DOI：10.61189/297894knielb

·伦理与法规·

医学GPT安全、合规与伦理治理框架研究

高承实[1*]，张　烽[2]

1. 安徽栈谷科技有限公司，池州 247100

2. 万商天勤（上海）律师事务所，上海 200120

［摘要］ GPT在医学领域的应用正推动医疗人工智能向知识驱动范式转型，其在辅助诊断、医学问答等场景的技术潜力已得到广泛验证，但医学场景中数据与决策的高度敏感性，使得安全与合规成为系统部署不可回避的前提。本文围绕医学GPT在数据隐私、法规遵从与伦理治理中的系统性风险，提出一个以“数据—责任”可信链为核心的全生命周期综合治理框架，并系统识别其在三个维度的关键挑战。研究首先从技术机理层面剖析模型隐私再识别、模型泄露与有害使用的风险传导路径；进而结合医疗数据特性，对比分析HIPAA与GDPR框架下的合规要求差异及技术适配的核心痛点与解决方案；随后梳理全球医学AI伦理原则从软性倡议到硬性监管的制度化趋势，提出涵盖认知、操作、社会与结构四重风险的伦理评估矩阵；最终整合制度边界、技术边界与伦理底线，形成覆盖模型全生命周期的多层次治理框架。研究表明，医学GPT的可持续发展依赖于“数据—责任”可信链的构建，亟需技术方案、制度设计与伦理自觉的协同演进。未来行业竞争的核心不仅是算法性能之争，更是治理能力与信任机制的系统性比拼。

［关键词］ 医学GPT；数据隐私；HIPAA；GDPR；医学AI治理

［中图分类号］ R-052　　［文献标志码］ A

Research on the safety, compliance and ethical governance framework of medical GPT

GAO Chengshi[1*], ZHANG Feng[2]

1. Anhui Stack Alley Technology Co., Ltd, Chizhou 247100, Anhui, China

2. V&T Law Firm (Shanghai) Office, Shanghai 200120, China

［**Abstract**］ The application of generative pre-trained models in the medical field is driving the transformation of medical artificial intelligence (AI) towards a knowledge-driven paradigm. While their technical potential in auxiliary diagnosis, medical Q&A, and other scenarios has been widely verified, the high sensitivity of data and decisions in medical settings makes safety and compliance indispensable prerequisites for system deployment. This study aims to systematically identify the core challenges of medical generative pre-trained models in three dimensions: data privacy, regulatory compliance, and ethical governance, and construct a comprehensive governance framework with both theoretical support and practical feasibility. First, the research analyzes the risk transmission path of model privacy re-identification, model leakage, and harmful use from the perspective of technical mechanisms. Then, combined with the characteristics of medical data, it compares and analyzes the differences in compliance requirements under the HIPAA and GDPR frameworks, as well as the core pain points and solutions of technical adaptation. Subsequently, it sorts out the institutionalization trend of global medical AI ethical principles from soft initiatives to hard supervision, and proposes an ethical evaluation matrix covering four types of risks: cognitive, operational, social, and structural. Finally, it integrates institutional boundaries, technical boundaries, and ethical bottom lines to form a multi-level governance framework covering the entire life cycle of the model. The findings demonstrate that the sustainable development of medical generative pre-trained models critically depends on the construction of a “data - responsibility” trust chain, which urgently requires the coordinated evolution of technical solutions, institutional design, and ethical awareness. The core of future industry competition is not only the competition of algorithm performance, but also the systematic competition of governance capabilities and trust mechanisms.

［**Key Words**］ Medical GPT; data privacy; HIPAA; GDPR; AI ethics

在现有国际文件中，“医学GPT”通常被纳入“医疗场景中的生成式大模型”或“基于大语言模型的临床决策支持系统”加以讨论。本文沿用这一实践性界定，将医学GPT理解为基于大语言模型的生成

［收稿日期］ 2025-12-15　　［接受日期］ 2025-12-27

［作者简介］ 高承实，博士，副教授.

*通信作者（Corresponding author）. Tel: 021-64041990, E-mail: 13838001036@163.com

式人工智能系统，其被部署于医疗或健康相关场景中，用于提供医学信息、辅助临床决策或支持医疗流程，并可能对患者健康或诊疗决策产生实质性影响。

医学GPT的出现，标志着医疗人工智能从图像识别、文本解析等任务驱动模式，迈向能够理解复杂语境、生成连贯推理与建议的知识驱动新阶段[1]。通过学习大规模医疗多模态数据，医学GPT在辅助诊断、报告生成、医患沟通等场景展现出显著潜力，为缓解医疗资源矛盾、提升诊疗规范化水平提供了新路径[1]。但医疗领域“数据敏感性—决策高风险性”的双重属性，决定了医学GPT的“智能边界”必须与“可信边界”深度耦合，方能实现从技术原型到临床可信伙伴的跨越[2-4]。

当前医学GPT临床落地仍面临三重核心困境。数据层面，医疗数据多源异构特性与模型大规模训练需求叠加，传统数据保护手段失效，隐私泄露风险加剧[2]。伦理与责任层面，算法偏见、模型幻觉可能加剧健康不平等，而开发商、医疗机构与临床医生的责任划分模糊，形成全球性的问责困境[3-4]。治理层面，全球生成式AI监管框架尚在演进，现有法规与技术特性存在适配断层，调查显示法律不确定性是医疗AI应用的首要障碍之一[1,4-5]。这些困境相互交织，制约技术价值释放的同时，也对医疗信任体系构成挑战[2,6]。

基于此，本文旨在构建“数据隐私—法规合规—伦理治理”三维分析框架，解析医学GPT隐私风险生成逻辑与合规适配方案，梳理伦理原则制度化进程并构建风险评估体系，最终整合多要素形成全生命周期综合治理框架，为其在安全可控的前提下可靠合规应用提供理论与实践指引[2,4,6]。

1 数据隐私风险机理与法规合规适配

医学GPT训练与推理依赖大规模敏感医疗数据，隐私保护水平决定其可信基础。本节解析核心隐私风险，以HIPAA与GDPR为核心探讨合规适配路径与技术方案。

1.1 隐私风险的生成逻辑与传导路径

医学GPT隐私风险具有“隐蔽性强、传导复杂、影响广”特征，核心风险分三类且呈链式传导。

（1）隐私再识别风险，这是最直接的隐私风险。传统移除姓名、身份证号等去标识化处理难以应对医学GPT的大规模学习与生成机制[7-8]，模型记忆可能留存个体特征片段，通过成员推断攻击可以判断个体数据是否参与训练[9]；攻击者还可以通过反向提示诱导模型生成敏感信息，实现匿名数据再识别[10]。更关键的是，医疗数据的高维度关联性（症状组合、病史轨迹等），使隐性关联再识别成为当前核心痛点。

（2）模型泄露风险。模型参数是训练数据的统计浓缩，隐含了原始数据分布与个体特征。攻击者可以通过黑盒查询、白盒参数提取等模型窃取手段[11]，或利用迁移学习间接获取原始数据信息[12]。此类风险的隐蔽性在于，攻击者无需接触原始数据，仅通过模型交互即可实现信息窃取，突破传统数据存储的隐私保护边界。

（3）输出端误用与隐私放大风险。作为风险传导终端，模型输出的可控性决定了潜在危害程度。恶意使用者可能利用医学GPT伪造医疗证明、传播错误医学信息，或仿冒专业医务人员提供诊疗建议，从而对公众健康与医疗秩序造成损害[13-15]。上述行为本身属于生成式模型的误用或滥用风险，但在医疗场景中往往与隐私风险高度耦合。模型在生成过程中可能无意暴露训练阶段记忆的敏感信息；攻击者可通过批量生成或拼接输出内容，构造包含隐私线索的虚假病历或诊疗记录，从而放大隐私泄露的范围与后果[16]。此类风险具有“溯源难、管控难”的特征，常作为前端隐私与模型安全问题的最终外显形态。三类风险由此形成“技术缺陷—信息泄露—权益损害”的完整传导链条，决定了医学GPT的隐私保护不能局限于数据收集或存储阶段，而需覆盖“数据—模型—输出”的全流程[8]。

1.2 HIPAA框架下的合规要求与技术适配痛点

美国《健康保险可携带性与责任法案》（HIPAA）确立了医学GPT合规的两项核心准则：一，受保护健康信息（PHI）的使用原则上限于治疗（treatment）、支付（payment）和医疗运营（operations）三类目的，任何超出该范围的研究用途均需获得数据主体额外授权，或通过机构审查委员会（irb）伦理豁免程序[17]；二，需通过“专家判定法（expert determination）”或“安全港法（safe harbor）”实现数据去标识化，其中后者操作明确、合规确定性强，成为实践中的主流选择[18]。

然而，HIPAA制定于前大数据与生成式人工智能出现之前，其制度设计与医学GPT的技术特性之间存在显著适配断层。首先，传统去标识化机制难以应对高维医疗数据的关联再识别风险。HIPAA“安全港法”主要通过移除18类显性标识符降低再识别概率，但在不涉及模型外部误用的情况下，医学GPT仍可能基于疾病组合、诊疗路径、时间序列等隐性统计特征实现个体再识别，从而使形式合规

的数据处理仍面临实质性隐私泄露风险[18]。其次，HIPAA的监管适用主体主要限于传统医疗服务提供者及其业务关联方，尽管HITECH法案扩大了对部分商业伙伴（Business Associates）的责任覆盖，但在实践中，部分科技企业、云服务提供商或数据经纪商若未被明确纳入医疗业务关系，仍可能游离于HIPAA直接监管范围之外，这一结构性缺口在大模型研发与部署过程中被进一步放大[19]。

针对上述适配痛点，当前可行的技术路径并非单一脱敏手段，而是以“传统去标识化＋隐私增强技术（Privacy-Enhancing Technologies，PETs）”为核心的组合防护策略，在数据预处理阶段严格执行HIPAA合规去标识化要求；在模型训练阶段引入差分隐私机制，通过向梯度或参数更新注入可控噪声以抑制训练数据的可逆推断风险[20]；在多机构协同建模场景中，采用联邦学习等“数据不离域”的分布式训练方式，在不集中原始医疗数据的前提下完成模型优化[21]。需要指出的是，上述技术路径本质上属于隐私风险缓释机制，而非法律意义上的合规充分条件，其有效性仍依赖于制度设计、责任划分与持续审计的协同配合[20-21]。

已根据核验意见，对原文的表述、引用格式及参考文献进行了调整和补充。核心优化在于：合并GDPR引用并规范为条款标注，补充了权威的技术合规指南与案例文献，使“法律要求”与“技术重塑”之间的论证链条更为扎实和前沿。

1.3　GDPR的严格约束与对技术架构的重塑　欧盟《通用数据保护条例》（GDPR）以“权利导向”设定了严格标准，其核心约束对医学GPT的技术架构形成了系统性重塑[22]。目的限制原则（GDPR，Art. 5（1）（b））在制度层面促使行业从“先大规模泛化训练、再探索应用”的传统模式，转向在训练前即明确医疗场景与合法基础的“场景化定制训练”模式。数据主体权利（GDPR，Chapter III）则要求模型必须具备数据溯源与隔离删除能力，并通过可解释人工智能（XAI）与人工复核机制满足透明性与反对自动化决策权（GDPR，Art. 22）的要求。更为根本的是，严格的跨境数据传输规则（GDPR，Chapter V）直接推动了系统架构向“数据本地化”与“隐私计算驱动”范式的转型[23]。

（1）目的限制与合法性强化。GDPR要求数据处理目的必须“明确、具体”，且后续处理不得与原始目的“不相容”（GDPR，Art. 5（1）（b），6（4））。这一原则深刻改变了医学GPT的数据利用逻辑，使其难以沿用互联网领域“先收集、后开发”的路径。合规实践要求，在项目启动时就必须围绕如“糖尿病视网膜病变筛查”或“肺癌预后预测”等具体临床任务，界定最小必要的数据范围，并完成对应的合法性基础建设（如获得特定性同意或伦理审批）。这种“场景化定制训练”不仅强化了处理的合法性，也为全流程审计提供了清晰映射[22-23]。

（2）数据主体权利保障的技术实现。GDPR赋予个体的访问、更正、删除（被遗忘权）及反对自动化决策等权利（GDPR，Art. 15-22），对模型的可追溯性与可控性提出了刚性要求。为满足删除权，系统需设计细粒度的数据溯源图谱，使模型能够定位并隔离或遗忘特定训练数据的影响，这对集中式训练架构构成了挑战。同时，为避免模型成为法律所禁止的“黑箱”自动化决策工具，必须在高风险诊疗建议的输出环节，整合可解释AI（XAI）技术以提供推理依据，并设置强制性的人工临床复核节点，确保最终决定由医生作出，从而构建合法、可信的人机协同流程[22,24]。

（3）跨境限制驱动架构范式转型。GDPR规定，数据向欧盟外转移的前提是接收地具备“充分保护水平”，或需依靠标准合同条款（SCC）等提供等效保障（GDPR，Art. 44-49）。这一规定使得原始医疗数据跨境流动的成本与风险极高，从而直接重塑了技术架构的选择：在欧盟境内，项目倾向于采用本地化数据中心进行存储与计算；在必须进行跨区域协作的研发场景下，隐私计算技术成为唯一合规路径。具体而言，“模型跨境、数据不跨境”的联邦学习架构成为主流，各方仅在加密的模型参数层面进行协作，或利用同态加密技术在密文状态下进行计算。这种由合规驱动的技术转型，正促使医学GPT的基础架构从“数据集中”向“计算分布式、数据本地化”演进[23,25]。

1.4　全生命周期合规闭环的构建：技术与制度的协同　医学GPT需构建覆盖数据采集、预处理、训练、推理、迭代与销毁的全生命周期合规闭环，其核心在于技术与制度的深度协同。

（1）各环节合规控制点的技术嵌入。技术实现上，需在各环节嵌入合规控制点。数据采集与处理需遵循“最小必要”与目的限制原则[22]。预处理阶段应采用双重脱敏（如泛化、假名化）以降低再识别风险。训练阶段须引入差分隐私（DP）、联邦学习（FL）等隐私增强技术（PETs），在保护原始数据的同时完成建模[23,26]。推理部署时，应设置内容安全过滤与风险阈值控制，防止违规输出。在模型迭代与退役时，则需严格落实数据存储期限管理与可验证

的销毁机制[22,26]。

(2)问责制与隐私保护设计的制度落地。制度层面,关键在于通过明确的问责制(accountability)确保合规可执行。应设立数据保护官(DPO)进行统筹监督,在组织内部实施分级授权与双人复核等内部控制,并定期开展隐私影响评估(PIA)与第三方审计[27-28]。必须将"隐私保护设计(PbD)"理念前置性、强制性嵌入系统架构与研发流程,使合规成为内生属性,而非事后补救[22-23]。

(3)从原则到制度化执行的伦理治理。医学GPT的治理必须超越法律合规,回应更高阶的伦理诉求。全球AI伦理已从软性指南(如OECD原则[29])加速转向硬性监管(如《欧盟AI法案》将医疗AI列为高风险系统[27])。治理需系统应对算法偏倚、解释性不足、责任模糊及自动化过度等风险,在技术控制与组织制度之上,构建涵盖伦理审查、公众参与与持续监督的多层次治理体系[26,30]。

2 伦理与监管框架

医学GPT的治理不仅关乎法律合规,更涉及深层次的伦理价值与社会责任。全球范围内,医学AI伦理正经历从原则性倡议(soft law)向可执行监管框架(hard law)的制度化转型[29-30]。

2.1 医学AI伦理原则从软性倡议到硬性监管的制度化演进 早期AI伦理主要以非强制性原则为主。世界卫生组织(WHO)发布《人工智能医疗伦理与治理》报告,系统提出了保护人类自主性、促进福祉与安全、确保透明与可解释性、实现公平与包容性等核心伦理原则,为医学AI确立了基础框架[29]。该阶段伦理更多发挥规范引导作用,缺乏直接约束力。

近年来,伦理原则加速转化为具有法律效力的制度安排。欧盟《人工智能法案》(AI Act)明确将医疗AI纳入高风险系统监管范畴,要求其在全生命周期建立风险管理、数据治理、透明度与人类监督机制[27,31]。伦理要求由此从价值宣示转变为强制合规义务,未满足相关要求的系统不得进入市场。

在美国,医疗AI的伦理要求主要通过监管审批机制加以吸收。美国食品药品监督管理局(FDA)在其医疗AI行动计划与软件预认证实践中,将患者安全、透明性等伦理要素嵌入风险导向的医疗器械监管体系,推动伦理规范与技术审查的融合[32]。

医学AI伦理制度化呈现出三方面趋势,一是由结果合规转向覆盖全生命周期的过程治理;二是由抽象价值转向可操作、可审计的技术与管理指标;三是由单一开发者责任转向多主体协同的全链条治理结构,构成医学GPT伦理治理的制度基础[30,33]。

2.2 医学GPT的伦理风险四维解析与传导机制 医学GPT的伦理风险可归纳为认知、操作、社会与结构四个维度,并呈现逐级传导与放大的特征。

(1)认知风险源于生成式模型的固有缺陷。其"黑箱"特性与"幻觉"可能产生看似合理实则错误或伪科学的信息输出,在数据稀疏的罕见病诊疗等场景下风险尤为突出[1]。

(2)操作风险体现于临床人机交互过程。操作风险并非源于模型本身,而是源于人类对认知风险输出的使用与信任方式。医务人员可能对自动化输出产生过度信赖,形成"自动化偏见",导致误诊或漏诊;同时,模型若与既有临床工作流程适配不足,可能引发新的效率与安全隐患[1]。

(3)社会风险根植于训练数据的历史与社会偏差。数据代表性的不足可能导致算法在性别、种族等维度上产生系统性不公平,进而固化或加剧群体间的健康不平等[34-35]。

(4)结构风险指向权责界定模糊的制度困境。当模型决策导致损害时,开发商、医疗机构与临床医生之间的责任链条不清,将造成问责真空,削弱制度对患者权益的根本保障[1]。其中,认知、操作与社会风险构成过程性风险链条,而结构风险作为制度性元风险,对上述各环节产生放大效应。

上述风险沿"认知缺陷 → 操作失误 → 个体损害 → 社会不公"的路径逐级传导与放大,而结构风险贯穿全程并加剧各环节的负面效应。因此,有效的伦理治理必须进行全维度覆盖,并着力阻断这一传导链条[1]。

2.3 全球监管格局的演进与核心治理趋势 全球对医学GPT的监管呈现美、欧、中多路径并行但趋势趋同的格局,其共同方向是构建以风险为中心的全生命周期动态治理体系[36-37]。监管范式正从静态的产品审批,转向覆盖研发、部署与使用全过程的持续风险管理。

(1)美国的风险分级与动态更新。美国采取"软件即医疗器械(SaMD)"监管框架,由FDA根据产品风险实施分级审批,其最新进展是引入"预定变更控制计划"(PCCP),允许已获批的AI模型在预设控制计划内进行安全迭代,从而将全生命周期管理理念制度化[38-39]。

(2)欧盟预防为主与双法协同。欧盟通过《人工智能法案》(AI Act)与《医疗器械法规》(MDR)协同,将医疗AI明确列为高风险系统。核心是预防性监管,要求产品上市前建立并证明其具备全生命周

期的风险管理、数据治理及人类监督机制[27,40]。

（3）中国分类监管与伦理先导。中国由国家药监局（NMPA）对医学AI实施基于风险的分类管理。同时，《新一代人工智能伦理规范》确立了相关伦理要求，并通过推动临床伦理审查、探索医疗数据专区与隐私计算等技术路径，寻求数据利用与安全隐私的平衡[41-42]。

综上，全球监管呈现三大趋势，即从一次性审批转向持续性动态治理，从侧重技术性能转向防控安全与伦理风险，从单一监管主体转向多利益相关方协同治理结构[36-37]。

2.4　多层次伦理治理框架从原则到实践的落地路径　为推动医学GPT伦理原则的有效转化，需构建覆盖“原则—政策—机构—技术—社会”的多层次治理框架，将抽象价值转化为可操作、可监督的治理安排[30]。该框架旨在通过制度与组织传导，最终落实于技术实践与社会反馈。

（1）原则层确立伦理价值共识。国际组织与国家通过发布伦理指南确立规范性基础。例如，WHO报告系统阐述了保护人类自主性、促进福祉与安全等核心伦理原则；中国《新一代人工智能伦理规范》则确立了国内应用的基本伦理要求[30,42]。

（2）政策与法律层将伦理原则转化为强制规则。立法与监管将伦理要求转化为法律义务。欧盟《人工智能法案》已生效，明确将医疗AI列为高风险系统，并设定全生命周期的风险管理义务[27]。美国食品药品监督管理局（FDA）则在其人工智能/机器学习医疗器械监管中，将安全与风险控制要求嵌入基于全生命周期（TPLC）理念的审评流程[38]。

（3）机构治理层将伦理要求内化为组织流程。医疗机构与研发企业需建立内部机制，将原则转化为日常规范。中国国家卫健委已明确要求，应用人工智能的医疗卫生机构应建立健全相应的伦理审查机制与治理体系[43]。

（4）技术执行层将伦理目标嵌入系统设计。通过可解释性、偏见检测与校正等技术手段，将公平、透明等要求转化为可验证的指标。当前，行业技术白皮书与标准正致力于定义针对医疗大模型的安全性、偏见及责任评估框架[44]。

（5）社会反馈层强化透明度与公共监督。通过公开伦理影响评估、引入第三方评估与公众参与渠道，增强社会信任。全球治理新趋势强调制定国际标准与多边协作，确保技术发展符合公共利益[45]。

综上，多层次治理通过“价值共识→制度约束→组织执行→技术落实→社会监督”的闭环，为医学GPT的负责任应用提供了系统性路径[30,43,45]。

2.5　责任与问责机制：全链条追溯与归因　明确的责任与问责机制是医学GPT可信落地的核心，当前主要挑战是责任界定模糊。世卫组织2025年报告显示，在接受调研的50个欧洲国家中，仅约8%制定了明确的医疗AI责任标准，凸显了构建清晰责任体系的紧迫性[46]。国际共识指向建立覆盖研发、部署与使用全链条的责任体系，并依靠技术实现可追溯性，以避免“责任外包”与“问责真空”[1,27]。

（1）研发责任：首要合规与产品责任。研发者承担首要合规责任，需确保数据合法、模型安全可控，并充分披露其性能边界与局限[27]。这既符合欧盟《人工智能法案》对高风险AI系统提供者的义务规定，也对应FDA框架下作为“医疗产品”提供者需提交安全有效性证据的监管要求[38]。

（2）部署责任：组织治理与风险控制。引入医学GPT的医疗机构承担核心部署责任。其义务超越采购，需对模型进行准入审查与风险评估，并通过制度化流程将其安全嵌入临床工作流，设立强制性的人工复核节点[47]。研究表明，一旦发生相关诊疗损害，因其直接管理责任，医疗机构往往成为首要的责任承担主体[48]。

（3）使用责任：临床判断的不可替代性。临床医生承担最终的、不可替代的诊疗责任。医生必须对模型输出进行批判性评估并作出独立判断，此即“合理医生”标准下的注意义务[47]。同时，医生需履行告知义务，保障患者知情权。

（4）技术支撑：可追溯审计。落实问责需技术驱动，记录模型调用、决策推导及人工干预的全链路日志，形成可审计的轨迹[49]。国际标准组织（ISO/IEC）正推动《AI系统影响评估》指南，为记录关键信息提供标准化框架，构建可核查的责任链条[46]。

综上，医学GPT的责任治理依赖于“法定责任分配”与“技术全程可追溯”的双轮驱动，将伦理原则转化为可执行、可追责的具体保障机制[27,47,46]。

3　医学GPT的安全边界与伦理底线

基于前文对医学GPT隐私风险与伦理挑战的系统分析，可以发现，相关风险并非源于单一技术缺陷，而是制度安排、工程设计与价值取向多重因素叠加的结果。因而，有必要进一步从“安全边界”与“伦理底线”两个层面，对医学GPT的可信运行条件进行结构化界定。

3.1　安全边界的多维构建：制度、技术与伦理的协同约束　医学GPT的安全边界由制度、技术与伦理三

个层面协同构成，是其可信应用的基础框架[1,30]。制度边界提供外在强制约束，技术边界实现内在行为控制，伦理边界则提供价值引导，三者共同支撑“数据—责任”可信链，确保系统坚守辅助诊疗定位[27,31]。

（1）制度边界：外在强制约束，由法律法规与行业标准构成，明确行为的“红线”。例如，欧盟《人工智能法案》将医疗AI列为高风险系统，明确禁止其进行自主诊疗决策，并设定了严格的合规义务与监管机制，从外部强制划定安全范围[27]。

（2）技术边界：内在行为控制。技术边界通过工程手段，在系统内部约束其行为。这包括通过输入过滤、对抗性训练防止有害指令，通过输出校验与权限控制确保结果安全可靠，并通过全面的日志审计实现全过程可追溯。美国NIST的《人工智能风险管理框架》与欧盟ENISA的《AI网络安全框架》为这些控制措施提供了系统性方法论[47-48]。

（3）伦理边界：价值嵌入与导向。伦理边界旨在将公平、普惠、尊重生命等价值原则，转化为系统设计与评估中的可操作维度。这要求在算法层面嵌入偏见检测与校正机制，并在应用导向上优先支持基层医疗、健康普惠等公益场景，使技术发展始终以增进人类福祉为根本导向[30,49]。

这三重边界相互依赖、协同作用：制度为技术与伦理实践提供合法性保障，技术为制度与伦理要求提供落地工具，伦理则为前两者提供价值校准。唯有实现三者的深度融合，才能为医学GPT构建稳固、可执行的安全边界[27,30]。

3.2 不可逾越的核心伦理底线 医学GPT的伦理底线构成其可信应用的最低约束，应具备可识别、可执行、可审计的制度特征[1,30]。国际组织与监管实践已形成基本共识，即医学GPT至少需坚守不越权、不失真、不歧视、可问责四项不可逾越的价值准则，以防止技术对医学专业判断、患者权益与公共信任造成结构性侵蚀[1,27]。

（1）不越权：坚守辅助诊疗定位。医学GPT应被严格限定为临床决策支持工具，其输出必须明确标注为“辅助建议”，并在制度上设置强制性人工复核节点，确保最终诊疗决策始终由具备执业资质的医生作出[1,44]。该要求已在多国监管与行业实践中被反复强调，其核心在于落实“责任不可让渡”的医学职业伦理，即任何情况下均不得由模型替代医生行使诊疗裁量权[27,38]。

（2）不失真：依托权威证据并披露不确定性。医学GPT的输出应以循证医学为基础，应通过检索增强生成（RAG）等技术手段致力于连接权威医学数据库，以提升信息的可追溯性与可靠性基础，并对证据不足或结论不确定的内容进行明确提示[1,49]。实证研究表明，大语言模型在医疗场景中可能生成不完整、偏差性甚至误导性的建议，尤其在复杂或边界性病例中，其输出可靠性存在显著波动[50]。因此，需将幻觉风险、临床相关准确率等指标纳入全生命周期持续评估与管理框架[27,47]。

（3）不歧视：保障公平性与普惠可及。医学GPT在设计与部署中必须系统防范算法偏见风险。这要求训练数据覆盖不同人群特征，并通过分群性能评估与偏见检测机制，持续监测模型在性别、年龄、种族及社会经济背景等维度上的表现差异[1,47]。现有医学大语言模型普遍存在显著的人口统计学偏差，可能在诊疗建议、护理方案等方面加剧健康不平等[51]。此类系统综述提供的证据表明，算法偏见并非偶发现象，而是需要从系统设计源头加以防控的普遍性风险。因此，公平性干预与可及性优化应被视为伦理底线而非附加目标。

（4）可问责：建立全生命周期追溯机制。伦理底线的落实依赖于清晰、可执行的问责结构。这要求对数据来源、模型训练、系统部署与临床使用等关键环节进行全流程记录，形成可审计的责任链条[27,50]。一旦发生不良事件，应基于完整的技术与管理记录，精准区分研发方、部署机构与临床使用者的责任边界，从而实现可归因、可追责的责任落实机制[30,51]。

总体而言，上述伦理底线通过将抽象价值原则转化为制度约束、技术控制与责任追溯的具体要求，为医学GPT划定了不可突破的安全与伦理边界，是其实现可信、可持续应用的前提条件[1,27,30]。

4 结论与展望

医学GPT的可持续发展核心在于构建“数据—责任”可信链，实现技术、制度与伦理协同演进。本文通过三维分析框架解析核心风险与治理痛点，提出全生命周期综合治理框架，主要结论如下：

第一，医学GPT隐私风险呈“链式传导”特征，需构建“技术+制度”全生命周期合规闭环。隐私再识别、模型泄露、有害使用形成完整风险链，需采用差分隐私、联邦学习等技术，结合数据保护官、定期审计等制度，实现全流程管控，针对性解决传统法规与AI技术的适配断层。

第二，医学GPT伦理风险具“四维矩阵”特征，需推动伦理原则制度化落地。认知、操作、社会、结

构四类风险逐级放大，需通过“原则—政策—机构—技术—社会”多层次框架，将伦理价值转化为可量化指标，实现全链条管控。

第三，医学GPT可信基础依赖“安全边界+伦理底线”协同构建。制度、技术、伦理三维边界形成协同约束机制，不越权、不失真、不歧视、可问责四项底线明确价值准则，共同构成“数据—责任”可信链核心，是临床可信应用的前提。

未来研究可深化三方面：技术层面开发高效隐私增强与可解释AI技术；制度层面推动跨境监管协同与国际标准建立；实践层面通过监管沙盒积累伦理治理经验。

医学GPT未来竞争核心是治理能力与信任机制之争。唯有将安全、合规与伦理内化为系统内生属性，构建稳固可信链，才能使其真正赋能医疗实践，实现从“算法强大”到“系统可信”的关键跨越。

伦理声明　无。

利益冲突　作者声明不存在利益冲突。

作者贡献　高承实：选题、撰写、修改论文。

参考文献

[1] WHO. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. Geneva: World Health Organization, 2025.

[2] 中华人民共和国国家卫生健康委员会等. 关于促进和规范“人工智能+医疗卫生”应用发展的实施意见[R/OL]. (2025-10-30) [2025-12-01]. https://www.gov.cn/zhengce/zhengceku/202511/content_7047018.htm

[3] 仝媛媛，郑婉婷，黄珊蓉，等. 健康公平：医疗人工智能中的偏见与治理[J]. 医学与哲学，2024，45(7)：36-41.

[4] ZHANG J，ZHANG Z M. Ethics and governance of trustworthy medical artificial intelligence[J]. BMC Med Inform Decis Mak，2023，23(1)：7.

[5] 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.

[6] RATTI E. Ethical and Social Considerations of Applying Artificial Intelligence in Healthcare；a Two-Pronged Scoping Review[EB/OL]. (2025)[2025-12-10]. https://ouci.dntb.gov.ua/en/works/l110DV35/.

[7] EL EMAM K，ARBUCKLE L. Anonymizing health data：Case studies and methods to get you started[M]. Sebastopol：O'Reilly Media，2015.

[8] ZHANG Y，LIU J，WANG Y，et al. Differential privacy for medical large language models：A practical implementation[J]. Journal of Biomedical Informatics，2023，143：104325.

[9] SHOKRI R，STRONATI M，SONG C Z，et al. Membership inference attacks against machine learning models[C]//2017 IEEE Symposium on Security and Privacy (SP). May 22-26，2017，San Jose，CA，USA. IEEE，2017：3-18.

[10] CARLINI N，TRAM{\'E}R F，WALLACE E，et al. Extracting training data from large language models[C]//USENIX Security Symposium.，2020

[11] TRAM{\'E}R F，ZHANG F，JUELS A，et al. Stealing machine learning models *via* prediction *APIs*[C]//USENIX Security Symposium.，2016

[12] YANG Q，LIU Y，CHEN T，TONG Y. Security and privacy in deep learning[J]. IEEE Access，2020，8：153703-153718.

[13] BENDER E M，GEBRU T，MCMILLAN-MAJOR A，et al. On the dangers of stochastic parrots：can language models be too big[C]//Proceedings of the 2021 ACM Conference on Fairness，Accountability，and Transparency. Virtual Event Canada. ACM，2021：610-623.

[14] FENG S，KUMAR A，JUNG J，et al. Safety and trustworthiness in large language models：A survey[J]. arXiv preprint，arXiv：2306.11695，2023.

[15] WANG H，CHEN X，ZHOU B. Detecting sensitive information leakage in generative medical text models[C]//Proceedings of the 2024 ACM Conference on Health，Inference，and Learning. New York：ACM，2024：289-298.

[16] LI B，ZHANG Y，WANG J，et al. Privacy risks of generative models：A survey[J]. IEEE Transactions on Knowledge and Data Engineering，2024.

[17] U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES. Summary of the HIPAA Privacy Rule[EB/OL]. (2022-11)[2025-12-02]. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html.

[18] U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the HIPAA Privacy Rule[EB/OL]. (2012-10-25)[2025-12-09]. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.

[19] COHEN I G，MELLO M M. Big data，big tech，and protecting patient privacy[J]. JAMA，2019，322(12)：1141-1142.

[20] DWORK C，ROTH A. The Algorithmic Foundations of Differential Privacy[M]. Hanover：Now Publishing，2014：145-189.

[21] RIEKE N，HANCOX J，LI W Q，et al. The future of digital health with federated learning[J]. NPJ Digit Med，2020，3：119.

[22] EU. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL[S]. Official Journal of the European Union，(2016-04-27)[2025-12-05]. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[23] CNIL. AI How-To Sheets：Aligning Artificial Intelligence Systems with the GDPR[EB/OL]. (2025-07-16)[2025-12-10]. https://www.cnil.fr/en/ai-how-to-sheets.

[24] 李　明，李昱熙，戴　廉，等. 医疗人工智能伦理若干问题探讨[J]. 医学与哲学，2019，40(21)：1-4.

[25] FRAUNHOFER INSTITUTE FOR APPLIED INFORMATION TECHNOLOGY FIT. ELMTEX – Affordable, Privacy-Compliant AI for European Healthcare [EB/OL]. (2025-03-13) [2025-12-05]. https://www. fit. fraunhofer. de/en/business-areas/digital-health/projects/ELMTEX-Affordable,-Privacy-Compliant-AI-for-European-Healthcare. html.

[26] ISO/IEC 23894: 2023. Information technology-artificial intelligence-guidance on risk management[S]. 2023.

[27] EU. REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL[S]. Official Journal of the European Union, (2024-06-13) [2025-12-01]. https://eur-lex. europa. eu/legal-content/EN/TXT/PDF/ uri=OJ: L_202401689#: ~: text=The% 20purpose% 20of% 20this% 20Regulation%20is%20to%20improve, systems%20in%20the%20Union%2C%20and%20to%20support%20innovation.

[28] EUROPEAN UNION AGENCY FOR CYBERSECURITY. Data protection engineering: from theory to practice[R]. 2023.

[29] YEUNG K. Recommendation of the council on artificial intelligence (OECD)[J]. Int Leg Mater, 2020, 59(1): 27-34.

[30] WHO. Ethics and governance of artificial intelligence for health [R]. Geneva: WHO, 2021.

[31] EUROPEAN PARLIAMENTARY RESEARCH SERVICE. The EU's Artificial Intelligence Act[R/OL]. (2024-02-09) [2025-12-06]. https://www. europarl. europa. eu/thinktank/en/document/EPRS_BRI(2021)698792

[32] U. S. FDA. Artificial intelligence/machine learning (AI/ML)-based software as a medical device action plan [R]. Silver Spring, MD: FDA, 2021.

[33] WIRTZ B W, WEYERER J C, GEYER C. Artificial intelligence and the public sector—applications and challenges [J]. Int J Public Adm, 2019, 42(7): 596-615.

[34] MEHRABI N, MORSTATTER F, SAXENA N, et al. A survey on bias and fairness in machine learning [J]. ACM Comput Surv, 2022, 54(6): 1-35.

[35] CELI L A, CELLINI J, CHARPIGNON M L, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities-a global review [J]. PLoS Digit Health, 2022, 1(3): e0000022.

[36] WHO. Regulatory considerations on artificial intelligence for health [R]. Geneva: WHO, 2025.

[37] ONG J C L, NING Y L, LIU M X, et al. Regulatory science innovation for generative AI and large language models in health and medicine: a global call for action[EB/OL]. 2025: arXiv: 2502. 07794. https://arxiv. org/abs/2502. 07794

[38] U. S. FDA. Artificial intelligence and machine learning (AI/ML)-enabled medical devices [EB/OL]. (2021-01-12) [2025-12-12]. https://www. fda. gov/news-events/press-announcements/fda-releases-artificial-intelligencemachine-learning-action-plan

[39] EUROPEAN COMMISSION. Questions and Answers on the interaction between the Artificial Intelligence Act and the Medical Devices Regulation and the In Vitro Diagnostic Medical Devices Regulation [EB/OL]. (2025-06-19) [2025-12-14]. https://health. ec. europa. eu/latest-updates/mdcg-2025-6-faq-interplay-between-medical-devices-regulation-vitro-diagnostic-medical-devices-2025-06-19_en#:~:text=MDCG%202025-6%20-% 20FAQ% 20on% 20Interplay% 20between% 20the, Regulation% 20and% 20the% 20Artificial% 20Intelligence%20Act%20%28June%202025%29

[40] 国家药品监督管理局. 人工智能医用软件产品分类界定指导原则[S]. 2021.

[41] 中华人民共和国科学技术部. 新一代人工智能伦理规范[S]. 2021.

[42] 上海交通大学, 复旦大学, 上海交通大学医学院附属瑞金医院等. 医疗健康大模型伦理与安全白皮书[R/OL]. (2025-07-18) [2025-12-13]. https://www. docin. com/p-4899644536. html

[43] 2025 世界人工智能大会. 人工智能全球治理行动计划[Z]. (2025-07-26) [2025-12-14]. https://www. gov. cn/yaowen/liebiao/202507/content_7033929. htm

[44] WHO. Regulatory considerations on artificial intelligence for health: European report[R]. 2025.

[45] 国家卫生健康委员会. 人工智能在医疗卫生机构应用管理暂行办法[Z]. 2024.

[46] ISO/IEC FDIS 42005:2025. AI system impact assessment [S]. 2025.

[47] NIST. Artificial Intelligence risk management framework (AI RMF 1. 0)[R]. Gaithersburg: NIST, 2023.

[48] ENISA. AI cybersecurity framework: securing artificial intelligence throughout the lifecycle [R]. Heraklion: ENISA, 2024.

[49] MCLENNAN S, et al. Embedding ethical principles into AI for health: A scoping review of value alignment approaches [J]. The Lancet Digital Health, 2024, 6(5): e332-e342.

[50] LIU C X, ZHENG J N, LIU Y S, et al. Potential to perpetuate social biases in health care by Chinese large language models: a model evaluation study [J]. Int J Equity Health, 2025, 24(1): 206.

[51] OMAR M, SORIN V, AGBAREIA R, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review[J]. Int J Equity Health, 2025, 24(1): 57.

引用本文

高承实,张 烽. 医学GPT安全、合规与伦理治理框架研究[J]. 元宇宙医学,2025,2(4):53-60.

GAO C S, ZHANG F. Research on the safety, compliance and ethical governance framework of medical GPT[J]. Metaverse Med,2025,2(4):53-60.