DOI：10.61189/117222wmzjug

·伦理与法规·

人工智能医学应用中的技术路径与法律边界——以强化学习与蒸馏技术为视角

卢叶婷[1*]，杨宇平[2]

1. 上海小度律师事务所，上海 200135

2. 香港城市大学，香港特别行政区 999077

［摘要］ 本文探讨了先进人工智能技术在医学领域的应用，特别关注强化学习与蒸馏技术的创新应用及应用中需注意的法律合规问题。以最新的大型语言模型（LLM）技术结合强化学习和蒸馏技术发展为例，分析了AI技术知识产权边界、医疗数据使用合规性以及医疗AI监管框架的关键问题。文章讨论了如何在推进医疗AI创新的同时，确保患者权益保护和医疗伦理规范遵守，为医疗AI的健康发展提供了理论参考。

［关键词］ 强化学习；蒸馏技术；人工智能；医学GPT

［中图分类号］ R-0　　［文献标志码］ A

Technical path and legal boundary in the medical application of artificial intelligence: from the perspective of reinforcement learning and distillation technology

LU Yeting[1*], YANG Yuping[2]

1. Shanghai ShallDo Law Firm, Shanghai 200135, China

2. City University of Hong Kong, Hong Kong Special Administrative Region 999077, China

［**Abstract**］ This paper discusses the application of advanced artificial intelligence technology in the field of medicine, with special attention to the innovative application of reinforcement learning and distillation technology and the legal compliance issues that need attention in the application. Using the latest large language model (LLM) technology combined with reinforcement learning and distillation technology developments as examples, the key issues of AI technology intellectual property boundaries, healthcare data use compliance, and healthcare AI regulatory framework are analyzed. This paper discusses how to promote the innovation of medical AI while ensuring the protection of patients' rights and interests and compliance with medical ethics, and provides a theoretical reference for the healthy development of medical AI.

［**Key Words**］ reinforcement learning; distillation; artificial intelligence; medical GPT

1　技术突破与法律边界之争

人工智能在医疗领域的应用正迅速扩展，特别是以大型语言模型（LLM）为代表的生成式AI技术展现出巨大潜力。中国AI公司DeepSeek（深度求索）通过创新使用强化学习技术，发布了一款出乎意料的高效且廉价的LLM并使用蒸馏技术（Distillation）使得蒸馏后的小模型（如Qwen、Llama系列）性能显著优于同类模型（如7B模型超越GPT-4o）。但近日，OpenAI公开表示他们注意到并正在审查DeepSeek可能不当地蒸馏他们模型或数据的迹象。然而，OpenAI自身也曾因不当获取未经授权的内容来构建ChatGPT而受到指控。

这些技术突破在为医疗领域带来巨大机遇的同时，也引发了诸多法律和伦理问题，尤其是在知识产权和数据合规方面。首先，关于AI技术的知识产权边界日益模糊。例如，用于训练这些高性能医疗AI模型的庞大数据集，其版权归属复杂，未经授权的使用可能引发侵权风险。以LLM为例，其训练数据可能来源于各种渠道，包括受版权保护的医学文献、临床记录等。如何界定这些数据的合理使用范围，以及在医疗AI模型训练中是否需要获得相应的授权，是亟待解决的法律问题。其次，医疗数据的合规使用是医疗AI应用的关键法律挑战。医疗数据的敏感性和隐私性要求极高，各国都出台了严格的数据保护法规。例如，欧盟的《通用数据保护

［收稿日期］ 2025-03-19　　［接受日期］ 2025-03-28

［作者简介］ 卢叶婷.

*通信作者（Corresponding author）. Tel：021-61187158，E-mail: luyeting@shalldolaw.com

条例》(GDPR)、美国《健康保险流通与责任法案》(HIPAA),中国的《中华人民共和国网络安全法》《中华人民共和国数据安全法》以及《个人信息保护法》等法律法规也对个人信息和敏感个人信息的处理进行了详细规定。医疗AI的开发和应用必须严格遵守这些法规,例如在数据使用前获得患者的充分知情同意,采取有效的匿名化和去标识化措施,确保数据安全,并满足跨境数据传输的合规要求。

因此,本文将以最新的大型语言模型技术结合强化学习和蒸馏技术在医学领域的发展为例,深入分析强化学习和知识蒸馏技术在医学领域的应用前景、挑战及合规路径,促进医疗AI的健康发展。

2 强化学习与知识蒸馏技术的基本原理与创新应用

强化学习通过奖励和惩罚机制训练智能体(Agent),使其在复杂环境中不断优化决策,目标是最大化长期回报。蒸馏技术将大型复杂模型(教师模型)的知识迁移到计算资源更为高效的小模型(学生模型)中,从而降低计算成本并提高模型的执行效率。这两种方法通常结合使用,共同提升学生模型的性能,特别是在资源受限的应用场景中,如医疗AI领域。

2.1 强化学习与知识蒸馏的技术原理 强化学习(reinforcement learning,RL)是指基于智能体(agent)通过与环境(environment)交互,基于奖励(reward)信号优化策略(policy),以最大化长期回报。智能体通过感知环境并执行动作,环境根据动作反馈奖励或惩罚,智能体依此更新策略(图1)。

强化学习的算法可以分为基于模型的强化学习(model-based RL)和无模型的强化学习(model-free RL),形成多层次的方法体系。无模型的强化学习又分为基于价值的强化学习(value-based RL)和基于策略的强化学习(policy-based RL)。在大型语言模型领域,基于策略的方法(如PPO算法)因其稳定性和可扩展性而被广泛采用(图2)。

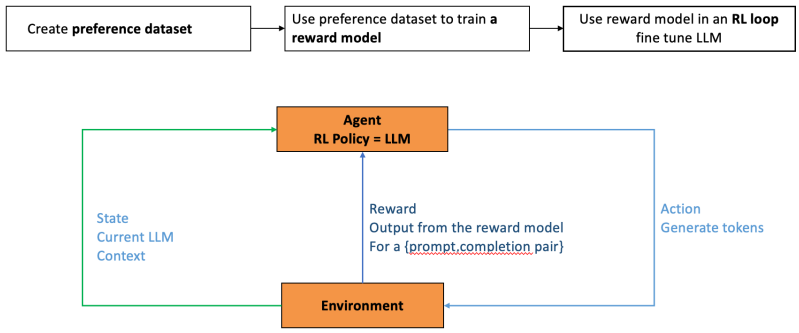

图1 强化学习流程示意图

RL:强化学习;LLM:大型语言模型。

在大模型训练实践中,蒸馏技术(Distillation)通过教师模型生成数据指导学生模型训练(如监督微调/SFT)。这类技术不仅限于传统软标签(Soft Labels)形式,还包括使用教师模型输出作为训练数据或辅助信号。这一技术使小型模型能够学习大型模型的决策能力,同时保持计算效率,为医疗等资源受限场景提供了可行解决方案。

2.2 Deepseek模型的技术创新 以Deepseek为例,DeepSeek-R1综合采用了强化学习和蒸馏技术等关键技术,实现了模型能力的突破性提升:(1)基于基座模型应用强化学习。训练流程首先基于基座模型(Base Model)直接应用强化学习,形成DeepSeek-R1-Zero,以探索模型在无监督微调(SFT)条件下自主进化推理能力的潜力。在RL训练过程中,涉及数据使用的主要环节在于创建偏好数据集(Preference Dataset)、奖励模型(Reward Model)和RL loop过程中的策略优化来提升模型性能(图3)。(2)蒸馏技术。Deepseek则通过蒸馏(Distillation)技术,将DeepSeek-R1的推理能力迁移至更轻量级的学生模型,使其在数学、编程和逻辑推理等任务上取得了显著进展(图4)。在AIME上,与OpenAI-o1-0912的性能相当,复刻OpenAI-o1的模型能力[20]。

这种技术路径不仅降低了计算资源需求,还为医疗AI的轻量化部署提供了可能,使先进AI能力在资源受限的医疗环境(如基层医院或便携医疗设备)中得到应用。

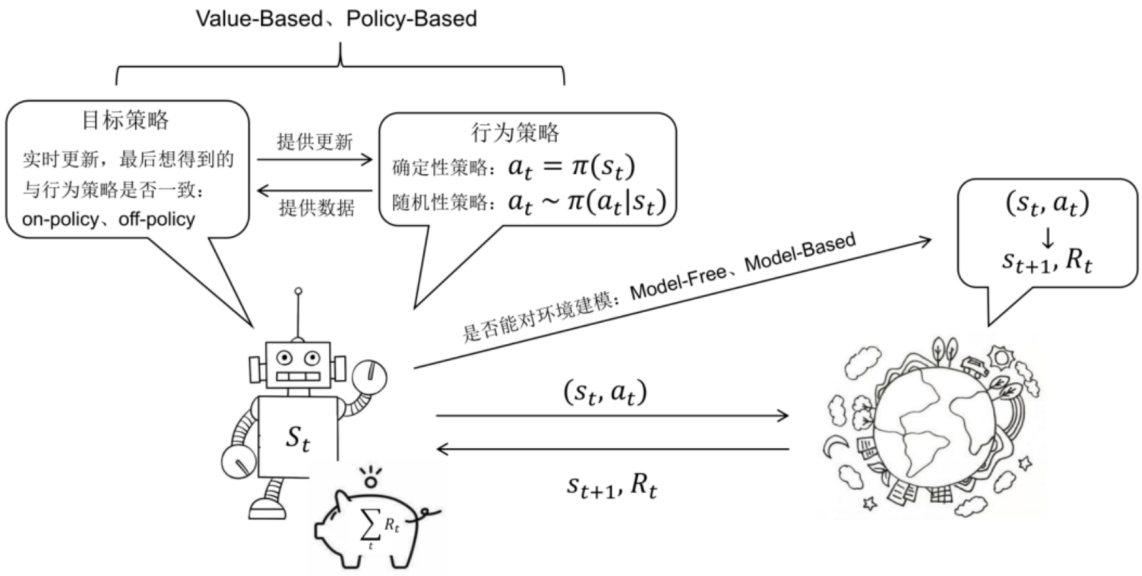

图 2　无模型的强化学习[1]

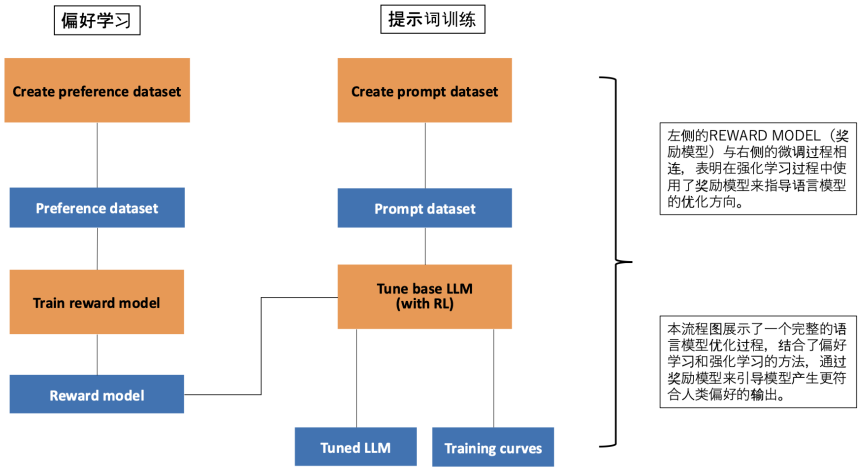

图 3　偏好学习和提示词训练结合强化学习（RL）优化语言模型（LLM）的训练过程

3　强化学习和蒸馏技术在医学领域中的应用

3.1　强化学习提高诊断决策　传统深度学习多采用监督学习，而强化学习允许模型通过“试错”和奖励机制不断改进决策策略，特别适合医学诊断这类复杂决策场景。以 DeepSeek 等大型语言模型为基础，结合强化学习可显著增强医学推理能力。

（1）精确诊断与风险评估：通过引入专家反馈作为“奖励信号”，让模型在反复试验中学会更准确地解读医学影像数据和各类检测指标（如基因表达、血液生化、生命体征等）。（2）优化随访与干预决策：在医疗诊断领域，强化学习可用于优化随访与干预决策流程。以肺癌诊断为例，研究人员将肺癌筛查过程建模为"最优停诊"问题：模型需要在每轮随访后决定是立即进行侵入性检查（如活检）确诊，还是等待下次影像随访以获取更多信息[3]。Yu 等人[4]的研究表明，强化学习驱动的决策模型可将不必要的侵入性检查减少 12.8%，同时保持诊断准确率不变。（3）个性化治疗方案生成：强化学习算法能够根据患者的病史、基因特征、治疗反应等因素，动态调整和优化治疗方案，实现精准医疗。例如，在糖尿病管理中，强化学习模型可根据患者的血糖波动、饮食、运动等实时数据，提供个性化胰岛素剂量建议，显著改善血糖控制效果。

这些应用表明，强化学习为垂类的医学大模型提供了持续改进诊断决策的机制，有望进一步降低疾病的误诊漏诊率，优化医疗资源分配。

3.2　模型微调与多模态适应　医疗场景的多样性和复杂性要求 AI 模型具备适应不同数据类型和临床环境的能力。通过知识蒸馏和微调技术，通用大模

型可以高效适应特定医学场景。(1)医学专业知识迁移:利用蒸馏技术,可将复杂高效的通用基座大模型的知识以及决策能力迁移到轻量级的垂类医学模型或将大型医学专家模型的专业知识迁移到轻量级医学模型中,使其在有限计算资源下也能提供专业医学决策支持。(2)多模态医学数据融合:在病理学领域,研究者开发了PathChat等多模态大模型,将预训练的视觉编码器与大型语言模型结合,利用大量病理图像-报告对进行联合微调,以理解和回应与病理学相关的复杂查询。PathChat在多种组织来源和疾病模型的多选诊断问题上展现出了最先进的性能,并在开放式问题和人类专家评估中表现出色[5]。(3)本地适应性调整:蒸馏和微调技术使模型能够进行本地化调整。Zhou等[6]研究提出,通过结合强化学习和微调技术,能够针对不同医院的影像设备差异和种族人群差异进行本地化调整,从而提高模型在特定临床环境下的鲁棒性和准确性。这些技术使AI模型能够在保留通用能力的同时,获得特定医学领域的专业能力,为实现AI辅助医疗的普及奠定基础。

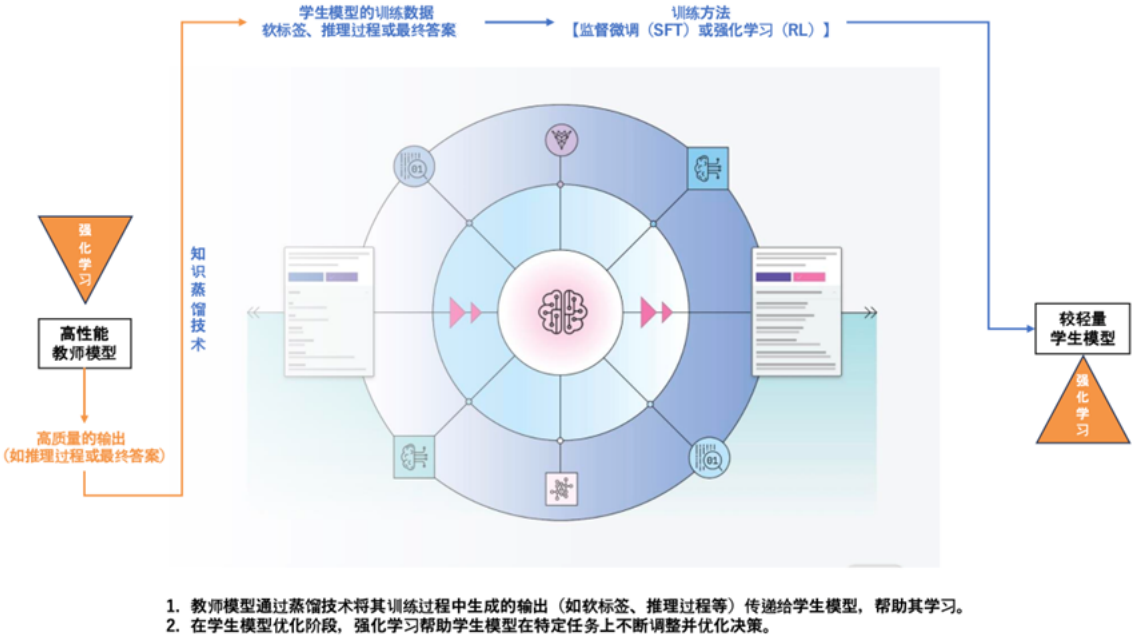

图4 强化学习和蒸馏技术结合的训练简化示意图[2]

4 伦理、法律及合规性考量

4.1 医疗AI的伦理原则与挑战 随着AI深入参与医疗决策,其引发的伦理问题日益受到关注,主要包括。(1)公平性与偏见防控:AI模型可能存在数据偏见问题。如果训练数据主要来自特定人群,模型在其他群体上的诊断准确性上可能下降,导致潜在的不公平诊断。MIT研究人员发现,人工智能模型在预测胸部X光图像中的种族和性别时,最准确的模型也显示出最大的"公平性差距"。这些模型在不同种族或性别的人群中诊断图像的准确性存在差异。研究还发现,这些模型可能在诊断评估中使用"人口统计捷径",导致女性、黑人和其他群体的错误结果[7]。(2)透明性与可解释性:医疗AI的决策直接影响患者生命健康,其"黑箱"特性可能导致医患对系统的信任危机,一方面医生无法验证AI诊断建议的可靠性(如肿瘤良恶性判断),可能引发医疗纠纷。另一方面,欧盟《人工智能法案》及中国《生成式人工智能服务管理办法》均要求高风险医疗AI必须具备可解释性,以符合责任追溯需求[8]。因此,当AI误诊时,难以界定开发者、医疗机构或算法的责任。(3)自主性与人类监督:虽然AI模型可以提供决策支持,但最终的临床判断应当由医疗专业人员做出。当前医疗AI部署需遵循"有意义的人类控制"(meaningful human control, MHC)原则(IEEE标准7010-2020),其核心要求包括:医生必须能在AI建议的任何阶段介入修正(如调整肿瘤分期判断阈值)和系统需记录医生与AI的交互痕迹(如修改次数、最终决策依据)[9]。

4.2 强化学习和蒸馏技术的法律风险分析

4.2.1 模型训练引发的法律风险 在实际应用强化学习(RL)时,奖励机制的设计和数据来源是关键因

素。在医疗场景中，如果训练数据存在偏差、噪声或不完整，或者奖励信号设置不当，可能会导致模型学习到错误的策略，甚至对患者造成伤害。此外，强化学习模型通常通过与环境的交互不断学习和改进，如果反馈循环中存在错误或偏差，可能会导致模型性能逐渐下降，甚至产生不可预测的风险。这在法律责任的认定上会变得更加复杂（表1）。

表1　强化学习奖励机制的法律风险分析

奖励机制	原理	数据来源	法律风险
基于规则判定奖励	奖励信号通过预设规则直接判定，如对输出的格式、逻辑结构或准确性进行评估。	使用标准答案的题目或任务，直接基于格式、结构或准确性评估模型输出。	使用标准答案的题目进行训练可能涉及版权问题，尤其是当这些答案受到版权保护时。
基于 Reward Model 的奖励	奖励来自专门训练的奖励模型，依赖于人类反馈或模拟数据，而非生成内容数据。	人类反馈（如问卷、众包平台）或模拟数据（通过其他模型生成的数据）来训练奖励模型。	奖励模型对源数据的依赖性较低，但仍可能涉及生成数据和闭源模型的版权或知识产权问题，尤其是未经授权的闭源数据使用。

医疗领域的强化学习面临额外挑战，因为涉及受保护的病人数据和医学知识。从法律角度看，使用病例数据训练RL模型需要确保：（1）数据已去标识化或匿名化处理；（2）获得伦理委员会批准；（3）符合法律规定的数据使用目的限制原则。

4.2.2　*数据使用引发的法律风险*　构建高性能的医疗AI模型，特别是像医学LLM这样的复杂系统，需要海量的数据进行训练。这些数据可能包括医学文献、临床记录、影像资料等，其中大部分可能受到版权保护。未经授权地使用这些数据进行模型训练，可能构成对版权所有者的侵权。例如，医学研究论文的版权通常归属于出版社或作者，若未经许可将其用于模型训练，则可能侵犯其复制权和信息网络传播权。（1）知识产权边界：教师模型的输出数据不仅用于直接训练学生模型，还可能被用于构建reward model或数据标注。在开源模型场景下，如基于Apache 2.0或MIT许可的模型（如Llama 2），这类使用通常被视为是合规的。然而，使用闭源模型（如专有医学模型或ChatGPT系列）的输出数据进行训练，可能违反服务协议或触发知识产权保护条款。（2）中间态模型使用的法律定性：将他人模型作为奖励模型（reward model）或评论模型（critic model）用于辅助训练的行为，其法律性质存在争议。核心问题在于此类使用是否构成对原始模型的“实质性复制”（表2）。

表2　中间态模型使用过程中的法律问题分析

关键要素	违反法律	条款核心	具体行为与法律条款对应
模型功能专有权益	《著作权法》	保护著作权人对作品的人身权和财产权。	若功能相似性被认定为对模型核心逻辑的“实质性复制”，可能侵犯《著作权法》第10条作者对复制权、改编权的权利。
数据与模型商业秘密	《反不正当竞争法》	禁止以不正当手段获取他人商业秘密	若该模型系商业秘密，未经授权使用他人模型输出数据训练自身模型的行为，可能构成反法第9条“侵犯商业秘密”。
合同约定权益	《民法典》	违反与原始模型开发者之间的许可协议（如限制用途条款）构成违约行为。	如果企业未遵循与原开发者的合同约定（如模型使用限制或未经授权使用数据），构成《民法典》第577条“违约行为”。
技术成果公平竞争权	《反不正当竞争法》	经营者应遵循自愿、平等、公平、诚信原则，禁止不正当手段竞争。	利用他人模型功能提升自身竞争力，则可能被视为违反商业道德及诚实信用原则。

在医疗领域，这一问题更为敏感，因为医学AI模型通常包含宝贵的专业知识和临床经验。使用知名医学模型的输出作为训练数据，可能构成对原始开发者知识产权和商业利益的侵害。

4.2.3 数据标注的授权边界 借用他人模型输出的生成数据进行标注与常规的数据标注方法的主要区别在于数据的来源和生成方式。若使用蒸馏技术，借用他人模型输出的生成数据进行标注，意味着依赖一个已有的模型（如教师模型）生成数据标签，这些标签并非直接来源于原始数据，而是通过模型的推理或输出产生的。这种区别引发了数据使用的知识产权边界问题，原因如下。(1)模型输出的知识产权：如果使用他人模型生成的数据进行标注，该数据实际上可能是基于他人模型的算法、架构和训练数据所产生的结果。虽然模型的输出不一定受版权保护，但其背后涉及的算法和训练方法可能受到版权保护或商业秘密保护。未经授权使用他人模型的输出，可能构成对其知识产权的侵害。(2)授权和使用范围：借用他人模型的输出时，需要确保使用行为在原始授权范围内。如果在未授权情况下或超出授权范围使用，可能侵犯模型提供者的版权或商业秘密。若输出数据涉及商业利益或竞争优势，可能还会引发不正当竞争等法律问题。(3)衍生性使用与复制问题：通过模型生成的数据进行标注，可能会被视为对原始模型的衍生性使用。如果这种使用在实质性上复制了模型的核心功能或获得了不正当的竞争优势，可能会面临法律的规制，特别是在不正当竞争法框架下（表3）。

表3 数据标注的授权边界

法律问题	适用法律	条款核心	具体行为与法律条款对应
数据标注的授权边界	《著作权法》	确保数据标注行为不超出原模型授权范围，尤其是在闭源模型的情况下。	若企业未经授权使用闭源模型的输出数据进行数据标注，可能触及《著作权法》中的复制权，构成对原模型开发者权益的侵犯。
开源模型的标注行为	《著作权法》《民法典》	开源模型的授权范围通常较宽。	在模型开源许可证的授权范围内，训练方被允许衍生使用（如Llama 2的Apache 2.0协议）。但需要关注开源模型的具体许可协议，若超出授权范围，则可能承担违约责任[18, 19]。
利用闭源模型输出的“搭便车”行为	《反不正当竞争法》	禁止通过不正当手段获取竞争优势，特别是在存在市场竞争关系的情况下。	低成本复用闭源模型输出数据并快速抢占市场份额，可能被认定为“搭便车”行为，违反《反不正当竞争法》第2条（诚信原则），并可能构成第6条的混淆行为。
教师模型输出的竞争影响	《反不正当竞争法》	若未经授权使用教师模型的输出并获得不当竞争优势，可按不正当竞争行为进行规制。	即使教师模型的输出不受版权保护，若衍生模型通过未经授权的输出数据获得不正当竞争优势，并对教师模型方造成实际损害，可能违反《反不正当竞争法》第2条（诚信原则），并构成第9条侵犯商业秘密行为。

4.3 医疗AI模型创新技术的合规分析

强化学习和蒸馏技术涉及复杂的数学模型和计算过程，这些过程往往难以直观理解和追溯。一方面，决策过程高度依赖于训练数据，数据的复杂性和多样性使得算法的行为更加难以解释。另一方面，强化学习中的决策过程是动态的，随着环境变化不断调整，这增加了对算法决策过程的解释难度（表4）[15-16]。

4.3.1 合规策略 面对上述法律风险，医疗AI开发者可采取以下合规策略：(1)知情同意和数据使用边界。医院在部署DeepSeek等大模型系统时，应将其作为诊疗流程的一部分向患者披露，取得患者的知情同意。这不仅涉及法律义务，也是对患者自主权的尊重。同时，模型训练所用的患者数据归属也需明确：通常患者数据由医院托管，第三方AI公司在使用时仅获得经过授权的“使用权”，不得将数据挪作他用或商业变现[14, 16]。(2)透明的授权机制：建立清晰的数据授权机制和模型使用协议，明确指出数据使用范围、目的和期限。特别是在使用开源模型或数据时，应详细审查相关许可证条款，确保遵守所有限制条件[15]。(3)遵循监管框架：在使用医疗AI系统时，必须严格遵守《药品监管人工智能典型应用场景清单》等监管框架和相关法律法规要求。具体而言，如果需要将模型输出用于临床决策，医院应先对模型进行充分验证，确保其性能达到临床应用标准，并建立健全的审核机制，对模型的每次

输出进行严密把关。针对AI辅助诊断过程中发生的事故责任划分问题，目前尚处于讨论阶段。尽管普遍认为在现行监管体系下，AI仅作为辅助工具，最终的诊断决策仍由执业医师负责，但随着AI系统自主性不断提升，未来法律可能需要进一步明确厂商、医院与医生在责任认定上的具体界定，以保障患者安全和医疗实践的合法性[10, 21]。

表4　算法数据使用和生成的合规性分析[8]

法律问题	适用法律	法律与技术	具体行为与法律条款对应	备注
技术合规性问题	《著作权法》《反不正当竞争法》	若学生模型在功能实现层面与教师模型形成实质性近似（substantial similarity），可能触发技术成果独创性认定争议。	可能构成《著作权法》第九条规定的演绎作品侵权风险，或违反《反不正当竞争法》关于技术成果保护条款。	需注意技术实现中的“避风港原则”适用性，以及是否属于合理使用范畴。
算法透明度问题	《著作权法》《个人信息保护法》	双重黑箱效应，导致算法决策路径难以逆向解析。这种技术不可解释性可能形成系统性合规漏洞。	违反《个人信息保护法》第24条算法决策解释权要求；可能触发《消费者权益保护法》第8条（知情权）和第20条（公平交易）。	Wachter et al. (2017)算法问责制研究；欧盟《人工智能法案》中对通用人工智能模型的规定。

4.3.2　三阶段合规审查流程　（1）数据源合规审查：建立数据谱系追踪系统（provenance tracking）。区分训练数据来源（患者授权/公共数据库/合成数据）。（2）训练过程合规控制：开源模型：遵守Apache 2.0等许可证约束条款。闭源模型：通过API日志审计确保合规使用。（3）模型部署合规验证：执行《医疗器械软件注册审查指导原则》全生命周期测试。嵌入可解释性模块（如LIME解释器）满足临床透明性要求[11]。

在引入DeepSeek等AI时，必须严格遵循医疗伦理准则和法律规范，通过制定明确的合规审查流程来约束AI的使用，保护患者权益和隐私。在强调技术创新的同时，确保“以人为本”的价值不被忽视。

4.3.3　基于区块链技术的医疗AI合规架构　针对医疗AI，特别是涉及强化学习的复杂模型，其数据溯源、模型验证和合规性监控面临诸多挑战。为了应对这些挑战，一种基于区块链的医疗AI合规架构被提出，该架构通过去中心化审计方案，旨在增强透明度、可信度和合规性。其主要组成部分包括：（1）数据溯源：利用区块链的不可篡改性，将训练数据哈希值写入以太坊智能合约，实现不可篡改记录[22]。（2）模型验证：通过零知识证明（zk-SNARKs）验证模型未使用侵权数据，同时保护商业秘密。这对于解决强化学习模型可能无意间学习到来自非法数据集的策略具有重要意义[22-23]。（3）合规自动化：为了实时监控医疗AI系统的数据使用是否符合相关的法律法规，例如PIPL或HIPAA或GDPR，该架构建议部署合规性检查链码（Chaincode）在区块链上。链码可以被编程为自动检查数据的使用是否符合预设的合规规则，例如访问权限控制、数据脱敏处理等[22-23]。

5　展　望

大模型如DeepSeek往往在开发阶段利用了高质量的数据和算力，但在面对不同医院的成像设备、医疗设备检测水平、不同人种的患者特征时，性能可能下降。如果训练数据不足以涵盖各种变异，模型在新环境下可能出现意外错误。因此，确保多样化的数据来源、在部署前进行本地验证和适应性调优非常重要[12]。然而，我们也必须谨慎推进，确保在隐私、安全、伦理合规方面构筑牢固的基础，同时不断完善模型的解释能力和鲁棒性。可以预见，随着医疗AI领域技术和监管的成熟，像DeepSeek这样的平台将成为临床医生的有力助手，共同揭开疾病诊断的新篇章，最终实现患者诊疗、医疗技术提升和公共健康的协同进步[13]。

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　卢叶婷：选题、撰写、修改论文；杨宇平：撰写、修改论文。

参考文献

[1]　CHRISTIANO P, LEIKE J, BROWN T, et al. Deep reinforcement learning from human preferences. In Proceedings

of the 31st International Conference on Neural Information Processing Systems (NIPS' 17)[C]. Curran Associates Inc., Red Hook, NY, USA, 2017: 4302-4310.

[2] IBM《什么是迁移学习》[EB/OL]. [2025-03-20], *https://www.ibm.com/cn-zh/topics/transfer-learning*.

[3] WANG Y F, ZHANG Q N, YING L, et al. Deep reinforcement learning for early diagnosis of lung cancer[J]. Proc AAAI Conf Artif Intell, 2024, 38(20): 22410-22419.

[4] YU C, LIU J M, NEMATI S, et al. Reinforcement learning in healthcare: a survey[J]. ACM Comput Surv, 2023, 55(1): 1-36.

[5] CHEN H Z, LIN R C, YU Y F. MetaPath Chat: multimodal generative artificial intelligence chatbot for clinical pathology [J]. MedComm (2020), 2024, 5(10): e769.

[6] PFEFFERLE A, PURUCKER L, HUTTER F. DAFT: data-aware fine-tuning of foundation models for efficient and effective medical image segmentation[M]//Medical Image Segmentation Foundation Models. CVPR 2024 Challenge: Segment Anything in Medical Images on Laptop. Cham: Springer Nature Switzerland, 2025: 15-38.

[7] TRAFTON A, Study reveals why AI models that analyze medical images can be biased - MIT EECS[EB/OL]. [2025-03-10]. *https://www.eecs.mit.edu/study-reveals-why-ai-models-that-analyze-medical-images-can-be-biased/*

[8] 《欧盟人工智能法案》[EB/OL]. [2025-03-10]. *https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689*.

[9] SCHIFF D, AYESH A, MUSIKANSKI L, et al. IEEE 7010: a new standard for assessing the well-being implications of artificial intelligence[C]//2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). October 11-14, 2020. Toronto, ON, Canada. IEEE, 2020: 2746-2753.

[10] 张建楠, 李莹莹, 周佳卉, 等. 人工智能独立医用软件监管研究[J]. 中国工程科学, 2022, 24(1): 198-204.

[11] JOHNS M, BAUM L, PRASSER F. Tracking provenance in clinical data warehouses for quality management[J]. Int J Med Inform, 2025, 193: 105690.

[12] 2025医疗AI算法临床验证指南[EB/OL]. [2025-03-12]. *https://editverse.com/zh-CN/%E5%8C%BB%E7%96%97%E9%AA%8C%E8%AF%81-2025/*.

[13] SCHREUDER A, SCHOLTEN E, GINNEKENET B al. Artificial intelligence for detection and characterization of pulmonary nodules in lung cancer CT screening: ready for practice?[J]. Translational lung cancer research, 2021(10)5: 2378-2388.

[14] HUTCHINSON B, SMART A, HANNA A, et al. Towards accountability for machine learning datasets: practices from software engineering and infrastructure[C]//Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Virtual Event Canada. ACM, 2021: 560-575.

[15] LIPTON Z C. The mythos of model interpretability[EB/OL]. [2025-03-15]. 2016: 1606.03490. *https://arxiv.org/abs/1606.03490v3*.

[16] DOSHI-VELEZ F, KIM B. Towards a rigorous science of interpretable machine learning [EB/OL]. [2025-03-15]. 2017: 1702.08608. *https://arxiv.org/abs/1702.08608v2*.

[17] LIPTON Z C. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery.[J]. Queue, 2018, 16(3): 31-57.

[18] Apache 2.0 License [EB/OL]. [2025-03-10]. *https://www.apache.org/licenses/LICENSE-2.0*

[19] MIT License [EB/OL]. [2025-03-10]. *https://opensource.org/licenses/MIT*.

[20] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. [2025-03-15]. 2015: 1503.02531. *https://arxiv.org/abs/1503.02531v1*.

[21] TOPOL E J. High-performance medicine: the convergence of human and artificial intelligence[J]. Nat Med, 2019, 25(1): 44-56.

[22] MUNYUZANGABO M, KHALIFA D S, GAFFEY M F, et al. Delivery of sexual and reproductive health interventions in conflict settings: a systematic review[J]. BMJ Glob Health, 2020, 5(Suppl 1): e002206.

[23] JHA P K, AKBARI H, KIM Y, et al. Nanoscale axial position and orientation measurement of hexagonal boron nitride quantum emitters using a tunable nanophotonic environment [J]. Nanotechnology, 2021, 33(1): Nanotechnologyvol. 33.

引用本文

卢叶婷,杨宇平.人工智能医学应用中的技术路径与法律边界——以强化学习与蒸馏技术为视角[J].元宇宙医学,2025,2(1):57-64.

LU Y T, YANG Y P. Technical path and legal boundary in the medical application of artificial intelligence: from the perspective of reinforcement learning and distillation technology[J]. Metaverse Med, 2025, 2(1): 57-64.