Med-SRAF: A Multi-Agent Framework for Medical Reasoning via Semantic Routing and Agentic Fusion

Xiao Li[1]*, Zhuo Chen[1]*, Jun Xia[4], Hongxin Xiang[5], Chao Wang[3], Wenjie Du[1,2]†, Yang Wang[1,2]†
[1]University of Science and Technology of China
[2]Suzhou Institute for Advanced Research, USTC, Suzhou, China [3]ByteDance, Inc.
[4]School of Engineering, Westlake University, Hangzhou, China
[5]College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
{lx_25225197,czchenzhuo, duwenjie,angyan}@mail.ustc.edu.cn
junxia@hkust-gz.edu.cn, xianghx@hnu.edu.cn, wangchao.hhhhh@bytedance.com

Abstract

While Retrieval-Augmented Generation (RAG) has become a standard paradigm for mitigating hallucinations in Large Language Models (LLMs), its effectiveness in complex medical reasoning remains limited. Existing RAG methods suffer from two main challenges: First, **Semantic Drift**: without explicit domain constraints, LLM-driven query decomposition often deviates from the original clinical intent, introducing substantial noise that degrades retrieval relevance. Second, **Concatenation Fallacy**: retrieved evidence from different semantic aspects is aggregated in a naive, unstructured manner, without modeling their inter-dependencies and potential conflicts, which ultimately undermines downstream reasoning. To address these challenges, we propose **Med-SRAF**, a multi-agent retrieval augmentation framework guided by medical domain knowledge. This framework reconstructs the traditional RAG process through two core mechanisms: (1) Intent-driven Semantic Routing, where a UMLS-based NavigationAgent dynamically maps queries to medical dimensions for strategic search space pruning; and (2) Evidence-based Agentic Fusion, where a FusionAgent resolves conflicts among dimension-specific evidence to build logically consistent reasoning chains. Extensive experiments on five widely used medical benchmarks show that Med-SRAF consistently outperforms existing general RAG baselines, achieving an average accuracy improvement of over **4.7%**, highlighting its effectiveness in robust and interpretable medical reasoning. Our code is at https://anonymous.4open.science/r/MultiAgent_RAG-F6DC.

1 Introduction

Recently, the rapid development of Large Language Models (LLMs) (Brown et al., 2020; Achiam et al., 2023; Touvron et al., 2023) has precipitated a fundamental paradigm shift in artificial intelligence, transforming Natural Language Processing (NLP) from task-specific pipelines to generalist reasoning engines. Trained on massive-scale corpora encompassing diverse domains, these models have demonstrated remarkable capabilities in language understanding, generation, and logical reasoning, reshaping how humans interact with information. However, despite their extensive pre-training and linguistic fluency, LLMs intrinsically suffer from a critical limitation: hallucinations (Ji et al., 2023; Zhang et al., 2025). That is, they may generate responses that are syntactically plausible and contextually coherent yet factually incorrect, unverifiable, or unsupported by reliable evidence (Lin et al., 2022). Hallucinations significantly impede the effectiveness of large language models, particularly in tasks that demand high accuracy and reliability, such as academic research and medical diagnosis (Nori et al., 2023; Aljohani et al., 2025).

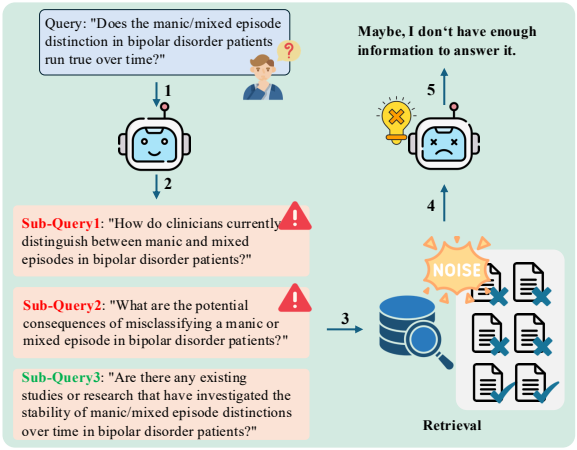

Figure 1: The traditional LLM-driven query decomposition struggled with semantic drift.

To mitigate hallucinations in Large Language Models (LLMs), Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has emerged as a stan-

* Both authors contributed equally to this research.
† Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 38009–38029

dard solution, aiming to ground model outputs in external, verifiable knowledge bases to bridge the gap between static parametric memory and dynamic real-world facts. By providing retrieved evidence related to the question, RAG significantly improves factual grounding and has been widely adopted across knowledge-intensive tasks. While RAG provides a promising foundation for mitigating hallucinations and ensuring factuality, applying it to complex medical reasoning remains fraught with challenges.

Medical reasoning often require precise intent alignment, current approaches often employ unconstrained Chain-of-Thought (CoT) (Wei et al., 2022) or Query Decomposition techniques (Khot et al., 2022) to break down queries. However, without domain-specific boundaries, these methods are prone to semantic drift. LLMs may decompose a clinical query into irrelevant sub-questions, introducing noise that distracts the retrieval process. Furthermore, unguided decomposition often leads to retrieval redundancy, where overlapping sub-queries exhaust the limited context window—a critical bottleneck for local LLM deployment. Existing frameworks lack a mechanism to strictly enforce clinical scope, resulting in reasoning paths that are computationally expensive yet diagnostically vague. To address this issue, we propose semantic routing to provide structured guidance. This approach introduces a UMLS-based (Bodenreider, 2004) constraint mechanism, which disentangles the reasoning space into five distinct semantic dimensions (e.g., Disease, Symptom). By mapping queries to disjoint sets of UMLS Semantic Types, our method effectively prunes irrelevant search paths before retrieval ensues. This significantly improves the effectiveness of the information retrieved by the query after decomposition and reduces redundancy and noise.

However, even when relevant information is successfully retrieved, a second fundamental limitation persists: the concatenation fallacy (Liu et al., 2024). Most multi-path retrieval systems adopt a naive strategy of mechanically concatenating retrieved segments into the context window. This approach implicitly assumes that the LLM will spontaneously deduce the latent inter-relationships among disparate pieces of information. This assumption is particularly problematic in medical reasoning, where meaningful conclusions often depend on explicit interactions across semantic dimensions—for instance, linking a *Pathological Mechanism* from one piece of evidence to a *Clinical Symptom* from another requires an explicit inferential bridge. Simply concatenating these as isolated facts creates information silos, leading to fragmented reasoning where the model possesses the puzzle pieces but fails to connect them. Moreover, this naive aggregation exacerbates the risk of information conflicts, where contradictory evidence retrieved from divergent paths is fed directly to the model without reconciliation, causing hallucination or indecision. To address this issue, we propose agentic fusion, which makes implicit logical chains explicit by establishing connections between knowledge and resolves conflicts across different dimensions. This allows large models to fully utilize the retrieved information, thereby generating more accurate responses.

Our main contributions can be summarized as follows:

- We establish a new paradigm of structured guidance grounded in five distinct semantic dimensions derived from the UMLS ontology.
- We propose a Semantic Routing Mechanism to specifically address semantic drift and retrieval redundancy, effectively pruning irrelevant search paths before retrieval ensues.
- We propose an Agentic Fusion Mechanism to overcome the concatenation fallacy inherent in traditional RAG systems, reconstructing latent logical chains to support more accurate and coherent medical inferences.
- We conduct extensive experiments on five widely used medical benchmarks. The results show Med-SRAF achieves an average accuracy improvement of over 4.7%.

2 Method

Overview. Figure 2 provides an overview of our Med-SRAF. To effectively address the dual challenges of semantic drift and concatenation fallacy highlighted in section 1, we introduce two novel core modules: the **NavigationAgent**, which implements semantic routing to strictly constrain the reasoning scope, and the **FusionAgent**, which executes agentic fusion to explicitly synthesize logical evidence chains. Orchestrated by a central coordinator, these modules collaborate with specialized Sub-Agents to transform the QA pipeline from a static retrieval task into a dynamic, structure-grounded decision process.

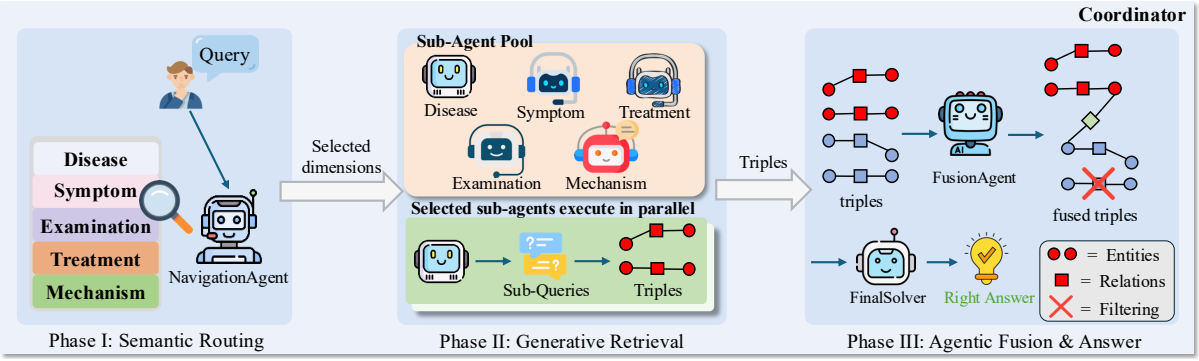

Figure 2: **The Overview of Our Framework**. Upon receiving a query, the NavigationAgent analyzes its clinical intent and dynamically selects relevant dimensions from the predefined dimensions. Subsequently, the central coordinator activates the corresponding Sub-Agents to execute retrieval tasks in parallel. This process involves dimension-specific query decomposition and evidence retrieval. Following this, the FusionAgent conducts evidence synthesis based on the evidence from the selected Sub-Agents to build logically consistent reasoning chains. Finally, the LLM is prompted to generate the final response based on these fused chains.

Problem Formulation. Given a biomedical query Q, which consists of a question stem q and a set of candidate options $op = \{o_1, o_2, \ldots, o_n\}$, the task is to identify the correct answer $Ans \in op$ through the three-phase workflow of Med-SRAF.

2.1 Phase I: Semantic Routing

Upon receiving the query Q, traditional LLM-driven query decomposition tends to be unconstrained and often deviates from the original clinical intent, causing the generation of irrelevant sub-queries called semantic drift. To mitigate it, we introduce semantic routing mechanism. Instead of allowing the LLM to freely decompose the query into arbitrary sub-queries, we explicitly constrain the action space to specific semantic dimensions grounded in the Unified Medical Language System (UMLS) (Bodenreider, 2004).

2.1.1 UMLS-based Dimension Mapping

Considering precise and controlled vocabulary are required for medical reasoning, and to prevent redundancy from query decomposition and ambiguity in open-ended retrieval, we ground our framework in the UMLS. We aggregate granular UMLS Semantic Types into five distinct semantic reasoning dimensions.

Specifically, we distill the vast UMLS Semantic Network into a quasi-orthogonal basis set. As illustrated in Figure 3, by explicitly pruning irrelevant categories (e.g., *Reptile*, *Geographic Area*) and aggregating clinically relevant semantic types (e.g., T047, T191), we construct five distinct reasoning dimensions, denoted as $\mathcal{D} = \{d_{dis}, d_{sym}, d_{exam}, d_{treat}, d_{mech}\}$:

Disease (d_{dis}): Focuses on diagnosis, etiology, and prognosis. It maps to UMLS types such as *Disease or Syndrome* (T047) and *Neoplastic Process* (T191), providing the foundational scope when specific pathological definitions and classification criteria are needed to answer the query.

Symptom (d_{sym}): Focuses on clinical manifestations and patient complaints. It captures phenomenological evidence by consolidating UMLS types such as *Sign or Symptom* (T184) and *Clinical Finding* (T033), representing both subjective complaints and observable signs.

Examination (d_{exam}): Distinct from symptoms, this dimension targets objective diagnostic investigations used to confirm or rule out hypotheses. It maps to UMLS types like *Diagnostic Procedure* (T060) and *Laboratory Procedure* (T059), covering imaging, pathology, and biochemical assays required for confirmation.

Treatment (d_{treat}): Focuses on all therapeutic interventions. Unlike systems that separate pharmacotherapy from surgery, we aggregate types like *Pharmacologic Substance* (T121) and *Therapeutic Procedure* (T061) here, addressing both medication management and surgical interventions uniformly.

Mechanism (d_{mech}): Focuses on "why" and "how" of medical phenomena. Mapping to UMLS types like *Pathologic Function* (T046) and *Molecular Function* (T044), this dimension is essential for answering deep reasoning questions regarding pathophysiology, mechanism of action (MOA), and metabolic pathways.

This structured taxonomy establishes a controlled action space for the NavigationAgent to

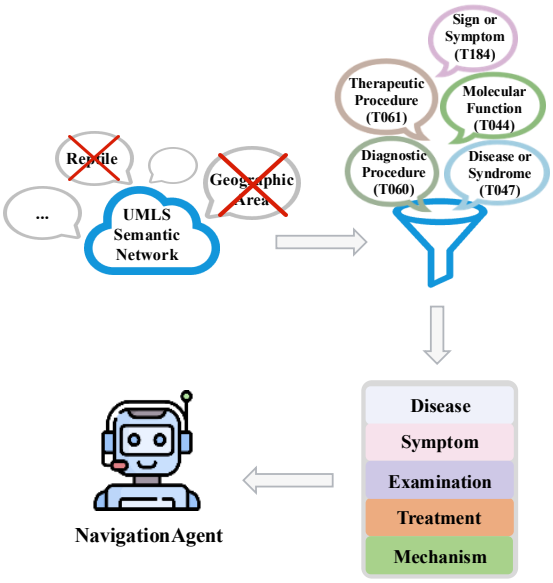

Figure 3: **Illustration of the UMLS-based Dimension Mapping strategy**. By explicitly pruning irrelevant categories and aggregating clinically relevant semantic types, the vast UMLS Semantic Network is distilled into five distinct semantic reasoning dimensions.

perform effective dimensional pruning. This process significantly reduces retrieval noise and ensures that the decomposition is both **comprehensive** (covering distinct clinical needs from diagnosis to underlying mechanisms) and **exclusive** (minimizing semantic overlap).

2.1.2 Intent-Aware Routing

Having established the static semantic space $\mathcal{D}$, the challenge shifts to how to map unstructured clinical queries into corresponding dimensions. The NavigationAgent acts as the brain in this process, analyzing the query Q to select the most relevant subset of dimensions $\mathcal{D}' \subset \mathcal{D}$.

$$\mathcal{D}' = \{d_i \in \mathcal{D} \mid P(d_i \mid Q) > \tau\} \quad (1)$$

where $P(d_i \mid Q)$ denotes the probability that dimension d_i is required to answer the query Q, and τ is a confidence threshold.

We designed a specialized prompt to instruct the NavigationAgent to analyze the query's latent needs. For example, consider a query regarding "clinical signs of third cranial nerve damage following orbital trauma". The NavigationAgent recognizes that the question focuses on phenotypic manifestations and diagnostic validation rather than therapeutic intervention or molecular pathology. Consequently, it activates the subset $\mathcal{D}' = \{d_{sym}, d_{exam}\}$ to retrieve information on specific signs like pupillary dilation and strabismus , while explicitly pruning $\{d_{treat}, d_{mech}, d_{dis}\}$. This scope locking strategy effectively filters out irrelevant noise (e.g., surgical repair techniques or metabolic pathways) before retrieval.

Upon receiving the semantic routing results $\mathcal{D}'$, the central coordinator dispatches the query Q to the designated specialized sub-agents. To optimize inference latency, these agents are orchestrated to execute Phase II (Generative Retrieval) in parallel, ensuring efficient and comprehensive evidence retrieval across the selected dimensions.

2.2 Phase II: Generative Retrieval

In long-text medical scenarios, directly using the original query for knowledge retrieval often faces problems like high contextual noise and blurred focus. To address this challenge, we adopt a generative retrieval paradigm, consisting of two consecutive sub-processes: Sub-Query Generation and Structured Triple Generation.

2.2.1 Sub-Query Generation

This module is designed to decompose Q into several short, precise search queries. Compared to typical query decomposition and rewriting, we designed a dimensional-aware generation strategy to ensure the generated sub-queries strictly adhere to the semantic dimensions selected during Phase I.

Specifically, for the received original question stem q, the options set $op = \{o_1, o_2, \ldots, o_n\}$, and the selected dimension $d \in D'$, the sub-query generation process is strictly guided by a dimension-specific prompt template P_d.

$$\begin{aligned} S_d &= \mathcal{G}(q, \mathcal{T}(op), P_d; \theta) \\ &= \{q_1^{(d)}, q_2^{(d)}, \ldots, q_k^{(d)}\} \end{aligned} \quad (2)$$

where $\mathcal{T}(\cdot)$ represents the function that converts structured options into a textual context, and S_d denotes the result set of k targeted verification sub-queries tailored to dimension d.

2.2.2 Structured Triple Generation

After the generation of dimension-specific sub-queries set S_d, the next task of the Sub-Agent is to utilize these sub-queries to obtain structured medical evidence. Here, we adopt a model fine-tuned on PrimeKG (Chandak et al., 2023; Su et al., 2024) to generate structured triples (h, r, t) for each sub-

query in S_d.

$$\mathcal{K}_d = \bigcup_{sub\_q \in S_d} \mathcal{E}(sub\_q; \Phi) = \{\tau_1, \tau_2, \ldots, \tau_m\} \quad (3)$$

where $\mathcal{E}(\cdot; \Phi)$ denotes the generative extraction operator based on model Φ, incorporating robust parsing logic to get structured evidence; and τ_i represents the resulting i-th standardized triplet (h, r, t).

After parallel generative retrieval, the system obtains a set of dimension-specific triplet knowledge sets $\{\mathcal{K}_d \mid d \in D'\}$. Then the set is flowed to FusionAgent for Agentic Fusion.

2.3 Phase III: Agentic Fusion

While the structured triples generated in Phase II provide reliable local evidence, they remain inherently discrete and dispersed across isolated semantic dimensions. Consequently, integrating this fragmented information and reconciling potential cross-dimensional conflicts is crucial. The FusionAgent acts as the core in this process, synthesizing these sets of triples $\{\mathcal{K}_d\}$ into a unified reasoning chain for more accurate and coherent medical inferences.

This process is not a simple union of sets, but rather a semantic reconstruction procedure. Specifically, we feed the raw triples from all selected dimensions (e.g., Disease, Treatment, etc.) together with the original query Q into the FusionAgent. We designed prompts that strictly instruct the FusionAgent to synthesize knowledge by adhering to three core principles:

- **Semantic Deduplication**: Normalize entities that convey the same medical meaning but are expressed differently (e.g., *myocardial infarction* vs. *heart attack*) into standardized medical terminology to ensure semantic connectivity.
- **Conflict Resolution**: When evidence from different sources is contradictory, the system prioritizes statements with higher granularity and definitive logical relations over broad or ambiguous claims.
- **Relevance Filtering**: Remove noisy triples that have no direct logical connection to the target unknown in Q.

$$\mathcal{G}_{\text{fused}} = \Psi(\{\mathcal{K}_d\}_{d \in D'}, Q; \mathcal{P}_{\text{fuse}}) \quad (4)$$

where $\Psi(\cdot)$ denotes the semantic synthesis operator executed by the FusionAgent; $\{\mathcal{K}_d\}_{d \in D'}$ represents the collection of knowledge sets retrieved from parallel sub-agents in Phase II; and $\mathcal{P}_{\text{fuse}}$ refers to the steering prompt encoding the core synthesis constraints—specifically semantic deduplication, conflict resolution, and relevance filtering. This operation transforms the isolated local evidence into $\mathcal{G}_{\text{fused}}$, a unified and consistent evidence landscape serving as the ground truth for final reasoning.

Notably, the FusionAgent employs an adaptive prompting strategy sensitive to the routing results of Phase I. In scenarios where just one dimension is selected (i.e., $|D'| = 1$), the FusionAgent's role shifts from Cross-Dimensional Synthesis to Single-Source Data Refinement. In this case, the fusion prompt $\mathcal{P}_{\text{fuse}}$ is reconfigured to emphasize precision within the dimension. Specifically, the instruction constraints focus on:

- **Denoising**: Strictly discarding ambiguous or "None" values.
- **Standardization**: Converting vague predicates into precise medical terminology.
- **Hallucination Prevention**: Ensuring that no unsupported relations are fabricated.

This adaptive mechanism ensures that even when cross-verification is unavailable, the output graph $\mathcal{G}_{\text{fused}}$ maintains high logical density and reliability.

After FusionAgent processing, the coordinator prompts the LLM to generate the answer Ans to the query Q based on the $\mathcal{G}_{fused}$.

3 Experiment

3.1 Experimental Setup

Datasets. Five multi-choice medical QA benchmarks are utilized for experimental evaluation: MMLU (Hendrycks et al., 2020), MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), BioASQ (Nentidis et al., 2023) and MedDDx (Su et al., 2024). More details are provided in Appendix A.

Baselines. 14 models are included as baselines in the evaluation, categorized into five groups: (1) General-purpose LLMs: LLaMA2 (7B/13B) (Touvron et al., 2023), LLaMA3 (8B), and LLaMA3.1 (8B) (Dubey et al., 2024); (2) Medically-tuned LLMs: MedAlpaca-7B (Han et al., 2023) and PMC-LLaMA-7B (Wu et al., 2024); (3) RAG-based models: Self-RAG (7B/13B) (Asai et al., 2023) and MedRAG (13B) (Xiong et al., 2024); (4) KG-based models: KG-Rank (13B) (Yang et al., 2024) and KGARevion (LLaMA3/3.1-8B) (Su et al., 2024); and (5) Proprietary LLMs: GPT-4 and Gemini-Pro.

Datasets	Multiple Medical QA				MedDDx		
	MMLU	MedQA	PubMedQA	BioASQ	Basic	Intermediate	Expert
Metrics	Acc. (std)	Acc. (std)	Acc. (std)	Acc. (std)	Acc. (std)	Acc. (std)	Acc. (std)
LLaMA2-7B	0.408 (.001)	0.371 (.003)	0.447 (.004)	0.577 (.002)	0.197 (.006)	0.199 (.004)	0.183 (.003)
LLaMA2-7B (COT)	0.406 (.001)	0.316 (.002)	0.477 (.003)	0.570 (.004)	0.263 (.016)	0.259 (.002)	0.231 (.012)
MedAlpaca-7B	0.598 (.003)	0.394 (.002)	0.345 (.014)	0.532 (.004)	0.356 (.012)	0.328 (.005)	0.295 (.005)
MedAlpaca-7B (COT)	0.601 (.007)	0.391 (.005)	0.294 (.005)	0.498 (.007)	0.370 (.018)	0.324 (.004)	0.297 (.014)
PMC-LLaMA-7B	0.227 (.005)	0.265 (.003)	0.182 (.006)	0.310 (.008)	0.095 (.002)	0.092 (.001)	0.068 (.005)
PMC-LLaMA-7B (COT)	0.231 (.003)	0.264 (.002)	0.180 (.005)	0.298 (.008)	0.086 (.012)	0.086 (.004)	0.060 (.009)
LLaMA2-13B	0.500 (.002)	0.394 (.004)	0.482 (.007)	0.554 (.004)	0.339 (.004)	0.345 (.010)	0.326 (.007)
LLaMA2-13B (COT)	0.498 (.004)	0.388 (.001)	0.478 (.007)	0.569 (.007)	0.301 (.006)	0.344 (.006)	0.320 (.009)
LLaMA3-8B	0.657 (.004)	0.549 (.003)	0.555 (.006)	0.664 (.014)	0.423 (.013)	0.358 (.003)	0.331 (.010)
LLaMA3-8B (COT)	0.684 (.001)	0.546 (.002)	0.563 (.004)	0.688 (.011)	0.439 (.014)	0.362 (.003)	0.355 (.009)
LLaMA3.1-8B	0.699 (.004)	0.605 (.003)	0.572 (.004)	0.685 (.009)	0.441 (.011)	0.401 (.003)	0.384 (.005)
LLaMA3.1-8B (COT)	0.706 (.002)	0.604 (.002)	0.567 (.004)	0.704 (.006)	0.452 (.018)	0.402 (.008)	0.383 (.009)
Gemini-pro	0.743 (.002)	0.610 (.003)	0.699 (.002)	0.798 (.004)	0.476 (.008)	0.492 (.003)	0.463 (.010)
GPT-4	**0.752** (.007)	**0.719** (.003)	**0.752** (.007)	**0.835** (.004)	**0.544** (.015)	**0.532** (.002)	**0.529** (.004)
Self-RAG (7B)	0.512 (.004)	0.409 (.002)	0.338 (.010)	0.598 (.006)	0.224 (.019)	0.200 (.004)	0.209 (.005)
Self-RAG (13B)	0.234 (.004)	0.264 (.002)	0.173 (.006)	0.612 (.014)	0.235 (.013)	0.253 (.002)	0.224 (.014)
KG-Rank (13B)	0.454 (.005)	0.364 (.004)	0.312 (.007)	0.505 (.015)	0.238 (.012)	0.253 (.003)	0.220 (.007)
MedRAG (13B)	0.534 (.003)	0.410 (.004)	0.493 (.004)	0.585 (.008)	0.340 (.012)	0.318 (.006)	0.326 (.005)
KGARevion (LLaMA3-8B)	0.702 (.004)	**0.599** (.003)	0.561 (.006)	0.740 (.005)	0.460 (.013)	0.394 (.004)	0.383 (.012)
KGARevion (LLaMA3.1-8B)	**0.728** (.006)	0.598 (.004)	**0.603** (.004)	**0.756** (.004)	**0.464** (.013)	**0.442** (.005)	**0.438** (.008)
LLaMA3-8B (Ours)	0.747(.003)	0.637(.003)	0.641(.008)	0.811(.008)	0.495(.023)	0.470(.006)	0.471(.012)
LLaMA3.1-8B (Ours)	0.769(.005)	0.640(.003)	0.652(.007)	0.835(.004)	0.503(.015)	0.499(.003)	0.475(.007)
Gemini-pro (Ours)	0.773 (.004)	0.669 (.003)	0.713 (.005)	0.868 (.005)	0.494 (.019)	0.520 (.003)	0.480 (.009)
GPT-4 (Ours)	0.793 (.003)	0.750 (.003)	0.804 (.005)	0.882 (.007)	0.605 (.008)	0.591 (.006)	0.573 (.007)
Improvement over best baseline	**+4.1%**	**+3.1%**	**+5.2%**	**+4.7%**	**+6.1%**	**+5.9%**	**+4.4%**

Table 1: The accuracy results of our method and all baseline methods on five multi-choice medical QA benchmarks. The value highlighted in **blue** represents the best result among General-purposed LLMs, Medically-tuned LLMs and Proprietary LLMs. The value highlighted in **red** marks the best result among RAG-based Models and KG-based Models. Gray represents the results of our method. std means the standard deviation under three runs.

All models are evaluated under consistent experimental settings to ensure fair comparison.

Evaluation. The evaluation is designed to include both multi-choice question answering and open-ended reasoning tasks. For multi-choice questions, given an input query with several candidate options, the model is required to select the correct answer from the provided set. For open-ended reasoning, the model generates free-form responses based on the input query, requiring coherent and logically consistent reasoning. Accuracy (Acc) is employed as the primary metric for the multi-choice setting, while open-ended responses are evaluated based on correctness and reasoning quality.

3.2 Multi-Choice Question-Answering

Table 1 presents the accuracy and standard deviation of our Med-SRAF and all baseline models across all five datasets. Results on five multi-choice medical QA datasets showed that our Med-SRAF achieves an average accuracy improvement exceeding 4.7%, consistently outperforming all baselines. On standard multiple medical benchmarks (MMLU, MedQA, PubMedQA and BioASQ), Med-SRAF achieves consistent accuracy gains over baselines, improving the accuracy by 4.1%, 3.1%, 5.2% and 4.7%, respectively. These benchmarks span a diverse range of medical knowledge, from fundamental biomedical concepts to complex clinical heuristics, providing a comprehensive evaluation of model generalization. The superior performance here validates the effectiveness of our Semantic Routing Mechanism. By mapping diverse queries into distinct semantic dimensions (Phase I), our framework effectively constrains the search space, mitigating the semantic drift highlighted in section 1. In addition, the results on MedDDX are particularly significant, improving the accuracy by 6.1%, 5.9% and 4.4%. MedDDX is specifically designed to assess fine-grained discrimination among semantically similar answer candidates, such as drugs with overlapping mechanisms of action. In this high-difficulty scenario, Med-SRAF significantly outperforms all baselines. This significant improvement is attributed to our Agentic Fusion Mechanism (Phase III). While baseline

Datasets	Medical QA (**Open-ended**)				MedDDx (**Open-ended**)		
	MMLU	MedQA	PubMedQA	BioASQ	Basic	Intermediate	Expert
Metrics	Acc. (ΔAcc.)	Acc. (ΔAcc.)	Acc. (ΔAcc.)	Acc. (ΔAcc.)	Acc. (ΔAcc.)	Acc. (ΔAcc.)	Acc. (ΔAcc.)
LLaMA3.1-8B	0.601 $_{(-0.098)}$	0.572 $_{(-0.033)}$	0.507 $_{(-0.065)}$	0.655 $_{(-0.030)}$	0.322 $_{(-0.119)}$	0.320 $_{(-0.081)}$	0.371 $_{(-0.013)}$
Self-RAG (13B)	0.204 $_{(-0.030)}$	0.220 $_{(-0.044)}$	0.173 $_{(-0.001)}$	0.581 $_{(-0.031)}$	0.192 $_{(-0.043)}$	0.215 $_{(-0.038)}$	0.216 $_{(-0.008)}$
MedRAG (13B)	0.514 $_{(-0.020)}$	0.392 $_{(-0.018)}$	0.475 $_{(-0.018)}$	0.592 $_{(+0.007)}$	0.354 $_{(+0.014)}$	0.308 $_{(-0.010)}$	0.309 $_{(-0.017)}$
KGARevion (LLaMA3.1-8B)	0.703 $_{(-0.025)}$	0.574 $_{(-0.024)}$	0.615 $_{(+0.012)}$	0.767 $_{(+0.011)}$	0.459 $_{(+0.015)}$	0.432 $_{(-0.010)}$	0.418 $_{(-0.020)}$
Med-SRAF (LLaMA3.1-8B)	**0.767** $_{(-0.002)}$	**0.634** $_{(-0.006)}$	**0.649** $_{(-0.003)}$	**0.838** $_{(+0.003)}$	**0.475** $_{(-0.028)}$	**0.492** $_{(-0.007)}$	**0.460** $_{(-0.015)}$

Table 2: The accuracy results of our method and baseline methods under open-ended reasoning settings. ΔAcc. denotes the difference in performance between the open-ended and multiple-choice reasoning settings.

models often failed due to the vector similarity of interfering options, our approach reconstructs explicit logical chains via structured triplets. This allows the model to capture subtle semantic distinctions and causal relationships that are often lost in unstructured text retrieval, which demonstrates exceptional robustness in scenarios requiring professional-grade differential diagnosis.

3.3 Open-Ended Reasoning

Table 2 consistently demonstrates that Med-SRAF outperforms all compared baselines across both general medical benchmarks (MMLU, MedQA, PubMedQA, BioASQ) and progressively more challenging diagnostic reasoning settings (MedDDx-Basic, Intermediate, and Expert). These results suggest that the proposed structured routing and fusion mechanisms generalize effectively beyond closed-ended benchmarks, with performance gains becoming more pronounced as reasoning complexity increases.

3.4 Ablation Study

The effect of semantic routing. To validate whether domain-specific semantic routing is superior to generic query decomposition strategies, we designed experiments with three different settings. Notably, to isolate the contribution of the routing phase, the FusionAgent was excluded from all settings, and the retrieved evidence was directly concatenated for the final solver:

(1) **Semantic Routing (Ours)**: The NavigationAgent analyzes the query intent and dynamically activates only a subset of relevant dimension-specific agents (e.g., selecting only Disease and Symptom) for sparse retrieval;

(2) **Generic Query Decomposition (w/o Semantic Routing)**: The NavigationAgent is replaced by a standard Chain-of-Thought (CoT) decomposition module without access to predefined medical dimensions;

(3) **Full Activation**: The routing mechanism is bypassed, and all five dimension-specific agents are forcibly activated for every query.

As shown in Table 3, the results underscore the critical role of the Semantic Routing Mechanism. Our method outperforms Generic Query Decomposition by an average of 3.4%. This improvement indicates that without domain-specific priors (i.e., the five medical dimensions), generic decomposition tends to generate divergent or less relevant sub-queries, failing to capture the precise medical intent. And our approach achieves a 3.3% accuracy gain over the Full Activation baseline. It shows our Semantic Routing Mechanism perform effective dimensional pruning to reduce the irrelevant noise and enhance reasoning precision.

The effect of Agentic Fusion. To evaluate the explicit conflict resolution and fusion mechanism (FusionAgent) outperforms simple context splicing in multi-source retrieval scenarios, we designed experiments with two different settings:

(1) **Med-SRAF (Ours)**: The retrieved triplets from selected sub-agents are processed by the FusionAgent, which applies semantic deduplication, conflict resolution (based on clinical specificity), and relevance filtering to generate unified evidence for the solver;

(2) **Med-SRAF (w/o Agentic Fusion)**: The FusionAgent is removed. All raw triplets retrieved by the active dimension agents are directly concatenated and fed into the solver without any intermediate logic or filtering.

As shown in Table 3, the accuracy improvement in our Med-SRAF exceeded 2.8% after incorporating the Agentic Fusion Mechanism. This demonstrates that naive concatenation of multi-source evidence is inadequate for complex medical reasoning, especially when there are conflicts among the retrieved information. By leveraging strategies such as conflict resolution and clinical specificity filtering, our mechanism functions as a rigorous logical consistency filter, ensuring the final solver operates

Datasets	Multiple Medical QA				MedDDx		
	MMLU	MedQA	PubMedQA	BioASQ	Basic	Intermediate	Expert
Metrics	Acc. (std)	Acc. (std)	Acc. (std)	Acc. (std)	Acc. (std)	Acc. (std)	Acc. (std)
Semantic Routing (Ours)	0.578 (.002)	0.441 (.003)	0.524 (.005)	0.649 (.007)	0.392 (.015)	0.362 (.001)	0.341 (.007)
Generic Query Decomposition (w/o Semantic Routing)	0.526 (.004)	0.409 (.003)	0.490 (.003)	0.581 (.003)	0.361 (.008)	0.347 (.005)	0.329 (.005)
Full Activation	0.534 (.005)	0.410 (.004)	0.495 (.006)	0.598 (.005)	0.341 (.010)	0.349 (.004)	0.329 (.011)
Med-SRAF (Ours)	0.615 (.004)	0.489 (.002)	0.545 (.008)	0.681 (.008)	0.423 (.009)	0.374 (.001)	0.362 (.005)
Med-SRAF (w/o Agentic Fusion)	0.578 (.002)	0.441 (.003)	0.524 (.005)	0.649 (.007)	0.392 (.015)	0.362 (.001)	0.341 (.007)

Table 3: The Results under Different Settings. All approaches above are evaluated with LLama2-13B.

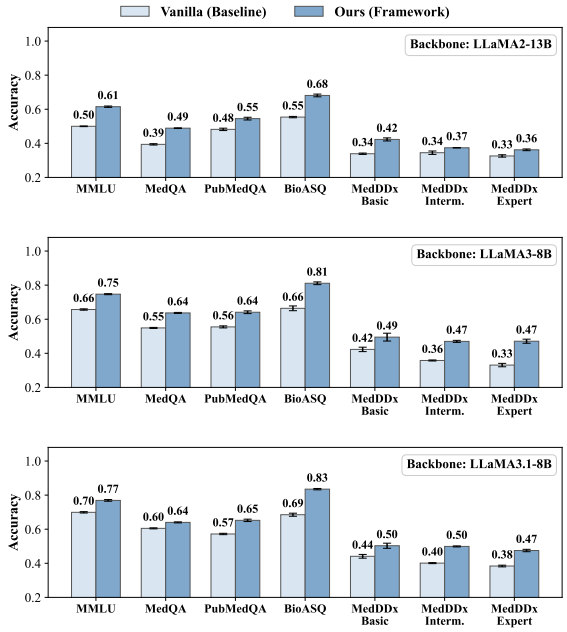

Figure 4: **Effect of Backbone Models.** Performance comparison across three backbone models (LLaMA2-13B, LLaMA3-8B, LLaMA3.1-8B).

on a coherent knowledge set rather than a noisy mixture of contradictory facts.

3.5 Evaluation with different backbones

To evaluate the versatility and robustness of our method, we implemented Med-SRAF across different Large Language Model (LLM) backbones. Specifically, we benchmarked the performance gain on three models: LLaMA2-13B, LLaMA3-8B and LLaMA3.1-8B. As illustrated in Figure 4, our Med-SRAF consistently enhances the performance of different backbone LLMs, improving average accuracy by over 7.8% with LLaMA2-13B, over 10.4% with LLaMA3-8B, and over 8.2% with LLaMA3.1-8B. These improvements underscore the adaptability of Med-SRAF across different model architectures, significantly boosting the reasoning capabilities of foundation models in knowledge-intensive medical QA tasks.

3.6 Case Study

To explicitly demonstrate how Med-SRAF mitigates semantic drift and concatenation fallacy, we present the analysis of a representative sample: *"Does the manic/mixed episode distinction in bipolar disorder patients run true over time?"*.

Failure of Generic RAG. As illustrated in the Figure 1, generic query decomposition lacking domain constraints exhibited significant semantic drift. The generic model generated irrelevant sub-queries such as *"What are the potential consequences of misclassifying a manic or mixed episode in bipolar disorder patients?"*. While linguistically related to the topic, it shifts the focus from *diagnostic stability* to *social or clinical consequences*, introducing noise documents that overwhelmed the context window and led to retrieval failure.

Success of Med-SRAF. As shown in Figure 5, our framework successfully solved this complex reasoning problem through its three-phase workflow:

- Phase I: Intent-Driven Routing. The NavigationAgent accurately identified the query's core intent as pertaining to "disease classification" and "clinical symptoms" definitions. Consequently, it activated only the Disease (d_{dis}) and Symptom (d_{sym}) dimensions, while explicitly pruning the irrelevant Treatment (d_{treat}), Examination (d_{exam}) and Mechanism (d_{mech}) dimensions. This strategic pruning physically blocked the ingestion of noise related to drug side effects or misdiagnosis consequences.

- Phase II: Generative Retrieval. Upon activation, the selected Sub-Agents executed retrieval tasks in parallel. The DiseaseAgent targeted the temporal stability aspect (e.g., *Prognosis and long-term outcomes...*), retrieving the contextual evidence (*Bipolar disorder, evaluated in, Longitudinal studies*). Simultaneously, the SymptomAgent focused on phenomenological distinctions, generating queries such as *What clinical manifestations differentiate manic from mixed episodes...*,

retrieving the definitive triplet (*Manic episodes, different from, Mixed episodes*).

- Phase III: Agentic Fusion. This case highlights the critical role of the FusionAgent in resolving inter-source knowledge conflicts. During parallel retrieval, the sub-agents retrieved two contradictory evidence triplets: From d_{sym}: (*Manic episodes, different from, Mixed episodes*). From d_{dis}: (*Manic Episode, characterized by, Mixed Episode*) A naive concatenation strategy would force the LLM to process these conflicting assertions simultaneously, likely leading to hallucination. However, the FusionAgent detected this logical contradiction. By applying a clinical specificity filter, the FusionAgent reasoned that the relationship "different from" in the symptom dimension offers a precise diagnostic distinction, whereas "characterized by" in the disease dimension was vague and contradictory in this context. Consequently, it discarded the latter and synthesized a coherent reasoning chain confirming that longitudinal studies support the distinct nature of these episodes, leading to the correct answer.

4 Conclusion

In this paper, we propose Med-SRAF, a multiagent framework designed to enhance the reliability of Retrieval-Augmented Generation in complex medical reasoning. By introducing intent-driven semantic routing and evidence-based agentic fusion, our framework effectively addresses the dual challenges of semantic drift during query decomposition and knowledge conflicts inherent in multi-path retrieval that plague general-purpose RAG systems. Extensive experiments and ablation studies on five multiple-choice benchmarks demonstrate the effectiveness of our Med-SRAF.

Beyond performance gains, this work offers a valuable perspective for high-stakes AI applications: integrating domain priors (such as medical reasoning dimensions) into agentic workflows significantly enhances system interpretability and trustworthiness. As our case studies demonstrate, Med-SRAF delivers not only accurate predictions but also transparent, verifiable reasoning traces.

Constrained by the textual nature of current benchmarks, our Med-SRAF presently targets textual reasoning. We plan to extend it to a multimodal framework by integrating visual data (e.g., medical imaging) directly into the routing and fusion pipelines and to explore more autonomous planning mechanisms to handle complex clinical queries with minimal human intervention.

Limitations

Although the significant performance improvements achieved by our Med-SRAF, there are still limitations needed to be addressed. Compared to the strongest baseline, Med-SRAF requires approximately 13.9% more processing time per query. However, considering that this moderate temporal cost yields an accuracy improvement of over 6.2%, we regard it as a justifiable trade-off for complex medical reasoning.

Acknowledgements

This work was partially supported by Natural Science Foundation of China (No.62502491; No.62072427), the Project of Stable Support for Youth Team in Basic Research Field, CAS (YSBR-005), Anhui Science Foundation for Distinguished Young Scholars (No.1908085J24), Jiangsu Natural Science Foundation (No. BK20191193). The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Science.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Manar Aljohani, Jun Hou, Sindhura Kommu, and Xuan Wang. 2025. A comprehensive survey on the trustworthiness of large language models in healthcare. *arXiv preprint arXiv:2502.15871*.

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. 2025. The distracting effect: Understanding irrelevant passages in rag. *arXiv preprint arXiv:2505.06914*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Ines Besrour, Jingbo He, Tobias Schreieder, and Michael Färber. 2025. Ragenta: Multi-agent retrieval-augmented generation for attributed question answering. *arXiv preprint arXiv:2506.16988*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, et al. 2025. Main-rag: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2607–2622.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

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):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. 2024. Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization. *arXiv preprint arXiv:2410.08815*.

Kevin Lin, Kyle Lo, Joseph Gonzalez, and Dan Klein. 2023. Decomposing complex queries for tip-of-the-tongue retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5521–5533.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Teng Lin, Yizhang Zhu, Yuyu Luo, and Nan Tang. 2025. Srag: Structured retrieval-augmented generation for multi-entity question answering over wikipedia graph. *arXiv preprint arXiv:2503.01346*.

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.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Salvador Lima López, Eulália Farré-Maduell, Luis Gasco, Martin Krallinger, and Georgios Paliouras. 2023. Overview of bioasq 2023: The eleventh bioasq challenge on large-scale biomedical semantic indexing and question answering. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 227–250. Springer.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Roxana Petcu, Kenton Murray, Daniel Khashabi, Evangelos Kanoulas, Maarten de Rijke, Dawn Lawrie, and Kevin Duh. 2025. Query decomposition for rag: Balancing exploration-exploitation. *arXiv preprint arXiv:2510.18633*.

Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clevert, and Marinka Zitnik. 2024. Kgarevion: an ai agent for knowledge-intensive biomedical qa. *arXiv preprint arXiv:2410.04660*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

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.

Rui Yang, Haoran Liu, Edison Marrese-Taylor, Qingcheng Zeng, Yuhe Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, et al. 2024. Kg-rank: Enhancing large language models for medical qa with knowledge graphs and ranking techniques. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 155–166.

Peiying Yu, Guoxin Chen, and Jingjing Wang. 2025. Table-critic: A multi-agent framework for collaborative criticism and refinement in table reasoning. *arXiv preprint arXiv:2502.11799*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.

Yunfei Zhong, Jun Yang, Yixing Fan, Lixin Su, Maarten de Rijke, Ruqing Zhang, and Xueqi Cheng. 2025. Reasoning-enhanced query understanding through decomposition and interpretation. *arXiv preprint arXiv:2509.06544*.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. 2025. Knowledge graph-guided retrieval augmented generation. *arXiv preprint arXiv:2502.06864*.

Appendix

A Datasets

We evaluate our system using five representative multi-choice medical QA datasets: MMLU (Hendrycks et al., 2020), MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), BioASQ (Nentidis et al., 2023), and MedDDx (Su et al., 2024). Notably, the MedDDx dataset is further categorized into three complexity levels: Basic, Intermediate, and Expert, to test the model's depth of clinical reasoning. Detailed statistics, including sample counts and option formats for each dataset, are summarized in Table 4, while representative examples are provided in Table 5.

Dataset	Count	Options
MMLU	1,089	A/B/C/D
MedQA	1,273	A/B/C/D
PubMedQA	500	Yes/No/Maybe
BioASQ	618	Yes/No
MedDDx-Basic	483	A/B/C/D
MedDDx-Intermediate	1,041	A/B/C/D
MedDDx-Expert	245	A/B/C/D

Table 4: The size and options of datasets

B Related Works

Structured RAG. Traditional Naive RAG typically retrieves the $top - k$ unstructured paragraphs based on semantic similarity (e.g., dense retrieval) (Lewis et al., 2020; Karpukhin et al., 2020). However, these approaches are highly susceptible to retrieval noise, meaning that irrelevant documents can affect the inference of large models (Amiraz et al., 2025). Furthermore, unstructured documents aggravate the "Lost-in-the-Middle" phenomenon (Liu et al., 2024). To mitigate these issues, recent research has shifted towards Structured RAG, introducing explicit symbolic representations (e.g., graphs, tables) to organize evidence. For instance, GraphRAG (Edge et al., 2024), StructRAG (Li et al., 2024), S-RAG (Lin et al., 2025) and KG-Guided RAG (Zhu et al., 2025) demonstrate that structured evidence enables more controllable aggregation and reliable multi-hop reasoning. Notably, KGARevion (Su et al., 2024) employs an agent-based "Generate-Review-Revise" framework to construct high-quality triplets. Despite these advancements, retrieving valid and helpful structured evidence remains a challenge, as most methods lack a high-level routing mechanism to lock the retrieval scope.

Multi-Agent RAG. To better support complex reasoning, recent work has extended conventional single-agent RAG into multi-agent frameworks, where multiple specialized agents collaborate to exploit complementary capabilities. (Chang et al., 2025) proposed MAIN-RAG to filter and score retrieved documents through multi-agent collaboration. RAGentA by (Besrour et al., 2025) iteratively filters retrieved documents by combining a multi-agent framework with hybrid retrieval, generating attributed answers with inline references. And, strategies like Multi-Agent Debate (Du et al., 2023) allows multiple agents to provide their own responses and reasoning processes, thereby reaching consensus through multiple rounds of discussion. Similarly, TableCritic (Yu et al., 2025) iteratively improves the collaborative critique and reasoning process among multiple agents until it converges to the correct solution. However, a critical limitation persists in the integration phase: most existing systems rely on majority voting or simple concatenation to aggregate evidence.

Query Decomposition. Decomposing complex queries into manageable sub-tasks is essential for effective medical reasoning. Recent works have explored various decomposition strategies. (Lin et al., 2023) breaks down the query into individual clues, retrieves information using a specialized search engine, and then integrates the retrieved information. And ReDI by (Zhong et al., 2025) leverages the reasoning and

Dataset	Sample
MMLU	A lesion causing compression of the facial nerve at the stylomastoid foramen will cause ipsilateral: A: paralysis of the facial muscles. B: paralysis of the facial muscles and loss of taste. C: paralysis of the facial muscles, loss of taste and lacrimation. D: paralysis of the facial muscles, loss of taste, lacrimation and decreased salivation.
MedQA	A junior orthopaedic surgery resident is completing a carpal tunnel repair with the department chairman as the attending physician. During the case, the resident inadvertently cuts a flexor tendon. The tendon is repaired without complication. The attending tells the resident that the patient will do fine, and there is no need to report this minor complication that will not harm the patient, as he does not want to make the patient worry unnecessarily. He tells the resident to leave this complication out of the operative report. Which of the following is the correct next action for the resident to take? A: Disclose the error to the patient and put it in the operative report. B: Tell the attending that he cannot fail to disclose this mistake. C: Report the physician to the ethics committee. D: Refuse to dictate the operative report
PubMedQA	Is oral endotracheal intubation efficacy impaired in the helicopter environment? A: yes. B: no. C: maybe.
BioASQ	Are there microbes in human breast milk? A: yes. B: no
MedDDx-Basic	Would you be able to provide information on whether the specified medication acts on genes or proteins that play a role in the signaling pathways of blood clotting receptors, particularly those associated with protein tyrosine kinase activator activity, given the importance of anticoagulation for specific medical conditions? A: Heparin Disaccharide Iii-S. B: Heparin Disaccharide I-S. C: Human blood group H type 1 trisaccharide. D: H TYPE II TRISACCHARIDE.
MedDDx-Intermediate	Could you suggest any medications effective for treating beriberi that are safe to use with Zopiclone? A: Niacinamide ascorbate. B: N-(4-phenoxyphenyl)-2-[(pyridin-4-ylmethyl)amino]nicotinamide. C: N-methylnicotinamide. D: Nicotinamide.
MedDDx-Expert	Which genes or proteins are expressed exclusively in the pericardium and not in either the dorsal or ventral regions of the thalamus? A: ADH1A. B: ADH1C. C: ADH4. D: ADH1B.

Table 5: Samples of five multi-choice medical QA datasets

comprehension capabilities of LLMs to break down complex queries into targeted sub-queries and enrich each sub-query with detailed semantic interpretations. (Petcu et al., 2025) combines query decomposition with information retrieval in an exploitation-exploration setting, where retrieving one document at a time builds a belief about the utility of a given sub-query and informs the decision to continue exploiting or exploring an alternative. Nevertheless, standard decomposition methods are often unconstrained, suffering from "Semantic Drift" in domain-specific applications.

C Implementions Details

C.1 Experimental Environments

Hadrware. All experiments were conducted on a high-performance computing server equipped with 8 NVIDIA A800 GPUs. During inference, we adopted a dynamic resource allocation strategy: we allocated 4 NVIDIA A800 GPUs for large-scale language models (e.g., LLaMA-2-13B, LLaMA-3, LLaMA-3.1) to ensure efficient and stable generation under high concurrency. For smaller-scale models, a single NVIDIA A800 GPU was sufficient to complete the inference process. This strategy allows for optimized resource efficiency and reduced computational overhead, ensuring experimental reproducibility while effectively balancing performance and cost.

Software. Our framework is implemented in Python 3.8+ and is designed for fully local deployment to ensure data privacy and reproducibility. We utilized the ollama framework as the unified inference backend, which serves quantized GGUF versions of LLMs and exposes an OpenAI-compatible API. This setup allows for seamless interaction via the `openai` Python library while avoiding cloud-based latency. To support the high concurrency required by our multi-agent architecture, we employed Python's `concurrent.futures` module to implement a `ThreadPoolExecutor`. This mechanism enables the *NavigationAgent* to dispatch tasks to downstream dimension agents (e.g., Disease, Symptom) in parallel, effectively utilizing the multi-GPU hardware setup. Additionally, PyTorch (v2.0.1) is used to load the external Knowledge Graph Augmentation model (KGARevion), and the `backoff` library is incorporated to implement exponential retry strategies, ensuring system stability during intensive inference tasks.

C.2 Baselines

C.2.1 General-purpose LLMs

We chose the LLaMA series as General-purpose LLMs in this paper. LLama2(7B, 13B) (Touvron et al., 2023), developed by Meta, is a series of open-source large language models that build upon the original LLaMA architecture. The 7B and 13B versions denote models with 7 billion and 13 billion parameters, respectively. These models are pre-trained on massive internet text corpora, demonstrating powerful language understanding and generation capabilities. Their open-source nature has led to their widespread use in academia and industry, making them an effective backbone for retrieval-enhanced generative (RAG) systems. Building upon this, LLaMA 3 (8B) (Dubey et al., 2024) represents a significant leap in performance, trained on a massive corpus of over 15 trillion tokens (approximately 7$\times$ that of LLaMA 2). Despite its smaller parameter size, it incorporates architectural optimizations such as Grouped Query Attention (GQA) and a larger vocabulary, enabling it to outperform many larger predecessors. Its successor, LLaMA 3.1 (8B), further extends the context window to 128k tokens. This long-context capability is particularly critical for our medical RAG framework, as it allows the model to comprehensively analyze extensive retrieved guidelines and case reports without truncation.

C.2.2 Medically-tuned LLMs

To assess the effectiveness of parametric knowledge injection versus our retrieval-based approach, we compared two representative domain-adapted models. MedAlpaca-7B (Han et al., 2023) represents the *instruction tuning* paradigm. Built upon the LLaMA architecture, it is fine-tuned on a diverse collection of medical tasks, including Q&A pairs and dialogue data, designed to align the model's behavior with clinical interaction patterns. In contrast, PMC-LLaMA-7B (Wu et al., 2024) adopts a *continued pre-training* strategy. By training on 4.8 million biomedical academic papers (S2ORC) and medical textbooks, it

aims to internalize raw domain knowledge directly into the model's parameters. Comparing these models allows us to evaluate the limits of static internal knowledge in handling complex, evolving medical queries.

C.2.3 RAG-based Models

These baselines represent mainstream solutions for mitigating hallucinations through external text retrieval. Self-RAG (7B/13B) (Asai et al., 2023) introduces an advanced self-reflection framework. Unlike standard RAG, it trains the model to generate "reflection tokens" that dynamically critique the necessity of retrieval and the relevance of retrieved documents. This serves as a state-of-the-art baseline for autonomous retrieval-augmented generation. MedRAG (13B) (Xiong et al., 2024) serves as a specialized benchmark for medical retrieval. It combines a robust retriever with authoritative medical corpora (e.g., PubMed, textbooks) to ground the model's responses. We include MedRAG to benchmark the performance of standard "Retriever-Reader" pipelines against our structured multi-agent approach.

C.2.4 KG-based Models

These models share our focus on structured knowledge but differ in their utilization mechanisms (Re-ranking vs. Verification). KG-Rank (13B) (Yang et al., 2024) integrates Knowledge Graphs (KGs) to enhance the retrieval stage. Instead of utilizing KGs for reasoning, it leverages the graph structure to re-rank retrieved documents, filtering out noise and prioritizing evidence that aligns with medical facts. KGARevion (LLaMA3/3.1-8B) (Su et al., 2024) represents a state-of-the-art KG-agent specifically designed for biomedical QA. Unlike standard retrieval, it employs a "Generate-Review-Revise" pipeline, where an LLM generates structured "Entity-Relation-Entity" triplets that are rigorously verified against a KG before answer generation.

C.2.5 Proprietary LLMs

To further contextualize the performance of our approach against state-of-the-art commercial systems, we include two strong proprietary models: GPT-4 and Gemini-Pro. These models are closed-source and accessed via API, representing the upper bound of current general-purpose LLM capabilities in real-world deployments. GPT-4 is a highly capable multimodal model trained with large-scale supervised learning and reinforcement learning from human feedback (RLHF). It demonstrates strong performance across a wide range of reasoning-intensive tasks, including medical question answering, making it a competitive baseline for both multi-choice and open-ended settings. Gemini-Pro, a deployment variant of the Gemini family, is designed for efficient and scalable inference while maintaining strong reasoning and knowledge integration capabilities. It benefits from large-scale multimodal pre-training and advanced architectural optimizations, enabling competitive performance across diverse domains. Including these proprietary models allows us to benchmark Med-SRAF against industry-level systems and assess whether our structured retrieval and reasoning framework can remain competitive even when compared to the most advanced closed-source LLMs.

D Data-Driven Justification for Optimality and Quasi-Orthogonal

To quantitatively verify that this 5-dimension taxonomy is an optimal, non-redundant basis set, we randomly sampled 100 queries and measured the semantic overlap (Jaccard similarity) among the sub-queries generated for different activated dimensions. The average cross-dimension token overlap is **21.4%**. We emphasize that this moderate overlap is not a sign of redundancy, but rather a necessary architectural feature. For instance, in a complex query evaluating the APACHE II score, the NavigationAgent co-activated the Examination and Disease dimensions. Our token-level analysis reveals a Jaccard similarity of **17.6%**. Crucially, this overlap consists almost entirely of **shared core entity anchors** necessary to prevent semantic drift (e.g., 'apache', 'emergency', 'patient'). Meanwhile, the **functional reasoning keywords (nearly 80% of the tokens) strictly diverge**: the Examination agent exclusively explores objective metrics (e.g., 'laboratory findings'), whereas the Disease agent targets pathological outcomes (e.g., 'mortality'). This empirically verifies that our taxonomy optimally prunes redundant paths while preserving essential entity constraints.

To prove that these 5 dimensions are sufficient to cover nuanced scenarios without missing context, we evaluated the NavigationAgent as a multi-label routing task on the 100-query subset. It achieved an exceptional **Recall of 98.8%**, indicating the framework almost never encounters a pathway outside its expressive capability.

E Routing Accuracy and Error Recoverability

Since there is no existing benchmark with dimension-routing ground truth, we conducted a new evaluation on a subset of 100 human-annotated queries randomly sampled from the benchmarks. We formulated this as a multi-label classification task and calculated the sample-averaged metrics. The NavigationAgent achieved a strong **Precision of 84.5%**, an exceptional **Recall of 98.8%**, and an **F1-score of 91.1%**. Crucially, this near-perfect recall empirically demonstrates that our system successfully captures almost all essential reasoning pathways, preventing critical semantic drift and fatal omissions at the very source. While the precision of 84.5% indicates a slight tendency toward a conservative routing strategy that occasionally over-activates unneeded dimensions (False Positives), we respectfully clarify that such routing errors are highly **recoverable** by design. Our framework operates precisely as a "High-Recall Routing + Precision Filtering" pipeline; any noisy or irrelevant triples retrieved from these over-activated dimensions are explicitly intercepted and discarded by the Relevance Filtering mechanism within the Phase III FusionAgent, ensuring that the routing stage acts as a robust, fault-tolerant safety net rather than a fragile single point of failure.

F Robustness against Wording and Ordering Variations

To address the gap in prompt sensitivity, we conducted a systematic prompt perturbation study focusing on the core NavigationAgent and FusionAgent. We evaluated the LLaMA3-8B backbone on a randomly sampled subset of 500 MedQA questions using three prompt configurations:

- $V_{original}$: The original prompt, including role-playing instructions and structured guidance.
- $V_{concise}$: A stripped-down version that removes all role-playing and explanatory scaffolding, retaining only the minimal core task description.
- $V_{rephrased}$: A semantically equivalent but lexically and syntactically rewritten version, preserving instruction complexity while altering wording and rule ordering.

Prompt Variant	**Avg. Acc. (%)**
Original	$\mathbf{76.1 \pm 0.1}$
Concise	74.6 ± 0.3
Rephrased	75.1 ± 0.1

Table 6: **Robustness across Prompt Variants.** Performance comparison of Med-SRAF under different prompt formulations.

The results indicate that Med-SRAF maintains **stable performance across prompt variants**, with all configurations falling within a narrow accuracy range. Removing auxiliary role-playing and explanatory scaffolding (*Concise*) leads to a modest performance drop, suggesting that such scaffolding provides incremental benefits but is not essential for correct system behavior. The *Rephrased* variant yields performance closer to *Concise* than to the original prompt, indicating that the observed robustness primarily stems from the **structural decomposition of reasoning roles**, rather than from specific wording or stylistic choices. Overall, these findings suggest that Med-SRAF does not rely on fragile prompt formulations, while richer prompt structure can offer additional, but non-critical, gains.

G Efficiency Analysis

To evaluate the computational cost of our framework, we compared the average inference latency per query of our Med-SRAF against the strongest baseline KGARevion. As shown in Table 7, our method incurs an average latency of 101.75 seconds per query, representing a moderate increase of approximately 13.9% compared to KGARevion (89.32 seconds). This additional temporal cost primarily stems from the Agentic Fusion phase. However, this investment yields a highly favorable return: in exchange for a 13.9% increase in latency, Med-SRAF achieves an average accuracy improvement exceeding 6.2%. Given the high-stakes nature of clinical diagnosis where decision precision is paramount, investing marginal computational time to secure such significant accuracy gains is a justifiable and necessary trade-off.

Method	Avg. Latency (s)	Avg. Acc. (%)
KGARevion	89.32	54.8
Med-SRAF (Ours)	101.75	**61.0**

Table 7: **Efficiency vs. Performance Trade-off.** Comparison of average inference latency (seconds per query) and average accuracy between the strongest baseline KGARevion and our Med-SRAF. Both methods utilize LLaMA3-8B as the backbone. Despite a marginal increase in latency, Med-SRAF achieves superior accuracy.

H Case Study Figure

Figure 5 provides a visualization of the successful reasoning for the case discussed in subsection 3.6. While this specific sample failed under traditional query decomposition due to Semantic Drift, Med-SRAF successfully derives the correct answer through its structured workflow.

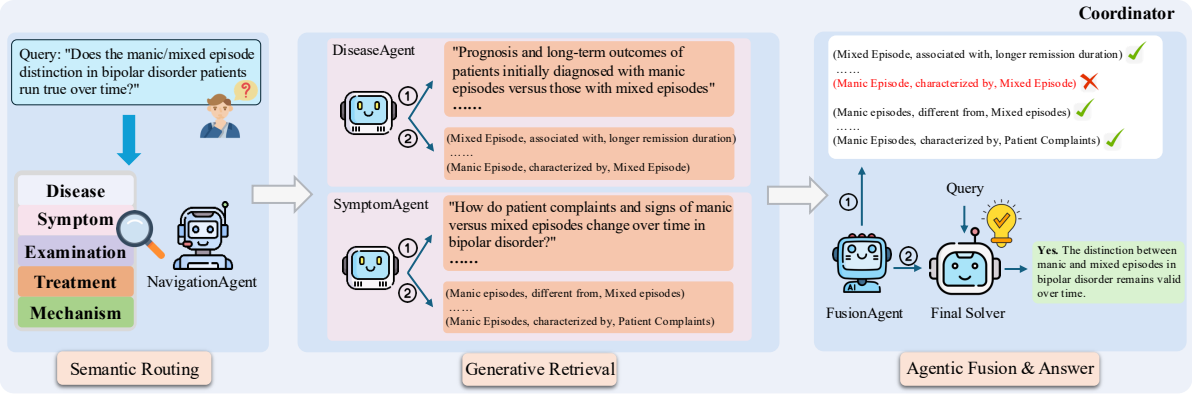

Figure 5: The same samples processed by SRAF.

I Real-world testing

To better contextualize the reported gains, we conducted additional evaluations on clinically collected case data that reflect realistic deployment conditions, including noisy, incomplete, and uncontrolled information.

Compared to curated benchmark datasets, such settings pose substantially greater challenges for retrieval-augmented methods, under which conventional RAG-based approaches often experience pronounced performance degradation. We therefore evaluated Med-SRAF against strong GPT-4–based baselines under identical settings. Due to patient privacy constraints, this dataset cannot be publicly released. The results are summarized in Table 8.

Med-SRAF achieves a substantial improvement over all baselines in this deployment-motivated setting. While we do not claim this evaluation to be a comprehensive clinical validation, the results provide evidence that the proposed framework is **more resilient to noise and evidence uncertainty**, which are common in real-world clinical workflows.

Method	Accuracy (%)
GPT4 + MedRAG	54.2
GPT4 + KGARevion	61.7
GPT4 + Med-SRAF (Ours)	**69.2**

Table 8: **Evaluation on real-world clinical case data.** Comparison of Med-SRAF with GPT-4–based retrieval-augmented baselines under realistic deployment conditions. The results show that Med-SRAF achieves superior performance in noisy and incomplete clinical scenarios.

J Regarding the use of LLMs

LLMs were employed solely for polishing the writing and enhancing readability. All scientific ideas, data, and analyses presented herein are human-generated and original.

K Prompts

Prompts for General-purpose LLMs

system_prompt:
You will receive a medical question. Please answer it. Return the answer strictly in JSON format, where the value is a single letter from the options (A/B/C/D).

user_prompt:
Question: {question}
Options: {options}
Answer:

Prompts for General-purpose LLMs with CoT

system_prompt:
You will receive a medical question. Please answer it. Think step by step. Return the answer strictly in JSON format, where the value is a single letter from the options (A/B/C/D).

user_prompt:
Question: {question}
Options: {options}
Answer:

Prompts for NavagationAgent

system_prompt:
You are a medical query analysis specialist. Your task is to analyze medical multiple-choice questions and select the most relevant dimensions for comprehensive analysis.

AVAILABLE DIMENSIONS:
1. disease: Questions focused on diagnosis, disease identification, etiology, prognosis, or risk factors.
2. symptom: Questions focused on clinical manifestations, patient complaints, signs, and presentation features.
3. examination: Questions focused on diagnostic investigations including laboratory tests, imaging, pathology, physical examination findings, and diagnostic criteria.
4. treatment: Questions focused on therapeutic decisions, medications, procedures, surgeries, or clinical management steps.
5. mechanism: Questions focused on biological, pathological, physiological, or pharmacological mechanisms, including mechanism of action (MOA), pathophysiology, metabolic pathways, enzyme regulation, or causal explanations.

IMPORTANT:
- You MUST return dimension names in LOWERCASE only: "disease", "symptom", "examination", "treatment", "mechanism"
- Choose 1–3 dimensions that best match the primary logic needed to answer the question.

user_prompt:
Question: {question}
Options: {options}
Selected_dimensions:

Prompts for Sub-Query Generation

system_prompt:
You are a medical research assistant.

user_prompt:
Task: Generate 1-3 specific search queries to help verify if the [Options] are the correct answer to the [Question]. The queries must link the clinical clues in the question to the provided options.
Question: {question}
Options: {options}
Dimension Requirement: {dimension_instruction}
Sub-Queries:

Prompts for Cross-Dimensional Knowledge Synthesis

system_prompt:
You are a strict Medical Evidence Integrator. Your goal is to MERGE these fragmented facts into a unified, high-precision knowledge set to answer the specific query.
[CORE PRINCIPLES]
1. **Deduplication**: Merge identical facts (e.g., [A, causes, B] and [A, leads to, B] -> merge into one).
2. **Conflict Resolution**: If Dimension A and Dimension B contradict, prioritize the source with higher clinical specificity or standard textbook definition.
3. **Relevance Filter**: STRICTLY discard any triplets that are irrelevant to the user's specific question.
4. **Connectivity**: If Dim A provides X->Y and Dim B provides Y->Z, preserve this chain to show X->Y->Z.

user_prompt:
[FUSION EXECUTION STEPS]
1. Analyze the User Query to understand the "Target Unknown".
2. Filter the Input Triplets: Keep only those that help differentiate options or confirm the diagnosis.
3. Merge & Standardize: Unify entity names to standard medical terminology (e.g., "stomach pain" -> "Abdominal Pain").
4. Output the final minimal set of fused triplets.

Question: {question}
Options: {options}
Formatted_Triplets: {formatted_triplets}
Fused_Triplets:

Prompts for Single-Source Data Refinement

system_prompt:
You are a Medical Data Analyst. Your goal is to CLEAN, STANDARDIZE, and REFINE triplets you received.
[CORE PRINCIPLES]
1. **NO HALLUCINATION**: Do NOT invent new triplets that are not supported by the input or standard medical consensus.
2. **Precision**: Convert vague relationships into precise medical predicates (e.g., change "related to" to "causes" or "is symptom of" if clearly supported).
3. **Filtering**: Remove noise, redundant entries, or triplets that contain empty/meaningless values (like "None").
4. **Consistency**: Ensure the triplets are logically consistent with the patient's presentation in the query.

user_prompt:
[REFINEMENT EXECUTION STEPS]
1. Review the Input Triplets against the User Query.
2. discard irrelevant noise.
3. Standardize entities (e.g., capitalize Disease names, unify symptom descriptions).
4. Return the refined list.

Question: {question}
Options: {options}
Formatted_Triplets: {formatted_triplets}
Fused_Triplets: