中图分类号：TP391　　单位代号：10280
密　　级：公开　　学　　号：23721623

上海大学

专业硕士学位论文

SHANGHAI UNIVERSITY PROFESSIONAL MASTER'S DISSERTATION

题目	面向长尾医学场景的检索增强方法研究及辅助诊断平台

作　　者　刘茂林
学科专业　软件工程
导　　师　王昊
完成日期　二〇二六年五月

姓　　名：刘茂林　　学号：23721623

论文题目：面向长尾医学场景的检索增强方法研究及辅助诊断平台

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学专业硕士学位论文质量要求。

答辩委员会签名：

主　　席：张伟

委　　员：

导　　师：

答辩日期：2026 年 6 月 1 日

姓　　名：刘茂林　　学号：23721623

论文题目：面向长尾医学场景的检索增强方法研究及辅助诊断平台

上海大学学位论文原创性声明

本人郑重声明: 所呈交的学位论文是本人在导师指导下, 独立进行研究工作所取得的成果。除了文中特别加以标注和致谢的内容外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他研究者对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名: 刘茂林

日期: 2026 年 6 月 9 日

上海大学学位论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即: 学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

(保密的论文在解密后应遵守此规定)

学位论文作者签名: 刘茂林　　导师签名:

日期: 2026 年 6 月 9 日　　日期: 2026 年 6 月 9 日

上海大学工学专业硕士学位论文

面向长尾医学场景的检索增强方法研究及辅助诊断平台

作　　者: 刘茂林
导　　师: 王昊
学科专业: 软件工程

计算机工程与科学学院
上海大学
2026 年 5 月

A Dissertation Submitted to Shanghai University for the
Degree of Professional Master in Engineering

Research on Retrieval-Augmented Generation Methods for Long-Tail Medical Scenarios and Auxiliary Diagnostic Platform

Candidate: Maolin Liu
Supervisor: Hao Wang
Major: Software Engineering

School of Computer Engineering and Science
Shanghai University
May, 2026

摘　要

大语言模型在自然语言处理领域取得了突破性进展，其在医疗等垂直领域的应用展现出了巨大的潜力。然而，由于医疗临床决策对准确性和逻辑严密性具有极高要求，LLMs 固有的幻觉问题成为了制约其落地的关键瓶颈。检索增强生成技术通过引入外部事实知识库，有效缓解了这一问题。但在面对错综复杂的真实医疗场景时，现有的 RAG 系统仍面临诸多严峻挑战：首先，传统的向量检索过度依赖表层语义匹配，极易在长尾复杂病理推理中导致推理漂移；其次，面对海量、异构的医学证据，系统缺乏有效甄别多源证据冲突与冗余噪声的机制；最后，学术界提出的诸多先进多智能体推理算法往往缺乏系统级的工程化整合与透明可视化的交互平台。针对上述痛点，本文以优化医疗检索增强生成方法为核心，从底层因果逻辑对齐、多源证据对抗质证以及系统平台工程化落地三个维度展开深入研究。本文的主要研究工作与创新成果如下：

（1）针对传统医疗检索受限于表层语义匹配、易导致引入噪音的问题，提出了一种基于因果感知伪问题生成与自适应检索迭代的医疗检索增强生成方法。该方法将结构因果模型引入非结构化医学文献的预处理阶段，生成具有明确逻辑指向的因果感知伪问题，从而将单纯的语义相似度匹配转化为基于病理逻辑的前向探测。此外，系统深度融合了 SNOMED-CT 与 UMLS 等权威医学知识图谱，提取标准病理逻辑链，并创新性地设计了因果图谱引导的自适应迭代检索机制。通过该机制能够检测非结构化文本与结构化图谱逻辑间的断层，动态触发补充查询，将传统 RAG 的单次静态召回升级为严密的动态循证补全过程，实验结果表明，该方法提升了模型在复杂长尾医学推理任务中的逻辑自洽性。

（2）针对多源医学证据间存在冲突与冗余、大语言模型难以甄别核心鉴别信息的问题，提出了一种基于对比假设和多智能体流的医疗问答检索增强生成优化方法。该方法首先构建竞争性候选假设与初始证据集，为复杂的临床表现发散推理边界并召回海量多源证据；其次，引入证据语义解耦与候选打分机制，对证据空间进行细粒度的特征级过滤，有效剥离缺乏临床鉴别力的共性干扰噪声；随后，设计了基于特征密度与语义重叠度的动态自适应路由模块，根据处理后医学证据特征进行候选

答案排序，并选择分发至对应的质证链路中；最后，构建多智能体定向质证与共识裁决工作流，驱动立场对立的智能体围绕候选疾病假设展开严格的对抗辩论。实验表明，该方法有效克服了冗余证据带来的推理漂移，显著提升了系统在复杂疑难病理鉴别中的准确率与特异性。

（3）针对现有医疗多智能体算法缺乏系统级整合与透明可视化交互的工程化痛点，设计并开发了整合多种检索方法的 AI 驱动医学研究平台（DeepMed Search）。该平台采用高可用、高并发的微服务架构，底层无缝打通了本地向量知识库（Milvus）、PubMed 权威文献接口及互联网搜索接口，彻底打破了医疗数据的信息孤岛。上层集成了异步多模态复杂文档解析引擎以及前述的因果检索与多智能体辩论机制。该平台将以往不可见的黑盒推理转化为包含智能路由调度、证据对比加权及流式对抗辩论的端到端白盒化工作流，为真实的临床辅助决策研究提供了一套高度透明、可信赖的智能化基础设施。

综上所述，本文围绕复杂医疗场景下的检索增强生成技术，从底层算法优化到全栈工程实践进行了全面、系统的探索。所提出的因果感知与多智能体验证方法，在多个权威医疗基准数据集上均取得了超越现有主流 RAG 架构的优异表现。本文的研究成果不仅丰富了垂直领域大语言模型的推理增强理论，更为安全、透明、可信的医疗人工智能辅助决策系统建设提供了更具价值的落地范式。

关键词：大语言模型；检索增强生成（RAG）；因果推断；多智能体系统；医疗辅助诊断

ABSTRACT

Large Language Models (LLMs) have achieved breakthrough progress in the field of Natural Language Processing (NLP), demonstrating significant potential in vertical domains such as healthcare. However, since clinical decision-making demands extreme accuracy and logical rigor, the inherent "hallucination" of LLMs has become a critical bottleneck for their practical application. Retrieval-Augmented Generation (RAG) technology effectively mitigates this issue by introducing external factual knowledge bases. Nevertheless, when faced with intricate real-world medical scenarios, existing RAG systems still encounter severe challenges: first, traditional vector retrieval relies excessively on surface-level semantic matching, which easily leads to "reasoning drift" in complex long-tail pathological reasoning; second, in the face of massive and heterogeneous medical evidence, systems lack mechanisms to effectively discriminate between multi-source evidence conflicts and redundant noise; finally, many advanced multi-agent reasoning algorithms proposed in academia often lack system-level engineering integration and transparent, visualized interaction platforms. Addressing these pain points, this thesis focuses on optimizing medical RAG methods and conducts in-depth research from three dimensions: underlying causal logic alignment, adversarial cross-examination of multi-source evidence, and the engineering implementation of the system platform. The main research work and innovations are as follows:

(1) To address the problem that traditional medical retrieval is limited by surface-level semantic matching and is prone to introducing noise, a medical retrieval-augmented generation (RAG) method based on causal-aware pseudo-question generation and adaptive retrieval iteration is proposed. This method introduces structural causal models into the preprocessing stage of unstructured medical literature to generate causal-aware pseudo-questions with clear logical directions, thereby transforming simple semantic similarity matching into forward probing based on pathological logic. Furthermore, the system deeply integrates authoritative medical knowledge graphs such as SNOMED-CT and UMLS to extract standard pathological logic chains, and innovatively designs an adaptive

iterative retrieval mechanism guided by causal graphs. This mechanism can detect gaps between unstructured text and structured graph logic and dynamically trigger supplementary queries, upgrading the single static recall of traditional RAG into a rigorous, dynamic evidence-based completion process. Experimental results show that this method enhances the model's logical consistency in complex long-tail medical reasoning tasks.

(2) To address the conflict and redundancy among multi-source medical evidence and the difficulty for LLMs to identify core differential information, an optimized medical QA-RAG method based on contrastive hypotheses and multi-agent flow is proposed. This method first constructs competitive candidate hypotheses and initial evidence sets to expand the reasoning boundaries for complex clinical manifestations and recall massive multi-source evidence. Second, it introduces evidence semantic decoupling and candidate scoring mechanisms to perform fine-grained feature-level filtering of the evidence space, effectively stripping away common confounding noise that lacks clinical discriminative power. Subsequently, a dynamic adaptive routing module based on feature density and semantic overlap is designed to rank candidate answers based on processed medical evidence features and dispatch them to the corresponding cross-examination paths. Finally, a multi-agent targeted cross-examination and consensus adjudication workflow is constructed, driving agents with opposing stances to engage in rigorous adversarial debates regarding candidate disease hypotheses. Experimental results show that this method effectively overcomes the reasoning drift caused by redundant evidence and significantly improves the system's accuracy and specificity in differentiating complex and difficult pathologies.

(3) To address the engineering pain points where existing medical multi-agent algorithms lack system-level integration and transparent visualized interaction, an AI-driven medical research platform (DeepMed Search) that integrates multiple retrieval methods is designed and developed. The platform adopts a high-availability, high-concurrency microservices architecture and seamlessly integrates local vector knowledge bases (Milvus), authoritative PubMed literature interfaces, and Internet search interfaces, completely breaking down "information silos" in medical data. The upper layer integrates an asynchronous multi-modal complex document parsing engine along with the aforementioned causal retrieval and multi-agent debate mechanisms. This platform transforms previously in-

visible black-box reasoning into an end-to-end white-box workflow comprising intelligent routing scheduling, evidence contrastive weighting, and streaming adversarial debate, providing a highly transparent and trustworthy intelligent infrastructure for real-world clinical auxiliary decision-making research.

In summary, this thesis conducts a comprehensive and systematic exploration of RAG technology in complex medical scenarios, from underlying algorithm optimization to full-stack engineering practice. The proposed causal-aware and multi-agent verification methods have achieved excellent performance on several authoritative medical benchmark datasets, surpassing mainstream RAG architectures. The research results of this thesis not only enrich the reasoning enhancement theory of LLMs in vertical domains but also provide a more valuable implementation paradigm for the construction of safe, transparent, and trustworthy medical artificial intelligence auxiliary decision-making systems.

Keywords: Large Language Models; Retrieval-Augmented Generation (RAG); Causal Inference; Multi-Agent System; Computer-Aided Medical Diagnosis

目　录

主要符号表

q, q_c	初始临床查询或复杂的患者主诉
$\mathcal{G}$	外部结构化医学知识图谱
$\mathcal{D}$	外部非结构化医学知识文献库
$\oplus$	字符串或上下文拼接操作符
$\mathbb{I}(\cdot)$	指示函数
$\mathcal{P}$	基础因果模式集合（涵盖单因单果、多因单果等）
$S(q)$	复杂查询 q 分解后的原子子意图集合
$\mathcal{Q}(q)$	实例化后的多视角因果伪问题集合
V_{seed}	初始种子节点集合
$\mathcal{R}_{diag}$	诊断相关关系集
$\mathcal{G}_{sub}$	针对特定查询提取的局部诊断相关概念子图
$\mathcal{L}_{cand}$	图谱挖掘的高置信度候选逻辑链集合
$\mathcal{C}^{(t)}$	第 t 轮迭代检索后的全局文本证据池
$s_{support}$	跨空间的综合证据支持度得分
τ_{logic}	触发自适应检索的逻辑自洽阈值
$\Delta^{(t)}(p)$	第 t 轮迭代检测出的逻辑断层（缺失边）集合
T_{max}	最大自适应迭代检索深度
$\mathcal{H}_{\text{comp}}(q_c)$	竞争性候选假设集合
h_k	大语言模型生成的第 k 个候选医学假设
r_k	候选假设 h_k 对应的具体推演理由
$\mathcal{E}_{\text{raw}}, \mathcal{E}$	初始召回的原始证据集与提纯后的去重证据集
M	外部证据片段与候选假设间的全局语义关

	联矩阵
$\tau_{\text{high}}, \tau_{\text{low}}$	用于证据语义解耦的高、低双重经验阈值
$\mathcal{E}_{\text{spec}}^{(k)}$	候选假设 h_k 专属的特异性（排他性）证据集合
$\mathcal{E}_{\text{shared}}$	缺乏真实鉴别力的全局共性干扰证据集合
$S_{\text{score}}(h_k)$	候选假设 h_k 基于解耦证据的综合置信度得分
$\lambda_{\text{spec}}, \lambda_{\text{shared}}$	特异性证据与共性证据的权重调节惩罚因子
Δ	头部候选假设与次优假设间的置信度得分残差
ω	候选假设推演理由间的语义重叠度
γ	候选假设的证据特异性密度
$\tau_{\text{base}}, \tau_{\text{dynamic}}$	自适应路由网关的基础阈值与动态判定阈值
$\pi(\cdot)$	系统的核心动态路由分流策略
$\mathcal{A}_{\text{pro}}, \mathcal{A}_{\text{opp}}$	定向质证中的正方专家智能体与反方专家智能体
$\mathcal{A}_{\text{judge}}$	负责逻辑结算的全局裁决者智能体
T	智能体差分质证的最大交互轮次
$\mathcal{H}_{\text{debate}}$	多智能体正反双方交叉质证的完整历史轨迹
D_{result}	系统最终输出的医学诊断结论
R_{chain}	包含关键鉴别特征与循证依据的解释性推理路径

第一章　绪论

1.1　研究背景和意义

大语言模型（Large Language Models, LLMs）[1]作为人工智能领域的突破性成果，凭借海量的训练数据与千亿级的参数规模，已展现出在自然语言理解与生成任务上媲美甚至超越人类的能力。如图 1.1 所示，最新大模型如 ChatGPT、Gemini 等模型如今能够处理多轮对话、文本摘要及逻辑推理等任务，为医疗、金融、法律等行业提供了全新的智能化解决方案。尽管 LLM 表现卓越，其在实际落地中仍面临严峻挑战：模型固有的幻觉问题[2]导致其容易生成似是而非的错误信息，且预训练数据的覆盖范围主要集中在通用语料，难以涵盖特定垂直领域深层的长尾知识与私有数据，这在对专业深度与准确性要求极高的领域（如临床诊疗、法律判决）成为了制约其应用的关键瓶颈。

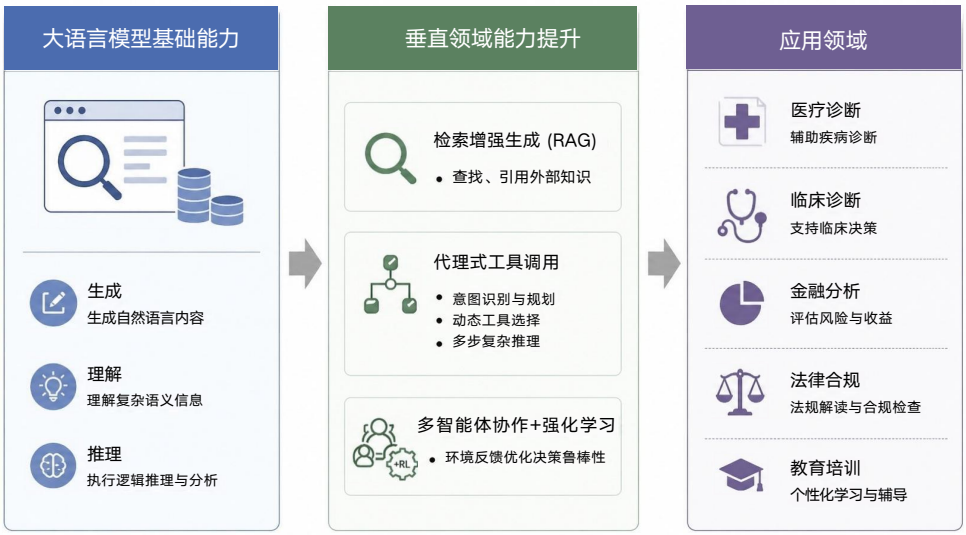

图 1.1　大语言模型基础能力、垂直增强技术与下游应用架构图

针对上述局限，检索增强生成（Retrieval-Augmented Generation, RAG）[3] 技术作为一种高效的工程化范式应运而生，为解决 LLM 的幻觉与知识更新问题提供了新的办法。RAG 技术利用检索组件从外部知识库中获取相关上下文，并将其与输入查询融合，引导 LLM 生成基于事实的回答以提升回复的可信度与准确性，在智能问答、私域知识库构建及辅助决策等场景中取得了显著效果。然而，现有的 RAG 架构并非

完美。一方面，主流检索模型主要依赖语义相似度进行匹配，容易引入噪声干扰。在处理专业性极强的问题时，误导性噪声会干扰大语言模型的判断；另一方面，面对需要多步推理的复杂长尾问题，单次检索往往会导致关键信息缺失，使得传统 RAG 系统在面对深层次的领域知识问答时，难以构建准确且完整的推理路径。在面对复杂医疗垂直领域时，上述局限性显得尤为突出。**长尾医学场景指的是在临床问答和医学推理中，由于疾病发生频率较低、训练语料覆盖不足、检索证据稀疏或症状表现高度混淆，导致模型难以通过常规语义匹配获得充分可靠证据的任务场景**。面向长尾医学场景的问答则具有两个特点：一是相关专业文献与知识分布相对稀疏；二是长尾疾病的外在临床表现往往与常见病存在较高的特征重叠，通常需要借助多跳的病理推演才能排查根本病因。这种特性使得传统 RAG 方法在医疗应用中暴露出明显的局限性：临床决策对事实准确性与逻辑可解释性要求严格，但现有的 RAG 架构通常依赖表层语义的相似度匹配[4-5]，难以有效解析医学知识中结构化的实体关联与深层因果网络[6]。当面对长尾医学查询时，由于相关证据较少，单次且单向的流水线式检索很容易被数量庞大的常见病文本所覆盖。在这种缺乏中间逻辑约束与动态验证机制的条件下，大模型容易被高频的共性特征所干扰陷入语义相关性陷阱，最终偏离正确的推演路径并给出具有潜在风险的误导性建议。因此，如何使 RAG 系统在长尾医疗场景中具备排除干扰噪声的鉴别能力与自适应纠错机制，已成为当前医疗大模型走向实际应用亟待解决的实际问题。

未来，垂直领域的 RAG 与 LLM 应用研究将朝着以下几个方向发展：Agentic RAG（代理式检索增强）[7]，即引入智能体架构，使模型具备自主规划、工具调用及多步迭代检索的能力，实现从被动查询到主动探索的转变；多智能体协作与强化学习优化[8]，通过引入环境反馈等方式，提升模型在复杂决策中的逻辑鲁棒性；以及领域知识深度融合[9]，研究如何更好地整合结构化与非结构化的知识结构、利用权威与非权威的知识生成可解释性更强的推理，以适应医疗、法律等特定领域的严谨性要求。

1.2 研究问题

本文研究的对象是面向医疗诊断和知识问答的检索增强生成技术与因果推理机制。由于医疗诊断过程具有高度的逻辑严密性以及证据链条的复杂性，现有的医疗

问答系统在实际应用中往往面临三大核心挑战：用户意图识别困难、多源异构知识融合困难以及长尾疑难病例幻觉风险高。现有的模型往往只关注文本语义的匹配，而忽略了医学知识中内在的因果逻辑以及诊断过程的迭代交互特性。因此，本文主要关注以下三个核心问题：

(1) 针对多源异构知识难以对齐的问题，如何在检索阶段组织利用不同结构的知识，以实现相关医学证据的精准捕获？

研究发现，尽管检索增强生成技术通过引入外部知识显著缓解了大模型的幻觉问题[3]，但在复杂医疗场景下，由于关键证据碎片化地分布于结构化知识图谱与非结构化医学文献中，现有的检索机制面临着相关性瓶颈。目前的方案大多未能充分挖掘异构知识结构间的互补价值：知识图谱擅长刻画实体的逻辑路径并蕴含潜在因果关联，而医学文献则蕴含更深层的语义细节。由于缺乏对不同知识结构的协同利用，系统在检索时难以打破拓扑逻辑与语义特征的壁垒，导致检索结果无法精准召回深层逻辑相关的关键证据。因此，如何利用因果逻辑整合结构化的医学图谱与非结构化的文献文本，实现不同类型数据的有效结合，从而提升医学证据的检索精度与召回率，是本文拟解决的第一个核心问题。

(2) 针对用户意图识别困难的问题，如何实现医疗查询意图与专业检索语料的逻辑对齐？

在真实的临床场景中，患者或医生的原始查询往往存在非标准化表述的现象，导致系统面临严重的意图识别困难。现有的 RAG 模型通常利用双编码器（Bi-Encoder）或交叉编码器（Cross-Encoder）计算查询与文档的相似度[10-11]。虽然这些模型在通用领域的开放域问答中取得了较好效果，但它们仅依赖浅层语义匹配，缺乏对医疗实体间隐性逻辑解析的能力，无法在底层实现复杂查询意图与专业医学语料库之间的精准映射。如果系统无法在检索初期准确对齐用户意图与知识空间，后续召回的证据将偏离用户需求。因此，如何深度解析复杂语义，构建查询意图与底层医学知识之间的因果映射以实现检索时的逻辑对齐，是本文需要解决的第二个核心问题。

(3) 针对长尾医疗问答中模型易受冲突证据干扰的问题，如何引入多智能体交互机制以弥补单向推理的不足？

虽然医疗大模型在处理常见病例时表现出较好的能力[12]，但在涉及长尾医学知识（如罕见病或非典型症状）时，由于相关语料稀疏且容易与常见病特征混淆，模型产生事实错误的风险较高。尽管部分研究引入了自我反思机制来提升系统的稳定

性[13]，但常规的单向检索与生成架构在面对多源且可能存在冲突的医学证据时依然缺乏有效的交叉验证手段，难以客观评估证据的可靠性。因此如何构建基于多智能体协作的推理框架以提高模型在长尾医疗问答中的鉴别能力与准确性，是本文拟解决的第三个核心问题。

1.3 研究内容

本文的主要研究内容依托图 1.2 展示的研究架构，总体分为底层逻辑因果对齐、多源证据对抗质证以及医疗辅助平台工程化实现三个维度。具体而言，本文围绕医疗大语言模型在复杂诊断中的检索与推理痛点，开展了以下三个方面的系统性研究与构建：

（1）为了解决医疗场景下用户意图识别困难、异构知识难以充分利用等问题，我们提出了一种基于因果路径约束与用户查询伪问题生成的多轮检索增强方法

针对医疗场景下查询意图复杂、异构数据难以协同应用的问题，本部分重点阐述如何构建一个实现因果逻辑对齐的混合检索预处理与动态召回系统。首先，设计基于大语言模型的文档解析管道，在非结构化医学文献的预处理阶段，利用权威医学知识图谱（如 SNOMED-CT、UMLS）提取标准病理逻辑链，并为非结构化知识切片生成具有明确指向性的因果伪问题。其次，构建支持结构化与非结构化数据协同利用的混合检索池，对查询进行问题分解并生成逻辑伪问题；同时优化传统的混合检索匹配算法，将因果伪问题检索纳入召回链路。最后，设计并实现图谱引导的自适应迭代检索策略，定义系统的动态查询触发条件与循证补全逻辑，使检索模块能够自动检测逻辑断层并执行补充召回，从而引入更多与用户需求逻辑相关的有效证

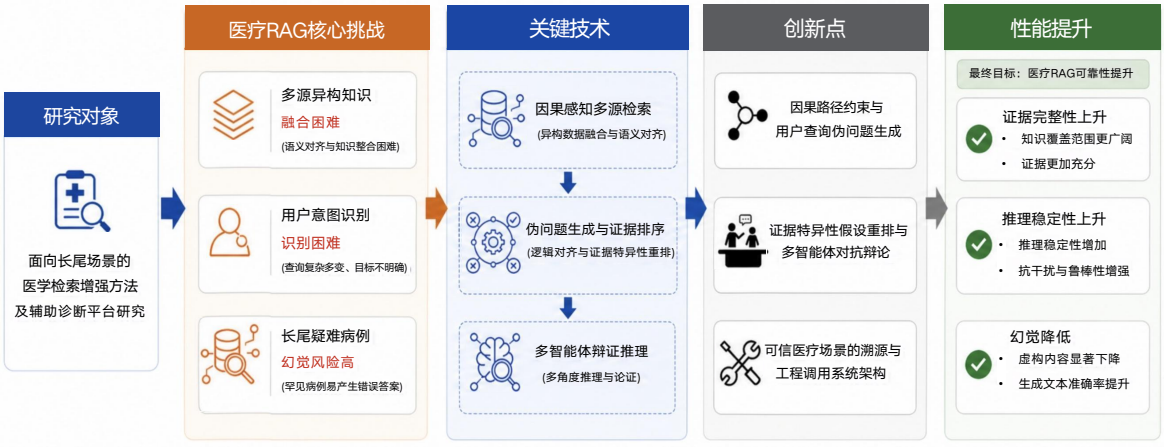

图 1.2 本文的总体研究架构，包括研究对象、研究内容、创新点等。

据。

（2）为了解决长尾医疗场景中模型易产生幻觉且缺乏自我纠错能力的问题，我们提出了一种基于证据特异性重排与多智能体协作交叉验证的长尾检索增强方法

针对长尾医疗问答中，单向生成架构易受共性干扰证据误导、缺乏有效验证环节的不足，本部分重点介绍如何构建证据解耦与多智能体协同的验证机制。首先，提出竞争性候选假设生成与证据特异性解耦方法，对初步检索到的文本中去除缺乏鉴别力的常见病共性特征，保留具有明确指向性的特异性证据。其次，设计多智能体质证流程，为不同的大语言模型实例分配对应的立场与角色并制定智能体间的交叉验证与共识规则，引导模型围绕高混淆性的候选假设与特异性证据展开辩论论证，从而实现对长尾问答下幻觉的有效缓解。

（3）为了解决医疗算法与实际临床业务流脱节、外部工具调度能力匮乏等问题，我们设计并实现了集成多种工具调用与复杂工作流的医疗辅助诊断研究平台

针对医疗辅助决策中工具调用离散化、推理黑盒化等工程难题，本部分重点介绍 DeepMed 平台的系统架构与功能实现。首先，采用微服务架构构建底层支撑系统，集成多源异构数据网关，实现对本地知识库和搜索引擎的管理调度。其次，在系统中引入工具调用机制，支持系统调用多模态解析器、医学搜索引擎及启发式规划与自我反思模块的能力，从而突破知识边界。最后，通过构建证据溯源与前端交互呈现模块，将检索、解耦、辩论等后台逻辑转化为白盒化的诊断链路，实现从算法逻辑到可解释辅助诊断系统的构建。

1.4 创新点

本文主要针对复杂医疗诊断与问答任务中大语言模型的推理漂移与相关证据验证等问题，从检索意图的因果逻辑对齐和多智能体交互验证反馈两大维度展开研究。本文的主要创新点具体表现在：

(1) 提出查询伪问题生成检索与因果约束的自适应迭代检索机制，实现检索证据与病理逻辑的深度对齐：第一个创新点是通过注入因果伪问题，实现检索意图的逻辑重构与对齐。现有的检索优化方案多使用生成常规伪问题或补充元信息来辅助更精准的检索，但这些方法往往缺乏对垂直领域深层逻辑的感知。在此基础上，本文提出了一种深度对齐医疗知识结构的因果伪问题生成方法，其目的是使系统在检索

时，能够按照更符合医学诊断规律的因果链条来定向召回相关文本证据，从而克服共性偏差等问题。本创新点的关键技术在于：我们首先提出了一种基于大语言模型的文档预处理方法，实现对非结构化医学文本的深层逻辑解析与伪问题生成；随后，我们构建了一套混合的多源异构（结构化知识图谱 KG 与非结构化文档）检索系统，将因果伪问题检索深度融入该系统中，并设计了自适应迭代循证策略。通过融合多种不同的检索方式，实现多源异构知识的优势互补，从而实现更精准、全面的检索效果。除此之外，我们将该系统在多个权威医疗问答数据集上进行评估，实验结果证明了本创新点可以有效解决语义与因果错位导致的推理失效难点，从而提升模型对复杂医疗知识的理解与推理能力。

(2) 提出基于证据特异性假设重排与多智能体对抗辩论的验证范式，实现长尾问答下对干扰证据的有效分辨：针对长尾医疗问答中证据逻辑冲突严重、传统单向线性架构缺乏中间纠错机制的问题，本创新点的关键技术在于：首先，针对复杂的临床查询构建竞争性候选假设空间，并引入证据语义解耦与特征过滤机制。通过对初始召回的证据提取特异性特征，从而有效剥离冗余信息中的共性干扰噪声。其次，设计了基于证据特征密度的动态自适应路由模块，将处理后的医学证据按逻辑属性分发至定向质证链路中。随后，构建了多智能体对抗辩论工作流，驱动具备对立立场角色的智能体围绕争议焦点展开质证与交叉验证。通过设计这种工作流，系统能够主动捕捉常规证据与特异性证据间的差异特征，有效模拟了医疗专家组在复杂疑难病例会诊中的递进式辨析过程。实验表明，该范式提升了模型在处理高冲突、高噪声长尾医疗任务时的逻辑自洽性和问答表现。

(3) 设计并实现了具备过程溯源与可视化交互的医疗辅助决策系统，提升了算法在实际应用中的可解释性：现有的医疗大模型研究多侧重于算法指标的提升，在系统实现时往往缺乏对推理步骤的展示，导致诊断过程不够透明。为此，本文开发了 DeepMed 医疗辅助决策平台。该平台基于微服务架构，实现了对外部检索工具和底层大模型的统一调度。其核心工作是将后台的异构图文检索、证据特异性解耦以及多智能体辩论等算法的执行过程，转化为前端页面上清晰、可读的步骤展示。这种系统设计不仅规范了各类诊断工具的调用流程，也为构建透明、可信的医疗人工智能辅助系统提供了一种具有实用价值的工程实现方案。

1.5 本文的组织架构

本文针对复杂医疗诊断任务中大语言模型的推理漂移与证据冲突问题展开研究。本文首先介绍医疗大模型与检索增强生成的研究背景和意义，梳理已有的主流 RAG 架构，分析医疗智能问答的复杂性以及引入因果逻辑推理的必要性。本文主要研究了基于医疗知识结构的因果伪问题生成及多源检索机制，以及基于证据特异性排序与多智能体辩论的长尾检索增强方法，设计并实现了一套集成搜索引擎与深度研究的医疗辅助诊断系统，最后总结了本文的研究工作并进行展望。

第一章绪论作为文章的开篇，主要介绍面向长尾医疗垂直领域的大语言模型与检索增强生成技术的研究背景和重要性。本章将概述现有 RAG 模型在处理复杂病例时所面临的共性偏差挑战和现有研究的局限性，进而引出本研究旨在解决的检索诱导推理漂移与证据链缺失等问题，并详细阐述本文的研究内容与创新点。

第二章将深入探讨本文相关的基础理论与相关工作。首先介绍了大语言模型与检索增强生成技术的发展脉络；紧接着分析了传统检索模型与现有医疗智能问答系统的研究现状及存在的不足；最后，介绍了结构化知识图谱范式以及多智能体协作理论的相关内容，为后续章节的方法设计提供坚实的理论支撑。

第三章中，研究将聚焦于基于因果伪问题和图谱路径约束的多轮检索方法。本章将针对传统检索易受语义噪声干扰的问题，探索如何通过大模型预处理技术生成符合医学知识规律的因果伪问题。通过结合结构化医学知识图谱（KG）与非结构化医学指南，旨在提出一种多源异构混合检索机制，使系统能够精准召回具备鉴别性的核心医疗证据。

第四章将深入研究基于证据特异性排序与多智能体协作交叉验证的复杂推理方法。针对大模型在处理冲突证据时极易产生逻辑幻觉且缺乏自我纠错能力的问题，本章首先探讨如何构建竞争性候选假设，并通过证据语义解耦与特异性过滤机制剥离共性干扰噪声。在此基础上，重点阐述多智能体辩论与共识裁决工作流的设计与实现。本章的研究目的是分辨共性干扰，并通过对抗辩论与交叉验证机制，赋予模型自主纠错能力，从而在复杂疑难病理的鉴别中有效抑制幻觉，提升推理结果的准确率与逻辑自洽性。

第五章将结合前两章的方法研究成果，设计并实现一套面向医疗领域的智能辅助诊断与检索系统。本章将详细介绍系统的整体架构与功能模块，展示如何将因果

检索机制、外部通用搜索引擎接口以及多智能体深度研究链路进行工程化集成，从而打造一个具备可靠性的实用化医疗信息检索与问答平台。

第六章将对全文进行总结，并对未来的研究方向进行展望。本章将回顾前述章节的主要发现和技术贡献，并客观讨论当前研究工作在实际应用中的局限性及未来改进的可能方向。通过总结，本文旨在为高可靠性医疗大语言模型的落地以及下一代主动式 RAG 系统的构建提供新的视角和工程化思路。

1.6 本章小结

本章作为全文绪论，综合阐述了本文的研究背景与意义，指出了医疗大模型在实际应用中面临的幻觉风险、异构知识割裂及推理黑盒化等严峻挑战。针对现有传统架构的局限性，本文提炼出检索精准捕获、冲突证据验证及系统工程落地三个核心难题，并对应提出了“基于因果伪问题构建和路径约束的多源异构知识并发检索机制”、“基于证据特异性解耦与多智能体对抗辩论的验证范式”以及“面向可信医疗的白盒化架构范式（DeepMed）”三大核心研究内容与技术创新。最后，本章简述了全文的章节组织脉络，为后续的基础理论介绍与核心方法论证奠定了清晰的框架基础。

第二章　相关理论与研究方法

2.1　大语言模型与检索增强生成（RAG）技术

2.1.1　大语言模型在医疗垂直领域的应用与局限

近年来，以 Transformer[14] 架构为基础的大语言模型在自然语言处理领域取得了突破性进展。得益于海量的训练数据与千亿级别的参数规模，以 GPT-4[15]、Gemini[16] 等为代表的通用大语言模型展现出了卓越的上下文理解、多轮对话交互以及文本生成能力。这些模型不仅刷新了各项自然语言处理任务的基准测试记录，也标志着人工智能从单一任务的感知智能向通用认知智能迈出了关键的一步。而随着通用大模型能力的日益成熟，研究人员开始积极探索其在垂直领域的应用潜力，医疗健康领域便是其中最具社会价值且极具挑战的落地场景之一[12,17-18]。针对通用大模型在专业医学知识上的欠缺，学术界与工业界相继推出了专门针对医疗领域进行预训练或微调的医疗大语言模型。早期的医疗语言模型主要**基于预训练模型进行领域自适应**。例如，Alsentzer 等人提出的 ClinicalBERT 模型[19]，通过在 MIMIC-III[20] 临床重症监护数据库上进行二次预训练，显著提升了模型处理电子病历等临床非结构化文本的能力。

随着生成式大语言模型的发展，模型在医疗专业领域的认知能力实现了质的飞

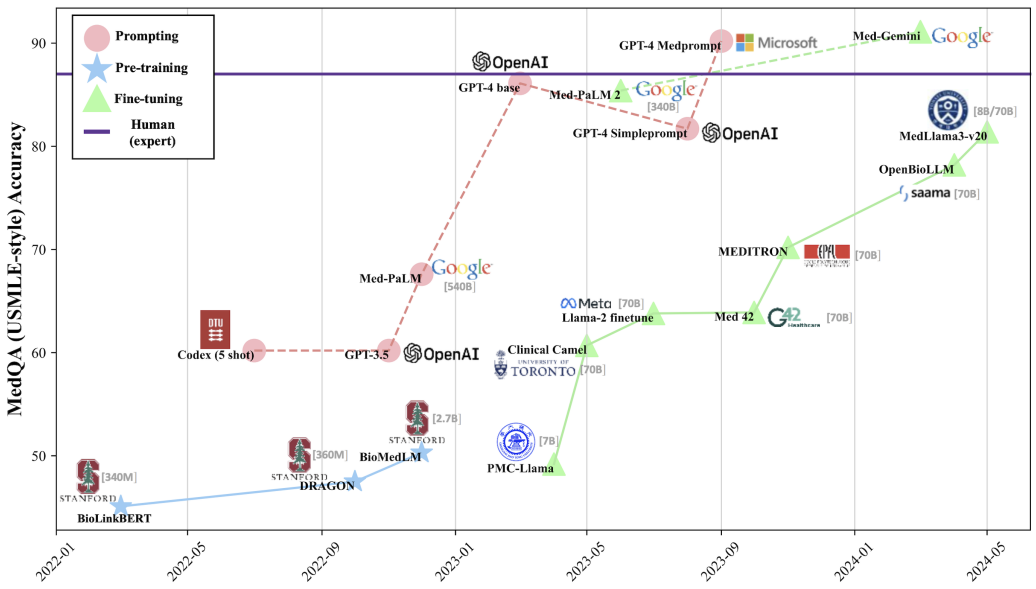

图 2.1　医学大语言模型在 MedQA (USMLE-style) 测试中的准确率随时间变化趋势[21]

跃。如图 2.1 所示，医学大语言模型在 MedQA[22]（USMLE 风格）客观评测中的表现呈现出惊人的跨越式增长：2022 年初的早期模型准确率尚不足 50%，而短短两年后的 2024 年中，顶尖大模型的表现已逼近甚至超越了人类专家的准确率基准线。从图中的技术路线分布可以清晰地看出，当前的研究重心已发生转移，**领域微调（Fine-tuning）与提示词工程（Prompting）**已成为提升医学大语言模型性能的核心手段。一方面，以 Medprompt[23] 为代表的提示词工程方法，通过引入动态少样本检索与自我生成的思维链等深度提示技术，在完全无需更改大模型底层参数的情况下，极大地激发了通用大模型的医疗认知潜力并达到了极高的准确率；另一方面，图中的发展轨迹也充分印证了**参数规模扩展结合领域微调**（如图中绿色三角所示）是当前构建专业医学大模型的另一条高效且成熟的技术路线。例如 Google 发布的 Med-PaLM 系列[12]、MEDITRON[24] 以及 MedLlama3[25] 等模型，均是通过在海量高质量医学语料上进行指令微调与参数更新，将其专业诊断性能推向了行业前沿。这些进展毋庸置疑地证明了生成式大模型在医疗知识编码与深层理解方面的巨大潜力。

在中文医疗大模型的研究方面，随着开源基座模型的普及，研究人员提出了多种**基于高质量医学知识库与多轮医患对话数据的微调方法**。例如，Wang 等人提出的华驼（Huatuo）模型[26]，通过注入中文医学知识图谱数据进行监督微调，提升了模型在医疗问答中的专业性；Li 等人开发的 ChatDoctor [27]则引入了疾病-症状数据库，以增强模型在模拟问诊场景下的表现。这些现有研究表明，通过特定领域的知识注入，大语言模型已能够胜任分诊导诊、病历摘要抽取及基础医学科普等常规任务。

然而，随着相关研究的深入[28]，学术界逐渐发现大语言模型在应对复杂临床诊断与长尾疑难病例时存在难以逾越的瓶颈。在医疗问责与幻觉风险方面，大语言模型极易放大训练数据中固有的社会与人口统计学偏见，并生成看似流利却违背医学事实的医疗幻觉内容[29]。这种现象在医疗领域尤为致命，因为任何诊断输出的偏差都将直接影响患者的生命安全；且由于生成式系统缺乏明确的责任追溯与底层纠错机制，在发生医疗失误时极难进行合理的责任界定[30]。在透明度方面，作为包含千亿参数的复杂黑盒神经网络，大语言模型固有的不可解释性极大地限制了其在严谨医疗领域的应用[31]。人类医生的决策往往依赖于清晰、可循证的病理因果证据链，而现有大模型在端到端生成时缺乏类似白盒的中间审查机制与推理透明度，难以让医疗专业人士信任其生成的决策路径。因此，如何突破大语言模型内部参数的局限，引入外部可信证据并施加医疗因果逻辑约束，成为了当前需要解决的核心科学问题。

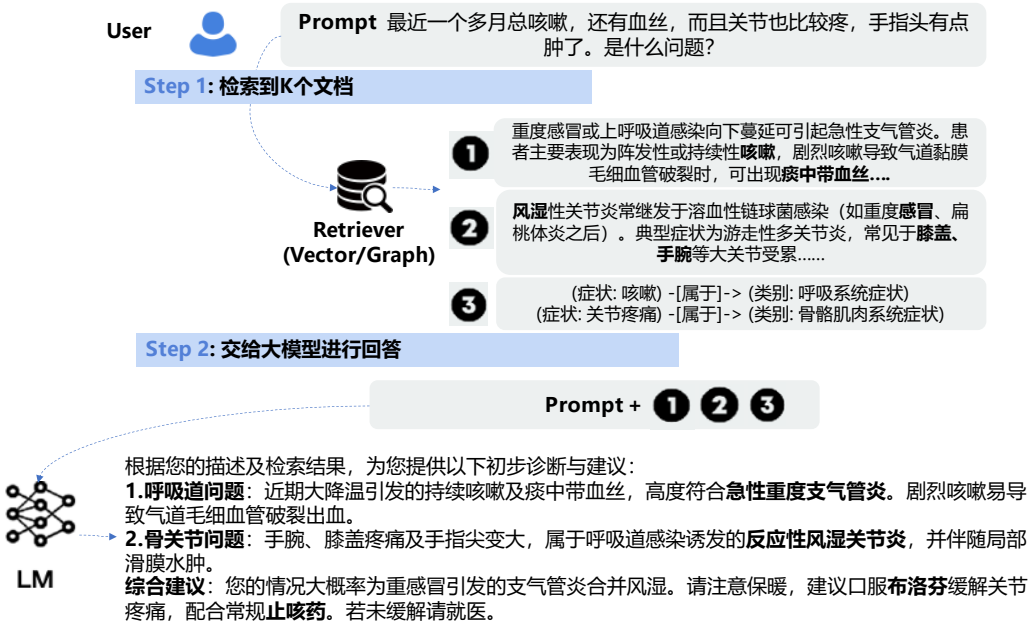

图 2.2　标准检索增强生成（RAG）的基础工作流程实例

2.1.2　检索增强生成（RAG）的基本范式

针对前文所述大语言模型在垂直领域面临的知识幻觉、长尾知识匮乏以及可解释性缺失等严峻挑战，检索增强生成技术作为一种极具前景的工程化范式应运而生。RAG 的核心思想最早由 Lewis 等人[3]系统性提出，其本质是将大模型内部封闭的参数化记忆与外部开放的非参数化记忆相结合。通过在模型生成回复之前，引入一个显式的外部知识检索阶段，RAG 能够将用户查询与动态更新的领域知识库进行实时对接，从而强制大语言模型基于外部召回的可靠事实依据进行回答。

从概率建模的角度来看，RAG 将检索到的文档 z 视为隐变量。给定输入查询 x，RAG 模型利用组件驱动的非参数化记忆（即检索器 $p_\eta(z|x)$）与参数化记忆（即生成器 $p_\theta(y|x,z)$）共同协作。根据 Lewis 等人的定义，生成目标序列 y 的概率分布可以表示为：

$$p_{RAG}(y|x) \approx \sum_{z \in top-k(p_\eta(\cdot|x))} p_\eta(z|x) p_\theta(y|x,z) \tag{2.1}$$

其中，$p_\eta(z|x)$ 负责从知识库中检索出与查询 x 最相关的 k 个文档切片，而 $p_\theta(y|x,z)$ 则负责在给定原始查询和检索上下文的情况下生成答案。这种架构允许模型在不重新训练参数 θ 的情况下，通过更新外部索引来实时扩展其知识边界。

标准的 RAG 架构（通常被称为朴素 RAG，Naive RAG）是一个典型的线性单次

流系统。如图 2.2 所示，以一个包含持续咳嗽、痰中带血、关节疼痛和手指肿胀等症状的医学查询为例，系统首先根据用户输入从向量库或知识图谱中检索出若干相关文档片段，包括呼吸道感染、支气管炎、风湿性关节炎以及症状类别等信息；随后将原始问题与检索结果共同拼接为增强提示词，交由大语言模型进行综合分析与回答生成。该流程体现了 RAG 的基本工作方式，即通过外部知识检索为生成模型提供上下文依据，使模型能够围绕召回证据组织回答。

（1）数据索引阶段（Indexing）：构建非参数化知识库

受限于大语言模型的上下文窗口，系统需先对海量多源异构文本（如医疗文献）进行清洗与分块。随后，利用预训练文本嵌入模型（如 BGE[32] 或 OpenAI 嵌入模型[33]）将离散文本块映射为高维稠密向量，并连同元数据持久化存储至向量数据库（如 FAISS[34] 或 Milvus）中，为后续相似度检索提供基础。

（2）信息检索阶段（Retrieval）：基于语义特征的知识召回

系统接收用户查询后，使用相同的嵌入模型将其转化为查询向量。接着，在向量库中通过计算空间距离（如余弦相似度）进行近似最近邻搜索，召回语义最相关的 Top-K 个文本块。此稠密检索机制突破了传统词频匹配（如 TF-IDF[35]、BM25[36]）的字面局限，能够有效捕捉查询与文档间的隐式语义关联。

（3）融合生成阶段（Generation）：上下文增强的答案合成

系统将用户原始查询与召回的 Top-K 文本块通过提示模板拼接，构建上下文增强提示词。大语言模型基于此外部事实证据扮演阅读理解与推理者角色，克服了对纯内部参数的依赖，从而生成准确、连贯且具备事实追溯能力的回答。

尽管标准 RAG 缓解了知识幻觉，但其检索-阅读-生成单次线性流水线在应对复杂垂直领域专业问答时暴露了三大缺陷：首先，**高度依赖初始查询质量**，面对模糊或多跳问题时，单次语义匹配易导致关键信息漏召或引入无逻辑关联的文档。其次，**缺乏知识过滤机制**，Top-K 召回结果无差别拼接使模型极易受噪声干扰，产生迷失在上下文中（Lost in the Middle）现象[37]。最后，**单向静态架构限制了复杂推理**，系统无法在证据不足时主动求证或自我纠错。因此，为提升复杂知识密集型任务的鲁棒性，学术界正推动 RAG 技术向具备查询重写、重排序及迭代交互的高阶检索增强范式演进。

2.1.3 检索增强技术的发展

随着单次线性检索架构在复杂知识密集型任务中的局限性日益凸显，学术界开始对检索增强生成技术进行深度的模块化解耦与流程重构，促使 RAG 技术向高阶范式（Advanced RAG）与模块化范式（Modular RAG）演进[38]。高阶检索增强技术不再局限于简单的文本召回与拼接，而是将优化重心向检索链路的两端以及大模型的交互机制延伸，主要涵盖检索前优化、检索后处理以及主动式迭代检索三个核心发展方向。

（1）检索前优化：查询重写与意图对齐（Pre-retrieval Optimization）

在面对复杂的用户提问时，原始查询往往存在语义模糊或关键信息缺失的问题。因此，检索前优化的核心目标是对用户的输入意图进行重构与扩展，以提高检索召回的准确率。当前主流的技术包括查询重写、查询扩展以及查询转换。

例如，Gao 等人提出了基于大语言模型的伪文档生成技术（如 HyDE[40]，Hypothetical Document Embeddings），该方法先让大模型根据原始问题生成一篇假设性回答（即伪文档），随后利用该假设性回答的稠密向量去知识库中进行检索。这种方法通过从问题空间向答案空间的映射，显著提升了语义匹配的精准度，但仅依赖模型自身生成的查询扩展会导致扩展内容与目标检索语料库之间出现对齐偏差。为了解决这一痛点，Lei 等人在此基础上进行了优化，提出了一种基于语料库引导的查询扩展方法（Corpus-Steered Query Expansion, CSQE）[39]。如图 2.3 所示，CSQE 方法

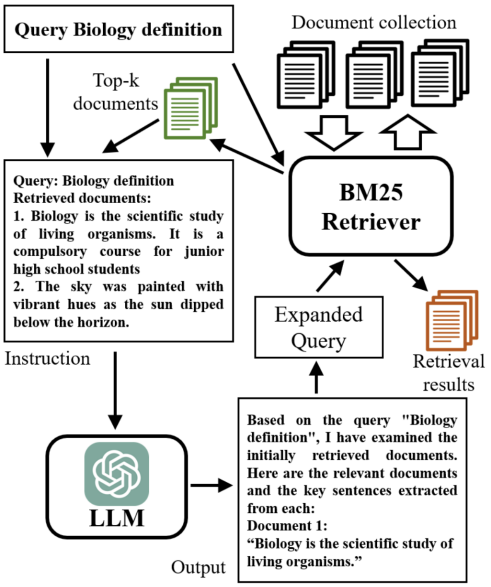

图 2.3　基于语料库引导的查询扩展（CSQE）工作流程图[39]

借鉴了传统的伪相关反馈机制[41]，旨在促进语料库中固有知识与扩展查询的深度融合。具体而言，该方法首先利用大语言模型强大的相关性评估能力，对初始检索召回的文档进行系统性审查，从中精准提取出与查询高度相关的关键句子。随后，系统将这些源自真实语料库的关键文本，与大模型基于自身知识生成的扩展内容相结合，共同重构为最终的查询扩展。通过引入语料库中真实存在的上下文作为事实锚点，CSQE 策略不仅大幅降低了幻觉生成的概率，还显著改善了原始查询与目标文档之间的相关性预测效果，在无需额外训练大模型的情况下即可实现检索性能的跃升。

此外，基于**查询路由（Query Routing）**的策略通过大语言模型的意图识别动态调度检索资源，主要体现在两个维度：一是**数据源结构路由**，系统借鉴 Toolformer[42] 等范式，自适应地将实体关系查询分发至结构化知识图谱，而将长文本释义路由至向量数据库；二是**链路复杂度路由**，如 Adaptive-RAG[43] 引入轻量级分类器预测查询复杂度，将其精准映射至三级处理链路：无须外部知识的常识问题直接触发大模型生成，单一事实查询进入单次检索流水线，跨文档多跳问题则触发检索-反思-再检索的多步迭代。这种动态路由机制不仅避免了简单问题因过度检索引入无关噪声，更提升了复杂问答的逻辑严密性，在最终回答的准确率与计算开销（如 Token 消耗和响应时间）之间实现了最优平衡。

（2）检索后处理：重排序与上下文压缩（Post-retrieval Processing）

针对检索阶段不可避免引入的噪声文档以及大语言模型面临的上下文窗口限制，检索后处理成为了提升生成质量的关键屏障。其中，重排序（Reranking）技术被广泛应用以解决初始召回精细度不足的问题。在初始的稠密检索阶段，为了保证海量数据的检索效率，通常采用双编码器架构分别对查询和文档进行独立编码，这种双塔结构导致两者在表征生成时缺乏深层的信息交互。

为了弥补这一缺陷，重排序阶段通常引入一个额外的交叉编码器模型进行细粒度的二次计算。以经典的基于 BERT 架构的重排序模型（如 monoBERT[44] 或当前开源前沿的 BGE-Reranker）为例，其核心技术原理是将原始查询与初步召回的单篇候选文档拼接为一个统一的输入序列（即构造为 [CLS] Query [SEP] Document [SEP] 的格式），而后整体送入深度 Transformer 网络中。这种输入范式打破了独立编码的隔离，使得模型能够在所有隐藏层中，通过自注意力机制实现查询词汇与文档词汇之间深度的、细粒度的双向特征交互。最终，模型通过顶层 [CLS] 标记的输出表征，预测出一个高精度的匹配分数，并据此对 Top-K 个文档重新排序。这种机制能够将

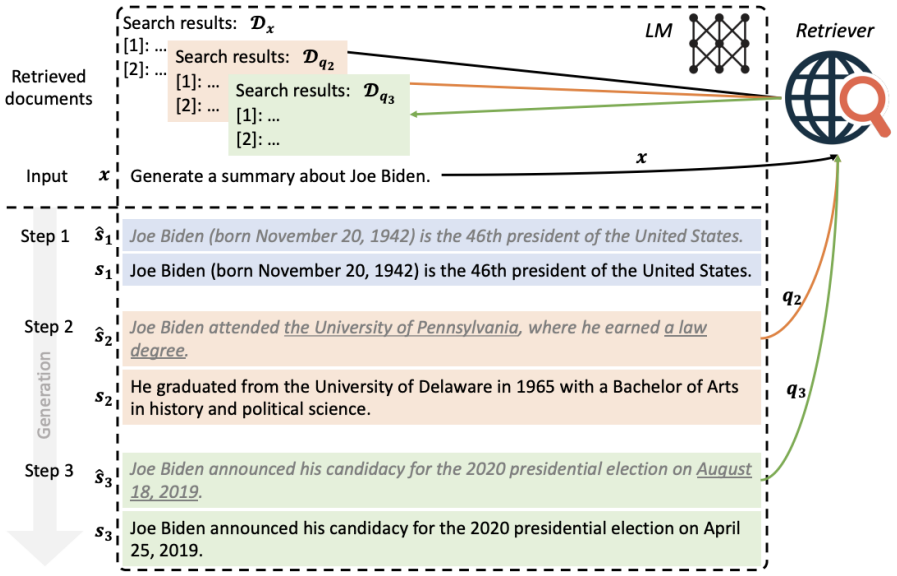

图 2.4 FLARE（前向主动检索）模型迭代与递归生成流程示意图[45]

最具价值的核心证据极其准确地推至上下文的前列，从而有效缓解了大语言模型迷失在上下文中现象。

（3）模块化与主动检索范式：迭代、递归与自适应交互（Modular & Active Retrieval）

为了突破单向检索无法应对复杂推理链条的瓶颈，RAG 架构开始向具备感知与决策能力的模块化主动检索范式演进。

- **迭代与递归检索（Iterative & Recursive Retrieval）：**现有的前沿模型打破了传统的一次性检索、后向生成的静态顺序，赋予了大语言模型在生成过程中动态获取知识的能力。如图 2.4 所示，以代表性工作 FLARE（Forward-Looking Active REtrieval）[45] 为例，该模型提出了一种前向主动检索机制。在生成长文本或解答复杂问题时，FLARE 会实时监控模型预测下一个句子中词汇的生成概率（即内部置信度）。当系统探测到即将生成的句子中包含低于预设阈值的低置信度词汇时，便会主动暂停当前生成；随后，系统会将这句临时生成但不确定的文本作为前向线索，反向构造查询语句触发二次或多次外部检索。在获取到相关的精准事实文档后，模型再结合这些新证据重新生成并覆盖原本低置信度的句子。这种边推理、边求证、按需检索的动态交互过程，不仅有效避免了无关检索对流畅生成的打断，更极大地增强了模型应对多跳（Multi-hop）复杂问答与长篇事实性文本生成的能力。

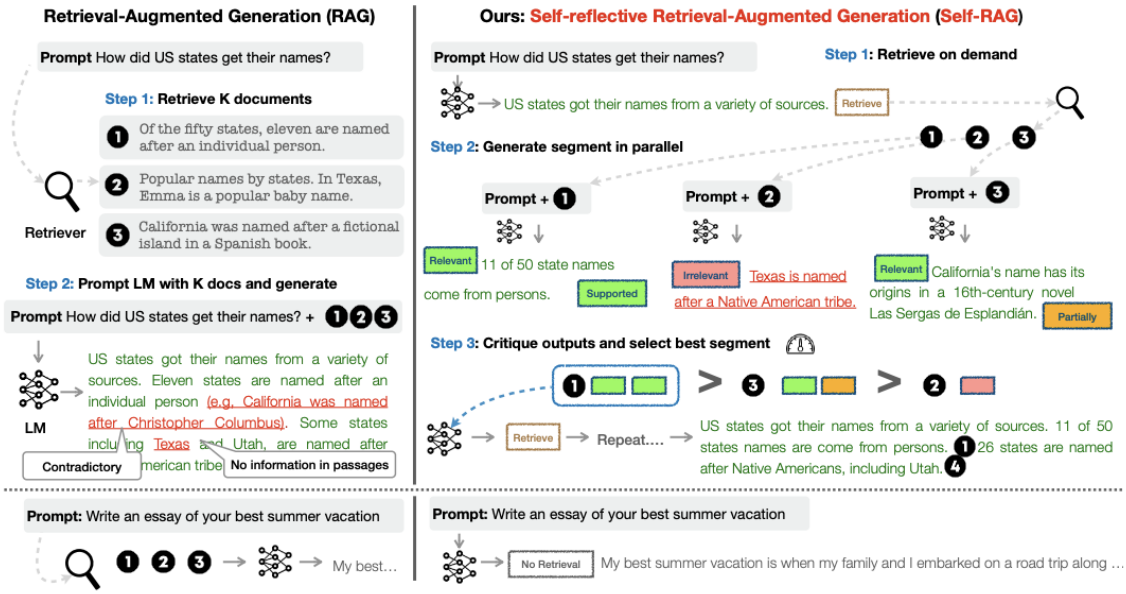

图 2.5　自反思检索增强生成模型（Self-RAG）工作流程示意图[13]

- **自适应与代理式检索（Adaptive & Agentic RAG）：**随着智能体（Agent）技术的发展，检索增强框架逐渐演变为由大语言模型充当中央处理器的自适应系统。以自反思检索增强生成模型（Self-RAG）[13] 为例，该框架通过引入特殊的反思标记（Reflection Tokens）和自我批判机制，实现了从被动生成到主动逻辑评估的范式转变。在推理阶段，Self-RAG 将过程划分为按需检索-并行生成-细粒度批判-动态解码四个环节。其核心决策逻辑在于对候选生成片段 y 进行加权评分，以选择最优推理路径。根据论文定义，给定输入 x 和检索文档 d，片段 y 的评分公式为：

$$\hat{y} = \arg\max_{y} \sum_{r \in \mathcal{R}} w_r \cdot \text{Softmax}(P(r^*|x, d, y)) \tag{2.2}$$

其中，$\mathcal{R}$ 代表反思标记集合（如 *IsREL, IsSUP, IsUSE*），r^* 表示每个标记中最理想类别的预测概率，w_r 为人为设定的超参数权重，用于平衡相关性、支持度与效用性。在整个推理工作流中，模型首先进行**自适应按需检索**，通过实时预测检索标记来评估自身参数化知识是否充足；若检测到知识盲区则主动触发外部检索器，从而有效避免针对简单问题过度检索而引入的噪声。对于召回的 Top-K 文档，系统随即进入**细粒度自我批判**阶段，模型会结合文档并行生成多个候选片段，并显式预测三种维度的反思标记：**IsREL**（评估文档相关性）、**IsSUP**（审查生成的文本是否完全由外部证据支撑，严防幻觉）以及 **IsUSE**（评

估片段的综合效用打分）。最终，系统采用**基于分数的分段束搜索**，利用上述评分公式对所有候选路径进行动态排序与剪枝，精准筛选出逻辑严密且事实准确的最优序列。这种将按需检索与评价分数权重解码耦合的代理式架构，显著提升了生成内容的事实准确性与逻辑一致性。

2.2 医疗领域信息检索

随着 RAG 技术的兴起，通用信息检索技术在开放域问答中也随之越来越被人们所重视，但医疗临床文本具有极高的复杂性。其不仅存在着严重的词汇鸿沟问题[46]（如患者口语化的心脏病发作与专业术语急性心肌梗死之间的语义对齐），还蕴含着海量嵌套的医学实体与复杂的并发症从属关系。为了提升专业临床文本的召回精度与逻辑严谨性，学术界针对医疗垂直领域发展出了一系列特定的检索增强技术。当前主流的技术路线经历了从符号化规则匹配到深度语义表征，再到结构化图谱推理的演进，本节将对这三大核心方法及其代表性工作进行深度剖析。

2.2.1 基于医学本体与自然语言处理流水线的规范化检索

早期的医疗垂直领域检索高度依赖于由医学专家精心构建的庞大医学本体库与受控词表，如统一医学语言系统（UMLS）[47]。

如图 2.6 所示，以 UMLS 为代表的系统构建了一个庞大的医学元叙词表，它不仅作为核心枢纽广泛整合了临床资源（如 SNOMED）、生物医学文献标签（如 MeSH），还交叉链接了解剖学、基因知识库（如 OMIM）等多个子领域的异构本体。这类技术的核心逻辑正是依托于这种高度结构化的网络，引入复杂的自然语言处理（NLP）

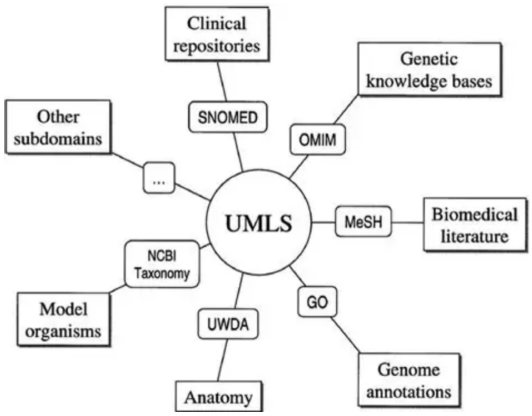

图 2.6 UMLS（统一医学语言系统）对多源生物医学本体与受控词表的整合架构

流水线与实体链接机制，对非结构化的用户查询与目标文档进行强制的概念规范化。

在具体的方法学实现上，此类系统通常包含一个多级级联的解析管道。以美国国家医学图书馆（NLM）开发的 MetaMap[48] 以及 Apache cTAKES（临床文本分析与知识提取系统）为代表的经典工具，展示了此类检索的底层工作流：首先，系统对用户的自然语言查询执行边界检测、分词、词性标注与浅层句法分析；随后，利用医学命名实体识别（NER）模块精准抽取疾病、解剖部位、药物等实体特征；接着，利用复杂的字符串匹配算法（如编辑距离计算）与形态学变体生成技术，将这些表层实体强行映射到标准本体库中的唯一概念标识符（Concept Unique Identifier, CUI）。进入检索阶段后，算法不再仅仅匹配原始字面词汇，而是基于 UMLS 的有向无环图（DAG）语义网络结构，自动沿边扩展这些标准概念的同义词、上位词与下位词来重构查询语句。

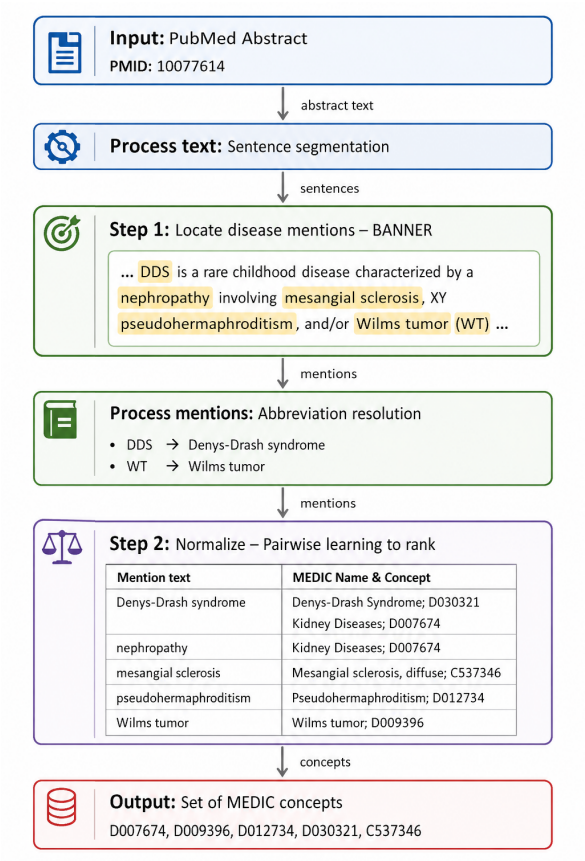

图 2.7 DNorm 系统处理医学文本的端到端概念规范化流水线架构[49]

为了更清晰地展示此类传统本体规范化方法的工作机制，如图 2.7 所示，以 Leaman 等人提出的经典 DNorm 系统[49] 为例。该系统采用典型的串行两段式流水线架

构：在第一阶段，系统首先利用 BANNER[50] 等前置命名实体识别工具，从输入的 PubMed 摘要中定位并抽取 DDS、nephropathy（肾病）等疾病提及，并执行缩写还原；在第二阶段，DNorm 创新性地引入了成对学习排序算法，将提取出的非标准表述精准映射到 MEDIC[51] 词典的标准概念与唯一标识符（如 D030321）上。这种机器学习算法的引入，极大地缓解了由非结构化表达带来的字面不匹配问题。然而，通过图 2.7 的工作流我们也可以直观地发现，此类重度依赖人工规则、先验知识与前置模块的检索方式存在三大致命弱点：第一，其串行流水线架构极易发生错误级联——如果在 Step 1 中前置 NER 模型漏抽或错划了实体边界，Step 2 的规范化算法将无从施展，错误会被直接放大并传递至最终的检索结果中；第二，本体库的构建与特征工程的更新维护成本极其高昂；第三，在面对口语化严重、缺乏显式实体边界或超出知识库覆盖范围的长尾罕见病查询时，往往表现出极低的鲁棒性，根本无法动态应对充满不确定性的真实临床场景。

2.2.2 适应医疗领域的稠密检索

随着深度预训练语言模型（PLMs）的爆发，研究重心全面转向面向医疗领域的稠密检索技术。与通用的双编码器（Bi-Encoder）架构不同，医疗稠密检索器通常需要基于特定领域的预训练基座进行权重初始化，以捕获专业医学词汇的深层分布式表征。

在基座模型的演进上，Lee 等人率先提出了 BioBERT[52]，在通用语料基础上继续使用 PubMed 摘要进行预训练；随后，Gu 等人提出的 PubMedBERT[53] 更进一步，彻底摒弃了通用语料，完全从头在 PubMed 摘要与 PMC 全文上进行预训练。这种策略使得模型在底层分词器与词汇表级别就实现了与医学语义的高度对齐，极大地减少了医学长专有名词被碎片化切分的情况。

更重要的是，医疗检索在高维向量空间中面临着极其严峻的困难负样本挑战。例如，1 型糖尿病与 2 型糖尿病、甲亢与甲减在字面上高度重合，但在病理机制与治疗指南上却截然不同。为了赋予模型对这种细微临床差异的判别能力，NCBI 团队在 2023 年提出了标志性的 MedCPT（Medical Contrastive Pre-trained Transformer）模型[54]。

如图 2.8 所示，MedCPT 创新性地引入了大规模用户搜索日志，并设计了一套将双编码器预训练与交叉编码器重排序紧密结合的两阶段对比学习架构。在第一阶

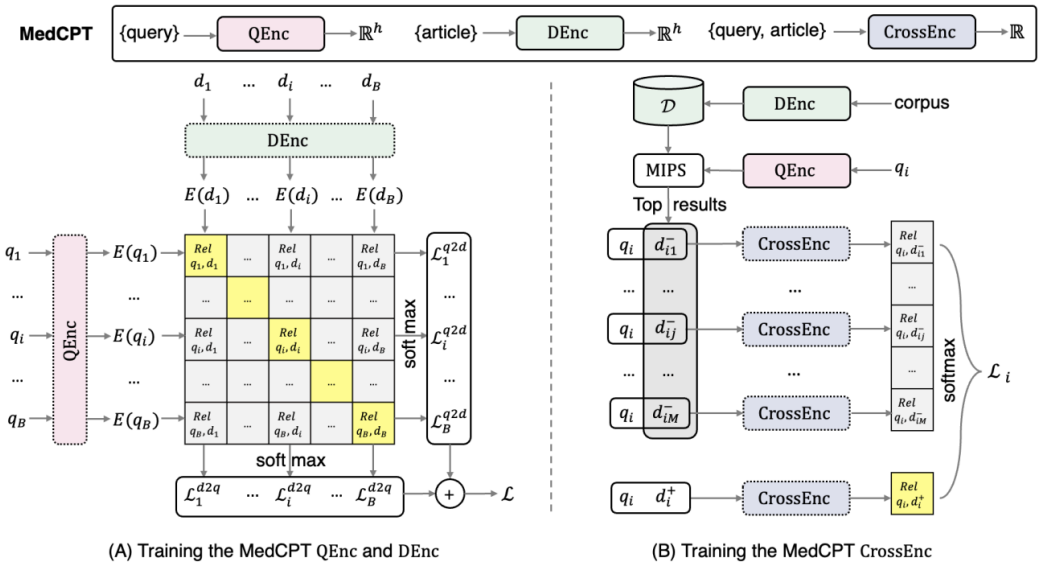

图 2.8 MedCPT 模型的两阶段对比学习训练流程概述[54]

段的表示学习中（图 A），系统对独立的查询编码器（QEnc）和文档编码器（DEnc）进行联合双塔训练，通过计算批次内所有查询与候选文档表征之间的密集相似度矩阵，并最小化对称交叉熵损失，使得模型能够在广阔的语义空间中初步拉近相关文献。然而，仅靠批次内的随机负样本难以区分高度相似的长尾医学实体。因此在第二阶段（图 B），系统利用首阶段训练好的双编码器在全量医学语料库上执行最大内积搜索，精准挖掘出字面高度相似但实际并不匹配的文档作为困难负样本。随后，系统将原始查询分别与这些困难负样本及真实正样本进行文本拼接，送入具有深层注意力交互能力的交叉编码器中，通过最小化联合的 Softmax 损失函数执行严苛的间隔惩罚训练。在最终的工程落地中，经过深层编码的高维医学向量被持久化存储于专用的向量数据库（如 Milvus 或 FAISS）中，利用分层导航小世界（HNSW[55]）等索引算法实现基于内积的毫秒级近似最近邻（ANN[56]）查询，而 CrossEnc 则作为高精度的顶层重排模块。这种双阶段范式使得医疗自适应检索器在海量医学文献的精准召回上达到了前沿水平，但其本质依然受限于分布语义学，无法进行跨文档的多跳显式逻辑推理。

2.2.3 融合结构化医疗知识图谱的子图检索

尽管稠密检索在特征匹配上取得了显著的进步，但这种方式仅能捕捉表层语义的高频共现，无法真正的感知深层病理因果与实体间的拓扑逻辑关系。在处理长尾复杂病例时，这种缺陷极易引发共性偏差与证据链断裂。为了弥补这一逻辑盲区，基

于图结构的信息检索技术在医疗领域日益受到重视，且相关研究强调了通过引入类型增强与多跳关系映射，能够有效提升复杂异构知识的融合效率与大规模问答的准确率[57-58]。

早期的图谱检索框架主要侧重于静态图谱的实体锚定与路径游走。为了在复杂的结构化网络中实现多跳推理，学术界进行了大量的前沿探索。例如，Yasunaga 等人提出的 QA-GNN 架构[59] 创新性地构建了文本语义与图谱拓扑的联合推理空间。该工作将语言模型对当前问答上下文的稠密表征作为一个特殊的上下文节点嵌入到局部知识图谱中，并与相关的实体节点建立跨模态链接。随后，系统利用图注意力网络（GAT[60]）在异构图上进行多轮的消息传递，动态计算图谱中每个节点相对于查询问题的注意力权重，从而在密集的拓扑网络中精准聚合并定位最关键的证据路径。

随着大语言模型推理能力的不断提升，相关研究逐渐由依赖图网络消息传递的范式，转向以大模型为核心的主动规划与推理机制。与此同时，为弥合结构化图谱与非结构化文本之间的信息鸿沟，图检索增强生成（GraphRAG）[61] 被提出用于支持复杂多跳推理。如图 2.9 所示，其流程分为索引与查询两阶段：在索引阶段，系统通过文本切块与多轮信息抽取构建局部图谱，并结合 Leiden 社区发现算法[62] 聚合高密度实体连接，自底向上生成节点、关系及社区层级的自然语言摘要，从而将图结构转化为具备语义表达能力的知识单元；在查询阶段，系统采用双轨机制，对全局查询利用 Map-Reduce 并行整合社区摘要，对局部查询则基于社区结构向下钻取，在实体子图中执行多跳推理。该聚类与摘要机制与医疗领域的 UMLS（统一医学语

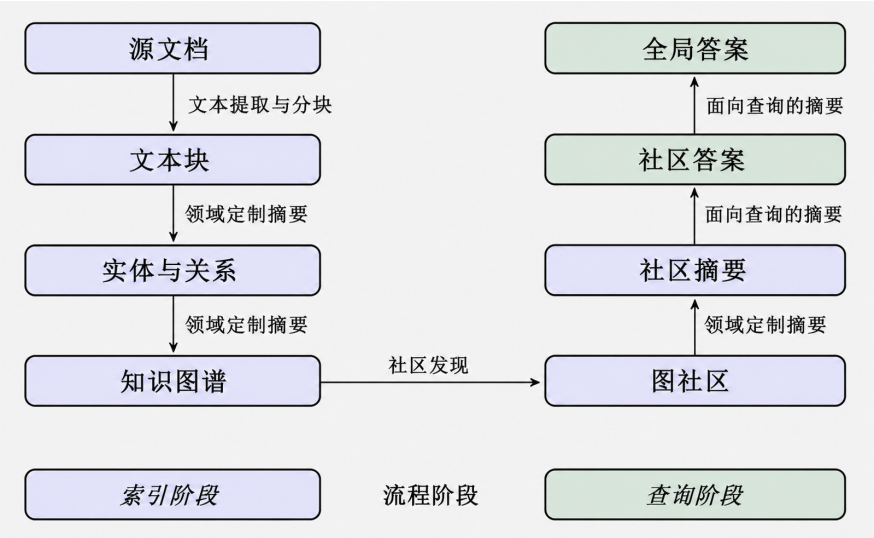

图 2.9　GraphRAG 范式的核心流水线[61]

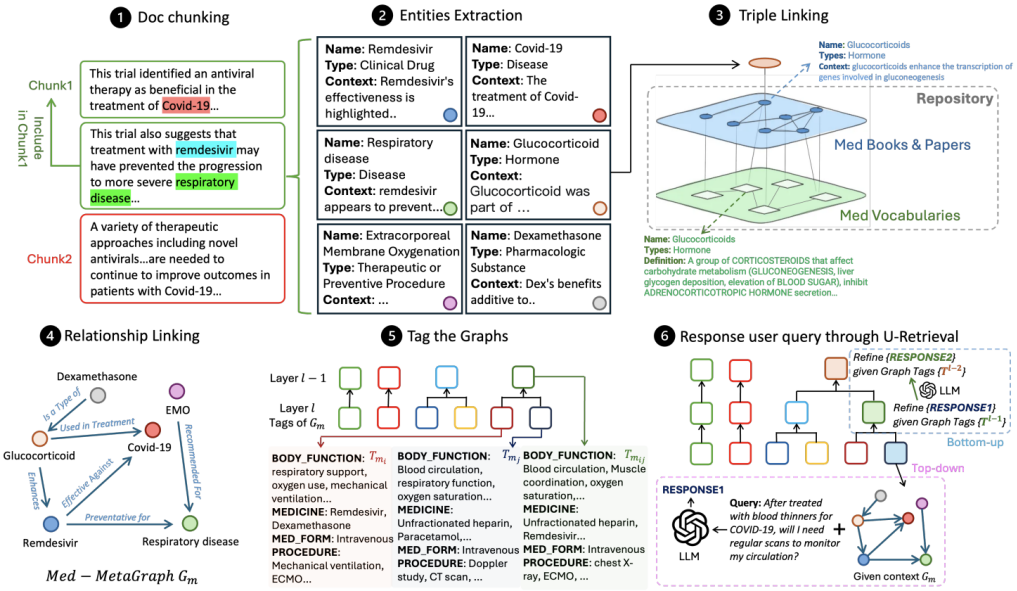

图 2.10 MedGraphRAG 的架构工作流[64]

言系统）等规范化词表具有良好的适配性。在开放域场景中，大模型抽取的图谱易受实体消歧问题影响，导致结构稀疏与社区碎片化[63]；而在医疗场景下，GraphRAG 可结合 UMLS 实体链接流程，利用 CUI 对异构表述进行对齐，将非标准症状映射为统一节点，从而提升子图连通性。高密度图结构在 Leiden 算法作用下，有助于将同一病理机制相关的症状、靶点及用药信息聚合至核心语义社区内，进而形成更符合循证医学逻辑的子图结构，为后续推理提供更稳定的语义与结构先验。

为将图检索架构更好地适配于医疗辅助诊断，Wu 等人提出了医疗专属的 **MedGraphRAG** 范式[64]。如图 2.10 所示，该方法在通用 GraphRAG 基础上进行了重构，避免仅依赖社区聚类导致医学概念层级被破坏的问题。具体而言，MedGraphRAG 面向医疗长文档构建具有严格医学逻辑的层级证据图：在图谱构建阶段，系统通过文档切块与实体抽取，引入权威医学词典与文献库作为事实锚点，完成三元组与关系链接，使非结构化病历中的表型特征能够准确映射至高阶医学概念，从而形成结构清晰、语义一致的医学元图。在图结构设计上，MedGraphRAG 引入多层级语义打标机制，将节点按医学本体归纳为身体机能、药物、用药形式和医疗程序等宏观类别，构建自下而上的层级拓扑结构。这一设计为后续检索提供了明确的语义组织方式。在推理阶段，MedGraphRAG 采用 U 型检索机制以增强长尾场景下的鲁棒性。系统首先执行自顶向下的路径筛选，在宏观语义层进行剪枝以过滤无关分支，降低语义共性噪声；随后在目标子图内执行自底向上的证据聚合，将底层体征、实体及其关系路

径逐步整合为紧凑上下文输入模型。该“宏观筛选—微观聚合”的检索策略，有效缓解了大模型在复杂证据环境中的推理负担，并为最终诊断提供了结构化、可追溯的证据支持。

然而，尽管 GraphRAG 与 MedGraphRAG 成功地在非结构化文本上动态建立起了全局的拓扑逻辑，但其图谱构建过程依然偏向于静态知识的整合。当系统同时检索到多篇结论相互矛盾的最新临床指南，或者图谱中的静态逻辑与患者的特异性动态体征发生冲突时，单一的生成式系统往往会陷入认知过载。因此，如何将图谱的结构化拓扑与非结构化指南的动态文本深度融合，并在检索系统内部引入能够处理冲突证据的多智能体辩论与迭代反馈机制，依然是当前构建高可靠性医疗检索系统亟待突破的困难。

2.3 高阶医疗 RAG 中的智能体验证与推理范式

尽管大语言模型在检索增强生成（RAG）阶段展现出强大的零样本涌现能力[65]，但在罕见病等长尾医学场景中，RAG 常召回碎片化甚至相互冲突的多源证据。由于单体模型在自回归解码时缺乏内部交叉验证，面对此类复杂上下文极易陷入认知过载与推理幻觉[66]。为突破单体模型的推理上限，学术界受“心智社会”理论[67]启发引入了多智能体协作范式。该范式通过精细化的角色分配，已在通用复杂逻辑推理任务中取得了极大成功（如 ChatDev[68] 与 AutoGen[69]）；同时，国内学者也相继提出了多智能体自主规划与协作评估框架[70-71]。这种角色分工逻辑与临床解决疑难重症的多学科联合会诊（MDT）制度高度契合[72]。因此，在医疗 RAG 系统中引入多智能体架构来模拟专家团队交互，其核心目的正是对长尾场景下召回的冲突证据进行辩论验证。通过这种内部对抗机制，系统能够有效清洗检索噪声、校准生成路径，成为提升医疗 RAG 鲁棒性的有效手段。

2.3.1 基于智能体的检索后证据协同评估

医疗多智能体系统在 RAG 框架中的初步应用，主要体现在对外部检索结果的多学科联合审阅（MDT）重构上。在这种范式下，底层的大语言模型结合召回的异构医学知识，被实例化为具备不同专长的虚拟专家智能体。

以 Tang 等人提出的 MedAgents 框架[73] 为代表，此类系统在完成初步的外部检索后并不急于直接生成答案，而是构建了一个多领域专家的动态协作网络，完美映

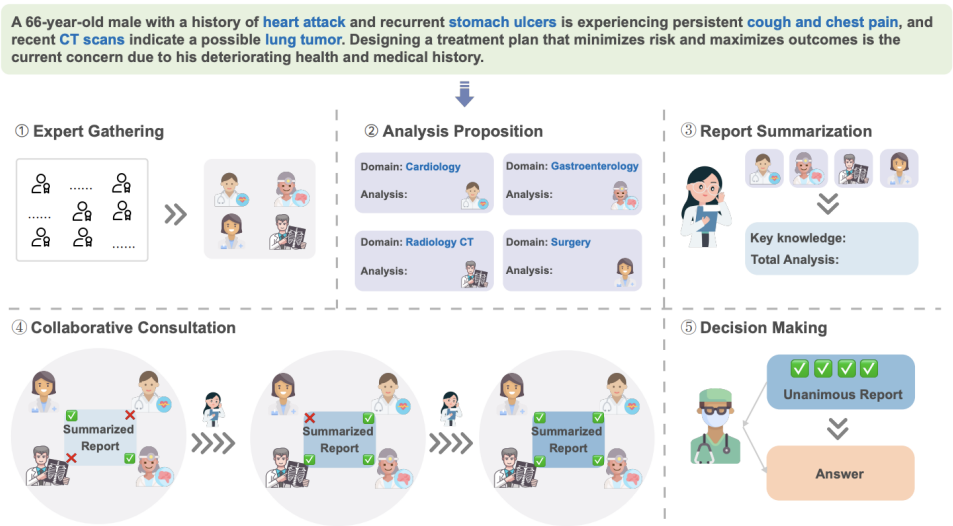

图 2.11　MedAgents 框架模拟多学科会诊的协作工作流[73]

射了真实的临床会诊流（如图 2.11 所示）：依次执行专家召集、独立分析，并对检索到的多源证据提取初步意见。随后进入核心的协同评估阶段，各智能体结合自身专长对同一批检索证据展开多轮交叉验证以消除逻辑分歧，最终输出达成共识的综合报告。这种基于角色解耦与迭代反馈的架构，有效降低了单体模型在处理长文本证据时的注意力偏移，显著提升了系统对复杂并发症相关知识的发现能力。

2.3.2　处理冲突证据的对抗性医学辩论与交叉验证

在应对长尾疑难重症时，常面临跨模态证据或医学指南间的底层逻辑冲突。针对智能体间矛盾的病理假设，简单的多数表决易导致诊断模糊。为此，医疗多智能体系统引入了对抗性辩论与交叉验证机制以解决深层次逻辑冲突，其典型落地范式如下：

DeepRare 框架[①]：在真实的罕见病诊断中，临床医生往往需要综合文本病历、结构化表型（HPO）甚至极其复杂的基因组测序（VCF）数据进行交叉验证。上海交通大学最新的研究提出了首个由大语言模型驱动的罕见病诊断智能体系统 DeepRare[74]。如图 2.12 所示，DeepRare 构建了分层协作架构，包括大语言模型驱动的中央主机、专职代理层（如表型/基因型分析器与检索模块）以及外部医学数据源。系统在推理前执行双轨信息采集，将患者输入解耦为表型（向量检索相似病例）与基因型（解析 VCF 并进行致病性评估）两条链路，并汇总至中央记忆库作为事实基础。其核心在于自反思诊断循环：在生成初步候选疾病后，中央主机触发内部博弈，通过外部

① https://raredx.cn

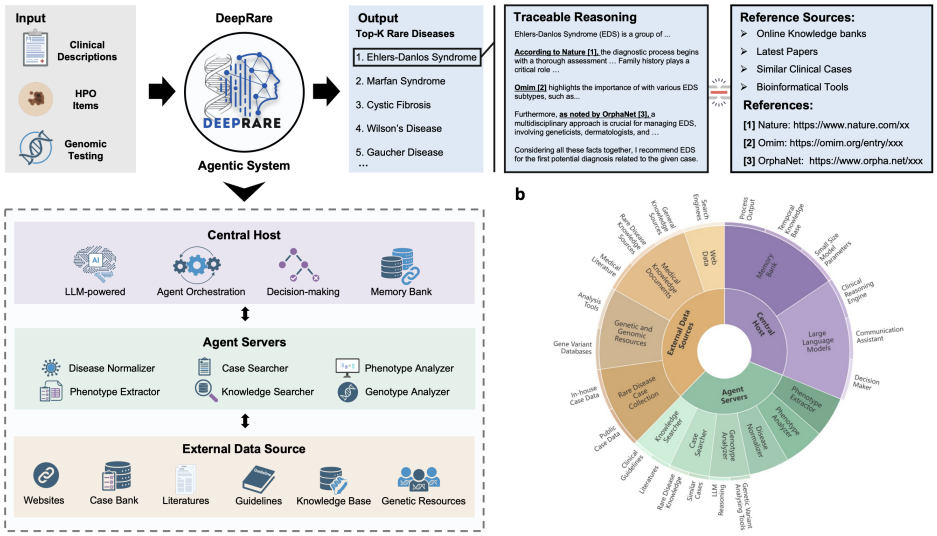

图 2.12　DeepRare 三层智能体架构与自反思诊断工作流[74]

证据检索与交叉质询对候选进行验证与证伪；若证据与患者特征冲突，则回溯并扩展检索空间重新推理。经过多轮迭代，系统输出带有可追溯推理链的诊断结果，每一结论均锚定至具体的 PubMed 文献或 Orphanet 条目。临床盲法验证显示，该机制的证据链准确率达 95.4%，在降低过度诊断与幻觉风险的同时，提高了推理过程的可解释性。

2.4　本章小结

本章围绕本文的理论基础与相关技术展开。首先，分析了大语言模型在医疗诊断中的应用现状及其面临的知识幻觉与可解释性问题，引出 RAG 技术，并梳理其从基础线性流程到高阶范式的演进路径。其次，结合医疗文本特点，对医学本体驱动检索、医疗自适应稠密检索及融合知识图谱的子图检索等方法进行了对比分析。最后，聚焦高阶 RAG 的生成与验证环节，讨论多智能体协同验证机制及其在处理长尾证据冲突中的作用。上述分析为后续因果感知图谱检索与多智能体对抗验证方法的设计提供了理论基础。

第三章　基于因果路径约束与用户查询伪问题生成的多轮检索增强方法

3.1　医疗复杂问答的因果失准问题与研究动机

在医疗人工智能领域，大型语言模型凭借强大的自然语言处理能力，正逐渐成为临床辅助决策的重要工具。然而，由于医疗场景极高的风险敏感性，LLMs 固有的幻觉问题成为了待解决的痛点。为此，检索增强生成被广泛引入，通过挂载外部专业医学知识图谱或文献库，以期大幅提升医疗问答的准确性与可解释性。然而，在真实的临床诊断中，疾病、症状与病理之间的映射并非简单的表层词汇共现，而是蕴含着深刻的隐式因果推理逻辑。为了直观地展示这一困境，图 3.1 对比了单纯依赖 LLM 与传统 RAG 架构在面对复杂医疗问答时的局限性。以急性心肌梗死为例，患者主诉可能仅为左肩放射性疼痛和出冷汗，这些表层症状的背后，隐藏着由心肌缺血导致交感神经兴奋及神经反射的深层因果链条。这个例子表明，医疗问答系统仅靠捕捉文本层面的语义相关性是远远不够的，这极易导致系统在检索阶段错失真正的病理链条，进而引入海量无关噪声并误导大模型的最终决策。因此，探究如何在 RAG 系统中实现从单纯的语义对齐向深层的因果逻辑对齐演进，对于提升复杂临床推理任务的可靠性具有极其重要的应用价值。

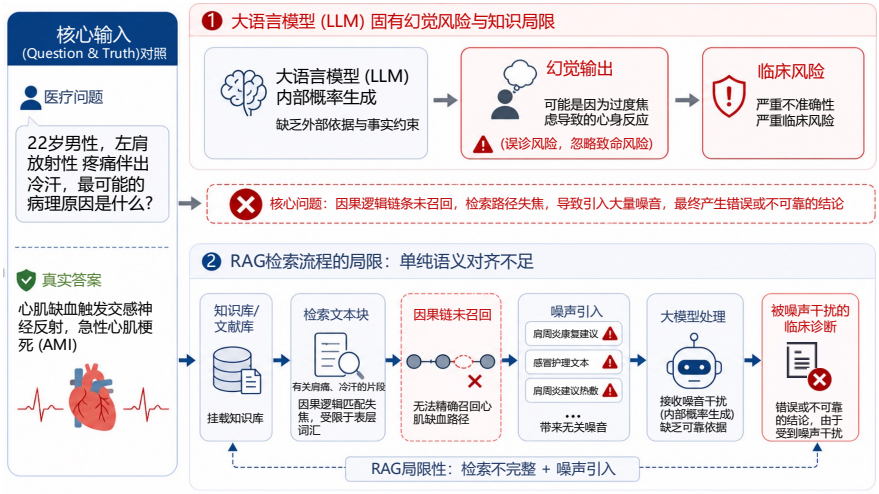

图 3.1　大语言模型与传统检索增强生成在医疗问答中的局限性对比。LLM 易因缺乏外部知识支撑而产生忽略致命风险的“幻觉”（上）；传统 RAG 受限于表层语义匹配，在检索阶段不仅难以对齐深层的真实因果证据，反而会引入海量无关噪声干扰最终的临床推断（下）。

面向医疗垂直领域的检索增强生成任务，广义上可以分为两类主流技术路线：1）基于向量检索的传统 RAG，这类任务主要依赖预训练文本嵌入模型，强调在海量医学文献中进行密集向量的语义相似度匹配；2）基于图谱检索的 Graph RAG，这类任务将结构化的医学知识图谱引入检索链路（如利用社区检测或子图匹配），专注于利用图结构来增强系统的多跳推理能力。其中，融合结构化知识与深度逻辑推理的复杂医疗问答场景为本文的主要关注点。

在医疗 RAG 领域，现有工作（如 MedCPT 及各类前沿的 GraphRAG 架构）在医学文献召回和结构化知识引入方面取得了较好进展。然而，在应对包含复杂临床因果关系的问答时，现有方法仍存在局限。观察现有的复杂医疗 RAG 系统，常出现两类逻辑失准现象：1）系统经常召回包含大量重叠医学术语、但因果方向完全错乱的文本块（例如将某药物的副作用误召回为某疾病的症状）；2）即使系统初步召回了部分正确的医学证据，由于缺乏对完整病理链路的校验，大量伴随而来的无关噪声文档极易误导大模型的最终诊断结论。这两类现象揭示了现有范式的深层缺陷：大语言模型在医疗场景中并非缺乏深层的因果推理能力，而是现有的检索机制过度依赖表面语义匹配，未能实现真正的因果逻辑对齐。这种由于证据对齐失败而引入的海量冗余噪声，导致系统在缺乏关键因果节点支撑的情况下，极易发生推理轨迹的严重漂移。在复杂的检索与匹配系统中，引入因果推断机制是消除数据混杂偏见、实现精准意图对齐的有效途径。[75]

基于上述背景，本文的研究动机为：提出一种基于因果路径约束与用户查询伪问题生成的医疗检索增强框架，旨在弥补传统检索方式在医疗领域的局限性，帮助模型在医学文本中更好地捕获完整的病理因果链路。为实现这一目标，本文分别从结构化和非结构化数据入手，设计了平行的检索与抽取路径：一方面，利用因果感知伪问题生成机制，将复杂的患者主诉转化为结构化的探寻线索，在非结构化文本中进行初始召回；另一方面，基于患者查询的核心实体，在医学知识图谱中提取标准逻辑子图。在此基础上，本文引入了自适应迭代检索机制。系统通过对比文本召回的证据与图谱中的标准逻辑路径，识别出证据链条中缺失的中间节点（即逻辑断层）。针对这些缺失环节，系统会自动生成新的补充查询并触发下一轮检索，不断完善证据链，直到文本证据能够覆盖图谱提示的因果路径或达到最大迭代次数。通过这种由图谱拓扑结构引导的迭代检索方式，本研究旨在将医疗检索从单次的语义匹配转化为动态的循证补全过程，从而提升大模型在复杂临床辅助决策中的准确性与

可靠性。

3.2 多轮因果约束检索增强模型整体框架

本章将系统介绍本文提出的基于因果路径约束与用户查询伪问题生成的多轮检索增强方法。该方法以“实现深层医疗因果逻辑对齐”为核心目标，通过预处理阶段的结构化约束、在线检索阶段的语义与逻辑双重对齐，以及图谱因果路径约束的自适应迭代机制，有效解决传统检索方法在医疗复杂语境的检索失准问题，从而帮助模型提供更相关、更有力的证据支持。整体系统架构如图 3.2 所示。本文提出的检索增强框架主要包含以下四个核心模块：

1. **因果伪问题预处理模块：**针对非结构化医学文献，在离线阶段引入医学先验模板（如一因多果、因果链条等典型推理模式），为切分后的文档片段生成具有明确逻辑指向的候选因果伪问题。该模块以此构建逻辑增强的外部知识索引，从底层数据层面弥补自然语言表述与隐式医学因果关系之间的表示鸿沟。
2. **复杂查询分解与混合索引模块：**在问答阶段，首先将非结构化的复杂患者主诉分解为多个原子查询意图，并将其映射至预处理阶段构建的因果伪问题空间。随后，依托多视角检索与混合分数校准机制，主动引导检索模型精准召回与查

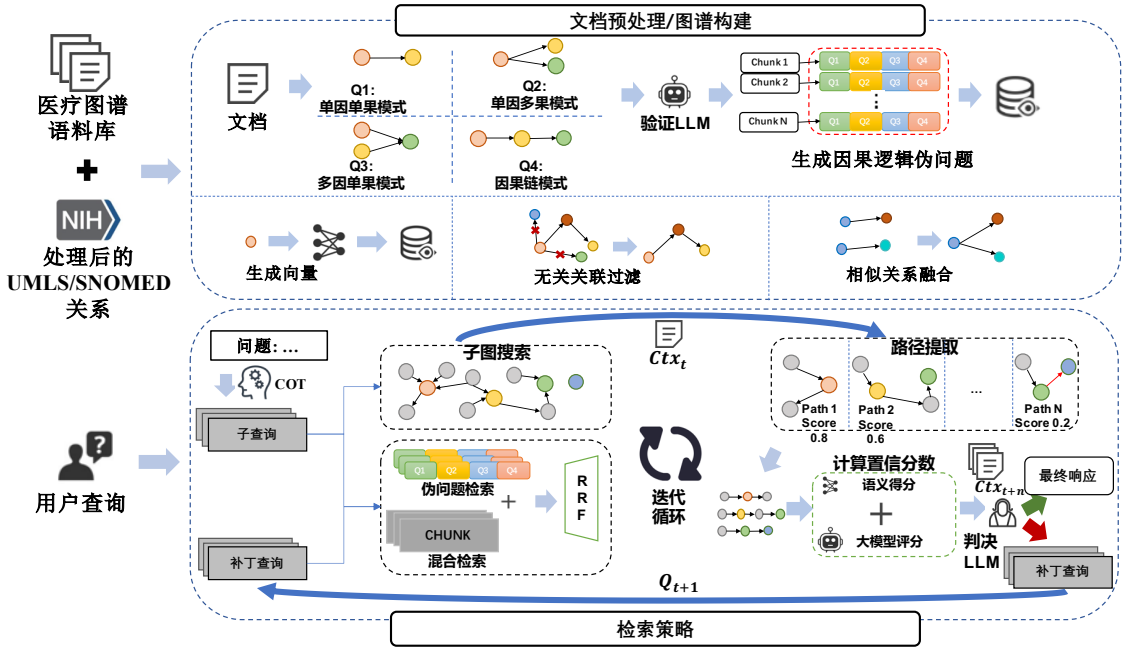

图 3.2 基于因果路径约束与用户查询伪问题生成的多轮检索增强方法总览。系统包含离线因果伪问题预处理、复杂查询分解与混合检索、结构化逻辑链挖掘，以及因果路径约束的自适应检索四大核心模块。

询具有深层因果关联的医学证据。

3. **局部子图与潜在逻辑链挖掘模块：**作为与文本检索并行的结构化召回路径，该模块依托外部医学知识图谱（如 UMLS 或 SNOMED-CT），搜索患者查询中的核心医学实体，并提取局部的相关概念子图。通过挖掘实体间的标准临床因果路径，为后续的推理提供结构化的事实依据。
4. **图谱因果路径约束的自适应迭代检索模块：**该模块负责对齐双路检索的结果。系统将文本召回的证据片段与知识图谱中提取的因果逻辑链进行跨结构比对，检测文本证据链中是否存在逻辑断层。一旦发现因果路径未闭环，系统将基于断层节点自动生成补充查询，并触发下一轮针对性的迭代检索，直至非结构化证据能够完整覆盖核心的病理逻辑链条。

3.2.1 数据来源及融合方法

在真实的临床辅助决策场景中，单一维度的文本匹配往往无法理清错综复杂的病理机制。因此，本文首先构建了融合非结构化文献与结构化医学图谱的多源外部知识库，为后续的因果逻辑链挖掘与多智能体辩论提供坚实的数据底座。为了保证图谱概念的广泛覆盖度以及临床层级关系的细粒度，本文融合了两个国际权威的医学标准术语集：

- **统一医学语言系统（UMLS）：** UMLS 是目前全球规模最大的医学概念元词表。本文抽取了其中经过语义网络严格定义的医学概念及其关联，以确保系统具备处理极其宽泛的医学实体（如罕见疾病、解剖结构、药理学物质）的能力。
- **系统医学术语命名（SNOMED-CT）：**鉴于 UMLS 的关联较为宏观，本文进一步引入了 SNOMED-CT 作为临床层级关系的补充。该图谱提供了极其细粒度的临床层级架构，特别擅长表达症状、体征与具体疾病之间的确切从属或并列关系。

在构建医疗外部知识库的过程中，不仅需要关注概念的覆盖面，更需要强调知识的纯净度与语义的一致性。针对 UMLS 与 SNOMED-CT 的异构特性，本文首先通过对三元组 (e_h, r, e_t) 进行分层采样，随后利用 GPT-5.1 作为领域专家进行启发式判别，自动化地剔除与临床因果推理无关的边缘关系。例如，通过过滤掉如 *has_subject_of_information* 等附属类的冗余关系，从而构建出精简且高质量的医疗候选三元组集。

为了进一步解决不同数据源间的同一实体不同命名问题，采用了基于语义相似度的融合策略实现节点归一化。具体而言，利用深度嵌入模型 BGE-M3 对所有节点进行语义编码，并在嵌入空间中计算节点 n_i 与 n_j 的余弦相似度 $sim(n_i, n_j)$。节点合并的准则定义为：$sim(n_i, n_j) > \tau_{ref}$，其中阈值 τ_{ref} 设置为 0.88。该阈值是基于 100 组人工标注的医学实体对进行实证分析后确定的平衡点，旨在确保语义对齐精准性的同时，有效减少因知识冗余导致的幻觉风险。

3.2.2 因果伪问题预处理模块

针对医疗领域查询中隐含的复杂推理需求，传统的密集检索方法往往因过度关注表面语义相似度而忽视了深层的逻辑关联，导致检索结果与真实的临床决策路径脱节。在真实的临床环境中，医生的认知过程高度依赖于“假设-演绎”范式——即根据初始线索生成初步的诊断假设，并进一步推演预期的临床表现以进行交叉验证。

为了给这种复杂的推理范式提供严谨的数学底座，本文引入了 Judea Pearl 的结构因果模型（SCM）。Pearl 在其理论中指出，传统的基于表示学习的检索方法本质上仅停留在第一层，即关联（$P(y|x)$）层面。这类方法只能捕捉文本序列的共现概率或表面语义的相似度，却无法区分虚假相关与真实因果。然而，真实的临床决策往往要求系统具备第二层干预（$P(y|do(x))$）的能力。SCM 通过引入有向无环图（DAG）和结构方程，明确刻画了医疗实体间的依赖拓扑（例如 $v_i = f_i(pa_i, u_i)$，其中 pa_i 为直接原因，u_i 为外生噪音）。如果检索系统缺乏这种拓扑感知能力，极易受到医疗文本中高频共现但无直接逻辑联系的混杂因素干扰，从而引发严重的语义漂移。

为了弥合传统检索在逻辑对齐上的根本缺陷，本文受上述理论的启发与约束，对伪文档增强策略（HyDE）进行改造，提出了一种因果感知查询扩展方法。该机制不再依赖大语言模型的无约束语义发散，而是将 SCM 的有向图拓扑与“假设-演绎”的动态思维相融合，将复杂的医疗推理严格抽象为具体的结构化引导模版。这种结合在增强语义匹配的同时，迫使检索过程服从底层的因果机理，从而提升逻辑对齐的概率。

为了系统化地刻画医疗逻辑，本文基于 SCM 的基础网络模体，定义了四种核心因果模式 $\mathcal{P} = \{P_1, P_2, P_3, P_4\}$，旨在精准覆盖临床推理中的典型拓扑结构：

- P_1（单因单果）：对应 SCM 中的直接因果链（$X \to Y$）。该模式用于描述如特定药物副作用等高度确定性的直接事实关系，旨在建立原子事实间的精准映

射，避免检索时的语义发散。

- P_2（单因多果）：对应临床“演绎”阶段的辐射式推演（$X \rightarrow \{Y_1, Y_2, \ldots, Y_n\}$）。针对一种核心病因引发多系统表现的综合征推理，确保检索系统能够召回单一节点的全部下游因果效应。
- P_3（多因单果）：对应临床“假设”阶段的聚合式溯源及 SCM 的共因结构（$\{X_1, X_2, \ldots, X_n\} \rightarrow Y$）。适用于涉及多项风险因素的复杂鉴别诊断，通过阻断混杂路径，有效提升高维临床线索下检索的特异性。
- P_4（因果长链）：对应 SCM 框架下的中介效应分析（$X \rightarrow M \rightarrow Y$）。该模式专用于模拟具有中间介导因素（$M$）的深层病理生理演变过程，为解决医疗长文本中的多跳推理提供了关键的逻辑锚点。

为了直观展示这一转化过程，表 3.1 列举了四种因果模式在实际医疗文本片段中的映射及伪问题生成案例：

表 3.1 用于因果伪问题生成的诊断模式示例

模式	**原始文本**	**因果伪问题**
p_1	Uncontrolled hypertension can rupture intracerebral vessels and cause hemorrhage.	Can hypertension directly cause intracerebral hemorrhage?
p_2	Hyperthyroidism increases metabolic activity and causes palpitations, tremor, heat intolerance, and weight loss.	Can hyperthyroidism explain these co-occurring symptoms?
p_3	Macrocytic anemia may result from vitamin B12 deficiency or hypothyroidism.	Do vitamin B12 deficiency and hypothyroidism cause macrocytic anemia?
p_4	Diabetes-related neuropathy and foot ulcers may progress to osteomyelitis.	Do neuropathy and foot ulcers from diabetes progress to osteomyelitis?

在离线预处理阶段，本文针对每一个原始文档块进行深度解构，强制其依据上述四种模式生成对应的结构化伪问题。这一转化过程成功将静态的文本知识重塑为动态的、具有明确因果指向性的逻辑单元，为下游的精准召回奠定了基础。

3.2.3 复杂查询分解与混合索引模块

在真实的医疗场景中，患者的主诉通常包含错综复杂的症状描述、既往病史与潜在诱因。如果直接将这类长文本作为查询输入到传统检索器中，极易导致语义重心的偏移，从而召回大量无关噪声。为了实现非结构化医学文本的高效、精准召回，本模块作为与知识图谱检索并行的独立分支，专门负责处理自然语言查询到非结构化证据的映射，其核心流程涵盖推理感知分解、因果视角实例化以及多视角混合分数校准。

面对复杂的初始查询 q，系统首先引入思维链（Chain-of-Thought, CoT[76]）策略，将其解构为一系列原子的、具有明确医学导向的中间推理步骤。我们将分解后的子意图集合定义为：

$$S(q) = \{s_1, s_2, \ldots, s_k\} \tag{3.1}$$

其中，每个 s_j 代表整体临床推理链条中的一个局部探寻节点（例如："患者的左肩放射痛是否由心肌缺血引起?"）。这种细粒度的分解有效剥离了复杂主诉中的冗余信息，为后续的靶向检索提供了清晰的语义锚点。

为了弥合自然语言与隐式病理机制之间的表示鸿沟，系统进一步将分解后的每一个子意图 s_j 映射至前文定义的四种基础因果模式 $\mathcal{P} = \{P_1, P_2, P_3, P_4\}$ 中。通过大语言模型作为模式对齐生成器 γ，针对每个 s_j 生成具备明确因果拓扑结构的伪问题。对于原始查询 q，最终实例化的多视角伪问题集合 $\mathcal{Q}(q)$ 可表示为：

$$\mathcal{Q}(q) = \bigcup_{j=1}^{k} \gamma(s_j, \mathcal{P}) \tag{3.2}$$

该集合为同一个临床意图提供了多个不同因果拓扑的 "探测视角"，从而引导检索模型捕捉符合特定病理推演路径的文本证据。

在获取多视角伪问题集合 $\mathcal{Q}(q)$ 后，系统进入基于因果模式匹配的非结构化文本多视角检索阶段。为了突破表层词汇重叠的局限，本模块采用 "模式对模式" 的严格对齐机制。对于每一个带有特定因果拓扑标记的伪问题 $\tilde{q} \in \mathcal{Q}(q)$，系统在离线构建的伪问题池 $\mathcal{Q}$ 中进行独立检索，并获取一个排序列表 $\pi(\tilde{q})$。我们将所有视角的检索列表集合定义为：

$$\mathcal{L}(q) = \{\pi(\tilde{q}) \mid \tilde{q} \in \mathcal{Q}(q)\} \tag{3.3}$$

对于候选文档块 c，我们令 $\mathcal{I}(c)$ 表示在集合 $\mathcal{L}(q)$ 中包含该文档块的检索列表的索引集。同时，令 $r_i(c)$ 和 $s_i(c)$ 分别表示文档块 c 在检索列表 $i \in \mathcal{I}(c)$ 中的排名和相似度得分。

由于不同视角的查询可能召回重叠的文档块，本文提出了一种结合排序共识与分数校准的融合机制，将多视角证据聚合为一个高置信度的证据池。首先，系统计算该文档块的秩共识项：

$$\text{AggRank}(c) = \sum_{i \in \mathcal{I}(c)} \frac{1}{K_{rrf} + r_i(c)} \tag{3.4}$$

其次，为了避免完全依赖相对排序而丢失底层向量空间的绝对语义相关性度量，系统通过平均相似度得分计算出归一化的相关性估计：

$$\text{AggSim}(c) = \frac{1}{|\mathcal{I}(c)|} \sum_{i \in \mathcal{I}(c)} s_i(c) \tag{3.5}$$

最终，文档块 c 的综合融合得分定义为：

$$\text{Fuse}(c) = \text{AggRank}(c) \cdot (1 + \alpha \cdot \text{AggSim}(c)) \tag{3.6}$$

其中，K_{rrf} 为平滑超参数，α 为平衡系数（除非另有说明，本文实验中统一设置 $K_{rrf} = 60$ 且 $\alpha = 0.1$）。文档库将依据 $\text{Fuse}(c)$ 进行降序排列，筛选出最终的证据池。这种混合机制有效过滤了仅在单一视角下偶然匹配的随机噪声，保留了能够支撑多维因果推理的高质量医学证据。

3.2.4 潜在逻辑链挖掘模块

作为与非结构化文本检索并行的结构化召回路径，本模块依托外部构建的医学知识图谱（如融合去噪后的 UMLS 与 SNOMED-CT），旨在为复杂的临床主诉挖掘出底层的标准病理逻辑链条。

针对复杂的患者主诉 q，系统首先利用医学实体识别工具提取出查询中的核心概念集合 $E_q = \{e_1, e_2, \ldots, e_m\}$。本方法采用了一种基于最大联合相似度的映射策略。系统不仅计算单个查询实体 e_i 与候选图谱节点 n_j 的局部语义相似度，还将其与查询全局特征向量 $\mathbf{v}_q$ 的上下文一致性进行耦合。对于每一个抽取出的临床实体 e_i，系统

通过求解以下优化目标，直接锚定其在图谱中的最优匹配节点 $\hat{n}_i$：

$$\hat{n}_i = \arg\max_{n_j \in \mathcal{G}} \left[\text{sim}(\mathbf{v}_{e_i}, \mathbf{v}_{n_j}) \cdot \text{sim}(\mathbf{v}_q, \mathbf{v}_{n_j}) \right] \tag{3.7}$$

这种乘法联合寻优机制起到了严格的逻辑与作用：它强制要求目标节点必须在局部字面和全局语境上同时达到最优契合。通过对所有抽取的查询实体执行该寻优操作，系统自适应地构建出初始种子节点集合 $V_{seed} = \bigcup_{i=1}^{m} \{\hat{n}_i\}$。

为了构建局部知识子图，系统以种子节点集 V_{seed} 为起点，在全局医疗知识图谱 $\mathcal{G}$ 中执行双向启发式扩展。为避免无约束扩展引发概念爆炸与语义噪声，本文引入了关系约束：预先定义并仅保留具有实质鉴别意义的逻辑关系集 $\mathcal{R}_{diag}$，对扩展路径进行动态剪枝。基于此，经过 k 跳扩展的局部子图 $\mathcal{G}_{sub}$ 形式化定义为：

$$\mathcal{G}_{sub} = \left\{ (v_i, r, v_j) \in \mathcal{G} \;\middle|\; v_i, v_j \in \mathcal{N}_k(V_{seed}) \wedge r \in \mathcal{R}_{diag} \right\} \tag{3.8}$$

其中，$\mathcal{N}_k(V_{seed})$ 表示种子节点集在拓扑约束下 k 跳范围内的可达节点集合（本文实验中默认设定最大搜索深度 $k = 2$）；$r \in \mathcal{R}_{diag}$ 则表示节点间的边必须属于预设的诊断关系类型。通过这种方式，系统能够有效排除如图谱中“层级归属”（Is-A）或“同义词”等非推理性边的干扰，从而确保提取出的 $\mathcal{G}_{sub}$ 既保留了核心的病理拓扑结构，又在医疗语义上具备极高的诊断关联性。

在获取局部子图 $\mathcal{G}_{sub}$ 后，系统通过路径枚举算法提取出连接不同种子节点的所有候选简单路径集合 $\mathcal{P}(q)$。为了确保这些拓扑路径具有真实的临床指导意义，系统引入大语言模型作为临床逻辑筛选器。具体而言，模型将基于医学先验知识对每条候选路径 $p \in \mathcal{P}(q)$ 在当前查询 q 语境下的合理性进行评估，最终输出经过逻辑验证的候选证据链集合：

$$\mathcal{L}_{cand} = \left\{ p \in \mathcal{P}(q) \;\middle|\; \mathbb{I}_{LLM}(p, q) = 1 \right\} \tag{3.9}$$

其中，$\mathbb{I}_{LLM}(p, q) \in \{0, 1\}$ 为大语言模型驱动的逻辑评估指示函数。当大语言模型研判路径 p 具备明确的病理因果必然性时，该函数取值为 1；否则取值为 0。

3.2.5 因果路径约束的自适应迭代检索模块

现有的混合检索范式通常直接拼接异构数据，忽略了证据间的逻辑一致性，易因证据链断裂而导致推理漂移。为此，本文提出因果路径约束的自适应迭代检索机制，将图谱挖掘的标准病理链作为逻辑骨架，扫描文本证据中的逻辑断层并动态触发定向补全，实现从单次召回向动态循证的跨越。

将迭代检索定义为多轮状态转移系统，设第 t 轮全局文本证据池为 $\mathcal{C}^{(t)}$（初始状态为 $\mathcal{C}^{(0)}$）。对于任一高置信度逻辑链 $p \in \mathcal{L}_{cand}$，其边集记为 $E(p) = \{e_1, \ldots, e_m\}$，其中 $e_i = (v_{head,i}, r_i, v_{tail,i})$。为精准衡量文本对逻辑链路的支撑能力，系统设计了融合局部语义蕴含与全局逻辑推理的双重评估机制。针对边 e_i，其综合证据支持度 $s_{support}$ 由两部分决定：

第一部分为基于文本嵌入模型计算的局部语义支持得分 s_{emb}，用于快速评估 $\mathcal{C}^{(t)}$ 是否包含支撑该推演环节的显式语义信息。系统分别对图谱边 e_i 与当前上下文 $\mathcal{C}^{(t)}$ 进行稠密向量编码，并通过计算两者的余弦相似度来量化文本证据的局部相关性：

$$s_{emb}(e_i, \mathcal{C}^{(t)}) = \text{CosineSim}\big(\text{Embed}(e_i), \text{Embed}(\mathcal{C}^{(t)})\big) \tag{3.10}$$

第二部分为基于大语言模型的深层逻辑校验得分 s_{LLM}，通过构建评估提示词触发大模型进行零样本因果判定，以挖掘文档中的隐式医学关系：

$$s_{LLM}(e_i, \mathcal{C}^{(t)}) = \text{LLM\_Evaluator}\big(\text{Prompt}(e_i, \mathcal{C}^{(t)})\big) \tag{3.11}$$

由于嵌入模型输出的相似度得分 s_{emb} 与大语言模型输出的逻辑校验得分 s_{LLM} 在数值分布上存在差异，系统在进行加权融合前，引入归一化映射函数 $\phi(\cdot)$ 将两者统一对齐至 $[0, 1]$ 概率区间。最终的跨空间证据支持度函数定义为两者的线性加权：

$$s_{support}(e_i, \mathcal{C}^{(t)}) = \alpha \cdot \phi_{\text{emb}}\big(s_{emb}(e_i, \mathcal{C}^{(t)})\big) + (1 - \alpha) \cdot \phi_{\text{LLM}}\big(s_{LLM}(e_i, \mathcal{C}^{(t)})\big) \tag{3.12}$$

其中，$\alpha \in [0, 1]$ 为平衡两者贡献的权重超参数（本实验设定 $\alpha = 0.5$）。归一化函数 $\phi(\cdot)$ 根据底层打分特性采用 Min-Max 缩放或平滑映射，确保了异构打分在逻辑阈值判定时的数学可比性。系统设定逻辑自洽阈值 $\tau_{logic} = 0.625$ 以平衡召回质量与计算成本。当 $s_{support} < \tau_{logic}$ 时，判定当前文本未能有效支撑该推演环节，据此提取第 t

轮迭代的逻辑断层集合：

$$\Delta^{(t)}(p) = \{e_i \in E(p) \mid s_{support}(e_i, \mathcal{C}^{(t)}) < \tau_{logic}\} \tag{3.13}$$

若 $\Delta^{(t)}(p) \neq \emptyset$，针对每一个缺失边 $e_{miss} \in \Delta^{(t)}(p)$，系统将其与原始主诉 q 融合，利用大模型构建具备上下文感知能力的补丁查询 q_{patch}。随后，复用前文的非结构化检索链路对 q_{patch} 进行定向召回，获取补偿文档片段集合 $\mathcal{D}_{new}$，状态转移方程更新为：

$$\mathcal{C}^{(t+1)} = \mathcal{C}^{(t)} \cup \mathcal{D}_{new} \tag{3.14}$$

迭代过程持续进行，直至 $\Delta^{(t)}(p) = \emptyset$（实现因果对齐）或达到最大迭代深度 T_{max}。最终输出的融合上下文 $\mathcal{C}^{(final)}$ 既具备图谱的严密因果骨架，又充实了文献的细节证据。

3.3 实验

3.3.1 实验设置

所有实验均在配备 8 块 NVIDIA RTX3090 GPU（每块显存 24 GB）、CUDA 版本为 12.4 的服务器上进行。对于闭源商业模型及较大参数的模型（如 GPT-5.1, DeepSeek-v3.2 等），本文通过 API 调用完成推理与评估；对于开源模型（如 Qwen3-8B, Qwen3-32B-30A），本文在本地服务器环境中部署运行。我们将本文方法与两类极具竞争力的基线检索与生成模型进行了对比：

- **非结构化检索基线（使用 MIRAGE[77]）**：包括稀疏检索器 BM25、通用领域密集检索器 Contriever 与 SPECTER、医疗垂直领域专用检索器 MedCPT，以及使用倒数秩融合策略的强混合检索基线 RRF-4。
- **结构化图谱基线（使用 GraphRAG-Bench[78]）**：包括目前前沿的结构化 RAG 系统，如基于树结构检索的 RAPTOR[79]、基于社区发现的 GraphRAG（包含 Global 与 Local 两种模式）、双层检索架构 LightRAG[80]，以及基于关联链接的 HippoRAG2[81]。

在整个实验过程中，我们使用统一的大语言模型来执行生成、路径提取与迭代验证任务，以保证实验对比的公平性。

- **非结构化检索配置**：采用 bge-m3[①] 模型获取文本和查询的密集嵌入特征。在离线预处理阶段，针对每个文档块生成 $N = 4$ 个因果伪问题。在线问答时，多视角的检索列表通过倒数秩融合（RRF）进行聚合（平滑参数 $K_{rrf} = 60$），最终截取综合得分排名前 $K = 5$ 的文本片段进入全局上下文。
- **自适应迭代配置**：系统采用轻量级自然语言推理（NLI）模型计算跨空间的证据支持度，并设定逻辑自洽阈值 $\tau_{logic} = 0.625$。当支持度低于该阈值时，触发补丁查询的自动生成与迭代检索。为控制推理时效性并防止无限循环，最大迭代深度设定为 $T_{max} = 2$。

在实验过程中，我们使用了如下的测试指标：

- **MIRAGE 测试集**：采用标准客观的准确率（Accuracy）进行衡量。
- **GraphRAG 测试集**：采用多维评价体系。其中，答案准确率（Answer Accuracy, Acc）定义为大模型事实判别得分（Fact Check, FC）与向量语义相似度（Semantic Similarity, SS）的加权融合：$Acc = \alpha \cdot FC + (1 - \alpha) \cdot SS$，本实验中设定平衡权重 $\alpha = 0.5$。此外，使用覆盖率（Coverage, Cov）评估生成答案对核心金标准实体的召回情况，并通过 ROUGE-L 评估生成文本的词汇重叠度与句法连贯性。

3.3.2 实验数据集说明

为了全面评估本文提出的框架在复杂医疗推理与检索中的有效性，我们在两个权威的医疗与图谱问答基准库上进行了评测：

- **MIRAGE 基准测试** [②]：这是一个大规模医疗 RAG 评测套件，包含 7,663 个问题，覆盖 MedQA、MedMCQA、PubMedQA、BioASQ 和 MMLU-Med 五个多样化的医学数据集。我们在其标准化的三个外部语料库上进行测试：医学教科书（Textbooks）、随机采样的 50 万篇 PubMed 摘要（PubMed 500K），以及 StatPearls 临床知识库。
- **GraphRAG-Bench**[③]：为评估系统的结构化图谱推理能力，我们采用了该综合测试台的医疗子集，主要从三个维度进行能力评估：事实检索（Fact Retrieval, FR）、复杂推理（Complex Reasoning, CR）和上下文摘要（Contextual Summarization, CS）。

① https://www.modelscope.cn/models/BAAI/bge-m3
② https://github.com/gzxiong/MIRAGE/tree/main
③ https://github.com/GraphRAG-Bench/GraphRAG-Benchmark

3.3.3 在非结构化文献检索基准上的主实验结果

表 3.2 所提出的方法在 MIRAGE-benchmark 上与主流医疗检索器的性能表现对比, 模型使用 Qwen3-8B。其中加粗数据为表现最好的数据，下划线数据为表现次好数据。

语料库	检索方法	MMLU	MedQA	MedMCQA	PubMedQA	BioASQ	Avg.
Textbooks	BM25	82.26	76.59	61.06	31.20	52.40	60.70
	Contriever	86.10	76.83	63.10	27.00	53.40	61.29
	MedCPT	86.08	75.45	62.20	34.80	59.06	63.52
	SPECTER	84.50	75.49	62.72	29.20	52.43	60.87
	RRF-4	84.98	77.12	62.18	33.20	55.02	62.50
	Our Method	**86.32**	**78.00**	**63.30**	**35.20**	**72.17**	**67.00**
StatPearls	BM25	82.87	73.52	**66.24**	30.20	58.74	62.31
	Contriever	85.35	**77.45**	62.53	29.60	58.90	62.77
	MedCPT	84.21	75.88	63.95	31.60	63.40	63.81
	SPECTER	**85.99**	75.73	61.75	26.80	54.53	60.96
	RRF-4	85.94	76.71	62.50	33.60	63.59	64.47
	Our Method	85.95	76.20	65.29	**33.80**	**74.11**	**67.07**
PubMed(500K)	BM25	85.28	76.73	59.58	31.80	64.40	63.56
	Contriever	83.79	75.41	60.11	27.80	60.19	61.46
	MedCPT	84.00	76.59	64.68	32.00	65.86	64.63
	SPECTER	85.52	75.40	61.77	31.40	55.02	61.82
	RRF-4	85.33	76.44	61.25	31.20	66.18	64.09
	Our Method	**85.86**	**77.22**	**65.29**	**33.20**	**68.45**	**68.35**

在对 MIRAGE 基准测试的全面评测中，本文提出的方法在 Textbooks、StatPearls 以及 PubMed(500K) 三种语料库背景下均展现出了显著的性能优势。如表 3.2 所示，本文方法在 PubMed(500K) 语料库背景下取得了 68.35% 的最高加权平均得分（Avg.），相较于该语料库下表现最强的基准模型 MedCPT 实现了 3.72 个百分点的绝对提升。在 Textbooks 语料库下，本文方法同样稳健地超越了 MedCPT 等主流医疗检索器。这种跨语料库的领先地位充分验证了本文框架在处理异构医疗知识源时的卓越适配能力，能够有效地为大语言模型提供高质量的上下文支撑。

进一步分析各个细分数据集的表现可以发现，本文方法在处理逻辑复杂的医疗任务时具有极强的鲁棒性。特别是在 BioASQ 这一极具挑战性的医学是非题任务中，本文方法在 StatPearls 场景下取得了 74.11% 的突破性成绩，相较于基准模型普遍存在的 50%–60% 性能瓶颈有显著提升。这表明通过引入多路径检索与图增强机制，本文方法能够更精准地捕捉医学概念间的深层语义对齐，从而弥补了传统语义检索在

表 3.3　所提出的方法在 MIRAGE-benchmark 使用不同大语言模型的性能表现比较

语料库	大模型	MMLU	MedQA	MedMCQA	PubMedQA	BioASQ	Avg.
Textbooks	Qwen3-8B	87.05	78.31	64.31	36.40	72.33	67.68
	Qwen3-32B-30A	89.07	82.40	70.48	51.00	80.42	74.67
	DeepSeek-v3.2	92.75	91.20	74.51	**56.00**	82.52	79.40
	GPT-5.1	**93.57**	**94.18**	**78.70**	54.80	**87.21**	**81.69**
PubMed(500K)	Qwen3-8B	86.40	79.50	66.98	35.20	76.21	68.86
	Qwen3-32B-30A	88.71	84.13	71.12	50.00	81.07	75.01
	DeepSeek-v3.2	93.57	90.10	76.19	**57.20**	83.29	80.07
	GPT-5.1	**94.03**	**93.87**	**83.60**	55.40	**88.52**	**83.08**
StatPearls	Qwen3-8B	86.04	76.36	65.48	34.60	74.27	67.35
	Qwen3-32B-30A	88.52	82.25	72.13	50.40	79.77	74.61
	DeepSeek-v3.2	92.38	91.44	75.20	**56.80**	82.04	79.57
	GPT-5.1	**93.94**	**94.03**	**78.99**	55.60	**87.22**	**81.96**

处理专业领域长链条逻辑时的不足。而在题量巨大的 MedMCQA 数据集上，本文方法展现出了优异的稳定性。不同于基准检索器（如 BM25）在不同语料库下表现出的剧烈性能波动，本文方法在三种知识源支持下均能维持在较高水平。这种稳定性证明了本文框架能够通过多路径特征融合，有效缓解检索过程对单一语料库分布的过度依赖，从而在面对大规模、多维度的医学考试题目时，依然能提供高质量的证据支撑。

表 3.3 进一步展示了本文提出的 RAG 框架在适配不同参数规模及架构的基座模型时的性能表现。实验结果显示，随着模型规模的提升和推理能力的增强，系统在所有评估维度上均表现出了显著的性能增益。在三种语料库场景下，GPT-5.1 作为基座模型时均取得了最优的全局加权平均分，最高达到 83.08%。这表明本文提出的多路径检索与图增强机制具有良好的模型通用性，能够充分挖掘顶尖语言模型在医学知识推理方面的潜力。

值得注意的是，尽管 GPT-5.1 在大多数指标上保持领先，但 DeepSeek-v3.2 在 PubMedQA 数据集上展现出了卓越的专业适配性。在 Textbooks、PubMed 和 StatPearls 三种检索背景下，DeepSeek-v3.2 在 PubMedQA 项上的得分（分别为 56.00%、57.20% 和 56.80%）均超越了 GPT-5.1。这一现象暗示 DeepSeek 系列模型在处理医疗专业摘要及证据提取任务时，可能具有更优的语义对齐能力或更针对性的知识储备。此外，从 Qwen3-8B 到 Qwen3-32B-30A 的性能跨越（平均提升约 7%）进一步印证了模型规模对复杂医学问答任务的关键影响。

综合来看，不同模型的表现趋势在三种语料库之间保持了高度的一致性，这再次证明了本文 RAG 框架的稳定性。无论使用何种规模的基座模型，系统均能通过精确的上下文注入有效提升回答的准确度。这种良好的模型兼容性为实际应用场景中的算力权衡提供了参考，即在资源有限的情况下，使用中等规模模型（如 Qwen3-32B）配合本文的增强策略，依然能够获得超越基础 RAG 配置的优异性能。

3.3.4 在结构化图谱推理基准上的对比实验结果

此外，我们在 GraphRAG-Bench 上进行了多维度对比实验。为确保公平性，所有基准模型均统一使用 Qwen2.5-14B 作为基座模型，并采用 GPT-4o-mini 作为客观评估器。如图 3.3 所示，我们提出的方法在各项核心指标上均展现出了稳健的性能。

在复杂推理准确率（CR_ACC）方面，我们的方法达到了 68.4%，相较于表现最强的基准模型 HippoRAG2（64.0%）提升了 4.4%。这一显著提升凸显了因果伪问题检索与因果链挖掘之间的协同效应，使模型能够基于深层证据链进行推理，而非仅仅停留在表层的语义匹配。

在事实检索准确率（FR_ACC）方面，我们的方法取得了 68.5% 的最高分，不仅略高于 HippoRAG2（64.5%），更是大幅领先于基于结构化的基准方法，如 GraphRAG (global)（40.0%）和 RAPTOR（50.5%）。这有力地验证了本文方法相较于传统基于聚类范式的优越性，进一步表明因果驱动的检索能够跨越聚类边界，捕捉到静态社区检测往往会遗漏的潜在因果证据。

在回答质量方面，本方法在复杂推理 ROUGE（CR_ROUGE）指标上达到了 39.3%，显著优于 HippoRAG2（33.0%）。这一结果表明，通过构建因果链，系统能够生成更

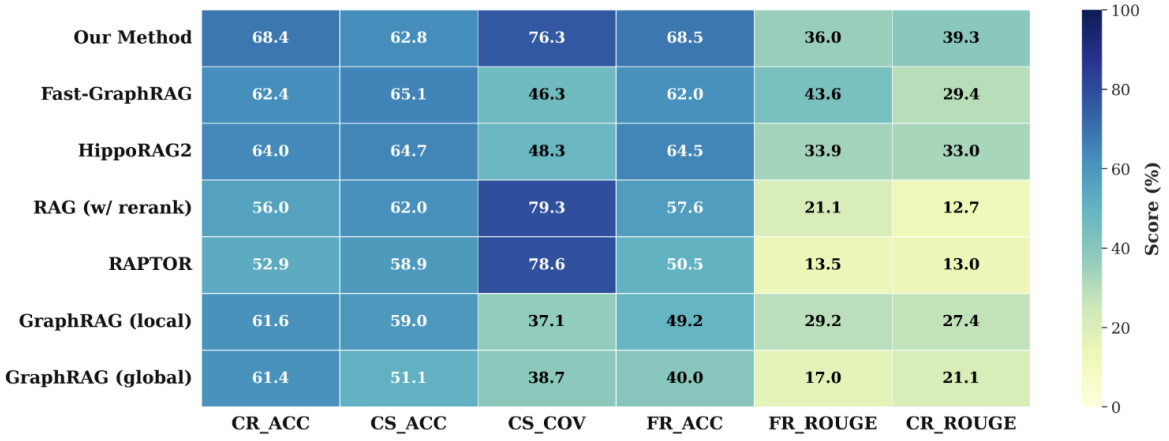

图 3.3 所提出的方法在 GraphRAG-Bench 上的表现

具连贯性、逻辑性且具备坚实语义支撑的解释。

3.3.5 消融实验与分析

为深入验证本文提出的因果图谱引导的自适应迭代检索机制的有效性，并探究迭代深度对最终生成质量的影响，我们针对检索步数配置开展了详细的消融实验。如表 3.4 所示，我们在高质量结构化语料（Textbooks）与包含海量噪声的非结构化语料（PubMed 500K）上，对比了四种设置：（i）**BM25 + Dense**：作为基线模型，仅使用标准的双路混合检索（BM25 + 向量检索），不包含因果图谱对齐与迭代机制；（ii）**Our Method**：引入本文提出的因果伪问题多视角检索，但不开启自适应迭代补全（即 0 step）；（iii）**Our Method (1 step)**：在基础架构上引入 1 轮逻辑断层检测与重检索；（iv）**Our Method (2 step)**：允许系统进行最高 2 轮的迭代检索。通过跨语料库与跨步数的对比，我们能够从经验上精确评估各核心组件及迭代策略对最终问答性能的贡献。

实验结果表明，在所有语料库下，仅使用 BM25+Dense 基线的配置性能均处于最低水平。当引入本文的基础因果对齐模块（Our Method）后，Textbooks 和 PubMed(500K) 上的平均准确率分别提升至 65.08% 和 64.25%。这一阶段的增益主要归功于“模式对模式”的匹配机制，它有效剔除了虽具语义相似度但缺乏因果逻辑支撑的无关段落，提升了初始召回证据的纯度。

当我们进一步引入迭代优化模块（1 step 与 2 step）后，系统性能展现出了更为丰富的动态变化，且在不同分布的语料库上呈现出差异化的表现：首先，在 Textbooks 语料库中，随着迭代步数的增加，系统平均准确率呈现稳步上升趋势（从 65.08% 提升至 1 step 的 66.10%，并最终在 2 step 达到峰值 66.65%）。由于医学教科书内容高

表 3.4　不同语料库背景下检索步数配置的性能对比消融实验

语料库	方法	MMLU	MedQA	PubMedQA	Avg.
Textbooks	BM25+Dense	84.96	76.90	31.40	64.42
	Our Method	85.86	76.98	32.40	65.08
	Our Method (1 step)	86.04	77.45	34.80	66.10
	Our Method (2 step)	**86.59**	**78.16**	**35.20**	**66.65**
PubMed(500K)	BM25+Dense	84.06	76.04	31.00	63.70
	Our Method	84.75	76.20	31.80	64.25
	Our Method (1 step)	85.40	**78.31**	32.80	**65.50**
	Our Method (2 step)	**85.77**	77.14	**33.00**	65.30

度规范、逻辑严密且噪声较少，多轮迭代能够安全、稳定地顺着知识图谱的骨架找回遗漏的中间介导证据，从而持续完善多跳逻辑链条，大幅提升了模型在复杂推理任务上的上限。

其次，在 PubMed(500K) 语料库中，性能演变呈现出明显的权衡特征。引入 1 轮迭代后，系统平均准确率达到了最高点 65.50%（特别是 MedQA 子任务达到了 78.31%），证明了重检索在真实文献检索中的巨大潜力。然而，当强制进行第 2 轮迭代时，平均性能并未继续增长，反而微降至 65.30%。我们将此归因于海量摘要文献中固有的高噪声与结论冲突：在极度嘈杂的环境中，过深的迭代容易打破原有的逻辑自洽，导致大模型在生成时吸入互斥的病理观点（即发生推理漂移），从而抵消了证据链补全带来的收益。

总体而言，表 3.4 的结果一致表明，我们的框架通过引入结构化知识图谱作为逻辑骨架，有效应对了长链条医疗问答中的“证据链断裂”问题。消融实验清晰展示了各模块在功能上的明确分工（基础模块提升证据纯度，迭代模块提升逻辑完整度）。

为了验证多模式伪问题生成模块在医疗知识检索增强（RAG）中的有效性，本文在 Qwen3-8B 模型上针对不同的因果逻辑模式进行了消融实验。如表 3.5 所示，全模式融合策略（All）在所有医疗基准测试中均取得了最优表现，尤其在涵盖面广的 MMLU-Med (85.86%) 和 MedMCQA (62.97%) 数据集上优势显著，这充分证明真实的医疗查询往往内含错综复杂的上下文，单一的因果逻辑模式无法满足开放域的检索需求。深入对比单一模式在不同类型任务中的表现可以发现，在侧重于复杂临床推理的任务（如 MedQA-US 和 BioASQ-Y/N）中，因果链模式（P4）表现出极强的竞争力，准确率分别达到 74.16% 和 54.69%，这是因为 P4 能够有效模拟“病因—病理变化—临床表现”的多跳演变路径，从而在长文本病例中精准锚定深层关联证据。与此

表 3.5 使用不同模式伪问题进行医疗知识检索的表现对比

模型	伪问题模式	MMLU-Med	MedQA-US	MedMCQA	PubMedQA	BioASQ-Y/N
Qwen3-8B	P1	82.19	72.66	59.53	28.40	52.27
	P2	84.48	73.06	61.22	28.60	52.10
	P3	84.66	72.98	61.34	26.60	50.81
	P4	83.20	74.16	57.09	29.20	54.69
	All	85.86	76.98	62.97	32.40	58.58

同时，在偏向广度基础医学知识问答的 MMLU-Med 中，单因多果（P2，84.48%）与多因单果（P3，84.66%）的表现均优于基础模式（P1，82.19%），表明发散性的伪问题有助于召回多维度的医学事实；然而，在包含大量直接事实性问答的 MedMCQA 数据集上，P4 模式的表现（57.09%）甚至出现了非正常下滑，低于 P1（59.53%），说明强制引入多跳因果链反而会导致语义偏移并引入检索噪声。此外，在侧重严谨医学文献检索的 PubMedQA 任务中，过度发散的生成策略同样暴露了局限性，P3 的准确率仅为全场最低的 26.60%，这是由于医学文献通常聚焦于单一变量或特定论证，P3 生成的多种潜在前置原因严重稀释了核心检索词的权重。综合来看，基础的线性检索策略（P1）信息覆盖率受限，而高阶的单向模式（P2, P3, P4）虽在特定分布下具有优势，但也伴随引入局部噪声的风险，本文提出的全模式融合方法（All）能够动态适配查询需求并有效取长补短，通过全面映射问题中的潜在因果网络结构，显著且鲁棒地提升了医疗垂直领域问答系统的整体检索准确率。

3.4 本章小结

本章针对医疗大语言模型在应对复杂临床推理时传统检索增强生成（RAG）过度依赖表层语义匹配的痛点，提出了一种基于因果路径约束与用户查询伪问题生成的多轮检索增强方法。该方法致力于在 RAG 链路中实现从语义对齐向深层因果逻辑对齐的跨越。

在具体实现上，本章详细设计了四个核心模块：首先，通过深度融合 UMLS 与 SNOMED-CT 构建高质量医学知识库，并结合结构因果模型（SCM）的四种基础拓扑，在离线阶段生成因果感知伪问题，实现文献逻辑化索引；其次，在线上问答阶段，引入复杂查询分解与混合索引机制，通过思维链解构与伪问题检索，精准定向非结构化医学证据；同时，通过潜在逻辑链挖掘模块在图谱中提取标准病理路径；最后，创新性地提出因果图谱引导的自适应迭代检索机制，利用大语言模型与自然语言推理技术跨空间检测逻辑断层，动态触发重查询，将单次召回升级为闭环的循证补全过程。

在实验验证方面，本章在 MIRAGE 和 GraphRAG-Bench 两个权威医疗基准库上进行了全面评估。实验结果表明，所提方法在多项核心指标上均显著优于现有的 BM25、MedCPT、RRF-4 等非结构化检索器以及 HippoRAG2 等结构化图谱基线。此

外，跨基座模型（如 Qwen3 系列、DeepSeek、GPT-5.1）的适配实验以及检索步数等消融实验，进一步证实了本框架在不同算力资源下的鲁棒性，以及自适应迭代机制在修复证据链断裂问题上的有效性。综上所述，本章提出的方法有效提升了大模型在复杂医疗辅助决策场景中的准确性、透明度与逻辑可靠性。

第四章　基于证据特异性假设重排与多智能体辩论的长尾检索增强方法

4.1　医疗长尾检索中的推理漂移问题与研究动机

在前文构建的图谱混合检索架构基础上，RAG 系统已初步具备从海量医学语料中召回相关证据的能力。然而，复杂医疗问答与推理在一些更加特殊的方向上仍面临着极为严峻的长尾分布与特征高度耦合挑战。例如有关罕见病的细微线索往往被海量常见病文献所掩盖，而中医的复杂证候也极易与表象相似的伪证相混淆。当系统依赖传统的相似度匹配进行检索时，极易捕获大量字面相关但缺乏特异性的通用医学描述。这些冗余的共性噪声会严重稀释关键的推理线索，导致大模型在推演阶段发生检索诱导的推理漂移。一旦模型顺应这些高频且似是而非的常见病证据进行生成，便会偏离正确的解析路径，最终得出事实性错误或不恰当的医学结论。

针对高噪声检索环境下的推理漂移问题，目前常见的优化方案多集中于后置的文档重排或基于单智能体的自我一致性采样。这些手段能够有效剔除与问题完全无关的噪声文本，在通用问答场景中的确取得了效果，但当将其直接应用于复杂的专业医疗问答时，则具有一定的逻辑局限性。大语言模型的内在机制倾向于在检索池中寻找与用户输入相匹配的证据。但在真实的医疗推理场景中，尤其是面对罕见病鉴别或中医疑难杂症分析时，不同的病理机制往往共享着大量的基础表征。如果系统仅仅执行浅层的相关性证实，则极易被那些文献数量庞大、描述详尽的常见证据所吸引，将不同疾病的通用共享特征误认为是解决问题的决定性证据。现有的单向生成模型无法主动提出具有排他性的反事实假设，因而极易被相似的病理或交叉证候所迷惑，难以从重重干扰中提取出真正具有决定性意义的因果证据；此外，单智能体的线性推理模式根本无法胜任高复杂度的交叉辨析任务。大语言模型的文本生成具有自回归特性，这意味着一旦模型在阅读检索上下文的初期，受到了某篇高相关度（但带有常见偏见）文档的误导，便极易产生认知偏差。在随后的推理中，单智能体不仅难以自我纠错，甚至会为了迎合早期的错误假设而强行扭曲后续证据的解读。这不仅无限放大了长尾知识缺失所引发的幻觉风险，也使得最终输出的医学问答结果完全是一个缺乏严密论证的“黑盒”，严重削弱了其在实际应用中的可靠性与

可解释性。

针对单纯语义检索和单向生成在复杂医疗问答场景下容易出现推理漂移与误判的问题，本文第四章提出了一种基于证据特异性假设重排与辩论质证的医疗推理机制。其核心目的在于解决传统 RAG 极易受共性文本干扰而产生常见偏见的缺陷，使系统具有主动证伪与排他性辨析的能力。本章首先设计了竞争性候选假设与证据集构建策略，针对复杂的初始医疗查询，系统主动提取出一组具有强竞争关系的疑似候选假设作为初始推理空间，并进行多源并发检索以构建对比证据库。为了避免全量候选假设导致的焦点发散问题，本文进一步提出了证据特异性解耦与候选打分算法。该算法通过计算检索文本与各候选假设的语义映射关联，将混杂的外部证据严格解耦为特异性证据与共性干扰证据，并据此对候选空间进行提权与加权打分。在此核心基础上，系统引入了基于特征密度与语义重叠度的动态自适应路由机制。通过计算推演理由的相似度与首选假设的证据特异性密度，构建了动态判定阈值：当首选假设具备高纯度特异性证据且优势显著时，系统触发早期退出机制直接输出结论以提升效率；当头部候选高度相似或陷入证据混淆区时，系统则拉高门槛、截断长尾噪声，仅保留最易混淆的锚点假设与其竞争假设。针对被锁定的两个核心候选答案，最终实现定向质证与共识裁决框架。系统针对特异性证据与共性干扰证据，强制正反方专家智能体围绕常见证据及特异性证据间差异展开质证，从而缓解检索诱导的推理漂移现象。

4.2 基于证据特异性与多智能体辩论的推理模型整体框架

本章将系统介绍本文提出的**基于证据特异性假设重排与多智能体辩论的长尾检索增强方法**。该方法以“突破高噪声检索环境下的相关性陷阱”为核心目标，通过前置的竞争性候选空间构建、证据特异性打分与动态自适应路由，以及在生成阶段引入具备排他性辨析能力的多智能体质证机制，从而有效抑制大语言模型在复杂医疗问答中易受“常见偏见”误导而产生的检索诱导推理漂移现象。整体系统架构如图 4.1 所示。本文提出的自适应医疗推理框架主要包含以下四个核心模块：

1. **竞争性候选假设与证据集构建模块：**针对长尾医疗场景下极易发生的“常见偏见”所导致的混淆问题，本模块改变了传统系统仅针对初始医学查询进行单一检索的被动局面。系统主动围绕复杂的医疗表征，提取出一组具有强竞争关系

的疑似候选假设作为初始推理空间，为后续的排他性辨析奠定基础。

2. **证据特异性解耦与候选打分模块：**根据竞争性假设，系统进行再检索以构建多源对比证据空间。本模块通过计算证据片段与各候选假设的语义映射关联，将混杂的外部证据逻辑划分为“特异性证据”与“共性干扰证据”。随后，系统利用加权打分算法对候选空间进行量化评估，精准识别并剥离缺乏真实鉴别力的通用共性特征，从而赋予特异性医疗表征更高的推理权重。
3. **基于特征密度与语义重叠度的动态自适应路由机制：**为防止静态分流阈值易引发相似候选混淆或误判的问题，本模块通过提取置信度最高的前两名假设，计算推演理由的“语义重叠度”与首选假设的“证据特异性密度”，构建包含相似度惩罚与特异性调节项的动态阈值函数。当首选假设具备高特异性证据且与竞争项差异显著时，系统动态降低门槛并触发早期退出机制直接输出结论；当头部候选高度相似或证据支撑薄弱时，系统拉高判定门槛，截断长尾噪声并仅保留最易混淆的候选对进入质证。
4. **差分质证与共识裁决模块：**针对被锁定的极易混淆候选对，系统将前期解耦出的特异性证据与共性证据分别分配给专属的正反方专家智能体。双方在严格的时序协议下，围绕常见病理或基础证候无法解释的特异性表征展开一对一交叉

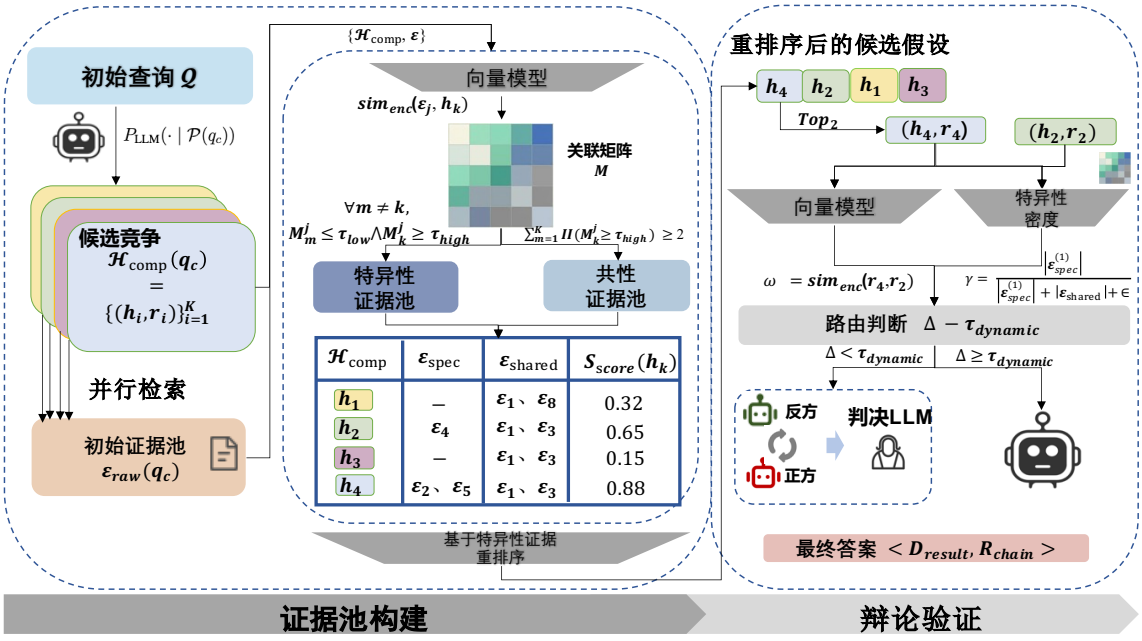

图 4.1　基于证据特异性假设重排与多智能体辩论的检索增强方法总览。系统主要包含竞争性候选假设空间构建、证据特异性解耦与候选打分、基于特征密度与语义重叠度的动态自适应路由，以及差分质证与共识裁决模块四大核心模块。

质证。最终，裁判模块根据正方能否利用特异性证据成功防御反方的共性攻击来进行逻辑结算，从而输出准确且具备深度因果解释力的最终医学解答。

4.2.1 竞争性候选假设与证据集构建模块

竞争性候选假设与证据集构建模块位于医疗推理流程的起始阶段，其核心目标是为复杂长尾医学问答与推理提供多维度的竞争锚点和无偏的证据空间。具体而言，该模块通过构造竞争性假设集 $\mathcal{H}_{\text{comp}}(q_c)$ 将推理的搜索空间从单一查询的线性推演扩展为包含高混淆度干扰项的对比空间，从而打破大语言模型极易陷入的常见偏见；同时，通过构造证据集 $\mathcal{E}(q_c)$ 将单一查询引发的局部检索结果转化为涵盖正反视角的全局上下文，为后续的证据解耦及质证提供支持。

考虑到真实医疗查询（如长尾罕见病鉴别或中医复杂证候分析）的模糊性与多义性，本模块利用大语言模型（LLM）的内在医学先验知识与指令遵循能力进行初步的推理发散。给定初始医疗查询 q_c，系统构造特定的医学提示模板 $\mathcal{P}(q_c)$，引导大语言模型自回归地并行生成包含 K 个可能的医学假设及其对应推演理由的竞争性候选集。该候选答案集可形式化定义为：

$$\mathcal{H}_{\text{comp}}(q_c) = \left\{ (h_i, r_i) \right\}_{i=1}^{K}, \quad (h_i, r_i) \sim P_{\text{LLM}}\big(\cdot \mid \mathcal{P}(q_c)\big), \tag{4.1}$$

其中，h_i 代表大语言模型输出的第 i 个医学假设或疑似证候标签；r_i 代表模型针对该候选生成的特定推演理由，即解释初始查询 q_c 中哪些具体表征支持了假设 h_i 的医学逻辑链。P_{LLM} 为大语言模型的条件生成概率分布，系统通过调整解码策略（如 Temperature 或 Top-p 采样）并截取生成概率最高的 K 个结果，以确保候选空间的多样性。

除了建立候选空间，本模块强制大模型同步生成推演理由 r_i。这不仅对候选假设进行了初步的逻辑自洽性检验，更重要的是，r_i 中包含的逻辑信息能够为后续的多源并发检索提供可靠的查询扩展线索。通过这种基于模型内生知识推演的候选锚点确立的预处理，系统实质上将庞大且无序的外部检索空间，提前收敛到了几个具有清晰医学争议点的竞争锚点上，从而有效限制了后续多智能体交互的计算规模。

为进一步打破传统 RAG 仅针对原问题进行检索的缺陷，本模块基于构建的假设空间进一步构建与查询相关的证据集 $\mathcal{E}(q_c)$。首先，执行跨假设空间的多查询并发检

索，聚合原始查询与各类竞争假设召回的初始文本片段，形式化为：

$$\mathcal{E}_{\text{raw}}(q_c) = \bigcup_{h \in \mathcal{H}_{\text{comp}}} \mathcal{R}(q_c \oplus h, \mathcal{D}), \tag{4.2}$$

其中 $\mathcal{R}(\cdot, \mathcal{D})$ 为检索函数，$\oplus$ 表示查询与假设的上下文拼接。这种结构化召回不仅获取了支持原始查询方向的证据，更强制系统吸纳了支持常见疾病或对立证候的反向线索。

通过这种结构化组织，最终提纯的证据集 $\mathcal{E}(q_c)$ 与竞争假设空间 $\mathcal{H}_{\text{comp}}(q_c)$ 共同构成了本模块的输出 $\{\mathcal{H}_{\text{comp}}, \mathcal{E}\}$。该结果作为具有对比证据空间，为后续 “证据解耦—动态自适应路由—定向质证” 流程提供了完整的输入基础和坚实的医学逻辑溯源依据。

4.2.2 证据特异性解耦与候选打分模块

在通过大语言模型获取竞争性假设空间 $\mathcal{H}_{\text{comp}}$ 并基于此完成多源检索后，本模块的核心任务是消除文本冗余与共性噪声，通过语义对齐与特征解耦量化各假设的真实因果鉴别力。传统生成模型在处理长文本时极易被高频共性表征引发注意力偏移，导致长尾排他性特征被掩盖。为此，本模块首先计算多源对比证据空间 $\mathcal{E}$ 中各文本片段 e_j 与候选假设 $h_k \in \mathcal{H}_{\text{comp}}$ 及其推理理由的语义映射关联度，构建全局关联矩阵 M，其中元素定义为：

$$M_{j,k} = \text{sim}_{\text{enc}}(e_j, h_k), \tag{4.3}$$

其中 $\text{sim}_{\text{enc}}(\cdot, \cdot)$ 为密集检索模型计算的余弦相似度。

基于关联矩阵 M，系统引入双重相似度经验阈值 $(\tau_{\text{high}}, \tau_{\text{low}})$ 将混杂的外部证据作逻辑物理隔离，解耦为特异性证据集合与共性干扰证据集合。对于任意候选假设 h_k，其专属的特异性证据集 $\mathcal{E}^{(k)}_{\text{spec}}$ 与全局共性证据集 $\mathcal{E}_{\text{shared}}$ 形式化定义为：

$$\mathcal{E}^{(k)}_{\text{spec}} = \left\{ e_j \in \mathcal{E} \mid M_{j,k} \geq \tau_{\text{high}} \ \wedge \ \max_{m \neq k} M_{j,m} \leq \tau_{\text{low}} \right\}, \tag{4.4}$$

$$\mathcal{E}_{\text{shared}} = \left\{ e_j \in \mathcal{E} \;\middle|\; \sum_{m=1}^{K} \mathbb{I}(M_{j,m} \geq \tau_{\text{high}}) \geq 2 \right\}, \tag{4.5}$$

其中指示函数 $\mathbb{I}(\cdot)$ 用于统计该证据片段同时强支持的候选假设数量。式 (4.4) 严密地筛出了仅支持特定医学假设而无法用其他竞争项解释的决定性表征（如罕见病的特

异性体征或中医特定的辨证眼目），而式 (4.5) 则捕获了缺乏排他性的通用共性线索。

在完成证据逻辑分层后，系统利用加权打分算法对候选空间 $\mathcal{H}_{\text{comp}}$ 中的每个假设 h_k 进行量化评估。为精准识别并剥离缺乏真实鉴别力的通用共性特征，系统赋予特异性证据较高的推理权重，候选假设 h_k 的综合置信度得分 $S_{\text{score}}(h_k)$ 计算如下：

$$S_{\text{score}}(h_k) = \lambda_{\text{spec}} \sum_{e_j \in \mathcal{E}_{\text{spec}}^{(k)}} M_{j,k} \; + \; \lambda_{\text{shared}} \sum_{e_j \in \mathcal{E}_{\text{shared}}} M_{j,k}, \tag{4.6}$$

约束条件为 $\lambda_{\text{spec}} > \lambda_{\text{shared}} > 0$。其中，$\lambda_{\text{spec}} = 1.0$ 与 $\lambda_{\text{shared}} = 0.3$ 分别代表排他性特征与共性干扰特征的权重惩罚因子。通过极其悬殊的权重分配，该公式实现了对长尾特异性临床表现的信号提纯与鉴别提权，而量化得分客观反映了每个竞争假设在剔除共性偏见后的真实因果支撑强度。

4.2.3 基于特征密度与语义重叠度的动态自适应路由模块

在获取候选假设的量化得分后，若将整个候选空间直接输入多智能体系统进行交互质证，不仅会使大语言模型的推理时间复杂度增加至 $\mathcal{O}(N^2)$，还会因引入过多无关假设导致模型注意力分散。为优化系统算力分配并降低推理误判风险，本模块提出了一种结合 “证据特异性密度” 与 “语义重叠度” 的动态路由机制。

具体而言，首先根据前述计算的综合置信度得分 S_{score} 对候选空间降序排列，选取首选假设 $h_{(1)}$ 与竞争假设 $h_{(2)}$，并计算两者的置信度得分差值 $\Delta = S_{\text{score}}(h_{(1)}) - S_{\text{score}}(h_{(2)})$。传统方法通常采用固定的经验阈值对 Δ 进行判定。但在实际复杂医疗推理中，若两个候选假设的病理或临床表征高度相似，即使得分差距较大，直接输出首选假设仍存在较高隐患；同理，若高分主要依靠缺乏特异性的共性线索累加，结论的可靠性也会降低。因此，本模块通过引入候选假设的内生特征，设计了动态判定阈值 τ_{dynamic}。

首先，系统基于大模型生成的推演理由 $r_{(1)}$ 与 $r_{(2)}$，计算两者的**语义重叠度** ω：

$$\omega = \text{sim}_{\text{sem}}(r_{(1)}, r_{(2)}), \quad \omega \in [0, 1]. \tag{4.7}$$

该指标用于量化两个假设在医学表征上的相似性。ω 越高，表明两者越容易混淆，系统应当相应提高判定门槛。

其次，基于前述证据解耦结果，定义首选假设的**证据特异性密度** γ：

$$\gamma = \frac{|\mathcal{E}^{(1)}_{\text{spec}}|}{|\mathcal{E}^{(1)}_{\text{spec}}| + |\mathcal{E}_{\text{shared}}| + \epsilon}, \quad \gamma \in [0, 1]. \tag{4.8}$$

其中 $|\cdot|$ 表示集合的基数，$|\mathcal{E}^{(1)}_{\text{spec}}|$ 与 $|\mathcal{E}_{\text{shared}}|$ 分别为首选假设专属的特异性证据数量与全局共性证据数量。γ 越高，说明该假设的得分更多来源于排他性的特异性证据，此时可适当降低结论输出的判定门槛。ϵ 为防止分母为零的平滑项。

综合上述特征，系统设计的动态阈值函数 τ_{dynamic} 形式化为：

$$\tau_{\text{dynamic}} = \tau_{\text{base}} \left(1 + \frac{\omega - \gamma}{2}\right). \tag{4.9}$$

其中 τ_{base} 为基础阈值。当竞争假设高度相似时（ω 较大），判定阈值升高，使系统更倾向于触发后续的质证环节；当包含较多特异性表征时（γ 较大），阈值降低，允许系统在证据充分时直接输出医学结论。

结合动态阈值，核心路由策略 $\pi(\mathcal{H}_{\text{sorted}})$ 定义如下：

$$\pi(\mathcal{H}_{\text{sorted}}) = \begin{cases} \text{Output}(h_{(1)}), & \text{if } \Delta \geq \tau_{\text{dynamic}} \\ \text{Debate}(h_{(1)}, h_{(2)}), & \text{if } \Delta < \tau_{\text{dynamic}} \end{cases} \tag{4.10}$$

该策略包含两条执行路径：当 $\Delta \geq \tau_{\text{dynamic}}$ 时，系统触发早期退出机制（Early Exit），跳过多智能体交互环节直接输出 $h_{(1)}$ 作为最终解答，从而大幅节省推理算力；当 $\Delta < \tau_{\text{dynamic}}$ 时，系统淘汰排名第三及之后的假设，仅保留最易混淆的候选对 $\{h_{(1)}, h_{(2)}\}$ 及其对应的解耦证据集，进入下游的质证模块。

4.2.4 定向质证与共识裁决模块

当候选空间的置信度分布触发动态路由网关的 “定向质证” 路径时，系统将截断长尾假设，仅将最具混淆性的首选假设 $h_{(1)}$ 与竞争假设 $h_{(2)}$ 引入本模块。为突破单一大语言模型在复杂医疗推理中的思维局限与 “幻觉” 偏差，本模块实例化了一个包含正方（Proponent）、反方（Opponent）与裁决者（Judge）的非对称多智能体交互架构，通过限定轮次的质证完成最终的共识收敛。

首先，系统根据前置的证据解耦结果，对正反方智能体进行非对称的上下文初始化与角色提示。正方智能体 $\mathcal{A}_{\text{pro}}$ 被锚定于首选假设 $h_{(1)}$，并被单独注入其专属特异性证据 $\mathcal{E}^{(1)}_{\text{spec}}$，其优化目标是最大化 $h_{(1)}$ 的成立概率；反方智能体 $\mathcal{A}_{\text{opp}}$ 则被锚定于

竞争假设 $h_{(2)}$，同时被注入反方特异性证据 $\mathcal{E}_{\text{spec}}^{(2)}$ 与全局共性证据 $\mathcal{E}_{\text{shared}}$，其目标是利用共性表征的模糊性对 $h_{(1)}$ 的证据链进行逻辑拆解与攻击。

在完成实例化后，系统启动多轮次的定向质证引擎。定义最大交互轮次为 T，在任意轮次 $t \in [1, T]$ 中，正反双方的推理状态更新可形式化为条件生成函数：

$$A_{\text{pro}}^{(t)} = \mathcal{M}_{\text{pro}}\left(h_{(1)}, \mathcal{E}_{\text{spec}}^{(1)}, A_{\text{opp}}^{(t-1)}\right), \tag{4.11}$$

$$A_{\text{opp}}^{(t)} = \mathcal{M}_{\text{opp}}\left(h_{(2)}, \mathcal{E}_{\text{spec}}^{(2)} \cup \mathcal{E}_{\text{shared}}, A_{\text{pro}}^{(t)}\right), \tag{4.12}$$

其中 $A_{\text{pro}}^{(t)}$ 与 $A_{\text{opp}}^{(t)}$ 分别代表正反双方在第 t 轮生成的质证论点，初始状态 $A_{\text{opp}}^{(0)}$ 为空，$\mathcal{M}$ 为智能体的推理映射函数。通过式 (4.11) 与 (4.12) 的交替执行，多智能体系统在受限的上下文空间内，针对 $h_{(1)}$ 与 $h_{(2)}$ 的鉴别边界展开深度的多步逻辑博弈。由于质证范围仅限于两个高度相关的假设，该方式能够确保模型注意力始终聚焦于核心差分特征上。

当交互轮次达到 T 或触发提前终止条件后，系统将完整的质证历史轨迹 $\mathcal{H}_{\text{debate}} = \left[A_{\text{pro}}^{(1)}, A_{\text{opp}}^{(1)}, \ldots, A_{\text{pro}}^{(T)}, A_{\text{opp}}^{(T)}\right]$ 打包传递给具备全局视野的裁决者智能体 $\mathcal{A}_{\text{judge}}$。

裁决者智能体被赋予最高决策权重，其内置指令要求其保持中立，并基于医学循证逻辑对 $\mathcal{H}_{\text{debate}}$ 中的推理链进行终审。裁决模型的最终输出定义如下：

$$\langle D_{\text{result}}, R_{\text{chain}} \rangle = \mathcal{M}_{\text{judge}}\left(h_{(1)}, h_{(2)}, \mathcal{H}_{\text{debate}}\right), \tag{4.13}$$

其中，$D_{\text{result}} \in \{h_{(1)}, h_{(2)}, \text{Uncertain}\}$ 为最终的医学结论，R_{chain} 为包含关键鉴别特征与循证依据的解释性推理路径。通过引入该裁决机制，系统不仅输出单一的分类标签，还同步生成了高度可解释的决策论证过程。

4.3 实验结果与多维场景分析

4.3.1 医疗基准与数据集介绍

为全面评估所提出的多智能体医疗鉴别与动态路由诊断系统，本研究选取了四个难度呈阶梯式分布的代表性基准数据集，全面覆盖常规病、长尾罕见病及中医辨证场景：

- **MedQA**：源自美国执业医师资格考试（USMLE），包含 1,272 条长文本微病历多选题。主要评估模型在专业临床场景下的逻辑推理与常规问答能力。

- **MedMCQA**：源自印度医学研究生入学考试的大规模多学科题库，本研究采用其验证集中的 4,183 条样本。重点检验系统处理高密度跨学科知识时的泛化能力及抗噪声干扰表现。
- **MedR-Bench**：基于真实临床病例报告构建，包含 1,453 个结构化病例（含 656 个罕见病样本）。专门用于评估系统在复杂真实场景中的决策能力，以及鉴别长尾疾病、缓解“检索诱导偏移”的有效性。在数据格式构成上，每个样本不仅保留了原始的完整临床报告文本，还提供了标准化的病例摘要、详尽的鉴别诊断逻辑思路等。
- **龙医临床诊疗数据集（Longyi-120）：**该数据集为本研究基于名老中医学术经验专著《龙医华章》自主构建。通过设计基于大模型的信息抽取工作流，从非结构化文献中精准结构化了 120 例真实名医临床医案。数据详尽涵盖了患者的四诊信息（包含现病史、症状及舌脉象）、中西医诊断与深层辨证结论。在本研究的中医诊断 RAG 评估实验中，我们将该数据集作为测试基准（Benchmark），并选用开源的**神农中医药数据集（ShenNong-TCM-Dataset**[①]**）**作为系统的外部知识库语料。神农数据集是一个基于中医药知识图谱抽取的开源高质量语料库，内含逾 11 万条数据，全面覆盖了中药、方剂、证候、疾病等核心医疗实体及其关联关系。利用该知识丰富的底座语料，旨在高标准评估本系统在应对真实长文本医案时，处理“同病异治”等中医异构逻辑以及相似证候鉴别时的知识召回率与推理鲁棒性。

在本文实验中，MedQA 与 MedMCQA 主要用于评估系统在常规医学问答场景下的基础推理能力；MedR-Bench 中的罕见病样本主要对应疾病频率长尾与证据稀疏长尾；Longyi-120 中医医案则主要对应知识体系长尾。通过这种划分，本文能够分别考察系统在常规病、罕见病以及中医复杂辨证场景下的适用性。

4.3.2 实验评估体系与基线模型设置

为了全面且客观地衡量系统在不同复杂度和不同医学体系下的推理性能，本研究针对选择题形式的常规医学基准与长文本生成形式的罕见病与中医医案基准，分别采用了不同的评估指标体系：

- **常规问答与罕见病推理评估（Accuracy）：**针对形式不同的西医基准测试，本

① https://huggingface.co/datasets/michaelwzhu/ShenNong_TCM_Dataset

实验采用统一的准确率（Accuracy, Acc）作为核心衡量指标，但在具体核验方式上有所区分：对于选择题形式的常规医学基准（如 MedQA 与 MedMCQA），通过严格比对模型输出选项与标准答案得出；对于基于真实长文本病历的罕见病鉴别任务（如 MedR-Bench），则重点核验模型最终输出的疾病名称实体是否准确命中真实的医学金标准。该指标直观且严格地反映了系统在常规临床问答与长尾疾病排查上的绝对精确度。

- **中医复杂医案推理评估（ROUGE-L 与加权综合准确率）：**考虑到中医真实医案（如 CMB-TCM 与真实病例集）通常要求模型输出包含具体证候、治法或方剂的自然语言文本，简单的选项匹配无法准确衡量其推理质量。本实验从以下两个角度综合评估中医推理的准确性：

  1. **ROUGE-L 分数：**用于衡量系统生成的中医解答与标准答案在字面序列上的最长公共子序列重叠度，客观反映文本级别的生成保真度。
  2. **加权综合准确率（Answer Accuracy, Acc_{hybrid}）：**由于中医表述存在“异词同义”现象，本研究引入了包含大模型事实核查的综合评估公式：

$$Acc_{hybrid} = \alpha \cdot FC + (1 - \alpha) \cdot SS \tag{4.14}$$

  其中，FC（Fact Check）为事实核查得分，由具备医学指令遵循能力的大语言模型作为中立裁判，对生成的证候与标准答案进行事实一致性判定；SS（Semantic Similarity）为基于向量的语义相似度得分，用于衡量生成内容与参考答案在隐式语义空间中的特征逼近程度。本实验设定平衡超参数 $\alpha = 0.5$，以确保事实严谨性与语义包容性的权重对等。

  3. **推理质量评分：**针对中医诊疗中辨证论治的复杂逻辑特性，本研究引入强基座模型作为专业领域裁判，对生成结果进行多维度的量化打分。裁判模型需结合标准医案，从辨证严密性（Dialectical Rigor, S_{DR}）与逻辑连贯性（Logical Coherence, S_{LC}）两个核心维度，分别给出 1 至 5 分的评分。最终的综合得分 $Score_{LLM}$ 由各维度得分的加权平均求得：

$$Score_{LLM} = w_1 \cdot S_{DR} + w_2 \cdot S_{LC} \tag{4.15}$$

  其中，w_1, w_2 为各评估维度的权重参数（本实验设定权重均等，即 $w_1 = w_2 = 0.5$）。该指标能够更直观地量化检索增强生成在医疗垂直领域中的

表 4.1 本方法结合不同大语言模型在 MIRAGE 基准上的性能表现。下划线数据表示在使用 Qwen3-8B 基座模型时，本方法与基线对比的更优结果。加粗数据表示本方法在不同基座模型中的最高分（按具体语料库和评估指标划分）。MedRAG 的分数取自所有 MedRAG 方法在相应指标上的最高得分。

语料库	方法	大模型	MedQA	MedMCQA	Avg.
Textbooks	MedRAG	Qwen3-8B	77.12	63.10	70.11
	Speculative-RAG		76.85	64.63	70.74
	CF-RAG		77.31	62.87	70.09
	Our Method	Qwen3-8B	<u>77.83</u>	<u>65.74</u>	<u>71.79</u>
		Qwen3-32B-30A	80.58	71.72	76.15
		GPT-5.1	**94.18**	**78.70**	**86.44**
PubMed (500K)	MedRAG	Qwen3-8B	77.45	66.24	71.85
	Speculative-RAG		77.63	65.81	71.72
	CF-RAG		76.95	66.57	71.76
	Our Method	Qwen3-8B	<u>79.01</u>	<u>67.42</u>	<u>73.22</u>
		Qwen3-32B-30A	85.69	69.33	77.51
		GPT-5.1	**93.87**	**83.60**	**88.74**
StatPearls	MedRAG	Qwen3-8B	<u>76.73</u>	64.68	70.71
	Speculative-RAG		75.89	65.23	70.56
	CF-RAG		75.45	64.21	69.83
	Our Method	Qwen3-8B	75.08	<u>67.30</u>	<u>71.19</u>
		Qwen3-32B-30A	80.97	70.52	75.75
		GPT-5.1	**94.03**	**78.75**	**86.39**

临床解释质量与逻辑可靠性。

4.3.3 在常规病场景下的对比实验结果

由表 4.1可见，通过综合评估 MedQA 与 MedMCQA 这两项核心医学问答指标，本研究能够全面反映各模型在不同外部医学语料库（Textbooks、PubMed 与 StatPearls）支持下的知识抽取与复杂推理能力。在使用相同的基础模型 Qwen3-8B 时，本方法在绝大多数测试中均显著优于传统的 MedRAG 基线，在整体表现上也超越了近期提出的复杂检索增强模型（Speculative-RAG[82] 与 CF-RAG[83]）。例如，在 Textbooks 和 PubMed (500K) 语料库上，本方法的平均得分（Avg.）分别达到了 71.79% 和 73.22%，

相较于 MedRAG 的 70.11% 和 71.85% 有了稳定的提升。相比之下，CF-RAG 等方法在引入反事实等复杂机制时极易受医疗高噪声干扰，导致其在 Textbooks 等语料上的提升幅度十分有限，甚至在部分子项上出现退化。这说明我们的抗漂移机制能够过滤外部检索知识中的噪声干扰，帮助模型在面对需要精准事实召回的医疗多选题时，做出更加可靠的上下文召回。

值得注意的是，在 StatPearls 语料库下，本方法结合 Qwen3-8B 的 MedQA 单项得分（75.08%）略低于 MedRAG 基线（76.73%）。不仅如此，Speculative-RAG（75.89%）与 CF-RAG（75.45%）在该测试中同样遭遇了性能滑坡，整体表现甚至不及最基础的 MedRAG。这一集体性的指标下探揭示了语料特性与模型推理机制之间的微妙关联：MedQA 主要由复杂的长题干临床病例组成，需要多步逻辑推理；而 StatPearls 语料本身偏向于高度凝练的临床诊疗速查手册。对于较小的 8B 模型而言，在严格的抗漂移约束或复杂的验证干扰下，当检索到的密集短文本未能完全覆盖长病例的全部特征时，模型可能会过度聚焦于局部片段，从而在一定程度上限制了其自身内部权重的全局推理惯性。这也进一步体现了在特定知识库下，过于复杂的 RAG 机制对小模型推理能力可能存在的负面影响。

此外，随着基座模型参数规模的扩大，本方法展现出了极强的性能扩展性。在引入 Qwen3-32B-30A 以及 GPT-5.1 后，模型的各项指标均实现了显著提升。特别是在 PubMed 语料库上，GPT-5.1 取得了所有测试中的最高平均分 88.74%，其中 MedQA 单项更是高达 93.87%。对比不同语料库的增益趋势可以发现，结构严谨、事实密度高的标准医学文献摘要（如 PubMed）与超大语言模型强大的深层阅读理解及交叉验证能力最为契合。大模型不仅能够提取事实，还能对多篇文献进行逻辑重构，将外部知识转化为解答医学难题的有效依据。

4.3.4 在罕见病场景下的对比实验结果

表 4.2 展示了本方法在 MedR-Bench 上的性能对比。在具体对比设定上，基线方法主要分为单轮（1-turn）与自由多轮（free-turn）交互模式。在这两种基线模式中，系统均引入了一个由大语言模型扮演的高级临床专家角色。在单轮设定中，诊断模型仅基于初始病历进行一次性推断；而在自由多轮设定下，诊断模型被允许向该专家主动发问，动态索取相关的实验室检验、影像学等额外辅助检查信息，直至模型认为收集到充分的证据并给出最终诊断。实验结果显示，允许模型通过多步追问获

表 4.2　本方法结合不同基座模型在 MedR-Bench 数据集上的性能对比。其中 “1 turn” 设定表示模型仅进行单轮前向推理；“free-turn” 设定表示允许模型进行多轮交互与迭代推理；“With GPT-4o” 设定表示引入 GPT-4o 作为高级医学专家，根据模型的初诊请求动态提供相应的辅助检查结果。

大模型	方法	All Disease Acc.	Rare Disease Acc.
QwQ-32B	With GPT-4o (1-turn)	63.74	64.15
	With GPT-4o (free-turn)	74.71	**73.93**
	Our Method	**74.74**	71.65
	Our Method (with 4o)	71.51	68.90
DeepSeek-v3.2	With GPT-4o (1-turn)	68.65	71.49
	With GPT-4o (free-turn)	74.12	76.03
	Our Method	**75.57**	**76.22**
	Our Method (with 4o)	71.03	71.49
DeepSeek-R1	With GPT-4o (1-turn)	71.79	70.67
	With GPT-4o (free-turn)	76.18	77.60
	Our Method	78.46	77.74
	Our Method (with 4o)	**80.52**	**79.57**

取信息的 free-turn 基线表现普遍优于 1-turn 设定，这印证了临床鉴别诊断中动态获取证据对锁定病因的重要性。本方法（Our Method）通过对比因果假设扩展和多智能体辩论机制，在不引入外部专家的情况下，就在 QwQ-32B 和 DeepSeek-v3.2 上取得了超越多轮交互基线的成绩（全疾病准确率分别为 74.74% 和 75.57%），说明针对证据空间中特异性证据的搜索确实能有效引导模型注意到正确的重点。而在罕见病诊断的表现上，本方法在 DeepSeek-v3.2 和 DeepSeek-R1 模型上取得了超越基线模型的表现 (76.22% 和 77.74%)，这能够证明我们提出针对多个候选竞争目标进行证据探索后，定向辩论的通路的确能够在罕见病等复杂场景下提升对常见证据的抗干扰能力，从而实现更加准确的判断。此外，实验观察到专家信号（GPT-4o）的影响具有模型依赖性：中等规模模型在引入动态专家反馈后，因难以处理新增的大量线索而导致性能下滑；而具备强推理能力的 DeepSeek-R1 则能通过有效的过滤产生正向协同效应，其全疾病准确率（80.52%）与罕见病准确率（79.57%）均达到全局最高，充分体现了本系统在处理长尾罕见病等复杂挑战时精准鉴别诊断的能力。

表 4.3　不同 RAG 方法在 Longyi-120 中医临床诊疗数据集上的性能对比。本实验统一采用开源的神农中医药数据集（ShenNong-TCM-Dataset）作为检索增强的外部知识库语料。

方法	ROUGE-L	Accuracy	LLM-Judge
Zero-Shot LLM			
Qwen3-8B	0.1614	0.5241	0.5742
Qwen3.5-9B	0.2005	0.5521	0.7034
Standard RAG			
BM25 (Qwen3-8B)	0.1967	0.5631	0.5942
BM25+Dense (Qwen3-8B)	0.1836	0.5367	0.6025
Advanced RAG Frameworks			
CRAG (Qwen3-8B)	0.1965	0.5812	0.6650
Speculative-RAG (Qwen3-8B)	0.2399	0.5849	0.6513
Speculative-RAG (Qwen3.5-9B)	0.2651	0.6032	0.7337
CF-RAG (Qwen3-8B)	0.2702	0.5897	0.6897
Our Method(Qwen3-8B)	**0.2651**	**0.6003**	**0.6949**
Our Method(Qwen3.5-9B)	**0.2603**	**0.6219**	**0.7347**

4.3.5　在中医诊疗场景下的对比实验结果

表 4.3 展示了各类基线模型与本文提出的推理框架在 Longyi-120 中医临床诊疗数据集上的综合评估结果。首先，从外部知识注入的必要性来看，对比零样本生成（Zero-Shot LLM）与标准检索增强（Standard RAG），引入神农中医药数据集作为外部语料后，BM25 (Qwen3-8B) 在字面命中率（ROUGE-L 提升至 0.1967）与准确率（Accuracy 提升至 0.5631）上均有基础提升，证明了外部医学知识的注入能有效缓解大模型的事实性幻觉。然而，当采用更为复杂的混合检索（BM25+Dense）时，其 ROUGE-L 与 Accuracy 反而出现了轻微下降（分别为 0.1836 与 0.5367）。这一反直觉的“检索退化”现象，有力印证了本文在引言中提出的核心痛点：在中医这种高度依赖隐式语义的场景中，直接基于语义相似度的检索极易引入大量包含“常见偏见”的干扰证候，导致模型陷入“相关性陷阱”，进而引发推理漂移。其次，观察前沿 RAG 框架在处理异构逻辑时的表现。Advanced RAG 框架系列（如 CRAG、Speculative-RAG 与 CF-RAG）通过对检索内容进行评估或反思，相较于标准 RAG 取得了明显的性能提升。其中，注重细粒度特征的 CF-RAG (Qwen3-8B) 在 ROUGE-L 上达到了 0.2702 的高分。但值得注意的是，这类方法在 Accuracy（最高 0.5897）与逻辑严密性 LLM-Judge（最高 0.6897）的提升上遭遇了明显瓶颈。这表明，仅依靠单线索的检索修正

表 4.4　本方法在 MedR-Bench 数据集上的消融实验结果（C 表示对比候选与多源检索模块，D 表示定向质证与共识裁决模块）。Token Cost/Q 表示单次查询的系统平均 Token 消耗估算量。

大模型	All Disease Acc. (%)	Rare Disease Acc. (%)	Token Cost/Q
Qwen3-8B (free-turn)	51.62	51.07	5.8K
Qwen3-8B + C	54.03	53.35	7.2K
Qwen3-8B + C + D	**54.43**	**54.64**	**12.5K**
QwQ-32B (free-turn)	70.12	68.21	6.3K
QwQ-32B + C	72.13	69.05	9.2K
QwQ-32B + C + D	**73.64**	**70.58**	**16.6K**
DeepSeek-v3.2 (free-turn)	71.50	70.80	6.7K
DeepSeek-v3.2 + C	73.23	72.56	7.4K
DeepSeek-v3.2 + C + D	**74.54**	**73.60**	**16.4K**

与反思，难以应对中医医案中类似“同病异治”的复杂排他性鉴别任务。模型虽然在字面上组合出了更像标准答案的文本，但在深层病机与理法方药的严密推演上仍显吃力。本文提出的自适应医疗推理框架展现出了优势。在同等 8B 参数规模下，本文方法取得了 0.6003 的 Accuracy 与 0.6949 的 LLM-Judge 评分，全面超越了本实验所选取的多种主流 Advanced RAG 框架。尽管其 ROUGE-L（0.2651）略低于 CF-RAG，但结合显著提升的 Accuracy 与 LLM-Judge 可以推断：本文系统并非机械地抄写检索片段，而是通过多智能体差分质证与共识裁决模块，输出了经过深度逻辑推演和因果溯源的定制化辨证分析。此外，当基座模型升级至 Qwen3.5-9B 时，本文方法的性能得到了进一步提升，Accuracy 达到全局最高的 0.6219，LLM-Judge 升至 0.7347。这充分验证了本文提出的证据解耦与自适应路由机制能够为大模型提供更为纯净的逻辑支撑，在真实临床辅助场景中展现出推理鲁棒性。

4.3.6　基座模型规模对系统效能影响的消融实验分析

本研究通过系统实验验证了所提出的多智能体医疗推理框架在不同参数规模的基座模型（Qwen3-8B、QwQ-32B 与 DeepSeek-v3.2）配置下的有效性，如表4.4。首先，从 MedR-Bench 数据集的全疾病准确率与罕见病准确率指标可以看出，实验结果呈现出符合客观规律的模型扩展（Scaling）梯度：8B 量级的小模型受限于自身相对薄弱的医学先验与逻辑推演能力，初始鉴别基线在 51% 左右徘徊；而 32B 及以上的大模型则能稳定在 70% 以上。尽管基础能力差异显著，但在引入本框架的模块后，

不同规模模型均取得了提升，体现出本框架在应对复杂医学特征场景下的稳健性。

进一步的消融实验揭示了不同规模模型对各模块的吸收与转化差异。对比候选与多源检索（C）模块通过发散式的多源检索提供了可靠的候选拓展线索，对所有模型均带来了直观的增益（例如 Qwen3-8B 从 51.62% 提升至 54.03%）。然而，定向质证与共识裁决（D）模块的效能发挥，对基座模型的内生推理能力提出了更高要求。在 Qwen3-8B 模型上，完整叠加 D 模块后，全疾病准确率仅微增至 54.43%，罕见病微增至 54.64%。这表明较小参数规模的模型在执行复杂的多智能体交叉辩论时，受限于自身的逻辑容量，无法有效提炼高质量共识。相比之下，在具备更强推理能力的 QwQ-32B 与 DeepSeek-v3.2 上，D 模块的效果更好。以性能上限最高的 DeepSeek-v3.2 为例，完整引入双模块后，不仅全疾病准确率提升至 74.54%，罕见病准确率更实现了较大幅度的跃升（至 73.60%），有效缓解了长尾疾病的识别落差。

结合上述实验，本研究在每次查询的系统资源开销（Token Cost/Q）与诊断准确性之间进行了合理的权衡分析。随着对比候选构建与定向辩论模块的叠加，各量级系统的整体 Token 消耗量均出现了预期内的明显上涨（例如 Qwen3-8B 增加至 12.5K，DeepSeek-v3.2 升至 16.4K）。结合前述的性能增益梯度可知，对于小模型而言，盲目上马多智能体复杂工作流的“算力性价比”偏低；但对于具备强基座能力的大模型，这种计算资源的集中投入成功换取了鉴别诊断能力的较大提升，实现了算力向逻辑的转化。最终的消融对比结果验证了所提出方法的可行性和稳定性，递进式的模块融合策略能够充分发挥大模型的多视角思辨能力，同时保留严密的医学约束，从而在真实临床辅助场景中实现更高的准确性与深度可解释性。需要指出的是，多智能体质证机制虽然能够增强推理过程的可解释性，但其最终效果仍然受到基座模型能力、检索证据质量和提示词稳定性的共同影响。此外，裁决者智能体本身也可能继承基座模型的偏见，对表述更充分或更符合常见病模式的一方给予更高权重。因此多智能体裁决仍不能替代真实临床诊断，其结果应被视为辅助性推理建议，最终仍需由专业医生结合临床检查进行判断。

4.3.7 长尾医疗检索的可解释性案例分析

如图 4.2 所示，本文从罕见病数据集 MedR-Bench 中选取了一个具有典型长尾特征的肺部感染病例作为可解释性案例分析对象。该病例的真实诊断为帕博利珠单抗相关肺炎接受大剂量糖皮质激素治疗后继发的肺诺卡菌病。该病例的难点在于，不

同候选诊断共享了“肺炎”“呼吸衰竭”等表面临床特征，但其因果解释并不相同。患者既往存在非小细胞肺癌和帕博利珠单抗治疗史，影像学曾提示机化性肺炎，因此普通检索容易召回免疫治疗相关肺炎或药物性肺损伤文献，并将其作为后续呼吸衰竭的主要解释。然而，患者在接受大剂量糖皮质激素治疗约两周后迅速恶化，这一时间线更符合免疫抑制后继发机会性感染的发生机制。因此，本病例真正具有鉴

[User Query]:
【患者信息】：62 岁女性，既往有 1 型糖尿病、心血管疾病史，以及 30 包年吸烟史。已诊断为非小细胞肺癌。
【主诉】：起初为轻度呼吸短促，随后进展为明显呼吸窘迫。
【现病史】：患者于 2020 年 8 月出现乏力；2021 年 12 月在接受手术切除及短程化疗后，被诊断为 4 期非小细胞肺癌。随后接受帕博利珠单抗治疗 13 个月，整体耐受良好。2023 年初的 CT 检查显示符合机化性肺炎的影像学表现，考虑为药物诱导性肺炎。患者接受大剂量糖皮质激素治疗，即泼尼松 50 mg 起始后逐渐减量，同时暂停免疫治疗。
糖皮质激素治疗开始两周后，患者在一次加勒比海邮轮旅行期间出现呼吸衰竭，需要气管插管和静脉-静脉体外膜肺氧合支持。之后通过医疗转运被送往加拿大。
【既往史】：1 型糖尿病、心血管疾病、30 包年吸烟史。
【个人史】：在糖皮质激素减量治疗开始 2 周后参加加勒比海邮轮旅行。
【体格检查】：初期无发热，也无明显阳性体征；随后在糖皮质激素治疗后出现严重低氧血症。初始 CT：双侧磨玻璃影和外周网状影，符合机化性肺炎表现。呼吸窘迫后的复查 CT：弥漫性小叶中心性磨玻璃结节、树芽征以及坏死性实变。
[Ground Truth]:
因帕博利珠单抗相关肺炎接受大剂量糖皮质激素治疗后，继发发生的肺诺卡菌病。

Naive RAG (COT + BM25 + Dense Retrieval)
【证据池召回】:
1. [PMID 30271171] 放射召回性肺炎是一种现象，即某种召回触发药物会在肺部诱发急性炎症反应，且炎症部位对应于既往接受过放射治疗的区域。既往已有报道称，多种细胞毒性抗癌药物和分子靶向药物治疗后可发生放射召回反应；然而，仅有少数报道描述了免疫检查点抑制剂诱发的放射召回性肺炎。 **（强误导证据）**
2. [PMID 30271171] 本文描述了一例 67 岁女性晚期肺腺癌患者，她在接受帕博利珠单抗治疗后发生了伴肺泡出血的间质性肺病。患者接受了 4 个疗程的帕博利珠单抗治疗，并获得部分缓解。她没有呼吸道症状；然而，胸部 X 线和计算机断层扫描显示磨玻璃影和铺路石样改变..... **（强误导证据）**
3. [PMID 34468195] 肺炎链球菌是社区获得性肺炎最常见的病原体，可导致严重的侵袭性感染，包括脑膜炎和菌血症。大环内酯类药物的广泛使用被认为会增加耐大环内酯肺炎链球菌的流行，从而导致肺炎链球菌肺炎患者治疗失败.....
【最终结论】 ✗
基于当前召回的证据池，大模型倾向于认为患者的急性呼吸衰竭主要由帕博利珠单抗相关肺毒性进展所致。

Our Method
【竞争假设】：1. 普通细菌性肺炎 **2. 肺诺卡菌病** 3. 帕博利珠单抗相关肺炎加重（高混淆性候选项）
【候选证据池】:
1. [PMID 19195205] 一名 76 岁男性自 6 岁起患有哮喘。近期因肺炎导致哮喘加重。他在当地诊所接受了超过一周的糖皮质激素和抗生素治疗。虽然肺炎有所改善，但哮喘仍未得到控制，因此被收入我院。为治疗哮喘，患者接受了约 2 周的糖皮质激素治疗。胸部 X 线和 CT 检查显示，双肺下叶出现多个结节。随后进行了支气管镜检查，并在双肺下叶支气管肺泡灌洗液中检测到诺卡菌属细菌，因此诊断为肺诺卡菌病..... **（候选2的特异性证据）**
2. [PMID 4079425] 本文报告了一例由帕博利珠单抗诱发的放射召回性肺炎病例。患者为非小细胞肺癌术后局部复发患者。该病例表明，即使典型放射性肺炎已经缓解，帕博利珠单抗仍然可能导致严重的放射召回性肺炎。 **（候选3的特异性证据）**
3. [PMID 3914372] 粒细胞减少是癌症患者易发生感染的主要因素，感染主要由细菌引起。大多数感染由革兰阴性需氧菌导致，例如大肠杆菌、铜绿假单胞菌和克雷伯菌属。这些病原体通常来源于患者自身的胃肠道、黏膜或皮肤菌群，但这些菌群常常会受到医院获得性病原体的影响而发生改变。 **（共性干扰）**
4. ....
【高混淆项】：肺诺卡菌病 vs 帕博利珠单抗相关肺炎加重
【关键辩论】：在进一步比较肺诺卡菌病相关证据后我发现，PMID 19195205 报道了患者在接受约 2 周糖皮质激素治疗后发生肺诺卡菌病，并通过支气管肺泡灌洗液检出诺卡菌属而确诊。该证据与本病例中的核心时间线高度一致；相比之下，帕博利珠单抗相关肺炎虽然能够解释既往 CT 的机化性肺炎表现，但不能充分解释“糖皮质激素治疗后短期内灾难性恶化”这一转折点。 （模型关注到了其中的关键矛盾点）
【最终诊断】 ✓
肺诺卡菌病，可能是帕博利珠单抗相关肺炎接受大剂量糖皮质激素治疗后继发的机会性感染。

图 4.2 长尾医疗问答场景下的案例分析

别意义的线索是“糖皮质激素治疗后短期恶化”这一因果转折。

在 Naive RAG 设置下，系统主要依赖 COT 后的子查询进行 BM25 与向量检索。检索结果优先召回了多篇与帕博利珠单抗肺毒性、放射召回性肺炎以及普通感染性肺炎相关的文献。这些证据与病例具有较强的表面相关性，但大多只能解释患者既往存在免疫治疗相关肺炎风险，无法充分解释其在接受大剂量糖皮质激素治疗后短期内迅速恶化这一关键转折。相比之下，本文方法首先构建竞争性诊断假设，将普通细菌性肺炎、肺诺卡菌病以及帕博利珠单抗相关肺炎加重同时纳入候选空间。随后，系统对候选证据池进行证据特异性解耦，将仅能支持广义感染风险或免疫治疗肺毒性的文献划分为共性干扰或强混淆证据，而将“糖皮质激素治疗后发生肺诺卡菌病”相关文献识别为肺诺卡菌病候选的特异性证据。

在后续的定向质证阶段，系统锁定“肺诺卡菌病”与“帕博利珠单抗相关肺炎加重”两个高混淆候选项进行对比推理。帕博利珠单抗相关肺炎能够解释患者既往影像学异常和免疫治疗背景，但无法合理解释糖皮质激素治疗后短期内病情灾难性恶化；相反，肺诺卡菌病虽然属于长尾机会性感染，但其证据链能够同时解释患者的恶性肿瘤背景、糖皮质激素诱导的免疫抑制状态以及治疗后迅速进展的低氧性呼吸衰竭。由此，模型最终将注意力从高频但非特异性的免疫相关肺炎证据，转移到更具鉴别意义的“免疫抑制后机会性感染”证据上，并推理出肺诺卡菌病这一更符合病例发展过程的诊断。

该案例表明，在长尾医疗问答场景中，普通 RAG 容易被表面相关但缺乏鉴别力的高频证据牵引，导致模型沿着常见诊断路径发生推理漂移。本文方法通过竞争性假设构建、证据特异性解耦和多智能体质证机制，使模型能够显式比较不同诊断假设的证据支撑强度，突出真正具有排他性解释力的关键线索，从而提升复杂医疗推理过程的可解释性与可靠性。

4.4 本章小结

本章针对复杂医疗问答中检索增强（RAG）系统容易被常见病偏见干扰，导致模型推理出错的问题，提出了一种结合证据解耦与多智能体质证的推理框架。在具体流程上，系统首先主动生成多个具有竞争关系的候选假设，以提供明确的对比方向；随后通过计算证据的特异性，将支持特定疾病的核心证据与通用的干扰证据分

离开来；接着引入动态路由机制，根据候选假设的相似度自动决定是否提前输出结论，从而在保证准确率的前提下节省算力开销；最后，针对那些极易混淆的候选假设，系统引入多智能体进行交叉辩论与裁决，确保最终得出的结论符合严密的医学因果逻辑。

为了验证该方法的有效性，本章在常规医学考试（MedQA、MedMCQA）、复杂真实病例（MedR-Bench）以及自主构建的中医医案（Longyi-120）四个不同难度的数据集上进行了评估。实验结果表明，该方法在不同参数规模的模型上均优于 MedRAG 和 Speculative-RAG 等现有基线模型，尤其是在罕见病鉴别和中医辨证推理中优势更加明显。此外，消融实验进一步表明，基座模型规模越大，多智能体质证机制带来的算力消耗越能有效转化为诊断准确率的提升。总体而言，本章提出的方法切实缓解了大语言模型在复杂医疗场景下的幻觉问题，为构建可靠、有理有据的辅助诊疗系统提供了务实的解决方案。

第五章　基于多源知识引擎融合的复杂医学场景辅助诊断平台

5.1　应用背景

近年来，以 GPT-4、Med-PaLM 为代表的大语言模型在自然语言理解与推理方面展现出了前所未有的能力，为医疗辅助诊断与文献问答系统带来了新的范式[84]。然而，大语言模型在应对垂直且高风险的医疗领域时，存在固有的局限性。一方面，模型内部的参数化知识无法实时更新，难以覆盖最新发表的医学顶刊研究；另一方面，大模型在缺乏外部事实约束时极易产生“幻觉”，这在容错率极低的医疗诊疗与临床研究场景中是不可接受的。为了解决这一问题，检索增强生成技术通过在生成阶段引入外部知识库作为事实依据，RAG 有效缓解了模型的幻觉问题，并提升了医学回答的可溯源性。尽管现有的 RAG 技术在通用领域取得了显著进展，但当其落地于复杂的真实医学研究场景时，当前的 AI 辅助工具与系统平台仍暴露出了以下三个核心痛点：

1. **信息信源的单一性与专业权威性的缺失：**目前的主流 RAG 系统往往局限于单一的检索途径——通常仅支持对本地私有文档的静态向量检索，或仅依赖通用的 Web 搜索引擎。然而，通用搜索引擎存在严重的信息同质化与质量参差不齐问题，而传统 RAG 又缺乏对 PubMed 等高度权威的专业学术数据库的动态直连能力。这不仅导致研究工作流被严重割裂形成“信息孤岛”，更使得检索到的医学证据在专业深度、时效性以及临床权威性上难以满足严谨的医学科研与辅助诊疗需求。
2. **单次浅层检索难以应对长尾与跨域的复杂医学推理：**正如本文前述章节所探讨，简单的语义相似度检索在处理常见病时表现尚可，但在面对长尾知识（如罕见病的鉴别诊断）或中西医跨域的复杂推理任务时，极易陷入“相关性陷阱”并引入大量共性噪声。现有系统大多采用“一次检索、直接生成”的浅层工作流，缺乏对复杂异构知识的深度解耦能力，在面对需要多跳逻辑推理的长尾医学难题时表现较差，难以构建包含因果逻辑的证据链。
3. **缺乏动态工具调用与 Agentic RAG 机制的工程化落地：**传统 RAG 系统本质上

是一种被动的“检索-阅读”流水线，缺乏自主规划与动态调用外部工具的能力。在真实的复杂医学诊疗场景中，医生往往需要进行多步骤的探索与迭代。近年来兴起的代理式检索增强（Agentic RAG）具备极强的优越性，能够赋予模型动态调用多模态工具（如 Paper Reader 深度长文解析、PubMed 专业精准搜索、Plan & Reflect 规划与反思、内部权威知识库比对以及多智能体辅助诊断等）的能力。然而，大量相关的学术算法仍停留在静态测试脚本阶段。业界急需一个工程化平台，将这种具备高自由度工具矩阵的 Agentic RAG 框架进行系统级集成，并将多智能体的黑盒推理过程转化为白盒化的透明可视化工作流。

针对上述核心痛点，本章设计并开发了一个高度工程化的 AI 驱动医学辅助诊断平台（DeepMed Search）。该平台参考了 Agentic RAG 的核心理念：在底层集成了稠密向量检索（针对本地私有知识库）、主流互联网搜索引擎以及专业医学数据库（PubMed），实现了高权威性多源信源的整合；在上层工程架构中，完整集成了基于大语言模型的动态工具调用路由、异步多模态文档解析队列，以及包含自主规划、迭代检索与交叉验证机制的多智能体深度研究（DeepResearch）工作流。本章将详细阐述该医学平台的整体架构设计、多源异构检索机制的整合策略，以及从 Agentic RAG 理论算法到系统落地的工程实现细节。

5.2 系统整体设计

5.2.1 系统整体架构设计

为了实现多源异构医学检索与多智能体推理机制，并解决海量医学长文本的高并发处理问题，本平台采用基于微服务理念的分布式四层架构。系统整体自上而下分为：用户界面与交互接口、核心业务调度层、多源检索与 AI 引擎层以及底层数据与持久化层。各层之间通过标准化的 API 与异步消息队列进行解耦，保证了系统的高可用性、可扩展性以及底层大模型切换的灵活性。系统总体架构如图 5.1 所示。

5.2.2 系统技术栈选型

为了将本文提出的理论算法转化为实际可用的生产力工具，本平台在技术选型上严格遵循高可用、高并发与模块化解耦的工程规范。首先，在系统表现层与全栈框架的构建上，本平台采用基于 React 的现代化 Web 框架 Next.js，并结合 TypeScript

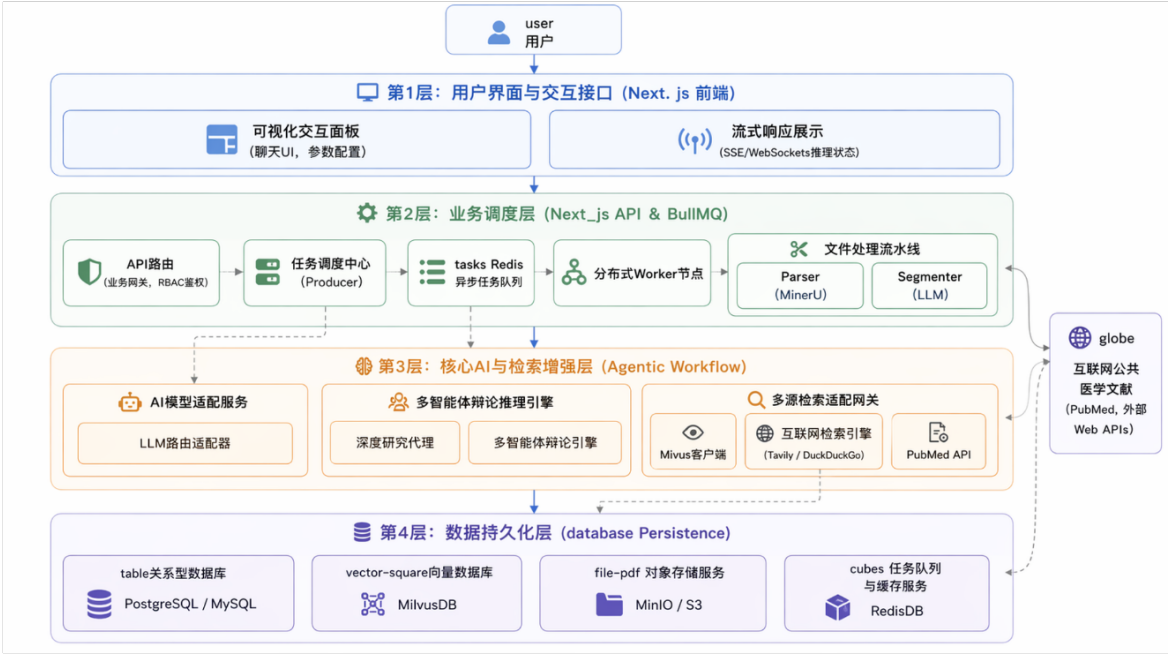

图 5.1 AI 驱动的医学辅助诊断平台 DeepMed 总体架构图

实现了严格的静态类型检查与代码安全。Next.js 的服务端渲染与边缘 API 路由不仅保障了系统可视化的流畅交互，还支撑了严密的基于角色的访问控制（RBAC）体系。

在底层数据架构方面，针对医疗平台多模态数据的异构特性，系统摒弃了单一的传统数据库方案，构建了关系型、向量化与对象存储相互协同的混合存储引擎。其中，PostgreSQL 结合类型安全的 Prisma ORM，专门负责持久化用户配置、知识库元数据及文档处理状态，确保了业务核心数据的强一致性（ACID）。面向海量医学文献的高维稠密检索需求，系统引入了高性能分布式向量数据库 Milvus，利用 HNSW 等高效索引算法存储并召回经大模型 Embedding 处理的文本特征向量。同时，考虑到原始医学病历、PDF 文献及解析生成的结构化 Markdown 文件属于典型的非结构化大文件，系统部署了兼容 S3 协议的 MinIO 对象存储服务作为统一的数据底座，实现多模态数据的高效解耦与安全存储。

针对海量医疗长文本解析与高维向量化所带来的阻塞等待问题，系统在业务调度层引入了基于 Redis 内存缓存的消息队列库 BullMQ。当研究人员批量上传文献或触发大规模重构任务时，API 网关仅作快速的状态拦截与记录，随后将诸如 PDF 逆向解析、大模型智能切分以及批量特征提取上库等重负载计算任务封装为异步作业，推入后台独立部署的 Worker 节点集群中排队消费。

为了给上层的质证推理提供精准、无偏的证据空间，系统接入了 MinerU 等先进的多模态结构化解析引擎，实现了复杂 PDF 向 Markdown 的精准还原。在检索源拓

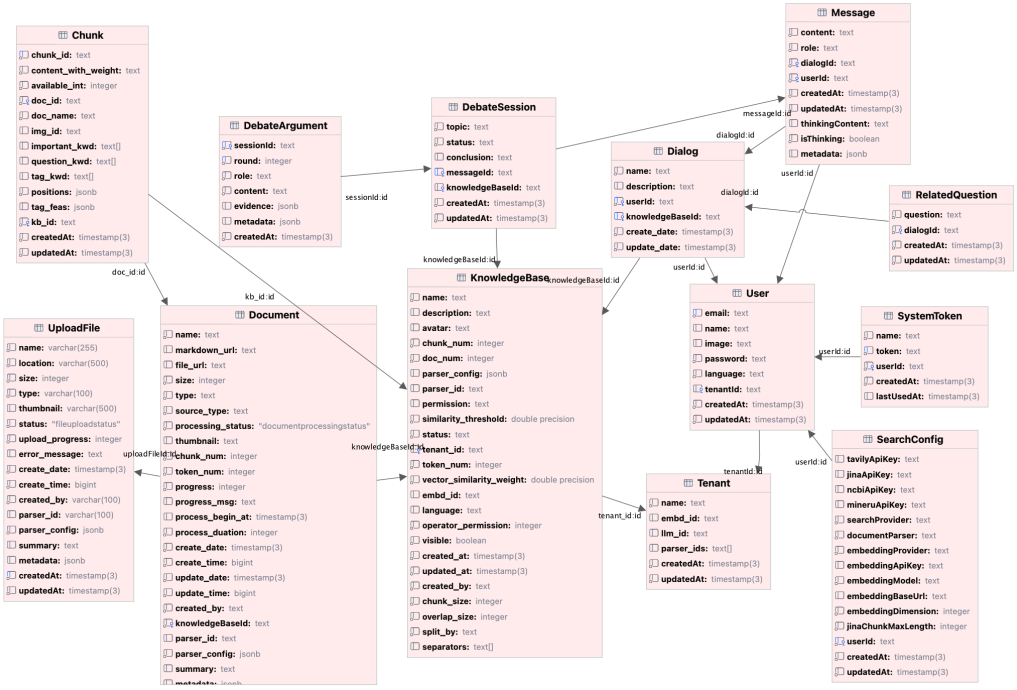

图 5.2　DeepMed Search 平台核心业务元数据实体关系图

扑上，本平台封装了统一的多源检索网关，不仅支持调用 PubMed E-utilities 接口获取权威的医学期刊库，还集成了 Tavily 与 DuckDuckGo 等互联网搜索引擎，为下游推理提供涵盖最新临床进展的证据视野。

5.2.3　系统数据库设计

如图 5.2 所示，数据库模型首先由三个核心实体表构成了完整的本地知识管理层级链路：KB（知识库表）作为顶层容器，封装了知识库的基本元数据（ID、名称、描述、大小），并通过 tenant_id（租户 ID）字段在物理层面严格隔离不同租户的数据，实现了多租户架构；Doc（文档表）用于追踪上传的医学非结构化大文件，通过 kb_id（知识库 ID）与 KB 建立 N:1 的关联，详细记录了文档的名称、路径、类型和异步解析状态；Chunk（数据块表）是 RAG 系统的核心检索单元，由 Doc 依据语义切割算法动态生成。Chunk 实体通过复合外键（kb_id, doc_id）同时关联 KB 和 Doc，其主要字段存储了具体的医学文本切片信息及溯源位置。

除此之外，针对系统的高阶 Agentic RAG 能力，数据库进一步扩展了多源搜索与推理辩论模块，以支撑复杂的医学辅助诊断：在多源搜索引擎的集成方面，系统设计了 SearchConfig（搜索配置表）。该表与具体的系统用户实体建立绑定关系，集中持久化了连接外部权威信源（如 NCBI/PubMed、Jina、Tavily）的 API 凭证，以及底层的

文档解析引擎（如 MinerU）和高维向量嵌入模型的环境参数；在深度推理与辩论方面，系统构建了 DebateSession（辩论会话表）与 DebateArgument（辩论论点表）的核心机制。DebateSession 与特定的对话消息（Message）强关联，负责管理一次复杂诊断推理的全局主题、运行状态与最终的裁判结论（Conclusion）；而 DebateArgument 则详细追踪了每一轮交锋中不同智能体角色（正方、反方、裁判）的具体论述内容。特别是该表中的 evidence 字段采用了高度灵活的 JSON 格式，专门用于记录“推理对齐的证据选择”（如关联的图谱节点或高特异性 Chunk ID）。

5.3 系统核心模块

为了满足真实临床辅助决策与医学文献分析的复杂需求，DeepMed 在底层架构与业务逻辑上进行了深度的模块化解耦。系统整体功能由四个核心模块构成：负责底层数据清洗与特征提取的知识库构建模块、提供标准化智能问答的 RAG 系统模块、打破信息孤岛的外部搜索引擎模块，以及执行复杂病理推演的深度研究模块。这四个模块共同构成了平台完整的智能化工作流。

5.3.1 知识库构建模块

知识库构建是整个辅助诊断系统的基础。在实际应用中，医学指南与临床病历多为包含双栏排版、统计图表及复杂分子式的 PDF 文件。直接进行暴力文本抽取会严重破坏医学上下文的逻辑关联。为此，本模块在代码底层接入了 MinerU 与 MarkItDown 等具备视觉多模态理解能力的解析引擎。当用户上传文件时，API 网关会将原始文件存入 MinIO 对象存储，并通过 BullMQ 消息队列发布异步解析任务。后台专属的 Worker 节点接管任务后，通过视觉模型精准识别版面布局，将复杂图文逆向还原为带有明确层级结构的纯净 Markdown 文本，从而避免了主线程阻塞并保障了高并发下的系统稳定性。

在获取结构化文本后，系统随即执行关键的文本切分与向量化入库。考虑到医学病理描述具有极强的长程依赖性，本系统摒弃了单一的按字符长度截断机制，开发了基于大语言模型的智能切割引擎。该引擎通过调用底层大模型通读段落上下文，自主识别出医学概念的完整语义边界，将长文本动态划分为语义自洽的知识块。随后，这些文本切片被送至专用的 Embedding 模型（如 Jina Embeddings）提取高维稠密向量，并批量持久化至 Milvus 向量数据库中。同时，系统在 PostgreSQL 关系型数

据库中严密记录了每个切片与原始文档的映射元数据，完成了从非结构化文献到可检索本地医学图谱的转换。知识库模块支持用户上传 PDF 等格式的文件对象，进行处理和切分之后，构建专属向量知识库，同时系统也支持用户对文档处理参数的自定义配置。

5.3.2 RAG 系统模块

在建立本地知识库的基础上，本平台实现了一套标准且高效的 RAG 问答系统，旨在为用户提供针对私有知识资产的精准对答。当用户在前端会话视窗中发起自然语言查询时，系统首先将查询文本转化为同维度的特征向量，随后通过 Milvus 客户端在指定的私有知识图谱中执行向量相似度检索，召回相关度最高的 Top-K 个文本切片。系统将这些切片作为上下文与用户的原始问题进行 Prompt 拼接，并提交给底层大语言模型进行推理生成。

为了赋予平台极强的算力适配弹性与医疗合规性，该模块开发了一套标准化的 LLM 服务路由适配器。该网关不仅原生集成了 OpenAI、Anthropic、Google 等主流商业 API，还通过标准化协议完美兼容了 DeepSeek、Qwen 等前沿模型，甚至允许医疗机构直连局域网内私有化部署的本地开源模型。在人机交互层面，系统依托服务器发送事件技术，将大模型的生成结果以极低的延迟逐字流式渲染在前端。同时，所有生成的医学结论均会在文本中自动生成脚注标号；用户点击标号即可溯源至具体的本地文档切片，提升了回答的严谨性与可信度。图5.3为 DeepMed 系统的 RAG 问答交互界面，支持用户选择不同模型和知识库进行问答，同时也会保存用户的过往会话历史记录。

为了验证本平台在专业长尾医疗场景（如中医诊疗）下的表现，我们在系统的知识库中接入了中医经典临床专著《龙医华章》。该书系统性地总结了海派中医名家的学术思想与疑难杂症诊治经验，其中详细阐述了陈以平教授在肾脏病领域的诊疗体系，特别是针对“血瘀”病理状态下“量变”与“质变”的演变规律，以及“新瘀”与“宿瘀”的精准辨证与用药法则。在面对此类高度专业化且对安全性要求极高的医学查询时，通用大语言模型的局限性十分明显。以“陈以平教授诊疗体系中新瘀与宿瘀的区分及用药差异”为例，如图 5.3 所示，未启用检索增强的原生 DeepSeek 模型出现了许多事实性幻觉，错误得出“新瘀宜温宜补”的结论并推荐附子、肉桂等温性药物，存在临床潜在风险。相比之下，DeepMed 平台精准检索了《龙医华章》

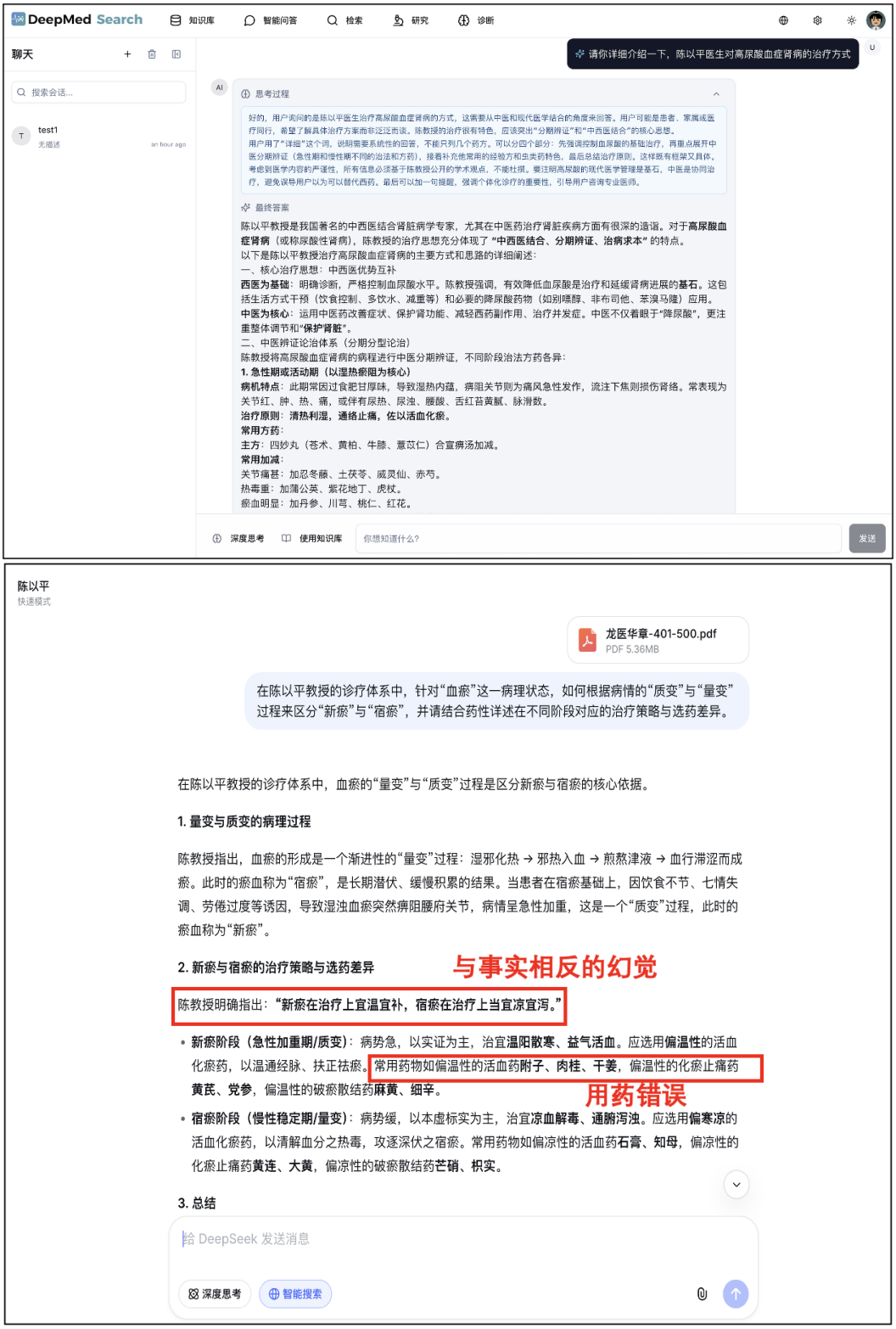

图 5.3　DeepMed RAG 系统模块与 DeepSeek RAG 系统模块对比

核心原文，底层模型在精准的文档解析支持和严密的事实约束下，准确输出了基于“量变与质变”的病理特征及符合原著的分期治疗与选药建议（如疏通血络、清热散寒等）。这一对比直观证明，本平台的 RAG 机制能克服大模型在专科长尾知识上的

图 5.4　DeepMed 搜索引擎模块界面

幻觉缺陷。

5.3.3　搜索引擎模块

医学研究要求极高的信息时效性与广度，单纯依赖本地私有数据库极易导致模型陷入“信息孤岛”，甚至在面对最新突发疾病时产生事实幻觉。为了打破这一局限，本模块在代码层面构建了一个多源且高度可配的外部检索协同网关。在垂直学术领域，系统深度封装了美国国家生物技术信息中心（NCBI）的 E-utilities 接口，允许系统直接将查询转化为专业的医学主题词，向全球最权威的 PubMed 数据库发起定向调用，实时获取经过同行评议的最新顶刊论文摘要与元数据。

在广域互联网通识层面，该网关平滑接入了 Tavily、DuckDuckGo 以及 Jina Search 等具备实时资讯抓取能力的搜索引擎 API。通过这套外部检索引擎，系统能够实时抓取开放网络上的医疗新闻、药监局公告及基础科普知识。在前端面板中，研究人员可以针对具体的研究场景，灵活配置各大检索源的开关状态。这种内外部协同的检索拓扑，使得平台能够将私有临床经验、前沿学术支撑与实时互联网资讯进行多维聚合，为下游的复杂推理提供极其丰满且无偏的全局证据池，如图5.4所示。

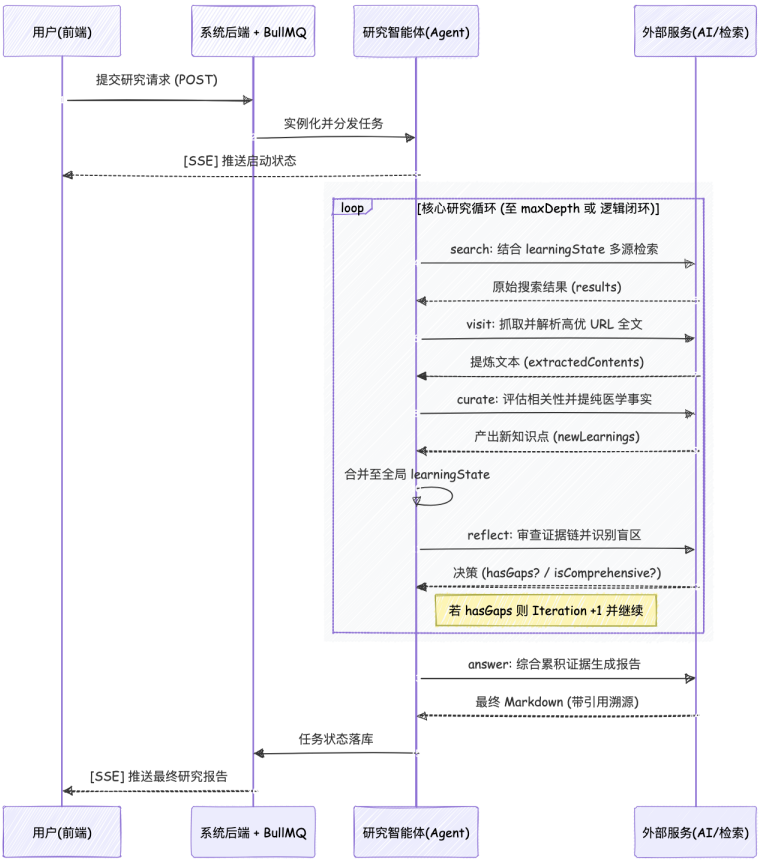

图 5.5 DeepMed 深度研究模块时序逻辑图

5.3.4 深度研究模块

传统 RAG 系统的“一次检索、单向生成”模式通常只能解决简单的显式事实问答，难以应对长尾疑难病症中需要多步因果推演的复杂逻辑。为此，本模块基于先进的代理工作流实例化了深度研究智能体（DeepResearch Agent），成为了整个平台的“核心大脑”。当用户输入一段症状错综复杂的病历并开启深度研究模式时，智能体不再急于输出最终结论，而是进入一个自主规划与动态深挖的循环闭环。

具体而言，该模块在代码底层的运行轨迹被严格划分为四个时序阶段，其后台 Worker、多智能体与外部检索引擎之间的状态流转与交互过程如图 5.5 的 UML 时序图所示。首先是规划阶段，编排代理自动对宏大的初始问题进行降维拆解，生成包含多项子任务的检索策略；其次是搜索阶段，系统并发调度前述的多源检索引擎，从本地库与 PubMed 中海量抓取候选片段；接着进入核心的策展阶段，智能体调用评估算法对每条候选证据的医学相关度与可靠性进行量化打分，精准剔除无效噪声；最后是反思阶段，智能体严格审视当前高质量证据链是否存在逻辑断层，如有缺失则

图 5.6　DeepMed 深度研究模块界面

自主生成新的衍生查询，开启下一轮深挖。在整个推演过程中，前端界面会通过流式通道将智能体的每一步计划制定、证据过滤理由及反思过程以可视化的形式逐步实时渲染。这种完全白盒化、具有自我纠错能力的动态推理机制，从根本上提升了平台在复杂医疗决策场景下的专业性与可靠性。图 5.6 展示了深度思考功能的整体过程。系统支持在有限 Token 预算限制下进行可控的深度研究，并会返回带有文献引用和网页引用的可解释性答案。

5.3.5　多源检索诊断模块

多源检索诊断模块是本平台执行高阶医学逻辑推演的核心工程落地。该模块旨在解决医疗检索增强生成系统在处理复杂长文本时，因检索证据与诊断逻辑不匹配而导致的“推理漂移”问题。为了夯实推理的数据基石，本模块首先利用国际权威公共医学文献库（PubMed）构建了大规模的专业医学知识库，并深度结合了 SNOMED CT 与 UMLS 等权威知识图谱进行因果关系抽取。通过将非结构化的海量语料与结构化的图谱相融合，系统构建了一个具备逻辑约束的因果关联网络。这一设计使得系统不仅能进行浅层的语义匹配，更能精准识别出临床表征与深层病理之间的隐式因果链路。

在具体的诊断执行逻辑上，基于上述因果网络构建了一套融合内省式验证机制的自主智能体推演工作流。该工作流严格遵循源自适应路由、因果内省式过滤与多

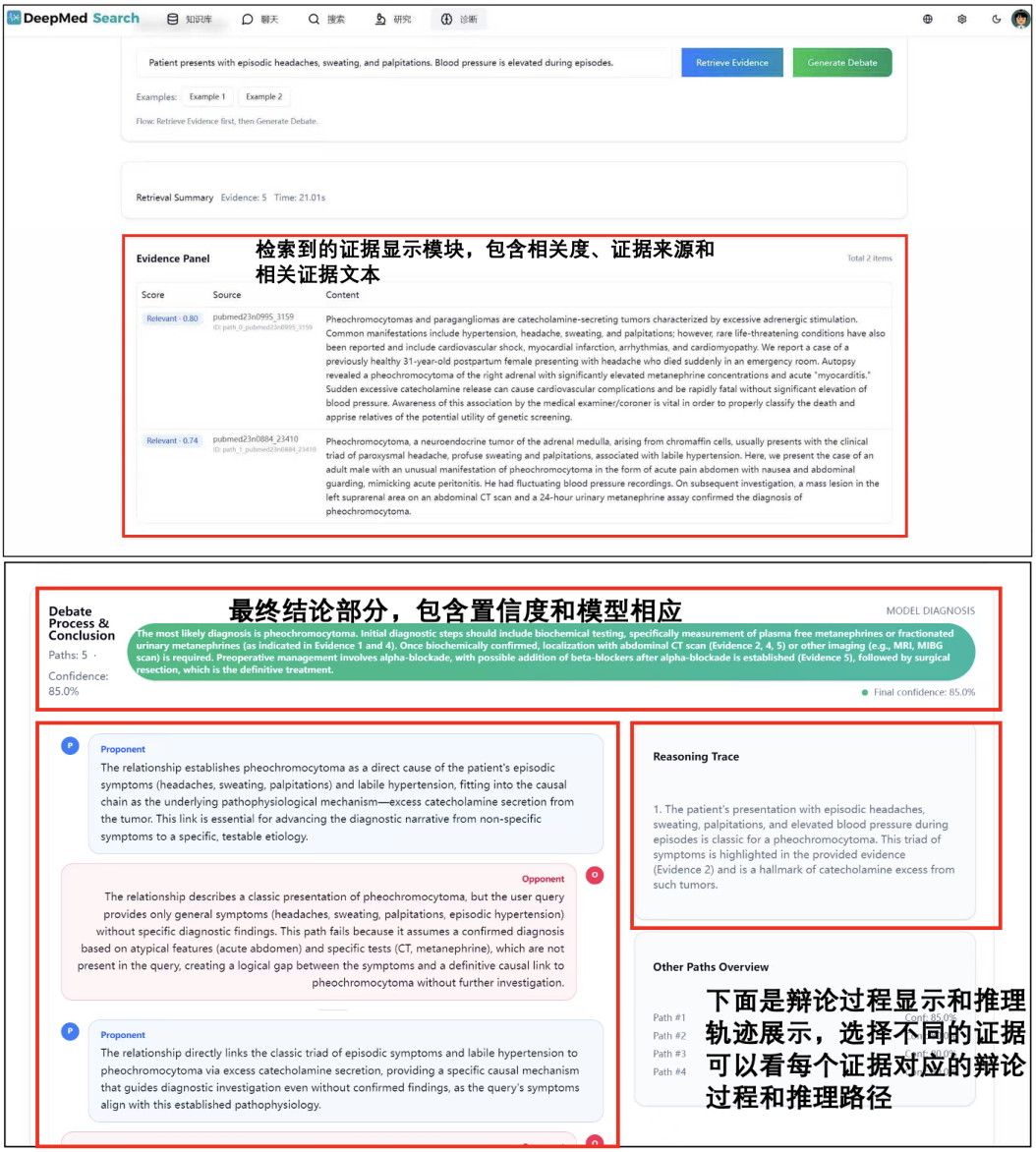

图 5.7　DeepMed 多源检索诊断模块界面

智能体合成三个核心阶段：首先，面对错综复杂的初始医学提问，编排智能体会对其进行原子化拆解，生成一系列子查询。随后，系统内置的源自适应路由中心会自主将检索任务分发至权威文献库、广域互联网以及本地私有知识图谱中。在此过程中，系统依托因果图谱路径辅助，精准识别医疗实体间的多跳因果关联，从而定向捕获多维度的深层证据。在获取海量候选证据后，系统进入因果内省式过滤阶段：为了剔除常见的共性偏见，系统主动执行对比假设发散，生成多种易混淆的干扰疾病假设（即发起因果伪问题查询）；紧接着，系统引入对比验证奖权机制（证据降权），通过动态计算各证据切片对目标假设与干扰假设的支撑度差异，对高价值的特异性证据予以保留，并大幅惩罚、剔除无临床鉴别力的共性噪声。最后，经过筛选的核心证

据链被送入辩论验证阶段。系统实例化立场对立的正反方智能体，围绕不同疾病假设下的逻辑一致性展开多轮交叉对抗质证，最终由法官智能体输出具备深度因果解释力的裁决。在本平台上，该推演全过程通过前端可视化面板进行动态呈现，提升了临床辅助决策的透明度与可信度, 如图5.7所示。

5.4 本章小结

本章详细阐述了一个整合了多种检索方法和搜索引擎的 AI 驱动医学辅助诊断平台（DeepMed Search）的设计与实现过程。针对当前医疗大模型应用中存在的信息信源单一、长尾复杂病理推理容易产生 “推理漂移”，以及多智能体框架缺乏工程化交互平台等核心痛点，本平台提出并落地了一套完整的全栈解决方案。

在整体架构设计上，系统采用了高可用、高并发的微服务分布式架构，结合 Next.js、PostgreSQL、Milvus 与 MinIO 混合存储，以及 BullMQ 异步队列，构建了稳健的底层技术底座。在此基础上，本章详细剖析了系统自下而上的核心功能模块：首先，通过知识库构建模块，系统实现了复杂医学多模态文档的保真解析与智能语义切分入库；其次，借助 RAG 系统模块和多源搜索引擎模块，平台打破了数据孤岛，实现了本地知识库、PubMed 权威学术库与广域互联网的集成；最后，系统工程化地实现了深度研究与多源检索诊断模块，通过引入对比验证降噪以及多智能体交叉对抗质证等高阶算法机制，将原本不可见的黑盒推理转化为结构化、白盒化、可溯源的端到端可视化工作流。

综上所述，本章所设计的系统不仅完成了从理论算法到工程化落地的跨越，更为真实临床决策与高难度的医学深层循证研究提供了一个透明、可信、高效的智能化基础设施平台。

第六章　总结与展望

6.1　总结

近年来，大语言模型在医疗问答领域展现出了巨大的潜力，但其固有的幻觉现象以及传统 RAG 系统存在的推理漂移问题，限制了大模型在严谨临床辅助决策中的应用。为此，本文聚焦于面向复杂医疗场景的检索增强生成方法优化及平台落地实现。通过融合因果知识图谱、多智能体协同框架以及多源搜索引擎，本文在提升医疗问答的逻辑严密性、证据可靠性及交互透明度方面开展了系统性的研究。本文的主要研究工作与核心贡献总结如下：

（1）针对传统医疗检索受限于表层语义匹配、无法实现逻辑对齐的问题，提出了一种基于因果路径约束与用户查询伪问题生成的多轮检索增强方法。

该方法打破了传统 RAG 依赖词法相似度的局限，将结构因果模型与医学先验逻辑模板引入非结构化文献的预处理中，生成具有明确逻辑指向的因果感知伪问题。系统通过挖掘外部知识图谱中的标准病理逻辑链，设计了自适应迭代检索机制。通过动态检测非结构化文本与图谱逻辑间的断层并触发重查询，本方法将传统 RAG 的单次静态召回升级为严密的动态循证补全过程。这一创新不仅从检索源头上阻断了逻辑不兼容的噪声，还提升了模型在复杂医学推理任务中的证据覆盖率与逻辑自洽性。

（2）针对多源医学证据间存在冲突与冗余、大模型难以甄别有效信息，容易产生幻觉的问题，提出了一种基于证据特异性假设重排和多智能体辩论的检索增强方法。

为了应对复杂临床表现背后的共性干扰因素，该方法模拟真实思考逻辑，主动构建竞争性候选假设以发散推理边界。通过引入证据语义解耦与特征级过滤机制，系统实现了对多维候选证据的细粒度打分，有效剥离了缺乏临床鉴别力的共性噪声。在此基础上，依托动态自适应路由模块，系统构建了立场对立的多智能体质证与共识裁决工作流。交叉校验机制的引入，使系统具备了更好的自我纠错能力，在有效抑制幻觉生成的同时，显著提高了辅助诊断的准确率与特异性。

（3）针对核心算法理论向实际应用转化的壁垒，设计并开发了高度可视化的 AI

驱动医学深度研究平台（DeepMed Search）。

为解决现有医疗多智能体算法缺乏系统级整合与透明交互的缺陷，本文依托高并发微服务架构，独立开发了 DeepMed 平台。底层打通了本地向量知识库（Milvus）、PubMed 权威文献库及 Web 搜索，打破了医疗数据的信息孤岛。上层则集成了多模态文档解析引擎（MinerU）与前述的因果检索及多智能体质证核心组件。该平台成功将大模型不可见的黑盒推理，转化为包含智能路由调度、证据对比以及对抗辩论的端到端白盒化工作流，为真实的临床辅助研究提供了一套更可信的智能化基础研究系统。

（4）构建了多维度的严谨评估体系，在多个医疗权威基准数据集上验证了本文所提架构的有效性与鲁棒性。

本文不仅在理论与系统架构上进行了创新，还开展了详细可靠的实证研究。通过在公开的医疗问答数据集以及复杂的罕见病诊断数据集上进行大规模实验，并与现阶段主流的 Hippo RAG、GraphRAG 等前沿基线模型进行全面的对比测试与消融实验。实验结果充分证明，本文所提出的底层因果逻辑对齐与多源证据对抗质证范式，能够有效突破现有 RAG 系统的性能瓶颈，在各个关键评价指标上均取得了更有竞争力的表现，论证了该架构在复杂医疗决策场景下的实际应用潜力。

6.2 展望

尽管本文在医疗大模型检索增强方法的优化及系统工程化落地方面取得了一定的成果，但在面对真实世界中极其复杂多变的临床医疗需求时，本系统在长尾知识覆盖、推理效率以及隐私合规部署等方面仍存在局限性。针对现有系统的不足与医疗人工智能的发展趋势，未来的工作将主要围绕以下四个方向展开深入系统性的探索：

（1）基于 AutoResearch 理念的医疗智能体自动化迭代

当前 DeepMed 平台的检索链路与多智能体工作流仍主要依赖人工规则设计与经验调参。由于真实临床病例在疾病类型、检查项目和证据来源上具有较强异质性，固定检索策略和统一推理流程难以适配所有场景。因此，后续工作可参考 AutoResearch [①] 中自动化实验迭代的设计思路，探索受约束的策略优化机制。具体而言，系统不直接开放对核心诊断逻辑的任意修改权限，而是将可调范围限定在安全边界较

① https://github.com/karpathy/autoresearch

清晰的工程参数上，例如检索 Top-k、BM25 与向量检索融合权重、知识图谱查询权重、证据过滤阈值、多智能体质证轮次以及不同任务模板的选择策略。每次策略调整后，系统均需先在固定验证集上进行离线评估，并从诊断准确率、证据覆盖率等指标进行综合比较。只有当新策略在关键指标上稳定优于原策略，且未引入明显安全风险时，才将其作为候选配置保留。同时，平台可进一步构建任务类型感知的模块化工作流。通过这种受控的模块化优化方式，系统能够在保持安全边界和人工可审核性的前提下，提高对不同医疗场景的适配能力与可维护性。

（2）面向长尾罕见病与中西医跨域知识的循证融合与解释性增强

本文对长尾罕见病场景中的检索诱导推理漂移问题进行了初步探索，但在中医与西医交叉的复杂临床场景中，不同医学体系在概念表达、证据形式和推理路径上仍存在差异。中医诊疗侧重舌象、脉象、证候等整体性表征，而西医罕见病诊断更多依赖标准化表型和遗传学证据。因此，后续工作可参考 DeepRare 等罕见病智能体系统的思路，探索面向中西医交叉场景的异构表型映射与可追溯推理框架。系统可对中医病例中的非结构化信息进行结构化整理，并结合医学词表、人工规则和大模型辅助，将可对应的症状和体征映射至人类表型本体（HPO）或相关西医检查指标。在完成特征对齐后，系统可围绕关键表型检索外部医学证据，并形成一整套有指标依据的分层推理链。通过这种受控的工具协同方式，系统有望在保持证据可追溯和人工可审核的前提下，提高对中西医交叉疑难病例及罕见病场景的适配能力。

（3）基于全栈开源工具的私有化部署与效率优化

本文引入的因果图谱检索与多智能体质证机制虽然提升了系统的逻辑严密性，但也成倍增加了推理步骤与系统耗时。在真实的医院信息化环境中，不仅要求响应速度快，且存在隐私数据保密的安全红线。因此，未来的工作将侧重工程落地，利用成熟的开源工具链，将当前的系统原型重构为可离线运行的私有化方案。在底层模型推理层面，计划在本地服务器上使用 vLLM 等开源高性能推理引擎，挂载 Qwen 等优秀的开源大模型。在业务架构与统一部署层面，计划引入 Dify 等企业级的开源大模型应用编排框架来重构诊断工作流。最终，将大模型推理服务、编排框架、底层的图数据库与向量数据库以及前端可视化界面，通过 Docker Compose 进行全栈容器化封装。通过这种方式，实现整套 DeepMed 系统能够离线一键部署，打造一个自主可控、低延迟的实用化辅助诊疗平台。

6.3 未来挑战

（1）多源异构知识的自适应协同与融合机制

当前 DeepMed 已能够协同利用医学知识图谱与非结构化文献数据，但在复杂临床推理场景中，不同来源证据的有效融合仍存在挑战。结构化数据准确但灵活性不足，文本数据信息丰富但存在噪声与冲突，现有方法多依赖静态检索与启发式聚合。未来需要构建上下文感知的动态融合机制，使模型能够根据具体病例自适应评估证据重要性并分配权重，从而减少无关信息干扰并提升推理质量。

（2）医学领域下的人机交互意图规约

医疗问答中，用户输入往往存在表达不规范和信息缺失的问题，直接用于检索与推理易导致结果偏差。因此，在进入下游流程前，有必要对输入进行规范化处理。系统应提取关键医学信息（如症状、病程、用药史等），并在信息不足或风险较高时优先进行澄清，而非直接生成诊断结果。该过程有助于为后续检索与推理提供清晰语义边界，从而提升系统的稳定性与可解释性。此外，未来还需要进一步研究多智能体质证过程中的错误传播控制、裁决者偏见抑制以及跨模型、跨提示词条件下的结论稳定性评估。

（3）中医数据缺乏与长尾场景下的幻觉问题

在中医等长尾医疗场景中，大模型易产生看似合理但不准确的生成结果。一方面，当前缺乏高质量的中医结构化数据；另一方面，中医理论体系与现代语料分布存在差异，增加了建模难度。未来需加强中医知识库建设，并提升模型在低资源场景下的对齐能力，以缓解幻觉问题。

插图索引

表格索引

参考文献

[1] ZHAO W X, ZHOU K, LI J, et al. A Survey of Large Language Models[A/OL]. 2023. arXiv: 2303.18223. https://arxiv.org/abs/2303.18223.

[2] ZHANG Y, LI Y, CUI L, et al. Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models[J/OL]. Computational Linguistics, 2025, 51(4): 1373-1418. https://aclanthology.org/2025.cl-4.9/. DOI: 10.1162/coli.a.16.

[3] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C/OL]//NIPS ’20: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada: Curran Associates Inc., 2020. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[4] 袁乐, 刘绍华, 王禹, 等. 大语言模型检索增强生成优化技术研究综述[J]. 计算机学报, 2026, 49(2): 383-422.

[5] 毕枫林, 张豈明, 张嘉睿, 等. 基于滑动窗口策略的大语言模型检索增强生成系统[J/OL]. 计算机研究与发展, 2025, 62(7): 1597-1610. DOI: 10.7544/issn1000-1239.202440411.

[6] ZAKKA C, SHAD R, CHAURASIA A, et al. Almanac —Retrieval-Augmented Language Models for Clinical Medicine[J/OL]. Nejm ai, 2024, 1(2): AIoa2300068. https://doi.org/10.1056/AIoa2300068.

[7] SINGH A, EHTESHAM A, KUMAR S, et al. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG[A/OL]. 2025. arXiv: 2501.09136. https://arxiv.org/abs/2501.09136.

[8] LIU S, CHEN T, LIANG Z, et al. LLM Collaboration With Multi-Agent Reinforcement Learning[A/OL]. 2025. arXiv: 2508.04652. https://arxiv.org/abs/2508.04652.

[9] SONG Z, YAN B, LIU Y, et al. Injecting Domain-Specific Knowledge into Large Language Models: A Comprehensive Survey[C/OL]//Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics, 2025: 25297-25311. https://aclanthology.org/2025.findings-emnlp.1379/. DOI: 10.18653/v1/2025.findings-emnlp.1379.

[10] KARPUKHIN V, OGUZ B, MIN S, et al. Dense Passage Retrieval for Open-Domain Question Answering[C/OL]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020: 6769-6781. https://aclanthology.org/2020.emnlp-main.550/. DOI: 10.18653/v1/2020.emnlp-main.550.

[11] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 3982-3992. https://aclanthology.org/D19-1410/. DOI: 10.18653/v1/D19-1410.

[12] SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge [J/OL]. Nature, 2023, 620(7972): 172-180. https://doi.org/10.1038/s41586-023-06291-2.

[13] ASAI A, WU Z, WANG Y, et al. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection[C/OL]//The Twelfth International Conference on Learning Representations. 2024. https://openreview.net/forum?id=hSyW5go0v8.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc., 2017: 6000-6010.

[15] OpenAI, ACHIAM J, ADLER S, et al. GPT-4 Technical Report[A/OL]. 2023. arXiv: 2303.08774. https://arxiv.org/abs/2303.08774.

[16] Gemini Team, ANIL R, BORGEAUD S, et al. Gemini: A Family of Highly Capable Multimodal Models[A/OL]. 2023. arXiv: 2312.11805. https://arxiv.org/abs/2312.11805.

[17] THIRUNAVUKARASU A J, TING D S J, ELANGOVAN K, et al. Large language models in medicine[J/OL]. Nature medicine, 2023, 29(8): 1930-1940. https://doi.org/10.1038/s41591-023-02448-8.

[18] MOOR M, BANERJEE O, ABAD Z S H, et al. Foundation models for generalist medical artificial intelligence[J/OL]. Nature, 2023, 616(7956): 259-265. https://doi.org/10.1038/s41586-023-05881-4.

[19] ALSENTZER E, MURPHY J, BOAG W, et al. Publicly Available Clinical BERT Embeddings[C/OL]//Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019: 72-78. https://www.aclweb.org/anthology/W19-1909. DOI: 10.18653/v1/W19-1909.

[20] JOHNSON A E W, POLLARD T J, SHEN L, et al. MIMIC-III, a freely accessible critical care database[J/OL]. Scientific data, 2016, 3: 160035. https://doi.org/10.1038/sdata.2016.35.

[21] LIU F, ZHOU H, GU B, et al. Application of large language models in medicine [J/OL]. Nature Reviews Bioengineering, 2025, 3(6): 445-464. https://doi.org/10.1038/s44222-025-00279-5.

[22] 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/OL]. Applied Sciences, 2021, 11(14): 6421. https://doi.org/10.3390/app11146421.

[23] NORI H, LEE Y T, ZHANG S, et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine[A/OL]. 2023. arXiv: 2311.16452. https://arxiv.org/abs/2311.16452.

[24] CHEN Z, CANO A H, ROMANOU A, et al. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models[A/OL]. 2023. arXiv: 2311.16079. https://arxiv.org/abs/2311.16079.

[25] QIU P, WU C, ZHANG X, et al. Towards building multilingual language model for medicine[J/OL]. Nature Communications, 2024, 15(1): 8384. https://doi.org/10.1038/s41467-024-52417-z.

[26] WANG H, LIU C, XI N, et al. HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge[A/OL]. 2023. arXiv: 2304.06975. https://arxiv.org/abs/2304.06975.

[27] LI Y, LI Z, ZHANG K, et al. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge[J/OL]. Cureus, 2023, 15(6): e40895. https://doi.org/10.7759/cureus.40895.

[28] HE K, MAO R, LIN Q, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics[J/OL]. Information Fusion, 2025, 118: 102963. https://doi.org/10.1016/j.inffus.2025.102963.

[29] ALANSARI A, LUQMAN H. Large Language Models Hallucination: A Comprehensive Survey[A/OL]. 2025. arXiv: 2510.06265. https://arxiv.org/abs/2510.06265.

[30] FAREED M, FATIMA M, UDDIN J, et al. A systematic review of ethical considerations of large language models in healthcare and medicine[J/OL]. Frontiers in digital health, 2025, 7: 1653631. https://doi.org/10.3389/fdgth.2025.1653631.

[31] SCHWARTZ I S, LINK K E, DANESHJOU R, et al. Black Box Warning: Large Language Models and the Future of Infectious Diseases Consultation[J/OL]. Clinical infectious diseases, 2024, 78(4): 860-866. https://doi.org/10.1093/cid/ciad633.

[32] XIAO S, LIU Z, ZHANG P, et al. C-Pack: Packed Resources for General Chinese Embeddings[C/OL]//Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval. 2024: 641-649. https://doi.org/10.1145/3626772.3657878.

[33] NEELAKANTAN A, XU T, PURI R, et al. Text and Code Embeddings by Contrastive Pre-Training[A/OL]. 2022. arXiv: 2201.10005. https://arxiv.org/abs/2201.10005.

[34] JOHNSON J, DOUZE M, JÉGOU H. Billion-Scale Similarity Search with GPUs [J/OL]. IEEE Transactions on Big Data, 2021, 7(3): 535-547. https://doi.org/10.1109/TBDATA.2019.2921572.

[35] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval [J/OL]. Information processing & management, 1988, 24(5): 513-523. https://doi.org/10.1016/0306-4573(88)90021-0.

[36] ROBERTSON S, ZARAGOZA H. The Probabilistic Relevance Framework: BM25 and Beyond[J/OL]. Foundations and Trends in Information Retrieval, 2009, 3(4): 333-389. https://doi.org/10.1561/1500000019.

[37] LIU N F, LIN K, HEWITT J, et al. Lost in the Middle: How Language Models Use Long Contexts[J/OL]. Transactions of the Association for Computational Linguistics, 2024, 12: 157-173. https://aclanthology.org/2024.tacl-1.9/. DOI: 10.1162/tacl_a_00638.

[38] GAO Y, XIONG Y, GAO X, et al. Retrieval-Augmented Generation for Large Language Models: A Survey[A/OL]. 2024. arXiv: 2312.10997. https://arxiv.org/abs/2312.10997.

[39] LEI Y, CAO Y, ZHOU T, et al. Corpus-Steered Query Expansion with Large Language Models[C/OL]//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers). St. Julian's, Malta: Association for Computational Linguistics, 2024: 393-401. https://aclanthology.org/2024.eacl-short.34/. DOI: 10.18653/v1/2024.eacl-short.34.

[40] GAO L, MA X, LIN J, et al. Precise Zero-Shot Dense Retrieval without Relevance Labels[C/OL]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics, 2023: 1762-1777. https://aclanthology.org/2023.acl-long.99/. DOI: 10.18653/v1/2023.acl-long.99.

[41] RUTHVEN I, LALMAS M. A survey on the use of relevance feedback for information access systems[J/OL]. The Knowledge Engineering Review, 2003, 18(2): 95-145. https://doi.org/10.1017/S0269888903000638.

[42] SCHICK T, DWIVEDI-YU J, DESSÍ R, et al. Toolformer: language models can teach themselves to use tools[C]//NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2023.

[43] JEONG S, BAEK J, CHO S, et al. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity[C/OL]// Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Mexico City, Mexico: Association for Computational Linguistics, 2024: 7036-7050. https://aclanthology.org/2024.naacl-long.389/. DOI: 10.18653/v1/2024.naacl-long.389.

[44] NOGUEIRA R, YANG W, CHO K, et al. Multi-Stage Document Ranking with BERT [A/OL]. 2019. arXiv: 1910.14424. https://arxiv.org/abs/1910.14424.

[45] JIANG Z, XU F F, GAO L, et al. Active Retrieval Augmented Generation[C/OL]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023: 7969-7992. https://aclanthology.org/2023.emnlp-main.495/. DOI: 10.18653/v1/2023.emnlp-main.495.

[46] FURNAS G W, LANDAUER T K, GOMEZ L M, et al. The vocabulary problem in human-system communication[J/OL]. Communications of the ACM, 1987, 30(11): 964-971. https://doi.org/10.1145/32206.32212.

[47] BODENREIDER O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology[J/OL]. Nucleic acids research, 2004, 32(suppl_1): D267-D270. https://doi.org/10.1093/nar/gkh061.

[48] ARONSON A R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program[C/OL]//Proceedings of the AMIA Symposium. 2001: 17-21. https://pmc.ncbi.nlm.nih.gov/articles/PMC2243666/.

[49] LEAMAN R, ISLAMAJ DOĞAN R, LU Z. DNorm: disease name normalization with pairwise learning to rank[J/OL]. Bioinformatics, 2013, 29(22): 2909-2917. https://doi.org/10.1093/bioinformatics/btt474.

[50] LEAMAN R, GONZALEZ G. BANNER: an executable survey of advances in biomedical named entity recognition[M/OL]//Biocomputing 2008. World Scientific, 2008: 652-663. https://doi.org/10.1142/9789812776136_0062.

[51] DAVIS A P, WIEGERS T C, ROSENSTEIN M C, et al. MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database[J/OL]. Database, 2012, 2012: bar065. https://doi.org/10.1093/database/bar065.

[52] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J/OL]. Bioinformatics, 2020, 36(4): 1234-1240. https://doi.org/10.1093/bioinformatics/btz682.

[53] GU Y, TINN R, CHENG H, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing[J/OL]. ACM Transactions on Computing for Healthcare, 2021, 3(1): 1-23. https://doi.org/10.1145/3458754.

[54] JIN Q, KIM W, CHEN Q, et al. MedCPT: Contrastive Pre-Trained Transformers with Large-Scale PubMed Search Logs for Zero-Shot Biomedical Information Retrieval [J/OL]. Bioinformatics, 2023, 39(11): btad651. https://doi.org/10.1093/bioinformatics/btad651.

[55] MALKOV Y A, YASHUNIN D A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs[J/OL]. IEEE transactions on pattern analysis and machine intelligence, 2020, 42(4): 824-836. https://doi.org/10.1109/TPAMI.2018.2889473.

[56] ARYA S, MOUNT D M, NETANYAHU N S, et al. An optimal algorithm for approximate nearest neighbor searching fixed dimensions[J/OL]. Journal of the ACM (JACM), 1998, 45(6): 891-923. https://doi.org/10.1145/293347.293348.

[57] 曹书林, 史佳欣, 侯磊, 等. 知识库问答研究进展与展望[J]. 计算机学报, 2023, 46 (3): 512-539.

[58] 何鹏, 周刚, 陈静, 等. 类型增强的时态知识图谱表示学习模型[J/OL]. 计算机研究与发展, 2023, 60(4): 916-929. DOI: 10.7544/issn1000-1239.202111246.

[59] YASUNAGA M, REN H, BOSSELUT A, et al. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering[C/OL]//Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies. 2021: 535-546. https://aclanthology.org/2021.naacl-main.45/. DOI: 10.18653/v1/2021.naacl-main.45.

[60] VELIčKOVIć P, CUCURULL G, CASANOVA A, et al. Graph Attention Networks [C/OL]//International Conference on Learning Representations. 2018. https://arxiv.org/abs/1710.10903.

[61] EDGE D, TRINH H, CHENG N, et al. From Local to Global: A Graph RAG Approach to Query-Focused Summarization[A/OL]. 2024. arXiv: 2404.16130. https://arxiv.org/abs/2404.16130.

[62] TRAAG V A, WALTMAN L, VAN ECK N J. From Louvain to Leiden: guaranteeing well-connected communities[J/OL]. Scientific reports, 2019, 9(1): 5233. https://doi.org/10.1038/s41598-019-41695-z.

[63] PONS RECASENS G, BILALLI B, QUERALT A. Knowledge Graphs for Enhancing Large Language Models in Entity Disambiguation[C/OL]//International Semantic Web Conference. Springer, 2024: 162-179. https://doi.org/10.1007/978-3-031-77844-5_9.

[64] WU J, ZHU J, QI Y, et al. Medical Graph RAG: Evidence-based Medical Large Language Model via Graph Retrieval-Augmented Generation[C/OL]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 28443-28467. https://aclanthology.org/2025.acl-long.1381/. DOI: 10.18653/v1/2025.acl-long.1381.

[65] WEI J, TAY Y, BOMMASANI R, et al. Emergent Abilities of Large Language Models [J/OL]. Transactions on Machine Learning Research, 2022. https://openreview.net/forum?id=yzkSU5zdwD.

[66] JI Z, LEE N, FRIESKE R, et al. Survey of Hallucination in Natural Language Generation[J/OL]. ACM computing surveys, 2023, 55(12): 1-38. https://doi.org/10.1145/3571730.

[67] MINSKY M. Society of mind[M]. New York: Simon and Schuster, 1986.

[68] QIAN C, LIU W, LIU H, et al. ChatDev: Communicative Agents for Software Development[C/OL]//Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers). 2024: 15174-15186. https://aclanthology.org/2024.acl-long.810/. DOI: 10.18653/v1/2024.acl-long.810.

[69] WU Q, BANSAL G, ZHANG J, et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation[C/OL]//First conference on language modeling. 2024. https://openreview.net/forum?id=BAakY1hNKS.

[70] 秦龙, 武万森, 刘丹, 等. 基于大语言模型的复杂任务自主规划处理框架[J/OL]. 自动化学报, 2024, 50(4): 862-872. DOI: 10.16383/j.aas.c240088.

[71] 王路桥, 周洋涛, 李青山, 等. 基于大语言模型的多智能体协作代码评审人推荐[J/OL]. 软件学报, 2025, 36(6): 2558-2575. DOI: 10.13328/j.cnki.jos.007326.

[72] TABERNA M, GIL MONCAYO F, JANÉ-SALAS E, et al. The Multidisciplinary Team (MDT) Approach and Quality of Care[J/OL]. Frontiers in oncology, 2020, 10: 85. https://doi.org/10.3389/fonc.2020.00085.

[73] TANG X, ZOU A, ZHANG Z, et al. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning[C/OL]//Findings of the Association for Computational Linguistics: ACL 2024. 2024: 599-621. https://aclanthology.org/2024.findings-acl.33/. DOI: 10.18653/v1/2024.findings-acl.33.

[74] ZHAO W, WU C, FAN Y, et al. An agentic system for rare disease diagnosis with traceable reasoning[J/OL]. Nature, 2026, 651(8106): 775-784. https://doi.org/10.1038/s41586-025-10097-9.

[75] 杨新新, 刘真, 卢思博, 等. 基于因果推断的推荐系统去偏研究综述[J]. 计算机学报, 2024, 47(10): 2307-2332.

[76] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//NIPS '22: Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2022.

[77] XIONG G, JIN Q, LU Z, et al. Benchmarking Retrieval-Augmented Generation for Medicine[C/OL]//Findings of the Association for Computational Linguistics: ACL 2024. 2024: 6233-6251. https://aclanthology.org/2024.findings-acl.372/. DOI: 10.18653/v1/2024.findings-acl.372.

[78] XIANG Z, WU C, ZHANG Q, et al. When to use Graphs in RAG: A Comprehensive Analysis for Graph Retrieval-Augmented Generation[C/OL]//The Fourteenth International Conference on Learning Representations. 2026. https://openreview.net/forum?id=i9q9xDMjG7.

[79] SARTHI P, ABDULLAH S, TULI A, et al. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval[C/OL]//The Twelfth International Conference on Learning Representations. 2024. https://openreview.net/forum?id=GN921JHCRw.

[80] GUO Z, XIA L, YU Y, et al. LightRAG: Simple and Fast Retrieval-Augmented Generation[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Findings of the Association for Computational Linguistics: EMNLP 2025. Association for Computational Linguistics, 2025: 10746-10761. https://aclanthology.org/2025.findings-emnlp.568/. DOI: 10.18653/v1/2025.findings-emnlp.568.

[81] GUTIÉRREZ B J, SHU Y, QI W, et al. From RAG to Memory: Non-Parametric Continual Learning for Large Language Models[C/OL]//Proceedings of Machine Learning Research: Vol. 267 Proceedings of the 42nd International Conference on Machine Learning. PMLR, 2025: 21497-21515. https://proceedings.mlr.press/v267/gutierrez25a.html.

[82] WANG Z, WANG Z, LE L, et al. Speculative RAG: Enhancing Retrieval Augmented Generation through Drafting[C/OL]//International Conference on Learning Representations. 2025. https://openreview.net/forum?id=xgQfWbV6Ey.

[83] QIN H, WEI C, CHEN Y, et al. Counterfactual Reasoning for Retrieval-Augmented Generation[C/OL]//The Fourteenth International Conference on Learning Representations. 2026. https://openreview.net/forum?id=9U51rOnGko.

[84] SINGHAL K, TU T, GOTTWEIS J, et al. Toward expert-level medical question answering with large language models[J/OL]. Nature medicine, 2025, 31(3): 943-950. https://doi.org/10.1038/s41591-024-03423-7.

攻读专业硕士学位期间取得的研究成果

一、论文

[1] **Maolin Liu**, Qi Yang and Hao Wang. QAlign-RAG: Causal-Aware Pseudo-Question Generation and Debate-Driven Verification for Medical RAG. In Proceedings of the 19th International Conference on Knowledge Science, Engineering and Management(KSEM 2026).（第一作者，CCF-C，第三章）

[2] **Maolin Liu**, Fanyu Xu, Ruoqing Xu, Jiahang Zhang, Hao Wang and Rui Wang. DeepMed Search: An Open-Source Agentic Platform for Medical Deep Research with Introspective Verification. The 35th International Joint Conference on Artificial Intelligence, Demo (第一作者，CCF-B Demo，第五章)

[3] Lingyi Meng, **Maolin Liu**, Hao Wang, Yilan Cheng, Qi Yang and Idlkaid Mohammed. Building from scratch: a multi-agent framework with human-in-the-loop for multilingual legal terminology mapping[J]. Artificial Intelligence and Law, 2025: 1-40.（第二作者，SCI 二区）

[4] Bin Chen, **Maolin Liu** and Hao Wang. MoE-TextDiffuser: Fine-grained control of font and color in visual text rendering[J]. Neurocomputing, 2026: 133131.（第二作者，SCI 二区）

二、专利

[1] 王昊, **刘茂林**, 杨旗. 一种基于多智能体协同与 RAG 技术的中医文献翻译系统及方法: 2025115555932[P].2025-10-29（专利受理中）

[2] 王昊, **刘茂林**, 杨旗. 面向药品说明书领域的视觉富文档信息抽取方法和装置: 2025115568398[P].2025-10-29（专利受理中）

[3] 王昊, **刘茂林**, 杨旗, IDLKAID MOHAMMED, 熊惠文. 一种融合阅读顺序与多模态的海运单信息抽取方法和装置: 2025115571757[P].2025-10-29（专利受理中）

三、软著

[1] 王昊; 杨旗; **刘茂林**; 熊惠文; 中文法律翻译多语种语料平台 V1.0, 2026SR0281940, 原始取得, 全部权利, 2025-12-01 (软件著作权)

[2] 王昊; **刘茂林**; 杨旗; DeepMed 医学文献检索与深度研究系统, 2025SR2356343, 原始取得, 全部权利, 2025-12-05 (软件著作权，对应第五章）

致　谢

转眼间，七年的上大时光即将画上句号。回首这段漫长而又充实的求学之路，我的每一步成长都离不开身边师长、同门与家人的支持与包容。临别之际，书不尽言，唯有将满腔感激凝于笔端。

首先，我要向我的导师王老师致以最深切的谢意。在我的整个硕士阶段，是王老师为我拨开科研路上的重重迷雾。无论是在前期小论文的选题构思和修改打磨中，还是在大论文的研究中，王老师都给予了我极其宝贵的指导与毫无保留的帮助。他严谨求实的治学态度、敏锐的学术洞察力以及在学术上精益求精的精神，让我受益匪浅。回顾这三年，王老师传授给我的不仅是做科研的方法，更是严谨踏实的做事态度。他让我学会了如何褪去浮躁、沉下心来解决问题，这将成为我未来人生路上宝贵的财富。

其次，我要感谢课题组的师兄师弟和室友。感谢你们三年来给予我的无私帮助、支持与鼓励。无论是在科研中遇到技术难题时的倾囊相授，亦或是日常生活中的默契相伴，都让枯燥的科研生活变得温暖而生动。这段并肩作战、共同攻坚克难的岁月，将是我人生中极为珍贵的财富。

我还要特别感谢我的家人。是你们无私的支持，为我筑起了一个坚固而温暖的避风港。感谢你们始终如一的理解与包容，替我扛下了生活的重责，让我能够毫无后顾之忧地驻足于象牙塔内，全身心地投入到学习与科研之中。你们默默的付出与期盼，是我面对困难时不断前行的最大动力。

最后，感谢上海大学计算机工程学院提供的平台，七年光阴，上大早已不仅仅是一所传道受业的学府，更是我心底如家一般温暖、熟悉的存在。愿母校与师友皆顺遂无虞，也愿自己在未来的道路上乘风破浪。

刘茂林

上海大学

2026 年 4 月 2 日