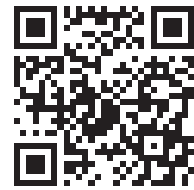


DOI: 10.61189/204742fpzdsf

· 专家述评 ·

医学智能体:体系架构、临床转化与可信应用

宋元林^{1,2,3,4}, 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. 复旦大学附属中山医院 AI+肺癌防治中心, 上海 200032
3. 上海呼吸物联网医学工程技术研究中心, 上海 200032
4. 上海市呼吸病研究所, 上海 200032

[摘要] 临床照护是在信息不完整、患者状态持续变化且资源受限的条件下,反复获取证据、更新判断并采取行动的序贯过程。现有医学人工智能(artificial intelligence, AI)仍主要完成识别、预测、内容生成等任务,因此,模型能力能否转化为安全、有效且可问责的临床方案,构成了医学智能体研究的中心科学问题。医学智能体并非仅能回答医学问题的大语言模型(large language model, LLM),而是由模型、知识、数据、记忆、工具、工作流、人工监督和治理机制共同构成的技术系统。本文综合截至2026年7月可获得的代表性研究、权威综述、报告规范和监管文件,提出“感知—知识—推理—规划—行动—反馈”六模块闭环架构,厘清其与医学大模型、临床决策支持系统、医疗辅助、多智能体系统和具身智能的区别,并贯通临床场景、证据路径与管理要求。现有研究显示,多智能体协作、检索增强生成、工具调用和反思机制可在部分受控任务中提高性能;但医学问答成绩对交互式临床表现的预测能力有限。2026年交互式基准表明,工具增强型智能体仅获得有限的准确性增益,却显著增加令牌消耗与时间延长。少量随机试验和真实工作流研究提示,生成式AI可能改善文书、试验预筛和部分边界明确的决策流程,但尚不足以支持无监督的自主诊疗。基于此,本文提出风险校准的四级自主性分层、六阶段证据路径和全生命周期治理框架。医学智能体应以“医生—智能体—工作流”联合系统的临床净获益衡量;近期目标并非脱离医生的完全自主,而是建立医生主导、患者参与、证据约束且责任可追溯的人机协同医疗。

[关键词] 医学智能体;大语言模型;多智能体系统;临床转化;人机协作;可信人工智能;全生命周期治理

[中图分类号] R 4 **[文献标志码]** A

Medical AI agents in health care: architecture, clinical translation, and trustworthy implementation

Song Yuanlin^{1,2,3,4}, Bai Chunxue^{1,2,3,4*}

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

[Abstract] Clinical care requires iterative evidence acquisition, revision of working hypotheses, and action under incomplete information, changing patient states, and finite resources. Yet medical artificial intelligence (AI) is still evaluated largely through recognition, prediction, content generation, or static question answering. The central scientific question is whether model capability can be translated into safe, effective, and accountable longitudinal clinical action. A medical agent is not merely a large language model (LLM), but a sociotechnical system integrating models, knowledge, data, memory, tools, workflows, human oversight, and governance. Drawing on representative studies, authoritative reviews, reporting guidance, and regulatory documents available through July 2026, this Review proposes a six-module closed-loop architecture—perception, knowledge, reasoning, planning, action, and feedback—clarifies boundaries with related AI systems, and links clinical use cases to evidence and governance requirements. Retrieval-

[收稿日期] 2026-08-10

[接受日期] 2026-09-13

[基金项目] 四大慢病重大专项(2024ZD0529300),国家自然科学基金(82570143、82170110),上海市科学技术委员会项目基金(20DZ2254400、21DZ2200600、20DZ2261200),上海市市级科技重大专项(ZD2021CY001)和上海市临床重点专科建设项目(shslczdk02201). Supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0529300), the National Natural Science Foundation of China (82570143 and 82170110), the Science and Technology Commission of Shanghai Municipality (20DZ2254400, 21DZ2200600, and 20DZ2261200), the Shanghai Municipal Science and Technology Major Project (ZD2021CY001), and the Shanghai Municipal Key Clinical Specialty (shslczdk02201).

[作者简介] 宋元林,博士,主任医师;E-mail: song.yuanlin@zs-hospital.sh.cn

*通信作者(Corresponding author).白春学,博士,主任医师、教授;E-mail: cxbai@fudan.edu.cn

augmented generation, tool use, reflection, and multi-agent collaboration have improved performance in selected controlled tasks. However, static medical question-answering scores are weak proxies for interactive clinical performance, and recent benchmarks show only modest accuracy gains despite substantial increases in token use and latency. A small number of randomized trials and real-world workflow studies suggest potential benefits for documentation, trial prescreening, and bounded decision processes, but do not support unsupervised autonomous diagnosis or treatment. We therefore propose four risk-calibrated autonomy levels, a six-stage evidence pathway, and lifecycle governance. Clinical value should be judged by the net benefit of the clinician - agent - workflow system. The near-term goal is not clinician-independent autonomy, but clinician-led, patient-engaged, evidence-constrained, and accountable human - agent care.

[Key Words] medical AI agent; large language model; multi-agent system; clinical translation; human - AI collaboration; trustworthy artificial intelligence; lifecycle governance

临床决策不是一次性分类或问答,而是在信息不完整、病情变化和资源约束下,持续采集资料、更新假设、选择检查、动态权衡疗效与风险、实施干预并观察结局的序贯过程。传统判别式模型主要完成预定义输入到输出的映射,如影像识别、风险预测或预警;基础模型和大语言模型(large language model, LLM)则以统一表征和自然语言接口扩展了多任务能力^[1-6]。但从“生成一个看似合理的答案”到“安全完成一项临床任务”,仍需要横跨信息、权限、时间和责任四重鸿沟。

智能体范式旨在缩小这一转化缺口:LLM不再只是终端输出器,而成为受约束的工作流编排器,围绕临床目标形成计划,调用检索、计算器、电子病历(electronic health record, EHR)、影像算法、处方核查或预约系统,并依据工具返回和患者状态重新规划^[7-9]。因此,系统表现不再仅由基座模型决定,还取决于知识可信度、数据完整性、记忆准确性、工具可靠性、权限边界和失败后的安全退出。相应地,潜在伤害也从错误文本延伸到延误检查、不当用药、错误转诊乃至物理干预。

现有研究同时呈现技术能力快速扩展与临床证据成熟度不足。2024年至2026年,多智能体会诊、交互式虚拟诊室、罕见病智能体和EHR沙箱相继出现;近期范围综述与综合调查仍一致指出,证据主要来自基准测试和受控环境^[10-11]。Agent Clinic显示,静态MedQA成绩与临床表现仅呈弱相关;另一项基准研究发现,智能体相较基座模型的性能增益有限,却消耗超过10倍的令牌并显著延长时延^[12-13]。MIRA在受控EHR沙箱中完成问诊、检查、诊断和治疗全流程,DeepRare则在多源罕见病任务中生成可追溯证据,拓展了可执行能力的边界^[14-15]。然而,这些发现主要来自回顾性数据、模拟患者和结构化环境,不能外推为真实世界自主诊疗的安全性或临床净获益。

由此产生的中心科学问题是:医学智能体能否

在信息不完整、任务连续且风险异质的真实临床环境中,把模型能力转化为可验证的患者与卫生系统净获益,同时维持可控自主性和可追责性?本文以此为主线,依次回答医学智能体是什么、如何工作、能够解决哪些临床问题、应如何评价以及如何安全落地,并以“临床问题—知识缺口—研究发现—机制解释—临床意义”的链条连接技术架构、临床证据与治理责任。

1 概念边界与医学智能体知识体系

1.1 从医学AI到医学智能体:操作性定义与概念边界 医学智能体的可操作性定义如下:围绕明确医疗目标,在限定环境、权限和监督条件下,持续感知患者及医疗情境,整合可追溯知识,开展推理与规划,调用工具或工作流执行行动,并根据结果反馈调整后续行为的AI系统。该定义要求同时具备目标性、状态性、序贯性、工具性、行动性和反馈性。若系统仅生成回答或摘要,不维持任务状态、不形成计划、不能调用外部工具,或不依据结果调整行为,则更宜称为医学LLM应用或对话式决策支持,而非完整智能体。

这一界定将评价单位由“模型”转向“系统”。同一基座模型配置不同知识库、提示模板、工具权限、记忆策略和人工确认规则,可形成风险不同的产品;临床有效性因此是模型、使用者、界面、组织流程和患者反馈共同作用的结果,不能由LLM单项成绩推定^[8-9]。医学智能体还具有技术与制度上的不对称性:技术上虽可主动发起动作,法律和伦理上却不因此取得医疗主体资格,责任仍须落实到可识别的开发者、部署机构和专业人员。

1.2 分类、能力边界与自主性 医学智能体可按临床目标、组织方式、运行环境和行动权限进行分类。目标包括知识服务、诊断、治疗决策、监测预警、连续管理、流程运营和科研智能体;组织方式包括单智能体、编排式多智能体和去中心化多智能体。

表 1 医学智能体与相关系统的边界比较

系统类型	核心功能	是否维持任务状态	是否自主规划/调用工具	典型输出或行动	主要监管关注
判别式医学 AI	分类、检测、预测	否或有限	否	风险分值、病灶框、预警	数据偏倚、泛化、校准
医学 LLM/大多模态模型	内容理解与生成	多为会话内	可选,非必要	问答、摘要、报告草案	幻觉、隐私、知识时效
临床决策支持系统 (CDSS)	规则或模型驱动的提示	有限	通常否	警报、推荐、医嘱提醒	警报疲劳、适用性
医疗副驾驶 (copilot)	协助完成局部任务	有限	在用户指令下调用	文书、检索、编码、草案	人工复核、工作流适配
医学智能体	目标驱动的多步骤任务	是	是	检索、计算、下单草案、协调流程	权限、行动安全、审计
多智能体系统 (MAS)	多角色分工、互审与协同	是,共享或分层	是	MDT 式意见整合、计划执行	错误传播、共识幻觉、责任分配
具身医学智能体	感知并作用于物理世界	是	是	设备控制、机器人操作、家庭干预	实时安全、物理伤害、紧急制动

CDSS: clinical decision support system; LLM: large language model; MAS: multi-agent system; MDT: multidisciplinary team.

环境包括患者端、医务人员端、医院信息系统端、居家端和设备端。对临床风险评估最可信及权威的参数不是智能体数量,而是行动与患者的距离、错误可逆性以及人工复核能否及时发生。

多智能体系统的潜在价值来自认知分工和相互校验,而非角色堆叠。MDAgents 根据任务复杂度动态选择单独或协作模式,在 10 个基准中的 7 个取得最佳表现,较既往方法最多提高 4.2%;加入主持

者复核和外部知识后,平均提高 11.8%^[16]。MedAgents 及罕见病多智能体研究同样观察到,角色化讨论可扩展鉴别诊断并提高一致性^[17-18]。但若各角色共享基座模型、训练数据和错误先验,表面辩论并不构成独立证据;角色增加还会扩大时延及成本。因此,复杂度、异构模型、外部证据和独立安全审查,比简单多数投票更能解释多智能体系统的净增益。

表 2 医学智能体的功能类型、适用场景与能力边界

类型	代表任务	可接受的默认自主性	必要的人类角色	不宜越过的边界
信息与知识智能体	指南检索、证据摘要、患者教育	信息生成	核验来源与情境适配	不把检索摘要直接当医嘱
诊断智能体	问诊、鉴别诊断、检查建议	建议/草案	采集体征、判断优先级、确认诊断	不在信息不全时排除高危病
治疗决策智能体	药物核查、方案比较、剂量计算	受控建议	权衡获益风险与患者偏好	不自主开具高风险处方
监测预警智能体	生理信号监测、恶化识别	自动监测与升级	响应警报、处置异常	不以模型静默替代临床观察
连续管理智能体	随访、康复、依从性支持	低风险任务自动化	设定目标、处理例外	不延误急症转诊
流程运营智能体	编码、排程、资源调度、试验预筛	规则内执行	审批例外、监督公平性	不因效率目标排斥复杂患者
多智能体临床团队	MDT 整合、复杂病例规划	监督式协作	主持、裁决冲突、承担责任	不以“共识”掩盖共同错误
具身智能体	机器人、呼吸机或家庭设备控制	极低至受控执行	实时监督与紧急接管	不在无冗余保护下闭环控制

2 体系架构与关键技术

2.1 “感知—知识—推理—规划—行动—反馈”六模块闭环 要把模型能力转化为可验证的临床行动,医学智能体可概括为“感知—知识—推理—规划—行动—反馈”六模块并形成闭环。感知层接入EHR、影像、病理、检验、生理信号、组学、可穿戴设备和患者报告结局,将多模态数据转换为带时间戳和质量标识的患者状态。数据完整性是首要的:过期用药、复制粘贴错误、单位混乱或缺失阴性信息,均可能产生后续的诊疗方案的混乱和错误。

知识层连接指南、系统综述、药品说明书、机构路径、知识图谱和患者历史。检索增强生成(retrieval-augmented generation, RAG)将参数化记忆与外部证据分离,为内容更新、引用溯源和本地化提供服务^[19]。但RAG本身不保证事实正确:来源质量不足、上下文截断或引文与结论不匹配,仍可产生“有引用支撑的错误”。因此,知识库必须实行版本管理、证据分级、有效期控制和撤回文献处置。

推理层完成问题表征、鉴别诊断、风险估计以及因果和反事实判断。链式推理在部分复杂任务

中可改善系统整体表现,但冗长的“思维链”存在掩盖真实机制和事后合理化的可能。临床可验证性应建立在关键证据、反证、不确定性、适用指南和可复核的中间结果上,而非单纯语言流畅度。

规划层把临床目标分解为有顺序、终止条件和回退策略的子任务。ReAct将推理与行动交替,Toolformer和Reflexion分别探索工具学习与基于反馈的自我改进^[20-22]。在医疗场景中,规划还必须同时考虑成本效益,花费时间和不可遗漏诊断;否则,智能体可能通过无上限检查提高表面准确率,却降低临床净获益。

行动层通过函数接口调用临床计算器、知识检索、影像算法、药物相互作用数据库、FHIR服务或设备。行动必须结构化和具有可执行性;“建议”“生成草案”“提交”和“执行”应对应不同权限。反馈层则判断任务是否完成、患者是否改善、是否发生异常,以及何时重新规划或请求人工接管。模型自检只能构成内部反馈,医师修正、患者结局和运行监测才提供外部约束,防止系统在自洽但错误的轨道上循环。

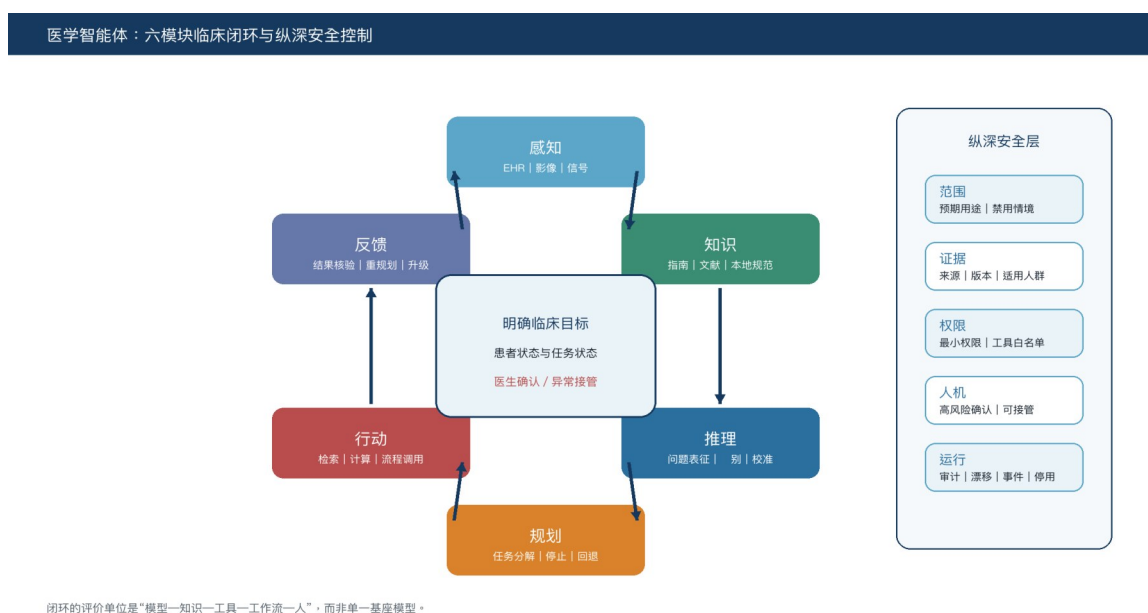


图1 医学智能体临床闭环与纵深安全架构

2.2 知识锚定、记忆、工具使用与互操作 知识锚定旨在把“模型可能知道什么”转化为“系统此刻能够证明什么”。高质量医学RAG至少应包括临床问题结构化、混合检索、重排序、证据去重、时间和人群过滤、冲突检测以及逐句引文核验。检索结果还需标注研究设计、适用人群、证据等级和发布日期,使

智能体区分指南推荐、观察性关联与专家意见。处方、剂量和禁忌证等高风险信息应优先由确定性数据库或规则引擎返回,而非由LLM自由生成。

记忆可分为短期工作记忆、患者纵向记忆、情景记忆和程序记忆,分别保存当前任务状态、跨次随访信息、既往行动及结果。

表 3 六模块闭环的输入、输出、典型失效与安全控制

模块	关键输入/输出	典型失效	优先控制
感知	多模态原始数据→患者状态	缺失、错配、单位/时间错误	数据质量门、身份核对、时效标签
知识	指南/文献/本地规范→证据包	召回偏差、过期知识、伪引用	权威源白名单、版本与证据等级
推理	状态+证据→候选解释与风险	幻觉、锚定、忽略反证	结构化问题表征、反证搜索、校准
规划	临床目标→有序任务图	目标偏移、过度检查、无限循环	成本/风险约束、停止条件、回退计划
行动	计划→工具调用/ workflow 变更	错工具、错参数、越权执行	最小权限、参数校验、关键动作双确认
反馈	工具结果/结局→验证与重规划	自我确认、异常未升级	独立验证器、异常停止、人工接管

长期记忆虽然支持连续照护,却也会累积“记忆污染”:一次错误摘要可能持续改变后续判断。安全机制应包括来源标记、有效期、患者可见与可纠正机制,以及敏感信息最小化存储;未经确认的模型推断不得写入永久病历。

工具调用使语言生成获得可执行性。每个工具均应具有机器可读的用途、参数类型、前置条件、风险等级和失败码;调用前校验身份、权限和参数,

调用后核对患者归属、单位、参考范围和时间。FHIR 与 SMART on FHIR 为 EHR 互操作提供标准化资源和授权框架^[23-24],但语法互通不等于语义一致,本地代码映射、字段缺失和 workflow 差异仍需端到端验证。外部网页、患者上传文件和跨机构消息必须置于非可信数据域,防止提示注入将内容伪装为系统指令。

表 4 关键技术的临床价值、局限及验证重点

技术	临床价值	固有限制	上线前验证重点
RAG	更新知识、提供来源、支持本地规范	召回失败与“有引文的错误”	召回率、引文蕴含、证据时效与冲突
知识图谱/规则引擎	约束关系、药物与禁忌核查	建设维护成本高、覆盖有限	术语映射、规则版本、边界病例
短期/长期记忆	连续管理、减少重复采集	污染、隐私和过时信息	来源、有效期、纠错和删除机制
工具/函数调用	精确计算、查询与流程执行	参数错误、越权、依赖服务故障	接口契约、沙箱、回滚、幂等性
反思/验证器	发现矛盾、提高鲁棒性	同源模型可能共同失误	异构验证、外部标准、停止阈值
FHIR/SMART	EHR 互操作与授权	本地实现和语义差异	端到端映射、权限范围、审计
联邦学习/隐私计算	跨机构协作与数据不出域	异质性、通信成本、泄漏仍可能	攻击测试、站点偏差、可撤回性

2.3 多智能体协作、具身连接与安全编排 多智能体架构通常由编排者、专业子智能体、工具执行者和安全监督者组成。编排者依据病例复杂度决定是否调用影像、病理、药学、指南或伦理角色;各角色以结构化格式提交结论、证据和不确定性;主持智能体处理冲突,独立安全智能体检查禁忌、遗漏和越权。AutoGen 等通用框架显示了对话式编排的可行性^[25],但医疗系统更需要有状态任务图而非无约束自由对话:每个节点都应规定输入、输出、责任、超时和失败处理。

虚拟环境可用于评估完整行动链,却不能消除模拟一现实差距。Agent Hospital 允许智能体通过模拟病例练习和反思^[26],但自生成患者、自生成标签和自我评分可能形成闭环偏差。2026 年,MIRA

在 EHR 沙箱中使用 11 类工具和 8.5 万余项选项,在 574 例急诊病例中取得 88.9% 的平均诊断准确率,并与医生进行诊疗质量对照评价^[14]。该研究表明,标准化全流程评测已可实施;但患者信息来自回顾性病历摘要,环境结构化程度高于真实病例,因此沙箱中的优势不能被表述为临床部署或无监督诊疗的证据。

具身智能体把软件行动扩展到机器人、监护设备、呼吸支持、给药泵和家庭设备,潜在伤害因而更直接且更难逆转。任何可能改变患者身体状态的闭环,都应由确定性控制器承担实时安全控制,LLM 仅负责高层意图解释或方案建议;关键参数须受到约束、双通道监测和独立紧急停止保护。网络中断、模型超时或上下文异常时,系统应回落到经验

证的设备默认模式,而不能继续推测性操作。

表5 单智能体、多智能体与具身智能体的架构取舍

架构	主要优势	新增风险	适宜任务	核心安全机制
单智能体+工具	结构简单、成本和时延较低	单点认知偏差、验证不足	文书、检索、有限流程	外部验证器、人工复核
编排式 MAS	专业分工、互审、按复杂度扩展	错误传播、协作成本、责任模糊	MDT、罕见病、复杂计划	角色独立、证据强制、主持裁决
去中心化 MAS	弹性强、可并行	行为不可预测、难审计	低风险科研探索	通信协议、预算和终止约束
EHR 沙箱智能体	可评估完整序贯流程	模拟—现实差距、标签泄漏	上线前压力测试	隔离环境、标准接口、对抗情境
具身智能体	连接感知与物理干预	实时物理伤害、网络/设备故障	康复、物流、受控设备辅助	硬约束、冗余传感、急停与接管

3 临床转化:从局部辅助到连续照护

3.1 筛查、分诊、诊断与治疗决策 筛查和分诊适合作为较早的临床入口,因为任务目标相对明确,可结合结构化问卷和规则,并把高风险结果直接升级人工。智能体可主动追问患者症状,整合年龄、暴露、共病和生命体征,调用风险计算器并推荐服务层级。其临床价值不应只用分类准确率衡量,还应检验是否缩短就诊路径、减少重复采集并降低时间敏感疾病漏分诊;同时报告低危人群不必要升级和高危人群漏分诊,并按语言、年龄、性别、地区和数字素养分层。

诊断任务更能检验智能体是否具备临床序贯能力,因为系统必须决定“下一步还需要知道什么”,而不是在信息完整的病例上作答。AgentClinic 允许医生智能体在 20 轮内与患者、测量和影像工具交互;在 MIMIC-IV 子集中,表现最佳的模型准确率仍仅为 42.9%。在 AgentClinic-MedQA 中,将交互轮数由 20 轮降至 10 轮时,诊断准确率由 52% 降至 25%^[12]。这说明信息采集策略和交互预算本身会影响诊断上限,也说明静态题库高分不能证明问诊充分、检查合理或资源使用适当。

对话式诊断 AI AMIE 在标准化模拟会诊中表现出较高的病史采集、鉴别诊断和沟通能力^[27-28]; Healthcare Agent 也在受控医疗咨询任务中验证了主动追问和工具调用增强通用 LLM 的可行性^[29]。DeepRare 把表型、基因数据和外部知识分配给专业工具,在 6 401 例、2 919 种罕见病的评测中,HPO 输入的 Recall@1 为 57.18%,新华医院外显子病例的多模态 Recall@1 为 69.1%,专家对推理链有效性与可

追溯性的同意率为 95.4%^[15]。这些结果支持“对话+工具+可追溯证据”这一技术路径,但主要结局仍是诊断排序,不能证明其可缩短真实诊断历程或改善治疗结局。

治疗决策比诊断排序更接近不可逆行动,因而需要更严格的证据和权限约束。智能体可以比较指南路径,核对剂量、过敏、肝肾功能和药物相互作用,并生成供医生与患者共同决策的方案;在病历、体格检查或偏好信息不完整时,不应直接生成可执行处方。围术期和危重症高度时间依赖,更适合从资料汇总、风险预警、清单核对和团队协调切入,而不宜早期开放自主治疗。高风险场景中,及时、可解释地请求接管,往往比提高自动化覆盖率更具有临床价值。

3.2 慢病连续管理、患者端与居家医疗 慢病照护的核心缺口通常不是缺少一次正确建议,而是院外状态不可见、依从性随时间下降和异常升级不及时。智能体可将间断随访转化为“连续感知—风险分层—个体化提醒—专业复核—疗效反馈”:整合可穿戴设备、家庭监测和患者报告结局,识别趋势而非孤立异常信号;按疾病路径安排随访、复查和教育;达到预设阈值时联系医护团队。COPD、哮喘、糖尿病、心衰、肿瘤随访和康复均有潜在应用,但临床净获益取决于能否降低急性事件而不增加警报负担、依赖和不公平。

患者端智能体可承担知识翻译、症状记录、用药提醒和共同决策准备,从而改善信息可及性;同样的语言流畅性、同理性和即时响应也可能形成“拟人化权威”,使患者高估系统能力,尤其是在精神健康、儿童、老年和低健康素养人群中。

表 6 代表性医学智能体研究及其可转化含义

研究/年份	场景与设计	关键结果	可支持的结论	不能外推的结论
MDAgents, 2024 ^[14]	10 个医学知识/诊断基准	7/10 最佳;最多提高 4.2%	复杂度自适应协作可提高部分任务性能	不能证明真实 MDT 效能
AMIE, 2025 ^[17-18]	标准化对话与复杂病例鉴别诊断	多项对话/鉴别指标优于对照	LLM 可优化问诊和诊断对话	不能替代真实体检与临床结局试验
MAC, 2025 ^[19]	302 例罕见病,多医生+监督者	四医生+监督者表现最佳	角色协作适合跨学科病例	同源智能体共识不等于真值
DeepRare, 2026 ^[13]	6401 例、表型/外显子/知识检索	多模态 Recall@1 69.1%;证据一致 95.4%	工具化、可追溯推理适合罕见病	尚不能证明缩短真实诊断时间
AgentClinic, 2026 ^[10]	多语种、多模态、序贯模拟	静态成绩仅弱预测交互表现	评价必须覆盖信息采集与工具使用	模拟患者不能替代真实患者
MIRA, 2026 ^[12]	>500 例 EHR 沙箱急诊全流程	574 例平均诊断准确率 88.9%	可在标准化沙箱评估完整行动链	不能据此批准无监督急诊诊疗

界面应明确系统身份、能力边界和紧急联系方式;高风险建议必须显示依据和升级条件,不能以笼统免责声明替代安全设计。世界卫生组织(WHO)关于大多模态模型的指导强调透明、问责、包容和部署后审计^[7,30]。

医院到家(hospital at home, HaH)把连续监测与真实照护流程连接,是检验智能体能否产生系统净获益的重要场景。白春学教授团队提出面向中

国的医院—社区—家庭一体化呼吸居家住院路径,强调同时评价临床安全、环境足迹、社会效益和经济价值^[31]。智能体更适合作为受约束的任务协调层,核对纳入标准、汇总家庭生命体征、提示护理任务并触发升级;家庭环境、照护者能力、网络稳定性和设备维护仍共同决定风险。对于数字化程度低或居住条件不足者,必须保留电话、社区上门和线下通道,避免把技术可用性转化为新的准入门槛。

表 7 连续照护场景的任务闭环、临床终点与升级触发

场景	智能体闭环任务	优先临床终点	必须升级人工的触发
COPD/哮喘	症状与吸入药监测、恶化识别、随访协调	急性加重、住院、吸入技术与生活质量	静息呼吸困难、低氧、救援药快速增加
糖尿病	血糖趋势、用药/饮食支持、低血糖预警	严重低血糖、HbA1c、依从性	意识改变、反复低血糖、酮症风险
心衰	体重/症状/生命体征趋势、利尿方案草案	再住院、容量状态、肾功能	低血压、胸痛、快速体重增加、肾损伤
肿瘤随访	症状与不良反应分级、复查安排	严重不良事件、剂量延误、患者体验	发热性中性粒细胞减少、出血、严重疼痛
居家住院	纳入核对、远程查房摘要、护理协调	转院、并发症、住院日、照护负担	生命体征恶化、家庭支持失败、设备故障
康复/老年	训练计划、跌倒/谵妄风险、照护者支持	功能、跌倒、再入院、照护者负担	新发神经症状、跌倒损伤、急性认知改变

3.3 医院运营、公共卫生与临床研究 医院运营智能体可处理病历质控、编码、床位匹配、转运、耗材和排班等任务。由于数据更结构化、直接临床风险相对较低,这些场景可作为较早的规模化入口;但价值必须以端到端流程而非局部准确率来衡量。若自动草案增加医师编辑时间,或在下游造成拒付、重复检查和沟通返工,则局部性能并未转化为

系统收益。优化目标也不能仅设为单位时间的工作量,否则复杂、高龄、残障或语言少数患者可能被算法性地视为“低效率对象”。

生成式 AI 文书和患者信息草拟是目前最接近真实工作流的证据领域。较早的盲评研究提示, AI 回复在质量和同理性评分方面具有潜力,但公共平台推理不能替代临床安全评价^[32]。临床摘要安全

框架在 12 999 个医师标注句子中观察到 1.47% 的幻觉率和 3.45% 的遗漏率,并通过流程优化降低严重错误^[33]。一项多中心质量改进研究^[34]中,263 名门诊医师及高级执业医务人员使用环境式 AI 文书 30 日后,自报倦怠比例由 51.9% 降至 38.8%;这一观察性结果受自愿参与、选择性使用、短期随访和缺少对照组等限制,不能单独归因于 AI。医师访谈仍报告文书过长、风格不合和编辑负担^[35-36]。系统应保留从音频或转录到最终病历的可追溯差异,并将医师修改用于质量监测,而不能默认把核查责任转移

给使用者。

临床试验智能体可解析方案、预筛 EHR、解释入排标准并安排随访联系。随机研究显示,LLM 辅助预筛可提高部分任务效率^[37];但时间性条件、隐含临床判断和招募公平仍需研究者确认。公共卫生智能体可聚合区域数据、发现异常信号并生成情景方案,却必须防范数据漂移、群体标签污名化和未经授权的个体再识别。科研智能体可支持文献检索、方案草拟、代码检查和数据字典生成,但研究问题、统计计划、证据解释和署名责任不能外包。

表 8 非直接诊疗场景的价值指标与治理重点

场景	主要价值指标	易被忽略的下游代价	治理重点
环境式文书	文书时间、下班后 EHR 时间、完整性	编辑返工、遗漏、患者知情与录音隐私	草案标识、差异审计、保留人工签署
病历质控/编码	编码准确、拒付率、质控时长	过度编码、错误传播至科研与支付	规则复核、抽样审计、申诉机制
床位/资源调度	等待时间、周转率、取消率	对复杂患者的不公平延迟	公平约束、例外通道、人工调度权
临床试验预筛	筛查时间、召回率、入组率	隐含标准误判、少数群体漏筛	时间逻辑验证、研究者确认、分层监测
公共卫生监测	信号提前量、阳性预测值、响应时间	群体污名、误警报与资源挤占	聚合最小化、阈值审议、跨源验证
科研智能体	检索覆盖、可重复性、协议完整性	虚假引文、分析路径依赖	人工方法学负责、代码/来源存档

4 评价框架与临床证据路径

4.1 从模型准确率到“人—智能体— workflow”系统评价 医学智能体的评价对象应是“人—智能体— workflow”联合系统,而非孤立模型。除医学知识准确率外,技术层还应测量计划成功率、工具选择和参数正确率、任务完成率、错误恢复、时延、令牌或能源成本以及跨版本稳定性。诊断任务应报告校准、鉴别诊断覆盖和检查利用;行动任务应按严重程度统计错误行动、近失事件和被安全机制阻断的事件。只报告最终答案会掩盖由错误过程偶然得到正确结果的风险。

临床安全评价应以潜在伤害而非表面文本差异组织错误分类。虚构、遗漏、矛盾、过期信息和不恰当确定性的后果不同,同一错误出现在患者教育与高风险给药中也具有不同严重度^[33,38-40]。研究应预先定义严重错误、可恢复错误、人工接管和停止条件,并由独立临床专家盲评。由 LLM 充当的自动评估者可用于大规模初筛,但不能作为唯一金标准,因为评估模型与被评模型可能共享偏差、语料

和失效模式。

人机协作评价的关键是信任校准,而非总体满意度。Goh 等^[41]纳入 50 名医师的随机试验发现,开放 GPT-4 辅助并未显著改善医师诊断推理,尽管模型单独表现较强;另一项横断面研究^[42]也观察到生成式模型在标准化临床推理任务中的高分。二者共同说明,模型能力不会自动转化为团队能力,界面呈现、使用者训练、时间压力 and 责任感均可能调节效果。研究应比较医生单独、智能体单独、医生+智能体和真实标准流程,并记录建议采纳、错误建议纠正及人工新增价值。

患者和卫生系统终点应由预期用途决定,包括死亡、并发症、诊断延误、症状控制、生活质量、患者体验、可及性、成本和机会成本。过程指标可支持早期迭代,却不能长期替代患者结局。公平性也必须贯穿各层:除模型误差外,还应检验数字接入、人工升级、服务等待和最终结局是否在群体之间产生系统性差异。只有当临床获益超过伤害、负担和机会成本时,模型性能才转化为临床净获益。

表 9 医学智能体多维评价核心指标集

维度	核心指标	智能体特异指标	建议分层
技术性能	准确性、校准、鲁棒性、时延	任务完成、工具/参数正确、计划长度、恢复率	模型版本、场景复杂度、机构
临床安全	严重错误、潜在伤害、漏诊/延误	错误行动、越权、被阻断事件、接管率	风险级别、疾病严重度、时间窗口
人机协作	工作负荷、决策时间、可用性	采纳/纠错率、信任校准、自动化偏倚	专业、年资、培训、时间压力
患者结局	死亡、并发症、症状、生活质量	因智能体升级/延误导致的结局	年龄、性别、语言、共病
卫生系统	成本、资源利用、可及性、扩展性	每成功任务成本、下游返工、能耗	城乡、机构等级、支付方式
公平与伦理	群体性能差异、隐私事件、申诉	升级机会、拒绝/退出、解释可获得性	受保护属性与交叉群体

4.2 六阶段转化证据路径与报告规范 围绕这一转化逻辑,智能体证据可分为6个递进阶段。第一阶段以静态单任务测试筛选基础能力不足;第二阶段进入虚拟病例、数字患者和对抗模拟,评价序贯决策、工具调用和极端情境;第三阶段开展回顾性多中心验证,锁定版本并检验外部泛化;第四阶段静默部署,系统运行但不影响照护,用于测量数据可得性、时延、漂移和潜在警报;第五阶段开展前瞻性临床研究或随机试验,评价人机团队和临床结局;第六阶段才进入规模化应用与持续监测。模型、知识库、提示、工具或工作流发生实质更新时,应根据影响返回相应阶段。

报告规范应按研究问题组合使用,不能被视为

临床放行凭证。CONSORT-AI和SPIRIT-AI分别扩展AI干预试验报告与方案^[43-44];DECIDE-AI聚焦早期真实临床环境中的人机决策支持^[45];TRIPOD+AI和PROBAST+AI分别用于预测模型报告和偏倚风险评估^[46-47];STARD-AI覆盖AI诊断准确性研究,并新增数据集实践、算法偏倚和公平性项目^[48];CLAIM适用于医学影像AI,MI-CLAIM规定临床AI模型的最小报告信息^[49-50];CHART规范健康建议聊天机器人研究^[51]。这些工具主要改善报告完整性或研究质量评估,不能替代产品验证和监管审批;医学智能体还需报告计划、工具、权限、记忆、人工接管、成本和版本依赖。



图 2 医学智能体从静态测试到持续监测的六阶段证据路径

4.3 人因、实施科学与卫生经济学 医学智能体属于复杂干预,其效果由技术、人员、组织和政策共同产生。部署前应绘制现有与目标工作流,明确谁在何时看到何种信息、可以采取何种行动以及异常由谁接收并处理。若系统要求医师在忙碌节点重复录入数据或逐句核对冗长文本,即使离线性能优良,也可能增加负担并被弃用。实施评价可借鉴社

会技术与实施科学框架,持续观察技术、任务、人员、组织和外部环境之间的相互作用^[52-54]。

人因设计需重点防范3种机制:自动化偏倚,即使用者在时间压力下接受错误建议;技能退化,即长期依赖系统后独立判断能力下降;责任稀释,即参与者均假定他人已完成核查。可通过先独立判断后显示建议、突出反证和不确定性等。

表 10 医学智能体六阶段临床转化与放行标准

阶段	主要目的	关键设计	最低放行标准
1 静态任务	验证基础能力	隔离测试集、困难/反事实病例	达到预设性能并完成错误分类
2 模拟与沙箱	验证序贯行动	虚拟患者、工具故障、提示攻击	高风险动作可阻断,循环可终止
3 回顾性外部验证	检验泛化	多中心、时间外、群体分层	校准与公平达标,版本锁定
4 静默部署	验证 workflow 可行性	实时数据、不呈现建议	数据覆盖、时延、漂移与负担可接受
5 前瞻性/随机研究	证明临床净获益	人机系统对照、患者/过程终点	获益超过伤害与机会成本
6 规模化与持续监测	维持真实世界效能	注册、审计、漂移与事件监测	性能稳定、事件可处置、更新受控

经典自动化分级和自动化偏倚研究,以及最新人机协作综述,均支持依据任务风险设计人工介入时点和否决权^[55-57]。解释应服务于具体决策;未经验证的机制解释可能增加而非减少不当信任^[58-59]。

卫生经济评价应覆盖完整路径,而不止采购价格,包括模型调用、算力、接口、知识维护、人工复

核、网络安全、培训、事件处理和下游资源利用。单位任务成本必须与成功完成率、患者结局及替代方案共同比较。若智能体通过增加检查提高诊断准确率,或仅把文书负担从医生转移给护士和患者,就不能推断效率改善。

表 11 部署前人因与经济性评价问题

领域	必答问题	可量化指标	不合格信号
工作流适配	是否减少真实端到端步骤?	点击/录入次数、总周转时间、返工率	局部节时但下游返工增加
信任校准	使用者何时接受或拒绝建议?	正确/错误建议采纳率、纠错率	对错误建议高采纳或对正确预警忽视
人工监督	复核者是否有时间和信息?	每例复核时长、接管成功率	“名义人工在环”但无法有效核查
培训与技能	能否独立工作并识别失效?	情景考核、无 AI 任务表现	技能退化、无法解释系统边界
总成本	是否计入维护和下游利用?	每成功任务/每 QALY 成本、预算影响	仅报告模型调用或人力节省
分配影响	获益与负担由谁承担?	群体净获益、等待时间、数字接入	效率提升伴随弱势群体可及性下降

5 可信应用:风险、自主性与全生命周期治理

5.1 风险谱:从幻觉到错误行动和系统性伤害 医学智能体风险沿“输入—认知—计划—工具—交互—系统”链条传播。需要注意的是它既继承了语言模型的虚构、歧视和滥用风险^[60-61],又因可规划和可行动而增加级联失效:输入侧包括病历错配、缺失、复制错误和对抗内容;认知侧包括幻觉、过度确定、锚定和知识过期;计划侧包括目标偏移、过度检查、遗漏紧急任务和无法终止;工具侧包括错选接口、参数或患者、越权写入和外部服务故障;交互侧包括自动化偏倚、患者过度依赖和责任稀释;系统侧包括模型更新、漂移、供应商中断和跨机构级联。不同于普通聊天机器人,智能体可把上游错误经后续动作固化,这类非预期后果早已在临床机器学习中受到警示^[62]。

多智能体协作也不会天然消除幻觉。若多个角色共享同一模型或知识源,一个错误可能被复述、投票并包装为“专家共识”;沟通链越长,来源和不确定性越容易丢失。2026 年一项智能体基准^[13]

显示,内部防护虽过滤了 89.9% 的幻觉,仍有部分幻觉未被阻断,且有限的性能增益伴随超过 10 倍的令牌消耗和超过 2 倍的时延。因此,安全评价应同时统计原始错误、被阻断错误和影响最终行动的错误,而不能只检查用户最终看到的文本。

提示注入在能够读取外部内容并调用工具的系统上尤为危险。恶意指令可嵌入网页、附件、影像报告或患者消息,诱导智能体泄露信息或越权执行。外部内容应始终被视为数据而非指令,系统提示、患者数据和工具返回需置于不同信任域,高风险工具不得把自由文本直接映射为执行参数。长期记忆同样可能被篡改,因此任何写入都应保留来源、审核状态和可撤销机制。

公平性风险也不限于训练数据代表性。历史 EHR 记录了既有诊疗差异,资源调度目标可能把过去的固化为未来规则^[63-65];语言能力、设备接入、数字素养和人工升级机会还会影响真实结局。评价必须区分统计公平、临床公平与程序公平:群体平均性能相近,并不意味着个体获得同等解释、申诉和人工照护。

表 12 医学智能体风险—控制矩阵

风险域	代表失效	潜在临床后果	预防控制	检测/缓解
数据与身份	患者错配、过期/缺失信息	错诊、错药、延误	主索引核对、时间与质量标签	异常一致性检查、人工核验
知识与推理	幻觉、伪引用、锚定	不当建议、遗漏高危鉴别	权威 RAG、反证与禁忌规则	引文蕴含检查、独立临床审阅
计划与目标	目标漂移、过度检查、循环	伤害、资源浪费、延误	预设目标/预算/停止条件	计划监控、超时和回退
工具与权限	错参数、越权写入/执行	错误医嘱或物理行动	最小权限、白名单、沙箱	双确认、回滚、审计告警
多智能体	共识幻觉、责任稀释	系统性错误传播	异构角色、证据强制、职责分离	独立安全智能体、冲突升级
网络与供应链	提示注入、服务/版本突变	泄露、行为改变、停摆	信任域隔离、供应商评估	红队测试、版本监控、降级模式
人与组织	自动化偏倚、技能退化	错误采纳、接管失败	界面与培训、独立首判	采纳/纠错监测、情景演练
公平与隐私	群体偏差、再识别	医疗差异、信任受损	数据最小化、代表性设计	分层审计、申诉与补救

5.2 自主性分级与最低安全要求 医学智能体的自主性应按行动权限分级,而非依据产品名称。本文建议四级:1级信息助手仅提供知识和教育;2级任务助手可整理病历、计算风险和生成草案,但关键结果由专业人员确认;3级工作流智能体可在预设权限内执行多步骤任务,高风险节点必须批准;4级为监督式多智能体或具身系统,可协调复杂流程,但医生持续监督并承担最终医疗责任。

自主级别应与任务风险、证据强度和监督能力同步匹配。同一功能在不同环境中的风险并不相同:药物相互作用查询属于信息任务,自动修改处

方则属于高风险行动;家庭提醒漏发与ICU给药参数错误的后果也不同。部署机构应建立逐项权限清单,明确读取、建议、生成草案、提交和执行能力,并默认关闭未经验证的权限。

所有级别均须满足最低安全基线:明确预期用途和禁用情境;实施数据与工具最小权限;保证来源可追溯;表达不确定性和信息缺口;对高风险动作进行人工确认;支持异常自动停止和人工接管;保留全过程审计;追踪版本、配置和知识库;报告安全事件;保障患者知情、退出和申诉。

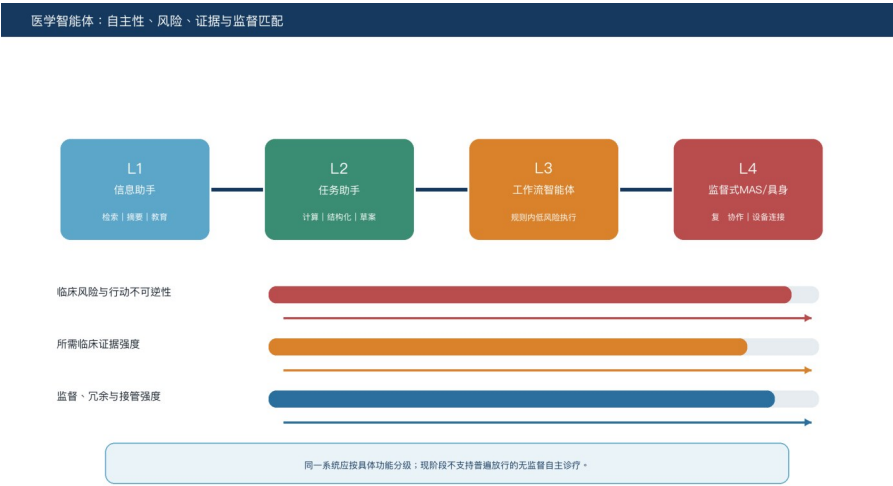


图 3 医学智能体自主性、风险、证据与监督的匹配关系

5.3 监管边界与全生命周期治理 监管边界取决于预期用途、目标使用者、输出是否可独立审查以及是否直接影响诊疗。负责任的临床转化应遵循“不

得造成伤害”的基本路线图,并将 AI 风险管理、医疗器械风险管理和软件生命周期过程纳入同一质量管理体系^[66-69]。

表 13 风险校准的四级医学智能体自主性

级别	能力与权限	例示	必要监督	建议证据门槛
L1 信息助手	读取、检索、解释;不改变流程	指南摘要、患者教育	使用者核验来源	静态+模拟安全测试
L2 任务助手	结构化整理、计算、生成草案	病历摘要、风险评分、处方核查草案	每例专业确认	外部验证+静默部署
L3 工作流智能体	规则内多步骤提交/低风险执行	随访安排、试验预筛、检查草案	高风险节点批准、持续监测	前瞻性研究,必要时随机试验
L4 监督式 MAS/具身系统	复杂流程协作或受控物理行动	MDT编排、居家闭环、机器人辅助	实时监督、冗余保护、紧急接管	临床结局试验+强化监管

L: level; MAS: multi-agent system。自主程度越高、行动越接近患者且越不可逆,所需证据和监督强度越高。

美国食品药品监督管理局(FDA)2026年临床决策支持软件指南区分非器械CDS与器械软件;软件面向患者、直接驱动诊疗,或专业人员不能独立审查推荐依据时,更可能落入器械监管范畴^[70]。2025年AI器械预定变更控制计划(predetermined change control plan, PCCP)指南要求预先描述计划修改、实施控制和影响评估^[71-72];国际透明度原则强调向使用者说明预期用途、性能、局限和更新^[73]。这些要求对频繁更新模型、提示和知识库的智能体尤为关键。

IMDRF 2025年良好机器学习规范强调多学科协作、代表性数据、独立测试、人机团队、清晰用户信息和部署后监测^[74]。欧盟《人工智能法案》将符合条件的医疗器械AI纳入高风险体系,要求风险管理、数据治理、技术文档、日志、透明度、人类监督和稳健性^[75]。中国《人工智能医疗器械注册审查指导原则》覆盖需求分析、数据、算法、验证和更新控制^[76];《关于促进和规范“人工智能+医疗卫生”应用发展的实施意见》提出全链条连续智能服务与安全可控^[77],生成式AI还须满足数据、内容、个人信息和算法治理要求^[78]。

由于智能体行为由多个持续变化的组件共同决定,治理不能止于一次性审批。治理对象应覆盖基座模型、系统提示、RAG语料、工具、权限、记忆、用户界面和人工流程。更新应按对临床行为的影响分级:不影响临床行为的小修订可在回归测试后发布;改变模型、知识范围、工具权限或预期用途的实质更新,应重新进行风险分析和相应临床评价。机构需维护版本清单、模型或系统卡、数据说明、风险管理文件、变更日志、事件响应和退役计划。FUTURE-AI的公平、普适、可追溯、可用、稳健和可解释六项原则可作为全生命周期组织框架^[52]。

5.4 实施就绪、可及性与责任闭环 责任闭环必须

明确6个问题:谁设定任务目标,谁批准数据和知识源,谁配置权限,谁确认关键动作,谁持续监测,谁处理异常并承担最终医疗责任。医院宜建立由临床、护理、药学、信息、工程、伦理、法务、安全、患者代表和管理者组成的管理委员会;高风险系统还应设独立安全负责人并赋予暂停部署权限。供应商合同需规定数据用途、模型更新通知、服务可用性、审计权、事件协作以及退出时的数据转移。

本地验证不可由外部基准替代,因为疾病谱、编码、设备、语言和工作流差异均可能改变性能。验证集应覆盖常见病例、高风险少见病例、边界输入、缺失数据、跨语言、网络中断和工具故障;上线初期宜采用影子或静默模式和选择有限科室。运行中应监测输入漂移、输出分布、接管率、严重错误、群体差异和用户变通行为,并预先设定暂停与回滚阈值。

可及性是临床有效性的组成部分,而非附加伦理指标。低资源机构可能缺少结构化EHR、稳定网络和维护人员,直接复制高资源方案会形成脆弱系统。更可行的路径是优先部署边界清晰、可靠且可测的小任务,采用开放标准、离线或低带宽降级、本地语言和人工替代通道。患者应知晓AI参与程度、数据用途、请求人工服务和纠错的方式;拒绝或退出AI不应降低其获得基本医疗的机会。

6 未来方向:从智能工具到可学习的人机协同医疗

6.1 连续、多智能体与具身化发展 未来智能体将由单次任务走向跨时间、跨场所的连续管理,由单模型走向按复杂度动态组建的专业团队,并与可穿戴设备、家庭传感器和机器人连接。关键进步不是增加对话轮数,而是形成可验证的执行计划:每项行动可改变患者或流程状态,后续决策依据真实反馈。

表 14 主要共识、报告规范与监管文件的适用范围

文件	年份	主要适用对象	对医学智能体的关键增量
WHO LMM 伦理与治理指南 ^[7]	2024	大多模态模型开发与使用	透明、问责、利益相关者参与、部署后审计
FUTURE-AI ^[41]	2025	可部署、可信医疗 AI 全生命周期	六原则与开发—部署—监测建议
DECIDE-AI ^[42]	2022	早期真实临床 AI 决策支持	人机交互、工作流、错误和实施情境
CONSORT-AI/SPIRIT-AI ^[43-44]	2020	AI 临床试验与方案	干预版本、输入输出、使用者和错误处理
TRIPOD+AI/PROBAST+AI ^[45-46]	2024/2025	预测模型报告与偏倚	数据、性能、公平、质量与适用性
STARD-AI ^[47]	2025	AI 诊断准确性研究	数据集、算法、偏倚、公平与可推广性
FDA CDS/PCCP ^[51-52]	2025/2026	CDS 及 AI 器械软件	可独立审查、预定更新、全产品生命周期
IMDRF GMLP ^[53]	2025	机器学习医疗器械	人机团队、独立测试、监测与再训练风险
EU AI Act ^[54]	2024	高风险 AI/通用 AI	风险管理、日志、人类监督、稳健和透明
中国 CMDE/卫健委文件 ^[55-56]	2022/2025	AI 医疗器械与医疗卫生应用	注册验证、场景落地、连续服务与安全可控

表 15 医疗机构部署就绪清单

领域	上线前最低要求	持续监测	暂停/回滚触发示例
临床用途	预期用途、禁用情境、责任人明确	适用外使用率、净获益	用途漂移或严重漏诊
数据/知识	本地数据映射、权威源、版本化	缺失率、时效、召回与引用质量	指南撤回、数据字段突变
模型/工具	锁定版本、权限清单、沙箱验证	时延、失败、越权与漂移	未通知更新、工具错误行动
人因/流程	培训、有效复核、接管和降级	采纳/纠错、工作负担、变通行为	无法复核、接管失败
安全/隐私	威胁建模、最小权限、事件预案	提示攻击、泄露、异常访问	数据泄露或密钥/权限失控
公平/可及	分层验证、替代通道、患者告知	群体性能、等待和退出情况	持续群体伤害差异
变更/退役	PCCP 或内部变更分级、备份方案	回归测试、事件和供应商状态	重大变更未再验证、服务不可用

2026 年提出的临床环境模拟器框架可用于在大规模部署前研究级联后果,但仍须以真实世界数据校准,且不能把模拟器自身规则当作医学真值^[79]。

多智能体研究需要从同质角色扮演转向异构协作。未来“数字 MDT”应组合不同训练来源的模型、确定性工具、知识图谱和人类专家,并按病例复杂度启动必要角色;关键结论必须绑定证据,冲突不能通过简单平均消失。罕见病、复杂肿瘤和药物安全可能较早受益于跨领域检索与工具协调,但能否改善患者结局并抵消时延和成本,仍缺乏直接证据。

LLM 不适合直接承担毫秒级闭环控制;较合理的架构是“智能体规划—经验证控制器执行—独立安全层监护—人类可接管”。家庭环境还需考虑误用、儿童接触、照护者疲劳和网络故障。把自主性限定于可逆、低风险且可监测的动作,可能比追求全面自主更早产生可验证净获益。

6.2 专病智能体、中国路径与知识体系建设 通用模型提供语言、推理和工具编排能力,专病智能体则把高质量知识、患者路径、专科经验和本地资源嵌入边界明确的任务。白春学教授团队从医学物联网和元宇宙医学提出“全面感知、可靠传输、智能处理”及“云一端”协同框架^[80-81]。

表 16 2026—2030 年医学智能体优先研究议程

方向	关键科学问题	推荐研究设计	成功判据
交互式诊断	能否在不完整信息下高效采集证据?	真实患者前瞻性、多中心比较	诊断、检查利用与时间共同改善
连续管理	长期记忆是否改善结局且不污染?	务实性随机试验、长期随访	急性事件下降、负担和公平可接受
异构 MAS	分工/互审何时优于单智能体?	因子试验、消融与成本分析	净增益超过时延、算力和新风险
行动安全	如何预测级联错误和近失事件?	EHR 沙箱、故障注入、红队测试	高危错误被阻断且接管成功
具身连接	高层规划如何与确定性控制分离?	硬件在环、分阶段临床试验	物理安全、实时性和降级可靠
人机团队	如何训练并保持信任校准?	界面/培训随机试验	错误建议采纳下降、正确纠错上升
生命周期学习	更新如何持续改进而不产生漂移?	注册队列、PCCP、连续监测	更新后效能不降且群体差异不扩大

在结构上与医学智能体的感知、通信、推理和行动闭环一脉相承。医学 GPT 研发的“四改”——由数据清洗转向数据精选、由简单咨询转向专家数字分身、由盲目相信转向质控核对、由单纯循证转向循证与临床经验融合——进一步强调知识质量和临床核验^[82-83]。

肺结节 BAIMGPT 白皮书及肺结节智能体进展将上述框架用于筛查、风险分层、路径推荐和长期随访^[84-85]。其临床意义不在于以 LLM 替代影像医生,而在于把结节影像、个体风险、指南、MDT 意见和随访执行连接为闭环,使管理对象从“单次影像中的结节”转向“患者未来肺癌风险”。下一阶段仍需通过多中心外部验证、前瞻性人机对照、患者结

局和卫生经济评价补足证据,并按国际报告规范透明呈现模型、数据、人工角色和错误。

中国可依托区域医疗中心、互联网医院、家庭医生、社区服务和医疗物联网,构建医院—社区—家庭协同智能体。国家卫生健康委等部门发布的 AI 医疗卫生应用文件覆盖医疗服务、基层公共卫生、健康产业和教学科研^[77,86],为场景设置提供政策基础。实施应避免“大模型先行、用途后找”,优先选择疾病负担明确、流程可测且风险可控的专病闭环;以国家和行业标准统一术语、接口、数据质量、评价指标和事件报告;建设覆盖中文、多方言和基层人群的数据资源;并以区域共享安全验证平台降低中小机构的准入成本。

表 17 中国专病医学智能体的建议研发—转化路径

环节	核心任务	建议产物	关键证据
临床定义	明确患者、场景、目标、禁用情境	预期用途与临床路径图	未满足需求、风险—获益假设
知识与数据精选	指南、病例、专家经验和本地规范治理	版本化知识库、数据说明书	代表性、来源、证据等级与时效
智能体构建	六模块闭环、工具和权限设计	系统卡、工具清单、审计结构	单元、对抗、故障和安全测试
质控核对	自动验证+专家复核+患者反馈	错误分类、红队集、纠错机制	严重错误率、阻断与接管能力
临床验证	多中心、静默、前瞻性和随机研究	注册方案、透明报告	人机团队、患者结局和公平性
区域实施	医院—社区—家庭协同	服务规范、培训与替代通道	可及性、成本、资源与环境影响
持续治理	版本、漂移、事件、更新与退役	监测面板、PCCP/变更计划	真实世界净获益持续稳定

7 结 论

医学智能体推动医疗 AI 由生成内容走向完成目标、由单点模型走向社会技术系统。本文的中心科学问题不是系统能否表现得像医生,而是能否在信息不完整、任务连续且风险异质的真实临床环境中,把模型能力转化为可验证的患者与卫生系统净获益,同时保持自主性可控、过程可追溯和责任可落实。2024 年至 2026 年的研究已在受控环境中显示多智能体协作、交互式问诊、可追溯罕见病推理和 EHR 沙箱全流程的技术可行性;同一批证据也表

明,静态成绩可能与临床能力脱节,工具增强带来的收益伴随高成本,模拟结果难以外推,语言错误还可能经计划和工具转化为行动伤害。

因此,可信转化应遵循三项相互依赖的原则。第一,以“医生—智能体—工作流”联合系统为评价对象,以临床净获益而非模型准确率作为最终标准;第二,使自主性、证据强度和监督能力随行动风险同步升级,对高风险任务保留有效的人工控制;第三,以全生命周期治理覆盖模型、知识、工具、记忆、界面和组织流程。医学智能体的成熟不以减少医生参与为标志,而在于使智能体承担可验证的数

据整合、知识检索、任务协调和连续反馈,让医生把注意力集中于价值判断、伦理权衡、复杂情境和患者沟通上。当前更可持续的方向,是医生主导、患者参与、智能体协作、证据约束且制度上可问责的连续、精准、安全的医疗。

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 宋元林:素材整理、稿件撰写;白春学:设计研究框架、文章审核。

参考文献

- [1] Topol E J. High-performance medicine: the convergence of human and artificial intelligence[J]. *Nat Med*, 2019, 25(1): 44–56.
- [2] Rajkomar A, Dean J, Kohane I. Machine learning in medicine [J]. *N Engl J Med*, 2019, 380(14): 1347–1358.
- [3] Moor M, Banerjee O, Abad Z S H, et al. Foundation models for generalist medical artificial intelligence[J]. *Nature*, 2023, 616(7956): 259–265.
- [4] Thirunavukarasu A J, Ting D S J, Elangovan K, et al. Large language models in medicine[J]. *Nat Med*, 2023, 29: 1930–1940.
- [5] Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine [J]. *N Engl J Med*, 2023, 388(13): 1233–1239.
- [6] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge[J]. *Nature*, 2023, 620(7972): 172–180.
- [7] World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models [M]. Geneva: World Health Organization, 2024.
- [8] Qiu J N, Lam K, Li G H, et al. LLM-based agentic systems in medicine and healthcare[J]. *Nat Mach Intell*, 2024, 6(12): 1418–1420.
- [9] Mehndru N, Miao B Y, Almaraz E R, et al. Evaluating large language models as agents in the clinic [J]. *NPJ Digit Med*, 2024, 7(1): 84.
- [10] Collaco B G, Haider S A, Prabha S, et al. The role of agentic artificial intelligence in healthcare: a scoping review [J]. *NPJ Digit Med*, 2026, 9(1): 345.
- [11] Xu G L, Li X Y, Chen Y X, et al. A comprehensive survey of AI agents in healthcare [J]. *J Biomed Inform*, 2026, 179: 105045.
- [12] Schmidgall S, Ziaei R, Harris C, et al. AgentClinic: a multimodal benchmark for tool-using clinical AI agents [J]. *NPJ Digit Med*, 2026, 9(1): 499.
- [13] Liu Y S, Carrero Z I, Jiang X F, et al. Benchmarking large language model-based agent systems for clinical decision tasks [J]. *NPJ Digit Med*, 2026, 9(1): 259.
- [14] Ferber D, Hilgers L, Höper C, et al. Towards autonomous medical artificial intelligence agents [J]. *Nature*, 2026, 655(8125): 1282–1291.
- [15] Zhao W K, Wu C Y, Fan Y J, et al. An agentic system for rare disease diagnosis with traceable reasoning [J]. *Nature*, 2026, 651(8106): 775–784.
- [16] Kim Y, Park C, Jeong H, et al. MDAgents: an adaptive collaboration of LLMs for medical decision-making [C]// *Advances in Neural Information Processing Systems* 37. December 10–15, 2024. Vancouver, BC, Canada. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 79410–79452.
- [17] Tang X R, Zou A N, Zhang Z S, et al. MedAgents: large language models as collaborators for zero-shot medical reasoning [C]// *Findings of the Association for Computational Linguistics ACL* 2024. Bangkok, Thailand and virtual meeting. Stroudsburg, PA, USA: ACL, 2024: 599–621.
- [18] Chen X, Yi H H, You M K, et al. Enhancing diagnostic capability with multi-agents conversational large language models[J]. *NPJ Digit Med*, 2025, 8(1): 159.
- [19] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. *Advances in Neural Information Processing Systems*, 2020, 33: 9459–9474.
- [20] Yao S, Zhao J, Yu D, et al. ReAct: synergizing reasoning and acting in language models [C]// *International Conference on Learning Representations*. 2023.
- [21] Schick T, Dwivedi-Yu J, Dessi R, et al. Toolformer: language models can teach themselves to use tools [C]// *Advances in Neural Information Processing Systems* 36. December 10–16, 2023. New Orleans, Louisiana, USA. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 68539–68551.
- [22] Shinn N, Cassano F, Gopinath A, et al. Reflexion: language agents with verbal reinforcement learning [C]// *Advances in Neural Information Processing Systems* 36. December 10–16, 2023. New Orleans, Louisiana, USA. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 8634–8652.
- [23] Mandel J C, Kreda D A, Mandl K D, et al. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records [J]. *J Am Med Inform Assoc*, 2016, 23(5): 899–908.
- [24] Health Level Seven International. FHIR Release 4 (v4.0.1) [EB/OL]. [2026–10–08]. <https://hl7.org/fhir/R4/>.
- [25] Wu Q, Bansal G, Zhang J, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation [EB/OL]. arXiv: 2308.08155, 2023.
- [26] Li J, Wang S, Zhang M, et al. Agent Hospital: a simulacrum of hospital with evolvable medical agents [EB/OL]. arXiv:2405.02957, 2024.
- [27] Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence [J]. *Nature*, 2025, 642: 442–450.
- [28] McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models [J]. *Nature*,

- 2025, 642(8067): 451–457.
- [29] Ren Z Y, Zhan Y B, Yu B S, et al. Healthcare agent: eliciting the power of large language models for medical consultation[J]. *npj Artif Intell*, 2025, 1: 24.
 - [30] World Health Organization. Ethics and governance of artificial intelligence for health [M]. Geneva: World Health Organization, 2021.
 - [31] Bai C X, Song Y L, Yang D W. Hospital at home: from international evidence to a China-oriented pathway—building an integrated hospital–community–home model for respiratory home hospitalization coordinated with cloud outpatient care, with evaluation of clinical safety, environment[J]. *Metaverse Med*, 2026, 3(1): 1–10.
 - [32] Ayers J W, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum[J]. *JAMA Intern Med*, 2023, 183(6): 589–596.
 - [33] Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation[J]. *npj Digit Med*, 2025, 8: 274.
 - [34] Olson K D, Meeker D, Troup M, et al. Use of ambient AI scribes to reduce administrative burden and professional burnout [J]. *JAMA Netw Open*, 2025, 8(10): e2534976.
 - [35] Shah S J, Crowell T, Jeong Y, et al. Physician perspectives on ambient AI scribes [J]. *JAMA Netw Open*, 2025, 8(3): e251904.
 - [36] Shah S J, Devon-Sand A, Ma S P, et al. Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden [J]. *J Am Med Inform Assoc*, 2025, 32(2): 375–380.
 - [37] Unlu O, Varugheese M, Shin J, et al. Manual vs AI-assisted prescreening for trial eligibility using large language models—a randomized clinical trial[J]. *Jama*, 2025, 333(12): 1084.
 - [38] 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: 26.
 - [39] Liu F L, Zhou H J, Gu B Y, et al. Application of large language models in medicine [J]. *Nat Rev Bioeng*, 2025, 3: 445–464.
 - [40] 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.
 - [41] Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial[J]. *JAMA Netw Open*, 2024, 7(10): e2440969.
 - [42] Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians [J]. *JAMA Intern Med*, 2024, 184(5): 581.
 - [43] Liu X X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension[J]. *Nat Med*, 2020, 26: 1364–1374.
 - [44] Cruz Rivera S, Liu X X, Chan A W, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension[J]. *Nat Med*, 2020, 26(9): 1351–1363.
 - [45] Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI [J]. *BMJ*, 2022, 377: e070904.
 - [46] Collins G S, Moons K G M, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods [J]. *BMJ*, 2024, 385: e078378.
 - [47] Moons K G M, Damen J A A, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods[J]. *Bmj*, 2025, 388: e082505.
 - [48] Sounderajah V, Guni A, Liu X X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence [J]. *Nat Med*, 2025, 31(10): 3283–3289.
 - [49] Mongan J, Moy L, Kahn C E Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers[J]. *Radiol Artif Intell*, 2020, 2(2): e200029.
 - [50] Norgeot B, Quer G, Beaulieu-Jones B K, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist[J]. *Nat Med*, 2020, 26(9): 1320–1324.
 - [51] Collaborative T C. Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement[J]. *Bmjmed*, 2025, 4(1): e001632.
 - [52] Lekadir K, Frangi A F, Porras A R, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare [J]. *BMJ*, 2025, 388: e081554.
 - [53] Kelly C J, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence[J]. *BMC Med*, 2019, 17(1): 195.
 - [54] Sendak MP, D'Arcy J, Kashyap S, et al. A path for translation of machine learning products into healthcare delivery [J]. *EMJ Innovations*, 2020, 4: 3–13.
 - [55] Strong J, Rogers H, Sun E, et al. Human-AI collaboration in healthcare: a scoping review [J]. *NPJ Digit Med*, 2026, 9(1): 768.
 - [56] Parasuraman R, Sheridan T B, Wickens C D. A model for types and levels of human interaction with automation[J]. *IEEE Trans Syst, Man, Cybern A*, 2000, 30(3): 286–297.
 - [57] Skitka L J, Mosier K L, Burdick M. Does automation bias decision-making [J]. *Int J Hum Comput Stud*, 1999, 51(5): 991–1006.
 - [58] Ghassemi M, Oakden-Rayner L, Beam A L. The false hope of current approaches to explainable artificial intelligence in health care[J]. *Lancet Digit Health*, 2021, 3(11): e745–e750.
 - [59] Amann J, Blasimme A, Vayena E, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary

- perspective [J]. BMC Med Inform Decis Mak, 2020, 20 (1): 310.
- [60] Weidinger L, Uesato J, Rauh M, et al. Taxonomy of risks posed by language models [C]//2022 ACM Conference on Fairness Accountability and Transparency. Seoul Republic of Korea. ACM, 2022: 214–229.
- [61] Ji Z W, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Comput Surv, 2023, 55 (12): 1–38.
- [62] Cabitza F, Rasoini R, Gensini G F. Unintended consequences of machine learning in medicine [J]. Jama, 2017, 318 (6): 517.
- [63] Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations [J]. Science, 2019, 366(6464): 447–453.
- [64] Perets O, Stagno E, Ben Yehuda E, et al. Inherent bias in electronic health records: a scoping review of sources of bias [EB/OL]. medRxiv, 2024.
- [65] Chen R J, Wang J J, Williamson D F K, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare[J]. Nat Biomed Eng, 2023, 7(6): 719–742.
- [66] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0) : NIST AI 100–1[S]. Gaithersburg: NIST, 2023.
- [67] ANSI/AAMI/ISO 14971: 2019 Medical devices—Application of risk management to medical devices [S]. Arlington: AAMI, 2019.
- [68] ANSI/AAMI/IEC 62304: 2006+A1: 2015 Medical device software—Software life cycle processes [S]. Arlington: AAMI, 2015.
- [69] Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care [J]. Nat Med, 2019, 25: 1337–1340.
- [70] US Food and Drug Administration. Clinical decision support software: guidance for industry and FDA staff [S]. Silver Spring: FDA, 2026.
- [71] US Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions [S]. Silver Spring: FDA, 2025.
- [72] US Food and Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency. Predetermined change control plans for machine learning-enabled medical devices: guiding principles [EB/OL]. [2026–10–08]. <https://www.fda.gov/medical-devices/software-medical-device-samd/ predeterminded-change-control-plans-machine-learning-enabled-medical-devices>.
- [73] US Food and Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency. Transparency for machine learning-enabled medical devices: guiding principles [EB/OL]. [2026–10–08]. <https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices>.
- [74] International Medical Device Regulators Forum. Good machine learning practice for medical device development: guiding principles: IMDRF/AIML WG/N88 FINAL; 2025[S]. 2025.
- [75] European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence [J]. Official Journal of the European Union, 2024, L series.
- [76] 国家药品监督管理局医疗器械技术审评中心. 人工智能医疗器械注册审查指导原则[S]. 2022年第8号, 2022.
- [77] 国家卫生健康委办公厅, 国家发展改革委办公厅, 工业和信息化部办公厅, 等. 关于促进和规范“人工智能+医疗卫生”应用发展的实施意见: 国卫办规划发〔2025〕30号[Z]. 2025.
- [78] 国家互联网信息办公室, 国家发展和改革委员会, 等. 生成式人工智能服务管理暂行办法[Z]. 2023.
- [79] Luo L Y, Kim S E, Zhang X M, et al. A clinical environment simulator for dynamic AI evaluation [J]. Nat Med, 2026, 32 (3): 820–827.
- [80] Yang D W, Zhou J, Chen R C, et al. Expert consensus on the metaverse in medicine[J]. Clin eHealth, 2022, 5: 1–9.
- [81] 白春学. 元宇宙医学之我见[J]. 中国医药导刊, 2023, 25 (1):1–5.
- [82] 白春学. 医学GPT研发应用之吾见[J]. 元宇宙医学, 2025, 2(1):6–14.
- [83] 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.
- [84] 白春学. 肺结节专家——BAIMGPT白皮书[J]. 元宇宙医学, 2025, 2(2):55–64.
- [85] 白春学. AI智能体赋能肺结节诊治研发进展及展望[J]. 元宇宙医学, 2026, 3(2):83–90.
- [86] 国家卫生健康委办公厅, 国家中医药局综合司, 国家疾控局综合司. 卫生健康行业人工智能应用场景参考指引[Z]. 2024.

引用本文

宋元林, 白春学. 医学智能体: 体系架构、临床转化与可信应用[J]. 元宇宙医学, 2026, 3(3): 159–175.

Song Y L, Bai C X. Medical AI agents in health care: architecture, clinical translation, and trustworthy implementation [J]. Metaverse Med, 2026, 3(3): 159–175.