。所谓严肃，是指医疗领域的错误容忍度非常低，对合规性要求非常高。同时医疗也是一类复杂应用场景，例如医生在对患者进行诊断时，不仅要考虑过往病史，还要通过场景判断患者所述是否属实。医疗专家多凭借自身丰富经验（通常属于隐性知识）解决问题。大模型难以仅从文字记载的数据中习得，因此与资深医疗专家水平仍有差距。同时，大模型为医疗领域带来新机遇。首先，大模型对患者与病情的认知在广度上远超人类专家。医疗诊断过程中，医生很难完整地对患者病史进行全面询问，往往在信息不完全前提下作出决策，而大模型能掌握相对完整的信息。其次，大模型对跨专科的病情认知能力远超人类专家。在临床实践中，亚专科医生往往过于专注于本领域的疾病谱，很难对其他亚专科疾病进行判断，可能产生漏诊和误诊。而医学大模型有望提示遗漏或者罕见病症诊断。

5.2　无法辅助解决真实工作场景中的复杂决策任务

以大模型为核心的人工智能技术在医学中的应用定位仍是辅助决策。ChatGPT 类大模型本质上是在开放环境中实现人机对话，但是开放聊天无法辅助解决真实工作场景中的复杂决策任务。例如在疾病诊断方面，医生会根据患者的收入和医保情况制定个性化诊疗方案。要胜任此类工作，智能机器仍需要丰富的专业知识、合理的角色定位、病情病势研判能力、复杂约束取舍能力、不完全信息下的推断能力等，大模型要从聊天能手变成医学专业助手仍然要经历漫长的优化过程。

5.3　幻觉问题

医疗关系到公众生命健康，对错误零容忍。任何人工智能技术都要有系统性的解决方案才能达到医学严苛的准确性与精度要求。当前生成式大模型是生成式人工智能（AI－generated content，AIGC）领域的主流，而生成式大模型本质上是概率模型，仍有产生错误的可能。例如，胃复安的主要成分是甲氧氯普胺，而不是 ChatGPT 回答的天然气孔草酮，见图 3。大模型看似严谨的回答可能存在基本事实错误，在应用时仍然要付诸极大的代价判断信息真伪，增加了应用成本。幻觉问题是大模型落地垂域应用不可避免的问题。

提问	ChatGPT回答
胃复安的主要成分是什么？	胃复安(Gastrodin)是一种用于治疗胃肠道问题的药物，其主要活性成分是“天然气孔草酮”(Gastrodigenin)，这是从中药材天麻中提取的化合物

图 3　ChatGPT 的幻觉问题

造成大模型幻觉问题的原因主要有 3 方面。首先，大模型先天能力不足，即训练不充分。大模型训练语料、轮次有限都会导致学习不充分、知识有限、推理受限等问题。其次，实际任务往往形态复杂，表达成复杂指令，大模型对复杂指令理解能力

有限，进而产生幻觉。再次，大模型对于特定领域的知识理解能力有限。由于专业领域知识相对薄弱、难获取，大模型对于某些专业知识的掌握能力有限，造成相应幻觉。

6 医学大模型发展建议

6.1 丰富通用大模型的医学专业知识

大模型的医学领域知识仍然相对匮乏。通用大模型具备宽广的知识底座，但深度不足。例如，由于缺乏心梗相关知识，大模型无法胜任诊疗任务，只能给出相对宏观的建议（如饮食和体重控制），见图 4。总体而言，通用大模型所训练的语料对医学知识的涵盖仍然相对有限。目前公开可以获取的医学语料主要包括医学书籍、文献。但是大量医学临床数据，如电子病历、医学检验检查等，由于具有较高的私密性而相对难以获取。要解决该问题，一方面要针对隐私敏感数据，私有化训练与部署医学大模型。同时进一步提升医学语料的体量与质量，尤其要注重医学隐私数据的清除、电子病历与医学检查单据的语料化等。另一方面，要注重医学领域结构化知识的利用。医疗领域已经建成很多大规模医疗、疾病、医药与健康知识库。应积极探索、有效利用结构化医学知识进行通用大模型向医学专业大模型的训练。本质上大模型所习得的也是一种知识，如果能充分利用知识库已经富含的知识，就可以降低大模型习得这些知识的门槛与成本，降低专业大模型对大规模高质量语料的需求。目前，通过设计领域知识特异的自监督任务、增强预训练过程，以引导大模型习得专业性较强的领域知识已取得一定研究进展，值得医学领域借鉴。

提问	ChatGPT回答
一名患者双侧膝关节疼痛两年，核磁共振显示双膝退行性骨关节病,部分关节软骨缺失,以内侧股胫间室为著,双膝关节滑膜稍增厚,关节周围软组织肿胀,少量渗液,右膝关节后方游离体形成。追问病史,患者3月前因“急性心梗”行CPI。根据上述病史,应该选择什么样的诊疗方式?	在这种情况下,根据患者的病史和核磁共振结果,可能要综合考虑多个方面来制定诊疗方案。请注意,我不是医生,以下只是一些建议供您参考…… 疼痛管理、非处方药物、物理疗法、关节保护、饮食和体重控制、药物治疗、手术干预、心血管健康管理……

图 4 缺乏医学专业知识的 ChatGPT 难以给出具有专业深度的答复

6.2 提升大模型对医学复杂数据的理解水平

目前大模型在自然语言处理与理解等任务方面取得较好效果。能够处理多模态数据的大模型近期也得到较快发展，但是其在理解专业图像、表格、复杂文档结构等方面仍然存在局限，如在处理基因/蛋白质序列、小分子结构、化学公式、病理图片、分析报告等专业数据方面能力有限。医学领域存在大量复杂诊断单据、医学影像、医学病历，呈现出多模态融合、布局复杂多样、手写体与印刷体混杂、富含数值表达、嵌套结构等特点。医学数据是一类极为复杂的数据，要在大模型训练和应用两个阶段应对医学数据的高度复杂性。这首先需要面向复杂医学数据的预训练机制开展相关研究，特别是结合医学数据特点设计相应的自监督学习机制，如在蛋白质与小分子结构预训练中应体现其化学、生物特性。其次，要在文本之外设计针对文档布局、多模态、数值表达的学习机制，以引导大模型习得文本之外的复杂数据语义。在这一过程中，不同模态数据之间的对齐尤为关键。

6.3 提升大模型对外部医学工具的规划与使用能力

仅凭大模型难以解决现实问题，仍需大模型之外的各类信息系统（如医学数据库、知识库、文献库等）、专业医疗系统（如医学影像系统、医学检测系统）和医疗设备（如手术机械臂）等工具协同。不同外部工具功能不同，API 也不同。为此，首先要进一步提升大模型的 API 理解与规划能力，将复杂任务拆解成原子步骤，并调用相应专业 API 完成执行的能力。当前 AgentGPT、MetaGPT 等开源项目均在推动大模型作为智能体调用多种 API 的能力。其次，需要提升大小模型协同、通用与专业模型协同水平。很多专业任务的医学系统是专业小模型，如医学影像检测模型可能是百万参数的卷积神经网络模型。可以将复杂任务分解为一系列细分任务，应用通用大模型完成通识类语言理解任务，应用专业小模型完成专业医学任务，最后再组合形成最终方案。发挥大小模型各自优势是未来以大模型为核心的智能化解决方案的基本思路。

6.4 提升大模型同理心与共情能力

医生与医疗机器人之间最大的区别在于医生不仅具备医学知识和专业技能，还具有作为社会人的同理心和情感。医患沟通不同于旨在传递信息的沟通。患者就诊时特别渴望医护人员的关爱和体贴，因而对其语言、表情、动作姿态、行为方式更加关注、敏感。例如患者家属伤心时，如果机器人询问是否要听一个笑话，将引起极大反感。再如在向患者传达信息时，“30% 的改善机会”可能比“70% 的失败机会”更易被接受。医生的同理心是医患沟通、传递治疗决策和取得患者及其家属配合的重要前提。作为智能体，大模型不仅需要专业知识和能力，更需要理解人类通过语言所传达的情感和心理状态，并作出合理、富有同理心的答复，才能胜任医疗应用。当前在生成内容时，需要对大模型适度提示，才能生成看似共情的回复。但是总体而言，大模型在稳定、更加人性化地进行共情对话，兼顾医疗答复的精准性与安慰性方面仍然有待进一步研究。

6.5 提升大模型解释能力

主流大模型是基于 Transformer 架构的深度神经网络。而深度模型的不透明、难理解、难解释是限制其应用的主要问题之一。医疗系统面向患者或医生，对可解释性有较高要求。患者需要关于医疗方案的更多解释，而非仅有结果。医学理论家一直尝试揭示疾病机理、药理，建立医学诊治理论体系，其本质上都是对可解释性的追求。医疗应用对大模型的可解释性要求较高，生成的解释必须合理合规（正确且完整）、有理有据（援引适当的专业权威指南、手册）、浅显易懂（以普通患者通俗易懂的方式）、逻辑清晰（前后一致，与患者个体信息相吻合）、层次分明（详略得当、条分缕析）。大模型的可解释性不仅在人机交互中具有重要作用，对大模型的安全性、可控性等问题也有决定性影响。某种程度上，正是大模型的“黑盒”特性导致其安全可控的困难。大模型的可解释性内涵十分丰富，需要从生成结果、生成过程、模型机理、训练过程、数据特性、参数影响等多个维度提升。

6.6 提升大模型可控编辑能力

医生和患者都希望寻求确定性结论，这对智能系统的可控性和可编辑性提出较高要求。大模型本质上是统计模型，其一旦习得某个事实便较难对其修改（更新）、删除，无法从根本上实现知识的可控编辑。但是在医学场景 1% 错误可能性都是难以接受的。与此高度相关的另一个问题是大模型的信念修正。目前大模型在信念方面会犯两种典型错误，一是随着用户不同的反馈而产生信念摇摆，二是虽然用户反馈了错误仍坚持其错误。这两种错误都关乎大模型的信念修正或者信念编辑问题，既要改变其错误信念，也要坚定其正确信念。为此要深入研究大模型和知识图谱深度结合的方案，利用知识图谱的可控、可编辑优点弥补大模型这方面能力的不足。

6.7 实现医学大模型持续更新与自我提升

医学领域的基本事实不断发展，对医学大模型的及时更新提出较高要求。新的医学进展（如新药、新治疗方案）及时融入大模型中，才能提醒医生采取更优的治疗方案。除了基于外部文献的更新之外，大模型能否根据专家反馈进行自我提升与迭代优化也十分重要。医学专家拥有丰富的专业知识和临床经验，其反馈信息对于提高医学模型的性能至关重要：一是可以帮助其修正错误，提高诊断和决策的准确性；二是可以帮助模型跟踪最新进展，持续进行优化和更新；三是可以识别模型在临床中可能出现的错误，帮助消除不确定性。

6.8 建立面向医疗应用的大模型评测体系

大模型经过客观评测才能付诸医学应用，应建立面向医疗应用的大模型评测数据集、指标体系和方法，以下 3 个问题值得特别关注。一是除了大模型应用评测之外，面向大模型训练过程的评测同样重要。医学大模型训练过程涉及众多因素，如自监督学习任务设计、不同类型数据配比、关键训练参数设定、医学大模型应用效果与其“炼制”工艺中

的关键参数关系等，均需要合理评测。二是训练数据集本身的评估评测。尤其要注重从用户隐私、政治偏见、性别/人种/地域歧视等方面建立数据集的安全性与合规性评价体系，从源头上确保大模型合规、安全。三是兼顾知识评测与能力评测。目前大模型评测多针对其掌握知识的程度。但医学大模型要在实际应中发挥价值必须能够胜任真实医疗场景的复杂任务，如读取病理片、分析检查报告、书写病历等，因此其评测要兼顾知识和能力两方面。

6.9　实现医学大模型持续演进

作为自治智能体的大模型，必须具备持续学习、进化演进能力。人类医生经过丰富的临床实践才能胜任工作，基于大模型的智能体也要通过虚拟环境成长、与物理环境交互模拟人类实践。在医学元宇宙等技术推动下，医学智能体虚拟成长环境得到完善。可以预见，大模型驱动下的智能体将在医学元宇宙等虚拟环境中快速进化。在这一过程中，要注重虚拟与现实环境的同步。目前通用的具身多模态大模型已经在日常生活场景中取得一定效果，利用大规模语言模型的规划和思考能力使其能够较好地操控机械身体完成日常工作，如冲调咖啡、清洗餐具等。未来，大模型能否操控专业的手术机器人也将成为重要研究课题。“身心一体”（具身化的大模型）的医学智能体一定程度上可以被视作完整的医生，具备在现实世界演进的可能性。当然，将虚拟环境和现实世界的反馈内化为驱动大模型演进的动力相关研究仍然面临挑战，大模型仍须经历持续演进才能接近医学专家水平，见图 5。

图 5　医学大模型的持续演进

6.10　提升大模型医学推理能力

大模型进一步提升医学推理能力才能实现更好的医疗决策。医学专家的疾病诊断过程有一套严密的诊断逻辑、排查过程、治疗策略、偏好习惯。这种专业推理能力是当前通用大模型所不具备的。除了专业推理，日常推理也是大模型需要进一步完善的。有研究表明，GPT4 在侦探类推理任务中的表现远逊于人类。疾病诊断中的推理本质是按照症状、检测指标等线索推断病因，与侦探类推理任务类似。专业医生在诊治过程中的推理过程往往融合直觉性质的日常推理与传统的形式推理，医生能够根据患者具体情况完成不同形式的推理过程。此外，还须要注意日常健康咨询与疾病诊治等方面中西医思维方式与推理方式的差异。

6.11　提升大模型安全性

相较于通用大模型，医学大模型对输出的安全性有更高要求，医学领域的规章制度、医学伦理合规、用药安全、计量准确等不同方面都对输出的要求极为苛刻。提升大模型的决策合规性尚待深入研究。总体而言大规模生成式语言模型的生成内容精准度与逻辑合理性等仍与人类专家存在显著差距。除此之外，医疗决策反射出不同社会文化与伦理，大模型要对普世价值、道德规范等有一定理解，才能辅助医生决策。

6.12　提升大模型可靠性

大模型本质上是一种统计模型，难以避免会产生事实错误、逻辑错误等问题，存在安全隐患，如较难针对药物剂量安排给出精准答案等，其可靠性有待进一步提升。但在医疗领域对错误零容忍。因此须要发展大模型的兜底方案，特别是人机结合的综合方案，由人类专家对大模型的错误进行审核与纠正。

6.13　提升大模型鲁棒性

大模型输出受到输入指令影响。这一问题在多

模态输入时可能更为严重，如患者的医学影像在不同色差下可能影响大模型产生不同决策结果。这种对输入扰动、输入噪音的敏感，是大模型待解决的问题。在疾病诊断过程中，患者可能使用模糊甚至错误的医学词汇描述病情，这就要求模型对非专业描述的输入具备一定理解能力，防止判断偏差和方差过大造成错误诊断。同时大模型还应该具备抗攻击能力，即模型在接受恶意注入的一些特殊提示词汇时，仍然能够进行正常的医学应用，给出符合事实的答案，这也是医学大模型的鲁棒性要关注的问题。

6.14 提升大模型公平性

大模型的公平性问题包含多个方面。训练数据分布不均所带来的大模型偏见是首要问题。大模型的学习语料往往存在较严重的分布偏差。医疗数据本身可能分布不平衡，例如地方病在不同地区语料中表达丰富性不同，欠发达地区医疗数据可能相对稀缺。这些都会造成大模型呈现出不公平特性。在我国医疗数字化发展基础较好，但是也要充分关注各地区间医疗数字化发展不平衡问题。

6.15 加强数据与用户隐私保护

ChatGPT 从海量数据中进行学习，其生成内容可能包含个人隐私信息，在生成问题答案过程中存在个人隐私泄露风险。大模型使用过程中用户提交的问题也可能涉及隐私。为此一方面应该在数据源头加强治理，确保患者隐私数据不参与训练。另一方面，在大模型使用过程中，模型服务方应在患者明确授权的前提下采集用户数据。大模型的隐私防护困难从根本上来讲与大模型内在机理的不透明高度相关。须要进一步加强大模型的内在机理与可解释研究以提升隐私安全防护水平。

6.16 提升大模型的人类价值对齐能力

在医学和伦理问题上专家之间可能存在不同观点。大模型面临的难题是如何在不同专家观点之间保持客观性，同时提供有用信息，帮助用户理解不同观点并作出判断。某些医学伦理问题存在“边缘地带”，例如安乐死、基因编辑等。这些问题可能涉及道德、法律、文化和宗教等多个维度，不同社会和文化背景下的人可能对这些问题有不同的看法。大模型需要避免倾向性，同时提供关于不同价值观和观点的信息，以帮助用户形成全面理解。一个具备共情能力的医学模型能够更好地理解患者的情感、痛苦以及需求，从而提供更贴近个体化的医疗建议和支持。医学模型应该能够从患者的言语、语气和表情中识别和理解情感状态，从而更好地回应患者的情感需求，对此大模型须要准确对齐医患偏好、价值观、道德倾向与情感诉求。

7 结语

以大规模生成式语言模型为代表的通用人工智能技术浪潮席卷全球，对各行业的高质量发展带来深远影响。医疗行业的智能化发展走上大模型驱动的技术路线势在必行。大模型为医疗智能化带来强大的通识知识基础、高级认知能力，是实现医疗认知智能的基础设施。然而大模型仍然存在幻觉等问题，难以直接胜任医学的复杂决策应用，需要多方面优化才能成为医疗行业高质量发展的先进生产力。未来应进一步发展以大模型为核心的数据驱动与以知识图谱为代表的知识驱动深度融合的双系统认知范式，进一步加强与重视医疗大模型的数据治理与认知评测[16]。

参考文献

1 KUNG T H, CHEATHAM M, MEDENILLA A, et al. Performance of ChatGPT on USMLE: potential for AI - assisted medical education using large language models [J]. PLoS digital health, 2023, 2 (2): e0000198.

2 SARRAJU A, BRUEMMER D, VAN ITERSON E, et al. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat - based artificial intelligence model [J]. Journal of the American medical association, 2023, 329 (10): 842 - 844.

3 AGBAVOR F, LIANG H. Predicting dementia from spontaneous speech using large language models [J]. PLOS digit-

al health, 2022, 1 (12): e0000168.

4 Reddit, r/ChatGPT. A boy saw 17 doctors over 3 years for chronic pain. ChatGPT found the diagnosis [EB/OL]. [2023 - 09 - 15]. https://www.reddit.com/r/ChatGPT/comments/16gfrwp/a_boy_saw_17_doctors_over_3_years_for_chronic/.

5 BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models [EB/OL]. [2023 - 09 - 15]. https://arxiv.org/abs/2108.07258.

6 ZHAO W X, ZHOU K, LI J, et al. A survey of large language models [EB/OL]. [2023 - 09 - 15]. https://arxiv.org/abs/2303.18223.

7 LIY X, LI Z H, ZHANG K, et al. ChatDoctor: a medical chat model fine - tuned on LLaMA model using medical domain knowledge [EB/OL]. [2023 - 09 - 15]. https://arxiv.org/abs/2303.14070v4.

8 THIRUNAVUKARASU A J, TING D S J, ELANGOVAN K, et al. Large language models in medicine [J]. Nature medicine, 2023 (7): 1 - 11.

9 陆玫. 国内首例 5G 超远程机器人肝胆手术：医生在杭州，患者在新疆 [EB/OL]. [2023 - 08 - 15]. https://www.thepaper.cn/newsDetail_forward_21946667.

10 DAVE T, ATHALURI S A, SINGH S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations [J]. Frontiers in artificial intelligence, 2023 (5): 1169595.

11 KHAN R A, JAWAID M, KHAN A R, et al. ChatGPT - reshaping medical education and clinical management [J]. Pakistan journal of medical sciences, 2023, 39 (2): 605.

12 LUO Y, ZHANG J, FAN S, et al. BioMedGPT: open multimodal generative pre - trained transformer for bioMedicine [EB/OL]. [2023 - 09 - 15]. https://arxiv.org/pdf/2308.09442.pdf.

13 LIN X, XU C, XIONG Z, et al. PanGu Drug Model: learn a molecule like a human [EB/OL]. [2023 - 09 - 15]. https://link.springer.com/article/10.1007/s11427 - 022 - 2239 - y.

14 LISANZA S L, GERSHON J M, TIPPS S W K, et al. Joint generation of protein sequence and structure with RoseTTAFold sequence space diffusion [EB/OL]. [2023 - 09 - 15]. https://www.biorxiv.org/content/10.1101/2023.05.08.539766v1.

15 JEONG J, TIAN K, LI A, et al. Multimodal image - text matching improves retrieval - based chest X - Ray report generation [EB/OL]. [2023 - 09 - 15]. https://arxiv.org/pdf/2303.17579.pdf.

16 肖仰华. 不要让大模型变成一场华丽的烟花秀 [EB/OL]. [2023 - 09 - 15]. https://m.thepaper.cn/newsDetail_forward_23000221.

2024 年《医学信息学杂志》征订启事

《医学信息学杂志》是国内医学信息领域创刊最早的医学信息学方面的国家级期刊。主管：国家卫生健康委员会；主办：中国医学科学院；承办：中国医学科学院医学信息研究所。中国科技核心期刊（中国科技论文统计源期刊），RCCSE 中国核心学术期刊（武汉大学中国科学评价研究中心，Research Center for Chinese Science Evaluation），美国《化学文摘》《乌利希期刊指南》及 WHO 西太区医学索引（WPRIM）收录，并收录于国内 3 大数据库。主要栏目：专论，医学信息技术，医学信息研究，医学信息资源管理与利用，医学信息教育等。读者对象：医学信息领域专家学者、管理者、实践者，高等院校相关专业的师生及广大医教研人员。

2024 年《医学信息学杂志》国内外公开发行，每册定价：15 元（月刊），全年 180 元。邮发代号：2 - 664，全国各地邮局均可订阅。也可到编辑部订购：北京市朝阳区雅宝路 3 号（100020）医科院信息所《医学信息学杂志》编辑部；电话：010 - 52328672，52328686，52328687。

《医学信息学杂志》编辑部