Linking Knowledge to Care: Knowledge Graph-Augmented Medical Follow-Up Question Generation

Liwen Sun[1*], **Xiang Yu**[2], **Ming Tan**[2], **Zhuohao Chen**[2],
Anqi Cheng[2], **Ashutosh Joshi**[2], **Chenyan Xiong**[1]
[1] Carnegie Mellon university [2] Amazon Health AI

Abstract

Clinical diagnosis is time-consuming, requiring intensive interactions between patients and medical professionals. While large language models (LLMs) could ease the pre-diagnostic workload, their limited domain knowledge hinders effective medical question generation. We introduce a Knowledge Graph-augmented LLM with active in-context learning to generate relevant and important follow-up questions, KG-Followup, serving as a critical module for the pre-diagnostic assessment. The structured medical domain knowledge graph serves as a seamless patch-up to provide professional domain expertise upon which the LLM can reason. Experiments demonstrate that KG-Followup outperforms state-of-the-art methods by 5% - 8% on relevant benchmarks in recall.

1 Introduction

Effective diagnostic performance relies not only on reasoning over explicit patient information but also on eliciting the right information. A significant portion of diagnostic errors stems from failures in comprehensive information gathering and history taking (Tu, 2025; Erin P. Balogh and Ball, 2015; Singh et al., 2013; Graber et al., 2013; Ely et al., 2011). Generating adequate follow-up questions can help reduce physician workload—from referrals and repeat visits to corrective actions—while lowering healthcare inefficiencies and costs, ultimately improving patient satisfaction, especially under time and resource constraints (Abimanyi-Ochom, 2019; Singh and Sittig, 2020; Schiff et al., 2009; Trowbridge et al., 2013).

In modern clinical practice, generating follow-up questions is still manual and time-consuming, limiting diagnostic efficiency. This work automates the process using LLMs that emulate physicians' inquiry strategies during medical encounters. However, existing LLMs often fail to identify information gaps across diverse symptoms (Xiong et al., 2024; Li et al., 2024a, 2025; Gatto et al., 2025). We introduce KG-Followup, a knowledge graph–augmented framework that leverages structured medical concepts to guide LLMs in generating clinically relevant and comprehensive follow-up questions (Chandak et al., 2023; Wu et al., 2024). The framework integrates EHR-guided concept retrieval, DDX-guided reasoning, and KG-informed active in-context learning to provide efficient, contextually grounded question generation—reducing clinicians' information-gathering burden and improving diagnostic efficiency.

To enable comprehensive evaluation across diverse real-world clinical scenarios, we introduce ClinicalInquiryBench, a novel benchmark specifically designed to assess an AI system's ability to generate clinically appropriate follow-up questions. ClinicalInquiryBench was developed through systematic transformation of publicly available physician-annotated clinical conversations (Arora et al., 2025).

Our experiments show that KG-Followup outperforms state-of-the-art methods on ClinicalInquiryBench and FollowupBench, achieving 70% and 80% recall with an adequate number of questions. KG-informed active ICL is more effective than random ICL in few-shot settings, and while prompting more questions from the LLM improves performance, KG-Followup achieves comparable results with far fewer. Case studies show that KG-Followup retrieves diverse symptom concepts, enabling comprehensive follow-up generation. Our main contributions can be summarized as follows:

- We introduce KG-Followup, a knowledge graph–augmented framework that guides LLMs to generate clinically relevant and comprehensive follow-up questions.
- We curate ClinicalInquiryBench, a benchmark

[*] Work done while interning at Amazon.

Findings of the Association for Computational Linguistics: EACL 2026, pages 846–853
March 24-29, 2026 ©2026 Association for Computational Linguistics

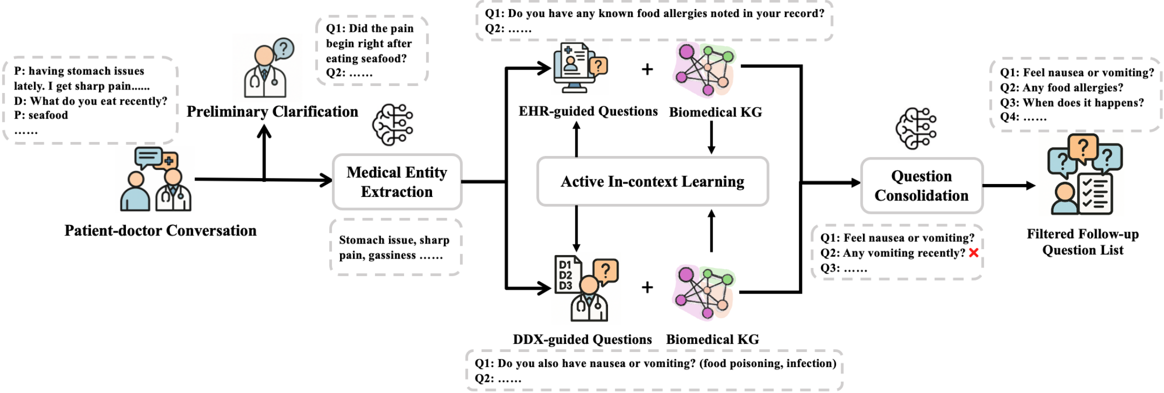

P: having stomach issues lately. I get sharp pain......
D: What do you eat recently?
P: seafood
......
Patient-doctor Conversation
Preliminary Clarification
Q1: Did the pain begin right after eating seafood?
Q2: ......
Medical Entity Extraction
Stomach issue, sharp pain, gassiness ......
Q1: Do you have any known food allergies noted in your record?
Q2: ......
EHR-guided Questions
+
Biomedical KG
Active In-context Learning
D1
D2
D3
DDX-guided Questions
Biomedical KG
Q1: Do you also have nausea or vomiting? (food poisoning, infection)
Q2: ......
Question Consolidation
Q1: Feel nausea or vomiting?
Q2: Any vomiting recently? ✗
Q3: ......
Q1: Feel nausea or vomiting?
Q2: Any food allergies?
Q3: When does it happens?
Q4: ......
Filtered Follow-up Question List

Figure 1: Our method generates preliminary questions from a patient's message and EHR, extracts medical entities via an LLM, applies (i) EHR-guided associative concept retrieval, (ii) DDX-guided reasoning path search, and (iii) KG-informed active in-context learning, and finally filters the results into a controlled set of follow-up questions.

with diverse clinical scenarios to directly evaluate LLMs' diagnostic question-seeking ability.

- We show that KG-Followup achieves state-of-the-art performance, reaching 70% and 80% recall on two benchmarks with a controlled number of questions.

2 Methods

2.1 Task Formulation

Given a patient-doctor conversation C and a language model F,

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

The goal is to produce a set $\hat{Q}$ where each $\hat{q}_i$ is a follow-up question to the conversational dialogue. As our focus is on static information gathering, all follow-up questions must be generated in a single pass from the conversation history.

2.2 Preliminary Clarification

Given the conversation C, we first prompt the LLM F to generate a preliminary set of follow-up questions $\hat{Q}_{\text{pre}}$. This step simulates an initial doctor consultation in which the LLM, relying solely on its internal knowledge, asks clarification questions about the patient's status before incorporating any external clinical knowledge.

2.3 EHR-Guided Questions via KG Linking Symptoms

To better identify the critical symptoms and diseases mentioned in a patient's messages, we first prompt the LLM to extract key medical entities, serving as a proxy electronic health record (EHR):

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

where $E = \{e_1, \ldots, e_n\}$ are the extracted clinical entities. These entities form the starting nodes in the medical KG, enabling more focused and informed reasoning about the patient's condition. We link extracted entities to KG nodes using string and embedding similarity, perform a breadth-first search with specific depth to build entity-specific subgraphs, and intersect them to identify shared clinically relevant concepts.

Because not all intersected KG concepts are relevant to the patient's case, we rank intersected KG concepts by relevance using the LLM and select the top-k_1 associative entities, then generate follow-up questions from the patient conversation C and ranked concepts:

$$\hat{E} = F_{\text{rank-entity}}(E, C), \tag{3}$$
$$\hat{Q}_{\text{ehr-kg}} = F_{\text{generate}}(C, \hat{E}). \tag{4}$$

This enriches question generation with KG-derived symptoms for more comprehensive, clinically relevant inquiries.

2.4 DDX-Guided Questions via KG Reasoning Diagnoses

To mirror the step-by-step reasoning physicians use to narrow down a diagnosis through strategic inquiry, we first conduct a differential diagnosis (DDX) with LLM based on the patient's current condition, producing a set of possible diagnoses.

Table 1: Major Results. / means the avg. number of generated questions. Active ICL refers to ICL with KG-informed hard examples.

Method	**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
Backbone: Claude Haiku			
Zero-Shot-U	0.52 / 7	0.52 / 7	0.54 / 10
Zero-Shot-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
+ Random ICL	0.72 / 20	0.72 / 20	0.77 / 40
+ Active ICL	**0.74 / 20**	**0.73 / 20**	**0.78 / 40**
Backbone: Claude Sonnet			
Zero-Shot-U	0.59 / 10	0.59 / 10	0.53 / 10
Zero-Shot-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
+ Random ICL	0.71 / 20	0.72 / 20	0.81 / 40
+ Active ICL	**0.75 / 20**	**0.77 / 20**	**0.82 / 40**

Table 2: Ablation study of different generation signals, evaluated using Claude Haiku on the FollowupBench.

Method	**Recall / No.**
Module: EHR-guided Generation	
Rationale from retrieved KG triplets	0.72 / 25
Retrieved similar KG concepts	0.71 / 25
Intersected concepts across traversed subgraphs	**0.72 / 26**
Module: DDX-guided Generation	
DDX rule-out questions	0.72 / 26
+ KG reasoning paths	0.75 / 34
Module: Question Consolidation	
KG-Followup w/o. consolidation	**0.79/51**
+ cluster merging	0.77/40
+ LLM selection	0.71/40

We then ask follow-up questions designed to eliminate both worst-case and best-case diagnosis:

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

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

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

where $D = \{d_1, ..., d_m\}$ are total possible diagnoses. The final question set is formed by taking the union of all targeted questions across the candidate diagnoses to eliminate different possibilities. Beyond the LLM's internal knowledge for DDX-based question generation, we integrate structured medical knowledge from the KG, linking each EHR entity to potential diagnoses through reasoning paths. Critical intermediate nodes along these paths provide additional context and medical grounding. Specifically, for each source entity–target diagnosis pair (e_i, d_j), we sample k_2 shortest reasoning paths $P_{i,j}$ in the KG. Since massive paths of the same length may exist, we then use the LLM to select the single most relevant path $\hat{P}_{i,j}$ based on the patient's context, pruning irrelevant KG reasoning paths:

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

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

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

where the complete path set $\hat{P}$ is aggregated from all traversed reasoning paths across source entity–target diagnosis pairs. We generate follow-up questions using the selected paths and their critical intermediate nodes as supporting knowledge:

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

These KG reasoning paths produce follow-up questions that are both clinically grounded and tailored to the patient's context. They also serve as a complement to DDX-only follow-up questions.

2.5 Active In-context Learning via KG-informed Hard Cases

KG not only helps identify additional symptoms to inquire about but also aids in finding challenging patient queries. Inspired by active learning, we treat these difficult cases as in-context learning (ICL) examples, providing them to the LLM alongside the original patient conversation to guide follow-up question generation. Specifically, when patient messages lack source symptoms or cannot be mapped to KG entities—making KG traversal infeasible—we treat them as hard queries and include them as ICL cases: $T = \{(C_1, Q_1) ...... (C_t, Q_t)\}$. The final question set is then constructed as:

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

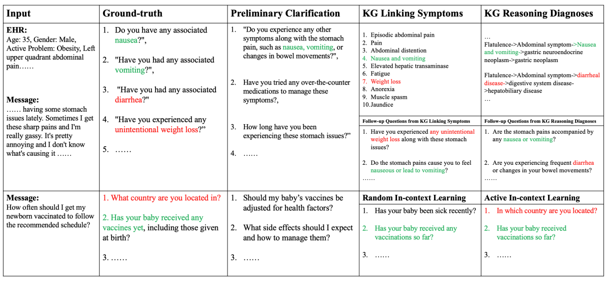
where each generation module is augmented with ICL using T to ensure effective handling of edge cases and broader coverage beyond explicit symptom questions.

2.6 Question Consolidation

To reduce redundancy among generated follow-up questions, we embed all questions with a medical encoder and apply K-means clustering. An LLM then refines multi-question clusters by merging overlaps and removing duplicates, yielding a concise set for efficient patient interaction.

Table 3: Ablation study of ICL example selection using Claude Sonnet on ClinicalInquiryBench.

Method	Dev	Test
Random	0.71 / 20	0.72 / 20
KG-informed Hard	**0.75 / 20**	**0.77 / 20**
Supervised Hard	**0.75 / 20**	0.75 / 20



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	… 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 Linking Symptoms	**Follow-up Questions from KG Reasoning Diagnoses**
			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? ……	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**	**Active In-context Learning**
			1. Has your baby been sick recently? 2. Has your baby received any vaccinations so far? 3. ……	1. In which country are you located? 2. Has your baby received vaccinations so far? 3. ……

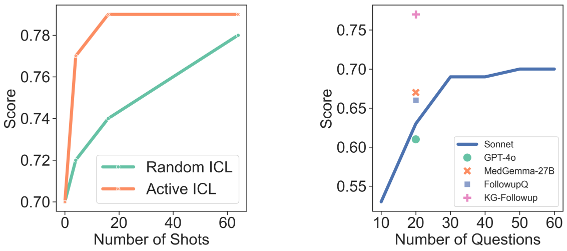
Figure 2: **Case Study.** Red and Green indicate questions generated from KG-retrieved symptom signals and the LLM's internal knowledge, respectively.

3 Experimental Setting

Datasets. We evaluate on two datasets: **Followup-Bench** (Gatto et al., 2025), an expert-curated set of 250 instances (avg. ∼9 questions, no weights), and **ClinicalInquiryBench**, derived from 5,000 HealthBench (Arora et al., 2025) cases filtered to 1,498 context-seeking instances, reformatted with question weights (avg. ∼5 questions). We split ClinicalInquiryBench into 250 dev and 1,248 test examples, sampling ICL examples from the dev set. Curation details are in Appendix 6.2.

Evaluation Metrics. Following Gatto et al. (2025), we use *weighted recall* as the main metric, comparing generated questions $\hat{Q}$ with ground truth Q via LLM-as-judge (Zheng et al., 2023). Since our goal is to capture those in the ground-truth set, precision would unfairly penalize clinically valid but unmatched questions. Details are in Appendix.

Baselines. We compare against: (1) **Zero-Shot-U**: LLM generates an unrestricted number of follow-up questions, serving as a proxy for preliminary clarification. (2) **Zero-Shot-k**: fixed k questions to study scaling with output size.(3) **FollowupQ** (Gatto et al., 2025): multi-agent system with clarification, EHR reasoning, and DDX modules (w/o. KG). (4) **Random ICL**: random in-context examples to show the benefit of KG-informed selection. We also evaluate **GPT-4o** (et al., 2024) and **MedGemma-27B** (Sellergren, 2025) under the Zero-Shot-k setting.



(a) ICL trend

(b) Zero-shot-k trend

Figure 3: Analysis of ICL and Zero-Shot-k trends with Claude Sonnet on ClinicalInquiryBench: Figure (a) shows the performance across shot numbers; Figure (b) shows the effect of controlled Zero-Shot-k.

4 Evaluation Results

Major Results. Results in Table 1 show that KG augmentation delivers consistent improvements over SOTA methods without external knowledge, exceeding them by over 5%. Incorporating KG-informed active ICL further increases performance by an additional 3% across both benchmarks.

Interestingly, Sonnet and Haiku perform comparably under Zero-Shot-k prompting. However, with KG augmentation, Sonnet surpasses Haiku by up to 4 points, indicating that Sonnet is more effective at leveraging external knowledge.

Ablation Study. In Table 2, We tested multiple KG augmentation strategies (e.g., RAG over triplets, concept retrieval) and found minimal differences. The DDX-guided module provided larger gains than the EHR-guided one. Using an LLM to trim questions reduced performance, while our clustering-based merging preserved quality and efficiently reduced the pool from 51 to 40, easing clinicians' workload.

For active ICL, we also test a supervised variant selecting hard cases (recall = 0) from the dev set in Table 3. Our unsupervised approach performs competitively, confirming the effectiveness of KG-informed hard case selection. As shown in Figure 3a, performance rises with more in-context examples, with few-shot active ICL outperforming random ICL, offering better token efficiency.

Case Study. In the first case (Figure 2), ground-truth questions contain diarrhea, while preliminary clarification only covers nausea and vomiting. KG reasoning reveals hidden symptom links (e.g., diarrheal disease, weight loss), enriching LLM-generated questions. In the last case, random ICL misses non-symptom queries like geographic location, whereas active ICL, guided by hard examples, enables generating edge follow-up questions.

Zero-shot-k Analysis. Figure 3b shows that performance improves with more LLM-generated questions but saturates beyond $k{>}40$. In contrast, our method achieves similar results with fewer questions, enhancing the efficiency of doctor–patient interactions.

5 Conclusion

This work tackles the follow-up question generation task through KG-Followup, a knowledge graph–augmented framework that enables LLMs to produce clinically grounded clarification questions. To support systematic evaluation, we curate ClinicalInquiryBench, a large-scale dataset annotated with follow-up question importance. Building on this resource, we introduce EHR-guided and DDX-guided generation modules enhanced with KG search and KG-informed active in-context learning. Experiments show that KG-Followup surpasses state-of-the-art methods by over 8% on ClinicalInquiryBench and 5% on FollowupBench in recall with an adequate number of questions.

Limitations

One limitation of this work lies in its dependency on pre-constructed biomedical knowledge graphs for question grounding. While the use of structured knowledge enhances the clinical relevance and accuracy of generated follow-up questions, it assumes that the knowledge graph is both comprehensive and up-to-date. In practice, many real-world EHR systems may not have access to such high-quality knowledge graphs, or may rely on domain-specific ontologies with limited coverage. This reliance may restrict the generalizability of KG-Followup to low-resource clinical settings or rapidly evolving domains where the KG lags behind current medical understanding.

Additionally, while KG-Followup demonstrates strong performance on ClinicalInquiryBench and FollowupBench, both benchmarks are constructed for offline evaluation and may not capture the full complexity of real-world deployment settings, such as clinical decision support tools or EHR-integrated applications. In practice, deployed systems must contend with incomplete or noisy data, diverse clinical workflows, and evolving patient contexts. Extending this work to support more diverse and dynamic patient contexts would be an important step toward real-world applicability.

Ethics Considerations

This work aims to enhance clinical question generation through knowledge graph–augmented language models, with the goal of supporting clinician decision-making. While the system operates on de-identified, publicly available datasets, ethical concerns remain regarding potential biases in both the language model and the underlying knowledge graph, which may influence the relevance or safety of generated questions. Additionally, there is a risk of over-reliance on automated suggestions in clinical workflows. To mitigate these concerns, any future deployment should include careful human oversight, domain-specific auditing, and alignment with ethical standards in healthcare AI, including fairness, transparency, and patient safety.

References

Bohingamu Mudiyanselage S. Catchpool M. et al. Abimanyi-Ochom, J. 2019. *Strategies to reduce diagnostic errors: a systematic review*.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. Healthbench: Evaluating large language models towards improved human health. *Preprint*, arXiv:2505.08775.

Abhinand Balachandran. 2024. Medembed: Medical-focused embedding models.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Nature Scientific Data*.

John W Ely, Jerome A Osheroff, Melisa L Chambliss, Mark H Ebell, and Mark E Rosenbaum. 2011. The value of clinical questions: identifying research priorities in primary care. *BMJ Quality & Safety*, 20(9):787–792.

Bryan T. Miller Erin P. Balogh and John R. Ball. 2015. *Improving Diagnosis in Health Care*, chapter 3, Overview of Diagnostic Error in Health Care. The National Academies Press.

OpenAI et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Joseph Gatto, Parker Seegmiller, Timothy Burdick, Inas S. Khayal, Sarah DeLozier, and Sarah M. Preum. 2025. Follow-up question generation for enhanced patient-provider conversations. *Preprint*, arXiv:2503.17509.

Mark L Graber, Robert M Wachter, and Christine K Cassel. 2013. Diagnostic error in internal medicine. *BMJ Quality & Safety*, 22(Suppl 2):ii21–ii27.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. 2025. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. *Preprint*, arXiv:2410.04585.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Jerome P. Kassirer and Richard I. Kopelman. 2010. *Learning Clinical Reasoning*, 2 edition. Lippincott Williams & Wilkins.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024a. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Preprint*, arXiv:2406.00922.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024b. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pages 28858–28888. 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, and Maarten Sap. 2025. Alfa: Aligning llms to ask good questions a case study in clinical reasoning. *Preprint*, 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, and 1 others. 2009. Diagnostic error in medicine: analysis of 583 physician-reported errors. *Archives of Internal Medicine*, 169(20):1881–1887.

Henk G. Schmidt and Remy M. Rikers. 2007. How expertise develops in medicine: knowledge encapsulation and illness script formation. *Medical Education*, 41(12):1133–1139.

Andrew et al. Sellergren. 2025. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Hardeep Singh, Andrew N D Meyer, and Eric J Thomas. 2013. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies. *JAMA Internal Medicine*, 173(6):418–425.

Hardeep Singh and Dean F Sittig. 2020. Advancing the science of measurement of diagnostic errors in healthcare: the safer dx framework. *BMJ Quality & Safety*, 29(10):874–880.

Robert L Trowbridge, Jason J Rencic, and Steven J Durning. 2013. Teaching clinical reasoning: a practical framework for classroom and bedside instruction. *Academic Medicine*, 88(2):182–188.

Schaekermann M. Palepu A. et al. Tu, T. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, pages 642, 442–450.

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, and Yuyin Zhou. 2025. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *Preprint*, arXiv:2504.00993.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *Preprint*, arXiv:2408.04187.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *Preprint*, arXiv:2402.13178.

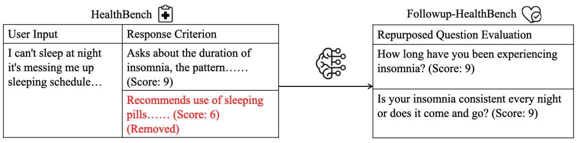
Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

6 Appendix

6.1 Related Work

Follow-up questions are a cornerstone of clinical reasoning, enabling physicians to clarify ambiguous details, uncover overlooked symptoms, and guide diagnostic decision-making.(Kassirer and Kopelman, 2010; Schmidt and Rikers, 2007). MediQA (Li et al., 2024b) extends traditional single-turn benchmarks such as MedQA (Jin et al., 2020) into more realistic, interactive formats. Unlike MedQA, where full patient context is given upfront, real-world decision-making begins with limited information. In such interactive settings, LLMs often fail to ask clarifying questions and tend to make premature or overconfident diagnoses.

To address this, FollowupQ (Gatto et al., 2025) curates a dataset of ground-truth follow-up questions authored by physicians and introduces a multi-agent system that generates personalized follow-ups based on patient messages and EHR data. The goal is to reduce ambiguity in medical conversations by helping clinicians collect the most relevant, case-specific information. However, existing datasets for AI-driven follow-up question generation remain limited in scale and scope, and they fail to capture the varying diagnostic importance of different questions. For instance, Followup-Bench (Gatto et al., 2025) contains only 250 instances and relies solely on static EHR information



User Input	Response Criterion
I can't sleep at night it's messing me up sleeping schedule…	Asks about the duration of insomnia, the pattern…… (Score: 9)
	Recommends use of sleeping pills…… (Score: 6) (Removed)

Repurposed Question Evaluation
How long have you been experiencing insomnia? (Score: 9)
Is your insomnia consistent every night or does it come and go? (Score: 9)

Figure 4: A curated instance illustration.

to motivate follow-up questions, without encoding their relative diagnostic importance. This restricts its ability to reflect realistic clinical reasoning and the nuanced value of different follow-ups.

Knowledge graph (KG)–augmented generation further advances medical question answering by allowing models to incorporate structured biomedical relationships, leading to more accurate and contextually grounded diagnostic reasoning. For example, PrimeKG (Chandak et al., 2023) provides a large-scale biomedical KG that supports knowledge-augmented inference. MedGraphRAG (Wu et al., 2024) leverages LLMs to organize retrieval-augmented generation (RAG) data into graph structures, showing strong potential for extracting holistic insights from long-form documents. MedReason (Wu et al., 2025) constructs supervised fine-tuning data from KG reasoning paths derived from GPT-4 to strengthen factual medical reasoning. KARE (Jiang et al., 2025) integrates community-level KG retrieval with LLM reasoning to improve healthcare predictions.

Building on these directions, our work explores how associative medical concepts from KGs can enhance follow-up question generation, enabling LLMs to produce clinically grounded and diagnostically useful questions.

6.2 Benchmark

Initially, we overview our source data, HealthBench, and then the curation process of ClinicalInquiryBench.

6.2.1 Preliminary

HealthBench (Arora et al., 2025) is an open-source benchmark for evaluating the performance and safety of LLMs in healthcare. It contains 5,000 multi-turn conversations between patients and healthcare professionals, with responses assessed against 48,562 unique rubric criteria curated by 262 physicians. These criteria span diverse medical contexts (e.g., emergencies, clinical data transformation, global health) and behavioral dimensions (e.g., accuracy, instruction following, communication), making HealthBench a comprehensive foundation for benchmark development. Building on this, we construct ClinicalInquiryBench by systematically filtering and transforming relevant instances from HealthBench.

Table 4: Theme coverage across HealthBench and ClinicalInquiryBench.

Theme	Curated Instances (%)	Original Instances (%)
Total examples	**1498** (100.0%)	**5000** (100.0%)
Global health	317 (21.2%)	1,097 (21.9%)
Hedging	348 (23.2%)	1,071 (21.4%)
Communication	99 (6.6%)	919 (18.4%)
Context seeking	312 (20.8%)	594 (11.9%)
Emergency referrals	132 (8.8%)	482 (9.6%)
Health data tasks	239 (16.0%)	477 (9.5%)
Response depth	51 (3.4%)	360 (7.2%)

ClinicalInquiryBench explicitly centers on the role of follow-up questioning in clinical conversations. Compared to prior benchmarks, it is larger in scale, enriched with broader categories of follow-up intent, and annotated with diagnostic importance. This design enables more rigorous evaluation of whether LLMs can generate context-aware, clinically valuable follow-up questions that mirror physician clarification strategies. ClinicalInquiryBench preserves all themes from HealthBench, ensuring no loss of clinical scenario coverage. Summary statistics are shown in Table 4.

6.2.2 ClinicalInquiryBench Curation

Here, we describe the transformation of HealthBench into a follow-up question benchmark, along with our problem formulation.

Rubric Filtering. We begin by selecting English-only instances from HealthBench. From this subset, we apply a filtering stage using Claude 4, designed to identify rubrics that explicitly seek additional clinical context—such as clarifying symptoms, narrowing differential diagnoses, or obtaining missing patient history. Entries with empty rubrics or rubrics unrelated to follow-up questioning are removed to maintain dataset relevance.

Follow-up Repurpose. Then, we repurpose the retained rubrics into explicit evaluation criteria for follow-up question generation. Using Claude 4, we reframe each rubric to emphasize the information-gathering objective, enabling the benchmark to assess whether a model can identify information gaps and generate questions that are both clinically relevant and diagnostically impactful. Importantly, each question in the benchmark is assigned a weight score derived from its original rubric score in HealthBench, reflecting its relative diagnostic

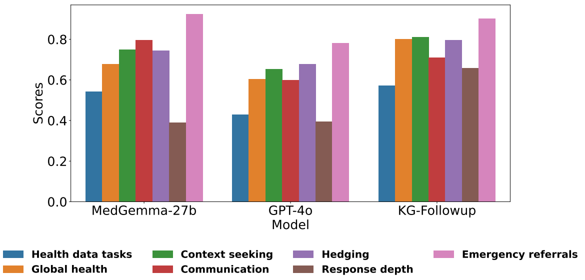


Figure 5: Weighted recall rates by theme across models. The KG-Followup framework is implemented using the Claude Sonnet backbone.

importance. This transformation shifts the rubric from a general conversational quality metric to a focused diagnostic questioning standard, producing a dataset tailored for evaluating LLM's capability to simluate a focused diagnostic inquiry. One sample is shown in Figure 4.

6.3 Implementation Details

For $\hat{Q}_{\text{pre}}$, we prompt the LLM to generate 20 questions for ClinicalInquiryBench and 40 questions for FollowupBench. For $\hat{Q}_{\text{ddx}}$, we prompt LLM to generates 2 worst-case and 2 best-case candidate diagnoses, followed by 2 follow-up questions for each diagnosis.For $\hat{Q}_{\text{ehr-kg}}$, we use the top-10 ranked intersected entities, and for $\hat{Q}_{\text{ddx-kg}}$, we sample 30 reasoning paths for each source entity–target diagnosis pair to augment generation. No fixed question number is enforced for $\hat{Q}_{\text{ehr-kg}}$ or $\hat{Q}_{\text{ddx-kg}}$. We use 4 ICL examples from the development set. Before consolidation, our framework generates an average of 30 questions on ClinicalInquiryBench and 50 on FollowupBench. We use MedEmbed-large-v0.1 (Balachandran, 2024) as the medical encoder, Claude Haiku and Sonnet as the LLM generator and judger, and PrimeKG (Chandak et al., 2023) as the external knowledge base. We will release source code after the paper gets accepted.

6.4 Evaluation Details

We primarily use Claude Sonnet as the evaluator, employing a list-wise prompt to assess whether the generated questions are present in the ground-truth question set.

6.5 Theme Analysis

We also present the model performance across different themes in Figure 5. KG-Followup achieves competitive results across all themes compared to other models. Among them, Health Data Tasks emerges as the most challenging category, likely due to its demand for precise structured reasoning and high factual accuracy—where even minor errors can significantly impact downstream clinical decisions.