DOI：10.61189/799037wwkyrc

·专家述评·

医学GPT研发的挑战与解决方案

白春学[1,2,3,4]

1. 复旦大学附属中山医院呼吸与危重症医学科，上海 200032
2. 上海市呼吸物联网医学工程技术研究中心，上海 200032
3. 上海市呼吸病研究所，上海 200032
4. 复旦大学附属中山医院AI+肺癌防治中心，上海 200032

［摘要］系统梳理医学GPT研发面临的关键挑战，结合国际最新综述、评价框架、伦理治理与监管文件，以及白春学教授《肺结节专家——BAIMGPT白皮书》，总结医学GPT从通用大模型走向临床可用系统的主要解决路径。基于近年发表于高影响力医学、数字医学与人工智能期刊的系统综述、方法学研究和真实工作流评价，同时结合WHO伦理治理文件与FDA人工智能器械生命周期建议，从事实可靠性、知识更新、数据治理、多模态整合、工作流适配、可解释性、偏倚公平性与责任边界等方面，对医学GPT研发的核心问题进行综合分析。现有证据显示，医学GPT的主要瓶颈并不只是一般意义上的“准确率不足”，而更集中体现为幻觉与事实错误风险高、对新增医学知识吸收有限、医学数据异质性强且标签稳定性不足、单一文本模型难以支撑真实临床中的多模态决策、离线高分与真实部署效果不一致、解释链与责任链不完整，以及偏倚、公平性和伦理治理问题突出。高质量研究进一步指出，当前LLM对信息顺序和信息量敏感，指令遵循性不足，尚未准备好承担自主临床决策。医学GPT研发的核心任务，不是单纯提升语言生成能力，而是将大模型重构为具备知识可靠性、流程适配性、可追溯性和制度可接受性的医学智能系统。现阶段医学GPT更适合作为知识增强工具和流程支持工具，而非替代临床主体的自动决策系统；未来应坚持“辅助决策而非替代决策”的基本边界，在强化数据治理、循证更新、真实世界验证和全生命周期治理的前提下稳步推进临床应用。

［关键词］ 医学GPT；大语言模型；临床决策支持；检索增强生成；数据治理；人机协同；专病智能体；BAIMGPT

［中图分类号］ R-0　　［文献标志码］ A

Challenges and solutions for the development of medical GPTs

Bai Chunxue[1,2,3,4]

1. Department of Respiratory and Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai 200032, China
2. Shanghai Respiratory IoT Medical Engineering Technology Research Center, Shanghai 200032, China
3. Shanghai Institute of Respiratory Diseases, Shanghai 200032, China
4. AI+Lung Cancer Prevention and Treatment Center, Zhongshan Hospital, Fudan University, Shanghai 200032, China

［**Abstract**］To systematically summarize the major challenges in developing medical GPT systems and, with reference to recent international reviews, evaluation frameworks, ethical and regulatory guidance, as well as Prof Chunxue Bai's BAIMGPT White Paper, to outline practical solutions for translating large language models into clinically usable systems. Recent high-impact systematic reviews, methodological studies, real-world workflow evaluations, and governance guidance were synthesized to examine the main issues in medical GPT development, including factual reliability, knowledge updating, data governance, multimodal integration, workflow adaptation, explainability, bias, fairness, and accountability. Current evidence indicates that the bottlenecks of medical GPT go well beyond imperfect accuracy. Major challenges include hallucinations and factual inconsistency, limited ability to absorb newly updated medical knowledge, heterogeneous clinical data and unstable labels, insufficient support for multimodal decision-making, weak adaptation to real-world workflows, incomplete explainability and accountability, and concerns regarding bias, fairness, and ethics. Current LLMs remain sensitive to information order and quantity and are not ready for autonomous clinical decision-making. The mission of medical GPT development is not simply to improve language generation, but to transform large models into trustworthy medical intelligence systems with reliable knowledge, workflow compatibility, traceability, and governance readiness. At present, medical GPT should be positioned as a tool for cognitive augmentation and workflow support rather than a substitute for clinical

［收稿日期］ 2026-01-05　　［接受日期］ 2026-03-01

［基金项目］ 四大慢病重大专项（2024ZD0529300）. Supported by Noncommunicable Chronic Diseases-National Science and Technology Major Project（2024ZD0529300）.

［作者简介］ 白春学，博士，主任医师、教授. E-mail：bai.chunxue@zs-hospital.sh.cn

judgment.

[**Key Words**] medical GPT; large language model; clinical decision support; retrieval-augmented generation; data governance; human-AI collaboration; disease-specific agent; BAIMGPT

生成式人工智能，尤其大语言模型的快速发展，使医学信息处理正由“数字化记录”走向“智能化组织”[1-2]。传统医疗信息系统的核心功能在于存储、检索与传输，而医学生成式预训练变换器（generative pre-trained transformer，GPT）进一步承担了理解、归纳、重写、解释与交互的任务，因而在临床文书生成、指南问答、患者教育、科研辅助和专病管理等方面展现出独特价值。近年综述[1-2]普遍认为，大语言模型（large language model，LLM）已成为生物医学与医疗人工智能的重要前沿之一，但其真正进入临床仍面临一系列系统性障碍。

需要强调的是，医学GPT并不是通用对话模型在医疗场景中的简单复制[1]。医学场景具有高风险、高责任、强时效与强监管等特点，任何看似轻微的事实偏差、知识过时、语义误导或流程脱节，都可能在真实诊疗中被放大为安全问题[3-5]。当前LLM在复杂临床决策中对信息顺序和信息量敏感，且难以自然嵌入既有工作流，因此尚不适合承担自主临床决策[3]。现有医疗LLM研究虽增长迅速，但评估口径并不统一，且真实工作流证据明显不足[5]。

在我国专病场景中，这一问题又具有更强的实践导向[6]。白春学教授在《肺结节专家——BAIMGPT白皮书》中提出，医学GPT不应停留在一般化医学问答，而应服务于肺结节等高价值专病场景中的筛查、风险分层、随访管理、患者教育和医患协同[7]。由此可见，医学GPT的先进性最终不取决于语言是否更自然，而取决于其能否与专病知识、循证证据、数字平台和真实工作流形成闭环[7]。这也正是本文讨论“挑战与解决方案”的现实出发点[7]。

1 幻觉与事实错误：医学GPT最根本的安全挑战

医学GPT研发首先必须解决幻觉与事实错误问题。医学内容不同于一般文本生成，其输出很可能被直接用于检查安排、随访计划、风险沟通甚至治疗建议，因此“表述流畅但事实不实”的回答比“明显答错”更危险。现有高质量研究反复显示，LLM在医学任务中的错误往往披着专业表达的外衣，具有较强迷惑性。

这一风险在临床决策场景中尤其突出。临床判断建立在不完整、非线性、带时间顺序且高度情境化的信息之上，远比医学考试题复杂[3,5]。当前LLM在面对这类任务时，不仅结果受提示变化影响显著，而且在长上下文中容易出现重点漂移与指令偏离，因此尚不足以承担自主临床决策[3]。这意味着，医学GPT的危险不只是“不会”，而是“会以足够像真的方式说错”。

从工程上说，这一问题决定了医学GPT不能被设计成无限制的自由生成系统，而必须被设计成受证据、规则和场景边界共同约束的输出系统。换言之，医学GPT的目标不是“像专家一样表达”，而是“在证据约束下稳定表达”。这一点构成后续检索增强生成（retrieval-augmented generation，RAG）、知识中台和规则护栏的理论基础[8]。

2 知识更新与版本管理：医学GPT不能停留在静态记忆

医学知识更新快，直接决定了医学GPT不能主要依赖预训练参数中的静态记忆。指南每年可能修订，药品适应证可能变化，支付规则和院内路径也在持续调整[8]。目前并没有证据表明“通用型临床LLM”能够稳定覆盖大范围临床任务，其部署仍然依赖具体任务与具体知识供给方式。

知识更新问题在医学中并不只是“新旧差异”，更是“版本与适用性问题”[8]。同一临床问题往往同时受到国际指南、国内共识、专科推荐、本院制度和可及性条件的共同影响。如果模型不能区分“何时、何地、针对何人群、依据哪一版证据”来作答，就容易出现表面正确、实际不可执行的回答。

因此，医学GPT研发必须引入动态知识中台。指南、药典、专病路径和院内制度应进行版本化管理，并通过检索增强生成在输出前被动态调用。只有这样，模型输出才不仅“像正确答案”，而且“是当前场景中的正确答案”。这一步看似是技术优化，实则是临床可用性的基本门槛。

3 数据异质与标签不稳：模型上限取决于数据治理质量

医疗数据看似丰富，真正可用于模型训练与部署的数据却常常充满噪声[9]。病历自由文本比例高，缩写体系杂糅，历史复制粘贴常见；影像、病理、检验和随访信息来自不同系统，字段不统一，时间

轴不完整[10]。当前医疗LLM研究在评估数据类型、任务设置、评价维度和应用场景方面缺乏统一性，这从侧面反映出基础数据质量与任务定义仍然不稳[5]。

更重要的是，医学任务的“标签”本身并不总是稳定[5,9]。许多临床问题没有唯一答案，而取决于疾病阶段、风险等级、资源条件和专家判断[9-10]。例如“是否进一步检查”“何时复查”“是否转诊”，都具有明显情境依赖。这使得医学GPT的研发不仅是模型学习问题，也是任务建模问题。若任务定义过粗、标签构建过度简化，模型即便取得较好离线表现，也可能在真实场景中迅速失效。

因此，数据治理不能被视为训练前的辅助步骤，而应被视为整个医学GPT工程的底层基础设施。术语标准化、模板统一、结构化抽取、时间轴重建、脱敏处理、专家标注和金标准验证集建设，都会直接决定模型输入边界与输出上限。没有高质量数据治理，所谓“高性能医学GPT”往往只是演示效果，而不是临床能力。

4　多模态整合不足：文本优势尚未自动转化为临床智能

当前多数医学GPT的优势主要体现在文本任务上，如摘要、改写、问答和说明生成。但临床决策从来不是单一文本任务，而是多模态、多时序、多角色共同作用的结果[10-11]。真实判断常常依赖影像、病理、检验、生命体征、既往治疗反应和患者报告结局的综合分析[10-11]。近年综述[1,10-11]普遍指出，医疗LLM的下一阶段关键并不只是更好的文本生成，而是与工具调用、检索系统及多模态模型的协同。

这一不足在专病场景中尤为突出。以肺结节管理为例，真正有临床价值的系统，不能只会解释“报告写了什么”，而应能结合结节影像学特征、既往CT变化、吸烟史、家族史与随访时间轴，生成符合风险分层逻辑的建议[7]。BAIMGPT之所以具有代表性，正是因为它强调肺结节筛查—评估—随访—教育的全流程整合，而不是单轮问答[7]。

因此，医学GPT若想从文本能力跨越到临床能力，必须从“纯语言模型”升级为“多模态专病智能体”。文本模型仍然重要，但更重要的是它能否调度其他模块，并把分散信息整合为结构化、可复核的临床支持输出。

5　工作流适配问题：离线成绩高，不代表真正能落地

医疗LLM研究中最常见的误区之一，是将题库成绩、基准表现或回顾性模拟结果视为真实临床可用性的充分证据[3-5]。事实上，医学GPT是否真正有价值，取决于它能否在门诊、病房、会诊、MDT和随访等工作流中减少重复劳动、缩短信息整理时间并降低遗漏[4-5,10]。当前LLM不仅容易受信息呈现顺序影响，而且难以自然适配既有工作流，这意味着离线“答得好”与真实“用得上”之间存在明显鸿沟[3]。

目前并无证据支持“一个通用临床LLM”可以跨场景稳定胜任广泛任务，因此更现实的策略是按具体任务、具体风险等级和具体场景部署模型[10]。也就是说，医学GPT研发的对象，不应只是模型，而应是“模型+界面+输入方式+复核路径+日志机制”的完整工作流单元[4,10]。

由此可见，医学GPT要想真正落地，关键不只是更高的能力分数，而是更低的部署摩擦[3-5,10]。它必须接得进现有系统、嵌得进现有流程、经得起临床复核，并最终为医生节省时间而不是增加新的操作负担[4,10]。

6　可解释性与责任边界：医学GPT必须被审计，而不只是被赞叹

临床信任并不来自模型“写得像医生”，而来自模型的输出“有据可查、可被复核、可被问责”[5,12-14]。医生真正关心的问题包括：模型依据了哪些数据？调用了哪一版指南？是否存在不确定性？最终由谁负责[12-14]？当前医疗LLM研究仍缺乏足够统一和稳健的评价框架，这意味着很多“高表现”结果还不足以直接转化为可审计的临床能力[5]。

与此同时，WHO关于人工智能健康治理的文件和FDA关于AI医疗器械软件功能生命周期管理的建议，都强调透明性、问责性、持续监测和风险管理的重要性[13-14]。对医学GPT而言，这些原则同样适用：高风险输出必须可追踪，知识来源必须明确，异常输出必须可回溯，人工监督必须被记录[13-14]。

因此，医学GPT真正需要的不是表面“会解释”，而是深层“能审计”[12-14]。一个临床可接受的系统，应能清楚展示证据来源、关键变量、规则校验结果以及人工确认节点。只有当证据链和责任链同时存在，医学GPT才可能逐步获得真实世界的临床信任。

7　偏倚、公平性与本土化：不能只追求平均表现

医学GPT的另一个深层问题，是偏倚、公平性和本土适配[5,13,15-16]。现有研究往往过度聚焦总体

性能，而对不同亚群、不同语言、不同医疗资源水平下的表现差异关注不足[5]。这意味着，一个“平均准确率较高”的模型，仍可能在某些特定人群或特定场景中持续输出不利结果。

对中国医学场景而言，本土化校准尤其关键[7]。国际高等级证据固然重要，但若模型忽视本土指南、基层医疗条件、患者沟通习惯和分级诊疗结构，就可能出现“证据上先进、执行上脱节”的情况[16]。BAIMGPT所强调的专病路径与本土流程整合，正提示医学GPT必须在国际证据与本土实践之间建立桥梁，而不能简单移植国外模型逻辑[7]。

因此，医学GPT的评价不应止于总体准确率，而应延伸到亚组稳定性、公平性、可及性、语言适配性和本土适应性[5,15-16]。公平性不是锦上添花的指标，而应成为研发起点中的基本原则[13,15-16]。

8 解决路径之一：知识中台+RAG+规则引擎

综合现有文献[1,6,8,10]，最可行的方向并不是继续依赖“更大的单体模型”，而是搭建知识驱动型系统架构。基础模型负责理解与生成，知识中台负责存放并更新指南、药典、共识和院内制度，RAG负责在回答前检索当前可用证据，规则引擎负责对剂量、禁忌证、危急值和高风险建议进行边界约束。临床部署应围绕任务需求选择LLM，而不是预设存在一个全能模型[10]。

这种架构的价值在于，它把医学GPT从“参数驱动的回答系统”转化为“知识驱动的临床支持系统”[1,6,10]。对医疗来说，真正重要的并不是模型记住了多少知识，而是它在给定场景中能否调用正确、最新、适用的知识，并把这些知识转换为可复核的输出[8,10,13-14]。

9 解决路径之二：人机协同与风险分级部署

现阶段，医学GPT更合理的定位是认知增强工具和流程支持工具，而不是替代医生的自动决策主体[3-5,10]。当前LLM尚未准备好自主临床决策，这一结论已得到高质量研究的明确支持[3]。因此，部署策略必须围绕风险分级展开[14]。

低风险任务如病历摘要、出院小结初稿、患者教育材料和文献综述初稿，可采用“模型生成—人工快速确认”[4,10]。中风险任务如分诊建议、编码辅助、随访提醒，可采用“模型提示—规则校验—人工签发”[10]。高风险任务如诊断、治疗选择和危急值处置，则必须坚持“模型辅助—专家裁决”[3,13-14]。这一模式并非保守，而是当前证据条件下最稳妥也最现实的实施路径。

10 解决路径之三：专病化、多模态与全生命周期治理

从未来趋势看，医学GPT最有前景的方向，不是做一个“什么都能聊一点”的通用医学机器人，而是围绕肺结节、慢阻肺、哮喘、睡眠呼吸障碍、肿瘤MDT等高价值场景，建设专病化、多模态、流程化的智能体[7,10-11]。按任务和场景部署，而非追求无边界泛化，更符合当前证据与临床现实[10]。

同时，医学GPT研发不应在“模型上线”那一刻结束[5,10]。上线后的知识更新、性能漂移、异常监测、采纳率变化和安全事件回顾，都属于全生命周期治理的一部分[13-14]。没有标准化的持续评估，任何早期高表现都不足以保证长期安全有效[5]。也正因此，未来真正领先的医学GPT，不会只是“更会写”，而会是“更能被管理、更能被复核、更能持续进化”[10,13-14]。

综上所述，医学GPT研发已经从“证明大模型能够进入医学”转向“证明大模型能够安全、稳定、可监管地服务医学”的阶段。现有高质量研究一致表明，其主要瓶颈并不只是一般性的准确率问题，而是幻觉、知识时效性不足、数据异质性高、多模态整合困难、工作流适配不足、可解释性不充分以及偏倚与治理风险等系统性难题。尤其需要强调的是，当前LLM仍不适合自主临床决策[3]。

因此，医学GPT的未来不应继续沿着“更大的模型、更高的考题得分”这一单线逻辑推进，而应转向以知识中台、RAG、规则引擎、人机协同和全生命周期治理为核心的医学智能系统。白春学教授提出的BAIMGPT路径也进一步说明，在我国专病场景中，专病化、流程化和本土化，是医学GPT走向高质量落地方案的重要方向[7]。总体而言，现阶段医学GPT更适合作为辅助决策与流程支持工具，而非替代临床主体的自动决策系统[3-5,10]。未来只有在数据治理、循证更新、真实世界验证与制度治理同步强化的前提下，医学GPT才可能真正从原型走向平台、从概念走向临床现实[13-14]。

伦理声明 无。

利益冲突 所有作者声明不存在利益冲突。

作者贡献 白春学：选题、撰写、定稿。

参考文献

[1] Zhou J X, Li H Y, Chen S Y, et al. Large language models in

biomedicine and healthcare[J]. npj Artif Intell，2025，1：44.

[2] Busch F，Hoffmann L，Rueger C，et al. Current applications and challenges in large language models for patient care：a systematic review[J]. Commun Med，2025，5(1)：26.

[3] Hager P，Jungmann F，Holland R，et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making[J]. Nat Med，2024，30(9)：2613-2622.

[4] Artsi Y，Sorin V，Glicksberg B S，et al. Large language models in real-world clinical workflows：a systematic review of applications and implementation[J]. Front Digit Health，2025，7：1659134.

[5] Bedi S，Liu Y T，Orr-Ewing L，et al. Testing and evaluation of health care applications of large language models：a systematic review[J]. JAMA，2025，333(4)：319-328.

[6] Omiye J A，Gui H W，Rezaei S J，et al. Large language models in medicine：the potentials and pitfalls：a narrative review[J]. Ann Intern Med，2024，177(2)：210-220.

[7] 白春学. 肺结节专家——BAIMGPT 白皮书[J]. 元宇宙医学，2025，2(2)：55-64.

[8] Mahla N，Jadhav K S，Ramakrishnan G. Exploring gradient subspaces：addressing and overcoming LoRA′s limitations in federated fine-tuning of large language models [J]. arXiv，2024：2410. 23111.

[9] Tam T Y C，Sivarajkumar S，Kapoor S，et al. A framework for human evaluation of large language models in healthcare derived from literature review[J]. NPJ Digit Med，2024，7(1)：258.

[10] Li H Y，Fu J F，Python A. Implementing large language models in health care：clinician-focused review with interactive guideline[J]. J Med Internet Res，2025，27：e71916.

[11] Wang X F，Xiong Z X，Zou K，et al. Reasoning-driven large language models in medicine：opportunities，challenges，and the road ahead[J]. Lancet Digit Health，2026，8(1)：100931.

[12] Bedi S，Jiang Y X，Chung P，et al. Fidelity of medical reasoning in large language models [J]. JAMA Netw Open，2025，8(8)：e2526021.

[13] World Health Organization. Ethics and governance of artificial intelligence for health：WHO Guidance [M]. Geneva：WHO，2021.

[14] U. S. Food and Drug Administration. Artificial intelligence-enabled device software functions：lifecycle management and marketing submission recommendations [EB/OL]. [2025-01-07] (2026-01-30). https://www. fda. gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing

[15] Omiye J A，Lester J C，Spichak S，et al. Large language models propagate race-based medicine [J]. NPJ Digit Med，2023，6(1)：195.

[16] Haltaufderheide J，Ranisch R. The ethics of ChatGPT in medicine and healthcare：a systematic review on Large Language Models (LLMs) [J]. NPJ Digit Med，2024，7(1)：183.

引用本文

白春学. 医学 GPT 研发的挑战与解决方案[J]. 元宇宙医学，2026，3(1)：11-15.

Bai C X. Challenges and solutions for the development of medical GPTs[J]. Metaverse Med，2026，3(1)：11-15.