科学通报 2025 年 第 70 卷 第 26 期: 4560 ~ 4568
评 述
《中国科学》杂志社
SCIENCE CHINA PRESS

卫生健康行业垂直大模型破茧之基石——构建行业专业多模态语料库

沈剑峰[1*], 黄茹[1], 闵栋[1,2], 车慧[1], 李宝山[1], 刘丽红[1,3], 张智[4], 程京[4,5*], 王杉[1,6*]

1. 中国医学装备协会智慧医院分会人工智能研究小组, 北京 100082
2. 中国信息通信研究院云计算与大数据研究所, 北京 100083
3. 北京大学人民医院信息中心, 北京 100044
4. 生物芯片北京国家工程研究中心大数据与人工智能研究院, 北京 102206
5. 清华大学生物医学工程学院, 北京 100084
6. 北京大学人民医院临床医学大数据研究中心, 外科肿瘤研究室, 北京 100044
* 联系人, E-mail:sjf_hz@126.com; jcheng@tsinghua.edu.cn; shanwang@pkuph.edu.cn

2025-02-22 收稿, 2025-06-29 修回, 2025-07-04 接受, 2025-07-15 网络版发表

摘要 在卫生健康行业的复杂应用场景中, 生成式人工智能大语言模型技术(大模型)的专业领域适应性限制了大模型在医疗卫生、金融等专业领域的广泛应用和创新能力, 目前自监督学习依赖的开放语料也难以满足医疗卫生领域高精度、高特异性的专业要求. 大模型通过卫生健康行业多模态语料的训练, 生成卫生健康行业垂直大模型. 该垂直大模型具备专业领域的知识和针对性解决医疗卫生问题的能力, 满足卫生健康行业的专业应用需求, 达到通用大模型无法替代的专业性、高效性和精度. 本文系统梳理国内外卫生健康行业的语料库建设模式、技术实现及其不足, 提出以"疾病–场景关联矩阵"为核心的标准化框架, 通过任务匹配机制实现疾病分类与医疗卫生场景之间的多维映射. 同时提出构建涵盖数据采集、质量评估、数据标注与隐私保护等关键环节的专业多模态语料库标准体系, 形成任务驱动、分级适配的多层次多模态语料资源结构. 通过建立质量控制与反馈闭环机制, 实现行业多模态语料库的动态优化与持续迭代, 为构建覆盖医疗卫生领域业务需求高质量数据的卫生健康行业多模态语料库提供系统性方法与理论支撑.

关键词 生成式人工智能, 大语言模型, 垂直大模型, 语料, 多模态语料库, 卫生健康行业

2022年11月, Open AI公司推出的ChatGPT引发了全球范围内的广泛关注. 2025年1月DeepSeek R1横空出世, 揭开了生成式人工智能(Generative Artificial Intelligence, Generative AI)大语言模型(large language model, LLM)(简称: 大模型)的新时代序幕和卫生健康行业大模型应用热潮. 生成式人工智能大模型(如ChatGPT系列、DALL·E2和Magic3D等)凭借其卓越的语言理解能力、多模态内容生成能力和复杂任务自动化处理能力[1], 被广泛用于文本创作[2]、语言翻译[3]、创意设计[4]与代码生成[5]等任务场景. 大模型发展遵循人工智能领域发展的"规模定律"(scaling law), 即模型能力与模型规模、数据集大小和训练所需计算量之间呈幂律关系, 随着这三个因素的指数级增长, 模型表现显著提升[6]. 在大模型的构建过程中, 算力、算法、数据(语料)和应用场景构成了四大核心基础要素, 为模型的研发与应用提供重要支撑. 其中, 算力为动力, 依托高性能硬件(如GPU (graphics processing unit)、TPU (tensor processing unit))支撑大规模的大模型训练与高

引用格式: 沈剑峰, 黄茹, 闵栋, 等. 卫生健康行业垂直大模型破茧之基石——构建行业专业多模态语料库. 科学通报, 2025, 70: 4560–4568
Shen J, Huang R, Min D, et al. The foundational cornerstone for healthcare AI models: constructing multimodal corpora in the health sector (in Chinese). Chin Sci Bull, 2025, 70: 4560–4568, doi: 10.1360/TB-2025-0185

效推理; 算法为核心, 决定大模型架构与能力; 数据(语料)为基础, 直接决定大模型语言能力、理解力、生成能力[7]; 应用场景则指引大模型设计和性能优化方向. 近年来大量实践证明, 大模型发展与用于训练的语料有密切关系[8]. 行业领域垂直大模型的训练和应用进一步验证了专业、规范语料的关键性和重要性. 为更好推进卫生健康行业垂直大模型的应用发展, 明确语料需求和规范极为重要.

1 语料是大模型训练应用的核心基础

语料(corpus)一词源起于语言学, 通常指一定数量和规模的文本资源集合. 在大模型发展过程中, 语料是指用于大模型预训练与微调的数据集合, 同时涵盖验证与评估过程中使用的独立测试数据. 语料涵盖多领域、多语种与多模态数据, 为大模型提供知识来源与模式学习基础, 能够全面提升大模型的语言表征能力, 具体包括语义理解、语法掌握与上下文关联建模等.

国内外人工智能大模型在语料的需求与构建方面存在差异. 国外语料建设更侧重于多语言、多领域的开放数据, 并强调隐私合规性, 如遵循《通用数据保护条例》(general data protection regulation, GDPR), 同时依赖开源社区. 国内受语言环境和政策影响, 更聚焦中文语境与特定行业, 强调数据安全与本地化管理, 语料多由企业与科研机构自建, 且偏重垂直领域应用(vertical large model, VLM)(表1).

国外大模型在语料构建方面注重开放性、多样性与高质量控制. 例如, Open AI的GPT-4o采用多模态训练数据, 包括图像、代码与文本等多领域高质量数据, 总训练数据量超过25万亿tokens (语言最小单元), 在多模态理解、安全性控制与复杂任务执行方面表现卓越[9]. Meta的LLaMA 3.3系列大模型采用约15万亿tokens的开放训练数据, 包含4倍于前代的代码数据量, 通过严格的质量控制确保数据高质量, 大模型在多语言处理、代码生成与指令遵循等任务中展现出色性能. Google的Gemini 2.0模型整合搜索、图像识别与地图导航等多元功能, 采用更大规模的训练数据, 并引入Flash架构(高效低延迟的模型执行结构)提升推理效率, 在日常任务处理与复杂场景应用中表现突出.

国内大模型更注重本地化与行业化应用. “深度求索”的DeepSeek-V2模型, 其训练数据集包含8.1万亿tokens的高质量多源语料库, 覆盖医疗、法律等行业领域数据, 在自然语言处理、代码生成、数学推理与对话

表 1 国内外生成式人工智能通用大模型与语料概况

Table 1 Overview of general-purpose generative AI models and corpora domestically and internationally

年份	模型名称	发布主体	参数规模	语料规模	语料来源	数据模态	语料特点
2020	GPT-3	OpenAI	1750亿	3000亿tokens	Common Crawl、Books、Wikipedia	文本	多领域、高质量文本语料
2022	GPT-3.5	OpenAI	1750亿	3000亿tokens	人工标注数据、对话优化数据集	文本	微调数据集、人类反馈强化学习(RLHF)
2023	GLM 130B	清华大学	1300亿	2万亿tokens	中英双语平衡语料	文本	跨语言能力、少样本学习
2023	DeepSeek-V2	深度求索	2360亿	8.1万亿tokens	医学、法律、金融等多领域数据集	文本、多模态	专业化语料占比高
2024	LLaMA 3.3	Meta	未公开	15万亿tokens	开放数据、筛选过滤	文本	高质量、去噪、多领域覆盖
2024	GPT-4o	OpenAI	1.5~2万亿	未公开	公共互联网、专有数据、多模态	文本、图像、代码	包含多模态数据, 覆盖多语言和领域
2024	文心一言4.0	百度	1万亿	未公开	中文语料库+知识图谱	文本	中文为核心, 多领域数据
2024	PaLM-2	Google	5400亿	3.6万亿tokens	多语言文本、科学、医疗	文本	多语言支持, 覆盖100+语言
2024	Claude 3 Opus	Anthropic	1.5万亿	未公开	开放语料、筛选数据	文本	关注伦理性、上下文长依赖
2025	Gemini 2.0	Google	未公开	未公开	多模态数据、开放语料库	文本、图像、音频	长上下文处理、特定任务应用
2025	Qwen 2.5 Max	阿里巴巴	未公开	未公开	跨模态数据(语言、视觉、声音)	文本、图像、音频	多模态融合, 多场景覆盖
2025	Kimi 1.5	月之暗面	未公开	未公开	公共互联网、专有数据、多模态	文本、图像、OCR	涵盖图表解析、视觉编码与推理
2025	Doubao1.5	字节跳动	未公开	未公开	公共互联网、专有数据、多模态	文本、图像	高质量、独立性、多模态
2025	DeepSeek-V3	深度求索	6710亿	14.8万亿tokens	多源大规模语料库(多领域)	文本、图像	多语言、多领域语料库

方面性能卓越. 随后发布的DeepSeek-V3、R1和Janus-pro大模型进一步展现了技术创新. 其中, V3大模型总参数量为6710亿, 每次推理激活370亿参数, 训练集包含14.8万亿tokens, 在数学推理、长文本处理与中文任务中表现突出; R1大模型通过强化学习训练展现出与OpenAI-o1相当的推理能力; Janus-Pro突破采用分离视觉编码的方法, 在多模态理解与生成任务上均取得出色表现. “字节跳动”的豆包大模型采用高质量多源数据, 涵盖中文互联网、专业文献和开源代码, 在中文理解和创意写作方面表现优异. “月之暗面”的Kimi大模型整合专业领域知识库与互联网数据, 通过强化知识推理能力, 在专业问答与逻辑分析任务中表现突出. “百度”的文心一言4.0大模型采用超大规模中文语料库, 融合多行业专业数据, 在中文创作、知识问答与多轮对话中具有显著优势. 阿里推出的Qwen 2.5 Max注重多模态数据(语言、视觉、声音)的融合, 语料针对电商、智能客服等场景优化, 大模型在多模态理解和场景化任务执行中表现独特. 清华大学的GLM系列(如GLM-130B)采用中英双语平衡语料, 总训练数据量约为4000亿tokens, 专注提升跨语言处理能力, 尤其在少样本学习任务中表现突出.

总体而言, 国内大模型强调语料行业专属性与中文语境覆盖范围, 强化领域适配能力. 明确语料在大模型训练应用中的核心地位, 有助于深入理解垂直领域语料在人工智能落地过程中的关键作用.

2 行业语料是行业垂直大模型的关键

通用大模型在处理通用任务时表现出色. 根据斯坦福大学基础模型研究中心(center for research on foundation models, CRFM)创建的评测平台“语言模型整体评估体系”(holistic evaluation of language models, HELM)排行榜(2024年12月16日数据), 通用大模型的总体准确率达94.3%. 然而, 在医疗、法律等专业任务中, 由于通用大模型在领域知识与语料覆盖上的不足, 其可靠性显著受限. 仍以HELM排行榜(2024年12月16日数据)为例, 其在医学任务中的准确率最高为86.3%.

目前, 通用大模型在专业领域普遍面临以下挑战: 准确性不足、偏见[10,11]、幻觉生成[12,13]与长尾知识覆盖缺失问题[14]. 造成此类问题的主要原因在于其训练语料中缺乏高质量的领域特定语料, 而该问题已成为制约通用大模型在行业落地应用中的关键瓶颈之一[15].

行业垂直大模型是指在特定领域或任务中, 基于大规模的行业语料训练或在基础大模型上进行定向微调, 从而对该领域任务进行高效与精准处理的大模型, 表2对比了通用大模型(如GPT系列)与垂直大模型(如医疗、法律、金融模型)在各自领域任务中的任务表现, 结果显示, 行业垂直大模型在专业任务上的性能普遍优于通用大模型, 体现出其在实际应用中的显著优势.

行业垂直大模型的发展不仅需要语料具备高度的专业化、结构化与规范性, 还需要具备动态更新机制以适应快速变化的行业知识. 同时, 其数据来源应体现鲜明的行业特征与场景关联性. 高质量行业语料资源是提升垂直大模型应用效果的关键要素.

3 卫生健康行业垂直大模型对语料提出更高要求

卫生健康行业的垂直大模型, 是指利用大模型技术, 结合卫生健康行业特定语料和任务进行训练或优化, 进而具备专业知识结构与任务处理能力的大模型[16]. 相较于其他行业, 卫生健康行业因其“生命至上”的特殊属性, 对人工智能应用中假阳性与假阴性结果的容忍度极低[17]. 这对大模型训练所依赖的高质量、多模态语料资源提出了更为严苛的要求.

卫生健康行业多模态语料库, 是指为支持垂直大模型在卫生健康行业研究与应用而专门构建的、覆盖大模型训练全流程的数据资源集合. 其数据来源于卫生健康行业中多源异构系统, 具有多模态(modality)与多维度特征, 覆盖预训练、指令微调、偏好优化到性

表 2 通用大模型与行业垂直大模型的领域任务性能对比(根据公开数据整理)

Table 2 Comparison of domain task performance between general-purpose and industry-specific large models (based on public data)

模型	MedQA	LawBench	FinEval
GPT系列(通用, 30%行业数据+70%通用数据)	60.2%	62.1%	70.5%
Med–PaLM(医疗, 80%医疗数据+20%通用数据)	67.2%↑	–	–
ChatLaw–33B(法律, 主要为法律领域语料)	–	78.4%↑	–
INF–Fin(金融, 51%金融数据+49%通用数据)	–	–	92.4%↑

能评估的全流程. 在模态类型上, 语料不仅包含结构化或非结构化文本, 还包括医学图像、音频、视频等多种数据形式, 如临床病历、医学影像资料、手术视频、患者听诊音频等. 这些异构、互补的信息源为大模型提供了更加丰富、全面与语义一致性强的语料基础, 从而有助于提升其在复杂医疗卫生任务中的表现与可解释性.

3.1 卫生健康行业人工智能语料的发展演变

随着深度学习与大模型技术的不断发展, 训练数据集对质量、规模、格式和多样性的要求显著提升. 在2010~2015年间, 人工智能技术开始逐步在卫生健康行业落地应用, 以传统机器学习为代表, 训练数据主要依赖于结构化数据(如临床诊疗记录、基因组测序数据)与非结构化数据(如医学影像、自由文本病历). 该阶段强调通过数据清洗、标准化处理与特征工程, 提升数据质量, 为模型构建与性能优化提供基础支持. 进入2015~2020阶段, 随着深度学习技术结合自然语言处理(natural language processing, NLP)的融合, 卫生健康行业人工智能迈入新阶段. 此时期, 高质量标注语料成为大模型构建的核心资源. 其中, 文本类数据(如医学文献、电子病历)广泛用于医学命名实体识别、疾病自动分类和临床诊断文本生成; 而对话数据、患者监测数据则用于为医疗问答系统、情感识别模型与健康管理应用提供语义基础. 此类语料在标注阶段对准确性要求极高, 以确保大模型在特定医疗任务中的理解与推理能力. 当前, 随着多模态学习与大模型架构的高速发展, 医学语料逐渐从单一模态拓展为多模态体系, 涵盖文本、图像、音频、时序信号等多种数据类型, 呈现出多源异构、多维结构特征.

3.2 国内外卫生健康行业人工智能语料库进展

在全球范围内, 卫生健康行业的大规模语料资源建设持续推进, 多个高质量数据库已成为推动医疗卫生人工智能研究的重要支撑.

国外, 美国构建了多个具有代表性的语料资源. 例如, VinDr-CXR数据集专注于胸部X光肺部疾病检测; RiTeK数据库支持复杂医学推理任务的建模与评估; MIMIC-Ⅲ和MIMIC-Ⅳ系列电子健康记录(electronic health records, EHR)数据集, 广泛应用于医疗自然语言处理与疾病预测建模等研究. 针对医学问答任务, MedQA、MultiMedQA、PubMedQA等数据集支持大模型基于权威医学文献与临床知识回答复杂问题. 此外, 英国的UK Biobank也持续扩展数据类型, 新增心血管磁共振影像与基因组注释数据, 为心血管疾病的人工智能研究提供丰富数据基础.

国内, 面向大模型训练的智慧诊疗语料库融合多种复杂疾病症状与诊疗措施, 为辅助诊疗与风险预警系统提供支撑. 上海市卫生健康委员会牵头发布的卫生健康行业语料库涵盖临床医学和公共卫生两大领域, 数据总量约3TB, 主要服务于传染病预警与慢性病管理需求. 上海交通大学构建的多语言医学语料库MMedC, 支持医学语言模型的多语言训练[18].

尽管现有卫生健康行业的语料库在规模与模态上不断丰富, 但在实际应用中仍面临人群多样性不足、实时性和动态性较弱、隐私合规难度高、跨领域数据融合能力有限等问题, 从而限制了在卫生健康行业精细化任务与个性化应用中的有效支持(表3).

表3总结了卫生健康行业典型语料库的关键信息. 从时间维度看, 2016年发布的MIMIC-Ⅲ至2024年的MMedC, 领域语料资源逐步积累. 从国家分布看, 美国主导了如MIMIC系列、MedQA等核心资源, 中国以MMedC为代表的数据库展现出强劲发展趋势. 这些数据库在数据规模、模态类型与应用方向上各具特色, 为多项医学研究和人工智能模型训练提供了支撑.

3.3 卫生健康行业人工智能语料库的语料特征

3.3.1 多源异构

卫生健康行业人工智能语料库的数据来源广泛, 包括公开数据集合(政策法规、医学书籍、文献期刊、指南共识)以及医疗发展领域不同时期、不同组织建设的多源异构系统中的专业性和高质量数据集. 多源异构系统来源包括医疗机构与卫生健康管理部门、商业保险机构与医保管理部门、科研机构(临床研究项目)、医药研发企业与监督管理部门(临床试验项目)、医药供销企业、大健康服务企业(体检中心、慢病中心等)等不同机构建设的相关信息系统. 卫生健康行业多模态语料库的语料来源于这些专业领域的信息系统数据, 按照人工智能模型使用要求(预训练数据集、指令微调数据集、偏好数据集、评估数据集、传统自然语言处理数据集)进行数据加工(不标注、弱标注或精细标注).

3.3.2 多维度[19]

卫生健康行业人工智能语料库的数据内容丰富、

表 3 卫生健康行业的共享数据库

Table 3 Shared databases in the healthcare and public health domain

年份	国家/地区	数据集名称	数据类型	模态	规模	特点	典型应用
2016	美国	MIMIC-Ⅲ	电子病历	文本	40000名患者	ICU临床数据, 覆盖多种医疗事件	临床预测、诊断建模
2017	美国	NIH Chest X-ray	医学影像	影像	112120张	大规模胸部X光影像, 14种疾病分类	肺部疾病检测
2018	美国	MedNLI	医学推理	文本	14049对	基于MIMIC-Ⅲ, 医学推理任务	推理能力评估
2018	美国	eICU	电子病历	文本	200000次入院	ICU数据, 多中心、多源异构	ICU临床决策
2019	美国	PubMedQA	问答数据集	文本	273000对	基于PubMed文献, 支持医学QA	医学问答系统
2020	越南	VinDr-CXR	医学影像	影像	18000张	高质量标注肺部X光影像	肺部疾病检测
2020	美国	MIMIC-IV	电子病历	文本	50000名患者	ICU临床数据, 更新版	ICU临床预测
2021	美国	MedQA	医学考试题集	文本	61000条	医学执照考试题目, 涵盖中、美、台地区	医学教育、模型测试
2023	美国	MultiMedQA	医学问答	文本	200000条	综合问答集, 整合多数据源	问答系统、推理任务
2024	中国	MMedC	多语言医疗文本	文本	255亿tokens	涵盖中、英、日等六种语言	多语言医学大模型
2024	美国	RiTeK	医学推理	知识图谱	15557条	专注复杂推理任务, 知识图谱数据	推理与知识整合
2024	美国	MedCalc-Bench	医学计算	文本	10055条	医疗计算任务, 评估计算能力	医疗计算任务评估

特征向量多, 包括病历(门急诊病历、住院病历、体检档案、实验室检查、影像学检查、病理检查、其他辅助检查、药物治疗、手术治疗、其他治疗、不良事件)、健康档案(基本信息、健康体检表、重点人群健康管理记录、其他卫生健康服务记录、居民健康信息卡、预防接种、传染病与突发公共卫生)、组学(基因组学、转录组学、蛋白质组学、代谢组学、糖组学、脂组学、免疫组学、影像组学、 超声组学、其他组学)、生物样本、随访、健康生活(生活质量与功能状态数据、健康指标与疾病管理效果、个人生活方式与行为数据、健康知识与心理健康数据)、卫生经济(人口与健康数据、医疗资源与服务数据、医疗费用与成本数据、医疗质量与效果数据、医疗保险与融资数据、经济与社会数据)、机构运营管理(战略与学科发展、财务管理、人力资源管理、医疗流程与质量管理、药品与耗材管理、医疗装备管理、后勤管理、科研管理、教学管理)、政策法规、医学书籍、文献期刊、指南共识、医疗常规等十三大类.

3.3.3 多模态[20]

包括通用类与专用医学类. 通用类语料的模态包括文本、图像、视频、音频、3D的五大主流模态, 以及扩展模态, 如时序数据、地理信息数据等类别. 专用医学类语料的模态包括临床病历、医学影像(放射、超声、内镜、显微成像)、数字病理、生理功能检测(如脑电、心电、肌电、皮肤电等)、生物声学信号(如心肺音、肠鸣音)、生命体征监测(离散型指标与连续流数据)、多组学数据、人体3D模型(如人体器官、医疗器械)等类别.

3.4 卫生健康行业人工智能多模态语料的问题与挑战

卫生健康行业作为与人类生命健康息息相关的核心领域, 其特殊性主要体现在两个方面: 一是生命科学体系的高度复杂性和动态演变特征; 二是疾病类型的多样性以及患者个体差异的显著性. 这一双重特性不仅增加了医学研究与诊疗决策中的不确定性, 也对人工智能模型在该领域的语料支撑与泛化能力提出了更高要求.

在此背景下, 构建面向多样化医疗场景的卫生健康行业大模型, 亟需建立涵盖广泛疾病类型、多源异构与多模态数据的高质量语料库, 以提升大模型在准确率、灵敏度与特异性等关键指标上的表现, 并支撑其独立应用于真实医疗场景的能力的预期标准. 然而, 当前卫生健康行业人工智能多模态语料构建面临以下四个问题.

(1) 卫生健康行业的语料极为重要又非常短缺. 一方面, 医疗数据开放程度有限, 数据采集难度较大; 另一方面, 隐私保护政策严格限制数据访问与共享[21], 进一步加剧了高质量语料的获取难度. 其中, 低频事件数据匮乏进一步影响模型预测灵敏度和特异性[22], 如在罕见病场景中, 数据的长尾分布导致模型性能显著受限, 可通过合成数据来增强模型对罕见病的识别能力, 减少漏诊和误诊风险, 提升大模型在长尾分布下的表现[23]. 随着医疗数据维度(即数据集中包含的变量或特

征)增加, 携带此类特征的特定组合人数减少甚至消失, 从而产生“数据集盲点”(dataset blind spots), 即没有观察到的特征空间(特征或变量所有可能组合的集合)[20]. 这些盲点区域限制了监督学习模型对实际医学决策场景的覆盖范围, 从而影响大模型的泛化能力.

(2) 数据质量问题成为语料构建中的关键瓶颈. 数据质量贯穿于数据采集、标注、存储和整合的全过程, 直接关系到大模型训练与推理的有效性. 在卫生健康行业, 数据质量问题表现为标注误差、噪声干扰、缺失值与数据异构性等多个层面. 医疗数据标注不仅对专业性与一致性要求极高, 细微偏差即可对大模型预测产生级联放大效应[24]. 此外数据噪声(如影像伪影、文本记录错误)与缺失值进一步增加数据处理难度, 尤其在病情动态变化情景下, 时间序列数据的不稳定性加剧高质量语料构建难度. 与此同时, 不同医疗机构、科室与设备采集的数据在格式、标注规范和时间戳等方面缺乏统一标准, 导致数据异构性显著, 这不仅增加了数据整合成本, 还削弱了语料库的一致性与可用性.

(3) 多模态医疗数据的异构性和整合难度大, 不同模态数据在采集设备、存储格式、时空维度等方面存在显著差异. 医疗数据涵盖文本、影像、信号和基因组等多种模态, 时间与空间对齐不准确可能导致信息丢失或错配, 不同模态间的信息冗余与冲突(如文本记录与影像诊断结论差异)[20]. 在跨模态学习任务中, 心电图信号的高频采样与分钟级护理记录之间的时间轴错位会导致大模型难以有效关联异质数据来源. 此外, 磁共振影像与基因组数据的维度差异可能导致大模型在特征整合过程中发生“特征覆盖”(feature overwriting)现象, 即高维模态中的次要特征(如基因组突变位点)被主导模态(如影像特征)覆盖, 从而削弱大模型对罕见特征的敏感性. 多模态数据内容差异增加了数据标准化和统一表征的难度[25]. 此外, 不同模态数据之间存在复杂的语义关联和互补性, 保持各模态特有信息且实现有效融合, 构建统一的语义理解框架, 是多模态医疗语料建设面临的重要挑战.

(4) 隐私性和安全性是语料库建设的关键因素, 现有隐私保护技术(如k-匿名化、联邦学习、合成数据)在保护敏感数据的同时, 往往会降低数据效用, 表现为典型的技术性两难问题[26]. 尽管联邦学习通过不直接共享原始数据方式降低隐私泄露风险, 但传输的模型梯度仍可能泄露部分训练数据, 攻击者可通过逆推梯度信息重建部分原始样本, 存在隐性知识泄露风险[27]. 合成数据(如基于StyleGAN的生成数据)在一定程度上缓解了数据稀缺问题, 但由于生成大模型在高频细节信息上的拟合不足或过拟合, 生成数据的统计特征可能偏离真实数据, 从而导致下游大模型预测结果偏差[28]. 此外, 医疗语料库通常包含患者姓名、身份证号、联系方式与病史记录等敏感信息, 这些信息一旦泄露, 不仅会给患者带来隐私侵犯, 还可能引发身份盗用与健康欺诈等安全风险.

4 展望

卫生健康行业多模态语料库是智能医疗与精准医疗的基石, 为新一代信息技术(如大模型)在行业中的深度应用奠定核心数据基础. 其构建路径如图1所示.

首先, 语料库的建设应由政府主导, 聚焦重大疾病与创新领域, 构建 “疾病-场景关联矩阵”标准化框架, 通过任务匹配机制打通疾病分类与场景需求. 其中“疾病”依据国际疾病分类(如ICD-10)并结合国内实际需求, 动态细化病种层级、诊断特征与多模态数据表现; “场景”参照《卫生健康行业人工智能应用场景参考指引》, 明确场景数据规范和标注规范.

在具体实施过程中, 以“疾病-场景关联矩阵”框架指引, 优先覆盖高发与重点病种(如“肺结节”), 通过矩阵框架映射至专项任务(如“肺结节AI辅助诊断”), 系统性指导任务生成、数据采集与标注分发. 建设过程中通过专业化团队分工、模态数据对齐与隐私合规控制, 确保语料建设过程的高质量、高效率. 框架应具备动态更新能力, 通过场景深化与任务持续迭代扩展“疾病谱”覆盖范围, 逐步构建从预训练通用语料到专科数据与专项任务语料的多层级语料资源体系.

其次, 构建覆盖数据采集、质量评估、数据标注与隐私保护等关键维度的语料库标准体系. 在数据采集方面, 应规范数据格式与接口协议, 优先采用HL7与FHIR等国际标准, 并融合疾病分类(如ICD-10)与UMLS等权威术语体系, 以确保异构数据在结构与语义层面的统一性. 在质量评估方面, 应建立由完整性、准确性、时效性与一致性构成的评价体系, 并结合自动化检测与人工复核机制, 提升语料的可信度与实用性. 在数据标注方面, 应制定层级清晰、规则明确的标注规范, 结合BioBERT、U-Net等模型实现自动化标注, 并引入专家审核环节, 构建高一致性的标准化标注体系. 在隐私保护方面, 应建立符合法规要求的数据脱敏与去标识化处理机制, 并通过加密存储与权限控制保障

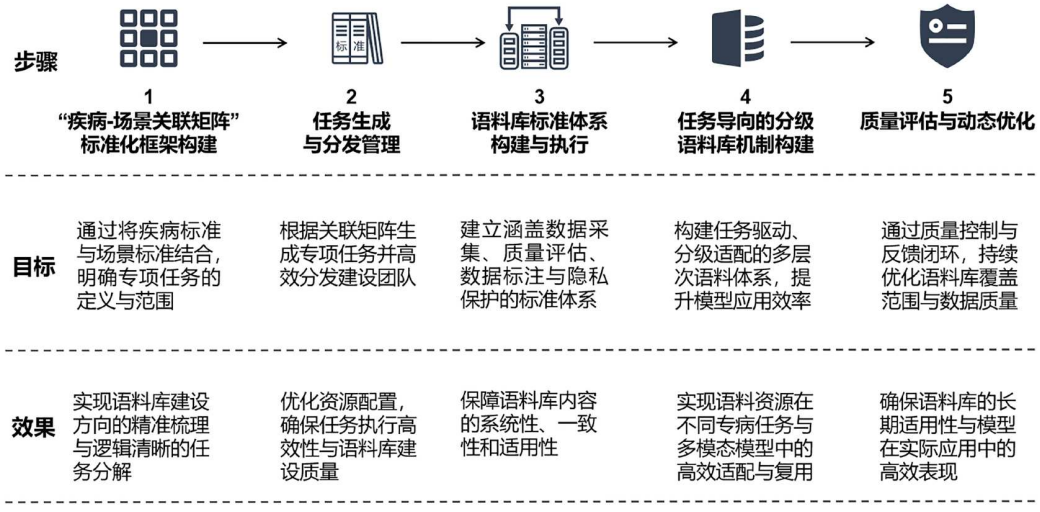

图 1 卫生健康行业人工智能多模态语料库的构建路径展望.该示意图展示了语料库构建路径, 包括: 步骤1, “疾病-场景关联矩阵”标准化框架构建; 步骤2, 任务分发与处理; 步骤3, 语料库标准体系构建与执行; 步骤4, 任务导向的分级语料库机制构建; 步骤5, 质量评估与动态优化

Figure 1 Prospects for constructing multimodal AI corpora in the healthcare and public health domain. This diagram illustrates the construction pathway of the corpus, including: (1) construction of a standardized framework for disease-scenario association matrix; (2) task assignment and processing; (3) development and implementation of a corpus standardization system; (4) construction of a task-oriented hierarchical corpus mechanism; (5) quality evaluation and dynamic optimization

数据全生命周期的安全性和可控性.

最后, 通过市场化开放机制和多方资源调动, 协同开展各专病、专项任务的语料库建设, 从特定专科入手, 如眼科、病理科等, 构建专科语料库, 逐步扩展到综合语料库. 同时在模态建设上, 先从文本数据入手, 再逐步扩展到多模态数据, 如医学影像、音频等, 建立多机构协同治理机制, 确保数据安全和隐私保护, 实现数据的共享和协同利用. 同时建立质量控制和反馈闭环机制以实现语料库的动态优化, 逐步构建全面深度覆盖、数据高质量、系统完整的卫生健康行业人工智能多模态语料库(图1), 为卫生健康行业垂直大模型开发、应用和发展奠定全新的、坚实的数据和专业基础.

参考文献

1 Drazen J M, Kohane I S, Leong T Y, et al. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med, 2023, 388: 1233–1239
2 Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. Science, 2023, 381: 187–192
3 Costa-jussà M R, Cross J, Çelebi O, et al. Scaling neural machine translation to 200 languages. Nature, 2024, 630: 841–846
4 Epstein Z, Hertzmann A, Akten M, et al. Art and the science of generative AI. Science, 2023, 380: 1110–1111
5 Romera-Paredes B, Barekatain M, Novikov A, et al. Mathematical discoveries from program search with large language models. Nature, 2024, 625: 468–475
6 Wu T, He S, Liu J, et al. A brief overview of ChatGPT: the history, status quo and potential future development. IEEE CAA J Autom Sin, 2023, 10: 1122–1136
7 Yu X, Zhang Z, Niu F, et al. What makes a high-quality training dataset for large language models: a practitioners’ perspective. In: Proc 39th IEEE/ACM Int Conf Autom Softw Eng. Sacramento: ACM, 2024. 656–668
8 Zhou L, Schellaert W, Martínez-Plumed F, et al. Larger and more instructible language models become less reliable. Nature, 2024, 634: 61–68
9 Sanderson K. GPT-4 is here: what scientists think. Nature, 2023, 615: 773
10 Drazen J M, Ferryman K, Mackintosh M, et al. Considering biased data as informative artifacts in AI-assisted health care. N Engl J Med, 2023, 389: 833–838
11 Hofmann V, Kalluri P R, Jurafsky D, et al. AI generates covertly racist decisions about people based on their dialect. Nature, 2024, 633: 147–154
12 Farquhar S, Kossen J, Kuhn L, et al. Detecting hallucinations in large language models using semantic entropy. Nature, 2024, 630: 625–630
13 Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. ACM Comput Surv, 2023, 55: 1–38

14 Kandpal N, Deng H, Roberts A, et al. Large language models struggle to learn long-tail knowledge. In: Proc 40th Int Conf Mach Learn. PMLR, 2023. 15696–15707
15 Kong Y, Nie Y, Dong X, et al. Large language models for financial and investment management: applications and benchmarks. JPM, 2024, 51: 162–210
16 Shah N H, Entwistle D, Pfeffer M A. Creation and adoption of large language models in medicine. JAMA, 2023, 330: 866–869
17 Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. npj Digit Med, 2021, 4: 65
18 Qiu P, Wu C, Zhang X, et al. Towards building multilingual language model for medicine. Nat Commun, 2024, 15: 8384
19 Acosta J N, Falcone G J, Rajpurkar P, et al. Multimodal biomedical AI. Nat Med, 2022, 28: 1773–1784
20 Topol E J. As artificial intelligence goes multimodal, medical applications multiply. Science, 2023, 381: eadk6139
21 Mehta M C, Katz I T, Jha A K. Transforming global health with AI. N Engl J Med, 2020, 382: 791–793
22 Shumailov I, Shumaylov Z, Zhao Y, et al. AI models collapse when trained on recursively generated data. Nature, 2024, 631: 755–759
23 Ktena I, Wiles O, Albuquerque I, et al. Generative models improve fairness of medical classifiers under distribution shifts. Nat Med, 2024, 30: 1166–1173
24 Rädsch T, Reinke A, Weru V, et al. Labelling instructions matter in biomedical image analysis. Nat Mach Intell, 2023, 5: 273–283
25 Baltrusaitis T, Ahuja C, Morency L P. Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell, 2018, 41: 423–443
26 Ziller A, Mueller T T, Stieger S, et al. Reconciling privacy and accuracy in AI for medical imaging. Nat Mach Intell, 2024, 6: 764–774
27 Zhu L, Liu Z, Han S. Deep leakage from gradients. Adv Neural Inf Process Syst, 2019, 32: 14747–14756
28 Murtaza H, Ahmed M, Khan N F, et al. Synthetic data generation: state of the art in health care domain. Comput Sci Rev, 2023, 48: 100546

Summary for “卫生健康行业垂直大模型破茧之基石——构建行业专业多模态语料库”

The foundational cornerstone for healthcare AI models: constructing multimodal corpora in the health sector

Jianfeng Shen[1*], Ru Huang[1], Dong Min[1,2], Hui Che[1], Baoshan Li[1], Lihong Liu[1,3], Zhi Zhang[4], Jing Cheng[4,5*] & Shan Wang[1,6*]

[1] *AI Project Team, Smart Hospital Branch, Chinese Society of Medical Equipment, Beijing 100082, China*
[2] *Institute of Cloud Computing and Big Data, China Academy of Information and Communications Technology (CAICT), Beijing 100083, China*
[3] *Information Center, Peking University People's Hospital, Beijing 100044, China*
[4] *Institute of Big Data and Artificial Intelligence, National Engineering Research Center for Beijing Biochips Technology, Beijing 102206, China*
[5] *School of Biomedical Engineering, Tsinghua University, Beijing 100084, China*
[6] *Surgical Oncology Laboratory, Clinical Big Data Research Center, Peking University People's Hospital, Beijing 100044, China*
* Corresponding authors, E-mail: sjf_hz@126.com; jcheng@tsinghua.edu.cn; shanwang@pkuph.edu.cn

The rapid advancement and widespread adoption of generative artificial intelligence (AI) technologies have demonstrated significant potential across a variety of sectors. However, in the medical domain-characterized by complexity, high precision, and specialized knowledge requirements-the application of general-purpose large language models (LLMs) is often constrained by limitations in domain adaptability. While general LLMs leverage self-supervised learning based on large-scale open-domain corpora, such data sources typically lack the granularity, specificity, and semantic precision necessary for healthcare and biomedical applications. Consequently, the efficacy and reliability of these models in clinical and healthcare-related scenarios remain limited.

In contrast, vertical domain-specific large models (often referred to as vertical foundation models) offer promising solutions to overcome these challenges. These models are designed with a focus on domain specialization, incorporating expert-curated corpora, fine-grained medical ontologies, and targeted task formulations. This specialized approach enables vertical models to achieve higher accuracy, better contextual understanding, and more effective task performance in scenarios where general models fail to deliver sufficient precision.

This paper conducts a comprehensive analysis of existing health and medical corpus construction practices, both domestically and internationally, identifying critical gaps in data structure, standardization, and adaptability to AI tasks. Building upon this analysis, we propose a standardized construction framework centered on a “Disease-Scenario Association Matrix”. This framework facilitates the multidimensional mapping between disease classifications-based on standards such as ICD-10 and adjusted for domestic needs-and medical application scenarios, which are guided by authoritative references such as the “Guidelines for AI Application Scenarios in the Health Industry” issued by national health authorities. The matrix serves as a dynamic, task-driven mechanism that connects disease characteristics to specific healthcare scenarios, enabling targeted corpus development for high-priority use cases.

To ensure data utility, quality, and long-term sustainability, we further introduce a corpus standards system encompassing four critical dimensions: data acquisition, quality evaluation, annotation protocols, and privacy protection. Data collection protocols are aligned with international standards such as HL7 and FHIR to ensure structural and semantic interoperability. Quality assessment frameworks are developed based on criteria such as completeness, accuracy, timeliness, and consistency, integrating both automated and manual validation mechanisms. Annotation systems are designed with hierarchical and rule-based structures, leveraging domain-specific pre-trained models such as BioBERT and U-Net for semi-automated labeling, followed by expert review for validation. Meanwhile, privacy protection is achieved through a combination of data de-identification, encryption, federated learning, and access control strategies to ensure full compliance with data governance and ethical standards.

Finally, a dynamic feedback and quality control loop is incorporated into the corpus development process to enable continuous updates, refinement, and expansion. By integrating these mechanisms into a multi-level, task-adaptive architecture, this framework lays the methodological and theoretical foundation for constructing a high-quality, multimodal AI corpus tailored to the unique demands of the healthcare sector. This corpus will support the development and deployment of vertical domain large models, unlocking new capabilities for intelligent diagnostics, clinical decision support, and precision medicine.

generative artificial intelligence, large language model, vertical large model, corpus, multimodal corpora, health industry

doi: 10.1360/TB-2025-0185