《中国卫生质量管理》第33卷 第5期(总第222期)2026年05月

医疗健康领域人工智能应用安全框架研究*

——鞠 鑫 姚永刚 赵芯蕊

【摘 要】 随着AI技术的快速发展,大模型在医疗健康领域的应用潜力日益凸显,医疗数据的高敏感性和隐私保护需求也对大模型全生命周期安全管控提出了更高要求。本研究围绕医疗健康领域AI大模型在私有云环境中的应用场景,提出了一套从大模型底座、运行环境到具体应用的全生命周期安全管控框架,主要包含数据采集、大模型训练、大模型推理和应用环节输入输出4个阶段的安全架构设计。通过引入多项国内外前沿大模型安全技术,包括数据隔离保护机制、模型攻击防护算法、隐私计算融合方案以及运行时动态风险监控等,全面提升了大模型安全性能,实现了医疗健康AI大模型安全风险闭环管理,为医疗健康领域AI应用安全管理提供了参考。

【关键词】 医疗领域;人工智能大模型;可信执行环境;风险检测大模型;安全框架

中图分类号:R197.3 文献标识码:B

Research on the Security Framework for Artificial Intelligence Applications in the Medical and Health Field/JU Xin, YAO Yonggang, ZHAO Xinrui. //Chinese Health Quality Management, 2026, 33(5): 07-11

Abstract With the rapid development of artificial intelligence (AI) technology, the application potential of large models in the healthcare field has become increasingly prominent. The high sensitivity of medical data and the demand for privacy protection have also imposed higher requirements on the full lifecycle security management and control of large models. Focusing on the application scenarios of AI large models in the healthcare field within private cloud environments, this study proposes a full lifecycle security management and control framework spanning from the foundation of large models and their operating environments to specific applications. It primarily encompasses security architecture design across four stages: data collection, large model training, large model inference, and input or output in application scenarios. By introducing multiple internationally and domestically leading security technologies for large models, including data isolation protection mechanisms, model attack prevention algorithms, privacy computing integration solutions, and runtime dynamic risk monitoring, the overall security performance of large models has been comprehensively enhanced. This achieves closed-loop management of security risks for AI large models in healthcare, providing a reference for AI application security management in the healthcare field.

Key words Medical Field; Artificial Intelligence Large Models; Trusted Execution Environment; Risk Detection Large Models; Security Framework

First-author's address Suzhou Health Information Center, Suzhou, Jiangsu, 215623, China

AI通过重塑诊疗流程、赋能精准医疗,正在引发医疗行业的系统性变革。在此过程中,作为生成式人工智能的核心技术,大模型因其强大的语言理解、数据分析及决策支持能力,逐渐成为智慧医疗重要推动力。相应地,大模型面临的安全威胁也日益显现,如模型被恶意利用生成虚假信息、隐私泄露问题以及模型的偏见和不公平等问题[1]。医疗健康行业因其数据的高度敏感性、严格的隐私保护要求以及复杂的法律法规环境,使得该领域大模型的安全性成为亟待解决的核心问题。为有效发挥医疗健康AI技术的积极作用,必须对之实行严格的评估与风险控制[2]。

苏州市医疗健康大数据中心在此背景下启动了AI能力中心规划,致力于探索医疗健康大模型应用创新路径及推动苏州市卫生健康行业

DOI:10.13912/j.cnki.chqm.2026.33.5.02

* 基金项目:2024苏州市科教强卫项目(编号:MSXM2024043)

苏州市卫生健康信息中心 江苏 苏州 215623

AI公共服务能力建设。该中心使用大模型私有化部署方式，在提升安全性的同时，可以建立公众对AI应用的信心[3]。本研究探讨了苏州市卫生健康AI能力中心在大模型安全管控中的创新实践，重点分析了医疗健康领域大模型的潜在安全风险，并提出了一套以全生命周期为核心视角的安全管理框架，以期为医疗健康领域人工智能应用提供安全支撑。

1　大模型安全研究背景

1.1　大模型安全问题分析

大模型的安全问题主要涉及以下几个方面：(1)数据隐私保护。大模型的训练通常需要大规模数据支持，包含敏感的个人信息数据，即使经过脱敏处理，大模型仍可能通过重构等手段泄露敏感信息。(2)模型鲁棒性。大模型易受对抗性攻击的影响，微小扰动输入即可显著改变模型输出，在医疗应用场景中可能导致诊断错误或决策失误。(3)复杂运行环境。大模型运行通常依赖复杂的软硬件环境，攻击者可通过系统漏洞或环境劫持影响模型的正常运行，甚至窃取关键数据或模型参数。(4)模型治理与合规。医疗健康行业需严格遵循法规要求，大模型的复杂性可能使其在合规性审查和治理方面面临挑战。

针对上述问题，当前国内外大模型安全研究重点聚焦以下方面：(1)数据安全与隐私保护技术。联邦学习和差分隐私技术被广泛应用于医疗领域，以减少数据泄露风险。其中，联邦学习框架已经成功用于医疗数据分析，但仍面临高计算成本和通信开销的问题。(2)对抗性攻击与防御。针对对抗性攻击的防御措施包括对抗性训练、模型正则化等。例如，百度文心一言、抖音云雀大模型、智谱AI的GLM大模型、百川智能的百川大模型、商汤的日日新大模型和科大讯飞星火大模型等，在增强大模型鲁棒性方面取得了显著进展，但这些方法在医疗场景中的适用性仍需进一步验证。(3)安全环境的构建。国内外研究机构和企业正在探索如何通过可信执行环境(trusted execution environment，TEE)、容器隔离和动态安全监控等技术提升大模型运行环境的安全性，如展讯、联发科和华为海思的CPU支持TEE并已将之用于保护大模型的运行安全，但该技术在大规模医疗数据场景下的性能优化仍是一个难题。(4)法规合规与伦理审查。欧美国家在制定AI应用安全标准方面起步较早，但在医疗大模型的具体法规指导上尚未形成统一规范。国内《数据安全法》《个人信息保护法》等法律框架为医疗AI的应用提供了安全性指导，但大模型的合规性验证体系仍需完善。

1.2　大模型安全在医疗健康领域的应用现状

在医疗健康领域，大模型的安全问题已引起广泛关注。国外部分医疗机构和企业正在探索大模型安全管理的实践路径，例如，美国的梅奥诊所和英国的国民医疗服务体系等正在尝试将大模型应用于医学影像分析和辅助诊断，同时构建数据隔离和隐私保护机制，以减少安全风险[4]。而国内该领域尚处于萌芽阶段，国内医疗机构和企业正在推动大模型在智慧医疗中的落地应用，尤其是在私有云环境下，重点加强数据隐私保护、模型对抗防御和运行时监控。例如，北京协和医院AI研究中心与国内AI技术企业合作，将大模型用于医学影像辅助诊断，同时构建了专用的隐私保护和安全防护机制；平安好医生通过构建私有云环境下的大模型系统，实现基于用户健康数据的智能问诊和健康管理，并在大模型运行时具有安全防护能力。尽管研究取得了一定进展，但大模型安全在医疗健康领域的应用仍面临诸多挑战，包括：大模型安全管控与性能的平衡、私有云环境下大规模部署的技术瓶颈、法规和伦理要求的动态适配等。随着医疗健康大数据中心的建设，探索创新的安全技术和管理框架，能够推动医疗健康大模型的落地应用。

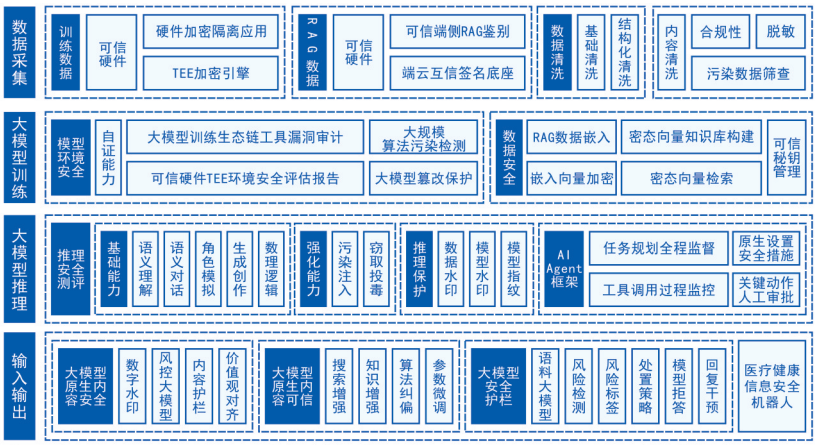
2　AI应用安全管理框架及其核心技术

在AI应用过程中，数据生命周期的每一个环节都可能造成个人、商业或国家敏感信息的泄露，从而危害个人隐私、社会安全与国家安全[5]。因此，研究提出了一套针对医疗健康行业的AI大模型全生命周期安全管理框架，覆盖数据采集、大模型训练、推理和应用输入输出的各个阶段，整体设计架构如图1所示。

为确保TEE在早期大模型应用中可靠运行，从而保障模型训练及推理过程的安全性，TEE架构必须从中央处理器(central processing unit，CPU)打通到图形处理器(graphics processing unit，GPU)。因此，算力平台硬件选型在第一阶段将根据应用需求以Intel SGX搭配NVIDIA H20和华为昇腾搭配鲲鹏为主要配置。

2.1　数据采集阶段安全防护

数据采集是大模型构建的首要



注：TEE（trusted execution environment）为可信执行环境；RAG(retriever-augmented generation)为检索增强生成。

图 1　苏州市医疗健康 AI 应用安全框架架构图

环节，需确保数据的隐私性和合规性。在数据采集阶段主要有 3 种防护措施：一是基于可信硬件的加密与隔离应用。通过利用支持 TEE 的服务器处理器，在数据采集、传输和存储过程中实现敏感数据的加密和隔离，以降低数据泄露风险。二是检索增强生成（retriever－augmented generation，RAG）数据安全管理。结合端云互信加密技术，确保 RAG 场景中数据传输和使用安全，通过向量加密与隐私保护，实现数据的“可用不可见”，避免数据被篡改或泄露。三是数据清洗与质量提升。当采集的数据在清洗过程中存在偏差时，用其所训练的大模型也会存在相应偏差，基于这些模型的医疗健康应用势必会复制并放大这种偏差[6]。因此，制订大模型数据的清洗与安全标准，才能为高质量的模型训练奠定基础。

2.2　大模型训练阶段安全防护

大模型训练是 AI 系统的核心环节，涉及数据、环境及算法等方面的安全保障。本阶段研究重点包括：(1)模型训练环境安全。针对复杂的训练生态链系统，开展全链路漏洞检测，可以防止因组件漏洞导致的用户数据泄露和服务中断。(2)算法污染检测。由于大语言模型使用了庞大的语料库，开发人员无法对所有数据进行审核，因此不可避免地会包含一些负面文本（如有毒数据和隐私数据），这些负面文本会直接影响大语言模型应用的安全性[7]。对此，可利用统计分析和异常检测技术识别中毒数据，清除污染影响并修复模型，建立模型恢复机制，从而降低污染带来的损失。(3)大模型篡改防护。采用加密和完整性验证技术，可防止大模型在训练和部署过程中被恶意篡改或注入后门，确保模型的可信性。(4)RAG 数据安全防护。大模型训练过程可能无意中记住并泄露敏感信息，从而带来潜在的风险，即使患者数据经过匿名化处理，也可能遭受对抗性攻击导致泄露，并且存在重新识别的风险[8]。在 RAG 场景中对嵌入向量进行加密，可确保语义匹配和检索过程的隐私性，保障数据在传输和存储过程中的保密性和完整性。

2.3　大模型推理阶段安全防护

推理阶段是大模型生成输出的关键环节，其安全性直接影响决策质量。因此，本阶段重点关注以下 4 个方面的防护：一是推理安全测评。通过开展基础安全能力、强化安全能力及推理保护能力的多维度评估，可确保模型在推理过程中的安全性。二是不良内容检测。结合自然语言处理分类模型与大语言模型技术，构建不良内容检测引擎，确保模型输出符合伦理与法规要求。三是提示注入攻击检测。训练专用检测模型，能够识别提示泄漏、反向诱导等多种攻击，防止提示注入攻击造成的数据泄露或模型异常。四是幻觉检测。大语言模型的幻觉问题是指模型在处理输入任务、维持输出语境连贯性以及与现实世界事实保持一致性时存在偏差或错误[9]。针对医疗场景中可能出现的模型幻觉现象，可通过使用幻觉检测引擎来识别。

本阶段所使用的核心技术主要有：(1)搜索增强能力集成。通过集成先进的搜索引擎技术，实时监测和过滤潜在的风险内容，能够确保大模型在处理输入时不会受到敏感或有害信息影响。(2)知识增强与指令微调。对模型进行指令微调训练，可以提升模型对敏感内容的识别和抵抗能力。(3)RAG 技术降低大模型内容安全风险。采用安全回复 Safety－RAG 技术，为大模型补充外部知识，可以提升模型对敏感主题、隐晦语义的识别能力。Safety－RAG 技术通过实时检索安全知识库，可以为大模型生成的内容提供额外约束和指引，确保输出内容的安全正向。(4)算法纠偏能力应用。通过监测模型输出中的偏差，并应用算法进行自动纠偏，能够确保模型在处理各种输入时保持输出的客观性和准确性。

2.4　输入输出环节全面安全管理

在应用环节，从输入的可信性与输出的安全性层面展开管理：(1)大模型内容安全护栏。通过建立多级检测机制，对用户输入和模型输出进行全程监控，确保内容的合法性与安全性。针对敏感或违规问

题，提供安全代答与主干模型输出结合的解决方案。(2)大模型原生安全加固。通过强化学习与指令微调，提高大模型对敏感内容的抵抗能力，确保输出质量与安全性的平衡。(3)模拟攻击验证能力。利用自主训练的攻击大模型生成大量攻击样本，进行自动化攻击测试，验证并优化防护系统。

本阶段所使用的核心技术有：(1)风险检测。对输入内容进行全面检测，识别潜在的高风险敏感内容，生成风险标签及处理建议，检测结果直接传递给干预回复模块或安全回复模块，形成后续处理链条。(2)执行干预回复。通过调用干预库返回预设答案，实现对高风险输入输出内容的直接覆盖处理。(3)使用安全代答。针对非直接屏蔽处理的敏感问题，调用安全回复大模型生成合规的替代答案，提升用户体验。(4)检测过滤。使用大模型内容过滤系统对通用大模型的输入提示词和输出结果进行检测与改写，确保大模型在生成内容时符合安全与合规要求。

3　应用效果

本研究提出的AI全生命周期安全防护框架已成功应用于苏州市卫生健康AI能力中心多个医疗健康场景，通过严格的安全管理流程与核心技术的实施，打通治理框架内各安全要素之间的连接链路，形成动态治理能力[10]。该框架投入运行近4个月，显著提升了医疗大模型应用的安全性、稳定性和可信性。

数据采集阶段，提升了数据隐私保护水平。通过引入TEE技术，敏感医疗数据在采集、传输和存储过程中得到了硬件级别的隔离与保护，数据泄露风险降低了90%以上；数据质量全面提升，数据清洗与质量提升模块显著改善了训练数据的规范性和一致性，剔除的污染数据比例高达15%，并确保数据合规性达到行业法规标准，有效保障了计算结果的准确性。

模型训练阶段，增强了训练平台漏洞防护能力。全链路漏洞检测系统有效发现并修复了大模型训练生态中的多种安全漏洞，其中，Web类漏洞12项，内存类漏洞5项，模型类漏洞3项，系统运行稳定性提升了35%。算法污染检测效果显著，通过模型行为监测和数据中毒检测，成功识别了20多种潜在算法污染风险，大模型在对抗性环境中的鲁棒性提升了约40%。

模型推理阶段，提升了推理内容安全能力。风险检测大模型和不良内容过滤系统在医疗场景中的识别准确率达到了97%，敏感内容误判率降低至1.5%。提示注入攻击防护能力得到了强化，通过红蓝对抗测试，框架成功防御了90%以上的提示注入攻击，有效保护了推理结果的安全性与系统敏感信息。幻觉检测引擎作用突出，在医疗问诊场景中，幻觉检测引擎对潜在幻觉现象的识别率超过95%，大幅度降低了模型输出错误信息的风险。

应用输入输出环节，保障了模型内容安全。输入风险检测与输出结果安全评估系统构建了双向防护屏障，模型内容违规率从原来的4.2%降至不足0.5%。原生安全加固有效性提升明显，通过强化学习和指令微调，模型对医疗专业问题的安全回答能力提升了20%，敏感问题处理的准确率提高至98%。

4　讨论

本研究围绕医疗健康行业AI大模型全生命周期进行了管理与防护，结合苏州市卫生健康AI能力中心的实际场景，提出了一套系统化的安全管理框架，从大模型的数据采集、模型训练、推理到应用输入输出环节，结合TEE、RAG、对抗性防御、隐私计算等前沿技术，系统性解决了医疗健康行业大模型不同阶段的安全问题。实现了多场景安全实践，成功应用于智慧医疗问诊、医学影像分析及多机构协作等实际场景，验证了该框架在数据安全、模型可信性和推理可靠性方面的能力，为行业应用提供了参考案例。

本研究的创新设计主要体现在：(1)大模型机密容器基于CPU的TEE保护向GPU的TEE协同工作对大模型训推运行时所在内存与显存中的数据实施硬件级隔离。(2)对于端侧传向AI能力中心端的RAG，通过同态加密和构建向量知识库，实现了RAG数据安全和应用保护。用户RAG数据的嵌入与加密采用同态加密技术，确保用户数据在存储和检索过程中始终处于加密状态[11]。(3)基于联邦学习框架实现大模型分布式训练和推理，通过细粒度的聚合机制，实现了步骤级安全聚合，与传统轮次级聚合方式相比，能够更频繁地更新和优化模型参数，从而加速模型收敛并提升学习效果[12]。

在技术架构实施过程中，实施人员也发现了一些需要完善的方面。其一，基于CPU的TEE技术能力在当前阶段技术成熟度是足够的，但在面向大模型运行时，环境GPU的TEE能力明显不足。例如目前全球范围内只有英伟达的H系列显卡对TEE有较为成熟的支持能力，国产显卡对TEE的支持多处于初期阶段，普遍不成熟。且截至2025年第一季度，英伟达H系列显

卡对TEE的适配依然无法有效支持八卡并行，只能支持单卡应用，这将对大模型业务算力环境带来较大影响。TEE能力是由CPU统一调度的，GPU的TEE能力需要对CPU进行适配，这就导致在对TEE支持上，CPU和GPU需要进行不同组合的适配，每一种组合的适配需要较大工作量，研发难度也较高，导致当前阶段无法大面积普及应用GPU的TEE能力。其二，基于联邦学习的大模型训练开发框架，从研究角度可以实现，但是在项目落地执行方面依然存在一定难度。主要表现在大模型应用软件开发商对于一种全新开发框架的引入，以及新框架在产品开发中的落地，都因成本较高存在被动性。因此，在整个方案的执行过程中，大模型机密容器在GPU的TEE支持与适配上，只能采用阶段性推进方法。训练数据高度敏感和高安全性的大模型应用，选择目前适配较为成熟的Intel SGX＋NVIDIA H20硬件架构，但对于大多数非高安全性的大模型应用，则采用国产算力卡硬件架构，待国产GPU的TEE能力成熟，再进行硬件架构的迭代。在分布式训练方面，选择一部分积极性较高大模型应用软件服务商，对基于联邦学习的分布式大模型开发框架进行尝试性引入，并根据实际工作情况不断完善方案，重点在于开发框架的快速引入和降低使用成本。基于以上两点，本方案着重将更多资源投向大模型内容安全、合规性安全以及大模型运行环境风险控制领域。当前，大模型面临着诸多安全威胁，而在应对这些威胁的手段中，聚焦于上述方向是可行性较高的选择。与此同时，这些技术节点本身具有显著风险属性，一旦运行异常，可能会引发严重后果，因此加大资源投入力度具有必要性与紧迫性。

而在面对更加复杂的医疗场景和快速发展的AI技术时，本研究安全管理框架仍需进一步优化和探索。一是优化安全知识图谱，使用知识图谱可以对整个攻击过程进行全面描绘，从而管理和利用大量的网络安全情报信息[13]，以应对不断演化的攻击方式。二是进一步研究数据共享和语义检索的隐私保护技术，提升数据交互的安全性和效率。三是大模型合规性与伦理保障，包括建立伦理委员会或专家团队，对模型的训练数据和算法进行审查。四是多模态模型的安全管理，通过创新工具与算法快速完成多样性病例数据的采集，从而实现对多模态真实世界数据的融合建模，辅助开展前瞻性与回顾性研究[14]。五是增加AI决策的透明度，在医疗健康领域，为增强深度学习模型的可解释性，可采用局部可解释性方法，通过制订可追溯、可解释的模型设计和决策规则，使决策过程透明化、可解释化[15]。

作者贡献：鞠鑫负责资料分析、论文撰写、研究选题及思路指导；姚永刚负责资料收集、论文修改、数据核对；赵芯蕊负责资料收集、数据核对。

利益冲突：所有作者声明本文无实际或潜在的利益冲突。

参考文献

[1] 付志远，陈思宇，陈骏帆，等. 大语言模型安全的挑战与机遇[J]. 信息安全学报，2024(5)：26－55.

[2] 肖非易，李　雪，李　睿，等. 医疗人工智能技术评估与监管的国际经验及启示[J]. 中国卫生质量管理. 2023，30 (7)：58－62.

[3] 杨晓姣，罗　仙，张　玲. 从美智库报告看人工智能大语言模型网络安全问题及对策[J]. 信息安全与通信保密，2024(8)：20－29.

[4] PRANAV R，EMMA C，OISHI B，et al. AI in health and medicine[J]. Nat Med，2022，28(1)，31－38.

[5] 林梓瀚，郭　丰. 人工智能时代我国数据安全立法现状与影响研究[J]. 互联网天地，2020(9)：20－25.

[6] 周吉银，刘　丹，曾圣雅. 人工智能在医疗领域中应用的挑战与对策[J]. 中国医学伦理学，2019，32 (3)：281－286.

[7] 王乔晨，吴振刚，刘　虎. 大语言模型应用的安全与隐私问题综述[J]. 工业信息安全，2024(5)：40－45.

[8] 孙　磊，汪安安，宋一敏，等. 大语言模型在临床医学领域的应用、挑战和展望[J]. 解放军医学院学报，2025，46(1)：50－60.

[9] 赵　月，何锦雯，朱申辰，等. 大语言模型安全现状与挑战[J]. 计算机科学，2024，51(1)：68－77.

[10] 皮　勇，张明诚. 总体国家安全观视域下人工智能安全风险治理研究[J]. 中国科技论坛，2023(6)：86－96.

[11] WANG C，CAO N，REN K，et al. Enabling secure and efficient ranked keyword search over outsourced cloud dat[J]. IEEE Transactions on Parallel and Distributed Systems，2012，23(8)：1467－1479.

[12] KAIROUZ EBP，MCMAHAN HB，AVENT B，et al. Advances and open problems in federated learning[J]. Foundations and Trends® in Machine Learning，2021，14(1)：1－2.

[13] 池亚平，吴　冰，徐子涵. 大语言模型在威胁情报生成方面的研究进展[J]. 信息安全研究，2024，10(11)：1028－1035.

[14] 潘伟华，许　健，穆嘉盛，等. 应用人工智能提升单病种质量管理水平[J]. 中国卫生质量管理，2022，29 (7)：19－21.

[15] 胡振生，杨　瑞，朱嘉豪，等. 大语言模型在医学领域的研究与应用发展[J]. 人工智能，2023(4)：10－19.

通信作者：
姚永刚：苏州市卫生健康信息中心副主任，正高级工程师
E－mail：leek0118@qq.com

收稿日期：2024－12－02
修回日期：2025－03－13
本文编辑：黄海凤、刘斯妤