MedCoRAG：通过混合证据检索与多专科共识实现可解释的肝病学诊断

Zheng Li[a], Jiayi Xu[a], Zhikai Hu[a], Hechang Chen[b],Lele Cong[c],Yunyun Wang[d,*],Shuchao Pang[a,e,*]

[a]南京理工大学网络空间安全学院，中国南京，*210094*[b]吉林大学人工智能学院，中国长春，*130015*[c]吉林大学中日联谊医院神经内科，中国长春，*130033*[d]吉林大学中日联谊医院麻醉科，中国长春，*130033*[e]麦考瑞大学计算学院，澳大利亚新南威尔士州悉尼，*2109*

arXiv:2603.05129v1 [cs.AI] 5 Mar 2026

摘要

准确且可解释地诊断肝脏疾病至关重要，但在真实临床环境中仍然充满挑战。现有用于临床诊断的人工智能方法往往缺乏透明度、结构化推理能力和可部署性。近期研究已开始采用大语言模型（LLM）、检索增强生成（RAG）和多智能体协作。然而，这些方法通常仅从单一来源检索证据，无法基于结构化临床数据支持迭代式、角色专门化的审议。为此，我们提出 MedCoRAG（即医学协作式检索增强生成），这是一个端到端框架，可根据标准化异常发现生成诊断假设，并通过联合检索和筛选 UMLS知识图谱路径与临床指南，构建患者特定证据包。随后，该框架执行多智能体协作推理：路由智能体根据病例复杂度动态分派专科智能体；这些智能体基于证据进行迭代推理，并在必要时触发有针对性的重新检索；全科智能体则综合所有审议结果，形成可追溯的共识诊断，模拟多学科会诊。基于 MIMIC-IV 肝脏疾病病例的实验结果表明，MedCoRAG 在诊断性能和推理可解释性方面均优于现有方法和闭源模型。

关键词：C　临床决策支持、大语言模型、检索增强生成、多智能体　　ts

1. 引言

在人工智能驱动的精准医学时代，从真实世界的电子健康记录（EHR）中准确、可解释地诊断肝脏疾病至关重要，但也极具挑战性。早期检测能够实现及时干预，对于防止不可逆损伤并显著改善预后至关重要。然而，这些疾病往往表现出模糊且相互重叠的症状 [1]。这种临床紧迫性要求方法不仅可靠，还必须透明，能够将复杂电子健康记录数据转化为可操作且可解释的诊断见解。

大语言模型在通用医学基准测试中展现出令人瞩目的能力 [2, 3] ，为临床人工智能提供了有前景的基础。然而，在 MIMIC-IV 等真实世界电子健康记录数据上应用时 [4], , 它们在肝病学领域面临着重大挑战：其知识是静态的，且可能已经过时，有时会给出自信却错误的诊断 [5, 6, 7]。更重要的是，它们的推理过程缺乏可追溯的逐步论证，因此难以使模型输出符合高风险肝脏疾病诊断所要求的可解释性标准。

检索增强生成（RAG） [8] 已成为一种将大语言模型的输出建立在外部证据基础上的策略。传统 RAG 依赖非结构化文本，但由于缺乏明确的医学关系，在多跳临床推理方面表现不佳。近期方法将医学知识图谱（KG） [9, 10] 融入其中，以支持结构化推理；通过利用概念之间的语义路径，这些方法的表现显著优于经典 RAG。然而，原始知识图谱路径往往包含无关或不合理的关联，而且这些方法仍无法纳入临床实践指南中的情境敏感指导，例如诊断标准或证据层级。与此同时，新兴的多智能体框架旨在通过部署专业智能体，在迭代对话中协作探讨鉴别诊断 [3, 11, 12]，从而模拟多学科会诊。然而，大多数系统 [13, 14] 依赖宽泛检索到的知识或内部知识，未能深入整合知识图谱和权威指南，因此达不到循证标准。此外，智能体的激活通常是静态的，无论病例复杂度如何 [15, 12], , 都会启用相同的一组专家，导致审议冗余或专业知识不足。值得注意的是，即使是先进的基于智能体的系统，也往往优先追求最终任务的准确性，而非可解释且符合临床医生认知的诊断推理，

*通讯作者电子邮箱：lizheng050427@163.com (Zheng Li), xujiayi041109@163.com (Jiayi Xu), huzhikai1115@163.com (Zhikai Hu), chenhc@jlu.edu.cn (Hechang Chen), congll18@mails.jlu.edu.cn(Lele Cong), wangyunyun@jlu.edu.cn (Yunyun Wang), pangshuchao@njust.edu.cn (Shuchao Pang)

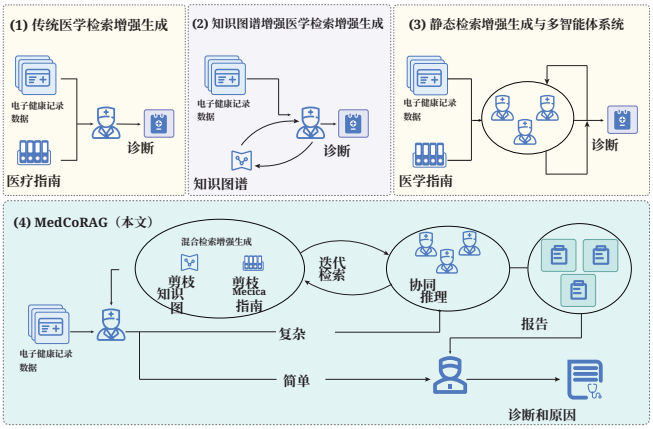

图1：医疗诊断推理框架对比概览

从而限制了人们对复杂肝胆决策的信任 [16, 17, 18]。

为弥补这些不足，我们提出 MedCoRAG（即医学协作式检索增强生成），这是一种混合检索增强生成与多智能体框架，旨在通过统一的证据综合来支持多学科临床推理。 [19] 如图 1所示，不同于依赖静态智能体团队或有限外部证据来源的既有方法，MedCoRAG 将指南约束的知识图谱剪枝与基于复杂度动态分派专科智能体相结合。我们首先将结构化电子健康记录数据转化为连贯的临床叙述，并生成一组初始诊断假设。针对每个假设，系统会检索临床指南摘录 [20, 21, 22] 和统一医学语言系统（UMLS） [23] 知识图谱路径，随后由大语言模型根据完整临床叙述和指南评估这些路径的临床连贯性，并据此对路径进行剪枝，从而为所有病例生成患者定制证据包。路由智能体根据临床叙述和异常发现评估病例复杂度。对于简单病例，通科智能体直接基于初始证据包综合得出诊断。对于复杂病例，系统根据临床背景动态分派相关专科智能体，例如肝病学或肿瘤学专科智能体；这些专科智能体基于证据包进行迭代推理，并在当前证据不足时触发定向检索。最终，通科智能体通过整体综合判定所有专家输入、审议历史和未解决的不确定性，得出单一且可追溯的共识诊断。

我们在来自 MIMIC-IV 的真实世界肝脏病例上评估了 MedCoRAG，结果表明，该方法能够生成精确且有证据支撑的诊断。该框架的动态路由和证据剪枝机制能够抑制虚假关联和冗余审议，从而实现聚焦且可解释的推理。

总而言之，我们的贡献如下：

- 我们提出 MedCoRAG，这是一种多专科检索增强生成智能体框架，通过协调按需专科智能体，在共享的、经指南剪枝的多跳知识图谱路径上进行迭代式、证据约束的诊断循环，从而动态模拟肝病学多学科团队会诊。
- 我们提出了符合多学科团队诊疗的混合推理方法，将剪枝后的知识图谱路径和临床指南摘录统一到单一证据空间中，并通过特定角色的专科视角进行联合解读，从而生成扎根于真实世界临床实践、具有可解释性且抗幻觉的共识诊断。
- 我们在来自 MIMIC-IV数据集的肝脏疾病病例上开展了全面实验。实验结果证明了 MedCoRAG 的有效性，其在各种诊断指标上均取得了较高性能。

2. 相关工作

医学检索与知识增强推理。标准检索增强生成通过将响应建立在医学文献或电子健康记录 [24], 之上来减轻大语言模型的幻觉，但其基于相似性的检索往往会返回不相关段落，且无法支持迭代式临床推理 [25]。近期研究从互补角度对此进行了改进：MedGraphRAG [26] 从学术论文、医学词典和临床指南中构建多层级知识图谱，并利用层次聚类生成结构化标签摘要，以实现连贯检索；MedRAG [27] 通过依据症状相似性而非诊断编码对疾病进行分组，提升了诊断对齐；KG-Rank [28] 利用知识图谱推导的实体路径相关性对检索段落重新排序，从而提升答案质量——在不修改大语言模型的情况下，ROUGE-L 提升超过 18%；而理由引导的检索增强生成 [29] 首先生成轻量级诊断理由，以引导单步检索，在无需微调的情况下提高精确率。尽管取得了这些进展，但每种方法都各自独立运行，没有一种方法能够联合整合结构化知识图谱路径、完整临床指南和自适应检索。我们的方法弥合了这一空白：联合检索由知识图谱导出的推理路径和指南摘录，然后应用领域感知剪枝，生成聚焦、可追溯且临床连贯的证据包。

临床协作多智能体系统。 早期多智能体框架通过零样本角色扮演 [11], 论证驱动的可解释性 [18], 或模拟临床环境 [15, 30]确立了基于角色的临床协作。近期工作转向以证据为基础且经过优化的工作流：ColaCare [14] 与LINS [13]在结构化电子健康记录或文献引用支持的链上协调智能体；TxAgent [31]专注于面向治疗规划的动态工具组合；MedAide [32]融合意图感知提取器以实现多方面推理。值得注意的是，MedAgent-Pro [33] 引入了一种推理型智能体工作流，可从多模态输入和临床指南构建可追溯的诊断路径，而MMedAgent-RL [34] 利用强化学习优化智能体协作策略，以提高诊断准确性。尽管取得了这些进展，大多数系统要么固定协作结构，要么将审议与统一的、

剪枝证据库解耦——从而限制了适应性与临床保真度。我们的方法通过基于异常发现动态路由专科医师，并在共享的、以临床指南为锚点的知识图谱上协调其强化学习指导的审议，来解决这一问题。

3. 方法

3.1. 总体架构

MedCoRAG通过结构化证据综合构建端到端诊断工作流程，如图 2 所示。系统从初始诊断假设集出发，首先检索临床指南摘录和多跳知识图谱路径，再结合完整的临床叙述对其进行联合剪枝，形成患者特定证据包。随后，路由智能体根据临床叙述和异常发现评估病例复杂度：对于简单病例，全科智能体直接依据初始证据给出诊断；对于复杂病例，则动态调度专科代理。这些专家基于共享证据迭代评估诊断假设，并在需要时触发由智能体引导的再次检索。最终，全科智能体通过对所有讨论进行整体综合判定，得出单一、可追溯且可用于临床决策的结论。

3.2. 核心组件

3.2.1. 异常发现与初步诊断

异常实体识别与标准化。 给定患者的病例描述 C，系统首先调用大语言模型提取候选异常实体：

$$\mathcal{E}_{\text{raw}} = \text{LLM}_{\text{NER}}(C), \tag{1}$$

其中，LLM_{NER} 表示一个经提示后从临床文本中识别与诊断相关异常的大语言模型。

为使这些实体与标准化医疗术语保持一致，系统会针对每个原始实体 $e \in \mathcal{E}_{\text{raw}}$查询知识图谱 UMLS，以获取一组候选标准化实体：

$$\mathcal{S}(e) = \text{KMatch}(e), \tag{2}$$

其中，KMatch(·)表示知识图谱实体匹配。

随后将原始实体e 及其候选匹配项 $\mathcal{S}(e)$呈现给大语言模型，由其选出匹配度最高的标准化实体，或指出不存在匹配项：

$$e_{\text{std}} = \text{LLM}_{\text{align}}(e, \mathcal{S}(e)) \in \mathcal{S}(e) \cup \{\varnothing\}. \tag{3}$$

最终的标准化异常实体集合定义如下：

$$\mathcal{E}_{\text{abn}} = \{e_{\text{std}} \mid e \in \mathcal{E}_{\text{raw}},\ e_{\text{std}} \neq \varnothing\}. \tag{4}$$

这确保了与底层知识图谱的语义一致性，从而能够进行精确的下游证据检索。

直接生成候选诊断。系统利用完整的病例描述 C和标准化异常发现 $\mathcal{E}_{\text{abn}}$生成初始诊断假设。这些输入被格式化为结构化临床提示，大语言模型直接生成一份简明的合理鉴别诊断列表。

形式上，初始假设列表生成如下

$$\mathcal{H}_{\text{initial}} = \text{LLM}_{\text{hypo}}(C, \mathcal{E}_{\text{abn}}). \tag{5}$$

输出采用有界序列的形式$[d_1, d_2, \ldots, d_K]$，且$K \leq K_{\max}$ ($K_{\max} = 4$)，以保持诊断聚焦，其中每个d_i 均为标准化的疾病名称。

3.2.2. 混合检索增强生成

该模块将每个候选诊断d_i 作为锚点，并同时检索两类互补证据：（1）来自临床指南语料库的权威陈述，以及（2）来自统一医学语言系统知识图谱的可解释推理路径。

临床指南检索与相关性过滤。 对于每个候选诊断d_i，通过将d_i 与 $\mathcal{E}_{\text{abn}}$中的所有异常发现相结合来构建复合查询，形成能够捕捉该假设临床背景的语义增强表示。该查询驱动对预索引临床指南数据库的两阶段检索过程。

在第一阶段，双编码器为查询和指南片段计算稠密嵌入，并使用余弦相似度检索最相关的前K 个片段（$K = 8$）。在第二阶段，交叉编码器基于上下文对齐对这些候选项重新排序，并选出前N 个片段（$N = 4$）。所得集合记为：

$$\mathcal{G}_i = \text{GRet}(d_i, \mathcal{E}_{\text{abn}}; N = 4), \tag{6}$$

其中，GRet(·) 表示两阶段指南检索流程。

知识图谱路径检索与指南引导剪枝。 对于每一对 (d_i, e_j)，其中 $e_j \in \mathcal{E}_{\text{abn}}$，系统会查询 UMLS知识图谱，检索从$e_j$ 到 d_i 且跳数不超过 3 的语义路径：

$$\mathcal{P}_{ij} = \text{KRet}(e_j, d_i; h_{\max} = 3), \tag{7}$$

其中，KRet(·) 表示知识图谱路径检索。

原始路径可能包含无关的语义关联。每条路径p 首先由大语言模型转述为一条自然语言陈述，通过中间概念连接e_j 与d_i 。为评估相关性，针对d_i 的转述路径（记为 $\mathcal{P}_i = \bigcup_{e_j \in \mathcal{E}_{\text{abn}}} \mathcal{P}_{ij}$）会分批处理（每批 8 条）。大语言模型仅依据完整病例描述 C 和排名靠前的指南摘录$\mathcal{G}^{\text{top}}{}_i \subseteq \mathcal{G}_i$ 评估每批路径，并给出二元判断。令 $\phi_{\text{rel}}(p;\, C, \mathcal{G}^{\text{top}}{}_i) \in \{0, 1\}$ 表示路径p 是否为患者的临床表现提供临床上连贯且有指南支持的解释。那么：

$$\mathcal{P}_i^{\text{valid}} = \left\{ p \in \mathcal{P}_i \,\middle|\, \phi_{\text{rel}}(p; C, \mathcal{G}_i^{\text{top}}) = 1 \right\}. \tag{8}$$

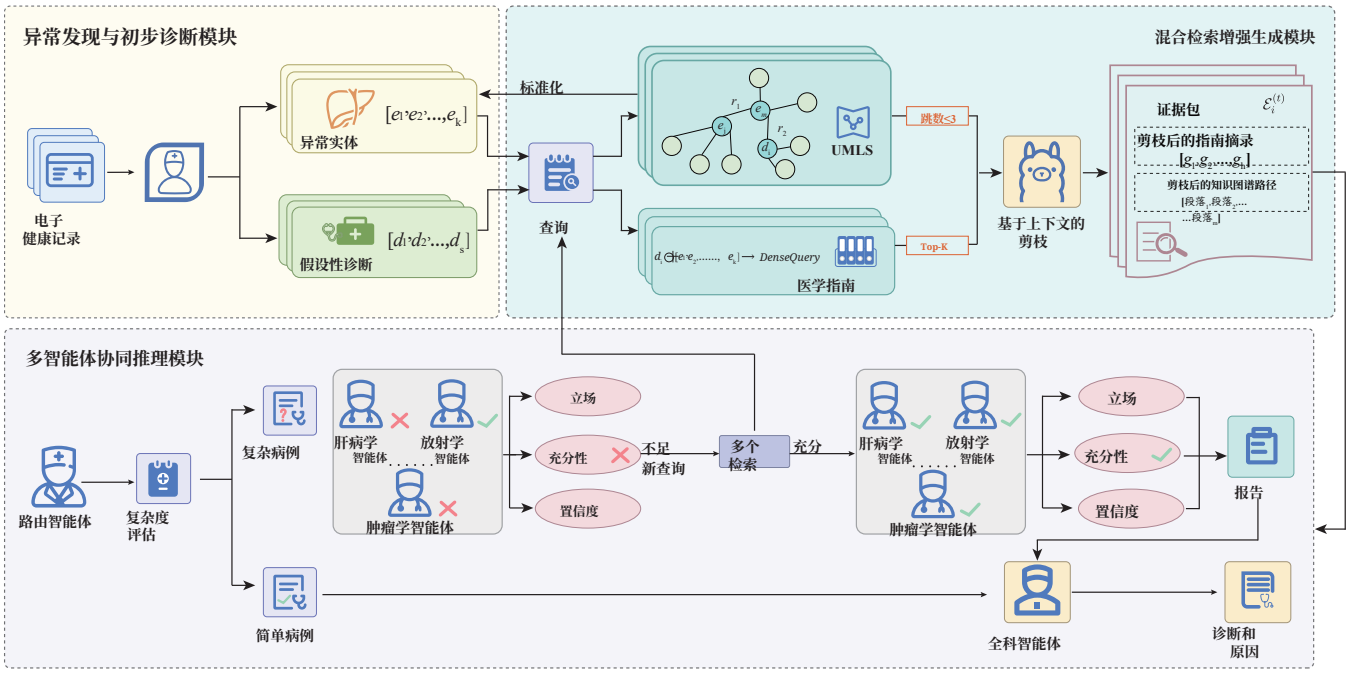

图2：MedCoRAG框架的整体架构，由三个核心组件构成。(1) 异常发现与初步诊断：从患者叙述中提取异常临床发现，并通过统一医学语言系统（UMLS）进行标准化，从而生成一组聚焦的初始诊断假设。(2) 混合检索增强生成：系统针对每个诊断假设检索临床指南摘录和UMLS知识图谱路径，然后结合完整的临床背景对其进行剪枝，形成连贯且患者特定的证据包。(3) 多智能体协作推理：路由智能体评估病例复杂度，以决定是否启用相关专科智能体，或将简单病例交由全科智能体处理；所有智能体基于共享证据进行迭代推理，并在需要时触发重新检索，最终通过全科智能体达成可解释的共识诊断。

最后，对于每个d_i，系统将$\mathcal{P}_i^{\text{valid}}$和$\mathcal{G}_i$汇总为结构化证据包$\mathcal{E}^{(0)}{}_i$，以支持后续的多智能体推理。

3.2.3. 多智能体协作推理

复杂度评估。 路由智能体处理完整的患者病例描述C——包括自由文本临床病史、体格检查记录以及叙述性实验室或影像学报告——并结合标准化异常发现$\mathcal{E}_{\text{abn}}$。这种整体表征能够检测诊断复杂性的细微指标，例如相互矛盾的发现、多器官受累、非典型表现和诊断不确定性。基于这一整合上下文，路由智能体应用叙述感知复杂度判别函数ϕ_{comp}，从而得出二元决策：

$$c = \phi_{\text{comp}}\left(C,\ \mathcal{E}_{\text{abn}},\ \mathcal{H}_{\text{initial}}\right), \tag{9}$$

其中$c \in \{0, 1\}$。当$c = 0$时，该病例被路由至全科智能体，由其根据预先检索的证据包$\mathcal{E}^{(0)}{}_i$直接综合得出最终诊断，从而绕过多智能体审议。当$c = 1$时，系统启动动态专科调度以进行协同推理。

对于复杂病例（$c = 1$），系统会执行动态专科调度。系统不会依赖固定的专科集合，而是根据临床背景动态选择相关专家。具体而言，调度函数ψ通过联合分析异常发现的语义类型和完整的临床叙述，识别出符合临床背景的专家子集：

$$\mathcal{A}_i = \psi\left(C,\ \mathcal{E}_{\text{abn}},\ d_i\right), \tag{10}$$

其中$\mathcal{A}_i$表示针对诊断而激活的专家集合d_i。例如，描述“黄疸、肌酐升高以及开始用药后出现斑丘疹”的病例将激活肝病学、肾脏病学和皮肤病学智能体。为了实现高效部署，我们采用知识蒸馏，将大模型（Qwen3-Max [35]）的临床推理能力迁移到学生模型（Llama-3.1-8B-Instruct [36]）中，由其在推理期间为专科智能体提供支持。

证据驱动的专科推理。 每个被调度的专科智能体$a \in \mathcal{A}_i$都会接收相同的推理上下文：

$$\text{Context}_i^{(t)} = \left(C,\ \mathcal{E}_{\text{abn}},\ d_i,\ \mathcal{E}_i^{(t)}\right), \tag{11}$$

其中$\mathcal{E}^{(t)}{}_i$表示第t轮的证据包（初始设为$t = 0$），其中包含通过第3.2.2节所述的混合检索增强生成流程检索到的知识图谱路径和指南摘录。在角色专属提示的引导下，每个智能体独立评估候选诊断d_i，生成立场$o_a^{(i)} \in$ {S、N、O}（支持、中立、反对）、置信度评分$c_a^{(i)} \in [0, 1]$以及证据充分性判断

$s_a^{(i)} \in$ {充分、 不足}，并附上明确以 $\mathcal{E}^{(t)}{}_i$ 中的条目为依据的理由。

如果有任何智能体报告了$s_a^{(i)}$ = 不足，协调智能体便会汇总这些信号，以决定是否启动额外的检索轮次。具体而言，令

$$\rho^{(t)} = \frac{1}{|\mathcal{A}_i|} \sum_{a \in \mathcal{A}_i} \mathbb{I}(s_a^{(i)} = \text{Ins}) \tag{12}$$

表示认为当前证据不足的智能体比例。当 $\rho^{(t)} > \tau_{\text{suff}}$ ($\tau_{\text{suff}} = 0$。5) 且$t < T_{\max}(T_{\max} = 3)$ 时，系统会启动定向二次检索：智能体协作制定查询，以解决特定的诊断不确定性，例如查询与生物标志物阈值相关的影像特征或指南标准；这些查询会与 C、 $\mathcal{E}_{\text{abn}}$、d_i以及已识别的知识缺口相结合，构建经过优化的检索输入；随后重新执行混合检索流程（第3.2.2)，以获取补充的知识图谱路径和指南摘录；最后，将新证据合并到更新后的信息包 $\mathcal{E}^{(t+1)}{}_i$ 中，供下一轮推理使用。这种闭环机制使检索与审议能够协同演进，直至诊断置信度趋于稳定或达到迭代上限。

共识形成与置信度校准。 在每次迭代 t 中，全科智能体综合所有专科智能体的输入——包括立场、置信度分数、基于证据的论证和充分性判断——为每个候选诊断 d_i生成临时共识报告。用于量化诊断可能性的校准分数 $s^{(t)}{}_i$ 计算如下:

$$s_i^{(t)} = \frac{1}{|\mathcal{A}_i|} \sum_{a \in \mathcal{A}_i} \mathbb{I}(o_a^{(i)} = \text{S}), \tag{13}$$

其中， $\mathbb{I}(\cdot)$表示指示函数。当 $s^{(t)}{}_i > \tau_{\text{high}}$ $(\tau_{\text{high}} = 0.9)$ 时，该过程提前终止，表明各方对 d_i具有强烈的共同支持。

最终诊断选择与输出。 全科智能体并非仅选择校准分数最高的诊断，而是进行全面的最终裁定。它整合完整的审议历史——包括不断演变的证据包、智能体间的共识与分歧、未解决的不确定性，以及患者特定背景（C、 $\mathcal{E}_{\text{abn}}$）——以生成单一且临床上连贯的最终诊断:

$$d_{\text{final}} = \mathcal{F}_{\text{final}}\left(\left\{(d_i, s_i^{(T)}, \text{Report}_i)\right\}_{d_i \in \mathcal{H}_{\text{initial}}}\right), \tag{14}$$

其中， $\mathcal{F}_{\text{final}}$ 表示全科医生的裁定函数。该决策同时附有一份全面且可追溯的共识报告，汇总所有迭代中的支持性与反对性证据，指出关键临床不确定性，并推荐可执行的后续检查。

MedCoRAG 的完整算法如算法1所示。

算法 1 MedCoRAG 诊断工作流程

输入： 病例描述 C

确保： 最终诊断d_{final}及其依据

$\mathcal{E}_{\text{abn}} = \{$大语言模型$_{\text{align}}(e,$ KMatch$(e)) \mid e \in$ 大语言模型$_{\text{NER}}(C)\}$

$\mathcal{H} =$ 大语言模型$^{\text{hypo}}(C, \mathcal{E}_{\text{abn}})$ {Top-K 重点假设 ($K \leq K_{\max}$)}

对每个$d_i \in \mathcal{H}$ **执行**

  $\mathcal{G}_i = \text{GRet}(d_i, \mathcal{E}_{\text{abn}}; N = 4)$

  $\mathcal{P}_i^{\text{valid}} = \{p \in \text{KRet}(e_j, d_i; h_{\max} = 3) \mid$ 大语言模型$^{\text{prune}}(p; C, \mathcal{G}^{\text{top}}{}_i) = 1\}$ {对所有$e_j \in \mathcal{E}_{\text{abn}}$}

  $\mathcal{E}_i^{(0)} \leftarrow (\mathcal{G}_i, \mathcal{P}_i^{\text{valid}})$

</style}结束循环

如果 $\phi_{\text{comp}}(C, \mathcal{E}_{\text{abn}}, \mathcal{H}) = 0$ **则**

  $d_{\text{final}} \leftarrow$ 通用智能体$(C, \mathcal{E}_{\text{abn}}, \mathcal{E}^{(0)}{}_i)$

  return d_{final} 并给出理由

结束条件判断

对于$t = 1$到 $T_{\max}$ **do**

  专家$a \in \mathcal{A}_i = \psi(C, \mathcal{E}_{\text{abn}}, d_i)$ 输出 $o_a^{(i)} \in$ {S, N, O} 以及$s_a^{(i)} \in$ {充分, 不足}

  计算 $\rho^{(t)} = \frac{1}{|\mathcal{A}_i|} \sum_a \mathbb{I}(s_a^{(i)} =$ 不足)

  如果 $\rho^{(t)} \leq \tau_{\text{suff}}$ **则**

    中断

  else

    通过智能体提出的查询和混合方法更新 $\mathcal{E}^{(t+1)}{}_i$ 检索

  结束条件判断

结束循环

$d_{\text{final}} = \mathcal{F}_{\text{final}}(\{(d_i,$ 报告$_i)\}_i)$

返回 d_{final} 及可追溯共识报告

4. 实验

我们在真实世界的肝病诊断任务上对 MedCoRAG 进行了全面评估，以考察异常实体驱动推理框架、多源知识整合和多智能体协作机制的有效性。

4.1. 数据集

我们从公开的 MIMIC-IV数据库中整理了一个临床数据集 [4], ，聚焦于确诊患有 13 种常见肝脏疾病之一的患者，这些疾病的标准化缩写列于表1中。为体现肝脏疾病的慢性和渐进性，我们保留每位患者的所有住院记录，从而重建其纵向病史。所有数据均已完全去标识化。

由于 MIMIC-IV 主要提供缺少叙述性背景的结构化表格记录，我们使用大语言模型，根据每位患者的纵向病程时间线，合成真实且背景信息丰富的临床叙述。随后，这些叙述被整理为医学问答对。最终数据集包含 3470 个问答样本，并在蒸馏阶段按疾病类别分层，以 7:3 的比例划分为训练集和测试集，确保全部 13 个类别均衡呈现。

表 1：疾病缩写

肝脏疾病	缩写
乙型肝炎	HBV
原发性胆汁性胆管炎	PBC
继发性肝癌	SLC
肝囊肿	肝囊肿
肝母细胞瘤	HB
肝硬化	LC
肝细胞癌	HCC
肝血管瘤	HH
肝衰竭	LF
自身免疫性肝炎	AIH
药物性肝损伤	DILI
非酒精性脂肪性肝炎	NASH
食管胃静脉曲张破裂出血	EGVB

4.2. 评估指标

我们使用四项标准指标评估诊断性能：召回率、精确率、F1分数和 F0.5分数。所有指标均按 13 个肝病类别加权平均报告，以考虑类别不平衡。

4.3. 基线模型

我们将 MedCoRAG 与一组涵盖不同模型规模和推理架构的全面基线方法进行比较，所有方法均使用相同的临床叙述，在同一测试集上评估。这些方法包括参数量最高达 8B 的医疗领域模型——Qwen3-Medical-GRPO-4B [37], OpenBioLLM-Llama3-8B [38], Bio-Medical-Llama3-8B [39], 和 Llama3-Med42-8B [40]；大型专有模型，包括 DeepSeek-V3.1-Think [41], Gemini-2.5-Pro [42], GLM-4.6 [43], 和 GPT-4o [44]；参数量从 14B 到 32B 的中型模型，包括 DeepSeek-R1-Distill-Qwen-32B [45], GPT-OSS-20B [46], Gemma3-27B [47], Qwen-QWQ-32B [48], 和 Phi-4-14B [49]；参数量低于 7B 的轻量级模型，即 ChatGLM3-6B [50] 和 Mistral-7B [51]；以及近期的检索增强生成与多智能体诊断框架——ColaCare [14], MedAgent-Pro [33], 和 MedAide [32]——这些框架均采用与 MedCoRAG 相同的基础模型（Llama-3.1-8B-Instruct [36]）实现。所有方法均按照相同的评估协议进行评估。

4.4. 实现细节

对于检索增强推理，我们整合了两种互补的外部知识源：（1）基于统一医学语言系统构建的结构化生物医学知识图谱；（2）由美国肝病研究学会、欧洲肝脏研究学会、亚太肝脏研究学会等主要专业学会发布的、涵盖肝脏疾病的 38 份权威临床指南组成的非结构化语料库。对于临床指南，我们使用 Qwen3-Embedding-8B [52], 进行稠密检索，将嵌入向量索引至 Milvus [53], , 随后使用 Qwen3-Reranker-8B [52]进行重排序。

我们使用 Llama-3.1-8B-Instruct 作为骨干大语言模型。为便于部署，我们将 Qwen3-Max [35] 的推理能力蒸馏到 Llama-3.1-8B-Instruct 中。具体而言，教师模型在 MedCoRAG 框架内模拟专科智能体行为以生成训练数据。学生模型通过监督学习并使用 LoRA [54] 微调三个轮次，采用初始化为5×10^{-5}的余弦衰减学习率，通过梯度累积实现有效批次大小为 8，最大序列长度为 11,000 个 token，全部以 bf16精度在单块 A800 GPU 上完成。

4.5. 主要结果

我们在标准化临床测试集上，对 13 类肝脏疾病进行诊断分类。MedCoRAG 的整体性能如表 2 所示。在所有评估方法中，MedCoRAG 在加权精确率、召回率、F1分数和 F0.5分数方面均取得了最佳表现。其性能优于参数规模最高达 8B 的专用医疗模型、大型专有语言模型、中小规模通用模型，以及近期的多智能体诊断框架。这些结果表明，MedCoRAG 通过将结构化证据综合与动态智能体协作相结合，能够支持在不同模型规模和推理策略下进行准确、可靠的临床诊断。

5. 分析

5.1. 诊断准确性与误分类分析

表 3 展示了 MedCoRAG 在各类疾病上的分疾病诊断指标，结果表明其在多种肝脏疾病中始终表现出色。对于具有明显临床或影像学特征的疾病，包括肝血管瘤、肝囊肿、继发性肝癌和药物性肝损伤，该模型均取得了较高的精确率和召回率。对于食管胃静脉曲张破裂出血等急性事件，该模型的预测精确率达到完美水平，表明其在给出关键诊断时具有较高置信度。图 3中的混淆矩阵显示，误分类主要发生在临床相关的疾病类别之间，例如肝硬化相关并发症和胆汁淤积性疾病，这反映了肝病学中已知的诊断相似性。这些结果表明，MedCoRAG 的推理与既有临床模式相一致，同时在多种肝脏疾病诊断中保持了稳健的准确性。

5.2. 诊断复杂性与推理模式

我们通过考察不同疾病的临床复杂性和推理深度，分析 MedCoRAG 如何适应不同的诊断需求。如图4所示，原发性胆汁性胆管炎和肝纤维化等疾病具有数量最多的异常实体，反映出其临床表现错综复杂；而肝血管瘤和乙型肝炎的临床表现则明显更为简单。相应地，图5显示，肝囊肿、肝纤维化、药物性肝损伤和继发性肝癌等疾病的诊断依赖于更长的知识图谱推理路径，这表明诊断过程中积极整合了临床指南和结构化医学知识中的多跳证据。这些结果共同表明，Med-CoRAG 会根据每个病例固有的复杂程度调整推理过程，在临床模糊性最大的情况下采用更深入的推理。

表2：不同模型类别的综合诊断性能（%）。每项指标的最佳结果以**粗体**显示。

类别	模型	精确率	召回率	F1分数	F0.5分数
医疗 (≤8B)	Qwen3-Medical-GRPO-4B [37]	69.68	58.93	61.07	59.06
	OpenBioLLM-Llama3-8B [38]	62.15	51.48	54.21	52.24
	Bio-Medical-Llama3-8B [39]	65.94	41.83	47.43	43.36
	Llama3-Med42-8B [40]	64.57	61.41	60.28	60.51
通用大型模型 (>100B)	DeepSeek-V3.1-Think [41]	79.59	76.98	77.61	77.05
	Gemini-2.5-Pro [42]	80.31	76.70	77.28	76.71
	GLM-4.6 [43]	80.76	75.36	76.55	75.53
	GPT-4o [44]	74.33	70.58	69.98	69.90
通用中型 (14–32B)	DeepSeek-R1-Distill-Qwen-32B [45]	78.49	74.59	74.90	74.35
	GPT-OSS-20B [46]	75.19	69.25	70.75	69.56
	Gemma3-27B [47]	73.89	60.84	65.92	62.61
	Qwen-QWQ-32B [48]	66.60	53.96	55.78	54.02
	Phi-4-14B [49]	62.70	54.44	55.68	54.26
通用小型 (≤7B)	ChatGLM3-6B [50]	56.57	40.97	41.97	40.89
	Mistral-7B [51]	60.86	40.02	41.52	39.12
智能体框架	ColaCare [14]	78.01	72.66	73.35	72.53
	MedAgent-Pro [33]	76.01	70.58	70.33	69.97
	MedAide [32]	77.98	73.23	74.10	73.26
Ours	MedCoRAG	**81.32**	**79.18**	**79.12**	**78.99**

表 3：MedCoRAG 的分疾病诊断指标（%）。

疾病	召回率	精确率	F1分数	F0.5分数
HBV	52.00	76.47	61.90	55.56
PBC	55.56	20.83	30.30	41.67
SLC	85.16	92.31	88.59	86.50
肝囊肿	87.76	93.48	90.53	88.84
HB	85.71	85.71	85.71	85.71
LC	67.44	62.59	64.93	66.41
HCC	90.45	74.19	81.52	86.65
HH	94.62	94.62	94.62	94.62
LF	15.79	75.00	26.09	18.75
AIH	69.49	54.67	61.19	65.92
DILI	82.00	88.74	85.24	83.27
NASH	69.35	87.76	77.48	72.39
EGVB	16.67	100.00	28.57	20.00

5.3. 消融研究

我们通过评估若干消融变体，考察 MED-CoRAG 各组件对诊断性能的贡献（见表4）：(1) w/oCG（不整合临床指南），智能体在推理过程中无法访问结构化临床指南；(2) w/o MA（不进行多智能体协作），由单个智能体生成诊断，不开展协作讨论；(3) w/o KG（不进行知识图谱基础化），排除医学知识图谱中的疾病实体关系；(4) w/o TD（不进行教师蒸馏），学生模型在没有基于教师的推理过程指导的情况下运行；此外

表4：消融研究结果（%）。每列最佳结果以**粗体**标出。CG：临床指南整合；MA：多智能体协商；KG：知识图谱基础化；TD：教师蒸馏。

变体	精确率	召回率	F1分数	F0.5分数
无临床指南	78.08	72.75	73.43	72.62
无多智能体协作	76.30	69.72	69.70	69.14
无知识图谱	77.96	73.14	73.81	73.03
无教师引导蒸馏	78.20	73.71	74.48	73.73
无知识图谱与临床指南	74.47	68.10	67.86	67.37
无教师引导蒸馏与多智能体协作	76.83	68.74	69.93	68.62
无知识图谱、临床指南与多智能体协作	69.30	57.50	55.32	55.62
MedCoRAG	**81.32**	**79.18**	**79.12**	**78.99**

作为组合消融实验，包括无知识图谱与临床指南、无教师引导蒸馏与多智能体协作，以及无知识图谱、临床指南与多智能体协作。

结果表明，移除任一单一组件都会导致所有指标的性能下降；当知识来源（知识图谱与临床指南）和多智能体协作同时缺失时，降幅最为显著（F1降至55.32）。值得注意的是，禁用多智能体协商（不含多智能体协作）相较完整模型使F1降低9.42个百分点，表明智能体交互在完善诊断假设方面发挥着关键作用。同样，省略临床指南或知识图谱会使F1分数低于74，说明结构化外部证据对于准确推理至关重要。不含教师引导蒸馏时的表现相对更强（F1：74.48），这意味着尽管教师引导蒸馏能够进一步提升性能，但核心增益来自权威知识与协作智能体的整合

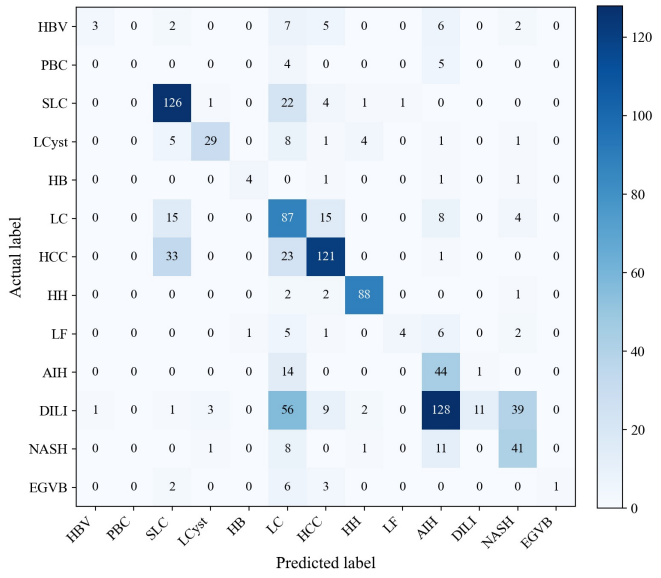

图3：MedCoRAG在13种肝病类别上的混淆矩阵。

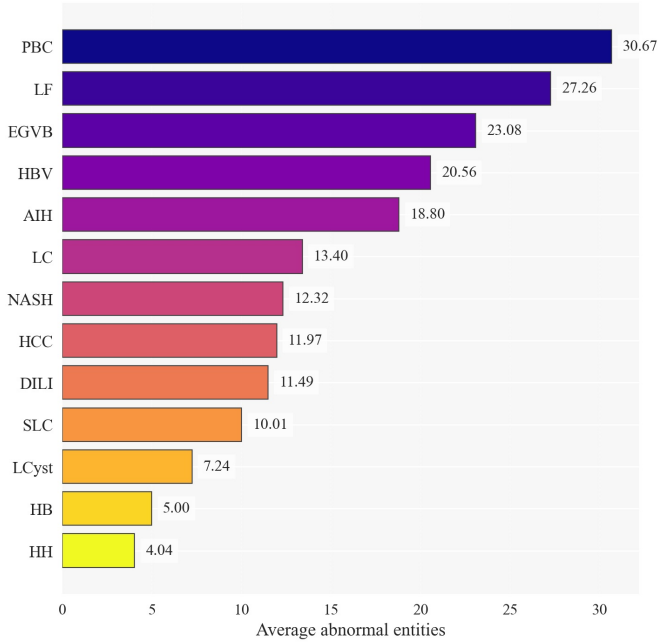

图4：不同肝脏疾病病例中每例的异常实体平均数量。数值越高，表明临床表现越复杂。

动态。

5.4. 部署成本与效率

使用Qwen3-Max构建由教师模型生成的训练数据集，需一次性支付24.53美元，此后无持续性开销。在推理时，MedCoRAG的延迟表现可预测：需要多智能体协作推理的病例平均耗时33.36秒，而由单个全科智能体处理的简单病例仅需9.95秒。

5.5. 案例研究：架构如何实现专家级诊断

我们展示一个具有代表性的病例，以突出MedCoRAG的模块化设计的优势。一名48岁女性因持续疲劳、瘙痒和黄疸就诊。实验室检查显示，碱性磷酸酶（ALP，340 U/L）、γ-谷氨酰转移酶（GGT，280 U/L）和免疫球蛋白M（IgM，3.8 g/L）显著升高，而免疫球蛋白G（IgG）水平正常。腹部超声未发现胆道梗阻，但显示轻度肝肿大。患者否认饮酒、近期更换药物或接触病毒性肝炎的情况。

该病例因原发性胆汁性胆管炎、自身免疫性肝炎和药物性肝损伤之间存在重叠而具有诊断挑战性。普通语言模型可能会根据疲劳和转氨酶升高而倾向于诊断自身免疫性肝炎或药物性肝损伤，却忽略胆汁淤积模式。

相比之下，MedCoRAG 运用其完整架构进行精准推理。首先，异常实体检测识别出碱性磷酸酶_升高、γ-谷氨酰转移酶_升高、免疫球蛋白M_升高和瘙痒，这些异常共同触发动态路由，将任务分配给自身免疫性肝病学智能体，同时抑制病毒学或肿瘤学等无关专家参与。

随后，已激活的智能体从欧洲肝脏研究学会检索到一段临床指南，其中指出：“中年女性若出现持续性胆汁淤积且仅有免疫球蛋白M升高，即使未检测到抗线粒体抗体，也高度提示原发性胆汁性胆管炎。”与此同时，知识图谱给出一条连贯的2跳路径：瘙痒 → 原发性胆汁性胆管炎 → 免疫球蛋白M_升高，依据既定的疾病语义将症状与血清学标志物联系起来。

多智能体协作进一步完善了诊断：免疫学智能体指出，免疫球蛋白G水平正常，不支持典型的自身免疫性肝炎；肝病学通才智能体则交叉验证胆汁淤积性酶谱，并确认不存在药物暴露，从而排除药物性肝损伤。系统最终以高置信度诊断为原发性胆汁性胆管炎，并提供可追溯的诊断依据以及抗线粒体抗体检测建议。

该病例展示了 MedCoRAG 各组件如何协同发挥作用：由异常驱动的路由确保临床相关性；混合检索将推理建立在指南和结构化知识之上；多智能体协商则支持细致的鉴别分析——整个过程无需依赖侵入性检查结果。这一集成工作流程支撑了该框架解决模糊病例的能力，其表现可媲美专家的临床判断。

5.6. 局限性与未来工作

MedCoRAG 在 MIMIC-IV 的肝脏病例中展现出出色的诊断性能和可解释推理，但仍存在若干局限。当前实现仅处理单个临床快照，未能对实验室指标趋势或影像变化等纵向信号进行建模。此外，它依赖与统一医学语言系统对齐的实体和静态指南，因此可能容易受到真实临床病历中歧义的影响。而且，所有评估均为回顾性研究，尚未在实时临床工作流程中得到验证。为解决这些问题，未来

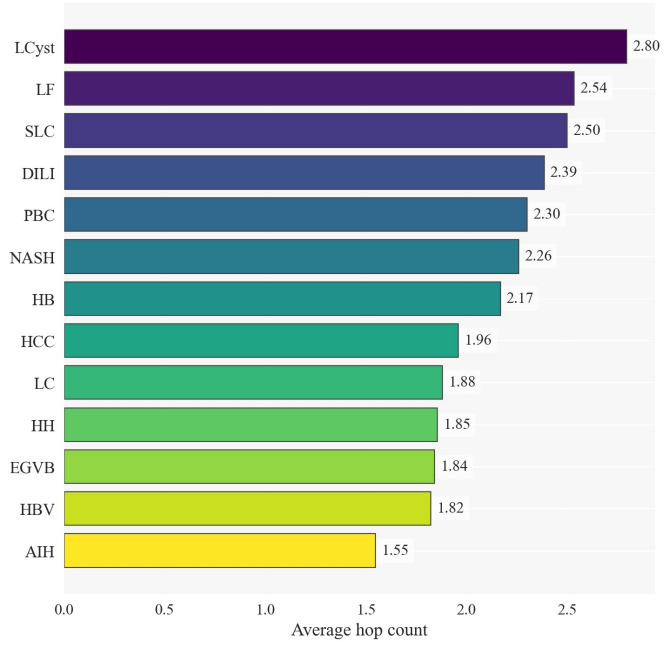

图5：诊断过程中使用的知识图谱路径的平均跳数。数值越高，表示推理复杂度越高。

的工作将纳入时序电子健康记录建模，以实现具备时间感知能力的诊断；将智能体框架扩展至肝病学以外的更多临床领域；通过改进自然语言理解，增强对非结构化文本的鲁棒性；开发轻量级机制，以便高效部署；并与临床合作伙伴开展前瞻性研究，评估其对决策支持和电子健康记录集成的现实影响。

6. 结论

我们提出了 MedCoRAG，这是一种用于可解释肝病诊断的混合检索增强生成与多智能体框架。该方法通过联合检索并剪枝医学知识图谱中的路径，以及从临床实践指南中提取的片段，构建针对患者的证据包。路由智能体根据异常临床发现动态激活相关专科智能体，使其能够围绕共享证据包开展迭代式、角色感知的审议。最终由全科智能体进行整体综合判定，形成一个同时以结构化知识和专家指导为依据、可追溯的诊断结果。在 MIMIC-IV 的真实世界肝脏病例上进行评估后，MedCoRAG 的表现优于多种基线方法，展现出更出色的诊断性能和推理可解释性。这项工作推动了更透明、以证据为基础且符合临床需求的人工智能发展，有助于实现可信赖的医疗决策支持。

CRediT作者贡献声明

Zheng Li：概念化、数据整理、形式分析、调查、方法论、项目管理、验证、可视化、撰写——初稿。**Jiayi Xu**：调查、撰写——审阅与编辑。**ZhikaiHu**：调查、撰写——审阅与编辑。**Hechang Chen**：监督、验证、撰写——审阅与编辑。**Lele Cong**：监督、验证、撰写——审阅与编辑。**YunyunWang**：形式分析、项目管理、监督、验证、撰写——审阅与编辑。**ShuchaoPang**：形式分析、经费获取、项目管理、监督、撰写——审阅与编辑。

利益冲突声明

作者声明，他们没有已知的可能影响本文所报告工作的竞争性经济利益或个人关系。

数据与代码可用性

MIMIC-IV数据库可通过PhysioNet公开获取：https://physionet.org/content/mimiciv/3.1。源代码将在录用后公开。

致谢

本研究得到国家自然科学基金（中国）（项目编号：62206128）和国家重点研发计划（中国）（项目编号：2023YFB2703900）的资助

参考文献

[1] C. Gan, Y. Yuan, H. Shen, J. Gao, X. Kong, Z. Che, Y. Guo, H. Wang, E. Dong, J. Xiao, 《肝脏疾病：流行病学、病因、趋势与预测》, Signal Transduction and Targeted Therapy 10 (1) (2025年) 33。[2] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis 等, 《借助大语言模型迈向专家级医学问答》, Nature Medicine 31 (3) (2025年) 943–950。[3] X. Tang, D. Shao, J. Sohn, J. Chen, J. Zhang, J. Xi- ang, F. Wu, Y. Zhao, C. Wu, W. Shi 等, 《Medagents-bench：复杂医学推理思维模型与智能体框架基准测试》, arXiv 预印本 arXiv:2503.07459 (2025年)。[4]A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow 等, 《MIMIC-IV：一个可免费访问的电子健康记录数据集》, Scientific Data 10 (1) (2023年) 1。[5] M. Griot, C. Hemptinne, J. Vanderdonckt, D. Yuksel, 《大语言模型缺乏可靠医学推理所必需的元认知能力》, Nature Communications 16 (1) (2025年) 642。

[6]P. Hager, F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. Makowski, R. Braren, G. Kaissis 等，《评估并缓解大语言模型在临床决策中的局限性》，Nature Medicine 30 (9) (2024年) 2613–2622。

[7] Y. Zhu, J. Gao, Z. Wang, W. Liao, X. Zheng, L. Liang, M. O. Bernabeu, Y. Wang, L. Yu, C. Pan 等，Clini-crealm：使用传统机器学习重新评估大语言模型在非生成式临床预测任务中的表现，arXiv 预印本 arXiv:2407.18525（2024年）。

[8] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, H. Wang，《大语言模型的检索增强生成：综述》，arXiv 预印本 arXiv:2312.10997 2 (1)（2023）。

[9] Y. Hu, Z. Lei, Z.Zhang, B. Pan, C. Ling, L. Zhao，GRAG：图检索增强生成（2025年）。arXiv:2405.16506。网址：https://arxiv.org/abs/2405.16506

[10] J. Wu、J. Zhu、Y. Qi、J. Chen、M. Xu、F. Menolascina、V. Grau，《医学图检索增强生成：通过图检索增强生成实现安全的医学大语言模型》，arXiv 预印本 arXiv:2408.04187（2024年）。

[11] X. Tang、A. Zou、Z. Zhang、Z. Li、Y. Zhao、X. Zhang、A. Cohan、M. Gerstein，《MedAgents：作为零样本医学推理协作者的大语言模型》，载于：计算语言学协会论文集：ACL 2024，2024年，第599–621页。

[12] Y. Kim、C. Park、H. Jeong、Y. S. Chan、X. Xu、D. McDuff、H. Lee、M. Ghassemi、C. Breazeal、H. W. Park，《MDAgents：面向医疗决策的大语言模型自适应协作》，《神经信息处理系统进展》第37卷（2024年），第79410–79452页。

[13] S. Wang, F. Zhao, D. Bu, Y. Lu, M. Gong, H. Liu, Z. Yang, X. Zeng, Z. Yuan, 等，LINS：一种通用医疗问答框架，用于提升大语言模型生成的回答的质量与可信度，Nature Communications 16 (1) (2025年) 9076。

[14] Z. Wang, Y. Zhu, H. Zhao, X. Zheng, D. Sui, T. Wang, W. Tang, Y. Wang, E. Harrison, C. Pan, 等，ColaCare：通过大语言模型驱动的多智能体协作增强电子健康记录建模，载于：ACM 2025年万维网大会论文集，2025年，第2250–2261页。

[15] S. Schmidgall, R. Ziaei, C. Harris, E. Reis, J. Jopling, M. Moor，Agentclinic：用于评估人工智能在模拟临床环境中表现的多模态智能体基准测试，arXiv预印本 arXiv:2405.07960 (2024年)。

[16] T. Hellingman, M. de Swart, J. Joosten, M. Meijerink, J.de Vries, J.de Waard, A.vanZweeden, B.Zonderhuis, G. Kazemier，《多学科专家小组评估结直肠癌肝转移患者治疗策略的价值》，《外科肿瘤学》35 (2020年) 412–417。[17] E. N. Smith, M. R. Bashir, A. Fung, B. D.Cash, M. Dixon, E. M. Hecht, B. M. McGuire, A. A. Pillai, G. K. Russo, R. T. Shroff, 等，《ACR适宜性标准®：原发性肝癌的分期与随访》，《美国放射学会杂志》22 (11) (2025年) S699–S712。

[18] S. Hong、L. Xiao、X. Zhang、J. Chen，Argmed-agents：基于论证方案、通过大语言模型讨论实现可解释的临床决策推理，载于：2024年IEEE生物信息学与生物医学国际会议（BIBM），IEEE，2024年，第5486–5493页。

[19] B. Pillay、A. C. Wootten、H. Crowe、N. Corcoran、B. Tran、P. Bowden、J. Crowe、A. J. Costello，多学科团队会议对肿瘤学临床环境中患者评估、管理和结局的影响：文献系统综述，《癌症治疗综述》42（2016年）56–72。

[20] E. A. F. T. S. O. T. Liver 等，《欧洲肝脏研究学会经颈静脉肝内门体分流术临床实践指南》，《肝病学杂志》83（1）（2025年）177–210。

[21] T. H. Taddei, D. B. Brown, M. Yarchoan, M. Mendiratta-Lala, J. M. Llovet，《重要更新：美国肝病研究学会关于肝细胞癌预防、诊断与治疗的实践指南》，《肝病学》（2025年）10–1097。

[22] J. P. Arab, L. A. Díaz, J. Rehm, G. Im, M. Arrese, P. S. Kamath, M. R. Lucey, J. Mellinger, M. Thiele, M. Thursz 等，《代谢功能障碍与酒精相关性肝病（代谢功能障碍与酒精相关性肝病）：酒精相关性肝病专家组立场声明》，《肝病学杂志》82 (4) (2025年) 744–756。

[23] O.Bodenreider，统一医学语言系统（统一医学语言系统）：整合生物医学术语，《核酸研究》32 Database issue(2004) D267–70。URL https://api.semanticscholar.org/CorpusID:205228801

[24] L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, J. Seidel，《医疗保健领域大型语言模型的检索增强生成：一项系统综述》，《PLOS数字健康》4 (6) (2025年) e0000877。

[25]G. Xiong, Q. Jin, Z. Lu, A. Zhang，《医学领域检索增强生成的基准测试》，载于：计算语言学协会 ACL 2024 论文集，2024年，第 6233–6251 页。

[26]J. Wu, J. Zhu, Y. Qi, J. Chen, M. Xu, F. Meno-lascina, Y. Jin, V. Grau，医学图检索增强生成：通过图检索增强生成实现循证医学大语言模型，载于：W. Che, J. Nabende,

E. Shutova、M. T. Pilehvar（编），《计算语言学协会第63届年会论文集》（第1卷：长篇论文），计算语言学协会，奥地利维也纳，2025年，第28443–28467页。 doi:10.18653/v1/2025.acl-long.1381。网址：https://aclanthology.org/2025.acl-long.1381/

[27] X. Zhao, S. Liu, S.-Y. Yang, C. Miao，《Medrag：通过知识图谱激发推理增强医疗副驾的检索增强生成》，载于：ACM 2025年 Web会议论文集，2025年，第4442–4457 页。

[28] R. Yang, H. Liu, E. Marrese-Taylor, Q. Zeng, Y. Ke, W. Li, L. Cheng, Q. Chen, J. Caverlee, Y. Matsuo 等，《KG-rank：利用知识图谱和排序技术增强大语言模型的医学问答》，载于：第 23 届生物医学自然语言处理研讨会论文集，2024年，第 155–166 页。

[29] J. Sohn, Y. Park, C. Yoon, S. Park, H. Hwang, M. Sung, H. Kim, J. Kang，《面向医学问答的理据引导检索增强生成》，载于：2025年计算语言学协会美洲国家分会会议：人类语言技术（第 1 卷：长篇论文）论文集，2025年，第 12739–12753 页。

[30] J. Li, Y. Lai, W. Li, J. Ren, M. Zhang, X. Kang, S. Wang, P. Li, Y.-Q. Zhang, W. Ma 等，智能体医院：一个由可演化医疗智能体构成的医院模拟体，arXiv 预印本 arXiv:2405.02957（2024年）。

[31] S. Gao, R. Zhu, Z. Kong, A. Noori, X. Su, C. Ginder, T. Tsiligkaridis, M. Zitnik，TxAgent：一种可借助工具库进行治疗推理的人工智能智能体，arXiv 预印本 arXiv:2503.10970（2025年）。

[32] D. Yang, J. Wei, M. Li, J. Liu, L. Liu, M. Hu, J. He, Y. Ju, W. Zhou, Y. Liu 等，MedAide：基于大语言模型的智能体协作实现医疗意图的信息融合与剖析，《信息融合》（2025年）103743。

[33] Z. Wang、J. Wu、L. Cai、C. H. Low、X. Yang、Q. Li、Y. Jin，《Medagent-pro：通过推理智能体工作流实现循证多模态医疗诊断》，arXiv预印本 arXiv:2503.18968（2025年）。

[34] P. Xia、J. Wang、Y. Peng、K. Zeng、X. Wu、X. Tang、H. Zhu、Y. Li、S. Liu、Y. Lu 等，《Mmedagent-rl：优化多模态医学推理的多智能体协作》，arXiv 预印本 arXiv:2506.00555（2025年）。

[35] A. Yang、A. Li、B. Yang、B. Zhang、B. Hui、B. Zheng、B. Yu、C. Gao、C. Huang、C. Lv 等，《Qwen3 技术报告》，arXiv 预印本 arXiv:2505.09388（2025年）。

[36] A. Dubey、A. Jauhri、A. Pandey、A.Kadian、A. Al-Dahle、A. Letman、A. Mathur、A. Schelten、A. Yang、A. Fan 等，《Llama 3 模型家族》，arXiv 电子预印本（2024年）arXiv–2407。

[37] T. Z. XIONG，Qwen3_医疗_grpo：通过组相对策略优化进行微调的医学领域大语言模型，访问日期：2025年11月15日（2025年6月）。网址：https://huggingface.co/lastmass/Qwen3_医疗_GRPO

[38]M. S. Ankit Pal，《OpenBioLLMs：推进面向医疗保健和生命科学的开源大语言模型》，https://huggingface.co/aaditya/OpenBioLLM-Llama3-8B，访问日期：2025年11月15日（2024年）。

[39]ContactDoctor，《Contactdoctor-bio-medical：一种高性能生物医学语言模型》，访问日期：2025年11月15日（2024年）。网址：https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B

[40]C. Christophe、P. K. Kanithi、T. Raha、S. Khan、M. A. Pimentel，《Med42-v2：一套临床大语言模型》，https://arxiv.org/abs/2408.06142，访问日期：2025年11月15日（2024年）。arXiv:arXiv:2408.06142。

[41]DeepSeek-AI，Deepseek-v3 技术报告 (2024年)。arXiv:2412.19437。网址：https://arxiv.org/abs/2412.19437

[42]G. Gemini 团队，《Gemini 2.5：以高级推理、多模态、长上下文和新一代智能体能力拓展前沿》，arXiv 预印本 arXiv:2507.06261（2025年）。

[43] A. Zeng、X. Lv、Q. Zheng、Z. Hou、B. Chen、C. Xie、C. Wang、D. Yin、H. Zeng、J. Zhang 等，《Glm-4.5：智能体、推理与编码（ARC）基础模型》，arXiv 预印本 arXiv:2508.06471（2025年）。

[44] O.（2024年），Gpt-4o 系统卡 (2024年)。arXiv:2410.21276。网址：https://arxiv.org/abs/2410.21276

[45] DeepSeek-AI，Deepseek-r1：通过强化学习激发大语言模型的推理能力 （2025年）。arXiv:2501.12948。网址：https://arxiv.org/abs/2501.12948

[46] OpenAI，gpt-oss-120b、gpt-oss-20b 模型卡（2025年）。arXiv:2508.10925。网址：https://arxiv.org/abs/2508.10925

[47] G. D. Gemma 团队，Gemma 3 技术报告 （2025年）。arXiv:2503.19786。网址：https://arxiv.org/abs/2503.19786

[48] Q. 团队，Qwq-32b：拥抱强化学习的力量 -ment学习 （2025年3月）。网址 https://qwenlm.github.io/blog/qwq-32b/

[49]M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan,S. Gunasekar, M. Harrison, R. J.Hewett, M.Javaheripi,P.Kauffmann等，Phi-4技术报告，arXiv预印本 arXiv:2412.08905（2024年）。[50] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng, H. Zhao等，Chatglm：从GLM-130B到GLM-4 All Tools的大语言模型系列，arXiv预印本 arXiv:2406.12793（2024年）。[51] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7B （2023年）。arXiv:2310.06825。网址 https://arxiv.org/abs/2310.06825[52] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, J. Zhou，Qwen3 嵌入：借助基础模型推进文本嵌入与重排序，arXiv预印本 arXiv:2506.05176（2025年）。[53] J. Wang, X. Yi, R. Guo, H. Jin, P. Xu, S. Li, X. Wang, X. Guo, C. Li, X. Xu等，Milvus：专用向量数据管理系统，载于：2021年数据管理国际会议论文集，2021年，第2614–2627页。[54] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen等，LoRA：大语言模型的低秩适配，ICLR 1 (2) (2022) 3。