大语言模型驱动的电子病历智能：应用、挑战与展望

赵 霞[1] 赵尔康[2] 程春雷[2] 李小华[1] 张海波[1] 姚佳璇[1]

（[1] 中国人民解放军南部战区总医院 广州 510010
[2] 江西中医药大学智能医学与信息工程学院 南昌 330004）

〔摘要〕 **目的/意义** 系统梳理大语言模型在电子病历领域的关键技术、应用现状与未来挑战。**方法/过程** 总结预训练、微调与检索增强生成等核心技术；分析信息结构化、临床文本生成与临床决策支持 3 大应用范式；剖析事实一致性、知识动态性等核心挑战。**结果/结论** 大语言模型在电子病历领域已形成 3 大核心应用范式，但规模化应用仍受技术与伦理限制。未来应向可信推理、自主智能体及协作网络生态 3 个方向演进，以释放智慧医疗潜力。

〔关键词〕 大语言模型；电子病历；临床决策支持；检索增强生成
〔中图分类号〕 R－058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2026.01.004

Large Language Model－driven Intelligence in Electronic Medical Records: Applications, Challenges and Prospects

ZHAO Xia[1], *ZHAO Erkang*[2], *CHENG Chunlei*[2], *LI Xiaohua*[1], *ZHANG Haibo*[1], *YAO Jiaxuan*[1]

[1] *General Hospital of Southern Theatre Command, PLA, Guangzhou 510010, China*; [2] *College of Intelligent Medicine and Information Engineering, Jiangxi University of Chinese Medicine, Nanchang 330004, China*

〔**Abstract**〕 **Purpose/Significance** To systematically review the key technologies, applications, and challenges of large language models (LLM) in the electronic medical record (EMR) domain. **Method/Process** The core technologies including pre－training, fine－tuning, and retrieval－augmented generation are summarized; three application paradigms such as information structuring, clinical text generation, and clinical decision support are analyzed; and the core challenges such as factual consistency and knowledge timeliness are analyzed. **Result/Conclusion** Three core application paradigms have been established for LLM in the EMR domain, but scalable adoption remains constrained by technical and ethical limitations. Future evolution should shifts towards trustworthy reasoning, autonomous agents, and collaborative ecosystems to unlock the potential of intelligent healthcare.

〔**Keywords**〕 large language model; electronic medical record; clinical decision support; retrieval－augmented generation

〔修回日期〕 2026－01－11

〔作者简介〕 赵霞，博士，发表论文 23 篇；通信作者：程春雷，博士。

〔基金项目〕 江西省自然科学基金项目（项目编号：20224BAB206102）；广东省医学装备学会科研基金项目（项目编号：YZXH2025KT02）。

1 引言

电子病历（electronic medical record，EMR）是驱动临床决策与医学研究的核心数字资产[1]。相关研究已从“数字化存档”向以自然语言处理（natural language processing，NLP）为核心的“智能化应用”演进，国内研究侧重利用 NLP 与实体识别技术挖掘数据价值，国际研究更聚焦风险控制与临床路径等决策支持应用。由于 EMR 中含有高度融合的非结构化文本、噪声及不规则时序信息，基于人工智能（artificial intelligence，AI）和深度学习挖掘其数据价值仍面临巨大挑战。早期机器学习方法以及后来的预训练模型（如 Med－BERT[2]、R－BERT[3]），均通过自监督学习掌握丰富的语言知识，但在具体临床任务微调阶段，仍难摆脱对大量精细标注数据的依赖，且受限于上下文长度，其在真实世界复杂场景下的应用效果受限。即便是深度学习模型，相较于传统机器学习方法，在此类数据上往往也只能取得有限的改进效果[4]。得益于海量通用语料训练[5]，大语言模型（large language model，LLM）凭借其少样本学习与多步逻辑推理能力，为处理 EMR 中的非标准化叙事文本提供了颠覆性范式，能够胜任从病历结构化到复杂诊疗建议的全链条任务。

本综述旨在系统梳理 LLM 在 EMR 领域的应用进展。首先分析 LLM 在医疗场景下的关键技术革新；其次探讨信息结构化、临床文本生成与临床决策支持 3 大核心应用范式；最后剖析事实一致性（“幻觉”）[6]、隐私合规等挑战，并对未来的医疗智能体（agent）与协作生态进行展望。

2 面向 EMR 的 LLM 适配技术

2.1 LLM 简介与技术演进

LLM 的颠覆性在于其无需领域特化训练即可展现出专家级的知识水平。以 GPT－4 和 Med－PaLM 2[7]为代表的顶尖模型，在美国医师执照考试[8]等高难度测试中准确率已超过 80%，显著优于早期经过专项微调的模型[9]。同时，其在医患沟通草稿撰写等真实临床任务中也展现出巨大的效率提升价值。

这一能力跃升，得益于 Transformer 架构[10]的持续演进。自然语言处理范式在结构上经历了从仅编码器到编码器－解码器，再到仅解码器架构的更迭，模型也由早期的 BERT、T5 等，逐步过渡至 GPT、LLaMA[11－13]等超大规模生成式基座。国内文心一言、通义千问和 DeepSeek 等优秀基座模型，在中文语境理解与医学逻辑推理上，展现出媲美国际顶尖水平的能力。这些生成式基座的成熟，为解决高复杂度 EMR 任务奠定了技术基础。

2.2 典型数据源

当前研究主要依赖两类数据源：一类是以 MIMIC、CBLUE 为代表的综合性数据库，凭借多模态与长时序特性成为复杂推理的验证标准；另一类是以 i2b2、CHIP、CCKS 为代表的特定任务数据集，侧重实体识别与关系抽取的精度评测，见表 1。然而，现有资源仍存在中英文分布不均、多轮对话数据稀缺等局限。

表 1 电子病历领域代表性数据集与评测基准

数据集名称	发布机构/来源	语言	类型/主要任务	特点/贡献
MIMIC－IV	麻省理工学院计算生理学实验室	英文	ICU 全景数据（文本、表格、时序）	全球使用最广的 ICU 多模态脱敏数据库，支持复杂决策研究
i2b2 Challenges	哈佛医学院生物医学信息中心	英文	NLP 任务（NER、RE、摘要等）	算法评测的国际金标准，标注质量极高
CBLUE	中国中文信息学会/阿里云天池	中文	综合评测基准（含分类、NER、RE、标准化以及光学字符识别等）	中文医疗 NLP 基准数据集，聚合多源数据，统一评测标准
CHIP	中国中文信息学会医疗健康与生物信息处理专业委员会	中文	术语标准化、文本结构化	国内规模最大的中文电子健康记录公开资源
CCKS[14－18]	中国中文信息学会语言与知识计算专业委员会	中文	知识图谱、医疗问答	侧重知识推理与语义计算

2.3 面向 EMR 的领域适配与增强技术

2.3.1 领域知识注入：增量预训练 针对电子病历中非标准缩写与“电报式”语言导致的高困惑度问题，增量预训练[19]成为标准范式。其在通用基座上利用 PubMed 文献、脱敏病历及医学教材进行二次训练，例如，BioMedGPT[20]与 ClinicalGPT[21]通过混合语料训练，为模型注入领域知识，能够准确解析“T2DM”（2 型糖尿病）、“3 +”阳性程度等专业表述的隐含语义。

2.3.2 临床指令对齐与模型轻量化 预训练初步赋予模型“医学知识”，为了弥补预训练知识通用、静态的不足，利用特定医学数据进行微调，以学习特定领域的医学知识，生成医学领域专用 LLM[22]，见表 2。为解决高质量医疗标注数据稀缺难题，知识蒸馏技术[23]被广泛采用，即利用强大的 GPT - 4 等通用大模型作为“教师”生成合成数据，指导小模型训练，典型应用包括 HuatuoGPT - II[24]、DeepSeek - R1 - 0528 - Qwen3：前者利用 ChatGPT 生成的合成数据解决冷启动问题，后者通过思维链（chain - of - thought，CoT）蒸馏路径实现复杂医疗推理能力迁移，使小模型具备类似医生的思考逻辑。此外，参数高效微调（parameter - efficient fine - tuning，PEFT）技术（如 LoRA、QLoRA[25]）结合蒸馏，使得在消费级显卡上训练 7B 至 13B 规模的轻量化模型成为可能，显著降低了 EMR 智能化的院内部署门槛。

表 2 代表性医学大语言模型及其训练范式

模型	基座模型	参数规模	核心数据源	微调方法
ChatDoctor	LLaMA	7B	110k 真实医患对话[26]	指令微调/监督微调
MedAlpaca	LLaMA	7B/13B	160k 医学问答对[27]	指令微调/监督微调
CPLLM	LLaMA	13B	109k 临床电子病历（MIMIC - IV）	全量微调
MedPaLM - 2	PaLM 2	340B	大规模医学问答（MultiMedQA）	指令提示微调/专家集成
HuatuoGPT - II	Baichuan/LLaMA	7B/13B/34B	混合数据（ChatGPT 生成数据 + 真实问答 + KG）	一步式领域适配/知识蒸馏
DeepSeek - R1 - 0528 - Qwen3	Qwen3	8B	推理思维链数据（DeepSeek - R1 - 0528 合成 CoT）	基于 DeepSeek - R1 - 0528 的思维链蒸馏

2.3.3 推理增强与事实核查：思维链与检索增强生成 提示工程指设计语言模型的引导语来获取更符合用户需求的信息的过程。通过这种设计，无需调整模型参数即可提升性能，使模型更易操作且更具推广价值[28]。在逻辑增强方面，提示工程已从简单的模板填充演进为 CoT[29]诱导。通过在提示词（prompt）中展示“主诉 - 查体 - 诊断”的推导步骤，CoT 能够激活模型的逻辑推理能力，显著提升其在疑难病例鉴别诊断中的准确性。在事实核查方面，检索增强生成（retrieval - augmented generation，RAG）[30]已成为解决模型“幻觉”的标准配置。通过将生成过程约束在最新指南或患者历史数据的证据框架内，有效规避了捏造药名或剂量等风险。为突破传统向量检索在多步推理上的局限，RAG 正向着结构化与智能化演进：Graph RAG[31]通过融合知识图谱捕捉疾病 - 药物间的深层因果关联；Light RAG 与 Adaptive RAG[32]引入按需检索机制优化推理效率。这些进阶范式结合多模态检索，正在重构复杂场景下的临床决策支持能力。

3 面向电子病历的 LLM 应用分析

3.1 从数据处理到临床决策

LLM 在 EMR 领域的应用呈现显著的“性能的二元性”。以 GPT - 4 为代表的通用大模型在开放式问答上表现卓越，但在追求精确边界的开放式、判别式任务中，如命名实体识别（named entity recognition，NER）、关系抽取（relation extraction，RE）等，往

往不及经过全量微调的专用小模型（如 BioBERT）。因此形成“专用模型”负责底层结构化，通用模型负责上层决策的协作架构，具体包括信息结构化、文本生成与决策支持 3 大范式。

3.1.1 范式 1：临床信息结构化 临床信息结构化旨在将非结构化的叙事文本转化为机器可读的结构化数据，是构建医学知识图谱（knowledge graph，KG）[33]的基石。其准确性直接决定整个 EMR 智能体系的上限。临床信息结构化的核心任务精准对应 KG 的构建流程，涵盖医学 NER、RE、事件抽取（event extraction，EE）等。然而，临床文本中普遍存在的非标准缩写、复杂的嵌套实体及不确定性表达，使其面临比通用领域更严峻的挑战。以最具代表性的 NER 任务为例，其技术路径演进，见表 3。该领域正经历从“特征工程”到“深度表示”再到“生成式抽取”的范式变迁。既往基于 Transformer 架构的预训练模型凭借强大的上下文表征能力，通过“序列标注”方式确立了其在该任务中的 SOTA 地位，特别是在高资源场景下具有极高的边界识别精度。随着生成式大模型的兴起，“基于指令的抽取”成为新趋势。不同于 BERT 的字符级分类，LLM 通过少样本提示可直接生成 JSON 格式的实体列表。虽然 BERT 在严格匹配指标上仍占优势，但 LLM 在低资源场景和开放式信息抽取中展现了惊人的泛化能力，为解决长尾医疗实体的结构化问题提供了新思路。

表 3 命名实体识别国内外研究现状

方向	方法	提出者	模型（架构）	特点/贡献
基于判别式模型的命名实体识别	引入字典信息	Zhang Y 等	树神经网络（Tree－LSTM）	开创性工作
		Sui D 等	协作图网络（CGN）	解决信息缺失问题
		Gui T 等	基于词典的图神经网络（LGN）	解决标注冲突问题
		Li X 等	扁平化格结构（FLAT）	支持长距离依赖
		Lewis P 等[30]	检索增强生成（RAG）	解决知识滞后与“幻觉”问题
	引入字形结构信息	Sun Z 等	中文 BERT（ChineseBERT）	多模态特征融合
		Hou Y 等	层级化命名实体识别（HiNER）	多类型实体处理
		Ning Q 等	多模态命名实体识别（IMNER）	多模态统一处理
	基于预训练与模型结构	AIsentzer E 等[34]	生物临床 BERT（BioClinical－BERT）	临床 NER 的经典 SOTA，高精度基座
		Zeng Z 等[35]	全局指针网络（GlobalPointer）	统一框架，高效解决嵌套实体问题
	基于对比学习的模型	Yu H 等	协作图对比学习（CGCL）	解决边界模糊问题
		Wang X 等	模板分类（TC）	支持嵌套实体
		Zhu Y 等	平面跨度对比学习	嵌套实体识别
基于生成式 LLM 的抽取	LoRA＋微调技术	Zhu Y 等[36]	Llama－3 指令微调版	有效区分细微实体
	LLM 自验证策略	Bian J 等[37]	GPT－NER 架构	资源匮乏和样本量较小的场景下表现出更强的能力

3.1.2 范式 2：临床文本生成与工作流程自动化 文本生成旨在利用 LLM 将医生从繁重的文书工作中解放出来，涉及病历生成、摘要生成与病历质控 3 大核心环节。病历内容生成经历了从被动记录到环境智能与主动协同的过程。传统病历录入依赖语音转写或模板填充，本质上仍是机械的信息搬运。LLM 的引入催生了“环境临床智能”[38]范式：模型能够实时理解复杂的医患对话，自动过滤闲聊噪音，并提取关键诊疗信息生成结构化病历。随着技术演进，应用模式正向着“临床智能体”跨越，模型不仅记录信息，更能通过工具调用，主动检索历史画像、提示鉴别诊断并

草拟医嘱，实现从“效率工具”到“协同伙伴”的质变。摘要旨在解决海量病历带来的信息过载问题。当前技术路径主要面临专用小模型微调（BART）与通用大模型提示（DeepSeek－R1、GPT－4）的权衡难题：前者通过在 BART、T5 等中等模型上进行全量微调，适应特定专业领域摘要生成任务[39－40]，成本低且格式可控，但依赖标注数据；后者利用 LLM 的少样本学习能力[41]，语义概括力强，但面临事实一致性挑战。因此，结合 RAG 进行事实核查，确保摘要的临床忠实度，是当前该领域的关注焦点。传统病历质控依赖规则匹配或 Text－CNN 等浅层深度学习模型，仅能解决必填项缺失等形式问题。难以处理跨段落的逻辑矛盾（如现病史与体格检查冲突），而 LLM 基于医疗知识学习，通过自然语言理解分析病历内容，模拟专家逻辑判断，可提升质控智能化水平[42]。LLM 的引入实现了从“形式”到“内涵”的质变。陈冯薏等[43]构建的儿科专科大模型，在复杂逻辑校验上 *F*1 达 0.86，显著优于传统 NLP 方法。该智能流程深度融入“采集－理解－比对－生成”业务闭环，将质控从滞后的“终末抽检”转变为实时的“环节干预”，对提升医疗质量具有革命性意义。

3.1.3 范式 3：智能推理与决策支持 LLM 应用的最高阶价值，在于构建新一代临床决策辅助系统（clinical decision support system，CDSS）[44]。该范式通过“神经符号协同”机制，实现了从数据洞察到辅助决策的跃迁。其架构基石是多模态数据语义计算。系统底层对应范式 1 的核心能力，利用 LLM 将非结构化的 EMR 文本、影像报告等映射为机器可理解的实体，并对齐至国际疾病分类等权威术语标准，将离散的原始数据转化为可计算的临床证据。该范式的核心引擎是图谱与模型的神经符号协同。突破传统规则库的局限，新架构融合了知识图谱（符号）的精确逻辑和 LLM（神经）的泛化推理：图谱提供精确、可解释的结构化医学知识（如疾病－症状－药物关系）；LLM 结合 RAG 技术执行多步跳跃推理，二者结合既保留了医疗逻辑的严谨性，又赋予了系统处理未见病例的灵活性。该范式的应用出口是可解释的临床洞见。例如，生成带有证据溯源的鉴别诊断列表或个性化治疗方案。这种“推理＋生成”的闭环，使 CDSS 真正进化为懂逻辑、能解释的“AI 临床助理”。

3.2 多维度的评估体系

电子病历领域的模型评估通常包括判别式任务（信息抽取）与生成式任务（文本生成）两个维度。然而，随着 LLM 在临床场景中的应用日益深入，单纯采用技术评估指标已难以满足临床实践对安全性的要求，促使评估维度向临床价值迁移。

3.2.1 信息抽取评估 对于 NER 和 RE 任务，学界普遍采用 *F*1 值作为核心指标，衡量模型预测结果在精确率与召回率之间的平衡性能。在 EMR 场景下，通常要求极其严格的边界匹配：只有当实体的起止位置与类型完全一致时，才被判定为正确。现有模型在标准数据集上已取得显著进展，基于结构增强的 GlobalPointer[35]在通用中文数据集 CLUENER 上 *F*1 达 0.796 6；得益于领域预训练，BioClinical－BERT[34]在生物医学 NER 任务中 *F*1 达到 0.91 的高分。在更复杂的因果关系抽取中，R－BERT 等在权威基准上 *F*1 达 0.892 5[45]，金方焱等[46]的融合模型在 SemEval－2010 Task 8[47]和金融数据集上，*F*1 值分别达到 0.729 8 和 0.757 4；黄俏娟等[48]提出的模型甚至在因果三元组抽取上 *F*1 达到 0.924 4。

3.2.2 临床文本生成评估 对于生成任务，ROUGE 体系是评测主流。不同范式在医患对话数据 IMCS－V2－MRG[49]上的表现，见表 4。数据揭示了关键现象：参数规模并非唯一决定因素，国产开源模型 Qwen 在微调后表现优于未微调的 GPT－4；采用“大小模型协同”的复合架构（REFLEX-ES）[41]取得了最佳效果。此外，GPT－4 与 BART 相近的 ROUGE 分数掩盖了二者在逻辑推理上的本质差异，进一步印证了引入 BertScore（语义相似度）及人工核查的重要性。

表 4 不同模型在 IMCS－V2－MRG 数据集上的性能对比

模型	方法	ROUGE－1（%）	ROUGE－2（%）	BertScore（%）	性能分析
BART＋FT	传统预训练＋全量微调	51.13	32.58	75.48	传统基线。受限于规模，语义理解上限较低
Qwen1.5－7B＋SFT	开源 LLM（7B）	53.60	34.31	78.03	开源模型微调效果优异，微调后性能超越 BERT，突显基座能力优势
GPT4＋ICL	闭源 LLM＋少样本提示	52.13	33.72	75.71	通用模型局限。缺乏领域微调，难以对齐特定文书规范
REFLEXES	大小模型协同＋迭代反思	58.42	39.86	78.75	SOTA 表现优异，结合大模型逻辑与小模型效率，效果最佳

3.2.3 临床价值评估 LLM 在“幻觉”和输入字面化解读方面仍面临挑战，缺乏人类医生常用的基于假设的视角[22]。鉴于单纯的技术指标无法衡量这些深层缺陷，评估重心逐渐转向临床价值与社会伦理维度。一是临床有效性与安全性方面，超越字面相似度，引入医学专家，依据诊疗指南对药物配伍、禁忌证规避等，进行逻辑自洽性审查，同时为解决人工评估的高成本问题，基于 LLM 的生成式文本自动评估方法[50]正逐渐成为趋势，用于在大规模测试中初筛临床逻辑错误。二是可解释性成为信任构建的关键，应重点考核证据回溯率，确保模型的推理路径符合病理生理学逻辑，而非单纯的概率拟合。三是评估框架还应纳入算法公平性检测，量化模型在不同人口学群体中的诊断方差，以防范技术红利转化为算法歧视。

3.3 LLM 在电子病历应用中的挑战、风险与伦理考量

尽管 LLM 在上述 3 大范式中展现了巨大潜力，但在“容错率为零”的医疗领域，从实验环境走向临床深水区仍要跨越技术可靠性与伦理合规性双重鸿沟。

3.3.1 技术瓶颈 一是事实一致性与“幻觉”风险。生成式模型的概率本质与临床对确定性的要求存在内生矛盾。无论是国外的 ChatGPT 还是国内的通用大模型，在未经医疗专项知识增强前，均面临此风险。例如，一项针对肾功能障碍患者临床决策支持系统的研究[51]发现，ChatGPT 在不同场景下的表现欠佳，其正确且完全一致的响应率不足 20%。这类伪造数值的事实性错误如果未被拦截，将酿成医疗事故。目前，结合 RAG 进行知识锚定仍是解决此问题的首选方案。二是评估基准的错位。一方面，MedQA（US－MLE）、MedMCQA 等现有基准虽覆盖广泛，但缺乏对可信度、有用性、可解释性和忠实度的综合考量。另一方面，当前评估体系过度依赖“闭卷”医学问答的高分，掩盖了模型在长病程逻辑推理等开放式任务中的短板。亟待建立涵盖真实世界复杂度的综合评测基准，以纠正这种“高分低能”的评估偏差。三是知识演进的滞后性。医学知识持续更新，而大模型的参数化知识难以动态更新。依赖过时知识的系统无法应对新发疾病或指南更新。这迫使技术路径必须从单纯的微调，向动态化 RAG 或更前沿的模型编辑[52]技术演进，以实现知识的准实时更新。

3.3.2 合规性与安全性挑战 医学界对使用 LLM 存在诸多担忧[53]，源于大模型在医疗领域落地必须坚持“安全、合规与自主可控”原则。一是数据隐私安全[54]，依据我国《数据安全法》与《个人信息保护法》，医疗数据属于敏感个人信息，严禁违规出境。直接调用 GPT－4、ChatGPT 等境外云端模型存在严重的数据合规风险，基于国产基座模型的私有化部署已成为其在国内医院落地的唯一可行路径。二是算法备案与全程监督，根据《生成式人工智能服务管理暂行办法》，医疗大模型在提供服务前需通过严格的算法备案与安全评估。如何确保模型输出符合社会主义核心价值观，确保算法公平性[55]，且在医疗建议上具备可追溯的责任主体与问责机制[56]，是产业化落地的关键。

4 未来展望

4.1 技术范式

针对“概率生成”导致的“幻觉”问题，未来技术范式将向“神经符号协同”演进[57]。该范式旨在融合 LLM 的语义理解能力和知识图谱的逻辑约束能力。一方面，通过将结构化医学知识显式注入推理路径，构建“可验证思维链”，强制模型在生成诊断时必须回溯至具体的病理生理学证据，而非单纯依赖统计共现，从而确保决策的医学可解释性。另一方面，为突破单一文本模态的信息校验局限，模型应具备跨模态的细粒度对齐能力，即将生命体征波形（时序）、影像学表现（视觉）与病程记录（文本）在语义层交叉验证，实现对患者病情的全息感知与核查。

4.2 应用模式

针对临床应用“割裂”与“被动”问题，未来应用模式将通过医疗智能体[58-61]技术实现“闭环”。这意味着模型将从单纯的“读写者”进化为具备工具学习能力的“临床操作者”。智能体将通过应用程序接口深度集成医院信息系统、检验信息系统等，在医生授权下自主执行数据检索、检查开单以及病历归档等操作，使人工操作不再烦琐。同时交互模式将从“医生问 - 模型答”转变为模型主动辅助（如主动追问病历缺项），并通过强化学习在人机协作中持续进化，真正融入临床工作流。

4.3 生态构建

针对数据孤岛与隐私安全的伦理悖论，未来的产业生态将向“国产化、轻量化、联邦化”转型。信创适配与自主可控是产业基石，为规避外部技术封锁风险，应加速医疗大模型与国产算力芯片（如华为昇腾、海光等）及深度学习框架的适配，构建软硬件全栈自主可控的各种医疗智能体。联邦微调[62]模式下，在不共享原始数据的前提下，利用 DeepSeek、Qwen 等开源基座，医疗机构可在本地训练极小的适配器参数并上传聚合，不仅降低了对高端算力的依赖，更确保数据在安全红线内，实现跨机构知识协同。此外，为应对责任归属的法律真空，必须建立基于红队测试的对抗性评估机制[63]，在模型上线前模拟各类恶意攻击与极端医疗场景，确立算法责任认定的技术标准，为 AI 医疗构筑坚实的伦理防线。

5 结语

LLM 重塑了 EMR 的价值挖掘模式。本研究系统梳理信息结构化、临床文本生成、智能决策支持 3 大核心范式。既有研究多集中于单模态任务与离线评测，在跨机构泛化能力、多模态时序数据精细对齐以及长尾罕见病知识覆盖等方面仍存在显著不足。展望未来，LLM 的演进将超越单一模态的性能提升，向“神经符号协同（可信推理）”“临床智能体（自主协同）”与“联邦微调生态（隐私协作）”的系统性变革迈进。这一进程终将推动 EMR 从被动式的数字档案库，进化为主动式的智慧诊疗引擎。

作者贡献：赵霞负责研究设计、论文撰写；赵尔康负责文献收集与整理、论文修订；程春雷、李小华负责提供指导；张海波、姚佳璇负责项目管理。
利益声明：所有作者均声明不存在利益冲突。

参考文献

1 KRUSE C S, STEIN A, THOMAS H, et al. The use of electronic health records to support population health: a systematic review of the literature [J]. Journal of medical systems, 2018, 42 (11): 214.
2 RASMY L, XIANG Y, XIE Z, et al. Med - BERT: pre - trained contextualized embeddings on large - scale structured electronic health records for disease prediction [J]. NPJ digital medicine, 2021, 4 (1): 86.
3 ABID W. The SADID evaluation datasets for low - resource spoken language machine translation of Arabic dialects [C]. Barcelona: The 28th International Conference on Computational Linguistics, 2020.
4 STEINBERG E, JUNG K, FRIES J A, et al. Language models are an effective representation learning technique for electronic health record data [J]. Journal of biomedical informatics, 2021, 113 (1): 103637.
5 RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring

the limits of transfer learning with a unified text – to – text transformer [J]. Journal of machine learning research, 2020, 21 (140): 1 – 67.
6 张铮, 刘晨旭. 大模型幻觉: 人机传播中的认知风险与共治可能 [J]. 苏州大学学报 (哲学社会科学版), 2024, 45 (5): 171 – 180.
7 SINGHAL K, TU T, GOTTWEIS J, et al. Toward expert – level medical question answering with large language models [J]. Nature medicine, 2025, 31 (3): 943 – 950.
8 JIN D, PAN E, OUFATTOLE N, et al. What disease does this patient have? A large – scale open domain question answering dataset from medical exams [J]. Applied sciences, 2021, 11 (14): 6421.
9 NORI H, KING N, MCKINNEY S M, et al. Capabilities of GPT – 4 on medical challenge problems [EB/OL]. [2025 – 09 – 12]. https: //doi. org/10. 48550/arXiv. 2303. 13375.
10 VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]. Long Beach: Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017.
11 LI L, ZHOU J, GAO Z, et al. A scoping review of using large language models (LLMs) to investigate electronic health records (EHRs) [EB/OL]. [2025 – 09 – 12]. https: //doi. org/10. 48550/arXiv. 2405. 03066.
12 DEVLIN J, CHANG M W, LE K, et al. BERT: pre – training of deep bidirectional transformers for language understanding [C]. Minneapolis: The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.
13 乔晋华, 马雪赟. LLaMA 人工智能大模型在高校未来学习中心应用的风险与规制 [J]. 农业图书情报学报, 2025, 37 (2): 37 – 48.
14 CHEN H, JI H, SUN L, et al. Knowledge graph and semantic computing: semantic, knowledge, and linked big data [EB/OL]. [2025 – 09 – 12]. https: //link. springer. com/book/10. 1007/978 – 981 – 10 – 3168 – 7.
15 PAN J Z, CHEN H, HU W, et al. Knowledge graph and semantic computing. language, knowledge, and intelligence [EB/OL]. [2025 – 09 – 12]. https: //link. springer. com/book/10. 1007/978 – 981 – 10 – 7361 – 8.
16 QU B, CHEN H, BI Z, et al. Knowledge graph and semantic computing: knowledge graph and cognitive intelligence [EB/OL]. [2025 – 09 – 12]. https: //link. springer. com/book/10. 1007/978 – 981 – 33 – 6612 – 1.
17 ZHAO S, PAN J Z, LUO X, et al. Knowledge graph and semantic computing: knowledge graph empowers new infrastructure construction [EB/OL]. [2025 – 09 – 12]. https: //link. springer. com/book/10. 1007/978 – 981 – 16 – 6471 – 7.
18 QI G, FENG Y, WANG Z, et al. Knowledge graph and semantic computing: knowledge graph empowers artificial general intelligence [EB/OL]. [2025 – 09 – 12]. https: //link. springer. com/book/10. 1007/978 – 981 – 99 – 7311 – 8.
19 COSSU A, CARTA A, PASSARO L, et al. Continual pre – training mitigates forgetting in language and vision [J]. Neural networks, 2024, 179 (11): 106492.
20 BOLTON E, VENIGALLA A, YASUNAGE M, et al. BioMedLM: a 2. 7 b parameter language model trained on biomedical text [EB/OL]. [2025 – 09 – 12]. https: //arxiv. org/abs/2403. 18421.
21 CHENG S W, CHANG C W, CHANG W J, et al. The now and future of ChatGPT and GPT in psychiatry [J]. Psychiatry and clinical neurosciences, 2023, 77 (11): 592 – 596.
22 ZHOU H, LIU F, GU B, et al. A survey of large language models in medicine: progress, application, and challenge [EB/OL]. [2025 – 09 – 12]. https: //doi. org/10. 48550/arXiv. 2311. 05112.
23 XU X, LI M, TAO C, et al. A survey on knowledge distillation of large language models [EB/OL]. [2025 – 08 – 29]. https: //doi. org/10. 48550/arXiv. 2402. 13116.
24 CHEN J, GUI C, OUYANG R, et al. Huatuogpt – vision, towards injecting medical visual knowledge into multimodal LLMs at scale [EB/OL]. [2025 – 08 – 29]. https: //doi. org/10. 48550/arXiv. 2406. 19280.
25 DETTMERS T, PAGNONI A, HOLTZMAN A, et al. QLoRA: efficient finetuning of quantized LLMs [C]. New Orleans: Advances in Neural Information Processing Systems 36 (NeurIPS 2023), 2024.
26 林致中, 王华珍. 基于 Transformer 交互指导的医患对话联合信息抽取方法 [J]. 计算机应用研究, 2024, 41 (8): 2315 – 2321.
27 HAN T, ADAMS L C, PAPAIOANNOU J M, et al. MedAlpaca—an open – source collection of medical conversational AI models and training data [EB/OL]. [2025 – 09 – 12]. https: //doi. org/10. 48550/arXiv. 2304. 08247
28 YANG X, LI T, SU Q, et al. Application of large language models in disease diagnosis and treatment [J]. Chinese medical journal, 2025, 138 (2): 130 – 142.
29 WEI J, WANG X, SCHUURMANS D, et al. Chain – of – thought prompting elicits reasoning in large language models [J]. Advances in neural information processing systems, 2022, 35: 24824 – 24837.
30 LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval – augmented generation for knowledge – intensive NLP tasks [C]. Online: Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020.
31 EDGE D, TRINH H, CHENG N, et al. From local to global: a graph RAG approach to query – focused summarization [EB/OL]. [2025 – 09 – 12]. https: //doi. org/10. 48550/arXiv. 2404. 16130.
32 GUO Z, XIA L, YU Y, et al. LightRAG: simple and fast retrieval – augmented generation [EB/OL]. [2025 – 09 – 12]. https: //doi. org/10. 48550/arXiv. 2410. 05779.
33 WU X, DUAN J, PAN Y, et al. Medical knowledge graph:

data sources, construction, reasoning, and applications [J]. Big data mining and analytics, 2023, 6 (2): 201 - 217.
34 ALSENTZER E, MURPHY J, BOAG W, et al. Publicly available clinical BERT embeddings [C]. Minneapolis: The 2nd Clinical Natural Language Processing Workshop, 2019.
35 ZENG Z, ZHANG Y, TANG H, et al. A domain knowledge graph construction method based on joint extraction of GlobalPointer [C]. Hangzhou: 2024 7th International Conference on Pattern Recognition and Artificial Intelligence (PRAI). IEEE, 2024.
36 ZHU Y, LIU Y. LLM - NER: advancing named entity recognition with LoRA + fine - tuned large language models [C]. Kyoto: 2025 11th International Conference on Computing and Artificial Intelligence (ICCAI). IEEE, 2025.
37 BIAN J, PENG Y, WANG L, et al. A survey on parameter - efficient fine - tuning for foundation models in federated learning [EB/OL]. [2025 - 09 - 12]. https://doi.org/10.48550/arXiv.2504.21099.
38 BALLOCH J, SRIDHARAN S, OLDHAM G, et al. Use of an ambient artificial intelligence tool to improve quality of clinical documentation [J]. Future healthcare journal, 2024, 11 (3): 100157.
39 倪海清, 刘丹, 史梦雨. 基于语义感知的中文短文本摘要生成模型 [J]. 计算机科学, 2020, 47 (6): 74 - 78.
40 席铁钧, 段宗涛, 曹建荣, 等. 面向长文本涉法舆情信息的混合式摘要方法 [J]. 中文信息学报, 2024 38 (7): 63 - 72.
41 钟博洋, 阮彤, 张维彦, 等. 基于大小模型结合与迭代反思框架的电子病历摘要生成方法 [J]. 计算机科学, 2025, 52 (9): 294 - 302.
42 周文粲, 陈洁, 冯艳芳, 等. 大语言模型病历质控与病程记录生成评估方法研究 [J]. 中国卫生信息管理杂志, 2025, 22 (2): 163 - 170.
43 陈冯薏, 朱珠, 俞刚, 等. 基于微调大语言模型的儿科电子病历质控系统研究与应用 [J]. 中国卫生信息管理杂志, 2025, 22 (5): 702 - 710, 727.
44 ABBAS Q, JEONG W, LEE S W. Explainable AI in clinical decision support systems: a meta - analysis of methods, applications, and usability challenges [J]. Healthcare, 2025, 13 (17): 2154.
45 WU S, HE Y. Enriching pre - trained language model with entity information for relation classification [C]. Beijing: The 28th ACM International Conference on Information and Knowledge Management, 2019.
46 金方焱, 王秀利. 融合 RACNN 和 BiLSTM 的金融领域事件隐式因果关系抽取 [J]. 计算机科学, 2022, 49 (7): 179 - 186.
47 HENDRICKX I, KIM S N, KOZAREVA Z, et al. Semeval - 2010 task 8: multi - way classification of semantic relations between pairs of nominals [C]. Uppsala: The 5th International Workshop on Semantic Evaluation, 2010.
48 黄俏娟, 曹存根, 陈志文. 模式与深度学习融合抽取因果事件三元组 [J]. 高技术通讯, 2024, 34 (9): 921 - 934.
49 CHEN W, LI Z, FANG H, et al. A benchmark for automatic medical consultation system: frameworks, tasks and datasets [J]. Bioinformatics, 2023, 39 (1): 817.
50 兰天, 马梓奥, 周杨浩, 等. 生成式文本质量的自动评估方法综述 [C]. 太原: 第 23 届全国计算语言学学术会议 (CCL 2024), 2024.
51 VAN NULAND M, SNOEP J J D, EGBERTS T, et al. Poor performance of ChatGPT in clinical rule - guided dose interventions in hospitalized patients with renal dysfunction [J]. European journal of clinical pharmacology, 2024, 80 (8): 1133 - 1140.
52 YAO Y, WANG P, TIAN B, et al. Editing large language models: problems, methods, and opportunities [EB/OL]. [2025 - 09 - 12]. https://doi.org/10.48550/arXiv.2305.13172.
53 SALLAM M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns [J]. Healthcare, 2023, 11 (6): 887.
54 施敏, 杨海军. 大语言模型数据隐私保护的难点与探索 [J]. 大数据, 2024, 10 (5): 168 - 176.
55 PIERSON E. Accuracy and equity in clinical risk prediction [J]. New England journal of medicine, 2024, 390 (2): 127 - 135.
56 NOVELLI C, TDDEO M, FLORIDI L. Accountability in artificial intelligence: what it is and how it works [J]. AI & society, 2024, 39 (4): 1871 - 1882.
57 WANG S, SUN X, LI X, et al. GPT - NER: named entity recognition via large language models [C]. Albuquerque: Findings of the Association for Computational Linguistics: NAACL 2025, 2025.
58 XI Z, CHEN W, GUO X, et al. The rise and potential of large language model based agents: a survey [J]. Science China information sciences, 2025, 68 (2): 121101.
59 WANG L, MA C, FENG X, et al. A survey on large language model based autonomous agents [J]. Frontiers of computer science, 2024, 18 (6): 186345.
60 LI G, HAMMOUD H, ITANI H, et al. Camel: communicative agents for "mind" exploration of large language model society [J]. Advances in neural information processing systems, 2023, 36: 51991 - 52008.
61 TANG X, ZOU A, ZHANG Z, et al. MedAgents: large language models as collaborators for zero - shot medical reasoning [C]. Bangkok: Findings of the Association for Computational Linguistics: ACL 2024, 2024.
62 ZHANG C, XIE Y, BAI H, et al. A survey on federated learning [J]. Knowledge - based systems, 2021, 216 (3): 106775.
63 包泽芃, 钱铁云. 大模型红队测试研究综述 [J]. 计算机科学, 2025, 52 (1): 34 - 41.