Dialogue to Question Generation for Evidence-based Medical Guideline Agent Development

Zongliang Ji[*1,2,3] JERRYJI@CS.TORONTO.EDU
Ziyang Zhang[*1,4] ZIYANG.ZHANG2@EMORY.EDU
Xincheng Tan[1] CARATAN@GOOGLE.COM
Matthew Thompson[1] MTHOMP@GOOGLE.COM
Anna Goldenberg[2, 3] ANNA.GOLDENBERG@UTORONTO.CA
Carl Yang[4] J.CARLYANG@EMORY.EDU
Rahul G. Krishnan[2, 3] RAHULGK@CS.TORONTO.EDU
Fan Zhang[1] ZHANFAN@GOOGLE.COM
[1] *Google Research*
[2] *University of Toronto, Canada*
[3] *Vector Institute, Canada*
[4] *Emory University, USA*

Abstract

Evidence-based medicine (EBM) is central to high-quality care, but remains difficult to implement in fast-paced primary care settings. Physicians face short consultations, increasing patient loads, and lengthy guideline documents that are impractical to consult in real time. To address this gap, we investigate the feasibility of using large language models (LLMs) as ambient assistants that surface targeted, evidence-based questions during physician–patient encounters. Our study focuses on question generation rather than question answering, with the aim of scaffolding physician reasoning and integrating guideline-based practice into brief consultations. We implemented two prompting strategies, a zero-shot baseline and a multi-stage reasoning variant, using Gemini 2.5 as the backbone model. We evaluated on a benchmark of 80 de-identified transcripts from real clinical encounters, with six experienced physicians contributing over 90 hours of structured review. Results indicate that while general-purpose LLMs are not yet fully reliable, they can produce clinically meaningful and guideline-relevant questions, suggesting significant potential to reduce cognitive burden and make EBM more actionable at the point of care.

Keywords: LLMs, Evidence-based medicine, Clinician-facing tools, Clinical Decision Support System

Data and Code Availability The data we use is not entirely shareable as the dataset is preparatory. We share some generated evidence-based medical questions in the shared code base. The prompts and partial data used for this study is contained in this GitHub repository [1].

Institutional Review Board (IRB) Our study does not require IRB.

1. Introduction

Evidence-based medicine (EBM) is central to the delivery of high-quality care. Its principles are operationalized through clinical guidelines, which summarize research evidence into structured recommendations designed to support everyday clinical decisions (Panteli et al., 2019). Despite their importance, primary care physicians (PCPs) frequently struggle to apply guidelines in practice. Consultation times are short, often less than 15 minutes in the United States (Neprash et al., 2021), and the format of guidelines as lengthy reference documents requiring precise keyword searches makes them impractical to use during real-world encounters (Pondicherry et al., 2023). As a

[*] Equal contribution. Work done while interning at Google.

[1] https://github.com/Jerryji007/Dialogue2Questions-ML4H2025

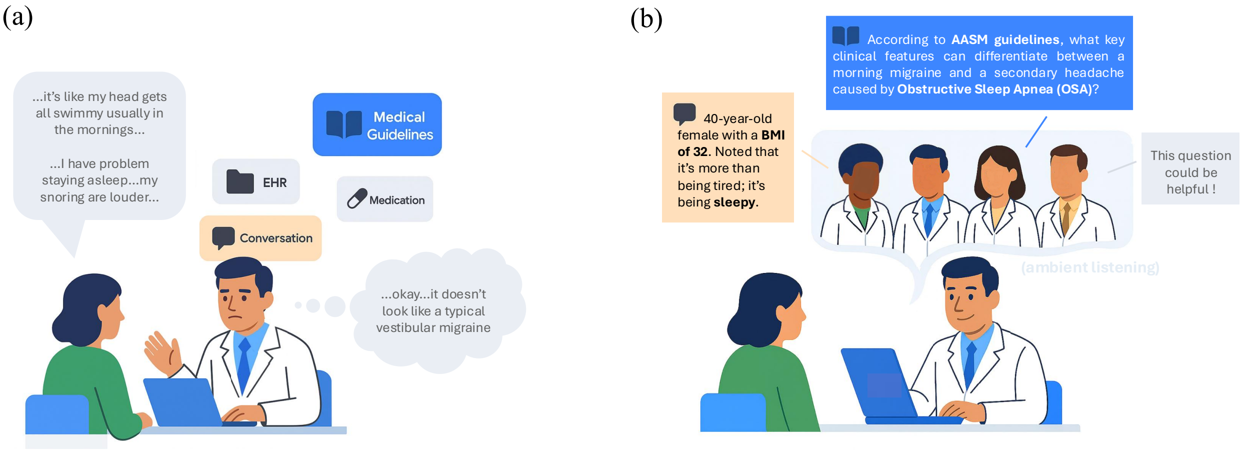

(a)
...it's like my head gets all swimmy usually in the mornings...
...I have problem staying asleep...my snoring are louder...
Medical Guidelines
EHR
Medication
Conversation
...okay...it doesn't look like a typical vestibular migraine
(b)
According to AASM guidelines, what key clinical features can differentiate between a morning migraine and a secondary headache caused by Obstructive Sleep Apnea (OSA)?
40-year-old female with a BMI of 32. Noted that it's more than being tired; it's being sleepy.
This question could be helpful !
(ambient listening)

Figure 1: **Illustration of an outpatient visit.** A 40-year-old female with a high BMI presents with headache and fatigue. (a) In a routine encounter, the PCP simultaneously juggles the patient's narrative, EHR, prior comorbidities, and guidelines, while considering differential diagnoses such as vestibular migraine. (b) With intelligent support, the PCP receives context-aware, evidence-based information that help structure reasoning without interrupting patient communication.

result, guideline-based decision-making often remains more aspirational than achievable at the point of care.

Consider the scenario in Figure 1a. A PCP must listen, extract key clinical details from the narrative, review the electronic health record (EHR), and begin differential diagnosis. Several guidelines may apply, but cross referencing them in real time creates substantial cognitive load. Figure 1b envisions a silent virtual expert that listens to the dialogue and raises targeted, evidence based questions, then uses retrieval to answer them by synthesizing guideline content. The ability to surface the right questions is not merely preparatory but decisive: it defines the retrieval space, shapes clinical reasoning, and directly reduces cognitive load when applying guidelines under time pressure. Question generation therefore constitutes a high-impact problem on its own, even before answer synthesis is integrated. This paper formalizes and studies the question generation in depth in the primary care outpatient setting.

Previous efforts to assist clinicians with guideline use include static mobile applications (Mitchell et al., 2020), electronic health record order sets (Bates et al., 2003), and computerized clinical decision support systems (CDSS) (Sutton et al., 2020). While these approaches improved accessibility, they often lacked context-awareness and imposed rigid workflows. More recent developments have shifted toward AI-driven assistants and large language models (LLMs), which allow clinicians to query guidelines in natural language (Sagheb et al., 2022; Lichtner et al., 2023) or receive retrieval-augmented responses that make guideline content more actionable at the point of care (Oniani et al., 2024; Ferber et al., 2024). Despite these advances, few systems have been explicitly designed or evaluated for proactive, real-time question generation during visits (Fast et al., 2024; Hager et al., 2024b).

In this work, we investigate the feasibility of using LLMs as ambient question generators that support physicians in practicing evidence-based medicine. Specifically, our contributions are as follows:

- We identify short primary care visits as a central barrier to practicing EBM and propose, for the first time, an ambient LLM that listens to the encounter, surfaces guideline oriented questions, and later retrieves and synthesizes answers. In this paper we address the first step by generating EBM questions from dialogue, instantiated with two prompting strategies: zero shot and multi stage.

- We curate an evaluation benchmark consisting of 80 de-identified real-world physician–patient dialogue transcripts, each presented at three different truncation lengths. To systematically assess question quality, we design five complementary metrics across different aspects and experimental tasks for evaluation.

- We conduct an extensive human evaluation with six experienced physicians, who collectively devoted more than 90 hours to assessing the generated questions. In parallel, we set up an LLM-as-judge evaluation for comparison. These experiments yield insightful findings on the necessity of such systems, the plausibility of an ambient assistant, the comparison of prompting methods, the variation across question types, and the promises & limitations of automated evaluation.

2. Related Work

Challenges in Practicing Evidence-Based Medicine Primary care physicians (PCPs) face longstanding structural barriers to applying evidence-based medicine (EBM). Time studies estimate that a solo PCP would require 7-18 hours per day for preventive and chronic care alone (Yarnall et al., 2003; Østbye et al., 2005), and nearly 27 hours per day to fully implement guideline recommendations for a typical patient panel (Porter et al., 2023). These demands intersect with a persistent workforce shortage, with the U.S. projected to lack over 50,000 PCPs by 2025 (Petterson et al., 2012). Consequently, many evidence-supported interventions are omitted in practice, with lack of time, difficulty accessing up-to-date guidance, and limited EBM training cited as key barriers (Zwolsman et al., 2012). At the point of care, roughly half of physicians' clinical questions remain unanswered due to workflow constraints (Del Fiol et al., 2014). Classic clinical decision support systems (CDSS) have shown benefits in narrow contexts (McGinn et al., 2013), but systematic reviews highlight mixed patient-level outcomes and barriers such as poor usability and alert fatigue (Garg et al., 2005; Kawamoto et al., 2005; Vasey et al., 2021).

LLMs in Healthcare and Clinical Support Recent advances in LLMs offer new opportunities for ambient, real-time support. LLMs have achieved expert-level accuracy on medical QA tasks (Singhal et al., 2025) and produced patient-facing answers judged more empathetic and higher quality than physicians' responses (Ayers et al., 2023). They have also been applied to clinical summarization, sometimes outperforming clinicians (Fraile Navarro et al., 2025; Van Veen et al., 2024). Retrieval-augmented approaches enhance grounding in guidelines (Shi et al., 2024). Emerging systems like MediQ (Li et al., 2024), HealthQ (Wang et al., 2025), and FollowupQ (Gatto et al., 2025) explore proactive question generation. Large-scale conversational agents (e.g., AMIE, KERAP) demonstrate diagnostic accuracy comparable to PCPs (Tu et al., 2025; Xie et al., 2025). Early trials show that GPT-4-based support can improve physician decision-making without worsening bias (Goh et al., 2025b,a), though limitations remain in nuanced clinical reasoning (Hager et al., 2024a). Overall, LLMs show promise as ambient assistants to reduce PCP cognitive load and integrate guideline-based reasoning into constrained visits.

Across prior work, EBM remains necessary yet underutilized in routine care, and existing tools have not effectively reduced the friction of guideline use during real encounters. Meanwhile, LLMs show promise in clinical NLP tasks. This work is, to our knowledge, the first to systematically assess whether an LLM can serve clinicians by proactively generating evidence-oriented, guideline-targeting questions during encounters, with structured, multi-physician human evaluation.

3. Problem Definition

Our long-term vision is to develop an ambient assistant that silently observes the primary care consultation and raises clinically meaningful questions grounded in EBM.

A downstream component, which is not part of this work, could then **retrieve and summarize the relevant guideline content** to support the physician. In this paper, we focus exclusively on the **generation of targeted, evidence-based questions** during the encounter.

We assume access to two main inputs. The first is the patient health record (PHR), represented by a structured intake questionnaire and basic clinical background, denoted as x_{phr}. The second is the dialogue between the patient and the physician, denoted as x_{dlg}. To approximate a realistic ambient setting, we model truncated dialogue contexts with a ratio parameter $r \in \{0.3, 0.7, 1.0\}$. The truncated dialogue is defined as

$$x_{dlg}^{(r)} = \text{First } r \times |x_{dlg}| \text{ tokens of the dialogue.}$$

This setting reflects two assumptions: the assistant may only have partial knowledge of the visit at a given time, and it may need to surface questions before the encounter is completed.

Given the pair $(x_{phr}, x_{dlg}^{(r)})$, the system generates a small set of questions: $Q = \{q_1, q_2, q_3\}$. We constrain the output to exactly three questions in order to maintain readability and align with the time-limited nature of outpatient consultations. The questions should not be trivial facts that a physician is expected to recall from memory; instead, they should require reference to guidelines or evidence-based resources.

4. Method

4.1. Backbone Model

The Gemini family of models has demonstrated strong performance, scalability, and flexibility in the fields of medicine and healthcare (Saab et al., 2024; Vedadi et al., 2025; Saab et al., 2025). In this work, we use Google's Gemini 2.5 (Comanici et al., 2025) as our backbone model . Gemini 2.5 contains Flash and Pro variants. For our Summarizer Agent, we use the Flash version for efficiency. For Question Generator Agent and Question Evaluator Agent, we use the Pro version for better quality.

4.2. Baseline: A Zero-shot Setting

As a baseline, we consider a simple zero-shot setting where the LLM directly takes the dialogue x_{dlg} and the patient questionnaire x_{phr} as input, to generate three candidate questions, without any few-shot exemplars or structured reasoning steps (Figure 2b). The prompt is presented in Appendix A.

4.3. Multi-stage Reasoning Framework

Summarizer Agent. The initial and most critical step in building an effective medical guidelines agent is to ensure all clinical information is accurately captured from the dialogue context, but this verbal communication is often unstructured and prone to information loss (Lange et al., 2024). Additionally, in real-world conversation, the patient and the doctor are not consistently communicating in professional ways. In many cases, their exchanges include low-density information or clinically irrelevant content, such as greetings or small talks, which does not directly contribute to identifying clinical conditions.

We design a *Summarizer Agent* to extract the essential information from such a verbose conversation and generate structured clinical documentation. It acts as the first stage of our multi-stage reasoning framework. Its objective is to listen to the dialogue, given the PHR, and generate a structured clinical summary (full prompt in Appendix A). Formally, the input to the LLM agent is a pair $(x_{phr}, x_{dlg}^{(r)})$. The output is a structured summary s, represented as a schema of clinical slots:

$$s = \{(k_1, v_1), (k_2, v_2), \ldots, (k_m, v_m)\}, \tag{1}$$

where each key k_i corresponds to a predefined clinical field such as ***chief complaint***, ***history of present illness***, ***medication and past history***, ***objective findings***, ***assessment***, and ***plan***. These fields draw on common clinical documentation practices such as the SOAP note structure (Podder et al., 2021), HL7 CDA (Dolin et al., 2001), and FHIR (Ayaz et al., 2021), which adopt similar section-based structures to ensure completeness and interoperability.

Question Generator Agent. Through consultation with more than 10 internal clinical experts[2], we identified scenarios where physicians typically consult guidelines: complex diagnoses, treatment initiation or adjustment, preventive screening, comorbidity management, and referral decisions. To reflect U.S. primary care practice, we also consider the distribution of common visit types and diagnoses (Ashman et al., 2023). A good generated question should be helpful, evidence-based, and insightful. Here is an example:

> *"A 40-year-old female with a history of migraines presents with new-onset morning headaches and fatigue. According to the American Academy of Sleep Medicine clinical practice guidelines, what key clinical features can help differentiate between a primary headache disorder and a headache caused by a strong clinical suspicion of Obstructive Sleep Apnea?"*

We follow a few-shot setting to ensure that the generated questions align with the high-quality *golden standard* questions verified by our experts. In particular, each question first introduces the necessary personal health background and key conditions (in blue) and then suggests a very specific guideline (in orange), acting as a *mental prompt* that exists only implicitly in the physician's mind, yet still consumes effort to think through or express. The LLM is prompted to analyze the given summary s generated by the *Summarizer Agent*, remain strictly grounded in the provided context to avoid hallucinations, and output exactly ten diverse questions across categories such as ***medication adjustment***, ***ordering tests***, ***medication details***, ***diagnosis***, ***follow-up***, and ***counseling***, with two expert-verified examples included in the prompt (see Appendix A).

Question Evaluator Agent. Not all ten generated questions are shown to physicians. To select

[2] These internal collaborators are different from the clinicians participated in the later evaluation study.

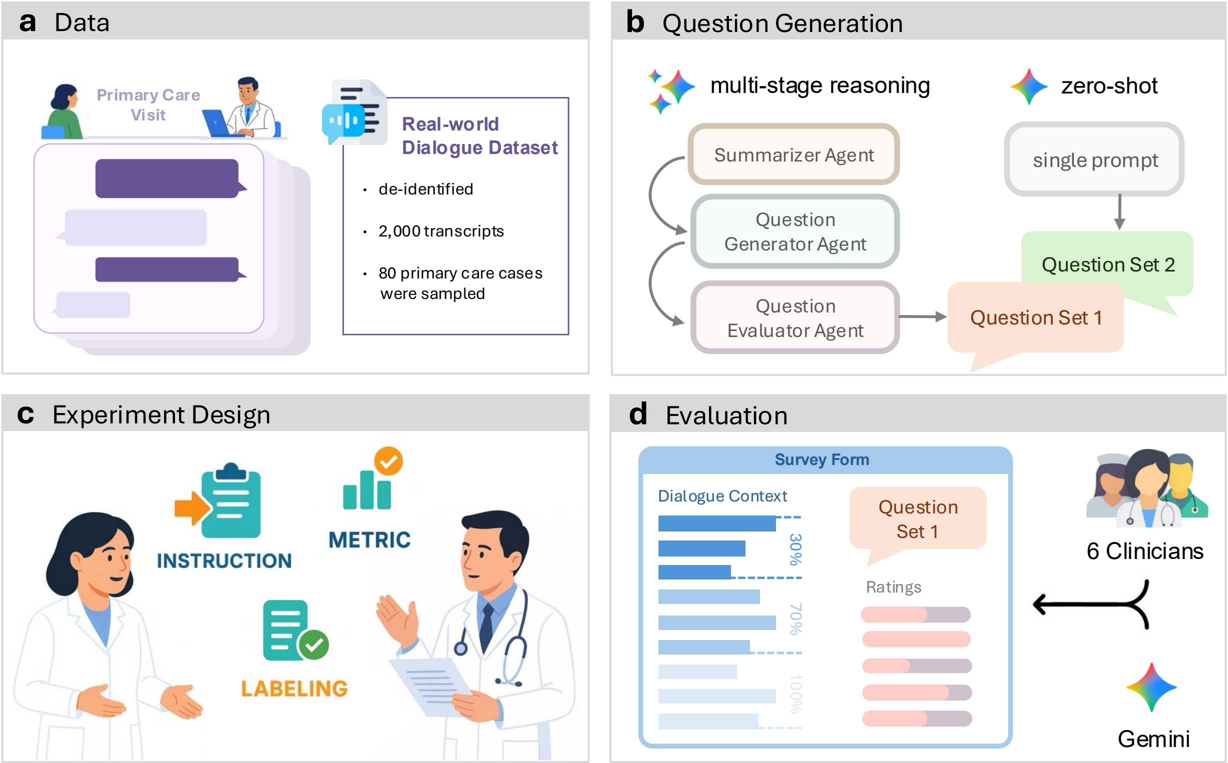


Figure 2: **Overview of our work.** (a) We sample 80 primary care cases from 2,000 real-world physician-patient dialogue transcripts. (b) We develop two methods to generate evidence-based medical questions from dialogue using Gemini 2.5. (c) We conduct a pilot study with more than 10 internal clinical experts regarding the evaluation metrics, pilot labeling, and rationale of designing the experiments. (d) We perform auto evaluation with LLM and human evaluation on 80 cases with 6 experienced clinicians.

the top-3 questions in terms of their quality, we run a candidates evaluation in which the *Question Evaluator Agent* rates each question (1.0–5.0) on seven predefined criteria (see their definition and prompt in Appendix A). Empirically, LLM evaluators often give uniformly high scores because the questions were generated by a state-of-the-art model, yet meaningful comparison requires more differentiated scoring. We employ a chain-of-thought (CoT) procedure inspired by Liu et al. (2025): for each candidate, the LLM briefly reasons about the strongest pros and cons (grounded in the seven criteria), then assigns scores for all criteria before proceeding to the next question. Let c_{ij} be the score for the i-th question under the j-th criterion; we select the top-3 by the mean score across criteria:

$$\text{Selected Questions} = \underset{\substack{Q\subseteq\{1,\ldots,10\}\\ |Q|=3}}{\arg\max} \sum_{i\in Q} \left(\frac{1}{7} \sum_{j=1}^{7} c_{ij} \right). \quad (2)$$

5. Experiment Design

5.1. Dataset and Preprocessing

We use a large de-identified dataset of medical dialogues previously adopted in AMIE (Tu et al., 2025). The corpus includes 2,000 U.S. clinical visit transcripts spanning various specialties and conditions (Figure 2a). For this study, we restricted to primary care, family medicine, and internal medicine, yielding 899 cases. After filtering the top and bottom 5% by length, 810 remained, with an average of 1,639 words and 150 turns. From these, we sampled 80 diverse encounters based on their quality with complete patient records as our evaluation set.

5.2. Human Evaluation Design

We conducted a structured human evaluation with six practicing clinicians, focusing on how generated questions could support evidence-based medicine in real consultations. To ensure clarity and consistency, we worked with an experienced collaborator to draft detailed rater instructions (Figure 2c, Appendix B). These instructions emphasized that the system acts as an "evidence-based guideline assistant," raising questions that clinicians might normally consult guidelines to answer.

Each evaluation consisted of a patient health record, a truncated dialogue ($r \in \{0.3, 0.7, 1.0\}$), and a set of three generated questions (Figure 2d). For every case, clinicians compared two model variants (zero-shot baseline and multi-stage framework), producing six total evaluations per case. Annotators assessed each set on five dimensions using a 7-point Likert scale: *Relevance*, *Guideline Navigation*, *Thought Alignment*, *Non-Redundancy*, and *Usefulness* (Table 1). These criteria were chosen after iterative discussions to reflect real-world value for PCPs. Beyond rating, annotators selected the single most useful question or indicated that no question was necessary, with the option to propose an alternative.

Metric	**Statement**
Relevance	The questions are relevant and highlight insightful aspects of the case.
Guideline Navigation	The questions guide me toward which specific guidelines for evidence I should consult.
Thought Alignment	The questions align with my own clinical reasoning or thought process, without challenging my judgment.
Non-Redundancy	The questions are not redundant and will not impose a cognitive burden during the visit.
Usefulness	I would find these questions genuinely helpful for saving time and improving my daily workflow.

Table 1: **Five metrics for evaluation.**

The evaluation was run as a survey over 80 cases[3]. Each case had a fillable questionnaire containing the patient health record, the truncated dialogue, and both model outputs. Clinicians used a central progress list linking to each questionnaire, which allowed annotators and the study team to track completion. This survey format minimized technical friction; pilot timing was 25 to 30 minutes per case, for a total of more than 90 hours of expert review.

We also supplied the same rater instructions to Gemini 2.5 Pro and used it as an automated evaluator on the same cases; we later compare these scores with the human ratings.

6. Results

We present our main findings below.

The AI System is perceived as a valuable tool by experienced clinicians. We find that experienced clinicians perceive our generated questions valuable in the context of routine patient visits. Across an evaluation of 80 cases, the proposed multi-stage reasoning framework received an average overall score of 5.63 on a 1-7 Likert scale, while the zero-shot baseline scored 5.54 (Figure 3a), which is a surprisingly strong result for an out-of-the-box LLM. On *Usefulness* specifically, scores are 5.70 (proposed) and 5.65 (zero-shot). These scores are both well above the mid-point, which shows clinicians mostly agree the system is helpful and valuable (Figure 6). We also measured the frequency with which clinicians commented *"no question needed in the given context"* or *"I don't find these questions useful"*. Across 1,440 data samples collected, fewer than 2% received such comments, a very low rate of outright rejection. This quantitative reception is mirrored in the feedback, as one participating physician (P5) summarized their experience by noting that the AI's questions were *"generally relevant and appropriate"*.

High-quality support is delivered consistently, even with partial context. We analyzed human expert ratings of question quality at 30%, 70%, and 100% of the dialogue (Figure 4). Scores are remarkably stable: a one way ANOVA finds no significant differences across metrics or stages (all p-value > 0.05). This indicates our approach generates consistently high quality questions across varying context lengths. Notably, generations at 30% context receive slightly higher scores across all five metrics, which aligns with pilot feedback that clinically decisive information often appears early and later turns can be redundant or noisy. This suggests the system surfaces nontrivial prompts even with limited context and is well suited for real time, ambient use throughout the visit rather than only post-hoc analysis tool.

Multi-stage reasoning can enhance clinical safety and guideline alignment. In high-stakes clinical environments, average performance is insufficient; reliability and safety are paramount. Our qualitative feedback underscored this point, as clinicians were highly sensitive to irrelevant or inaccurate suggestions. One physician (P5) noted that such prompts can "take up valuable time in a visit," while another (P1) found unclear recommendations to be a significant barrier. These "hallucinated" or poorly supported suggestions pose a direct risk to clinical

[3] The clinician evaluation interface is shown in Appendix C.

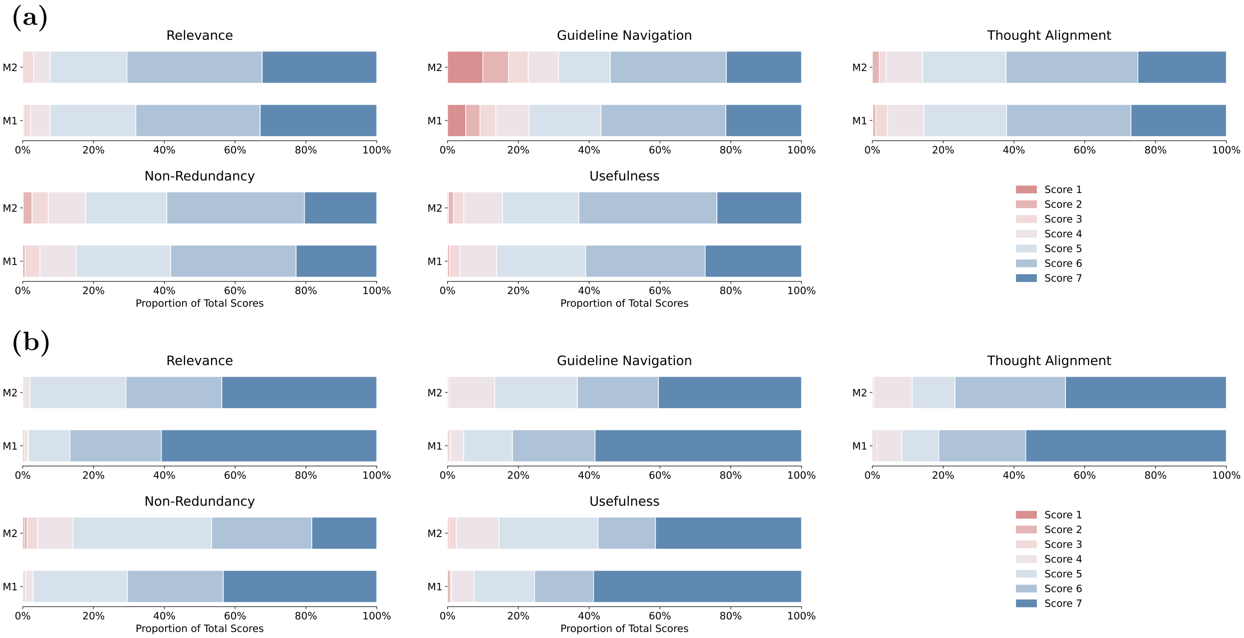


Figure 3: **Stacked bar plots of (a) the human evaluation by six clinicians and (b) the automated evaluation by Gemini on 80 dialogue transcripts.** M1 refers to the multi-stage reasoning method and M2 refers to the zero-shot baseline. Each method is rated on a Likert-scale from 1 – Strongly Disagree to 7 – Strongly Agree.

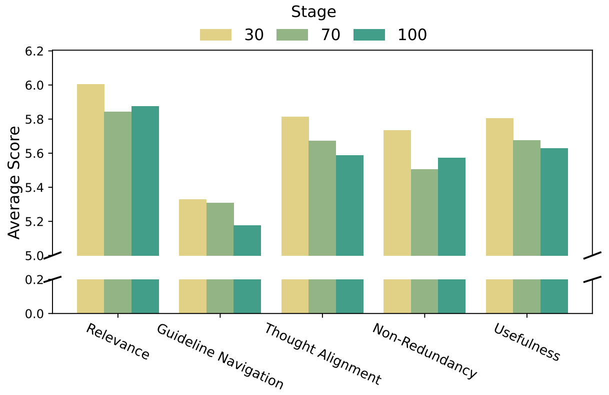


Figure 4: **Averaged scores of the proposed framework rated by PCPs at 30%, 70%, and 100% dialogue context.**

utility and user trust. The zero-shot baseline produced low-quality or unsupported guideline citations (scores of 1-2 on Guideline Navigation) in 17.22% of its generations versus 9.17% for the proposed reasoning framework. Beyond this improvement, our method also yields small but consistent gains on four of five dimensions: Guideline Navigation (+6.72%), Non-Redundancy (+1.51%), Usefulness (+0.98%), and Thought Alignment (+0.32%). Relevance is essentially unchanged (–0.23%), as shown in Figure 3a. Based on these findings, we conclude the multi-stage reasoning framework, while more complex than a zero-shot approach, is a valuable and preferable strategy for improving the reliability and safety in clinically-oriented tasks.

Clinicians preferences for question types evolve during the encounter. As mentioned in Section 4.3, the questions generated through our multi-stage reasoning framework fall into six distinct categories and offer diverse types of support for physicians. We compute each proportion as the number of times clinicians marked a question of that type as the “best question” divided by the total number of valid clinician annotations at that context level. In Figure 5, the empirical pattern can be summarized in two points: (1) Across contexts clinicians show a persistent preference for questions that directly support management decisions, primarily medication adjustment and ordering tests. These categories dominate the distribution in both sparse and complete contexts (medication adjustment consistently above 25%). (2) A stage-dependent shift centered at 70% context. At this stage, the dialogue introduces richer details, which prompts the clinician to shift focus toward specific follow-up actions (from 6% to 15%) and to begin forming preliminary judgments and diagnostic reasoning (from 8% to 15%), as clinicians seek to resolve residual uncertainty. Once the full context is present, preference returns to management-focused questions, and follow-up queries decline.

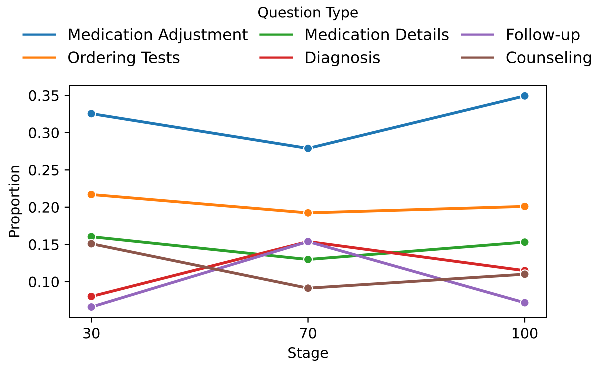


Figure 5: **Proportional trend of question types preferred by PCPs at 30%, 70%, and 100% dialogue context.** Only questions generated by the proposed framework are included. The proportions may be biased by an uneven initial distribution of question types.

LLM-as-judge reveals useful signal but with clear limitations. We want to study whether there is an alternative or less expensive solution to reliably replicate clinicians' evaluative priorities and safety judgments when assessing guideline-oriented questions.

We instruct Gemini-2.5 Pro to evaluate two sets of our generated questions, following the exact same setting. We collect the results in a similar stacked bar plots on the same 80 dialogue transcripts in Figure 3b. LLM-as-judge shows our method outperforms the zero-shot baseline on all five metrics: Relevance +5.17%, Guideline Navigation +7.49%, Thought Alignment +2.87%, Non-Redundancy +11.93%, and Usefulness +7.46%. Overall, the LLM assigns higher average scores (6.28 and 5.88 for ours and zero-shot, respectively).

Human evaluators and the LLM-as-judge agree on the direction of effect: both mark the proposed framework as preferable to the zero-shot baseline. In that sense the automated judge may offer a usable *relative* comparison signal. However, the automatic scores diverge substantially from clinician ratings when examined quantitatively. See details in Appendix D and E. We highlight the following findings:

1) *Systematic optimism.* Despite directional agreement, the LLM consistently inflates the magnitude of improvement. Automatic scores are also noticeably higher than human evaluation across multiple dimensions. Such bias suggests that while the LLM is a promising rapid proxy for comparative evaluation, its absolute judgments should be interpreted with caution.

2) *Safety alignment risk.* The LLM-as-judge does not reliably detect certain classes of evidence- or guideline-related errors (*i.e.,* low-quality outputs identified as false positives) that human reviewers do.

3) *Deployment plausibility.* These observations imply a mixed deployment outlook: LLM can scale well and relatively agree with clinicians on the direction of effect, which is valuable for rapid, large-scale iteration; but its optimistic bias and hallucinations reduce its suitability as a standalone assessor for clinical safety. Human experts therefore remain the gold standard in such evidence-based, context-aware evaluation task.

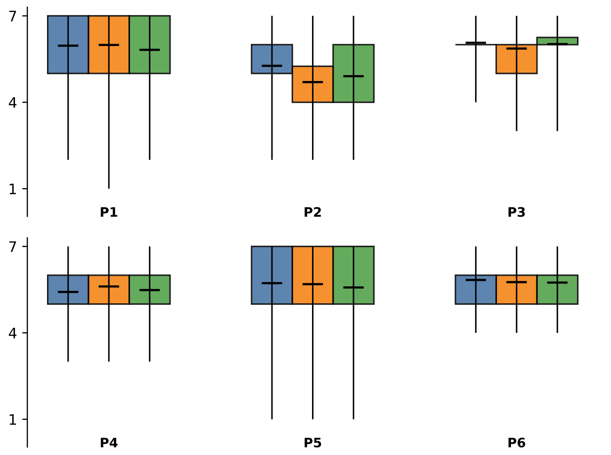


Figure 6: **Clinicians exhibit different rating styles.** On the X-axis, P1 to P6 means participated clinicians. The Y-axis, shows the range of Likert score for the evaluation result. Each group of box plots shows a clinician's ratings of the multi-stage reasoning method across 40 annotated cases, evaluated under 30%, 70%, and 100% dialogue context.

Recruiting multiple clinicians to evaluate each case is crucial because clinical experts vary despite receiving the same rater instructions and having similar practices. All of our clinicians are primary care doctors or internists with substantial experience (average years in practice: 16.5, IQR: [13, 21]) and are currently practicing. Nevertheless, their rating styles differ markedly (Figure 6): P1 and P5 show wide interquartile ranges (5 and 7) and often assign low ratings of 1 or 2, whereas P4 and P6 rarely go below 3 or 4 and concentrate their scores around the mean. These differences highlight that clinicians bring individual tendencies and backgrounds to the task, so relying on a single annotator risks bias; multiple independent evaluations are essential for robust and reliable assessment in human-in-the-loop studies.

7. Limitations

Several limitations should be considered before real-world deployment of our system.

Cost. Expert evaluation is resource-intensive; our 90 hours of physician review cost more than $10k, which is not sustainable at scale. While API calls are cheaper than in-person visits, multi-stage prompting significantly increases token usage compared to zero-shot baselines, creating trade-offs between quality and expense.

Latency. Our framework adds noticeable processing time (around 60 seconds to generate 3 questions in the multi-agent framework) due to API calls. Although acceptable offline, real-world use would also require audio transcription and integration into busy workflows. Such delays could limit the practicality of real-time ambient support.

Generalizability. Our evaluation is limited to 80 primary-care cases assessed by PCPs, so the results may not generalize to other specialties or broader case mixes. Many specialties pose different EBM challenges, for example, dermatology and radiology rely heavily on visual inputs, where text-centric LLMs may not meet without multimodal integration.

8. Discussion and Future Work

We explore whether LLMs can act as ambient assistants that surface evidence-based prompts during primary care encounters. Our multi-stage reasoning framework produces useful questions that align with guideline navigation approved by clinicians, while also revealing patterns in question-type preferences and partial alignment with LLM-as-judge. Overall, proactively surfacing guideline-relevant questions may ease cognitive load and support evidence-based reasoning in time-constrained visits, though real-world use remains limited by cost, latency, and privacy, with human experts as the gold standard for safety.

Future work can explore (1) *proactivity*, by learning when to surface questions, what types to ask, and when to abstain in order to minimize burden, and (2) *question answering*, by grounding responses in trusted guideline sources and delivering concise, actionable recommendations with provenance for safe clinical use.

References

Jill J Ashman, Loredana Santo, and Titilayo Okeyode. Characteristics of office-based physician visits by age, 2019. 2023.

Muhammad Ayaz, Muhammad F Pasha, Mohammed Y Alzahrani, Rahmat Budiarto, and Deris Stiawan. The fast health interoperability resources (fhir) standard: systematic literature review of implementations, applications, challenges and opportunities. *JMIR medical informatics*, 9(7):e21929, 2021.

J. W. Ayers, A. Poliak, M. Dredze, E. C. Leas, Z. Zhu, J. B. Kelley, D. J. Faix, A. M. Goodman, C. A. Longhurst, M. Hogarth, and D. M. Smith. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.*, 183(6):589–596, 2023. doi: 10.1001/jamainternmed.2023.1838.

David W Bates, Gilad J Kuperman, Samuel Wang, Tejal Gandhi, Anne Kittler, Lynn Volk, Cynthia Spurr, Ramin Khorasani, Milenko Tanasijevic, and Blackford Middleton. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *Journal of the American Medical Informatics Association*, 10(6): 523–530, 2003.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

G. Del Fiol, T. E. Workman, and P. N. Gorman. Clinical questions raised by clinicians at the point of care: a systematic review. *JAMA Intern. Med.*, 174 (5):710–718, 2014. doi: 10.1001/jamainternmed.2014.368.

Robert H Dolin, Liora Alschuler, Calvin Beebe, Paul V Biron, Sandra Lee Boyer, Daniel Essin, Elliot Kimber, Tom Lincoln, and John E Mattison. The hl7 clinical document architecture. *Journal of the American Medical Informatics Association*, 8 (6):552–569, 2001.

Dennis Fast, Lisa C Adams, Felix Busch, Conor Fallon, Marc Huppertz, Robert Siepmann, Philipp Prucker, Nadine Bayerl, Daniel Truhn, Marcus Makowski, et al. Autonomous medical evaluation for guideline adherence of large language models. *NPJ Digital Medicine*, 7(1):358, 2024.

Dyke Ferber, Isabella C Wiest, Georg Wölflein, Matthias P Ebert, Gernot Beutel, Jan-Niklas Eckardt, Daniel Truhn, Christoph Springfeld, Dirk Jäger, and Jakob Nikolas Kather. Gpt-4 for information retrieval and comparison of medical oncology guidelines. *Nejm Ai*, 1(6):AIcs2300235, 2024.

D. Fraile Navarro, E. Coiera, T. W. Hambly, Z. Triplett, N. Asif, A. Susanto, A. Chowdhury, A. Azcoaga Lorenzo, M. Dras, and S. Berkovsky. Expert evaluation of large language models for clinical dialogue summarization. *Sci. Rep.*, 15:1195, 2025. doi: 10.1038/s41598-024-84850-x.

A. X. Garg, N. K. Adhikari, H. McDonald, M. P. Rosas-Arellano, P. J. Devereaux, J. Beyene, J. Sam, and R. B. Haynes. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA*, 293(10):1223–1238, 2005. doi: 10.1001/jama.293.10.1223.

Joseph Gatto, Parker Seegmiller, Timothy Burdick, Inas S. Khayal, Sarah DeLozier, and Sarah M. Preum. Follow-up question generation for enhanced patient-provider conversations. In *Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, page 25222–25240, 2025. URL https://aclanthology.org/2025.acl-long.1226.pdf.

E. Goh, B. Bunning, E. C. Khoong, R. J. Gallo, A. Milstein, J. H. Chen, et al. Physician clinical decision modification and bias assessment in a randomized controlled trial of AI assistance. *Commun. Med.*, 5:59, 2025a. doi: 10.1038/s43856-025-00781-2.

E. Goh, R. J. Gallo, E. Strong, Y. Weng, H. Kerman, J. H. Chen, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nat. Med.*, 31(4):1233–1238, 2025b. doi: 10.1038/s41591-024-03456-y.

P. Hager, F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, G. Kaissis, D. Rueckert, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.*, 30:2613–2622, 2024a. doi: 10.1038/s41591-024-03097-1.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30 (9):2613–2622, 2024b.

K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ*, 330(7494):765, 2005. doi: 10.1136/bmj.38398.500764.8F.

Silvan Lange, Nils Krüger, Maximilian Warm, Johanna Buechel, Orsolya Genzel-Boroviczény, Martin R Fischer, and Konstantinos Dimitriadis. Lost in translation: Unveiling medical students' untold errors of medical history documentation. *The clinical teacher*, 21(4):e13749, 2024.

S. S. Li, V. Balachandran, S. Feng, J. Ilgen, E. Pierson, P. W. Koh, and Y. Tsvetkov. MediQ: Question-asking LLMs for adaptive and reliable clinical reasoning. *arXiv preprint arXiv:2406.00922*, 2024.

Gregor Lichtner, Claudia Spies, Carlo Jurth, Thomas Bienert, Anika Mueller, Oliver Kumpf, Vanessa Piechotta, Nicole Skoetz, Monika Nothacker, Martin Boeker, et al. Automated monitoring of adherence to evidenced-based clinical guideline recommendations: design and implementation study. *Journal of Medical Internet Research*, 25:e41177, 2023.

Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang'Anthony' Chen. Proactive conversational agents with inner thoughts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2025.

T. G. McGinn, L. McCullagh, J. Kannry, M. Knaus, A. Sofianou, J. P. Wisnivesky, and D. M. Mann. Efficacy of an evidence-based clinical decision support in primary care practices: a randomized clinical trial. *JAMA Intern. Med.*, 173(17):1584–1591, 2013. doi: 10.1001/jamainternmed.2013.8980.

James Mitchell, Ed de Quincey, Charles Pantin, and Naveed Mustfa. The development of a point of care clinical guidelines mobile application following a user-centred design approach. In *International Conference on Human-Computer Interaction*, pages 294–313. Springer, 2020.

Hannah T Neprash, Alexander Everhart, Donna McAlpine, Laura Barrie Smith, Bethany Sheridan, and Dori A Cross. Measuring primary care exam length using electronic health record data. *Medical care*, 59(1):62–66, 2021.

David Oniani, Xizhi Wu, Shyam Visweswaran, Sumit Kapoor, Shravan Kooragayalu, Katelyn Polanska, and Yanshan Wang. Enhancing large language models for clinical decision support by incorporating clinical practice guidelines. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 694–702. IEEE, 2024.

T. Østbye, K. S. H. Yarnall, K. M. Krause, K. I. Pollak, M. Gradison, and J. L. Michener. Is there time for management of patients with chronic diseases in primary care? *Ann. Fam. Med.*, 3(3):209–214, 2005. doi: 10.1370/afm.285.

Dimitra Panteli, Helena Legido-Quigley, Christoph Reichebner, Günter Ollenschläger, Corinna Schäfer, and Reinhard Busse. Clinical practice guidelines as a quality strategy. *Improving healthcare quality in Europe*, page 233, 2019.

S. M. Petterson, W. R. Liaw, R. L. Phillips, D. L. Rabin, D. S. Meyers, and A. W. Bazemore. Projecting us primary care physician workforce needs: 2010-2025. *Ann. Fam. Med.*, 10(6):503–509, 2012. doi: 10.1370/afm.1461.

V Podder, V Lew, and S Ghassemzadeh. Soap notes.[updated 2021 sep 2]. *StatPearls [Internet]. StatPearls Publishing. Available from: https://www. ncbi. nlm. nih. gov/books/NBK482263*, 2021.

Neha Pondicherry, Hope Schwartz, Nicholas Stark, Jaskirat Dhanoa, David Emanuels, Malini Singh, and Christopher R Peabody. Designing clinical guidelines that improve access and satisfaction in the emergency department. *JACEP Open*, 4(2): e12919, 2023.

J. Porter, C. Boyd, M. R. Skandari, and N. Laiteerapong. Revisiting the time needed to provide adult primary care. *J. Gen. Intern. Med.*, 38(1):147–155, 2023. doi: 10.1007/s11606-022-07707-x.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Khaled Saab, Jan Freyberg, Chunjong Park, Tim Strother, Yong Cheng, Wei-Hung Weng, David GT Barrett, David Stutz, Nenad Tomasev, Anil Palepu, et al. Advancing conversational diagnostic ai with multimodal reasoning. *arXiv preprint arXiv:2505.04653*, 2025.

Elham Sagheb, Chung-Il Wi, Jungwon Yoon, Hee Yun Seol, Pragya Shrestha, Euijung Ryu, Miguel Park, Barbara Yawn, Hongfang Liu, Jason Homme, et al. Artificial intelligence assesses clinicians' adherence to asthma guidelines using electronic health records. *The Journal of Allergy and Clinical Immunology: In Practice*, 10(4):1047–1056, 2022.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Y. Shi, S. Xu, T. Yang, Z. Liu, T. Liu, Q. Li, X. Li, and N. Liu. MKRAG: Medical knowledge retrieval augmented generation for medical question answering. *arXiv preprint arXiv:2309.16035*, 2024.

K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, V. Natarajan, S. Azizi, A. Karthikesalingam, Y. Liu, et al. Toward expert-level medical question answering with large language models. *Nat. Med.*, 31(5):943–950, 2025. doi: 10.1038/s41591-024-03423-7.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and

Karen I Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17, 2020.

T. Tu, M. Schaekermann, A. Palepu, K. Saab, J. Freyberg, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642:442–450, 2025. doi: 10.1038/s41586-025-08866-7.

D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C. P. Langlotz, J. Hom, S. Gatidis, J. Pauly, and A. S. Chaudhari. Clinical text summarization: adapting large language models can outperform human experts. *Nat. Med.*, 30(4):1134–1142, 2024. doi: 10.1038/s41591-024-02855-5.

B. Vasey, S. Ursprung, B. Beddoe, P. Taylor, N. Marlow, G. Bilbro, P. Watkinson, and P. McCulloch. Association of clinician diagnostic performance with machine learning-based decision support systems: a systematic review. *JAMA Netw. Open*, 4 (3):e211474, 2021. doi: 10.1001/jamanetworkopen.2021.1474.

Elahe Vedadi, David Barrett, Natalie Harris, Ellery Wulczyn, Shashir Reddy, Roma Ruparel, Mike Schaekermann, Tim Strother, Ryutaro Tanno, Yash Sharma, et al. Towards physician-centered oversight of conversational diagnostic ai. *arXiv preprint arXiv:2507.15743*, 2025.

Ziyu Wang, Hao Li, Di Huang, Hye-Sung Kim, Chae-Won Shin, and Amir M. Rahmani. Healthq: Unveiling questioning capabilities of llm chains in healthcare conversations. *Smart Health*, 36:100570, 2025. doi: 10.1016/j.smhl.2025.100570.

Yuzhang Xie, Hejie Cui, Ziyang Zhang, Jiaying Lu, Kai Shu, Fadi Nahab, Xiao Hu, and Carl Yang. Kerap: A knowledge-enhanced reasoning approach for accurate zero-shot diagnosis prediction using multi-agent llms. *arXiv preprint arXiv:2507.02773*, 2025.

K. S. H. Yarnall, T. Østbye, K. M. Krause, K. I. Pollak, M. Gradison, and J. L. Michener. Primary care: Is there enough time for prevention? *Am. J. Public Health*, 93(4):635–641, 2003. doi: 10.2105/AJPH.93.4.635.

S. Zwolsman, E. te Pas, L. Hooft, M. Wieringa-de Waard, and N. van Dijk. Barriers to gps' use of evidence-based medicine: a systematic review. *Br. J. Gen. Pract.*, 62(600):e511–e521, 2012. doi: 10.3399/bjgp12X652382.

Appendix A. Prompt Template

We provide our prompt templates described in Section 4. Table 2, 3, 4 are the prompts for three agents in the multi-stage reasoning framework. Table 5 is the prompt for the zero-shot setting.

Appendix B. Human Evaluation Instructions for Clinician Raters

Background. This evaluation focuses on an Evidence-Based Medical Guideline Agent, an AI tool designed to raise useful, insightful, and reference-oriented questions that reflect the problems physicians face when consulting medical guidelines. The purpose of developing this agent is to assist primary care physicians during their outpatient visits. In these brief encounters, physicians are expected to make informed decisions and ensure their practice aligns with current medical guidelines.

Task. The task is to evaluate the questions generated by different AI agents based on real-world physician-patient dialogues. Assuming you are the physician in the dialogue, you will need to evaluate two sets of questions generated by two AI agents for each dialogue. The same case will be evaluated under three dialogues of varying lengths. You will assess each set of questions against a rubric built on five key dimensions: Relevance, Guideline Navigation, Thought Alignment, Non-Redundancy, and Usefulness.

You should first begin by carefully reading the provided dialogue context and patient's questionnaire. Next, you will be provided with a set of three distinct questions. For this set of questions as a whole, you will use a 7-point Likert scale to score each metric. You are required to select the question that you believe is the most appropriate and beneficial within the given context. You can: (1) Choose a preferred question, (2) Select none, if you don't need questions under such context, (3) Provide custom question / question type / any idea (optional).

Appendix C. Evaluation Interface for Human Evaluation

We used a simple web based survey workflow. A central progress tracker listed the 80 cases, one row per case, with the assigned clinician, a completion checkbox, and a link to the case questionnaire (Figure 7). Both annotators and the study team could see progress and access the corresponding questionnaire from this list.

Each case questionnaire had six pages. The six pages corresponded to two model variants at three dialogue context levels (Figure 8). Raters saw the 30 percent context twice, the 70 percent context twice, and the 100 percent context twice. Each page began with the patient health record and the relevant portion of the patient clinician dialogue. When content had already been shown at an earlier page, a clear indicator stated that the patient record or the earlier dialogue had been presented, and a jump marker allowed the rater to scroll directly to the new content for that stage (Figure 8).

Below the context, each page displayed three candidate questions labeled A, B, and C. Raters completed five seven point Likert items aligned with our metrics of Relevance, Guideline Navigation, Thought Alignment, Non Redundancy, and Usefulness. A final item asked the rater to select the single most useful question, with options A, B, C, no question needed, or Other. Choosing Other opened a free text field where raters could propose an alternative question or leave comments.

To mitigate order effects, approximately half of the cases presented the two model variants in reversed order at each context level. The interface kept wording and layout identical across pages and cases, and required a response for each item before advancing, which reduced missing data and kept the rating experience consistent across annotators.

Appendix D. LLM-as-judge Results using Open-source Model MedGemma-27B

We additionally report results from the state-of-the-art open-source medical model MedGemma-27B (text-only) (Sellergren et al., 2025) using the same automatic evaluation pipeline used for Gemini 2.5 Pro in Table 6. Due to resource constraints, we do not include a full human evaluation of MedGemma. Therefore, Table 6 is not to benchmark model performance, but to demonstrate that our question generation pipeline is model-agnostic and can be reproduced with open-source or emerging models by the community.

Appendix E. Correlation Analysis

To quantify the alignment between automatic LLM-based scoring and human evaluation, we compute Spearman correlations between the LLM-as-judge scores and human physician ratings for all five metrics in Table 7. The correlations are consistently weak (ρ

Prompt: You are an expert medical assistant specializing in clinical documentation. Your task is to analyze the provided doctor-patient dialogue and transform it into a structured, clinically relevant summary in JSON format. Base your summary ONLY on the information present in the dialogue. Do not infer conditions or details that are not explicitly stated or directly implied. Use concise, clear, professional language. Specific numerical information in the dialogue must be fully recorded, such as medication dosage, BMI, age, etc.

Analyze the dialogue provided below in the <dialogue> section and the EHR from the <questionnaire> section. Populate the following JSON structure based on these field definitions:

- chiefComplaint: The primary reason for the visit, as stated by the patient. Keep it brief.

- historyOfPresentIllness: A detailed narrative of the chief complaint. Include descriptions of symptoms (quality, timing, severity), pertinent negatives (symbols the patient denies having), and any other related complaints explored during the visit.

- medicationHistory: List all medications mentioned, including the ones in the conversation and patient's past history. Note their reported effectiveness if mentioned.

- personalHistory: List known chronic conditions (Past Medical History) and key lifestyle factors (Social History) like smoking status.

- objectiveFindings: List all objective, observable findings from the doctor's physical exam and any mentioned vitals.

- doctorsAssessment: The final diagnoses or clinical conclusions reached by the doctor.

- plan: The course of action decided upon. Include any new prescriptions, tests ordered, and patient education provided in the dialogue.

<DIALOGUE>content</DIALOGUE>

<QUESTIONNAIRE>questionnaire</QUESTIONNAIRE>

Table 2: Prompt for the summarization agent.

Prompt: You are an experienced primary care physician shadowing a peer during their patient visit. Your primary role is to be helpful and supportive, using your clinical experience as a 'second set of eyes' on the case. The purpose is to generate search/reference-oriented questions that reflect the problems physicians face when consulting medical guidelines.

Your task is to analyze the provided clinical summary and generate exactly 10 distinct questions. The questions must be grounded in the details of the provided context. The questions should cover the following categories:

* Medication Adjustment
* Ordering Tests
* Medication Details (e.g., dosage, use, adverse effects)
* Diagnosis (e.g., differential diagnosis)
* Follow-up
* Counseling

Instructions:

- Purpose: The goal is to create questions that are search/reference-oriented. They should be the kind of questions a doctor would use to query a clinical evidence database or guideline repository.

- Structure and Style: Each of the 10 questions you generate must be a detailed, single-paragraph clinical vignette. Frame them from the professional perspective of a physician and use standard medical terminology.

- Embed Context: Each question must contain all the necessary context for it to be a standalone query. An agent answering the question should not need the original summary. Follow this structure within the question:

  1. Patient Profile: (e.g., "A 62-year-old male...")
  2. Relevant History & Diagnoses: (e.g., "...with hypertension and a history of gout...")
  3. Current Clinical Status: (e.g., "...on Lisinopril 20mg daily...")
  4. The Core Guideline Question: (e.g., "according to the 2017 ACC/AHA guidelines")

- Formatting: Do not include the question category (e.g., "Medication Adjustment", "Ordering Tests") in the question text itself. Generate only a numbered list of questions. Do not provide answers. Each question must be a single, focused query. Do not chain multiple questions together using conjunctions like 'and' or 'or'. Ensure each string in the list poses one, and only one, question.

- Quantity: Ensure there are exactly 10 questions generated from the single summary provided.

- Tone and Persona: Frame each question as if you are a helpful, collaborative clinical partner. Your goal is to gently remind or prompt for deeper thinking, not to test the user. Use collaborative phrasing where appropriate. The tone should be supportive and aim to reduce, not increase, the user's cognitive load.

EXAMPLE 1: A 6-year-old child has had persistent coryza, mild cough, and low-grade fever for 12 days. What, if any, diagnostic tests are recommended by guidelines at this point for an uncomplicated but prolonged URI in a child, before considering empiric antibiotics for suspected sinusitis?

EXAMPLE 2: A 58-year-old female with Type 2 Diabetes, hypertension, and hyperlipidemia is currently on metformin 1000mg twice daily, lisinopril 10mg daily, and atorvastatin 40mg daily. Her latest A1c is 8.5%, and her LDL-C is 110 mg/dL. As per the latest ADA guidelines, what is the recommended second-line agent to add for glycemic control, considering her cardiovascular risk factors?

IMPORTANT NOTE: The two examples above are for style and format reference only. The clinical details within them are not related to the actual case provided below. Your task is to generate 10 new questions based only on the following summary.

<SUMMARY>summary</SUMMARY>

Table 3: Prompt for the question raising agent.

Prompt: You are a meticulous AI Quality Assurance specialist. Your task is to evaluate a batch of candidate questions against a provided clinical dialogue summary. For each question, you will first reason about its pros and cons, and then assign a score for each of the seven distinct criteria, finally calculating an average score.

Instructions:

1. Process Sequentially: Evaluate each of the questions in the 'candidateQuestions' list independently.
2. Reason Before Rating: For each question, you must first perform the positive and negative reasoning steps before assigning scores.
3. Adhere to the Scale: You MUST score each criterion on a scale of 1.0 to 5.0, using increments of 0.5.
4. Strictly Follow the Output Format: Your final output must be a single, valid JSON object.

<SUMMARY>summary</SUMMARY>

<QUESTIONS>generated_questions</QUESTIONS>

For EACH question in the 'candidateQuestions' list, perform the following steps in order:

1. Positive Assessment (Pros): First, reason about why the question is good and useful for a doctor. From the seven criteria below, select the top two that best argue FOR the question's quality and briefly state why.

2. Negative Assessment (Cons): Second, reason about why the question is not good enough or could be improved. From the seven criteria below, select the top two that best argue AGAINST the question's quality and briefly state why.

3. Detailed Scoring: Third, based on your balanced assessment from the steps above, assign a score from 1.0 to 5.0 for ALL seven of the following criteria.

* relevance: How relevant and important is the question to the patient's chief complaint and core clinical problems identified in the summary?

* expectedImpact: How likely is the question to save the doctor time, uncover critical missing information, or directly improve the quality of the clinical decision?

* originality: Is the question redundant? Does it ask something already answered or made obvious by the summary?.

* factualAccuracy: Is the premise of the question factually consistent with the summary?

* comprehensiveness: Is the question itself well-formed and complete enough to be answerable, or is it too vague?

* clarityAndConciseness: Is the question clear, professional, and easy to understand?

* collaborativeTone: How well is the question framed to be supportive, non-confrontational, and respectful of the physician's expertise? (A low score means the question sounds like a test or a command).

4. Mean Score Calculation: Finally, calculate the 'meanScore' by averaging the seven individual scores from Step 3. Round the result to two decimal places.

5. Format the Output: Consolidate all the information for the question into a single JSON object as specified below and add it to the 'evaluationResults' list. Repeat for all questions.

OUTPUT FORMAT:

Your entire output must be a single JSON object. Do not add any text before or after it.

Table 4: Prompt for the question evaluation agent.

Prompt: You are an experienced primary care physician shadowing a peer during their patient visit. Your primary role is to be helpful and supportive, using your clinical experience to provide useful and meaningful questions. Your questions should be search/reference-oriented questions that reflect the problems physicians face when consulting medical guidelines.

Please analyze the provided clinical dialogue (given under <DIALOGUE> and <QUESTIONNAIRE> section) and generate exactly 3 distinct questions. The questions must be grounded in the details of the provided dialogue. Each question should be a standalone, context-rich clinical vignette. Each question must be a single, focused query. Do not chain multiple questions together using conjunctions like 'and' or 'or'. Ensure each string in the list poses only one question. Generate only a numbered list of questions. Do not provide answers.

<DIALOGUE>content</DIALOGUE>

<QUESTIONNAIRE>questionnaire</QUESTIONNAIRE>

OUTPUT:

Your entire output must be a single, valid JSON object with no additional text before or after it. The JSON object must contain a single key, "questions", with a list of exactly three strings.

Table 5: Prompt for the zero-shot baseline setting.

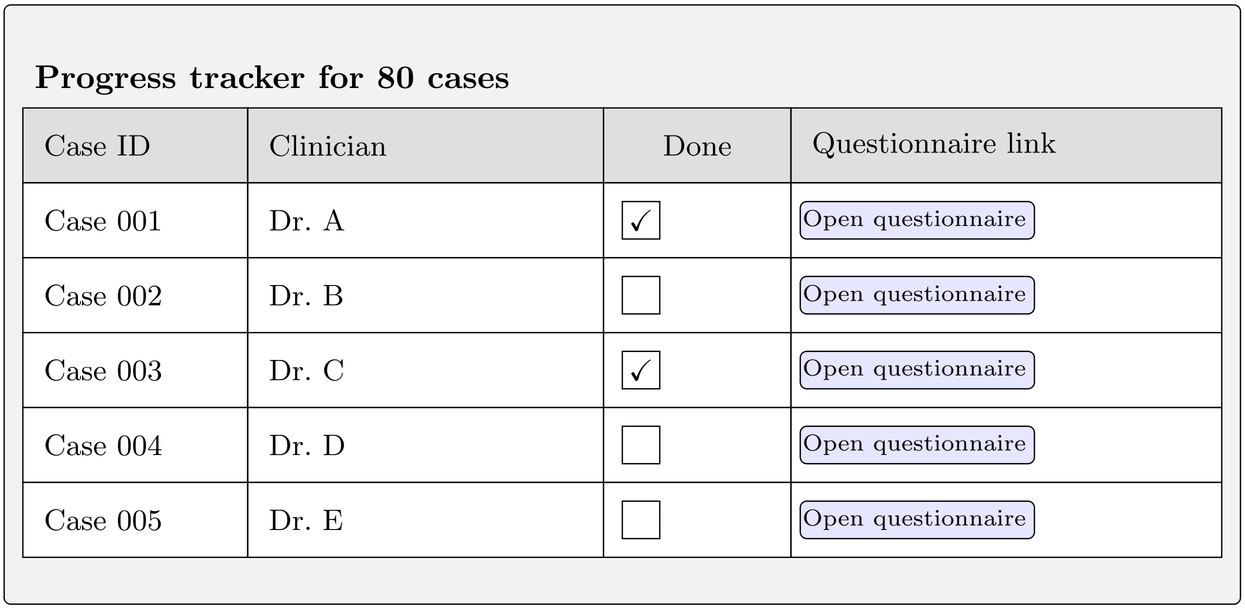
Table 6: LLM-as-judge evaluation results across five dimensions by MedGemma-27B (text-only).

Metric	Zero-shot	Multi-stage
Average	6.36	6.08
Relevance	6.6	6.2
Guideline Navigation	6.8	5.8
Thought Alignment	6.4	6.2
Non-Redundancy	5.6	6.2
Usefulness	6.4	6.0

close to 0) and mostly non-significant across both the zero-shot baseline (M2) and the multi-stage pipeline (M1). When averaged across all stages and methods (Table 8), the correlations remain near zero. It indicates that LLMs do not reliably approximate clinician judgment in this setting. This analysis further supports our decision to treat LLM-as-judge only as a supplementary signal rather than a replacement for human evaluation.

Table 7: Spearman correlation between LLM-as-judge scores and human physician ratings across five evaluation metrics, reported separately for the zero-shot baseline (M2) and the proposed multi-stage pipeline (M1).

Metric	ρ	p
Relevance-M1	0.0718	0.2677
Relevance-M2	0.0519	0.4237
Guideline Navigation-M1	0.1190	0.0657
Guideline Navigation-M2	0.0212	0.7439
Thought Alignment-M1	0.0804	0.2145
Thought Alignment-M2	-0.0335	0.6055
Non-Redundancy-M1	-0.0060	0.9262
Non-Redundancy-M2	-0.0363	0.5759
Usefulness-M1	0.0722	0.2652
Usefulness-M2	-0.0401	0.5367



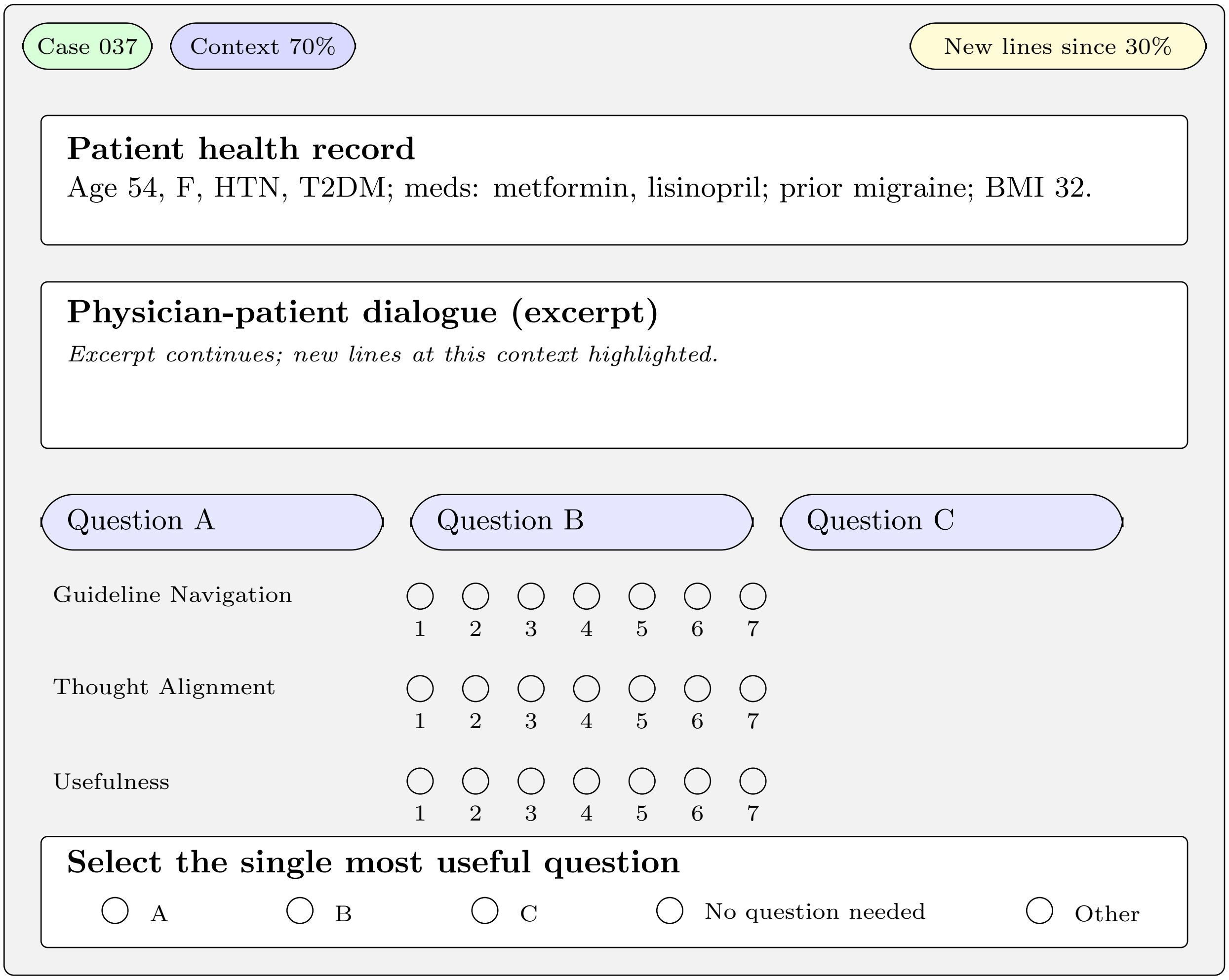
Progress tracker for 80 cases

Case ID	Clinician	Done	Questionnaire link
Case 001	Dr. A	☑	Open questionnaire
Case 002	Dr. B	□	Open questionnaire
Case 003	Dr. C	☑	Open questionnaire
Case 004	Dr. D	□	Open questionnaire
Case 005	Dr. E	□	Open questionnaire

Figure 7: Progress list with one row per case, a completion checkbox, and a link to the questionnaire.

Table 8: Average Spearman correlation between LLM-as-judge scores and physician ratings across five evaluation metrics, computed by aggregating correlations over all dialogue stages and both generation methods.

Metric	Average Spearman Correlation
Relevance	-0.013329
Guideline Navigation	-0.025142
Thought Alignment	0.042234
Non-Redundancy	-0.048712
Usefulness	0.014104



Case 037 Context 70% New lines since 30%

Patient health record

Age 54, F, HTN, T2DM; meds: metformin, lisinopril; prior migraine; BMI 32.

Physician-patient dialogue (excerpt)

Excerpt continues; new lines at this context highlighted.

Question A Question B Question C

Guideline Navigation	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Thought Alignment	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Usefulness	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7

Select the single most useful question

○ A ○ B ○ C ○ No question needed ○ Other

Figure 8: Sample questionnaire page at a given context level with three candidate questions, Likert items, and a best question choice.