

DB4403

深圳市地方标准

DB4403/T XXX—XXXX

卫生健康领域生成式人工智能应用指南

Guidelines for implementing generative artificial intelligence in
healthcare

(送审稿)

XXXX-XX-XX 发布

XXXX-XX-XX 实施

深圳市市场监督管理局 发布

目 次

目 次 II

前 言 IV

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 原则 1

 4.1 合规性 1

 4.2 安全性 1

 4.3 透明性 2

 4.4 公平性 2

 4.5 先进性 2

 4.6 可控性 2

5 全生命周期流程 2

6 启动阶段 2

 6.1 成立管理小组 2

 6.2 制定管理制度 3

 6.3 需求分析和目的确定 3

7 选型阶段 3

 7.1 初步模型选择 3

 7.2 评估 3

 7.3 评估结论 6

8 部署阶段 6

8.1 部署前调整 6

8.2 部署后评估验证 7

8.3 上线测试 7

9 运维阶段 7

9.1 持续监测 7

9.2 更新 8

9.3 定期评估 8

9.4 终端用户 8

9.5 审计追溯 8

10 下线阶段 8

附 录 A （资料性） 应用场景分类及潜在风险 10

参考文献 12

前 言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由深圳市卫生健康委员会提出并归口。

本文件起草单位：深圳市卫生健康发展研究和数据管理中心、深圳市前海蛇口自贸区医院、深圳市福田区妇幼保健院。

本文件主要起草人：

卫生健康领域生成式人工智能应用指南

1 范围

本文件提供了卫生健康领域生成式人工智能应用的原则、全生命周期流程的启动、选型、部署、运维、下线各阶段的指南。

本文件适用于指导深圳市各机构在卫生健康领域应用生成式人工智能的活动。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 35273 信息安全技术 个人信息安全规范

GB/T 39725 信息安全技术 健康医疗数据安全指南

GB 45438 网络安全技术 人工智能生成合成内容标识方法

GB/T 45654 网络安全技术 生成式人工智能服务安全基本要求

YD/T 6090 面向互联网的医疗人工智能辅助决策 基于病理图像的辅助决策系统算法指标和测试方法

3 术语和定义

下列术语和定义适用于本文件。

3.1

生成式人工智能 generative artificial intelligence

具有文本、图片、音频、视频等内容生成能力的模型及相关技术。

3.2

服务提供者 service providers

以平台、系统、接口等形式面向公众或特定用户提供生成式人工智能服务的组织。

4 原则

4.1 合规性

确保利用人工智能技术生成的内容导向正确，符合国家相关法律法规以及行业规范和标准，尊重社会价值观和道德标准。

4.2 安全性

服务提供者在数据全生命周期实施分类分级保护，收集、存储、使用、加工、传输、提供、公开的过程中应提供相关措施确保数据安全和自主可控，不得危害他人身心健康，不得侵害个人隐私权、信息权等权益。

4.3 透明性

在应用生成式人工智能技术的过程中，所有参与者都应了解并明确披露人工智能内容生成及使用情况，包括但不限于底层数据集、数据来源和数据处理方法等。

4.4 公平性

在训练数据选择、算法设计、模型生成和优化、使用过程中仔细评估和审查潜在的数据和内容偏差来源，同时减少人工智能生成内容在文化或者语言上的不公平现象。

4.5 先进性

选择具有先进人工智能生成内容技术和算法模型的服务提供者，即该服务提供者所具有的算法核心指标和权威评测得分在行业内达到领先水平。

4.6 可控性

建立动态熔断及人工介入机制，对所有应用场景，特别是高风险场景（如临床诊断建议、治疗方案生成）、紧急医疗场景（如急诊、重症监护）等具有最终决策权，确保人工智能始终处于人类控制之下，以保障医疗质量和患者安全。

5 全生命周期流程

卫生健康领域应用生成式人工智能实行全生命周期管理，包括：启动阶段、选型阶段、部署阶段、运维阶段以及下线阶段。

6 启动阶段

6.1 成立管理小组

管理小组由具有不同专业及背景的人员组成，包括机构管理人员、医务人员、人工智能技术人员、信息技术人员、数据安全专家、医学伦理专家、厂商、法律法规专业人士等。其中医务人员宜具有丰富临床经验，覆盖内科、外科、妇科、儿科等核心科室。

小组职责为确保安全有效地使用生成式人工智能，具体包括但不限于：

- a) 确保机构内生成式人工智能应用遵守法律法规；
- b) 维护伦理准则；
- c) 保护数据隐私与安全；
- d) 与相关方协商合作；

- e) 确定责任部门和人员；
- f) 建立生成式人工智能应用管理制度；
- g) 进行生成式人工智能应用前评估，评估内容相关信息，见表 1；
- h) 提供应用审批；
- i) 监测模型运行；
- j) 监督模型的维护和更新；
- k) 定期进行应用期间评估，评估内容相关信息，见表 1；
- l) 进行终端用户的使用指导及反馈收集。

6.2 制定管理制度

制定管理流程及方案，内容涵盖评估项中安全与伦理要求，并有日常报告和解决流程、变更管理流程、意外事件管理流程；跨机构合作时，规范数据共享的权责归属和安全保障措施；实施生成式人工智能在多角色、多权限环境下的信息安全隔离机制，及检测到异常输出时的紧急熔断机制。

6.3 需求分析和目的确定

管理小组进行需求分析，确定内容生成式人工智能应用需求、目的及具体任务，包括收集用户需求及各相关方需求和数据，开展文献研究。

7 选型阶段

7.1 初步模型选择

基于需求分析结果及目的，筛选生成式人工智能模型，并确认所选模型已通过国家互联网信息办公室备案。

7.2 评估

评估生成式人工智能的安全与效用、风险、卫生经济学3个维度，各维度评估内容及要点如下：

表 1 生成式人工智能应用评估框架

一级维度	二级维度	评估内容	具体评估内容
安全与效用评价	安全	医疗安全	模型应用涉及的医疗临床安全情况，包括场景是否为诊疗核心环节，是否直接影响临床决策或涉及生命支持场景，错误后果的严重性、可控性、影响范围。
		数据安全	模型涉及的健康、医疗信息的安全管理，防止数据被未经授权的访问、篡改或泄露，包括数据加密、数据备份、数据访问控制等安全措施。
		隐私安全	具有保护个人敏感信息免受非法访问或泄露的措施，包括数据脱敏、数据掩码、模型加密、个人信

一级维度	二级维度	评估内容	具体评估内容
			息使用和处理条款等隐私保护措施，并明确告知用户数据收集、使用、共享的目的和方式等，个人信息的使用及存储符合 GB/T 35273 的规定。
		问责	明确责任方及责任归属。责任方能够回溯使用情况及问题环节，并采取纠正措施。
		透明度	服务提供者充分披露和全面记录研发信息、模型概要信息、使用情况信息等，并符合 GB/T 45654 中 6.2 和 6.3 的规定。模型生成的意见可解释，显示推理过程及文献依据。模型生成内容同时采用显式标识（如界面提示“AI 生成内容需医生确认”）和隐式标识（如嵌入不可篡改的元数据），模型生成内容标识方式符合 GB 45438 的规定。数据合规证明符合 GB/T 39725 的规定。
	效用	准确性	模型生成的内容真实、精确、无误，包括准确率（临床指南依从性、医学事实回答等）、精确率、召回率、时效性、完整性、语义一致性等。
		逻辑性	模型生成的内容直接明确，与用户的问询保持一致，无不必要或不相关信息。在不同时间或不同措辞下回答同一问题，逻辑自洽，推理稳定。
		响应	处理数据和生成预测时所消耗的计算资源及时间。包括吞吐量、延迟时间、并发数、生成速度、浮点运算次数。
		适用性	模型应用场景与其设计和开发目的相匹配。
		覆盖度	模型应用覆盖的广度及深度，包括模型服务的人口数量，涉及的医疗设施数量及种类、医疗环境类型等。
		整合度	模型与现有卫生健康服务信息系统进行整合的工作量及复杂度，包括机构信息化水平、算力资源、数据交换、工作流程协同等。
		终端用户能力	终端用户个人的技能水平，包括患者、医务人员和

一级维度	二级维度	评估内容	具体评估内容
			非医务工作人员的模型使用能力，医学素养等。
		员工用户体验	患者及普通民众用户的体验感受及满意度，包括：信息准确性、反馈及时性、信任度、接受度、医患互动、诊疗过程舒适度等。
		公众用户体验	患者及普通民众用户的体验感受及满意度，包括：信息准确性、反馈及时性、信任度、接受度、医患互动、诊疗过程舒适度等。
风险评价	技术应用	公平	模型生成的内容公平客观、无系统性偏见、不对某些社会群体产生刻板印象及不利影响或伤害，如民族、地域、年龄或性别偏见。通过偏见检测，确保群体差异敏感度。
		自主	维护患者的尊严、自主权和知情同意权，同时能防止医务人员对人工智能的过度依赖和技能退化。
		人文关怀	能保持必要的具有人文关怀的医患互动和情感支持，能够维持具有人文关怀的医患互动模式，提供有效的情感支持。
	技术内在	算法	对算法模型稳健性、泛化性、鲁棒性、可解释性、可信任性等进行测试。
		训练数据	所使用训练数据的科学性、有效性、合规性以及制定不良训练数据限制规则和防范毒性训练数据的沾染。
		无害性	模型对潜在有害问题，包括仇恨言论、暴力、欺诈、违法内容等，能做出拒答，不会提供间接帮助。
卫生经济学评价	成本	费用成本	为部署、运行模型应用所投入的资金总和，包括硬件采购、租赁、软件授权、数据获取、云服务等货币支出。
		时间成本	模型应用的规划、开发、测试、部署、维护及人员使用培训全过程所消耗的时间总量。
		人力成本	参与模型应用研发、运维、数据处理等相关工作人

一级维度	二级维度	评估内容	具体评估内容
			员薪酬、福利等各类人力投入的总和。
	效益	经济效益	模型带来的工作流程优化和工作效率的改善及相应的费用降低，包括节省的临床、行政、科研等时间、经费。
		功能效果	模型应用的功能效果，包括诊断准确性、病案编码准确率、问诊效率、并发症预警准确率、药物研发效率、医患沟通效率等。
		健康效益	模型应用带来的健康改善结果，包括卫生健康服务使用者健康、生活质量、用药依从性、治愈率等。

7.3 评估结论

根据评估给出结论，分为不应用，条件性应用，一般性应用三种，内容及要点如下：

表 2 评估结论

评估结论	定义	判断参考
不应用	在当前条件下，不适宜应用于特定领域或场景。	经过全面综合评估，认定该技术在应用中存在显著不足或重大潜在风险，未能达到预设的性能、安全等要求；或各项评估结果均较差。
条件性应用	可应用，需在特定条件下应用，和/或制定特定的操作规程或使用限制。	多数评估项达到基本要求和需求，并显示技术具有应用潜力，但部分评估结果显示存在一些限制或不足，风险不高。可通过明确设定应用条件或限制应用范围，规避存在的不足。
一般性应用	可应用，应用时满足卫生健康领域生成式人工智能的基础通用要求。	各项评估结果均良好，满足应用条件。

8 部署阶段

8.1 部署前调整

根据评估结论，对分类为条件性应用和一般性应用的模型，评估结果较差及不足的项目进行优化，并根据机构实际情况进行本地化适配调整，优化管理及微调模型。

微调模型宜使用符合行业要求的卫生健康专用标准化数据集进行训练，使用的数据集应具有多样性和代表性，避免模型输出偏见，以确保模型输出结果的客观性和公正性。使用机构内部数据进

行微调的，数据样本量及样本分布参考YD/T 6090-2024, 6.1.2.2和6.1.2.3，且模型仅限于内部应用，不与外部机构共享。

微调模型后进行备案，按照《互联网信息服务算法推荐管理规定》履行算法备案和变更、注销备案手续。

8.2 部署后评估验证

与现有信息系统整合，宜使用标准接口。整合后，对模型进行再次测试及评估完成不少于 200 例的验证，所用验证案例满足机构的特点及需求。确认满足所有部署条件后，方能正式部署。

操作日志保存期不少于10年。

8.3 上线测试

正式上线前进行压力测试、集成测试、安全测试及应急演练，确保模型性能及安全运行。

压力测试验证模型在使用高峰时段的响应速度及数据处理准确性，无数据重复或丢失。

集成测试验证模型与其他系统的接口和协议的交互是否正常，数据传输是否准确。模拟真实医疗场景，检查模型在完整业务流程中的表现，确保各环节无缝衔接。验证模型输入输出数据在不同子系统间的传递是否一致，避免因集成导致的数据格式错误或丢失。

安全测试模拟医疗数据泄露场景，如异常批量下载时，验证系统的实时预警及熔断能力。

应急演练模拟急诊等场景下模型输出错误建议时，手动否决及切换人工决策的流程顺畅性。

9 运维阶段

9.1 持续监测

根据建立的管理机制，持续监测模型性能，并跟踪运行中安全风险，记录运行全流程情况，确保全程可追溯，用于模型评估、倒查应用过程情况等工作。

监督生成式人工智能行为和影响，确保其服务运行在人类授权和控制范围内，防止过度依赖人工智能决策。监控及记录未采纳人工智能决策的情况，并对决策不一致进行分析，具备在遭遇事故时及时切换到人工或传统系统等的能力。

加强网络保障措施，持续监测网络安全、供应链安全等方面的能力，以尽量减低网络攻击、数据盗窃或泄漏的风险，并确保业务连续性。

保持与服务提供厂商的沟通交流，做到对模型应用的信息同步，并存储交流记录。

建立反馈机制，设置便捷的反馈入口，并鼓励社会公众、机构专业人员、技术人员提供反馈意见。

9.2 更新

定期更新模型，确保其性能保持最新。符合最新法规及行政主管部门要求、最新的医学知识、用户需求。更新后进行测试和评估，确保更新后模型性能不降低、运行效率不降低、模型输出结果稳定，并在其性能下降时对其进行重新训练。

硬件更新，确保硬件与模型匹配。

持续改进管理方案及流程，确保与当前技术、法规、最佳实践和用户需求保持一致。

9.3 定期评估

根据表1，定期进行核查、评估，并收集反馈；可引入第三方伦理委员会或独立审计机构，持续改进和完善模型的功能。

9.4 终端用户

建立明确的使用和服务条款，明确双方权利与义务，知识产权与责任归属。向用户披露使用的背景和目的，并建议用户不要超出预期的使用范围。告知用户在使用模型的过程中其数据将被用于后续的训练、迭代，并取得用户的知情同意授权。

定期全面、系统培训员工用户，内容涵盖以下方面：

- a) 技术基础：生成式人工智能的基本概念、核心原理和在卫生健康领域中的功能应用；
- b) 决策机制：生成式人工智能的决策过程及其局限性；
- c) 人的监督意识：人与生成式人工智能交互时对其进行监督和控制的重要性；
- d) 使用适当性：生成人工智能的适宜使用时机，以及在不同临床情境中可能遇到的潜在限制；
- e) 信息可靠性：幻觉和信息准确性等风险及有效的预防和应对措施；
- f) 预防偏见：自动化决策过程中识别和避免偏见及技术应用的公平和公正；
- g) 隐私保护：病人隐私保护意识；
- h) 防止技术滥用：防止技术滥用和恶意使用；
- i) 法规政策：最新政策和法规相关要求；
- j) 技能增强：有效和高效地使用生成式人工智能方法。

9.5 审计追溯

审计追溯需包含以下方面：

- a) 算法审计：每季度进行模型权重分布检测，提供算法备案表（含模型架构图、训练数据分布说明）；
- b) 数据审计：实施数据安全审计；
- c) 伦理审计：年度偏见审查；
- d) 行为审计：操作日志区块链存证。

10 下线阶段

模型下线后，模型与数据彻底清理、模型权限回收、法律合规闭环、后续风险监控。

附 录 A
(资料性)
应用场景分类及潜在风险

表A. 1提供了卫生健康领域生成式人工智能应用场景分类及潜在风险

表 A. 1 应用场景分类及潜在风险

一级分类	二级分类	使用对象	潜在风险
医疗健康管理	临床专病智能辅助决策	医护人员	误诊 过度依赖 技能退化 偏见 医患互动及人文关怀减少 知情同意
	智能就医咨询	患者	隐私泄露 内容不准确、不可靠
	临床用药智能辅助	医护人员	数据质量缺陷 人机协同 责任界定 知情同意
	智能预问诊	患者	自动化偏见 隐私泄露
	患者用药指导智能辅助	医护人员 患者	不准确、不完整或虚假陈述 操纵 隐私泄露 减少临床医生与患者之间的互动 认知上的不公正 在卫生系统外提供医疗服务的风险
	智能患者院后管理	患者	内容不准确、不专业 隐私泄露
	智能随访	医护人员 患者	数据安全与隐私保护 技术适应性问题 系统集成与标准化问题
	商业健康险智能设计	产业人员	设计方案不贴合实际 合规风险 公平性风险 数据安全
	智能医院管理	医院管理人员	生成方案不可靠 隐私泄露
	智能病历辅助生成	医护人员	病历内容模板化

			生成内容不符合实际诊疗操作 数据泄露
	智能医疗文书质控辅助	医护人员	质控标准僵化 误判合规文书 隐私泄露 技能退化
	中医临床智能辅助诊疗	医护人员	忽视患者体质禁忌 过度依赖
基层公卫服务	健康管理	公众	隐私泄露 生成方案不可靠
	公共卫生	公众	责任界定风险 数据安全 自动化偏见 信息误导 技术风险
	养老托育	公众	生理安全的误判与疏漏 医疗建议越界 隐私泄露 责任界定风险
健康产业发展	药物研发	科研人员	数据隐私与合规风险 数据偏见 数据质量问题 知识产权模糊性
	医用机器人	患者	无法识别紧急病情，错过最佳时间 隐私信息过度采集
医学教学科研	医学教学	医护人员	助长自动化偏见 错误或虚假信息有损医学教育质量 学习数字技能的新负担 内容过时 信息冲突、误导学习 知识产权纠纷 过度依赖
	医学科研	科研人员	无法让算法对内容负责 算法编码偏见 生成不存在的信息和/或参考文献 破坏科学研究的关键原则，如同行评议 加剧科学知识获取的差异 数据错误误导科研

参 考 文 献

[1] 国家互联网信息办公室 中华人民共和国国家发展和改革委员会 中华人民共和国科学技术部 中华人民共和国工业和信息化部 中华人民共和国公安部. 生成式人工智能服务管理暂行办法: 国家广播电视总局令 (2023) 15 号. 2023-07-10.

[2] 国家互联网信息办公室 工业和信息化部 公安部 国家广播电视总局. 人工智能生成合成内容标识办法: 国信办通字 (2025) 2 号. 2025-03-07.

[3] 国家互联网信息办公室 工业和信息化部 公安部. 互联网信息服务深度合成管理规定: 第 1 2 号. 2022-11-25.

[4] GB/T 45288.2 人工智能 大模型 第 2 部分: 评测指标与方法

[5] World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models[R]. Geneva: WHO 2024.

[6] U.S. Department of Health & Human Service. Trustworthy AI (TAI) Playbook. [R]. 2021.
