# SEMA-RAG: A Self-Evolving Multi-Agent Retrieval-Augmented Generation Framework for Medical Reasoning

**Yongfeng Huang**[1]* **Ruiying Chen**[2]* **James Cheng**[1]†
[1]CSE, The Chinese University of Hong Kong [2]Wuhan University of Technology
{yfhuang22,jcheng}@cse.cuhk.edu.hk 355227@whut.edu.cn

## Abstract

Retrieval-Augmented Generation (RAG) is widely employed to mitigate risks such as hallucinations and knowledge obsolescence in medical question answering, yet its predominantly single-round, static retrieval paradigm misaligns with the multi-stage process of clinical reasoning. This compressed workflow induces two structural deficiencies: question-to-query translation often lacks clinically grounded semantic interpretation, and retrieval lacks iterative sufficiency feedback, making it difficult to form reliable evidence chains. We argue that both issues stem from a deeper cause—overloading a single reasoning chain with heterogeneous tasks of interpretation, exploration, and adjudication—and that the remedy is to reconstruct the workflow via task decoupling and dynamic multi-round exploration. To this end, we propose **SEMA-RAG**, a **S**elf-**E**volving **M**ulti-**A**gent RAG framework for medical question answering, which assigns these roles to three specialist agents: the **Interpreter Agent** for clinical schema interpretation, the **Explorer Agent** for sufficiency-driven self-evolving retrieval, and the **Arbiter Agent** for evidence adjudication and answer selection. Across five benchmarks and five LLM backbones, SEMA-RAG improves the strongest baseline by **+6.46** accuracy points on average, measured per backbone.

## 1 Introduction

In recent years, large language models (LLMs) have shown specific capabilities in understanding and reasoning about medical knowledge when applied in healthcare (Kung et al., 2023; Omar et al., 2024). However, they remain prone to hallucinations and outdated information in high-stakes clinical settings (Omiye et al., 2024; Roustan and Bastardot, 2025). Retrieval-Augmented Generation (RAG), which incorporates external authoritative evidence to support the generation process, has been widely adopted to mitigate these risks (Lewis et al., 2020).

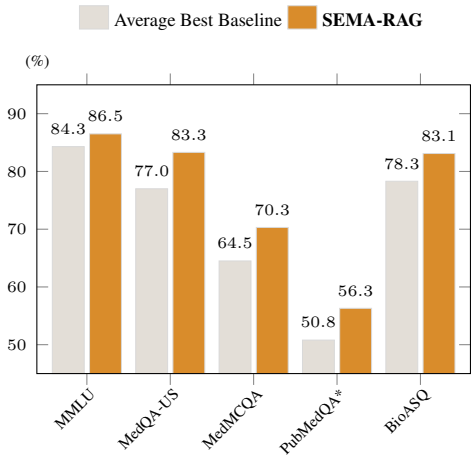


Figure 1: Benchmark-level accuracy averaged over five LLM backbones.

However, standard RAG frameworks typically treat retrieval as a static, single-round auxiliary step, misaligning with the multi-stage process of clinical reasoning: clinicians often first interpret patient narratives as searchable clinical questions, then progressively gather and verify information to address evidence gaps, and weigh and integrate redundant or contradictory evidence to ultimately form judgments based on relatively robust evidence (Linn et al., 2012; Yazdani et al., 2017). In contrast, single-round static RAG compresses this process into a single retrieval and generation step. This is akin to requiring clinicians to simultaneously analyze, retrieve, evaluate, and diagnose immediately upon receiving initial medical records, without adjusting their reasoning as new evidence emerges. This typically leads to two structural flaws: *(i)* **the translation from question to query lacks clinical semantic interpretation**, making implicit constraints difficult to articulate explicitly (Soldaini

*These authors contributed equally to this work.
†Corresponding author.



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 18423–18442

et al., 2017); and *(ii)* **the retrieval process lacks mechanisms for sufficiency assessment and feedback**, hindering self-evolving iterative convergence under insufficient evidence and thus weakening reliable evidence-chain formation (Mallen et al., 2023; Shi et al., 2023).

These shortcomings, we argue, are not independent issues but rather symptoms of a deeper problem: **overloading heterogeneous tasks into a single reasoning chain**. When question interpretation, evidence exploration, and answer adjudication are tightly coupled, the cognitive load increases and the steps become interdependent, making it hard for the model to promptly adjust retrieval and reasoning when evidence is insufficient or conflicting (Wang et al., 2023; Liu et al., 2024). Thus, the key is not to intensify single-round reasoning, but to restructure RAG to better match the phased clinical workflow by extending single-round queries into multi-round iterative exploration. After each retrieval round, the system evaluates whether the evidence covers key constraints and then chooses the next action: terminate exploration and proceed to decision integration if sufficient, or generate targeted follow-up queries to fill gaps if insufficient. This mechanism, which continuously updates the direction of queries and retrieval based on the evaluation results of each round, enables the system to adjust and converge as evidence accumulates progressively. In this sense, the process constitutes a form of **intra-test-time self-evolution**, in which the system adaptively updates its query and retrieval trajectory during task execution while remaining tightly coupled to the current problem instance (Gao et al., 2026). For simplicity, we use the term "self-evolving" in the remainder of the paper to refer to this intra-test-time setting.

To this end, we propose **SEMA-RAG** (**S**elf-**E**volving **M**ulti-**A**gent RAG). This framework simulates clinical workflows through task decoupling and role specialization, decomposing complex clinical reasoning into three collaborative modules: **Interpreter Agent (I-Agent)** maps unstructured inputs to structured clinical semantics; **Explorer Agent (E-Agent)** implements self-evolving, evidence-sufficiency-driven retrieval for convergent exploration; **Arbiter Agent (A-Agent)** performs comprehensive adjudication based on closed-loop evidence.

We evaluate SEMA-RAG on five medical question-answering benchmarks. As shown in Figure 1, SEMA-RAG consistently outperforms representative baselines in terms of average accuracy on each benchmark when averaged over multiple underlying LLMs. Across five benchmarks and five LLM backbones, it improves the strongest baseline by an average of **+6.46** accuracy points, validating convergent evidence-chain construction via task decoupling, role specialization, and evidence sufficiency-driven self-evolving retrieval. Our main contributions are as follows:

- We propose **SEMA-RAG**, a multi-agent RAG framework for medical question answering, which models clinical reasoning processes via role division and collaboration.
- We develop a self-evolving **Explorer Agent** that updates queries based on evidence gaps, steering retrieval toward medical reasoning objectives.
- We validate SEMA-RAG on five medical Q&A benchmarks across multiple underlying LLMs, achieving consistent improvements over baselines.

## 2 Preliminaries

### 2.1 Task Formulation of Medical RAG

Given a medical question $Q$, the system selects the final answer $\tilde{y}$ from the discrete candidate set $\mathcal{Y}$ ($y \in \mathcal{Y}$). Under question-only retrieval conditions, the system may only retrieve evidence from the medical corpus $\mathcal{C}$. The core RAG consists of a retrieval operator $\mathrm{Ret}(\cdot)$ and a generation operator: $C = \mathrm{Ret}(Q)$, and predicts accordingly:

$$\tilde{y} = \arg\max_{y \in \mathcal{Y}} p(y \mid Q, C).$$

### 2.2 Multi-Agent Roles and Abstraction

We employ role-based division of labor, with three agents collaborating to complete the medical question answering process: **I-Agent** handles question interpretation, **E-Agent** manages evidence exploration, and **A-Agent** oversees answer adjudication.

Three agents share the same underlying language model, differentiated solely by role-specific prompts. We denote the output of the shared LLM conditioned on a role prompt $Pmt_r$ and an input $X$ as $\mathrm{Agent}_r(Pmt_r, X)$, where $X$ may be a set containing multiple elements.

## 3 Method

Figure 2 illustrates the overall SEMA-RAG framework, which comprises three role-based agents with responsibilities described below.

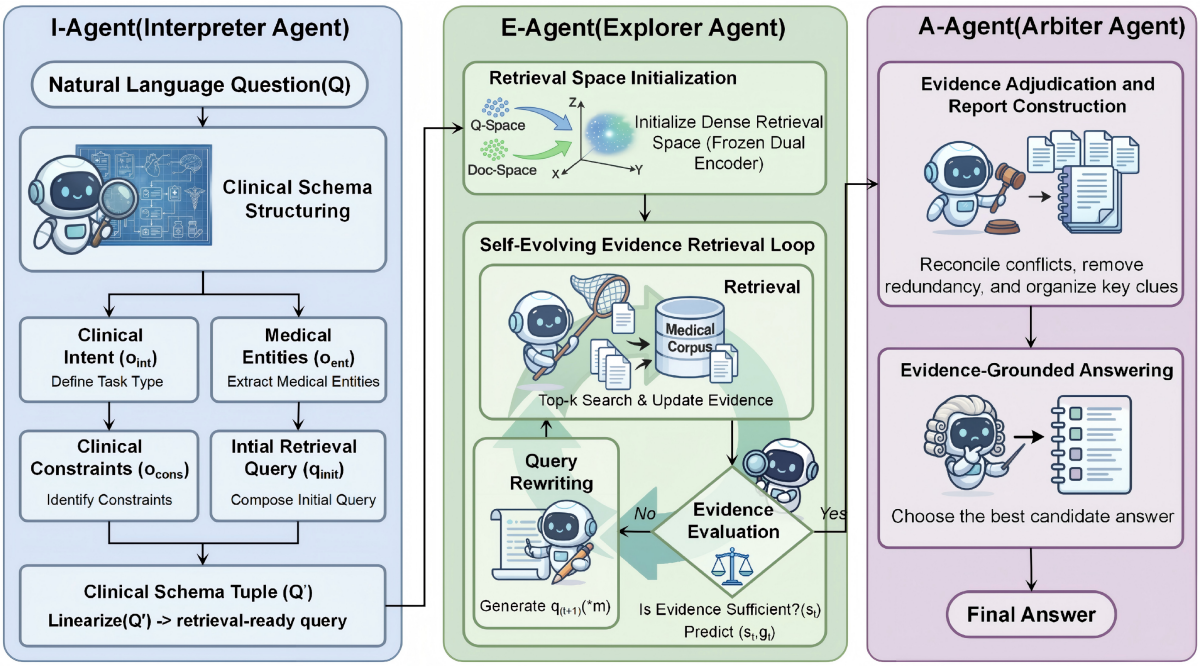


Figure 2: Overview of **SEMA-RAG**: (i) **I-Agent** structures the input question $Q$ into a clinical schema tuple $Q'$ for retrieval; (ii) **E-Agent** conducts sufficiency-driven self-evolving multi-round retrieval to obtain a converged evidence set $C^*$; (iii) **A-Agent** adjudicates evidence into a traceable report $R$ and selects the final answer grounded in $R$.

### 3.1 I-Agent as a Question Interpreter

I-Agent does not merely rephrase the input medical question $Q$; instead, it semantically structures $Q$ and projects it onto an explicit **Clinical Schema**. This process externalizes latent clinical intent and key constraints, providing stable anchors for subsequent retrieval and reasoning.

Specifically, I-Agent produces a clinical schema tuple $Q'$ with four components: *(i)* clinical intent $o_{\text{int}}$, describing the implied task type (e.g., diagnosis, treatment, dosage); *(ii)* medical entities $o_{\text{ent}}$, identifying core medical objects (e.g., diseases, drugs); *(iii)* clinical constraints $o_{\text{cons}}$, specifying applicability conditions (e.g., pregnant, renal impairment, adult); and *(iv)* an initial retrieval query $q_{\text{init}}$, a concise, search-oriented question distilled from the above schema.

Formally, I-Agent maps $Q$ to the schema tuple:

$$\begin{aligned} Q' &= \langle o_{\text{int}},\, o_{\text{ent}},\, o_{\text{cons}},\, q_{\text{init}} \rangle \\ &= \text{Agent}_{\text{I}}(Pmt_{\text{I}},\, Q). \end{aligned}$$

To make the schema tuple usable by the dense retriever, we further linearize $Q'$ into a retrieval-ready query string:

$$\begin{aligned} \hat{q}_{\text{init}} &= \text{Linearize}(Q') \\ &= \text{Concat}(q_{\text{init}}, \oplus, o_{\text{int}}, \oplus, o_{\text{ent}}, \oplus, o_{\text{cons}}), \end{aligned}$$

where $\oplus$ denotes a semicolon separator. Here, $\text{Linearize}(\cdot)$ is a parameter-free function without field-specific weights or additional control tokens. It preserves $q_{\text{init}}$ as the core query while explicitly incorporating $o_{\text{int}}, o_{\text{ent}}$, and $o_{\text{cons}}$, making clinically important but implicit constraints more visible to the retriever and reducing semantic drift in the initial retrieval stage. The resulting query $\hat{q}_{\text{init}}$ is used to initialize E-Agent, while $Q'$ remains the clinical anchor for subsequent coordination.

### 3.2 E-Agent as a Knowledge Explorer

E-Agent begins with the linearized schema query $\hat{q}_{\text{init}}$ generated by I-Agent and progressively completes the evidence through a self-evolving iterative retrieval process, ultimately constructing the final evidence set $C^*$.

**Retrieval Space Initialization** We construct a dense vector retrieval space based on the medical corpus $\mathcal{C}$. Using a parameter-frozen medical dual encoder, we map queries and documents to the same vector space, where $E_{\text{qry}}(\cdot)$ and $E_{\text{doc}}(\cdot)$ denote the query and document encoders, respectively. Given a query $q$, its Top-$k$ candidate documents (passages/chunks) are retrieved based on

vector similarity as follows:

$$\mathrm{TopK}(q) = \underset{D \in \mathcal{C}}{\text{Top-k}} \left\langle E_{\mathrm{qry}}(q),\, E_{\mathrm{doc}}(D) \right\rangle .$$

where Top-k returns the $k$ documents with the largest similarity scores.

**Self-Evolving Evidence Retrieval Loop** Using the linearized schema query $\hat{q}_{\mathrm{init}}$ as the initial query, we set $\mathcal{Q}_1 = \{\hat{q}_{\mathrm{init}}\}$ and $C_0 = \varnothing$, where $\mathcal{Q}_t$ is the query set for the $t$th retrieval round and $C_t$ is the accumulated evidence set after round $t$. Each retrieved document $D_i$ is associated with a deterministic document identifier $\mathrm{ID}(D_i)$, which is retained throughout the pipeline for exact deduplication and source tracing. At round $t$, E-Agent performs retrieval for each query in $\mathcal{Q}_t$ and updates the evidence set:

$$\mathcal{D}_t = \bigcup_{q \in \mathcal{Q}_t} \mathrm{TopK}(q),$$

$$C_t = C_{t-1} \cup \{D_i \in \mathcal{D}_t : \mathrm{ID}(D_i) \notin \mathrm{IDs}(C_{t-1})\}.$$

Conditioned on clinical anchors $Q'$, the current textual query set $\mathcal{Q}_t$, and the evidence set $C_t$, E-Agent predicts a sufficiency flag $s_t$, a gap description $g_t$, and the next query set $\mathcal{Q}_{t+1}$:

$$[s_t,\, g_t,\, \mathcal{Q}_{t+1}] = \mathrm{Agent}_{\mathrm{E}}\big(Pmt_{\mathrm{E}},\, [Q',\, \mathcal{Q}_t,\, C_t]\big)\,,$$
$$\mathcal{Q}_{t+1} = \left\{ q_{t+1}^{\langle 1 \rangle}, \ldots, q_{t+1}^{\langle m \rangle} \right\}.$$

where $s_t \in \{0, 1\}$ indicates evidence sufficiency: if $s_t = 1$, the evidence is sufficient and we set $\mathcal{Q}_{t+1} = \varnothing$; otherwise ($s_t = 0$), evidence gaps remain, and $g_t$ identifies missing conditions or reasoning steps, from which $\mathcal{Q}_{t+1}$ generates $m$ candidate follow-up queries targeting these gaps.

When $s_t = 0$, the generated $\mathcal{Q}_{t+1}$ is issued in the next round to retrieve additional evidence, and the results are incorporated into the update of $C_{t+1}$.

Iteration terminates when $s_t = 1$, $t = T_{\max}$, or stagnation occurs (i.e., $\mathcal{Q}_{t+1} = \varnothing$). Upon termination, we obtain the closed evidence set, record the actual number of iterations $T \le T_{\max}$, and store the self-evolving trajectory:

$$C^* = C_T,\ \tau = \{[\mathcal{Q}_1, C_1], \ldots, [\mathcal{Q}_T, C_T]\}\,.$$

### 3.3 A-Agent as an Evidence Arbiter

A-Agent adjudicates evidence by organizing the converged set $C^*$ into a traceable evidence report and generating a discrete answer from it.

**Evidence Adjudication and Report Construction** Given redundant and potentially conflicting evidence, A-Agent first adjudicates $C^*$ by removing irrelevant or duplicated content, identifying consistencies and conflicts, and organizing supporting and refuting clues into a structured evidence report $R$. For traceability, we retain the original document identifier for each retrieved document $D_i \in C^*$, forming the source set

$$\mathcal{S}^* = \{(\mathrm{ID}(D_i), D_i) \mid D_i \in C^*\}.$$

A-Agent then generates the evidence report

$$R = \mathrm{Agent}_{\mathrm{A}}(Pmt_{\mathrm{adj}}, [Q, C^*, \mathcal{S}^*]),$$

where $R$ explicitly organizes key conclusions relevant to the question along with their source indices, provides a reconciled synthesis of conflicting evidence, and offers a stable basis for final answer selection.

**Evidence-Grounded Answering** Upon obtaining the evidence report $R$, A-Agent performs discrete answer selection over the candidate answer set $\mathcal{Y}$:

$$\tilde{y} = \mathrm{Agent}_{\mathrm{A}}(Pmt_{\mathrm{ans}}, [Q, R]),$$

where $\tilde{y}$ is the final predicted answer.

## 4 Experiments

### 4.1 Experimental Setup

#### 4.1.1 Evaluation Benchmarks

To systematically evaluate SEMA-RAG's performance and generalisation capabilities across diverse medical question-answering scenarios, we select five widely used datasets from the MIRAGE benchmark (Xiong et al., 2024): three medical examination datasets (MMLU-Med (Hendrycks et al., 2021), MedQA-US (Jin et al., 2020), MedMCQA (Pal et al., 2022)) and two biomedical research QA datasets (PubMedQA* (Jin et al., 2019) and BioASQ-Y/N (Tsatsaronis et al., 2015; Krithara et al., 2023)). Together, they cover general medical knowledge, clinical examinations, and biomedical literature inference. Following MIRAGE's filtering and preprocessing pipeline, we retain only discrete biomedical classification questions, use PubMedQA* (with the original evidence context removed), and apply question-only retrieval for all tasks.

| Model | MMLU-Med | MedQA-US | MedMCQA | PubMedQA* | BioASQ-Y/N | Average |
|---|---|---|---|---|---|---|
| `deepseek-v3.1` (DeepSeek-AI et al., 2025) | | | | | | |
| + CoT (Wei et al., 2022) | 88.15 | 77.53 | 71.69 | 38.40 | 80.10 | 71.17 |
| + MedCPT (Jin et al., 2023) | 85.12 | 73.84 | 62.66 | 43.20 | 76.38 | 68.24 |
| + MedRAG (Xiong et al., 2024) | 88.61 | 77.14 | 67.99 | 44.60 | 78.48 | 71.36 |
| + *i*-MedRAG (Xiong et al., 2025) | 85.86 | 74.78 | 65.65 | 50.60 | 80.58 | 71.49 |
| + SEMA-RAG (Ours) | **91.46** | **89.95** | **75.09** | **59.20** | **82.85** | **79.71** |
| `kimi-k2` (Team et al., 2025b) | | | | | | |
| + CoT (Wei et al., 2022) | 84.39 | 77.85 | 72.08 | 53.60 | 85.76 | 74.74 |
| + MedCPT (Jin et al., 2023) | 89.81 | 80.68 | 73.85 | 50.20 | 81.39 | 75.19 |
| + MedRAG (Xiong et al., 2024) | 91.37 | 81.54 | 73.20 | 52.60 | 85.60 | 76.86 |
| + *i*-MedRAG (Xiong et al., 2025) | 91.28 | 81.78 | 74.13 | 54.60 | 83.17 | 76.99 |
| + SEMA-RAG (Ours) | **91.46** | **86.41** | **76.07** | **55.80** | **88.67** | **79.68** |
| `qwen3-coder-plus` (Yang et al., 2025) | | | | | | |
| + CoT (Wei et al., 2022) | 89.26 | 76.90 | 73.06 | 47.20 | 81.72 | 73.63 |
| + MedCPT (Jin et al., 2023) | 87.42 | 75.26 | 67.44 | 46.60 | 75.89 | 70.52 |
| + MedRAG (Xiong et al., 2024) | 89.44 | 81.54 | 69.26 | 49.20 | 72.33 | 72.35 |
| + *i*-MedRAG (Xiong et al., 2025) | 89.26 | 77.38 | 70.26 | 48.60 | 82.52 | 73.60 |
| + SEMA-RAG (Ours) | **92.10** | **86.17** | **74.23** | **56.00** | **83.01** | **78.30** |
| `gemini-2.0-flash` (Google, 2025) | | | | | | |
| + CoT (Wei et al., 2022) | 58.22 | 65.12 | 41.33 | 40.20 | 68.45 | 54.66 |
| + MedCPT (Jin et al., 2023) | 62.35 | 70.54 | 44.90 | 42.80 | 70.06 | 58.13 |
| + MedRAG (Xiong et al., 2024) | 74.29 | 83.19 | 50.87 | 44.20 | 72.65 | 65.04 |
| + *i*-MedRAG (Xiong et al., 2025) | 65.47 | 77.69 | 46.78 | 51.20 | 77.99 | 63.83 |
| + SEMA-RAG (Ours) | **80.99** | **90.42** | **71.60** | **59.20** | **88.19** | **78.08** |
| `glm-4.0-flash` (GLM et al., 2024) | | | | | | |
| + CoT (Wei et al., 2022) | 68.14 | 50.43 | 48.22 | 40.00 | 58.90 | 53.14 |
| + MedCPT (Jin et al., 2023) | 73.00 | 58.21 | 50.63 | 42.20 | 60.84 | 56.98 |
| + MedRAG (Xiong et al., 2024) | **77.59** | 60.88 | 52.59 | 46.80 | 62.78 | 60.13 |
| + *i*-MedRAG (Xiong et al., 2025) | 73.28 | 53.57 | 51.47 | 48.40 | 64.72 | 58.29 |
| + SEMA-RAG (Ours) | 76.68 | **63.79** | **54.34** | **51.20** | **72.98** | **63.80** |

Table 1: Accuracy (%) comparison of **SEMA-RAG** and baselines on five medical QA benchmarks across different LLMs. Bold indicates the best result within each model block and underline indicates the second-best.

#### 4.1.2 Models and Baselines

To assess SEMA-RAG's robustness across backbones, we instantiate the framework on five publicly accessible LLMs: **deepseek-v3.1** (DeepSeek-AI et al., 2025), **kimi-k2** (Team et al., 2025b), **qwen3-coder-plus** (Yang et al., 2025), **gemini-2.0-flash** (Google, 2025), and **glm-4.0-flash** (GLM et al., 2024). These models originate from different providers, encompass diverse pretraining configurations and capability focuses.

We selected three representative methods for comparison to characterize performance differences across no retrieval, single-round retrieval, and iterative retrieval. The no-retrieval setting employs **CoT** (Wei et al., 2022), relying solely on model-internal knowledge for chain-of-reasoning. The single-round retrieval setting employs **MedCPT** (Jin et al., 2023) as the medical domain retriever, further contrasting it with **MedRAG** (Xiong et al., 2024)'s retrieval-fusion framework. The iterative retrieval setting utilizes ***i*-MedRAG** (Xiong et al., 2025), which generates subsequent queries to drive multi-round retrieval and accumulate evidence.

#### 4.1.3 Implementation Details

Following *i*-MedRAG (Xiong et al., 2025) and MedRAG (Xiong et al., 2024), we retrieve from Textbooks (Jin et al., 2021) and StatPearls (StatPearls Publishing, 2025) on all benchmarks. We use MedCPT (Jin et al., 2023) as the dense retriever and perform FAISS-based retrieval over these corpora.

All methods are evaluated in a zero-shot setting. Unless stated otherwise, SEMA-RAG uses $T_{\max} = 2$, $k = 16$, and $m = 3$. We set temperature to 1.0 for I/E-Agent and 0.0 for A-Agent. Other baselines follow their official settings.

### 4.2 Main Results: Consistent and Significant Improvements

Table 1 reports results on five medical QA benchmarks across five underlying LLMs. Overall, **SEMA-RAG achieves the best average accuracy within every model block**, indicating that the improvement is backbone-agnostic rather than tied to

a particular LLM.

A consistent pattern is that the gains are larger on benchmarks where premature commitment under incomplete evidence is costly. This matches SEMA-RAG's self-evolving exploration: it checks evidence sufficiency, performs targeted follow-up retrieval when needed, and only then moves to adjudication, leading to more reliable evidence chains.

Overall, this consistent and substantial advantage is not accidental. It directly stems from SEMA-RAG's successful simulation of expert clinical reasoning by decoupling interpretation, exploration, and adjudication, thereby alleviating the cognitive overload bottleneck diagnosed in the Introduction. To unpack the source of these gains, we next present a core component analysis.

## 4.3 Core Component Analysis

In this section, we quantitatively evaluate the contributions of the three roles in SEMA-RAG via role-wise removal analysis. Specifically, we compare the full framework with three variants: (i) **w/o I-Agent**, which removes the question interpretation module and directly uses the raw question as the initial retrieval query; (ii) **w/o E-Agent**, which removes the self-evolving retrieval mechanism and performs only a single round of static retrieval based on the linearized schema query generated by I-Agent; and (iii) **w/o A-Agent**, which removes the answer adjudication module and directly generates answers from the final evidence set. The results are summarized in Table 2.

| I-Agent | E-Agent | A-Agent | MedQA-US | PubMedQA* |
|---|---|---|---|---|
| ✗ | ✓ | ✓ | 85.47 | 54.20 |
| ✓ | ✗ | ✓ | 83.58 | 50.80 |
| ✓ | ✓ | ✗ | 86.49 | 53.60 |
| ✓ | ✓ | ✓ | **89.95** | **59.20** |

Table 2: Role-wise removal results of SEMA-RAG on MedQA-US and PubMedQA* (deepseek-v3.1).

### 4.3.1 Resolving Ambiguous Queries with I-Agent

Table 2 shows that removing the I-Agent consistently degrades performance, indicating that question-to-query translation benefits from clinically grounded semantic interpretation. Without this step, queries often remain underspecified and fail to surface implicit constraints, which increases the chance that retrieval drifts toward generic, symptom-level evidence rather than the decision-critical clinical setting.

As shown in Table 4 (Step 1), I-Agent makes *hospital day 7* retrieval-actionable, anchoring evidence in the late-onset inpatient context and enabling option-level discrimination for *S. aureus*. By extracting a structured clinical schema and making such key conditions retrievable, I-Agent keeps retrieval aligned with the intended clinical scenario and provides a cleaner substrate for subsequent exploration and adjudication.

### 4.3.2 Achieving Dynamic Reasoning with E-Agent

Table 2 shows that removing E-Agent causes the largest performance drop, highlighting that the core gain comes from a self-evolving, sufficiency-driven closed-loop retrieval rather than static retrieval. Without E-Agent, the system loses explicit sufficiency feedback and is more prone to stop with partially covered evidence, leaving key constraints unresolved.

To configure this loop, we vary the maximum exploration depth $T_{\max}$ with $m = 3$. Figure 3 suggests that most of the benefit is captured within two rounds, with performance peaking around $T_{\max} \in \{2, 3\}$; beyond that, deeper exploration saturates and can introduce noise. Notably, the two-round setting already outperforms our reproduced *i*-MedRAG baseline (which uses three fixed retrieval rounds), suggesting that the gain comes from self-evolving, sufficiency-driven closed-loop exploration rather than simply increasing the number of iterations.

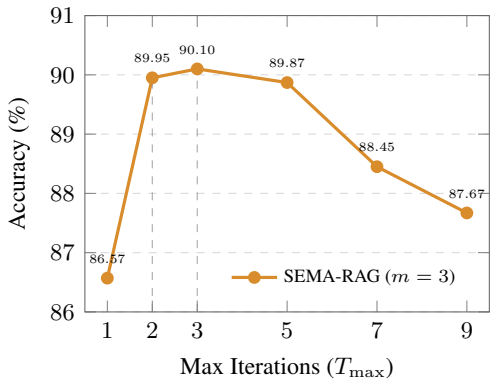


Figure 3: Impact of max iterations $T_{\max}$ (fix $m = 3$) on MedQA-US (deepseek-v3.1).

## 4.4 Further Analysis

### 4.4.1 Synergy of the Multi-Agent Architecture

Table 2 shows that the full SEMA-RAG consistently performs best across both MedQA-US and

PubMedQA*. Ablating any single agent leads to a clear drop, suggesting that the gains do not come from one isolated module but from role-specialized collaboration that mirrors the staged clinical workflow: I-Agent anchors clinically grounded interpretation, E-Agent drives sufficiency-driven evidence completion, and A-Agent consolidates and adjudicates evidence for option selection.

#### 4.4.2 Impact of Query Breadth

Building on the depth analysis in Figure 3, we examine how the per-round query breadth $m$ affects E-Agent's exploration. With $T_{\max} = 2$ fixed, Table 3 shows a clear monotonic trend: increasing $m$ improves accuracy, but the gains quickly taper. This suggests that expanding the query set helps cover complementary evidence gaps in early exploration, while additional branches beyond a moderate breadth tend to introduce overlapping or low-yield retrievals. We therefore use $T_{\max} = 2$ and $m = 3$ as the default setting in subsequent experiments, balancing coverage and efficiency.

### 4.5 Qualitative Case Study: A Head-to-Head Comparison

Table 4 shows a representative MedQA-US case comparing MedRAG and SEMA-RAG, where the decisive cue is the temporal constraint (hospital day 7), pointing to a hospital-acquired rather than community-acquired etiology.

In the *Baseline* block, MedRAG relies on a single round of static retrieval. Its evidence remains centered on generic pneumonia cues and aspiration risk, without explicitly anchoring retrieval to the hospital day 7 condition. This shifts the evidence toward typical community-acquired pathogens and leads to an incorrect selection of Streptococcus pneumoniae (Option A), which mismatches the inpatient, late-onset setting in the question.

In contrast, *SEMA-RAG* makes the temporal constraint retrieval-actionable and carries it through to option selection. The **I-Agent** surfaces hospital day 7 as a key constraint, the **E-Agent** detects the missing distinction between community- and hospital-acquired spectra and performs targeted follow-up retrieval, and the **A-Agent** consolidates the resulting evidence and maps it to the candidate set, yielding the correct choice Staphylococcus aureus (Option D). Overall, the case shows how task decoupling with sufficiency-driven self-evolving retrieval helps form a more reliable evidence chain and prevents premature decisions under insufficient evidence.

### 4.6 Cost and Efficiency Analysis

Table 5 compares methods in terms of both accuracy and inference cost. Calls denotes the number of LLM invocations per question, while Retr. counts the number of vector retrieval operations; Time reports the average end-to-end latency per question; Tok./Q denotes the average total token consumption per question. Because SEMA-RAG employs sufficiency-driven early stopping, we report per-question averages over the MedQA-US set.

| **Method** | **Calls** *(#)* | **Retr.** *(#)* | **Time** *(s)* | **Acc.** *(%)* | **Tok./Q** *(#)* |
|---|---|---|---|---|---|
| CoT | 1.0 | 0.0 | 2.5 | 77.53 | 713.7 |
| MedRAG | 1.0 | 1.0 | 3.2 | 77.14 | 2264.9 |
| *i*-MedRAG | 3.0 | 9.0 | 8.8 | 74.78 | 21516.6 |
| **SEMA-RAG** | **4.8** | **3.4** | **9.5** | **89.95** | **19488.4** |

Table 5: Efficiency comparison on MedQA-US (deepseek-v3.1).

The results reveal three main patterns. First, SEMA-RAG consistently improves decision quality over single-round baselines. Second, these gains come with a moderate overhead relative to single-pass methods, as expected from multi-agent, multi-round inference. Third, compared with the iterative baseline *i*-MedRAG, SEMA-RAG achieves a more favorable accuracy–efficiency trade-off, indicating that sufficiency-driven early stopping allocates additional computation more effectively than fixed-step iteration. Taken together, these results suggest that SEMA-RAG improves decision quality in a practically affordable regime, making the added overhead worthwhile for high-stakes medical QA.

## 5 Related Work

### 5.1 Retrieval-Augmented Medical Reasoning

Early medical LLMs systems largely depended on parametric knowledge or single-round retrieval (Luo et al., 2022; Singhal et al., 2023). Methods such as MedRAG (Xiong et al., 2024) and MedCPT (Jin et al., 2023) improve domain retrieval with medical dual encoders and RAG pipelines, yet still follow a retrieve-once-then-answer pattern, which often falls short on questions requiring multi-hop evidence integration and clinically constrained reasoning (Jiang et al., 2023; Trivedi et al., 2023).

Recent work has shifted from one-shot retrieval to iterative RAG (Gao et al., 2024; Zhao et al.,

**MedQA-US Question 0024**

**Context**: A 62-year-old patient has been hospitalized for a week due to a stroke. On hospital day 7, he develops a fever (38.4°C) and purulent cough. Vitals: HR 88, RR 20, BP 110/85. Physical exam: right basal crackles. Chest X-ray: new right-sided consolidation.
**Labs (Selected)**: WBC 8,900/mm³ (Neutrophils 72%, Bands 4%), Hb 16 g/dL, Platelets 280,000/mm³.
**Question**: What is the most likely causal microorganism?
**A.** *Streptococcus pneumoniae* **B.** *Mycobacterium tuberculosis* **C.** *Haemophilus influenzae* **D.** *Staphylococcus aureus*

**BASELINE: MEDRAG**

**Retrieved Evidence**:
"Community-acquired pneumonia (CAP) is most frequently caused by *Streptococcus pneumoniae*..." ; "Stroke patients are at risk for aspiration pneumonia..."
**Analysis & Answer**:
The patient presents with typical signs of pneumonia. Given that *S. pneumoniae* is the most common cause, it is the likely pathogen...
**Prediction**: A ✗

**OURS: SEMA-RAG**

**1. I-Agent (Interpretation)**
*Structured Schema*:
**Clinical Intent**: Infectious etiology & pathogen identification
**Medical Entities**: stroke; HAP/aspiration pneumonia; right-basal crackles; new right consolidation; fever + purulent cough
**Clinical Constraints**: 62y; hospital day 7 ; post-stroke aspiration risk; neutrophil predominance
**Initial Query**: "hospital day 7 post-stroke pneumonia right consolidation most likely causative organism"
**2. E-Agent (Exploration)**
***Iteration 1***:
**Evidence**: "...Post-stroke patients have high risk of aspiration... pneumonia causes include anaerobes and streptococci..."
**Gap**: Evidence does not distinguish pathogens based on hospitalization duration (day 7).
**Sufficiency**: $s_1 = 0$ (*Insufficient*) $\rightarrow$ **Next Query**: "most likely pathogen hospital-acquired pneumonia vs community-acquired"
***Iteration 2***:
**Evidence**: "...Hospital-acquired pneumonia (HAP) is defined as pneumonia $\geq$ 48h after admission..."
**Key Find**: "For late-onset HAP ($\geq$ 5 days), common pathogens include *Staphylococcus aureus* (MRSA) and *Pseudomonas*..."
**Sufficiency**: $s_2 = 1$ (*Sufficient*)
**3. A-Agent (Adjudication)**
**Report**: ... Hospital day 7 indicates HAP rather than CAP under standard definitions; combined with the candidate set, the evidence most consistently supports *S. aureus*. Among the provided options, *S. aureus* is the only matching HAP pathogen.
**Prediction**: D ✓

Table 4: A case of how SEMA-RAG helps deepseek-v3.1 find the correct answer on MedQA-US (Question 0024) by making the key clinical constraint explicit and retrieving decision-critical evidence, while MedRAG's single-round retrieval leads to a misleading rationale.

| Variant | MedQA-US (%) |
|---|---|
| SEMA-m1 ($m = 1$) | 86.72 |
| SEMA-m2 ($m = 2$) | 89.00 |
| SEMA-m3 ($m = 3$) | **89.95** |

Table 3: Effect of query breadth $m$ (fix $T_{\max} = 2$) on MedQA-US (deepseek-v3.1).

2024). In general domains, Self-RAG (Asai et al., 2023) and CRAG (Yan et al., 2024) use self-reflection to trigger re-retrieval, while in healthcare *i*-MedRAG (Xiong et al., 2025) iteratively refines queries via follow-up questions. However, without explicit clinical intent and constraints modeling, iterations may devolve into shallow rewrites, yielding inefficient retrieval, drift, and uncontrolled expansion (Zhao et al., 2024; Zhu et al., 2026).

### 5.2 Agentic Collaboration in Medicine

Multi-agent systems enhance complex-task solving through role specialization and coordination (Guo et al., 2024; Zong et al., 2024; Tran et al., 2025) (e.g., CAMEL (Li et al., 2023), MetaGPT (Hong et al., 2024)). ReAct (Yao et al., 2023) further couples reasoning with tool use, allowing agents to act and retrieve information during inference. In healthcare, MedAgents (Tang et al., 2024) and Agent-Hospital (Li et al., 2025) similarly show that multi-role clinical collaboration improves diagnosis and decision quality.

However, prior healthcare multi-agent work

largely centers on deliberation under the assumption that key evidence is already in-context, leaving evidence acquisition unsystematic (Chen et al., 2025; Gorenshtein et al., 2025; Wang et al., 2025). In particular, gap identification, sufficiency-driven termination, and evidence adjudication/integration are often missing, weakening reliable external evidence grounding and closed evidence-chain formation in real clinical tasks (Li, 2025; Amugongo et al., 2025).

## 6 Conclusion

We propose **SEMA-RAG**, a self-evolving multi-agent framework for medical question answering that restructures retrieval-augmented generation according to the staged process of clinical reasoning, with the **I-Agent** for clinical schema interpretation, the **E-Agent** for sufficiency-driven evidence exploration, and the **A-Agent** for evidence adjudication and final answer selection. Across five medical QA benchmarks and five LLM backbones, SEMA-RAG consistently outperforms strong baselines, improving the strongest baseline by an average of **+6.46** accuracy points, while ablations verify the necessity of the interpret–explore–adjudicate loop for reliable evidence-chain construction. Additional experiments further support its robustness across retrievers, smaller models, and more open-ended interactive settings. These findings suggest that medical RAG should move beyond static single-round retrieval toward more adaptive and reliable evidence construction.

## Limitations

Although we extend the evaluation beyond discrete-choice medical QA with additional open-ended and multi-turn benchmarks, the current study is still limited to benchmark-based settings rather than realistic clinical workflows such as longitudinal EHR reasoning or record-grounded decision support.

Our framework also depends on the quality and coverage of the retrieval corpus. If critical evidence is missing, outdated, or only partially retrieved, the self-evolving loop may still converge to incomplete grounding. In addition, the current sufficiency criterion is not explicitly designed for option-level separability or generative completeness.

Finally, SEMA-RAG introduces additional inference cost due to role specialization and multi-round exploration. Although this overhead is more efficient than fixed-step iterative baselines, it remains higher than single-round methods. The current design also lacks explicit relevance filtering during evidence accumulation, making performance sensitive to stopping criteria and exploration hyperparameters.

## References


Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877.

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. *arXiv preprint*. ArXiv:2505.08775 [cs].

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv preprint*. ArXiv:2310.11511 [cs].

Xiaolan Chen, Jiayang Xiang, Shanfu Lu, Yexin Liu, Mingguang He, and Danli Shi. 2025. Evaluating large language models and agents in healthcare: key challenges in clinical applications. *Intelligent Medicine*, 5(2):151–163.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. DeepSeek-V3 Technical Report. *arXiv preprint*. ArXiv:2412.19437 [cs].

Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, Hongru Wang, Han Xiao, Yuhang Zhou, Shaokun Zhang, Jiayi Zhang, Jinyu Xiang, Yixiong Fang, Qiwen Zhao, Dongrui Liu, and 8 others. 2026. A Survey of Self-Evolving Agents: What, When, How, and Where to Evolve on the Path to Artificial Super Intelligence. *arXiv preprint*. ArXiv:2507.21046 [cs].

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint*. ArXiv:2312.10997 [cs].

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, and 39 others. 2024. ChatGLM: A Family of Large Language Models from

GLM-130B to GLM-4 All Tools. *arXiv preprint*. ArXiv:2406.12793 [cs].

Linlu Gong, Ante Wang, Yunghwei Lai, Weizhi Ma, and Yang Liu. 2025. The Dialogue That Heals: A Comprehensive Evaluation of Doctor Agents' Inquiry Capability. *arXiv preprint*. ArXiv:2509.24958 [cs].

Google. 2025. Gemini 2.0 flash: Model card. Accessed 2025-12-28.

Alon Gorenshtein, Mahmud Omar, Benjamin S Glicksberg, Girish N Nadkarni, and Eyal Klang. 2025. AI Agents in Clinical Medicine: A Systematic Review.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: a survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *arXiv preprint*. ArXiv:2009.03300 [cs].

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. *arXiv preprint*. ArXiv:2308.00352 [cs].

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

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. *arXiv preprint*. ArXiv:2009.13081 [cs].

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. *arXiv preprint*. ArXiv:1909.06146 [cs].

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. BioASQ-QA: A manually curated corpus for Biomedical Question Answering. *Scientific Data*, 10(1):170.

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. In *Advances in Neural Information Processing Systems*, volume 36, pages 51991–52008. Curran Associates, Inc.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2025. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. *arXiv preprint*. ArXiv:2405.02957 [cs].

Xinzhe Li. 2025. A review of prominent paradigms for LLM-based agents: Tool use, planning (including RAG), and feedback learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9760–9779, Abu Dhabi, UAE. Association for Computational Linguistics.

Andrew Linn, Carole Khaw, Hugh Kildea, and Anne Tonkin. 2012. Clinical reasoning: A guide to improving teaching and practice. *Australian Family Physician*, 41(1-2):18–20. January/February issue.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. *Briefings in Bioinformatics*, 23(6):bbac409. ArXiv:2210.10341 [cs].

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of*

*the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Mahmud Omar, Girish N. Nadkarni, Eyal Klang, and Benjamin S. Glicksberg. 2024. Large language models in medicine: A review of current clinical trials across healthcare applications. *PLOS Digital Health*, 3(11):e0000662.

Jesutofunmi A. Omiye, Haiwen Gui, Shawheen J. Rezaei, James Zou, and Roxana Daneshjou. 2024. Large Language Models in Medicine: The Potentials and Pitfalls: A Narrative Review. *Annals of Internal Medicine*, 177(2):210–220.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. *arXiv preprint*. ArXiv:2203.14371 [cs].

Dimitri Roustan and François Bastardot. 2025. The Clinicians' Guide to Large Language Models: A General Perspective With a Focus on Hallucinations. *Interactive Journal of Medical Research*, 14:e59823.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Luca Soldaini, Andrew Yates, and Nazli Goharian. 2017. Learning to reformulate long queries for clinical decision support. *J. Assoc. Inf. Sci. Technol.*, 68(11):2602–2619.

StatPearls Publishing. 2025. Statpearls [internet]. NCBI Bookshelf. NCBI Bookshelf ID: NBK430685. Accessed 2026-01-04.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025a. Gemma 3 Technical Report. *arXiv preprint*. ArXiv:2503.19786 [cs].

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, and 150 others. 2025b. Kimi K2: Open Agentic Intelligence. *arXiv preprint*. ArXiv:2507.20534 [cs].

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. 2025. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. *arXiv preprint*. ArXiv:2501.06322 [cs].

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, and 3 others. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025. A survey of LLM-based agents in medicine: How far are we from baymax? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10345–10359, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2025. Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 30:199–214.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective Retrieval Augmented Generation. *arXiv preprint*. ArXiv:2401.15884 [cs].

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs].

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint*. ArXiv:2210.03629 [cs].

Shahram Yazdani, Mohammad Hosseinzadeh, and Fakhrolsadat Hosseini. 2017. Models of clinical reasoning with a focus on general practice: A critical review. *Journal of Advances in Medical Education & Professionalism*, 5(4):177–184.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint*. ArXiv:2506.05176 [cs].

Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K. Qiu, and Lili Qiu. 2024. Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely. *arXiv preprint*. ArXiv:2409.14924 [cs].

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2026. Large Language Models for Information Retrieval: A Survey. *ACM Transactions on Information Systems*, 44(1):1–54. ArXiv:2308.07107 [cs].

Chang Zong, Yuchen Yan, Weiming Lu, Jian Shao, Eliot Huang, Heng Chang, and Yueting Zhuang. 2024. Triad: A Framework Leveraging a Multi-Role LLM-based Agent to Solve Knowledge Base Question Answering. *arXiv preprint*. ArXiv:2402.14320 [cs].

# Appendix

## A Algorithm Pseudocode

Algorithm 1 formally describes the inference flow of SEMA-RAG, highlighting the interaction between agents.

**Algorithm 1** Inference Process of SEMA-RAG

**Require:** Question $Q$, corpus $\mathcal{C}$, max iterations $T_{\max}$
**Ensure:** Final answer $\tilde{y}$

**Stage 1: I-Agent**
$Q' = \langle o_{\text{int}}, o_{\text{ent}}, o_{\text{cons}}, q_{\text{init}} \rangle \leftarrow \text{Agent}_{\text{I}}(Pmt_{\text{I}}, Q)$
$\hat{q}_{\text{init}} \leftarrow \text{Linearize}(Q')$
$C_0 \leftarrow \varnothing$
$\mathcal{Q}_1 \leftarrow \{\hat{q}_{\text{init}}\}$
**Stage 2: E-Agent**
**for** $t = 1$ to $T_{\max}$ **do**
  $\mathcal{D}_t \leftarrow \bigcup_{q \in \mathcal{Q}_t} \text{TopK}(q)$
  $C_t \leftarrow \text{Dedup}(C_{t-1} \cup \mathcal{D}_t)$
  $[s_t, g_t, \mathcal{Q}_{t+1}] \leftarrow \text{Agent}_{\text{E}}(Pmt_{\text{E}}, [Q', \mathcal{Q}_t, C_t])$
  **if** $s_t = 1$ **then**
    **break**
  **end if**
  **if** $\mathcal{Q}_{t+1} = \varnothing$ **then**
    **break**
  **end if**
**end for**
$T \leftarrow t$
$C^* \leftarrow C_t$
**Stage 3: A-Agent**
$\mathcal{S}^* \leftarrow \{(\text{ID}(D_i), D_i) \mid D_i \in C^*\}$
$R \leftarrow \text{Agent}_{\text{A}}(Pmt_{\text{adj}}, [Q, C^*, \mathcal{S}^*])$
$\tilde{y} \leftarrow \text{Agent}_{\text{A}}(Pmt_{\text{ans}}, [Q, R])$
**return** $\tilde{y}$

## B Dataset Details

We evaluate SEMA-RAG on five standard medical question-answering benchmarks from the MIRAGE suite: MMLU-Med, MedQA-US, MedMCQA, PubMedQA*, and BioASQ-Y/N. Table 6 summarizes the dataset sizes and answer formats used in our experiments, and we briefly describe each benchmark below.

### B.1 MMLU-Med

The MMLU-Med dataset is a subset of the Massive Multitask Language Understanding (MMLU)

benchmark (Hendrycks et al., 2021). It covers six distinct subtasks: *Clinical Knowledge*, *Medical Genetics*, *Anatomy*, *Professional Medicine*, *College Biology*, and *College Medicine*. The questions are designed to measure knowledge acquired during preclinical and clinical medical training, formatted as 4-choice multiple-choice questions.

### B.2 MedQA-US

MedQA-US (Jin et al., 2020) is derived from the United States Medical Licensing Examination (USMLE). It represents highly complex clinical case studies that require multi-hop reasoning and domain-specific knowledge to solve. Following the standard setting in MIRAGE, we use the 4-option English version of the dataset. The questions typically present a patient vignette followed by a query about diagnosis, prognosis, or pharmacology.

### B.3 MedMCQA

MedMCQA (Pal et al., 2022) is a large-scale dataset collected from Indian medical entrance examinations (AIIMS and NEET-PG). It covers a wide range of 21 medical subjects, including surgery, pediatrics, and pharmacology. The questions vary significantly in difficulty and length, testing both memorization of medical facts and application of concepts in clinical scenarios.

### B.4 PubMedQA*

PubMedQA (Jin et al., 2019) is a biomedical research question-answering dataset. The task requires answering "Yes", "No", or "Maybe" to a research question based on a provided abstract. **PubMedQA*** refers to the setting where the original context (abstract) is removed, forcing the model to retrieve external evidence to answer the question. This setting tests the system's ability to find relevant biomedical literature to support a scientific conclusion.

### B.5 BioASQ-Y/N

BioASQ-Y/N is a subset of the BioASQ Task B benchmark (Tsatsaronis et al., 2015; Krithara et al., 2023). It consists of biomedical questions that require a strict "Yes" or "No" answer. These questions are expert-constructed and reflect real-world information needs of biomedical researchers. The task is challenging because it often involves specific gene-disease associations or protein interactions that require precise fact-checking.

| Dataset | #Samples | Task |
|---|---|---|
| MMLU-Med | 1089 | 4-choice MCQ |
| MedQA-US | 1273 | 4-choice MCQ |
| MedMCQA | 4183 | 4-choice MCQ |
| PubMedQA* | 500 | 3-choice Y/N/M |
| BioASQ-Y/N | 618 | 2-choice Y/N |

Table 6: Statistics of the medical QA datasets from MIRAGE used in our experiments.

## C Retrieval Corpus Details

Following *i*-MedRAG (Xiong et al., 2025) and MedRAG (Xiong et al., 2024), we employ a hybrid retrieval corpus that combines medical textbooks with point-of-care clinical summaries, covering both foundational concepts and practical clinical knowledge.

**Textbooks** This component is the released medical textbook collection used in prior medical QA benchmarks (Jin et al., 2021). It contains widely used reference textbooks spanning core biomedical sciences and clinical specialties, and is particularly helpful for queries requiring standard definitions, canonical mechanisms, and established medical principles.

**StatPearls** StatPearls (StatPearls Publishing, 2025) is a point-of-care clinical review resource that provides high-yield summaries across diseases, diagnostics, and treatments. In our setup, we use the publicly available StatPearls articles (e.g., via the NCBI Bookshelf releases) as in prior work, which complements textbooks with concise, practice-oriented evidence for retrieval.

## D Error Analysis

To probe where SEMA-RAG can still fail, we analyze a representative MedQA-US error (Question 0060) in Table 7. This case is informative because the system retrieves the correct biochemical cue at the *class* level, yet still makes an incorrect discrete choice among the remaining candidates.

Here, **I-Agent** identifies the decision anchors from both presentation and assay. The clinical picture indicates severe sepsis with a pelvic infectious focus, accompanied by DIC-like abnormalities. The phenol-heating assay reveals a phosphorylated *N*-acetylglucosamine dimer with multiple fatty acids, which strongly suggests a Lipid A–type

| Error Case (MedQA-US Q0060) |
|---|
| **Question (brief)**: Septic shock with pelvic infectious focus; phenol/90°C assay indicates a Lipid A–like motif ⇒ Gram-negative signal. |
| **Options**: |
| **A.** Coagulase-positive, Gram-positive cocci |
| **B.** Encapsulated, Gram-negative coccobacilli |
| **C.** Spore-forming, Gram-positive bacilli |
| **D.** Lactose-fermenting, Gram-negative rods |
| **1) I-Agent**: Extracts anchors (pelvic source, shock/DIC-like labs) and treats the biochemical clue as the key discriminator. |
| **2) E-Agent**: Retrieves evidence consistent with an LPS/Lipid A signature and correctly narrows the class to **Gram-negative**, ruling out **A/C** … |
| **3) A-Agent**: Commits among the remaining candidates without enforcing *option-separating* evidence (rod vs. coccobacillus; lactose fermentation) … |
| **Prediction**: B ✗ **Ground Truth**: D |

Table 7: A representative failure where retrieval supports only *class-level* elimination, while *option-level* discrimination remains under-supported

structure. Guided by these anchors, **E-Agent** retrieves evidence linking Lipid A to LPS and therefore to Gram-negative organisms, which is sufficient to eliminate the Gram-positive distractors.

The failure arises when moving from class identification to option selection. The retrieved evidence supports ruling out the Gram-positive options, but it does not provide *option-separating* signals within the remaining Gram-negative candidates. When the evidence report lacks an explicit bridge from the clinical setting to the discriminative phenotype expected in blood culture, **A-Agent** can be pulled toward salient surface descriptors and commit to an unsupported candidate.

This exposes a limitation of the current sufficiency and adjudication design. The loop may stop once it reaches a correct coarse conclusion, even though an additional round is still needed to uniquely determine the answer option. A practical implication is that sufficiency should be judged against *option-level separability*, not only against class-level plausibility. A simple fix is to tighten E-Agent's stopping rule by requiring evidence that supports one remaining option while directly excluding the other plausible candidates. If this condition is not met, the system should issue a final follow-up query targeting discriminative attributes and then re-adjudicate. We leave such option-aware sufficiency calibration to future work.

# E Additional Implementation Details

**Computational Resources** All experiments were executed in an API-based inference setting. The dense retrieval index was constructed and queried locally, while all LLM calls were served by the corresponding model providers.

**Retriever Settings** We use the MedCPT Query Encoder and Article Encoder to embed queries and passages, respectively. We then perform FAISS-based dense retrieval over Textbooks and StatPearls and globally rank the retrieved candidates.

# F Additional Robustness and Transfer Experiments

## F.1 Retriever Robustness

To examine whether SEMA-RAG depends strongly on a specific retriever, we further compare the domain-specific retriever MedCPT with the general-purpose retriever **qwen3-embedding-4b** (Zhang et al., 2025) on MedQA-US, using deepseek-v3.1 as the backbone. Table 8 shows that SEMA-RAG consistently improves over the corresponding single-round retrieval baseline under both retrievers. Meanwhile, MedCPT remains the stronger default choice in this clinical setting, suggesting that domain-specific retrieval is still advantageous for medical QA.

| Method | qwen3-embedding-4b | MedCPT |
|---|---|---|
| CoT | 77.53 | 77.53 |
| MedRAG | 76.43 | 77.14 |
| SEMA-RAG | **87.43** | **89.95** |

Table 8: Accuracy (%) on MedQA-US with different retrievers (deepseek-v3.1).

## F.2 Smaller-Model Robustness

To assess whether the gains of SEMA-RAG rely mainly on strong large-scale backbones, we further evaluate the framework on MedQA-US using a much smaller open-source model, **gemma3:4b** (Team et al., 2025a). All other settings remain the same as in the main experiments. Table 9 shows that although the absolute performance of all methods drops on the smaller model, SEMA-RAG still maintains a clear advantage over the strongest single-round baseline. This result suggests that task decoupling remains beneficial even when the base model has weaker instruction-following ability.

| Method | Acc. (%) |
|---|---|
| CoT | 51.77 |
| MedRAG | 56.01 |
| SEMA-RAG | **60.41** |

Table 9: Results on MedQA-US using gemma3:4b as the backbone.

### F.3 Beyond Discrete QA: Open-Ended and Interactive Settings

To examine whether SEMA-RAG generalizes beyond discrete-choice medical QA, we further evaluate it on two benchmarks covering open-ended generation and multi-turn medical dialogue. In these experiments, we keep the I-Agent and E-Agent unchanged, and only adapt the final stage of the A-Agent to generate free-text outputs instead of selecting a discrete option. This setting isolates the contribution of the evidence loop under more open-ended output formats.

**HealthBench.** HealthBench is an open-ended health-domain benchmark designed to assess response quality under clinician-authored rubrics (Arora et al., 2025). To test whether SEMA-RAG can transfer to grounded free-text generation with minimal modification, we evaluate deepseek-v3.1 on 500 randomly sampled English questions from HealthBench main using the official scoring script, and report the overall score together with three rubric dimensions: accuracy, completeness, and instruction following.

As shown in Table 10, SEMA-RAG consistently outperforms MedRAG across all reported metrics. This suggests that the evidence completion loop remains effective when the target output shifts from option selection to grounded free-text generation, providing a stronger factual basis for open-ended responses.

| Method | Avg. | Acc. | Comp. | Instr. Follow. |
|---|---|---|---|---|
| MedRAG | 26.87 | 31.34 | 30.17 | 41.66 |
| SEMA-RAG | **33.64** | **37.22** | **38.30** | **47.58** |

Table 10: Results on HealthBench for grounded open-ended response generation using deepseek-v3.1.

**MAQuE.** MAQuE is a multi-turn medical dialogue benchmark that evaluates response quality in simulated clinical communication settings (Gong et al., 2025). Compared with single-turn QA, it places greater emphasis on maintaining robust, relevant, and contextually appropriate responses across iterative interactions.

To test whether SEMA-RAG remains effective in interactive settings, we evaluate deepseek-v3.1 on 200 randomly sampled MAQuE test cases using the official evaluation script, and report four communication-oriented metrics: Accuracy, Robustness, Relevance, and Empathy. Table 11 shows that SEMA-RAG maintains clear gains over MedRAG on all four metrics. This improvement suggests that sufficiency-driven exploration remains beneficial in interactive clinical scenarios, where the system must sustain grounded response generation over multiple turns.

| Method | Acc. | Rob. | Rel. | Emp. |
|---|---|---|---|---|
| MedRAG | 52.50 | 64.86 | 74.00 | 66.40 |
| SEMA-RAG | **61.50** | **75.38** | **82.00** | **72.00** |

Table 11: Results on MAQuE for multi-turn clinical response generation using deepseek-v3.1.

Overall, these results suggest that the SEMA-RAG framework generalizes beyond discrete medical QA to more open-ended and interactive forms of clinical response generation.

## G Prompt Templates

To facilitate reproducibility, we provide the system instructions used for the three agents in SEMA-RAG. Note that the A-Agent uses two prompts for evidence adjudication and final answer selection, respectively.

```
I-Agent Prompt (Clinical Schema Interpreter).

Role:
You are an expert clinician.

Goal:
Given an unstructured medical question, extract an explicit Clinical Schema that makes the implied intent and constraints searchable. Focus on what must be retrieved and do not answer the question itself.

Input:
Medical Question: {research_topic}

Task:
Identify:
1. clinical intent (task type),
2. core medical entities (salient concepts from the question),
3.  key constraints (time course, demographics, setting, comorbidities, severity, contraindications, risk factors, anatomical or functional qualifiers),
4. a concise retrieval query aligned with the schema (q_init).

Key Instructions:
- Entities should be small, focused, and primarily grounded in the question itself.
- Include only the most retrieval-relevant concepts; avoid broad, redundant, or unnecessary enumeration.
- Merge obvious synonyms, near-duplicates, or simple morphological variants into one canonical medical expression when possible.
- Prefer the main clinical concept, condition, mechanism, finding, test, treatment, population, anatomical target, or other decision-critical concept that is necessary for retrieval.
- If the question provides candidate answers or options, do not mechanically include all of them as entities; include an option only if needed for retrieval or candidate discrimination.
- For multiple-choice, judgment, or open-ended questions, center the schema on the stem and its decision-critical medical concepts rather than listing answer choices.
- Constraints should capture only decision-relevant qualifiers explicitly stated or strongly implied by the question.
- Preserve key medical relations when they are essential for retrieval, such as derivation, origin, cause, association, indication, or contraindication.
- q_init should retrieve the knowledge needed to answer the question, remain neutral, and avoid prematurely inferring a conclusion.
- q_init should be short, medically precise, and should not simply concatenate all entities or options.
- Use precise medical terminology.
- Do not add explanations, rationale, or extra keys.

Output JSON:
{
"intent": "<short clinical task type>",
"entities": ["<entity1>", "<entity2>"],
"constraints": ["<constraint1>", "<constraint2>"],
"q_init": "<one concise neutral search-style query>"
}
```

Figure 4: Prompt template for the I-Agent clinical schema interpreter.

```
E-Agent Prompt (Self-Evolving Explorer).

Role:
You are an evidence sufficiency auditor and query refiner for medical question answering.

Goal:
Determine whether the current retrieved evidence is sufficient to answer the medical question under the given Clinical Schema. Do not answer the question itself.

Input:
Clinical Schema: {clinical_schema}
Current Query Set: {query_list}
Retrieved Evidence Summaries: {summaries}

Key Instructions:
- Assess whether the current evidence sufficiently covers the key intent, entities, and constraints in the Clinical Schema.
- Judge sufficiency based on whether the evidence is enough to support final answer selection, or to distinguish among competing candidate answers when relevant.
- Evidence may be relevant yet still insufficient; do not mark sufficiency = 1 unless the evidence is adequate for confident answer selection.
- If the evidence is insufficient, identify the single most important missing fact, missing distinction, or unresolved clinical criterion.
- Generate 1 to 3 follow-up queries that directly target this gap.
- Follow-up queries must be specific, self-contained, non-redundant, and explicitly grounded in the Clinical Schema.
- For questions with candidate answers, prioritize queries that help distinguish among candidates rather than broad background expansion.
- Prefer targeted refinement over broad exploratory expansion.
- Do not repeat an existing query unless revision is necessary.
- If the current evidence is already sufficient, return no follow-up queries.

Rules:
- If sufficiency = 1, set "gap" to "N/A" and "queries" to [].
- If sufficiency = 0, "gap" must be specific, concrete, and decision-relevant rather than generic.
- Queries should target missing clinical distinctions, time conditions, population constraints, contraindications, severity, mechanisms, diagnostic criteria, or option-level discrimination when relevant.
- Return JSON only.

Output JSON:
{
"sufficiency": 0 or 1,
"gap": "<short concrete description of the most important missing evidence>",
"queries": ["<query1>", "<query2>", "<query3>"]
}
```

Figure 5: Prompt template for the E-Agent self-evolving explorer.

```
A-Agent Prompt (Phase 1: Evidence Adjudicator).

Role:
You are a medical evidence adjudicator.

Goal:
Synthesize the final retrieved evidence into a concise, traceable report that can support final answer selection. Do not directly answer the question. Only organize, adjudicate, and summarize the evidence.

Input:
Medical Question: {research_topic}
Clinical Schema: {clinical_schema}
Final Query Set: {query_list}
Retrieved Evidence Summaries: {summaries}

Key Instructions:
- Review the retrieved evidence in light of the medical question and Clinical Schema.
- Focus on the most decision-relevant evidence and remove redundancy.
- Identify which evidence directly supports a candidate conclusion, which evidence conflicts with it, and which evidence is only background, indirect, or weakly relevant.
- When multiple pieces of evidence overlap, merge them into one concise statement.
- When evidence is incomplete, uncertain, indirect, or conflicting, make that explicit rather than resolving it prematurely.
- Preserve traceability by attaching source identifiers or summary indices whenever available.
- Every claim in the report must be supported by the provided summaries; do not infer unsupported medical facts.
- Do not introduce external medical knowledge.
- Do not perform final answer selection.
```

Figure 6: Prompt template for the A-Agent evidence adjudicator.

```
A-Agent Prompt (Phase 1: Evidence Adjudicator) (continued).

Rules:
- Keep the report concise, traceable, and decision-oriented.
- Prefer evidence that is directly relevant to the question over general background knowledge.
- If there is no real conflicting evidence, return an empty list for "key_conflicting_or_limiting_evidence".
- If source identifiers are unavailable, use summary indices or short summary labels consistently.
- Do not repeat the same evidence across multiple fields unless necessary.
- Return JSON only.

Output JSON:
{
"question_focus": "<one short sentence stating what must be decided>",
"key_supporting_evidence": [
{
"claim": "<concise evidence-supported statement>",
"source_ids": ["<source1>", "<source2>"]
}
],
"key_conflicting_or_limiting_evidence": [
{
"claim": "<concise conflicting, uncertain, or limiting statement>",
"source_ids": ["<source1>", "<source2>"]
}
],
"evidence_synthesis": "<short integrated synthesis of what the evidence supports, what remains uncertain, and what
distinction matters most for final answer selection>"
}
```

Figure 7: Prompt template for the A-Agent evidence adjudicator.

```
A-Agent Prompt (Phase 2: Evidence-Grounded Answering).

Role:
You are a medical AI assistant.

Goal:
Answer the multiple-choice medical question using the provided evidence adjudication report.

Input:
Medical Question: {research_topic}
Evidence Adjudication Report: {adjudication_report}

Key Instructions:
- Select exactly one final answer: A, B, C, or D.
- First rely on the evidence adjudication report.
- If the report contains relevant evidence, choose the option best supported by that evidence.
- If the report is incomplete, weak, or lacks directly relevant evidence, use medical knowledge to reason and choose the
most appropriate answer.
- Do not output reasoning, JSON, code blocks, or any extra text.

Output Format:
Final Answer: [A/B/C/D]
```

Figure 8: Prompt template for the A-Agent evidence-grounded answerer.