将知识融入诊疗：知识图谱增强的医疗随访问题生成

Liwen Sun[1*]，**Xiang Yu**[2],**MingTan**[2],**Zhuohao Chen**[2],**Anqi Cheng**[2], **Ashutosh Joshi**[2],**Chenyan Xiong**[11] 卡内基梅隆大学 [2]
Amazon HealthAI

arXiv:2603.01252v1 [cs.CL] 1 Mar 2026

摘要

临床诊断耗时较长，需要患者与医疗专业人员进行深入交流。尽管大语言模型（LLM）可以减轻诊前工作负担，但其领域知识有限，因而难以有效生成医疗问题。我们提出一种结合主动上下文学习的知识图谱增强型大语言模型，用于生成相关且重要的后续问题，称为 KG-Followup，可作为诊前评估的关键模块。结构化的医学领域知识图谱能够无缝补充专业领域知识，供大语言模型进行推理。实验表明，在相关基准测试的召回率方面，KG-Followup 比最先进的方法高出 5% - 8%。

1 引言

有效的诊断表现不仅取决于对患者明确提供的信息进行推理，也取决于能否获取恰当的信息。相当一部分诊断错误源于未能全面收集信息和采集病史（Tu 等，2025年；Balogh 等，2015年；Singh 等，2013年；Graber 等，2013年；Ely 等，2011年）。提出恰当的后续问题有助于减轻医生的工作负担——减少转诊、重复就诊和纠正性处置——同时降低医疗服务低效和成本，最终提升患者满意度，尤其是在时间和资源受限的情况下（Abimanyi-Ochom 等，2019年；Singh 和 Sittig，2020年；Schiff 等，2009；Trowbridge 等，2013）。

在现代临床实践中，生成随访问题仍依赖人工，既耗时又限制了诊断效率。本研究利用大语言模型自动化这一流程，使其能够模拟医生在医疗就诊过程中的问诊策略。

[*]在亚马逊实习期间完成的工作。

然而，现有大语言模型往往无法识别不同症状之间的信息缺口（Xiong 等人，2024年；Li 等人，2024a，2025年；Gatto 等人，2025年）。我们提出 KG-Followup，这是一种知识图谱增强框架，利用结构化医学概念引导大语言模型生成具有临床相关性且全面的随访问题（Chandak 等人，2023年；Wu 等人，2024年）。该框架整合了电子健康记录指导的概念检索、鉴别诊断引导的推理以及知识图谱指导的主动上下文学习，从而高效生成基于上下文的问题，减轻临床医生收集信息的负担并提高诊断效率。

为实现对多样化真实临床场景的全面评估，我们推出了 ClinicalInquiryBench，这是一项专门用于评估人工智能系统生成临床上恰当的后续问题能力的全新基准测试。ClinicalInquiryBench 通过对公开可用的医生标注临床对话进行系统性转换而构建（Arora 等，2025年）。

实验结果表明，KG-Followup 在 ClinicalInquiryBench 和 FollowupBench 上优于最先进的方法，在问题数量适当的情况下，召回率分别达到 70% 和 80%。在少样本设置下，知识图谱引导的主动上下文学习比随机上下文学习更有效。尽管向大语言模型提示更多问题能够提升性能，KG-Followup 只需少得多的问题便能取得可比的结果。案例研究表明，KG-Followup 能够检索多样化的症状概念，从而实现全面的随访生成。我们的主要贡献总结如下：

- 我们提出 KG-Followup，这是一种知识图谱增强框架，可引导大语言模型生成临床相关且全面的后续问题。

- 我们整理构建了 ClinicalInquiryBench，这是一个基准测试

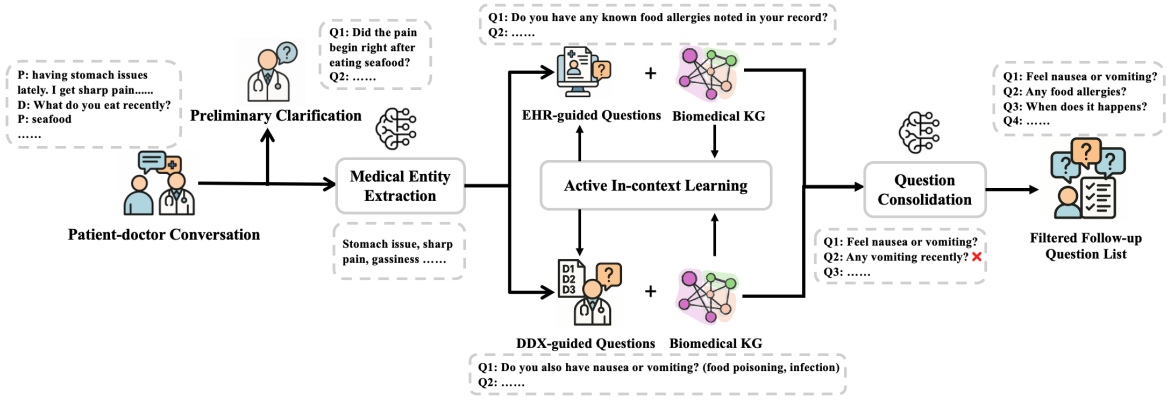


图 1：我们的方法根据患者消息和电子健康记录生成初步问题，通过大语言模型提取医学实体，应用 (i) 电子健康记录引导的关联概念检索、(ii) 鉴别诊断引导的推理路径搜索，以及 (iii) 知识图谱指导的主动上下文学习，最后筛选结果，得到数量可控的后续问题。

涵盖多样化临床场景的基准测试，用于直接评估大语言模型提出诊断问题的能力。

- 我们展示了 KG-Followup 实现了最先进的性能，在两个基准测试上以数量可控的问题达到 70% 和 80% 的召回率。

2 方法

2.1 任务定义

给定一段医患对话 C 和一个语言模型 F，

$$F_{\text{generate}}(C) = \hat{\mathcal{Q}} = \{\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_n\}. \quad (1)$$

$\hat{}$ 目标是生成一个集合 $\mathcal{Q}$，其中每个 $\hat{}q_i$ 都是对话的后续问题。由于我们关注的是静态信息收集，所有后续问题都必须根据对话历史一次性生成。

2.2 初步澄清

给定对话 C，我们首先提示大语言模型F生成一组初步的后续问题 Q_{pre}。此步骤模拟初次医生问诊：大语言模型仅依靠内部知识，在纳入任何外部临床知识之前，就患者的状况提出澄清问题。

2.3 通过知识图谱症状链接生成电子健康记录引导的问题

为了更好地识别患者消息中提及的关键症状和疾病，我们首先提示大语言模型提取关键医学实体，作为电子健康记录（EHR）的替代表示：

$$E = F_{\text{extract}}(C) = \{e_1, e_2, \ldots, e_n\}, \quad (2)$$

其中 $E = \{e_1, \ldots, e_n\}$ 是提取出的临床实体。这些实体构成医学知识图谱中的起始节点，使我们能够更有针对性、更充分地推理患者的病情。我们通过字符串匹配和嵌入相似度将提取出的实体链接到知识图谱节点，按指定深度执行广度优先搜索以构建实体特定子图，并对这些子图取交集，从而找出共同的临床相关概念。

由于并非所有相交知识图谱概念都与患者病例相关，我们使用大语言模型按相关性对相交知识图谱概念进行排序，并选出排名前 k_1 个关联实体，然后根据患者对话 C 和排序后的概念生成后续问题：

$$\hat{E} = F_{\text{rank-entity}}(E, C), \quad (3)$$

$$\hat{Q}_{\text{ehr-kg}} = F_{\text{generate}}(C, \hat{E}). \quad (4)$$

这借助知识图谱衍生症状丰富问题生成，从而提出更全面、具有临床相关性的问询。

2.4 经由知识图谱推理诊断的鉴别诊断引导问题

为了模拟医生通过策略性询问逐步缩小诊断范围的推理过程，我们首先根据患者当前状况，利用大语言模型进行鉴别诊断（DDX），得出一组可能诊断。

表 1：主要结果。/ 表示生成问题的平均数量。主动上下文学习是指采用知识图谱信息辅助的困难示例进行上下文学习。

方法	ClinicalInquiryBench		FollowupBench
	Dev	Test	
GPT-4o	0.61 / 20	0.61 / 20	0.67 / 40
MedGemma-27B	0.67 / 20	0.67 / 20	0.74 / 40
骨干：Claude Haiku			
零样本 U	0.52 / 7	0.52 / 7	0.54 / 10
零样本 k	0.63 / 20	0.63 / 20	0.72 / 40
FollowupQ	0.63 / 20	0.65 / 20	0.73 / 40
KG-Followup	0.70 / 20	0.70 / 20	0.77 / 40
+ 随机上下文学习	0.72 / 20	0.72 / 20	0.77 / 40
+ 主动上下文学习	**0.74 / 20**	**0.73 / 20**	**0.78 / 40**
主干模型：Claude Sonnet			
零样本 U	0.59 / 10	0.59 / 10	0.53 / 10
零样本-k	0.64 / 20	0.64 / 20	0.72 / 40
FollowupQ	0.65 / 20	0.66 / 20	0.74 / 40
KG-Followup	0.72 / 20	0.72 / 20	0.81 / 40
+ 随机上下文学习	0.71 / 20	0.72 / 20	0.81 / 40
+ 主动上下文学习	**0.75 / 20**	**0.77 / 20**	**0.82 / 40**

然后，我们提出后续问题，旨在排除最坏和最好的诊断情况：

$$D = F_{\text{best}}(C) \cup F_{\text{worst}}(C), \tag{5}$$

$$\hat{Q}_{d_i} = F_{\text{eliminate}}(C, d_i), \quad d_i \in D \tag{6}$$

$$\hat{Q}_{\text{ddx}} = \hat{Q}_{d_1} ... \cup ... \hat{Q}_{d_m}, \tag{7}$$

其中 $D = \{d_1, ..., d_m\}$ 为所有可能的诊断。最终问题集通过对所有候选诊断的针对性问题取并集而形成，以排除不同的可能性。除大语言模型基于鉴别诊断的问题生成所依赖的内部知识外，我们还整合来自知识图谱的结构化医学知识，通过推理路径将每个电子健康记录实体与潜在诊断联系起来。这些路径上的关键中间节点提供了额外的上下文和医学依据。具体而言，对于每个源实体–目标诊断对（e_i, d_j），我们在知识图谱中采样 k_2条最短推理路径$P_{i,j}$。由于可能存在大量长度相同的路径，我们随后使用大语言模型根据患者的上下文选择唯一最相关的路径$P_{i,j}$，并剪除无关的知识图谱推理路径：

$$P_{i,j} = \texttt{shortest-path}(e_i, d_j, G, k_2), \tag{8}$$

$$\hat{P}_{i,j} = F_{\text{rank-path}}(P_{i,j}, C), \tag{9}$$

$$\hat{P} = \hat{P}_{1,1} ... \cup ... \hat{P}_{n,m}, \tag{10}$$

ˆ 其中，完整路径集 P 是从所有源实体–目标诊断对遍历得到的推理路径中汇总而成。我们使用选定的路径及其关键

表 2：不同生成信号的消融研究，使用 Claude Haiku 在 FollowupBench 上进行评估。

方法	召回率 / 数量
模块：电子健康记录引导生成	
基于检索到的知识图谱三元组的推理依据	0.72 / 25
检索到的相似知识图谱概念	0.71 / 25
遍历子图中的交集概念	**0.72 / 26**
模块：鉴别诊断引导的生成	
鉴别诊断排除问题	0.72 / 26
+知识图谱推理路径	0.75 / 34
模块：问题整合	
KG-Followup（无整合）	**0.79/51**
+ 聚类合并	0.77/40
+ 大语言模型选择	0.71/40

中间节点作为支撑知识来生成后续问题：

$$\hat{Q}_{\text{ddx-kg}} = F_{\text{generate}}(C, \hat{P}). \tag{11}$$

这些知识图谱推理路径生成的后续问题既有临床依据，又符合患者的具体情境。此外，它们还可作为仅基于鉴别诊断的后续问题的补充。

2.5 基于知识图谱辅助的困难案例的上下文学习

知识图谱不仅有助于识别需要进一步询问的其他症状，还有助于找出具有挑战性的患者查询。受主动学习启发，我们将这些困难案例视为上下文学习（ICL）示例，并将其与原始患者对话一同提供给大语言模型，以指导后续问题生成。具体而言，当患者消息缺少源症状，或无法映射到知识图谱实体——导致知识图谱遍历不可行时——我们将其视为困难查询，并将其作为上下文学习案例纳入：$T = \{(C_1, Q_1) ... ... (C_t, Q_t)\}$。最终的问题集构建如下：

$$\hat{Q} = \hat{Q}_{\text{pre}} \cup \hat{Q}_{\text{ehr-kg}} \cup \hat{Q}_{\text{ddx}} \cup \hat{Q}_{\text{ddx-kg}}, \tag{12}$$

其中每个生成模块均使用 T 进行上下文学习增强，以确保有效处理边缘案例，并覆盖明确症状问题之外的更广范围。

2.6 问题整合

为减少生成的后续问题之间的冗余，我们使用医学编码器对所有问题进行嵌入，并应用 K-均值聚类。随后，大语言模型通过合并重叠内容并删除重复项来优化多问题簇，从而生成简洁的问题集，便于高效开展患者互动。

表 3：在 ClinicalInquiryBench 上使用 Claude Sonnet 进行 ICL 示例选择的消融研究。

方法	Dev	Test
随机	0.71 / 20	0.72 / 20
知识图谱辅助的困难样本	**0.75 / 20**	**0.77 / 20**
监督式困难样本	**0.75 / 20**	0.75 / 20

3 实验设置

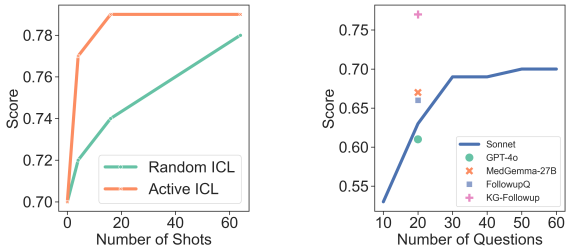
数据集。我们在两个数据集上进行评估：**Followup-Bench**(Gatto 等人，2025年)，这是一个由专家整理、包含 250 个实例的数据集（平均 ∼9 个问题，不含权重）；以及 **ClinicalInquiryBench**，该数据集源自 5,000 个 HealthBench（Arora 等人，2025年）案例，经筛选后保留 1,498 个寻求上下文信息的实例，并重新整理为带有问题权重的格式（平均 ∼5 个问题）。我们将 ClinicalInquiryBench 划分为 250 个开发集样例和 1,248 个测试集样例，并从开发集中抽取上下文学习样例。整理详情见附录 6.2。

评估指标。遵循 Gatto 等人（2025）的做法，我们采用加权召回率作为主要指标，通过大语言模型评审（Zheng 等人，2023）比较生成的问题 Q 与真实标注Q 。由于我们的目标是涵盖真实标注集合中的问题，精确率会不公平地惩罚临床上有效但未匹配的问题。详情见附录。基线方法。我们比较了以下方法：(1) 零样本 U：大语言模型不受限制地生成任意数量的随访问题，作为初步澄清的替代方案。(2) 零样本-k：生成固定数量的 k 个问题，以研究输出规模的影响。(3) Fol-lowupQ（Gatto 等人，2025）：包含澄清、电子健康记录推理和鉴别诊断模块的多智能体系统（不含知识图谱）。(4) 随机上下文学习：使用随机上下文示例，以展示基于知识图谱的选择所带来的益处。我们还评估了 GPT-4o（等人，2024）和 MedGemma-27B（Sellergren，2025）在零样本-k 设置下的表现。

4 评估结果

主要结果。 表 1 显示，知识图谱增强相较于不使用外部知识的最先进方法，能够持续提升性能，领先幅度超过 5%。融入知识图谱指导的主动上下文学习后，在两个基准测试上的性能还可进一步提升 3%。

有趣的是，在零样本-k 提示下，Sonnet 和 Haiku 的表现相当。然而，经过知识图谱增强后，Sonnet 的表现最高超出 Haiku 4 分，这表明 Sonnet 更有效利用外部知识。



(a) 上下文学习趋势

(b) 零样本-k 趋势

图 2：使用 Claude Sonnet 在 ClinicalInquiryBench 上分析上下文学习和零样本-k 趋势：图 (a) 展示不同示例数量下的性能；图 (b) 展示受控零样本-k的影响。

消融研究。 在表2中，我们测试了多种知识图谱增强策略（例如，基于三元组的 RAG、概念检索），发现它们之间的差异很小。鉴别诊断引导模块带来的提升大于电子健康记录引导模块。使用大语言模型精简问题会降低性能，而我们的基于聚类的合并方法则保留了质量，并高效地将问题池从 51 个缩减至 40 个，从而减轻了临床医生的工作负担。

对于主动上下文学习，我们还在表 = 0中测试了一种有监督的变体，该变体从开发集选择困难案例（召回率 3）。我们无监督的方法表现出色，证实了知识图谱引导的困难案例选择的有效性。如图 2a所示，随着上下文示例增多，性能不断提升；少样本主动上下文学习优于随机上下文学习，令牌效率也更高。

案例研究。 在第一个案例中（图 3），真实标注问题包含腹泻，而初步澄清仅涉及恶心和呕吐。知识图谱推理揭示了隐藏的症状关联（例如腹泻病、体重减轻），从而丰富了大语言模型生成的问题。在最后一个案例中，随机上下文学习会遗漏地理位置等非症状查询，而在困难示例的引导下，主动上下文学习能够生成边缘追问。

零样本k 分析。 图 2b 显示，随着大语言模型生成的问题增多，性能有所提升，但超过 k>40后便趋于饱和。相比之下，我们的方法只需更少的问题便能取得相近的效果，从而提高医患互动的效率。

5 结论

本研究通过 KG-Followup 这一知识

Input	Ground-truth	Preliminary Clarification	KG Linking Symptoms	KG Reasoning Diagnoses
EHR: Age: 35, Gender: Male, Active Problem: Obesity, Left upper quadrant abdominal pain…… **Message:** …… having some stomach issues lately. Sometimes I get these sharp pains and I'm really gassy. It's pretty annoying and I don't know what's causing it ……	1. Do you have any associated nausea?", 2. "Have you had any associated vomiting?", 3. "Have you had any associated diarrhea?" 4. "Have you experienced any unintentional weight loss?" 5. ……	1. "Do you experience any other symptoms along with the stomach pain, such as nausea, vomiting, or changes in bowel movements?", 2. Have you tried any over-the-counter medications to manage these symptoms?, 3. How long have you been experiencing these stomach issues?" 4. ……	1. Episodic abdominal pain 2. Pain 3. Abdominal distention 4. Nausea and vomiting 5. Elevated hepatic transaminase 6. Fatigue 7. Weight loss 8. Anorexia 9. Muscle spasm 10.Jaundice **Follow-up Questions from KG Linking Symptoms** 1. Have you experienced any unintentional weight loss along with these stomach issues? 2. Do the stomach pains cause you to feel nauseous or lead to vomiting? ……	… Flatulence->Abdominal symptom->Nausea and vomiting->gastric neuroendocrine neoplasm->gastric neoplasm Flatulence->Abdominal symptom->diarrheal disease->digestive system disease->hepatobiliary disease … **Follow-up Questions from KG Reasoning Diagnoses** 1. Are the stomach pains accompanied by any nausea or vomiting? 2. Are you experiencing frequent diarrhea or changes in your bowel movements? ……
Message: How often should I get my newborn vaccinated to follow the recommended schedule?	1. What country are you located in? 2. Has your baby received any vaccines yet, including those given at birth? 3. ……	1. Should my baby's vaccines be adjusted for health factors? 2. What side effects should I expect and how to manage them? 3. ……	**Random In-context Learning** 1. Has your baby been sick recently? 2. Has your baby received any vaccinations so far? 3. ……	**Active In-context Learning** 1. In which country are you located? 2. Has your baby received vaccinations so far? 3. ……

图3：**案例研究**。红色和绿色分别表示根据知识图谱检索到的症状信号和大语言模型内部知识生成的问题。

图谱增强框架解决随访问题生成任务，使大语言模型能够生成具有临床依据的澄清问题。为支持系统评估，我们整理了 ClinicalInquiryBench，这是一个大规模数据集，并标注了随访问题重要性。基于这一资源，我们引入了由知识图谱检索和知识图谱指导的主动上下文学习增强的电子健康记录引导型及鉴别诊断引导的生成模块。实验表明，在问题数量适当的情况下，KG-Followup 在 ClinicalInquiryBench 上的召回率比最先进的方法高出 8% 以上，在 FollowupBench 上高出 5%。

局限性

本研究的一个局限在于依赖预先构建的生物医学知识图谱来为问题提供依据。结构化知识的使用提高了临床相关性和生成的随访问题的准确性，但这也意味着知识图谱必须全面且及时更新。在实践中，许多真实世界的电子健康记录系统可能无法访问此类高质量知识图谱，或者可能依赖覆盖范围有限的领域专用本体。这种依赖可能会限制 KG-Followup 在低资源临床环境或快速发展的领域中的普适性，因为在这些领域中，知识图谱可能滞后于当前的医学认知。

此外，尽管 KG-Followup 在 ClinicalInquiryBench 和 FollowupBench 上表现出色，但这两个基准测试都是为离线评估而构建的，可能无法反映现实部署环境的全部复杂性，例如临床决策支持工具或电子健康记录集成应用。在实践中，部署的系统必须应对不完整或含噪数据、多样化的临床工作流程以及不断变化的患者情境。将这项工作扩展到支持更多样、动态的患者情境，将是迈向现实应用的重要一步。

伦理考量

本研究旨在通过知识图谱增强的语言模型提升临床问题生成能力，以支持临床医生决策。尽管该系统处理的是去标识化且公开可用的数据集，但语言模型和底层知识图谱中的潜在偏见仍会引发伦理问题，并可能影响生成问题的相关性或安全性。此外，在临床工作流程中过度依赖自动化建议也存在风险。为减轻这些问题，未来任何部署都应包含严格的人工监督、领域专门审查，并符合医疗人工智能的伦理标准，包括公平性、透明度和患者安全。

参考文献

J. Abimanyi-Ochom, S. Bohingamu Mudiyanselage, M. Catchpool，及另外 1 位作者。2019年。减少诊断错误的策略：系统综述。*BMCMed Inform Decis Mak*。

Rahul K.Arora、Jason Wei、Rebecca SoskinHicks、Preston Bowman、Joaquin Quiñonero-Candela、Foivos

Tsimpourlas, Michael Sharman, Meghan Shah, An- dreaVallone, Alex Beutel, Johannes Heidecke, 和 Karan Singhal。2025年。Healthbench：评估大语言模型，促进人类健康。预印本，arXiv:2505.08775。

Abhinand Balachandran。2024年。Medembed：医学领域专用嵌入模型。

Erin P. Balogh、Bryan T. Miller 和 John R. Ball。2015年。改善医疗保健中的诊断，第3章，《医疗保健中的诊断错误》概述。美国国家科学院出版社。

Payal Chandak、Kexin Huang 和 Marinka Zitnik。2023年。构建知识图谱以推动精准医学。《自然·科学数据》。

John W Ely、Jerome A Osheroff、Melisa L Cha mbliss、Mark H Ebell 和 Mark E Rosenbaum。2011年。临床问题的价值：确定初级保健中的研究优先事项。《英国医学杂志·质量与安全》，20(9)：787–792。

OpenAI 等。2024年。GPT-4o 系统卡。预印本，arXiv:2410.21276。

Joseph Gatto、Parker Seegmiller、Timothy Burdick、Inas S. Khayal、Sarah DeLozier 和 Sarah M. Preum。2025年。用于增强患者与医疗提供者对话的后续问题生成。预印本，arXiv:2503.17509。

Mark L Graber、Robert M Wachter 和 Christine K Cassel。2013年。内科学中的诊断错误。《英国医学杂志·质量与安全》，22（增刊2）：ii21–ii27。Pengcheng Jiang、Cao Xiao、Minhao Jiang、Parminder Bhatia、Taha Kass-Hout、Jimeng Sun 和 Jiawei Han。2025年。通过知识图谱社区检索实现推理增强的医疗保健预测。预印本，arXiv:2410.04585。Di Jin、Eileen Pan、Nassim Oufattole、Wei-Hung Weng、Hanyi Fang 和 Peter Szolovits。2020年。这位患者患有什么疾病？来自医学考试的大规模开放领域问答数据集。预印本，arXiv:2009.13081。Jerome P. Kassirer 和 Richard I. Kopelman。2010年。学习临床推理，第2版。Lippincott Williams & Wilkins。Shuyue Stella Li、Vidhisha Balachandran、Shangbin Feng、Jonathan S. Ilgen、Emma Pierson、Pang Wei Koh 和 Yulia Tsvetkov。2024a。MediQ：通过提问的大语言模型与可靠交互式临床推理基准测试。预印本，arXiv:2406.00922。Shuyue Stella Li、Vidhisha Balachandran、Shangbin Feng、Jonathan S. Ilgen、Emma Pierson、Pang Wei Koh 和 Yulia Tsvetkov。2024b。MediQ：通过提问的大语言模型与可靠交互式临床推理基准测试。载于神经信息处理系统进展，第37卷，第28858–28888页。Curran Associates, Inc.

Shuyue Stella Li、Jimin Mun、Faeze Brahman、Pedram Hosseini、Bryceton G. Thomas、Jessica M. Sin、Bing Ren、Jonathan S. Ilgen、Yulia Tsvetkov 和 Maarten Sap。2025年。Alfa：引导大语言模型提出优质问题——临床推理案例研究。预印本，arXiv:2502.14860。

Gordon D Schiff、Osman Hasan、Sujan Kim、Ramona Abrams、Karen Cosby、Bruce Lambert、Arthur S Elstein、Sarah Hasler、Ngoy Kabongo、Nada Krosnjar及其他1人。2009年。医学诊断错误：对583例医师报告错误的分析。《内科学文献》，169(20):1881–1887。

Henk G. Schmidt和Remy M. Rikers。2007年。专长如何在医学中发展：知识封装与疾病脚本形成。《医学教育》，41(12):1133–1139。

Andrew et al. Sellergren。2025年。 Medgemma技术报告。 预印本，arXiv:2507.05201。

Hardeep Singh、Andrew N D Meyer和Eric J Thomas。2013年。门诊医疗中诊断错误的发生频率：基于三项大型观察性研究的估计。《美国医学会杂志·内科学》，173(6):418–425。

Hardeep Singh 和 Dean F Sittig。2020年。推进医疗保健中诊断错误测量科学：Safer Dx 框架。*BMJ Quality&Safety*，29(10):874–880。

Robert L Trowbridge、Jason J Rencic 和 Steven J Durning。2013年。《临床推理教学：课堂与床旁教学的实用框架》。《学术医学》，88(2)：182–188。T. Tu、M. Schaekermann、A. Palepu 等。2025年。迈向对话式诊断人工智能。《自然》，第642卷，442–450页。Juncheng Wu、Wenlong Deng、Xingxuan Li、Sheng Liu、Taomian Mi、Yifan Peng、Ziyang Xu、Yi Liu、Hyunjin Cho、Chang-In Choi、Yihan Cao、Hui Ren、Xiang Li、Xiaoxiao Li 和 Yuyin Zhou。2025年。Medreason：通过知识图谱引导大语言模型生成事实性医学推理步骤。预印本，arXiv:2504.00993。Junde Wu、Jiayuan Zhu、Yunli Qi、Jingkun Chen、Min Xu、Filippo Menolascina 和 Vicente Grau。2024年。医学图谱检索增强生成：通过图检索增强生成实现安全的医学大语言模型。预印本，arXiv:2408.04187。Guangzhi Xiong、Qiao Jin、Zhiyong Lu 和 Aidong Zhang。2024年。医学领域检索增强生成的基准测试。预印本，arXiv:2402.13178。Lianmin Zheng、Wei-Lin Chiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric P. Xing、Hao Zhang、Joseph E. Gonzalez 和 Ion Stoica。2023年。使用 MT-Bench 和 Chatbot Arena 评判大语言模型评审。预印本，arXiv:2306.05685。

6 附录

6.1 相关工作

后续问题是临床推理的基石，能够帮助医生澄清含糊不清的细节、发现被忽略的症状，并指导诊断决策。（Kassirer 和 Kopelman，2010年；Schmidt 和 Rikers，2007）。MediQA（Li 等，2024b）将 MedQA（Jin 等，2020年）等传统单轮基准测试扩展为更贴近现实的交互式形式。与 MedQA 不同，MedQA 会预先提供完整的患者背景信息，而现实世界中的决策往往从有限的信息开始。在这类交互场景中，大语言模型经常未能提出澄清性问题，并倾向于过早下诊断或做出过度自信的诊断。

为解决这一问题，FollowupQ（Gatto 等人，2025年）整理了一个由医生撰写的真实后续问题数据集，并引入了一个多智能体系统，可根据患者消息和电子健康记录数据生成个性化的后续问题。其目标是帮助临床医生收集最相关、针对具体病例的信息，从而减少医疗对话中的歧义。然而，现有用于 AI 驱动的后续问题生成的数据集在规模和范围上仍然有限，而且未能体现不同问题的诊断重要性。例如，Followup-Bench（Gatto 等人，2025年）仅包含 250 个实例，且仅依靠静态电子健康记录信息来生成后续问题，并未对这些问题的相对诊断重要性进行编码。这限制了它反映真实临床推理以及不同后续问题细微价值的能力。

知识图谱（KG）增强生成进一步推动了医学问答的发展，使模型能够整合结构化的生物医学关系，从而实现更准确、基于上下文的诊断推理。例如，PrimeKG（Chandak 等人，2023年）提供了一个大规模生物医学知识图谱，支持知识增强推理。Med-GraphRAG（Wu 等人，2024年）利用大语言模型将检索增强生成（RAG）数据组织成图结构，展现出从长篇文档中提取整体性见解的巨大潜力。MedReason（Wu 等人，2025年）根据 GPT-4 得出的知识图谱推理路径构建监督微调数据，以强化事实性医学推理。KARE（Jiang 等人，2025年）将社区级知识图谱检索与大语言模型推理相结合，以提升医疗预测能力。

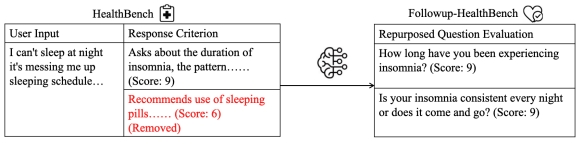


图4：一个精选实例示意图。

在这些研究方向的基础上，我们探讨知识图谱中的医学关联概念如何助力后续问题生成，使大语言模型能够提出有临床依据且具有诊断价值的问题。

6.2 基准测试

首先，我们概述源数据 Health-Bench，然后介绍 ClinicalInquiryBench 的整理流程。

6.2.1 初步

HealthBench（Arora 等人，2025年）是一个用于评估大语言模型在医疗保健领域表现与安全性的开放式基准测试。它包含患者与医疗专业人员之间的 5,000 段多轮对话，其回复依据 262 位医生制定的 48,562 条独特评分标准进行评估。这些标准涵盖多种医疗情境（如急症、临床数据转换、全球健康）和行为维度（如准确性、指令遵循、沟通），使 HealthBench 成为开发基准测试的综合基础。在此基础上，我们通过系统筛选并转换 HealthBench 中的相关实例，构建了 ClinicalInquiryBench。

ClinicalInquiryBench 明确聚焦于临床对话中后续提问的作用。与以往的基准测试相比，它规模更大，涵盖更广泛的后续提问意图类别，并标注了诊断重要性。这一设计使评估更加严谨，能够衡量大语言模型是否可以生成结合上下文且具有临床价值的后续问题，并反映医生澄清策略。ClinicalInquiryBench 保留了 HealthBench 的所有主题，确保临床场景覆盖范围不受损失。摘要统计数据见表4。

6.2.2 ClinicalInquiryBench 的构建

在此，我们将介绍如何将 Health- Benchinto 转化为随访问题基准测试，以及我们的问题形式化定义。

评分标准过滤。 我们首先选取仅包含英文的

表4：HealthBench 和 Clin-icalInquiryBench 的主题覆盖情况。

主题	精选实例（%）	原始实例（%）
示例总数	**1498** （100.0%）	**5000** (100.0%)
全球健康	317 (21.2%)	1,097 (21.9%)
模糊表达	348 (23.2%)	1,071 (21.4%)
沟通	99 (6.6%)	919 (18.4%)
寻求上下文	312 (20.8%)	594 (11.9%)
紧急转诊	132 (8.8%)	482 (9.6%)
健康数据任务	239 (16.0%)	477 (9.5%)
回答深度	51 (3.4%)	360 (7.2%)

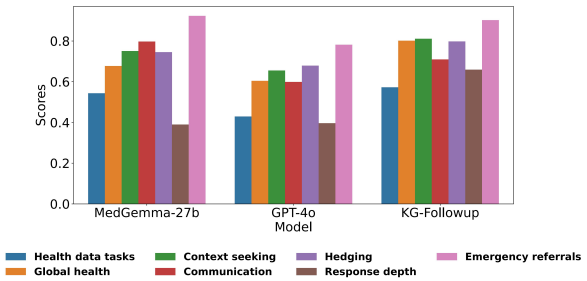


图5：各模型按主题划分的加权召回率。KG-Follo wup 框架采用 Claude Sonnet 作为基础模型实现。

HealthBench 实例。随后，我们使用 Claude 4 对这个子集进行筛选，以识别明确要求补充临床背景的评分标准，例如澄清症状、缩小鉴别诊断范围或获取缺失的患者病史。为确保数据集的相关性，我们会移除评分标准为空或与后续提问无关的条目。

随访重构。 然后，我们将保留的评分细则重构为后续问题生成的明确评估标准。我们使用 Claude 4 重新表述每项评分细则，突出信息收集目标，使基准测试能够评估模型是否能够识别信息缺口，并生成兼具临床相关性和诊断影响力的问题。重要的是，基准测试中的每个问题都根据其在 HealthBench 中的原始评分细则分数赋予权重，以体现其相对诊断重要性。这一转变将评分细则从一般的对话质量指标转变为聚焦诊断的提问标准，从而生成一个专门用于评估大语言模型模拟聚焦诊断询问能力的数据集。图中展示了一个示例4。

6.3 实现细节

对于 Q_{pre}，我们提示大语言模型为ClinicalInquiryBench生成20个问题，为FollowupBench生成40个问题。对于 Q_{ddx}，我们提示大语言模型生成2个最差情况和2个最佳情况的候选诊断，然后针对每个诊断生成2个后续问题。对于 $Q_{\text{ehr-kg}}$，我们使用排名前10的相交实体；对于 $Q_{\text{ddx-kg}}$，我们为每个源实体–目标诊断对采样30条推理路径以增强生成。对于 $Q_{\text{ehr-kg}}$ 或 $Q_{\text{ddx-kg}}$，不强制固定的问题数量。我们使用来自开发集的4个上下文学习示例。在整合之前，我们的框架在ClinicalInquiryBench上平均生成30个问题，在FollowupBench上平均生成50个问题。我们使用MedEmbed-large-v0.1（Balachandran，2024年）作为医学编码器，Claude Haiku和Sonnet作为大语言模型生成器与评判器，以及PrimeKG（Chandak等人，2023年）作为外部知识库。我们将在论文被录用后发布源代码。

6.4 评估细节

我们主要使用 Claude Sonnet 作为评估器，并采用列表式提示来评估生成的问题是否出现在真实问题集中。

6.5 主题分析

我们还在图中展示了模型在不同主题下的表现5。与其他模型相比，KG-Followup 在所有主题下都取得了有竞争力的结果。其中，健康数据任务是最具挑战性的类别，这可能是因为它要求进行精准的结构化推理并具备较高的事实准确性——即使是微小的错误，也可能对下游临床决策产生重大影响。