**• 专论：医疗健康大语言模型 •**

**编者按：** 近年来人工智能经历了从判别式到生成式的重要转变。大语言模型作为生成式人工智能的代表，展现出强大的语言理解和生成能力。在医疗健康领域，大语言模型已应用于智能问诊、影像解读、智慧病案、康复管理等方面，在处理复杂医疗数据和执行多样化任务中展现出巨大潜力，有助于显著提高诊断准确率、病例处理效率，并为医学教育提供支持。然而，医疗健康大语言模型的实际部署仍面临诸多挑战，包括"幻觉"、安全性、可解释性、公平性、隐私担忧以及可追责性等，有待进一步深入研究。本期专论着眼于医疗健康大语言模型，所载论文详细阐述了医疗健康大语言模型的发展、应用、挑战和对齐要求，以及大语言模型在临床实践中的应用，为推动医疗健康大语言模型安全、有效、公平应用提供参考。

# 医疗健康大语言模型的发展、挑战与对齐*

刘 通[1,2] 黄成凤[3] 张 雯[4]

（[1] 之江实验室发展战略与合作中心 杭州 311121
[2] 浙江省哲学社会科学实验室－之江实验室智能社会治理实验室 杭州 311121
[3] 之江实验室数据枢纽与安全研究中心 杭州 311121 [4] 浙江树人学院管理学院 杭州 312028）

〔摘要〕 **目的/意义** 系统回顾和分析医疗健康大语言模型的发展、应用前沿、面临挑战以及对齐问题的研究现状，以期为推动其安全、有效和公平应用提供参考。**方法/过程** 通过文献分析阐述大语言模型在医疗健康领域的发展和应用情况，分析其面临的挑战，从技术层面和社会层面提出应对策略。**结果/结论** 医疗健康大语言模型应该尤其注意模型的准确性、安全性、可解释性、公平性、隐私性和可追责性，以完成模型对齐，满足伦理要求，提升患者福祉。

〔关键词〕 医疗健康大语言模型；人工智能监管；大模型对齐

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2024.09.001

**Development, Challenges and Alignments of Large Language Models in Healthcare**

*LIU Tong*[1,2], *HUANG Chengfeng*[3], *ZHANG Wen*[4]

*[1]Development Strategy and Cooperation Center, Zhejiang Lab, Hangzhou 311121, China; [2]Zhejiang Laboratory of Philosophy and Social Sciences －Laboratory of Intelligent Society and Governance, Zhejiang Lab, Hangzhou 311121, China; [3]Research Center for Data Hub and Security, Zhejiang Lab, Hangzhou 311121, China; [4]School of Management, Zhejiang Shuren University, Hangzhou 312028, China*

〔Abstract〕 **Purpose/Significance** The paper systematically reviews and analyzes the development, cutting－edge applications, challenges and alignment issues of large language models (LLMs) in healthcare, so as to provide references for promoting the safe, effective

〔修回日期〕 2024－09－05

〔作者简介〕 刘通，博士，博士后，发表论文6篇；通信作者：黄成凤。

〔基金项目〕 浙江省哲学社会科学规划课题（项目编号：124NDQN203YBM）；广东省深圳市科创委基础研究项目（项目编号：RKX20220705152810024）。

and equitable application. **Method/Process** Through literature analysis, the development and application of LLMs in healthcare are expounded, the challenges faced are analyzed, and countermeasures are proposed from the technical and social aspects. **Result/Conclusion** LLMs in healthcare should prioritize truthfulness, safety, interpretability, fairness, privacy and accountability to realize model alignment, meet ethical requirements, and improve patients' welfare.

〔**Keywords**〕 large language models (LLMs) in healthcare; artificial intelligence (AI) regulation; LLMs alignment

## 1 引言

近年来人工智能（artificial intelligence，AI）经历了从判别式到生成式的重要转变。判别式 AI 主要通过学习输入与输出之间的关系，执行识别、分类、分析和预测任务。生成式 AI 能够创造新内容，如文本、图像、代码等，通过学习数据的联合概率分布来模拟人类的创造性过程[1]。这种转变极大地拓展了 AI 的应用范围和创新潜力。在此背景下，大语言模型（large language models，LLMs）作为生成式 AI 的代表，展现出强大的语言理解和生成能力。这些模型基于 Transformer 架构，通过基于海量文本数据的自监督学习，习得复杂语言模式[2]。LLMs 能够执行多种自然语言处理任务，逐渐被应用于医疗、教育、科学研究等领域。然而，LLMs 的核心功能是预测给定序列的下一个最可能出现的词语，即使出现了智能“涌现”，也无法改变 LLMs 与人类学习模式不同这一根本事实，这种工作机制限制了模型对真实世界的理解。

在医疗健康领域，国外出现了 Med－PaLM2、BioMedLM 等医疗健康大语言模型（以下简称医疗健康大模型），其中 Med－PaLM2 在美国医师执照考试问题集 MedQA 上的表现相比之前的 AI 模型有巨大提升[3－4]。国内也出现了华佗 GPT、医联 MedGPT 等中文医疗健康大模型。医疗健康大模型利用语言、视觉乃至多模态融合的医学数据进行“预训练－微调”，在不同医疗任务中表现出良好的智能化水平和泛化能力，为用户提供高效、准确、个性化的服务。例如，医疗健康大模型使用规划、推理能力，利用患者的话语和病历数据提取有用信息，提出专业的诊断建议[5]。然而，医疗健康大模型仍存在亟待解决的问题，如模型产生的“幻觉”、缺乏可解释性、数据隐私和安全问题，以及确保模型输出的公平性和去偏见等。这些挑战不仅涉及技术层面，还涉及伦理、法律和社会接受度等方面，需要跨学科合作来解决。使 AI 的发展符合人类的意图、目标、偏好和道德准则，就是 AI 的对齐问题。医疗是高风险领域，上述挑战直接关系到患者生命安全和社会安全治理。因此，确保医疗健康大模型与人类价值观、医疗伦理的对齐是重要任务。

本文旨在系统性地回顾和分析医疗健康大模型的发展、应用前沿、面临的挑战以及对齐问题的研究现状，以期为未来研究提供有价值的参考，推动医疗健康大模型安全、有效和公平的应用。

## 2 大语言模型在医疗健康领域的发展

医疗健康大模型是基于 Transformer 架构或者混合了 Transformer 架构的深度学习模型，即通用 LLMs 的领域增强或者调优后的版本[6]。医疗健康大模型的改进、扩展和能力提升主要由 3 个要素驱动：医疗健康大模型技术发展、医疗健康大模型落地应用和政策法律层面的监管和治理。

### 2.1 技术发展方面

医疗健康大模型主要在数据架构和 AI 代理两方面推进。对于数据架构，医疗健康大模型通过在大量医学数据集上进行预训练和调优[7]，或者连接外部知识库[8－9]，提高基础模型在医疗文本和多模态任务中的准确性[10－11]。MedPaLM 等模型还通过采取更先进的训练架构，进一步提升其在医疗文本处理中的性能和可靠性[4]，增强模型设计和运行过

程的透明性，这在医疗领域尤为重要[11]。对于 AI 代理，自动或协同工作的医疗 AI 代理在诊断和决策支持、药物发现、医疗管理等领域具有重要作用。在诊疗之外，AI 代理尤其能够在模拟医疗的社会影响方面发挥作用，对公共卫生和政策制定具有重要意义。例如，美国斯坦福大学开发的 AI 代理模拟社区中的群体行为[12]；清华大学构建了更全面的模拟医院环境 Agent Hospital[13]。

### 2.2 应用落地方面

医疗健康大模型已应用于智能问诊、影像解读、智慧病案、康复管理等领域。从分诊前、诊中和诊后 3 个阶段分析医疗健康大模型的落地应用情况。诊前阶段，医疗健康大模型提供咨询和导诊服务，通过非医学术语和口语化的表达展开与患者的多轮问答，提供就诊建议。例如，ChatDoctor[14]、DoctorGLM[15] 能够向患者提供诊前帮助，与医院数据库对接后，模型自动提取信息生成患者病历，提高诊断效率。诊中阶段，医疗健康大模型能提升临床决策支持系统的智能化水平，帮助患者有序就诊，改善就诊环境。例如，通义医疗行业大模型、PULSE 中文医疗健康大模型能够解读多来源、多模态的医疗数据，优化就诊流程并提高运营效率。NYUTron 能够进一步预测 30 天再入院率、住院死亡率、住院时间等指标[16]。此外，医疗健康大模型能够帮助医生决策，提高诊断效率。例如 SAM－Med2D [17]、PANDA[18] 能够辅助医生进行图像分析判断。诊后阶段，医疗健康大模型通过统一管理医疗数据，帮助患者随访、医保支付、排队领药、获取报告和健康管理等。例如，天津市基层数字健共体与睿医 AI 平台结合，在高龄人群慢病诊治中，通过分析历史数据生成个性化方案，提升就医效率。

### 2.3 政策监管方面

随着 LLMs 被整合到各个领域的众多人工智能系统中，全球范围内都强调必须加强监管，同时明确支持可信赖的创新，但暂未出现专门监管医疗人工智能的政策。欧盟于 2024 年出台了世界上首部全面的《人工智能法案》，对不同等级的 AI 系统分类管理。与欧盟采取的综合性监管制度不同，美国在联邦层面还未专门推出一部人工智能综合监管法律，更多呈现出“一事一议”的特点。2023 年我国发布了《生成式人工智能服务管理暂行办法》，是我国现阶段促进生成式人工智能健康发展和规范应用的主要法规。

## 3 医疗健康大语言模型面临的挑战

医疗健康大模型具有巨大潜力，但仍然面临挑战。例如 LLMs 的“幻觉”始终存在，即产生不正确或误导性的信息[19]，对患者信心和信任造成很大干扰。对此，国外 AI 研究机构提出不同治理方向，例如美国 Meta 公司提出 5 条标准：隐私（privacy）、公平（fairness）、稳健（robustness）、透明（transparency）和可追责（accountability）；美国微软公司提出公平（fairness）、安全可靠（reliability & safety）、隐私安全（privacy & security）、包容性（inclusiveness）、透明性（transparency）和可追责（accountability）；美国谷歌公司则关注公平（fairness）、隐私（privacy）和安全（safety）等。各机构应对挑战的角度和方式也不相同，例如，美国 OpenAI 公司使用红队网络和人工反馈进行对齐，美国 Anthropic 公司则倾向于通过建立 AI 原则指导训练和使用。

本文将医疗健康大模型所面临的挑战分为技术型挑战和社会型挑战。技术型挑战来源于 LLMs 结构本身，包含 3 个具体问题：“幻觉”或准确性问题、安全性问题、可解释性问题。社会型挑战是社会形态和社会问题在 AI 层面的反映，也包含 3 个具体问题：公平问题、隐私问题和可追责性问题。与之对应，本文提出应对上述挑战的对齐方法同样分为技术和社会两个层面，见图 1。

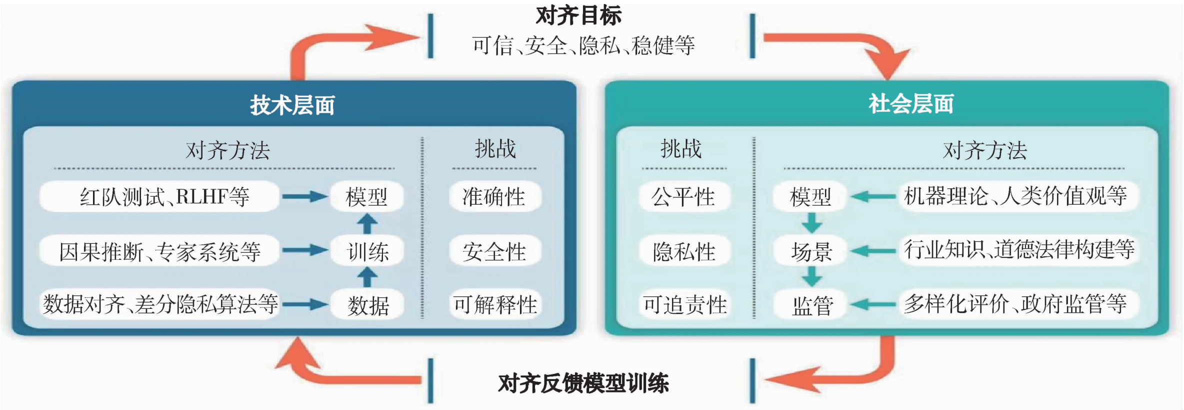


图1 医疗健康大模型对齐框架

## 3.1 准确性

准确性或信息准确性（truthfulness）直接影响LLMs应用[20]。其代表问题是“幻觉”，即生成的输出包含不准确或非事实信息。“幻觉”分为内在“幻觉”和外在“幻觉”，内在“幻觉”指LLMs产生了逻辑上与事实信息相矛盾的输出，如数学公式错误计算；外在“幻觉”指LLMs的输出无法被证实或者证伪，如LLMs伪造论文引用的参考文献。在医疗场景中，医生诊疗建议具有相当程度的信任属性，即患者缺乏足够医疗知识，只有相信医生才会继续采取诊疗措施[21]。当LLMs错误地使用了知识或者使用了错误的知识，却顺畅地输出语言内容，在缺乏信息检验和修正的情况下，患者可能会难以辨别信息的准确性或者危险性，导致误诊、不适当治疗和有害治疗的结果。

## 3.2 安全性

安全性（safety）是LLMs在广泛应用时所面临的问题，虽然LLMs应该被设计成安全和无害的，但其安全方面具有脆弱性，如存在“越狱”风险[22-23]，即在人为操纵下，LLMs的输出可以突破限制。LLMs的安全性与稳健性（robustnesss）息息相关，即LLMs的输出结果在各种条件下保持稳定[24-25]。LLMs要确保能应对意外中断和对抗攻击，安全性方面的失败有可能显著影响用户交互内容和下游行为[26]，产生有毒内容，并进一步用于虚假宣传、诱骗欺诈、舆论操纵、仇恨引导等[27]。在医疗场景中，若医疗健康大模型的安全性不足，患者输入指令的轻微变化都有使输出结果的巨大改变，使患者丧失信心，甚至产生健康威胁。医生同样对医疗系统的输出有严格要求，模型安全性不足将降低医生的接受度。

## 3.3 可解释性

LLMs等深度学习模型在可解释性（interpretability）或者透明性（transparency）方面尚存在问题。相比于经典统计模型，深度学习框架下的模型决策是典型的“黑箱”过程[28]。LLMs拥有巨量参数，其输出流程更加复杂，输出结果不具有解释性，导致在医疗领域的应用推广受阻[29-30]。在医疗场景中，一方面，医疗健康大模型的训练数据中混杂着医疗数据和公开数据，模型使用者并不清楚是哪部分数据和参数作出了决策，继而担忧决策的可靠性；另一方面，LLMs的训练往往非常困难，闭源模型不会公布代码，训练过程难以复现，可能导致用户对模型透明性产生疑虑。医疗领域对于AI输出结果的因果性有严格要求，一旦模型无法合理解释输出，医疗健康大模型就无法被大规模应用。

## 3.4 公平性

公平性（fairness）或者其反面的偏差（bias）是重要原则，可确保LLMs的设计、训练和部署不会导致有偏见或歧视性的结果，并公平地对待所有用户和群体[31]。越来越多国家在实施法律框架中要求人工智能应用应当遵守公平原则[32]。LLMs训练

数据集中往往存在大量歧视和偏见，主要针对社会边缘人群[33]，导致LLMs难以保证输出的公平性。应该尽可能降低LLMs的偏见，以防进一步被再生成和传播为更加有毒的内容，例如仇恨言论和人身攻击等。医疗场景中，医疗保健体系存在的不平等的系统性差异和偏见可能进一步被LLMs的不公正输出和衡量方式放大。此外，患者对医疗健康大模型建立信任后，一旦遇到有毒性和偏差的内容，可能会产生更大的风险和伤害。

### 3.5 隐私性

隐私问题（privacy）在LLMs发展过程中受到巨大关注。因为LLMs具有记忆能力，可能通过提示词攻击、梯度重建攻击等方式泄漏个人隐私信息[34-35]。LLMs的隐私保护应该帮助维护人类的自主性、身份和尊严[36]。LLMs的训练数据集通常是通过爬取得到的海量互联网数据，包括通话记录、地址等个人信息，这会增加用户对隐私问题的关切程度[37]。一方面，在医疗场景中，用于训练医疗健康大模型的数据中可能包含健康状况、个人病史、临床信息等涉及患者隐私的数据，一旦泄漏会产生不可挽回的损失。另一方面，医疗健康大模型的训练离不开医学相关数据，不经过监管的训练过程可能会导致模型更容易被攻破，进而导致隐私信息外漏。

### 3.6 可追责

可追责（accountability）一直是AI领域的热门话题，LLMs的发展增加了对输出结果的问责难度。有学者[38]认为可追责与可解释性在AI领域是同义词，也有学者[39-40]认为AI的可追责性应该被视为一种供应商承担程序输出结果的措施。AI模型的可追责能力应该被视为让设计者、拥有者或AI算法使用者承担责任的一系列预防或减轻风险的策略[41]。在医疗场景中，目前已有研究[42-43]关注LLMs可追责性，虽然医疗健康大模型能够提供医疗决策支持，但也可能造成限制个人选择权等新问题和风险。在医学临床和科研任务中，大模型的参与导致权属和责任不明，因为其开发涉及链条较长，包括模型开发者、数据提供者、模型使用者等，如果医疗健康大模型输出内容出现问题，则难以确定责任归属。

## 4 医疗领域大模型的对齐要求

### 4.1 保证模型输出信息的准确和安全

医疗健康大模型首先应该注重模型输出信息的准确和安全，即模型预测结果与真实结果的一致性。在医疗场景中，可反映为诊断的准确率、治疗建议的适当性等。安全性关注医疗健康大模型是否符合人类的伦理标准、是否能避免对人类的伤害等[44-45]。医疗健康大模型的安全性对齐不仅要求较高准确性，还要求能够对模型可能产生的有害行为进行预防，例如避免给出自我伤害的建议、色情或暴力内容、攻击或暴力行为建议、寻求非法商品或服务的指引。医疗健康大模型在安全性和准确性方面的对齐方式有两个层面。一是技术层面，医疗健康大模型可以使用训练时校正[46]，添加推理过程，使用外部资源进行检索增强以缓解“幻觉”[47-48]，组织防御能力的红队测试[49]、人类反馈强化学习[50-51]等。二是社会层面，医疗健康大模型需要专门的、符合价值观的数据集进行训练[52]，将适当的道德价值观输入其中，形成机器伦理[53-54]。由于道德标准是随时间和空间不断变化的[55]，研究人员应该开发特定的用户界面和交互模式，使人类能够在必要时进行干预和反馈[56-57]。

### 4.2 提高模型的可解释性或者透明性

医疗健康大模型应该对可解释性或者透明性进行对齐，确保医疗上下游使用人员都能够理解医疗健康大模型的能力边界，以获取信任。深度学习具有“黑箱”属性，研究人员通过信息和社会维度特征对可解释性进行研究[58]。医疗健康大模型可解释性对齐的方式分为两个层面。一是技术层面，可解释性的对齐可以分为事后可解释性和内在可解释性[44]，事后可解释性技术目标是理解神经网络结构和单元及其与输出结果的因果关系，例如提取知识在模型内部的表示方式，以编辑神经网络的权

重[59]；内在可解释性则通过替换神经网络模块，帮助改进模型的性能和可解释性，如测试新的模型架构[44,60]。二是社会层面，模型的透明度涉及模型细节披露，应该与相关行业要求相一致，反映行业上下游各参与者之间动态利益关系，应当建立专门标准进行规范和评估。

## 4.3 促进社会公平和消除偏见

保证不出现偏见和歧视是重要的 AI 伦理，医疗健康大模型需要公平对待所有用户和群体[31]。中国、美国和欧盟都将公平、公正、非歧视等原则加入人工智能伦理规范或者指导意见中。然而，LLMs 训练数据中累积的人类偏见和歧视不可避免地影响输出内容[61-62]。在医疗领域，不公正可能导致个人健康受损和医疗资源分配不合理问题，因此需要格外重视。医疗健康大模型公正性的对齐方式可以分为两个层面。一是技术层面，数据集方法是最基本和直接的方法[63]。模型训练一方面应当对数据集进行挑选，确保数据覆盖多样的人群、文化和观点，减少偏见和歧视，另一方面需要识别和修正数据集中的偏见，对显性和隐性偏见进行人工标注[64-65]，或者使用专家系统类型的算法 Hatecheck[66]、BBQ（bias benchmark for QA）[67]等进行识别。二是社会层面，公平性指标应该被准确定义并细化，避免“刻板印象”和“选择偏好”等名词在具体事务中的重叠和干扰[68]。在人工智能背景下，人类感知公平变得更为复杂[69-70]，医疗健康大模型应该进行定期的公平性评估，使用相关量化指标确保模型在不同人群、场景中的表现保持公平。

## 4.4 增强隐私保护与数据安全

隐私保护与数据安全是医疗健康大模型对齐的另一个要求，其在数据收集和使用阶段面临着不同威胁。在数据收集阶段，隐私问题主要存在于收集过程及数据存储，包括投毒攻击和对抗攻击等；在模型使用阶段，反复查询、反转攻击和提示词攻击等诱导大模型吐露信息的攻击手法是引发隐私担忧的主要来源[71-72]。通过医疗健康大模型的隐私对齐确保模型的设计、训练和部署过程中充分保护个人隐私，防止数据泄漏或被滥用，这对于赢得用户信任、遵守法律法规、确保医疗服务的安全和效率至关重要。对齐也分为两个层面。一是技术层面，在模型训练时，可以使用差分隐私[73]、数据合成[74]、同态加密[75]等方法对数据集进行隐私化处理；在落地阶段，可以通过本地部署、设置访问权限等方法保护隐私。二是社会层面，医疗健康大模型需要遵守隐私法律和规定，确保所有操作，包括数据收集、处理、存储和分析活动，均符合相关国家和地区的隐私法律和政策，如《中华人民共和国个人信息保护法》和欧盟《通用数据保护条例》等。此外，医疗行业应该规范 AI 模型训练和使用，建立数据治理框架，包括数据访问控制、数据使用审计和伦理审查机制，确保数据处理活动透明且可追溯。

## 4.5 建立责任与监管框架

医疗健康大模型发展直接关系人类健康，其系统误差可能导致严重医疗事故，因此合理的责任框架与监管政策具有重大意义。只有建立完善的责任监督机制，医疗健康大模型才能在安全范围内发展推广。医学大模型引入责任和监督的对齐，意味着必须确保技术进步与伦理标准、法律法规以及患者利益一致，继而推动医疗领域的可持续发展和技术创新。医疗健康大模型可以从以下 4 个层面进行对齐。一是构建符合行业标准的法律框架，建立和遵循一套明确的行业标准、应用指南和法律规范，以确保 AI 系统开发和应用符合社会伦理标准和法律要求。二是结合基准测试和现实场景构建合理的技术验证标准，当前技术人员通过基准测试集，如 MedQA、PubMedQA 和 MultiMedQA 等衡量模型能力，但是医疗健康大模型更应该在现实医疗场景中进行测试和验证，将其中涉及的技术与实际医疗状况进行交叉比对，提升模型的有效性和实用性。三是医疗健康大模型监管应该是防风险和促发展并重，一方面考虑到医疗行业的特殊性更加谨慎地开展 AI 应用，另一方面 AI 与医疗行业的结合是大势所趋[3,18]，注重风险和发展的平衡是出台相关政策

的重点。四是医疗健康大模型的对齐应促进公众主动参与、提高 AI 素养和建立社会共识，以确保技术发展与社会价值观、文化背景保持一致。这些措施体现了政府监管在 AI 治理中实现全面对齐的重要性，也体现了技术创新同时促进社会福祉和保护个人权利的最终目标。

## 5　结语

AI 的快速进步，尤其是 LLMs 的发展，为医疗健康领域带来了巨大的变革潜力。LLMs 基于大量数据进行训练得到庞大的参数规模，在多种任务中展现出卓越性能，并在医疗领域得到快速推广。LLMs 在医疗健康领域的发展正逐渐从理论研究转向实际应用，包括但不限于诊断支持、临床笔记生成、医学教育以及疾病预测等方面。医疗健康大模型在处理复杂医疗数据和执行多样化任务方面展现出巨大潜力，有望显著提高诊断准确率、病例处理效率，并为医学教育提供支持。然而，尽管医疗健康大模型应用前景广阔，但其实际部署仍面临诸多挑战，包括“幻觉”、安全性、可解释性、公平性、隐私担忧以及可追责性。为了应对这些挑战，应确保医疗健康大模型在不同维度进行对齐，使模型与医疗伦理、患者福祉、专业医疗标准保持一致。模型对齐不仅涉及技术层面的改进，如通过人工反馈强化学习、隐私计算等技术方法，也需要社会层面的努力，如明确责任框架和监管政策、提升公众 AI 素养等。

未来医疗健康大模型发展应更加注重解决现实世界的临床问题，提供更加精准、个性化的医疗服务。医疗健康大模型不仅要在技术上进行创新，如通过融合更多类型的医疗数据（如时间序列、视觉和音频数据）构建多模态大模型，还需要在伦理、法律和社会接受度等方面进行深入探讨和有效管理。跨学科合作将成为推动医疗健康大模型发展的重要驱动力，医学专家、数据科学家、伦理学家和法律专家等的紧密合作将确保医疗健康大模型的发展既能满足临床需求，又能符合伦理标准和社会期待。通过技术创新与社会伦理的紧密结合，医疗健康大模型在提高医疗服务质量、促进医学研究和保障患者福祉方面将发挥更大作用。

**作者贡献：**刘通负责文献收集与分析、论文撰写；黄成凤负责研究设计、提供指导；张雯负责提出研究选题、提供材料支持、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　WU J, GAN W, CHEN Z, et al. AI－generated content (AIGC): a survey [EB/OL]. [2023－03－26]. https://arxiv.org/pdf/2304.06632.

2　SHANAHAN M. Talking about large language models [J]. Communications of the ACM, 2024, 67 (2): 68－79.

3　SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge [J]. Nature, 2023, 620 (7972): 172－180.

4　SINGHAL K, TU T, GOTTWEIS J, et al. Towards expert－level medical question answering with large language models [EB/OL]. [2023－05－16]. https://arxiv.org/pdf/2305.09617.

5　TU T, PALEPU A, SCHAEKERMANN M, et al. Towards conversational diagnostic AI [EB/OL]. [2024－01－11]. https://arxiv.org/pdf/2401.05654.

6　OMIYE J A, GUI H, REZAEI S J, et al. Large language models in medicine: the potentials and pitfalls [EB/OL]. [2023－08－31]. https://arxiv.org/pdf/2309.00087.

7　HE K, MAO R, LIN Q, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics [EB/OL]. [2024－01－11]. https://arxiv.org/pdf/2310.05694.

8　PAIS C, LIU J, VOIGT R, et al. Large language models for preventing medication direction errors in online pharmacies [J]. Nature medicine, 2024, 30 (6): 1574－1582.

9　GAO Y, LI R, CROXFORD E, et al. Large language models and medical knowledge grounding for diagnosis prediction [EB/OL]. [2023－11－27]. https://www.medrxiv.org/content/10.1101/2023.11.24.23298641v1.full.pdf.

10　NAZI Z A, PENG W. Large language models in healthcare and medical domain: a review [EB/OL]. [2023－12－11]. https://arxiv.org/pdf/2401.06775.

11　KIM C, GADGIL S U, DEGRAVE A J, et al. Transparent medical image AI via an image－text foundation model grounded in medical literature [J]. Nature medicine, 2024, 30 (4): 1154－1165.

12 PARK J S, O'BRIEN J, CAI C J, et al. Generative agents: interactive simulacra of human behavior [EB/OL]. [2023-04-07]. https://arxiv.org/pdf/2304.03442.

13 LI J, WANG S, ZHANG M, et al. Agent hospital: a simulacrum of hospital with evolvable medical agents [EB/OL]. [2024-05-05]. https://arxiv.org/pdf/2405.02957.

14 LI Y, LI Z, ZHANG K, et al. Chatdoctor: a medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge [EB/OL]. [2023-03-24]. https://arxiv.org/pdf/2303.14070.

15 XIONG H, WANG S, ZHU Y, et al. Doctorglm: fine-tuning your Chinese doctor is not a herculean task [EB/OL]. [2023-04-03]. https://arxiv.org/pdf/2304.01097.

16 JIANG L Y, LIU X C, NEJATIAN N P, et al. Health system-scale language models are all-purpose prediction engines [J]. Nature, 2023, 619 (7969): 357-362.

17 LIU M, DING J, XU J, et al. MedBench: a comprehensive, standardized, and reliable benchmarking system for evaluating Chinese medical large language models [EB/OL]. [2024-01-24]. https://www.arxiv.org/pdf/2407.10990.

18 CAO K, XIA Y, YAO J, et al. Large-scale pancreatic cancer detection via non-contrast CT and deep learning [J]. Nature medicine, 2023, 29 (12): 3033-3043.

19 JI Z, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation [J]. ACM computing surveys, 2023, 55 (12): 1-38.

20 BAI Y, JONES A, NDOUSSE K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback [EB/OL]. [2024-04-12]. https://arxiv.org/pdf/2204.05862.

21 CHEN Y, LEE S. User-generated physician ratings and their effects on patients' physician choices: evidence from Yelp [J]. Journal of marketing, 2024, 88 (1): 77-96.

22 BHARDWAJ R, PORIA S. Red-teaming large language models using chain of utterances for safety-alignment [EB/OL]. [2023-08-18]. https://arxiv.org/pdf/2308.09662.

23 WANG W, TU Z, CHEN C, et al. All languages matter: on the multilingual safety of large language models [EB/OL]. [2023-10-02]. https://arxiv.org/pdf/2310.00905.

24 CHEN P Y, LIU S. Holistic adversarial robustness of deep learning models [J]. Proceedings of the AAAI conference on artificial intelligence, 2023, 37 (13): 15411-15420.

25 WANG B, XU C, WANG S, et al. Adversarial glue: a multi-task benchmark for robustness evaluation of language models [EB/OL]. [2024-01-10]. https://arxiv.org/pdf/2111.02840.

26 WELBL J, GLAESE A, UESATO J, et al. Challenges in detoxifying language models [EB/OL]. [2023-09-15]. https://arxiv.org/pdf/2109.07445.

27 WEIDINGER L, MELLOR J, RAUH M, et al. Ethical and social risks of harm from language models [EB/OL]. [2024-02-22]. https://arxiv.org/pdf/2112.04359.

28 BAUER K, VON ZAHN M, HINZ O. Expl (AI) ned: the impact of explainable artificial intelligence on users' information processing [J]. Information systems research, 2023, 34 (4): 1582-1602.

29 LIAO Q V, VAUGHAN J W. AI transparency in the age of LLMs: a human-centered research roadmap [EB/OL]. [2023-08-08]. https://arxiv.org/pdf/2306.01941.

30 NANDA N, CHAN L, LIEBERUM T, et al. Progress measures for grokking via mechanistic interpretability [EB/OL]. [2023-10-19]. https://arxiv.org/pdf/2301.05217.

31 WANG L, SONG M, REZAPOUR R, et al. People's perceptions toward bias and related concepts in large language models: a systematic review [EB/OL]. [2024-03-03]. https://arxiv.org/pdf/2309.14504.

32 FJELD J, ACHTEN N, HILLIGOSS H, et al. Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI [J]. Berkman Klein center research publication, 2020 (1): 3518482.

33 YANG Z, YI X, LI P, et al. Unified detoxifying and debiasing in language generation via inference-time adaptive optimization [EB/OL]. [2023-01-02]. https://arxiv.org/pdf/2210.04492.

34 BROWN H, LEE K, MIRESHGHALLAH F, et al. What does it mean for a language model to preserve privacy [EB/OL]. [2023-02-14]. https://arxiv.org/pdf/2202.05520.

35 GUO S, XIE C, LI J, et al. Threats to pre-trained language models: survey and taxonomy [EB/OL]. [2023-02-14]. https://arxiv.org/pdf/2202.06862.

36 SUN L, HUANG Y, WANG H, et al. Trustllm: trustworthiness in large language models [EB/OL]. [2024-08-26]. https://arxiv.org/pdf/2401.05561.

37 KHOWAJA S A, KHUWAJA P, DEV K, et al. ChatGPT needs spade (sustainability, privacy, digital divide, and ethics) evaluation: a review [J]. Cognitive computation, 2024, 16 (5): 2528-2550.

38 BEZUIDENHOUT L, RATTI E. What does it mean to embed ethics in data science? An integrative approach based on microethics and virtues [J]. AI & society, 2021, 36 (3): 939-953.

39 BEERKENS M. An evolution of performance data in higher education governance: a path towards a 'big data' era

[J]. Quality in higher education, 2022, 28 (1): 29 – 49.
40 SHIN D, RASUL A, FOTIADIS A. Why am I seeing this? Deconstructing algorithm literacy through the lens of users [J]. Internet research, 2021, 32 (4): 1214 – 1234.
41 MEMARIAN B, DOLECK T. Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: a systematic review [EB/OL]. [2024 – 05 – 01]. https://www.sciencedirect.com/science/article/pii/S2666920X23000310?via%3Dihub.
42 GUO E, GUPTA M, SINHA S, et al. NeuroGPT – X: towards an accountable expert opinion tool for vestibular schwannoma [EB/OL]. [2023 – 02 – 26]. https://www.medrxiv.org/content/10.1101/2023.02.25.23286117 v1.full.pdf.
43 KIM J K, CHUA M, RICKARD M, et al. ChatGPT and large language model (LLM) chatbots: the current state of acceptability and a proposal for guidelines on utilization in academic medicine [J]. Journal of pediatric urology, 2023, 19 (5): 598 – 604.
44 JI J, QIU T, CHEN B, et al. AI alignment: a comprehensive survey [EB/OL]. [2024 – 05 – 01]. https://arxiv.org/pdf/2310.19852.
45 SHEN T, JIN R, HUANG Y, et al. Large language model alignment: a survey [EB/OL]. [2023 – 09 – 26]. https://arxiv.org/pdf/2309.15025.
46 ROIT P, FERRET J, SHANI L, et al. Factually consistent summarization via reinforcement learning with textual entailment feedback [EB/OL]. [2023 – 05 – 31]. https://arxiv.org/pdf/2306.00186.
47 ASAI A, WU Z, WANG Y, et al. Self – rag: learning to retrieve, generate, and critique through self – reflection [EB/OL]. [2023 – 10 – 17]. https://arxiv.org/pdf/2310.11511.
48 SHUSTER K, POFF S, CHEN M, et al. Retrieval augmentation reduces hallucination in conversation [EB/OL]. [2024 – 04 – 15]. https://arxiv.org/pdf/2104.07567.
49 PEREZ E, HUANG S, SONG F, et al. Red teaming language models with language models [EB/OL]. [2024 – 02 – 07]. https://arxiv.org/pdf/2202.03286.
50 ACHIAM J, ADLER S, AGARWAL S, et al. GPT – 4 technical report [EB/OL]. [2024 – 03 – 04]. https://arxiv.org/pdf/2303.08774.
51 OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback [EB/OL]. [2024 – 03 – 04]. https://arxiv.org/pdf/2203.02155.
52 AWAD E, DSOUZA S, KIM R, et al. The moral machine experiment [J]. Nature, 2018, 563 (7729): 59 – 64.
53 TOLMEIJER S, KNEER M, SARASUA C, et al. Implementations in machine ethics: a survey [J]. ACM computing surveys (CSUR), 2020, 53 (6): 1 – 38.
54 WINFIELD A F, MICHAEL K, PITT J, et al. Machine ethics: the design and governance of ethical AI and autonomous systems [J]. Proceedings of the IEEE, 2019, 107 (3): 509 – 517.
55 GRAHAM J, MEINDL P, BEALL E, et al. Cultural differences in moral judgment and behavior, across and within societies [J]. Current opinion in psychology, 2016, 8 (4): 125 – 130.
56 CADARIO R, LONGONI C, MOREWEDGE C K. Understanding, explaining, and utilizing medical artificial intelligence [J]. Nature human behaviour, 2021, 5 (12): 1636 – 1642.
57 MELZNER J, BONEZZI A, MEYVIS T. Information disclosure in the era of voice technology [J]. Journal of marketing, 2023, 87 (4): 491 – 509.
58 FELZMANN H, FOSCH – VILLARONGA E, LUTZ C, et al. Towards transparency by design for artificial intelligence [J]. Science and engineering ethics, 2020, 26 (6): 3333 – 3361.
59 VILONE G, LONGO L. Explainable artificial intelligence: a systematic review [EB/OL]. [2023 – 10 – 12]. https://arxiv.org/pdf/2006.00093.
60 KLAMBAUER G, UNTERTHINER T, MAYR A, et al. Self – normalizing neural networks [EB/OL]. [2023 – 09 – 07]. https://arxiv.org/pdf/1706.02515.
61 LIU Y, GAUTAM S, MA J, et al. Investigating the fairness of large language models for predictions on tabular data [EB/OL]. [2024 – 04 – 02]. https://arxiv.org/pdf/2310.14607.
62 ZHENG C, ZHOU H, MENG F, et al. Large language models are not robust multiple choice selectors [EB/OL]. [2024 – 02 – 22]. https://arxiv.org/pdf/2309.03882.
63 CELIKYILMAZ A, CLARK E, GAO J. Evaluation of text generation: a survey [EB/OL]. [2023 – 05 – 18]. https://arxiv.org/pdf/2006.14799.
64 ROSENTHAL S, ATANASOVA P, KARADZHOV G, et al. SOLID: a large – scale semi – supervised dataset for offensive language identification [EB/OL]. [2023 – 09 – 24]. https://arxiv.org/pdf/2004.14454.
65 ZAMPIERI M, MALMASI S, NAKOV P, et al. Predicting the type and target of offensive posts in social media [EB/OL]. [2023 – 04 – 16]. https://arxiv.org/pdf/1902.09666.
66 RÖTTGER P, VIDGEN B, NGUYEN D, et al. HateCheck: functional tests for hate speech detection models [EB/OL]. [2023 – 05 – 27]. https://arxiv.org/pdf/2012.15606.
67 PARRISH A, CHEN A, NANGIA N, et al. BBQ: a hand – built bias benchmark for question answering [EB/OL]. [2023 – 04 – 16]. https://arxiv.org/pdf/2110.08193.

（下转第18页）

ment feedback [EB/OL]. [2024-06-05]. https://arxiv.org/abs/2306.00186.

59 CHERN I C, WANG Z, DAS S, et al. Improving factuality of abstractive summarization via contrastive reward learning [EB/OL]. [2024-06-07]. https://arxiv.org/abs/2307.04507.

60 DHULIAWALA S, KOMEILI M, XU J, et al. Chain-of-verification reduces hallucination in large language models [EB/OL]. [2024-06-10]. https://arxiv.org/abs/2309.11495.

61 XIE Q, SCHENCK E J, YANG H S, et al. Faithful AI in medicine: a systematic review with large language models and beyond [EB/OL]. [2024-06-10]. https://doi.org/10.1101/2023.04.18.23288752.

62 JIN D, PAN E, OUFATTOLE N, et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams [J]. Applied sciences, 2021, 11 (14): 6421.

63 PAL A, UMAPATHI L K, SANKARASUBBU M. Medmcqa: a large-scale multi-subject multi-choice dataset for medical domain question answering [C]. online: Conference on Health, Inference, and Learning, PMLR, 2022.

64 LIN S, HILTON J, EVANS O. Truthfulqa: measuring how models mimic human falsehoods [EB/OL]. [2024-06-10]. https://arxiv.org/abs/2109.07958.

65 LI J, CHENG X, ZHAO W X, et al. Halueval: a large-scale hallucination evaluation benchmark for large language models [C]. Singapore: The 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

66 LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks [C]. online: Advances in Neural Information Processing Systems 33, 2020.

67 VALIÑA L G, MASTROLEO I. The ethical and scientific challenges of ChatGPT in health: utopianism, technophobia and pragmatism [EB/OL]. [2024-06-11]. https://doi.org/10.31219/osf.io/phtwz.

68 HE K, MAO R, LIN Q, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics [EB/OL]. [2024-06-11]. https://arxiv.org/abs/2310.05694.

69 TIAN S, JIN Q, YEGANOVA L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health [J]. Briefings in bioinformatics, 2024, 25 (1): bbad493.

70 STOKEL-WALKER C. ChatGPT listed as author on research papers: many scientists disapprove [J]. Nature, 2023, 613 (7945): 620-621.

（上接第 9 页）

68 KIVIKANGAS J M, FERNÁNDEZ-CASTILLA B, JÄRVELÄ S, et al. Moral foundations and political orientation: systematic review and meta-analysis [J]. Psychological bulletin, 2021, 147 (1): 55-94.

69 KORDZADEH N, GHASEMAGHAEI M. Algorithmic bias: review, synthesis, and future research directions [J]. European journal of information systems, 2022, 31 (3): 388-409.

70 RHUE L. The anchoring effect, algorithmic fairness, and the limits of information transparency for emotion artificial intelligence [J]. Information systems research, 2024 (1): 28.

71 LIU Y, DENG G, XU Z, et al. Jailbreaking ChatGPT via prompt engineering: an empirical study [EB/OL]. [2024-02-28]. https://arxiv.org/pdf/2305.13860.

72 WEI A, HAGHTALAB N, STEINHARDT J. Jailbroken: how does LLM safety training fail [EB/OL]. [2023-07-05]. https://arxiv.org/pdf/2307.02483.

73 DWORK C, KOHLI N, MULLIGAN D. Differential privacy in practice: expose your epsilons [J]. Journal of privacy and confidentiality, 2019, 9 (2): 1-22.

74 PEREZ E, RINGER S, LUKOŠIŪTĖ K, et al. Discovering language model behaviors with model-written evaluations [EB/OL]. [2023-12-19]. https://arxiv.org/pdf/2212.09251.

75 ACAR A, AKSU H, ULUAGAC A S, et al. A survey on homomorphic encryption schemes: theory and implementation [J]. ACM computing surveys (CSUR), 2018, 51 (4): 1-35.