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*

颜见智[1,2] 何雨鑫[1] 骆子烨[1] 胡 晗[1] 范士喜[3] 汤步洲[1,2]

([1] 哈尔滨工业大学(深圳) 深圳 518055 [2] 鹏城实验室 深圳 518055
[3] 深圳职业技术大学 深圳 518055)

〔摘要〕 **目的/意义** 为快速适应新型人工智能技术发展,精准把握医疗人工智能发展方向,亟须系统地分析和梳理生成式大语言模型在医疗领域的潜在典型应用和面临的挑战。**方法/过程** 调研分析文献与公开报道,梳理总结生成式大语言模型在医疗领域不同任务中的应用尝试和评估结果。**结果/结论** 生成式大语言模型在医疗领域的应用逐渐增多,为医疗服务、医学研究和教育等方面提供智能辅助,同时也面临诸多挑战,如其本身存在的幻觉问题,以及数据隐私保护、伦理、结果可控性和算法可解释性等问题。

〔关键词〕 生成式大语言模型;医疗;人工智能
〔中图分类号〕R-058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2023.09.003

Generative Large Language Models in the Medical Domain: Potential and Typical Applications and Challenges

YAN Jianzhi[1,2], *HE Yuxin*[1], *LUO Ziye*[1], *HU Han*[1], *FAN Shixi*[3], *TANG Buzhou*[1,2]

[1] *Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China*; [2] *Peng Cheng Laboratory, Shenzhen 518055, China*; [3] *Shenzhen Polytechnic University, Shenzhen 518055, China*

〔**Abstract**〕 **Purpose/Significance** In order to quickly adapt to the development of new artificial intelligence (AI) technologies and accurately grasp the development direction of medical AI, it is urgent to systematically analyze and sort out the potential and typical applications and challenges of generative large language models in the medical field. **Method/Process** The paper investigates and analyzes the literature and public reports, systematically summarizes the attempts and evaluation results of generative large language models in different tasks in the medical field. **Result/Conclusion** The application of generative large language models in the medical field is gradually increasing, covering almost all medical informatics tasks from medical information extraction, text classification, information retrieval, question and answer and dialogue to physician examination, medical record generation, medical result prediction, drug development and medical image analysis, etc., which can provide intelligent assistance for medical services, medical research and education. At the same time, the application of generative large language models in the medical field also faces many challenges, such as the hallucination problem, data privacy protection, ethics, controllability of results and interpretability of algorithms.

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

〔修回日期〕 2023-09-20
〔作者简介〕 颜见智,博士研究生;通信作者:范士喜,博士;汤步洲,博士,研究员,博士生导师。
〔基金项目〕 科技创新 2030——“新一代人工智能”重大项目(项目编号:2021ZD0113402);国家自然科学基金项目(项目编号:62276082)。

1 大语言模型简介

语言模型是计算语言学范畴概念，用于建模任意字词序列属于自然语言的概率。在深度学习流行之前，语言模型多是基于统计的 N - gram 语言模型；随后，基于人工神经网络的语言模型（神经概率语言模型）逐渐占据主导地位。当神经概率语言模型的有效参数规模达到一定量级就成为大语言模型。

2003 年 Bengio Y[1] 在其论文 *A Neural Probabilistic Language Model* 中首次提出神经概率语言模型，基于词嵌入向量和多层感知机计算文本中每个词的条件概率。受限于当时的算力资源，该工作并没有得到太多重视。2013 年 Mikolov T[2] 延续 Bengio 的思想提出 word2vec，利用各种高效的损失设计成功实现在包含 16 亿词的语料库上的预训练。2015 年 Dai A M 等[3] 提出基于长短期记忆神经网络（long - short term memory，LSTM）的语言模型，提出先利用语言模型任务在大规模语料上进行预训练，再在下游任务微调的思路。

然而前馈神经网络的表示能力较弱，循环神经网络又难以高效并行训练，因此一种基于注意力机制的神经网络 Transformer[4] 开始受到关注。从 2018 年起 Transformer 几乎成为神经概率语言模型的标配，研究者也习惯将这类语言模型称为预训练语言模型（pre - trained language models，PLMs）。根据模型架构不同，可以将 PLMs 分为以下 4 类：一是基于双向编码器的 PLMs，如双向编码器表征（bidirectional encoder representation from transformers，BERT）[5] 及其变种。二是基于单向解码器的 PLMs，如生成式预训练 Transformer（generative pre - trained transformer，GPT）系列模型[6]、PaLM 系列模型[7]、LLaMa 系列模型[8] 以及 BLOOM/Z[9]。三是基于编码器 - 解码器的 PLMs，如 BART[10]、T5 系列模型[11] 和 UL2[12]。四是基于混合掩码解码器的 PLMs，如 XLNet[13]、UniLM[14] 以及 GLM 系列模型[15]。目前这 4 类 PLMs 中只有后 3 类能够成功完成量变到质变的跃迁，成为大语言模型；而 BERT 等基于双向编码器的 PLMs 则止步不前。原因也许在于去噪自编码这一预训练目标较简单，不需要大规模参数模型就能完成得很好。单纯的去噪自编码任务无法充分激发预训练语言模型的潜力。

而自回归生成的预训练目标则更具挑战性。模型需要在对世界进行高质量建模的同时具备强大的推理能力。研究结果表明只有当模型参数规模达到一定量级，模型才会涌现出这些能力[16]。没有大语言模型就无法很好地完成自回归生成；没有自回归生成这样一个高难度的预训练目标，大语言模型就没有产生的必要。因此，当提及大语言模型（large language model，LLM）时，实际上是指生成式大语言模型。

虽然 LLM 能涌现语义理解、文本生成和逻辑推理能力，但还无法较好地服从人类指令，其生成内容也不一定符合人类价值理念。因此在完成 LLM 的预训练后，一般会进行有监督微调（supervised fine - tuning，SFT）[17] 和基于人类反馈的强化学习（reinforcement learning from human feedback，RLHF）[18]，使 LLM 能够服从人类指令并生成符合人类价值观的内容，成为实用的人工智能助手。此处有监督微调采用的上下文指令学习样本[19] 是一种特殊的提示形式[20]。LLM 经过有监督微调，甚至能掌握调用外部工具的能力，具备成为人机交互的统一接口、重塑现代信息处理系统的潜力。

2 生成式大语言模型在医疗领域的潜在典型应用

生成式 LLM 因其出色的语义理解、文本生成和逻辑推理能力，正在被尝试应用于多个领域。在医疗领域，从基础的医疗信息抽取、医疗实体标准化，到常用的文本分类、信息检索、问答和对话等应用，再到医疗领域特有的医师考试、病历生成、医疗结果预测、药物研发和医学影像分析等任务均有尝试，取得了令人惊喜的结果。国内外代表性生成式大语言模型的基本特点及性能测试情况如下。

2.1 ChatGPT

ChatGPT（chat generative pre - trained transform-

er）是 OpenAI 于 2022 年 11 月 30 日发布的一款基于人工智能技术的聊天机器人，基于含有 1 750 亿参数的生成式大语言 GPT－3.5 模型[21]开发，能与用户以问答的形式进行自然语言交互，为用户提供通用、有用信息和建议。尽管 ChatGPT 没有专门针对医疗领域进行微调，但也具有良好的医疗领域任务处理能力。ChatGPT 基本能通过美国职业医师资格考试（United States Medical Licensing Exam，USMLE），并能提供较好的解释[22]；未能通过中国国家医师资格考试（Chinese National Medical Licensing Examination，CNMLE），但已表现出很大潜力[23]。在基础生命支持（Basic Life Support，BLS）和高级心血管生命支持（Advanced Cardivascular Life Support，ACLS）考试中，BLS 成绩较好，ACLS 成绩较差，但均未通过[24]。ChatGPT 也可应用于临床试验人员招募，经过临床实体识别、否定信息识别、关键词抽取和临床试验检索等系列提示学习指令，获得优于传统检索和基于 BERT 关键信息抽取检索方法的性能[25]。2023 年 3 月 14 日 OpenAI 发布 GPT－4，更新之后的 ChatGPT（即 ChatGPT 4.0）能力得到很大提升。在 USMLE 问题上正确率达 90%[26]；在 CNMLE 的中英文数据集和中国全国医学研究生入学考试的中文数据集上均获得超过 80% 的分数，明显优于前一版本[27]。尽管两个版本的 ChatGPT 在回答语言流畅性方面性能出色，但在错误回答方面依然存在较大比例的幻觉，开放领域幻觉现象更为明显。在出院小结逻辑一致性和小组学习语言流畅性与满意度的小规模测试中，ChatGPT 3.5 不能满足出院小结逻辑一致性要求，ChatGPT 4.0 能在 60% 的情况下满足。两个版本 ChatGPT 在小组学习任务上的语言流畅性和满意度达到 100%。

2.2 Med－PaLM

2023 年 5 月 10 日谷歌发布新一代人工智能大语言模型 PaLM 2 以及基于 PaLM 2 的医疗领域变体 Med－PaLM[28]。Med－PaLM 可以检索医学知识、回答问题、生成有用的模板和解码医学术语，以及从图像（如 X 光胸片）中解读信息。在 MedMCQA 数据集上，Med－PaLM 获得 72.3% 的分数，超过 Flan－PaLM14% 以上，但略低于 GPT－4。在 PubMedQA 数据集上，Med－PaLM 获得 75.0% 的分数，低于 BioGPT－Large 的 81.0%。在 MMLU 临床主题上，Med－PaLM 在 6 个主题中的 3 个上表现最佳，而 GPT－4 在其他 3 个上表现更好。在 1 000 多个实际医疗场景问答中，Med－PaLM 在 9 项基准测试中有 8 项表现良好，相较于人类医生回答更受认可；72.9% 的回答被认为与医生回答一致。Med－PaLM 在 MedQA 数据集上的测试结果很好，但医学领域应用关乎人的健康，仅通过简单的基准测试难以全面评估模型的生成事实性和回答安全性。因此，除了在 MedQA 数据集上的直观定量对比，还进行了人工评估，选取 1 066 个消费者医疗问题，在 9 个与临床效用相关的属性（如事实性、医学推理能力和低风险性）上，Med－PaLM 在 8 个属性上给出了比医生评分更高的回答。

2.3 Galactica

大部分现有语言模型是基于爬虫爬取、未经整理的大规模语料训练构建的，而 Galactica[29]大模型是在大量且精心构造的人类科学知识语料库上训练得到的。所使用语料库包括 4 800 余万篇论文、教科书和讲义、数百万种化合物和蛋白质、科学网站、百科全书等。Galactica 在 MedQA 数据集上的准确率达到 44.4%，在 PubMedQA 数据集上达到 77.6%，在 BioASQ 数据集上达到 94.3%。

2.4 GatorTronGPT

为了研究医学领域的生成式大语言模型，并评估其在医学研究和医疗保健领域的实用性，佛罗里达大学研究团队整理了其附属医院包含 820 亿 token、去隐私信息的临床文本，以及包含 1 950 亿 token 的 Pile 数据集，将之一起用于训练 GatorTronGPT[30]。该模型使用 GPT－3 架构从头开始训练，在医疗信息抽取、文本相似度计算等任务上均超过以往最佳性能。在 PubMedQA 数据集上取得 77.6% 的准确率，在 MedQA 数据集上取得 45.1% 的准确率，在 MedMCQA 数据集上取得 42.9% 的准确率。

2.5 PubMedGPT

斯坦福基础模型研究中心和 MosaicML 联合开发了一种经训练可以解释生物医学语言的大型语言模型 PubMedGPT[31]。其采用 Pile 数据集的 PubMed Abstracts 和 PubMed Central 部分训练得到。在 MedQA 数据集上的准确率达到 50.3%，在 PubMedQA 数据集上达到 74.4%，在 BioASQ 数据集上达到 95.7%。在使用较少训练数据的情况下获得良好性能。

2.6 PMC－LLaMA

PMC－LLaMA[32] 是上海交通大学于 2023 年 4 月发布的医学大语言模型。其基于 LLaMA－7B 模型，在 480 万篇生物医学学术论文数据集基础上微调得到。在 3 个生物医学问答数据集（USMLE、MedMCQA 和 PubMedQA）上对比全量参数微调和 PEFT 微调两种方式。与 LLaMA－7B 相比，全量参数微调得到的 PMC－LLaMA 在 USMLE 和 MedMCQA 上均获得明显的性能提升，在 PubMedQA 上则没有提升；PEFT 微调得到的 PMC－LLaMA 在 3 个数据集上均获得明显的性能提升。通过 GPT－4 评价，PMC－LLaMA 比 LLaMA 在 zero－shot 任务上能提供更多和输入相关的上下文，表现出对医学背景知识更深入的理解能力。受限于设备性能，PMC－LLaMA 仅在 480 万篇生物医学论文数据集上训练了 5 轮，模型训练可能并不充分，暗示 PMC－LLaMA 还存在很大潜能。

2.7 MedGPT

MedGPT 是医联于 2023 年 5 月 25 日发布的国内首款基于 Transformer 框架的医疗大语言模型。模型从医疗知识图谱中获取大量准确、结构化的医疗知识，并使用经过整理的近 20 亿条真实世界中的医患沟通对话、检验检测和病历信息进行训练，使用 800 万条高质量结构化临床诊疗数据进行微调，最后通过医生的真实反馈进行强化学习。MedGPT 率先实现使 AI 大模型与真实患者连续自由对话的功能，能够整合多种医学检验检测模态能力，支持医疗问诊中的多模态输入和输出。问诊结束后，MedGPT 还能给患者开具合适的医学检查项目，再根据问诊和检查结果，为患者设计治疗方案，实现全流程覆盖的智能化诊疗。医联抽取 532 名复诊患者档案进行信息脱敏，并进行模拟首诊实验，结果显示 MedGPT 的诊断结果与线下门诊的原有诊断吻合率超过 97.5%，充分证明 MedGPT 的诊断能力。MedGPT 能从多轮问诊中收集足够信息，逐步得出诊断结论，诊断的准确率和安全性较高，已达到主治医师水平。

2.8 山海大模型

山海大模型是云知声于 2023 年 5 月 24 日发布的通用领域大模型，已进入有序迭代阶段。其能快速积累特定领域的专业知识，通过语料的不断迭代升级突破专业能力，在医疗领域的性能也十分优异。为提供更加全面、专业的医疗知识支持，山海大模型学习了大量医学文献、医学教材和病历数据，得到医疗基座模型。2023 年 6 月在 MedQA 任务上的准确率提升到 87.1%，超越了 Med－PaLM；临床执业医师资格考试提升至 523 分（总分 600 分），超过 99% 的考生。同年 7 月 28 日迎来新一轮迭代升级，并在当月的全球大模型综合性考试评测（C－Eval）中跻身榜单前 10 名。在同年 8 月 24—27 日举办的第十七届全国知识图谱与语义计算大会上，云知声团队通过大赛官方提供的训练数据对医疗基座模型进行指令微调，并采取数据增强、思维链等技术手段不断优化模型表现，再利用模型融合技术构建 UNIGPT－MED 比赛模型，在 PromtCBLUE 医疗大模型评测中夺得 AB 双榜冠军。同年 8 月 28 日山海大模型再次迭代升级，参数规模达到千亿级。山海大模型 2.0 在预训练阶段使用海量的医学病历、医学教材、临床指南和医学文献等数据，并在对齐阶段使用人机结合方法构建近百万级的病历理解、医学考试和医学知识问答等指令学习数据。当月实测性能在全球大模型综合性考试评测（C－Eval）中超越 GPT－4，以平均 70 分的成绩位列第 3 名。

2.9 添翼医疗大模型

添翼医疗大模型是东软于 2023 年 6 月发布的医

疗领域大模型，与飞标医学影像标注平台 4.0、基于 Web 的虚拟内窥镜等多款“AI + 医疗行业应用”相结合，形成在“AI + 医疗领域”的“1 + N”组合，加速推动了东软“AI + 领域应用”的人工智能生态图谱战略布局。医生能通过自然语言与添翼交互，快速准确地完成医疗报告与病历、医嘱开立。添翼能成为患者全天私人专属医生，提供全面的诊后健康饮食、营养与运动建议等。

2.10 百度灵医

百度灵医（灵医 bot）是基于百度文心大模型，融合全国超 800 家医院、4 000 多家基层诊疗机构的智慧医疗服务经验，推出的医疗领域对话机器人。此外，灵医 bot 所使用医学知识图谱包含万级医学专业书籍、亿级权威专家审校的科普内容；训练数据来自超百万条经三甲医院主任医师带队的医学专家队伍标注、评估和整理的医学数据；涵盖长/短医疗文本分类、医疗问答、医患对话和病历生成、冲突检测、因果关系推理、病灶检测、分割与分类等高质量标注语料。面向医疗领域从业者，灵医 bot 能对自有知识内容进行快速问答，提供病历生成、辅助治疗、病历质控等服务。面向患者，灵医 bot 升级了智能分导诊、预问诊等功能，提升病因分析、危急情况识别、检验检查识别、口语表达识别的及时性和准确性。2023 年 7 月 20 日百度“灵医智惠”与固生堂联合举办了大模型战略合作启动仪式，促成了国内中医药领域首个大模型应用落地，并在同年 9 月 19 日正式发布。

2.11 Deepwise MetAI

Deepwise MetAI 是深睿医疗于 2023 年 4 月推出的智慧影像和大数据通用平台，也是国内首个融合计算机视觉、自然语言处理、深度学习等技术构建的平台。以深睿自主研发的通用医学影像理解模型 DeepWise - CIRP Model 为支撑，将影像科日常应用产生的数据结构化，进而实现影像处理、打印、诊断、会诊、教学、科研一站式全周期智能管理，并实现跨越呼吸系统、心血管系统、神经系统、运动系统、女性关爱等多个领域图文并茂的 AI 生成式结构化报告。Deepwise MetAI 在科研和市场需求领域均获得认可。在科研方面，2023 年 6 月 16 日深睿医疗与香港大学、四川大学华西医学院、澳门科技大学合作开展关于多模态数据的医学诊断研究，使用 IRENE 深度学习框架在多模态数据上训练医学诊断模型，显著改善 4 种疾病（支气管扩张、气胸、间质性肺疾病和结核病）的诊断效果[33]。

2.12 ClouD GPT

ClouD GPT 是智云健康于 2023 年 5 月发布的慢性病管理领域的首个大语言模型，由 ClouDr Machine Learning Infrastructure 基础平台提供智能诊断技术，并成为智云医疗大脑的一部分。经过大量、专业的医学数据训练，ClouD GPT 能够应对不同模式下的复杂情况。目前智云健康已在医院及互联网医院的软件即服务（software as a service，SaaS）中安装应用 ClouD GPT，主要用于临床辅助决策。在医院 SaaS 方面，ClouD GPT 能够全面分析患者病情，为同类疾病提供预警及建议治疗方案，协助医师更快、更精准地确立诊疗方案。在互联网医院 SaaS 方面，ClouD GPT 能够协助医生及药师进行处方质量控制，并提升医生诊疗方案的效率及准确性。此外，得益于智云医疗大脑，ClouD GPT 还可以应用于 AI 药物和器械研发，为慢性病数字医疗提供多项关键技术。例如，在心血管疾病领域成功研发了“ClouDTx - CVD”数字疗法，是首个公开发表的在心血管疾病治疗领域采用数字疗法干预血脂的临床研究。

2.13 其他

国内已发布的其他医疗领域大模型，包括以开源通用预训练大语言模型为基座的哈尔滨工业大学的本草（原名华佗）、香港中文大学（深圳）的华佗等，以华为鲲鹏生态下自研通用预训练大语言模型脑海为基座的鹏城实验室的扁鹊等。

3 生成式大语言模型在医疗领域应用面临的挑战

3.1 缺乏统一评估

医学依赖于专家知识和经验，生成式大语言模型依赖于数据，医疗专家知识和经验往往蕴含在医疗数据中，这为生成式大语言模型缓解医疗资源短缺提供了可能性。未来生成式大语言模型在医疗领域应用前景广阔，但模型评估仍存在诸多挑战。虽然已有在公开数据集上的模型评估、基于 ChatGPT 4.0 的自动评估，甚至还有专业医生的人工评估，但这些评估均存在规模小、不全面、封闭和难以复制等问题。目前，尽管已经涌现出各种各样的生成式大语言模型，但由于缺乏统一评估标准，不同模型的性能难以客观全面地进行比较，这也导致不同研究结果难以互相验证和重现，从而大大降低模型可信度。

3.2 幻觉

幻觉指大模型在处理常识问题时，生成的内容在语义或句法上符合逻辑，但内容不正确或无意义[34]。医疗领域错误或不准确的信息可能对患者健康产生严重影响。因此，应用生成式大语言模型时准确性和可靠性至关重要。评估和减少生成式大语言模型在医疗领域中的幻觉是确保模型高准确性和可靠性的关键。为此，研究者最近提出了一些基准数据集。例如 Med - HALT[35]，包括创新的检测方式，并涵盖多国医疗检查，可以评估 Text - Davinci、GPT - 3.5、LLaMa - 2、MPT 和 Falcon 等 LLMs 的性能。总体而言，面向医疗领域的幻觉数据集仍然匮乏，这一情况可能是由医疗数据隐私和安全性导致的。

3.3 数据隐私保护

医疗数据通常包含敏感信息。在使用生成式大语言模型时，必须确保数据的隐私和安全得到充分保护，以防止数据泄露和滥用。否则可能会引发敏感信息滥用、患者对医疗机构信任度降低、医患矛盾激化等一系列重大问题。一是在数据合规性方面，医疗数据通常受到法规（如美国《健康保险携带和责任法案》(Health Insurance Portability and Accountability Act, HIPPA）和欧盟《通用数据保护条例》（General Data Protection Regulation, GDPR）等）的约束，需要确保生成式大语言模型在训练和应用时符合这些法规，包括数据访问控制、审计跟踪、数据脱敏等合规性措施。在一些情况下，医疗领域需要多个组织之间共享数据以进行合作研究。确保这些共享数据的隐私和安全性是一个复杂的问题，需要设计安全的数据共享协议和技术。

3.4 伦理

为了确保生成式大语言模型的开发和应用符合道德准则和法规，建立相应伦理审查和监管机制将有助于提高医疗 AI 大模型系统的可信度。应建立专门的伦理审查委员会，对生成式大语言模型数据收集、存储和处理，数据中偏见影响的评估等方面进行全面跟踪监管，以确保生成式大语言模型的合法性、道德性和可信度。

3.5 结果可控性

与通用领域相比，医疗领域因其特殊性，对生成式大语言模型的结果可控性要求更高，以确保其合理性、安全性和符合医疗实践规范。但生成式大语言模型的高度复杂性和黑盒性质，使其生成的结果难以有效控制和管理。缺乏结果可控性表现在算法本身可控难度大，以及可能引发的医疗严重后果和法律法规风险等多个方面。

3.6 算法可解释性

深度学习模型可解释性差的问题至今仍难以解决。就医疗生成式大语言模型而言，难以解释其决策过程以及模型的错误或不当行为会带来以下问题。首先，医疗专业人士和患者难以理解模型为何作出特定的医疗决策或提供特定的诊断建议。医生可能会不信任和否定模型建议，以作出最佳治疗决策。同时患者希望了解为什么模型提供特定医疗建议，可解释性的缺乏会导致患者对治疗方案不信

任。其次，可解释性不足可能导致模型的错误无法被及时发现和修正。如果模型产生不准确结果或者基于不当数据进行决策，但无法解释为何会出现这种情况，就可能延误患者治疗或带来不当医疗建议。

3.7 跨领域迁移能力

一是不同领域数据具有不同特点和分布。医疗数据可能包含丰富的患者病历、医学图像和实验数据，不同医疗领域的数据特征和分布截然不同。例如将一个肺部疾病模型迁移到眼科领域可能会面临数据不匹配问题。模型需要适应新领域数据，需要大量标记数据和领域适应技术。每个医疗细分领域都有其独特的临床实践和标准，跨领域迁移需要将领域专业知识整合到模型中，以确保生成的结果与特定领域最佳实践相符。二是医疗大模型跨领域迁移能力受到伦理和法律法规的影响。不同领域的医疗数据可能受到不同的伦理和法规约束。将模型迁移到新领域需要确保其符合新领域法规要求，尤其是涉及患者隐私和数据保护的问题。三是跨领域迁移面临风险管理问题。跨领域迁移可能伴随一定风险，包括模型性能下降、不准确的结果以及患者安全等问题。

4 医疗领域生成式大语言模型未来发展方向

4.1 建立统一评估体系

短期内发展医疗大语言模型的首要任务是建立统一评估体系。理想的评估体系应具备以下 6 个特点。一是全面性，能够全面涵盖各个科室，并覆盖诊疗全流程（导诊、首诊、复诊、复健、预防）。二是可重复性，能够重复实施，并对相同模型给出一致的评估结果。三是区分性，能够对不同水平的模型给出有区分度的评估结果。四是权威性，评估应由权威机关主持，评测内容应高度保密，评估流程应高度透明，评估结果应具有一定法律效力。五是时间可扩展性，生物医学处于高速发展中，人类对疾病的认知和诊疗方式亦不断进步，为了体现模型掌握最新医学知识的能力，需要每隔一段时间对评测内容进行更新。六是多维度、多粒度，评估结果不单要体现模型综合水平，还要具体反映模型在用户友好性、事实性、内容一致性等不同维度的能力，并能够细化指出模型的具体事实性错误或其他扣分项。目前可以从现有基础任务数据集（如英文的 I2B2、N2C2、PubMedQA[36]、MedMCQA[37] 和 USMLE[38] 等，中文的 CBLUE[39]、CNMLE、CMB[40] 等）开始，逐渐组成多层次、多维度的评估体系和数据集矩阵。

4.2 多模态

虽然目前大语言模型已经能够在文本信息处理上取得不错效果，但文本只是医学信息的一种模态，医学信息还包括视觉、听觉、医学影像、基因组学等其他重要模态。这些非文本模态信息一方面是医患交互的重要接口，另一方面能够为大语言模型提供丰富的真实世界语境，约束大语言模型的生成内容。因此，探索医疗多模态大语言模型是必由之路。

4.3 与知识图谱深度结合

医疗知识图谱涵盖细粒度、高质量的人类医学专家知识，恰好能与生成式大语言模型形成互补。因此将大语言模型与知识图谱相结合有可能解决大语言模型的幻觉问题，提高生成内容的可控性和可解释性。然而目前知识图谱在大语言模型领域的应用主要还停留在信息检索方面，如何将大语言模型输入、输出的文字和知识图谱细粒度地对齐以实现文本生成过程与知识图谱的深层次耦合是未来值得探索的方向。

4.4 个性化医疗

随着个性化医疗的发展，大语言模型在医疗领域的应用也将更加个性化。未来，可以根据患者的个性化需求和特征，定制开发适用于不同场景和人群的大语言模型应用，如个性化健康管理、个性化药物研发等。为了实现这一目标，一方面可以尝试将患者的既往病历或体检报告等医疗记录作为大语言模型的上下文（这需要模型能有效支持非常规的

上下文长度)；另一方面，可以尝试采用更细粒度的数据分析和挖掘技术，显式地挖掘患者的个性化特征和需求，为个性化医疗提供精准支持。

5 结语

生成式大语言模型在医疗领域的应用前景广阔，但仍存在亟待解决的关键问题，有待进一步深入研究和持续改进。未来，需要学术界和企业界继续加强相关研究和探索，快速推动生成式大语言模型在医疗领域的应用和发展。

参考文献

1 BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. Journal of machine learning research, 2003, 3 (2): 1137 - 1155.

2 MIKOLOV T, CHEN K, CORRADO G S, et al. Efficient estimation of word representations in vector space [C]. Scottsdale: International Conference on Learning Representations, 2013.

3 DAI A M, LE Q V. Semi - supervised sequence learning [C]. Montreal: International Conference on Neural Information Processing Systems, 2015.

4 ASHISH V, NOAM S, NIKI P, et al. Attention is all you need [C]. Red Hook: The 31st International Conference on Neural Information Processing Systems (NIPS'17), 2017.

5 DEVLIN J, CHANG M, LEE K, et al. BERT: pre - training of deep bidirectional transformers for language understanding [C]. Minneapolis: The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

6 RADFORD A, NARASIMHAN K. Improving language understanding by generative pre - training [EB/OL]. [2023 - 09 - 13]. https://cdn.openai.com/research - covers/language - unsupervised/language_understanding_paper.pdf.

7 CHOWDHERY A, NARANG S, DEVLIN J, et al. PaLM: scaling language modeling with pathways [EB/OL]. [2023 - 09 - 13]. https://arxiv.org/abs/2204.02311.

8 TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models [EB/OL]. [2023 - 09 - 13]. https://arxiv.org/abs/2302.13971.

9 SCAO T L, FAN A, AKIKI C, et al. BLOOM: a 176B - parameter open - access multilingual language model [EB/OL]. [2023 - 09 - 13]. https://arxiv.org/abs/2211.05100.

10 LEWIS M, LIU Y, GOYAL N, et al. (2019). BART: denoising sequence - to - sequence pre - training for natural language generation, translation, and comprehension [C]. Seattle: Annual Meeting of the Association for Computational Linguistics, 2020.

11 RAFFEL C, SHAZEER N M, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text - to - text transformer [EB/OL]. [2023 - 09 - 13]. https://arxiv.org/abs/1910.10683.

12 TAY Y, DEHGHANI M, TRAN V Q, et al. UL2: unifying language learning paradigms [EB/OL]. [2023 - 09 - 13]. https://arxiv.org/abs/2205.05131.

13 YANG Z, DAI Z, YANG Y, et al. XLNet: generalized autoregressive pretraining for language understanding [C]. Vancouver: The 33rd International Conference on Neural Information Processing Systems (NIPS'19), 2019.

14 DONG L, YANG N, WANG W, et al. Unified language model pre - training for natural language understanding and generation [C]. Vancouver: The 33rd International Conference on Neural Information Processing Systems (NIPS'19), 2019.

15 DU Z, QIAN Y, LIU X, et al. All NLP tasks are generation tasks: ageneral pretraining framework [EB/OL]. [2023 - 09 - 13]. https://arxiv.org/abs/2103.10360v1.

16 WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models [EB/OL]. [2023 - 09 - 13]. https://arxiv/abs/2206.07682.

17 GUPTA P, JIAO C, YEH Y, et al. InstructDial: improving zero and few - shot generalization in dialogue through instruction tuning [C]. Abu Dhabi: Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022.

18 OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback [EB/OL]. [2023 - 09 - 13]. https://arxiv.org/abs/2203.02155.

19 BROWN T B, MANN B, RYDER N, et al. Language models are few - shot learners [EB/OL]. [2023 - 09 - 13]. https://arxiv.org/abs/2005.14165.

20 LIU P, YUAN W, FU J, et al. Pre - train, Prompt, and Predict: asystematic survey of prompting methods in natural

language processing [EB/OL]. [2023 - 09 - 13]. https://arxiv. org/abs/2107. 13586.

21 OpenAI. GPT - 4 Technical Report [EB/OL]. [2023 - 09 - 13]. https: //arxiv. org/abs/2303. 08774.

22 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.

23 WANG X, GONG Z, WANG G, et al. ChatGPT performs on the Chinese National Medical Licensing Examination [J]. Journal of medical systems, 2023, 47 (1): 86.

24 NINO F, LUCIJA G, GREGOR Š, et al. Can ChatGPT pass the life support exams without entering the American heart association course [J]. Resuscitation, 2023, 185: 109732.

25 PEIKOS G, SYMEON S, PRANAV K, et al. Utilizing ChatGPT to enhance clinical trial enrollment [EB/OL]. [2023 - 09 - 13]. https: //arxiv. org/abs/2306. 02077.

26 LEE P, BUBECK S, PETRO J. Benefits, limits, and risks of GPT - 4 as an AI chatbot for medicine [J]. The New England journal of medicine, 2023, 388 (13): 1233 - 1239.

27 WANG H Y, WU W Z, DOU Z, et al. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI [J]. International journal of medical informatics, 2023, 177: 105173.

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

29 TAYLOR R, KARDAS M, CUCURULL G, et al. Galactica: alarge language model for science [EB/OL]. [2023 - 09 - 13]. https://arxiv. org/abs/2211. 09085.

30 PENG C, YANG X, CHEN A, et al. A study of generative large language model for medical research and healthcare [EB/OL]. [2023 - 09 - 13]. https: //arxiv. org/abs/2305. 13523.

31 VENIGALLA A, FRANKLE J, CARBIN M. Biomedlm: a domain - specific large language model for biomedical text [EB/OL]. [2023 - 09 - 13]. https: //crfm. stanford. edu/2022/12/15/biomedlm. html.

32 WU C, ZHANG X, ZHANG Y, et al. PMC - LLaMA: further finetuning llama on medical papers [EB/OL]. [2023 - 09 - 13]. https: //arxiv. org/abs/2304. 14454.

33 ZHOU H Y, YU Y, WANG C, et al. A transformer - based representation - learning model with unified processing of multimodal input for clinical diagnostics [J]. Nature biomedical engineering, 2023 (7): 1 - 13.

34 MAYNEZ J, NARAYAN S, BOHNET B, et al. On faithfulness and factuality in abstractive summarization [EB/OL]. [2023 - 09 - 13]. https: //arxiv. org/abs/2005. 00661.

35 UMAPATHI L K, PAL A, SANKARASUBBU M, et al. Med - halt: medical domain hallucination test for large language models [EB/OL]. [2023 - 09 - 13]. https://arxiv. org/abs/2307. 15343.

36 JIN Q, DHINGRA B, LIU Z, et al. Pubmedqa: a dataset for biomedical research question answering [EB/OL]. [2023 - 09 - 13]. https: //arxiv. org/abs/1909. 06146.

37 PAL A, UMAPATHI L K, SANKARASUBBU M. Medmcqa: a large - scale multi - subject multi - choice dataset for medical domain question answering [J]. Proceedings of machine learning research, 2022, 174: 248 - 260.

38 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.

39 ZHANG N, BI Z, LIANG X, et al. CBLUE: a Chinese biomedical language understanding evaluation benchmark [EB/OL]. [2023 - 09 - 13]. https: //arxiv. org/abs/2106. 08087.

40 WANG X, CHEN G, SONG D, et al. CMB: a comprehensive medical benchmark in Chinese [EB/OL]. [2023 - 09 - 13]. https: //arxiv. org/abs/2308. 08833.