

Med-Chain: An Adaptive Clinical-Chain Retrieval Framework for Evidence-Grounded Medical Question Answering

Fucheng Niu¹, Songtao Jiang¹, Qinxin Wu¹, Sylvia Chung Yan Shan¹,
Tianxiang Hu¹, Jian Wu¹, Zuozhu Liu^{1,*}

¹Zhejiang University, Zhejiang, China

*Corresponding author

Abstract

Large Language Models (LLMs) have shown strong potential in medical question answering, but their reliance on static parametric knowledge limits their ability to handle rapidly evolving clinical evidence. The retrieval system mitigates this issue by incorporating external medical knowledge, but flat retrieval can return lexically relevant but clinically misaligned evidence. We propose Med-Chain, a training-free medical evidence retrieval framework that reformulates retrieval as inspectable and repairable clinical evidence paths. Med-Chain first routes each question through a clinical topic tree to progressively narrow retrieval from broad domains to specific guideline sections. It then verifies whether the retrieved path matches the clinical intent and performs level-aware refinement to repair mismatches. Experiments on seven medical QA benchmarks show that Med-Chain consistently improves backbone LLMs and outperforms existing medical Retrieval-augmented generation baselines, highlighting the value of clinically structured and repairable evidence retrieval.

1 Introduction

Large Language Models (LLMs) have shown considerable promise for medical question answering (Jiang et al., 2025; Zhang et al., 2023; Luo et al., 2025). However, medical knowledge is inherently dynamic (Chow et al., 2023; Densen, 2011). Clinical guidelines, diagnostic criteria, and treatment recommendations are continuously revised as new evidence emerges (Akl et al., 2017). This makes it difficult for LLMs to rely solely on static parametric knowledge, which may become outdated and provides limited transparency regarding the evidence behind a prediction (Lewis et al., 2020). Continuously updating LLMs through domain-specific training is often impractical, as it requires expensive medical data and computation.

Question: Patient: 27-year-old male IV drug user, found unresponsive in ED, attends methadone clinic. Initial Vitals: BP 97/48 mmHg, Pulse 140, RR 29, O₂ 98%. Initial Labs: Glucose 589 mg/dL, K 6.3, HCO₃ 17. Post-treatment Vitals: BP 117/78, Pulse 100, RR 23, O₂ 98%. Post-treatment Labs: Glucose 90 mg/dL, K 4.3, HCO₃ 19. Problem: Persistent metabolic acidosis. What is the next step in management?
Gold Answer: Insulin, potassium, IV fluids, and glucose

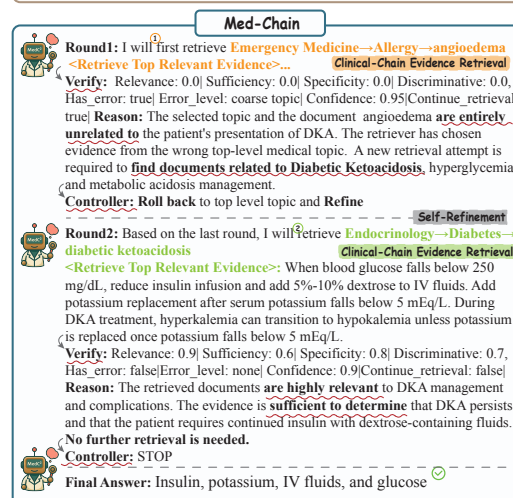
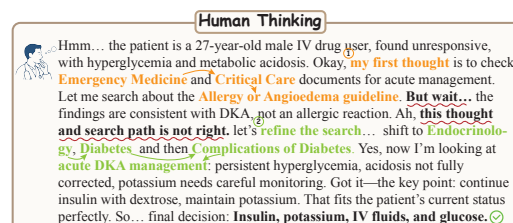
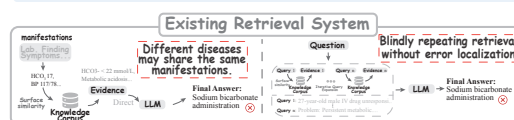


Figure 1: Comparison between existing retrieval system, human thinking and Med-Chain. Clinicians narrow evidence search from broad hypotheses to specific evidence and revise the search direction when a mismatch is found. Inspired by this process, Med-Chain constructs and repairs clinical evidence paths through clinical-chain retrieval and adaptive retrieval refinement.

Evidence retrieval therefore offers a natural solution by augmenting LLMs with external medical knowledge (Lewis et al., 2020; Xiong et al., 2024a). However, external retrieval alone

does not guarantee reliable medical reasoning. Most retrieval pipelines rely on flat top- k lexical or embedding-based similarity (Robertson and Zaragoza, 2009; Karpukhin et al., 2020; Lewis et al., 2020), which may retrieve passages that appear relevant at the surface level but are clinically misaligned (Sohn et al., 2025; Xiong et al., 2024a). In medical QA, evidence can share symptoms, laboratory findings, or treatment terms with the question while referring to the wrong disease, patient population, care setting, disease stage, or decision point. Such misaligned evidence can mislead the LLM and degrade answer quality (Sohn et al., 2025).

The bottleneck of medical evidence retrieval is not merely retrieving evidence, but retrieving evidence along the correct clinical path. Existing medical evidence retrieval frameworks (Zhao et al., 2025; Xiong et al., 2024b; Frisoni et al., 2024; Li et al., 2026; Wu et al., 2024) mainly improve what evidence is retrieved or how retrieved passages are filtered, but they rarely expose the clinical trajectory that leads to the evidence or diagnose which level of the retrieval process has failed. In clinical practice, evidence use is organized around focused questions, evidence acquisition, critical appraisal, and application to patient context. As illustrated in Figure 1, clinicians search in a coarse-to-fine and self-correcting manner: they narrow from broad clinical domains to specific guideline sections, while revising the search when the evidence fails to match the patient presentation or decision point. This motivates a medical retrieval framework that retrieves over interpretable clinical chains and refines them when misalignment occurs.

Based on this motivation, we propose Med-Chain, a training-free medical retrieval framework that reformulates medical evidence retrieval from flat passage matching into inspectable and repairable clinical evidence paths. Med-Chain has two core components. First, *Clinical-Chain Evidence Retrieval* (C²R) routes a medical question through a guideline-derived clinical topic tree, progressively narrowing retrieval from broad clinical domains to specific guideline sections. Second, *Adaptive Retrieval Refinement* (ARR) verifies the retrieved evidence path and selectively repairs it when the evidence is insufficient or clinically mismatched. Rather than blindly repeating retrieval, ARR localizes errors to different levels of the clinical chain and performs level-aware re-

finement, such as rolling back to a broader domain or revisiting a subtopic. As shown in Figure 1, the retrieval narrative records the evolving evidence path, while the verifier and controller determine whether the current evidence is sufficient or whether retrieval should be redirected to a more appropriate clinical topic. To validate the effectiveness, we conduct extensive experiments covering both general medical knowledge and expert-level clinical reasoning. Results show that Med-Chain achieves strong performance across benchmarks and model families, highlighting the effectiveness of clinical-chain evidence retrieval and adaptive retrieval refinement.

Our contributions are summarized as follows:

- 1) We identify clinical evidence misalignment as a path-level failure mode in existing medical retrieval frameworks, where retrieved passages may be lexically relevant but clinically mismatched with the target disease, patient context, or decision point.
- 2) We introduce Med-Chain, a training-free framework that organizes medical retrieval into inspectable clinical evidence paths and supports level-aware repair when retrieval is misaligned.
- 3) Experiments on seven medical QA benchmarks show that Med-Chain consistently improves multiple backbone LLMs and outperforms existing medical RAG baselines.

2 Related Work

2.1 Medical Large Language Models

Medical foundation models improve medical understanding, generation, and question answering through domain-specific pretraining, instruction tuning, and multimodal adaptation. (Jiang et al., 2025; Lin et al., 2025; LASA Team et al., 2025; He et al., 2024; Zhao et al., 2026; Wu et al., 2026; Liu et al., 2025; Wang et al., 2025). Representative systems, including Med-PaLM 2 (Singhal et al., 2025), HuatuoGPT (Zhang et al., 2023), HuLuMed (Jiang et al., 2025), and MedDr (He et al., 2024), show that adapting LLMs with medical corpora or medical instructions can substantially improve performance on clinical reasoning benchmarks.

However, these methods mainly rely on parameter-centric adaptation. In dynamic clinical settings, guidelines, diagnostic criteria, and treatment recommendations are continuously revised, making it costly to keep parametric knowledge up

to date. Moreover, parameterized knowledge provides limited evidence grounding and traceability. These limitations motivate Med-Chain that provides external and inspectable medical evidence.

2.2 Medical Evidence Retrieval

Medical evidence retrieval improves knowledge-intensive generation by conditioning LLMs on retrieved external evidence (Lewis et al., 2020; Asai et al., 2024; Yan et al., 2024). In medical question answering, Retrieval-Augmented Generation (RAG) is particularly useful because clinical guidelines, biomedical literature, and medical textbooks provide evidence-grounded knowledge beyond static parametric memory (Zhao et al., 2025; Li et al., 2026; Wu et al., 2024). Existing medical retrieval systems, such as MedRAG (Zhao et al., 2025), i-MedRAG (Xiong et al., 2024b), and MedRGAG (Li et al., 2026), improve medical QA by retrieving curated medical evidence, generating follow-up queries, filtering evidence, or incorporating structured medical knowledge.

However, most existing methods still treat retrieval primarily as passage selection or passage-level filtering. In medical QA, the critical failure often occurs earlier: the system may enter the wrong clinical topic, subtopic, disease stage, patient population, or decision point, even if the retrieved passages appear lexically relevant. Such clinical-path misalignment is difficult to diagnose from isolated passages alone. Med-Chain addresses this limitation by modeling retrieval as an explicit clinical chain over a guideline-derived topic hierarchy and using adaptive retrieval refinement to repair errors at the level where the chain becomes misaligned.

3 Method: Med-Chain

In this section, we introduce Med-Chain, including Guideline Corpus Collection (Section 3.2), Clinical-chain Evidence Retrieval (Section 3.3) and Adaptive Retrieval Refinement (Section 3.4).

3.1 Overview

Med-Chain is a training-free evidence retrieval framework for medical question answering. It addresses the clinically misaligned evidence problem in medical retrieval frameworks, where retrieved passages appear lexically relevant but fail to support the intended clinical decision. The core mechanism of Med-Chain is to maintain an explicit retrieval state and iteratively refine it when

the evidence is insufficient or mismatched. As shown in Figure 2, the framework first constructs clinical evidence paths over a hierarchical medical guideline corpus, then uses a self-refinement retrieval harness to diagnose retrieval errors and update the retrieval state, and finally summarizes the refined evidence into a compact context for answer generation.

Formally, given a medical question q , answer options $O = \{o_1, \dots, o_K\}$, and a hierarchical medical guideline corpus $\mathcal{D}$, Med-Chain initializes retrieval by calling the clinical-chain retriever:

$$(\mathcal{P}_0, \mathcal{E}_0) = R_{\text{chain}}(q, \mathcal{D}; H_0)$$

where $\mathcal{P}_0$ denotes the ranked clinical paths, $\mathcal{E}_0$ denotes the retrieved evidence chunks, and $H_0 = \emptyset$ is the initial refinement history.

At refinement round t , let τ denote the index of the most recently completed retrieval state. The verifier evaluates the current state:

$$v_t = V(q, O, \mathcal{P}_\tau, \mathcal{E}_\tau).$$

A deterministic controller maps this diagnosis to a refinement action:

$$a_t = \pi(v_t).$$

The diagnosis and action are first written into a temporary updated history:

$$H_t = H_{t-1} \cup \{(v_t, a_t)\}.$$

If refinement is required, the same clinical-chain retriever is re-invoked with the current paths, action, and updated history as feedback:

$$(\mathcal{P}_t, \mathcal{E}_t) = R_{\text{chain}}(q, \mathcal{D}; \mathcal{P}_\tau, a_t, H_t),$$

After refinement, we set $\tau \leftarrow t$. The process terminates when the controller chooses STOP or the refinement budget is reached; if refinement stops, τ remains the index of the most recently completed retrieval state. After termination at round $\tau \leq T$, the final evidence $\mathcal{E}_\tau$ is summarized into a compact context $\mathcal{C}_\tau$, and the answer is generated by:

$$\hat{y} = G_{\text{final}}(q, O, \mathcal{C}_\tau).$$

3.2 Medical Guideline Corpus Collection

We construct a multi-source medical guideline corpus from publicly available clinical guidelines, screening recommendations, disease descriptions,

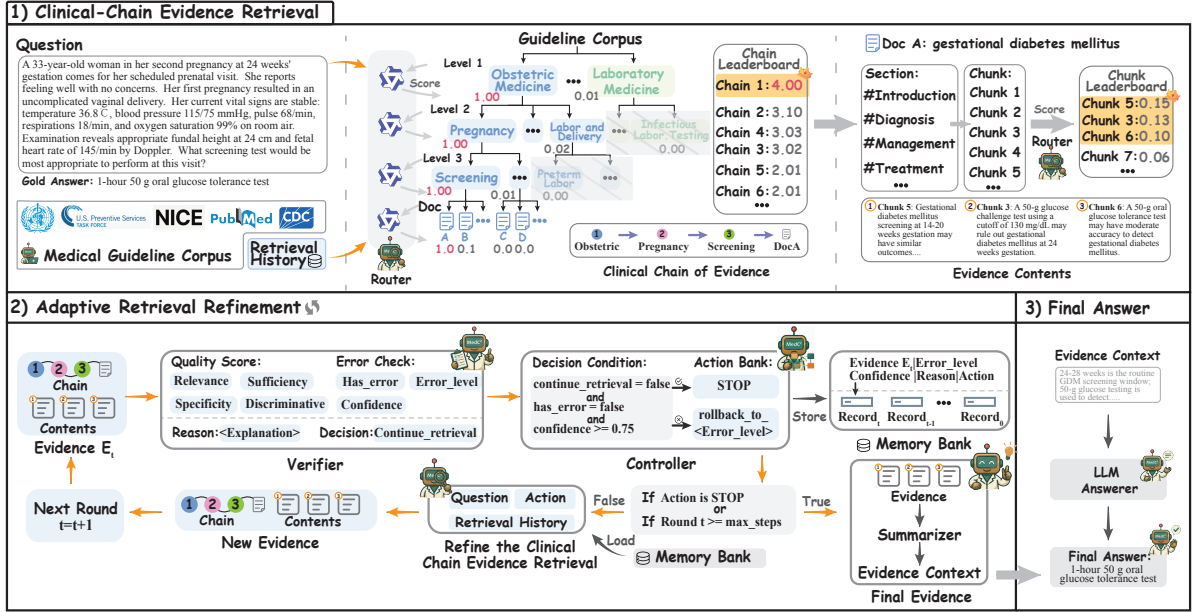


Figure 2: Overview of Med-Chain. The framework follows a retrieve–refine–answer mechanism: (1) *Clinical-Chain Evidence Retrieval* performs coarse-to-fine routing over a hierarchical guideline corpus, producing explicit clinical evidence paths and top evidence chunks; (2) *Adaptive Retrieval Refinement* verifies evidence quality, detects retrieval errors, selectively re-invokes retrieval with history-aware feedback, and summarizes the refined evidence into a compact context; and (3) *Final Answer Generation* uses the evidence context to produce the final answer.

and evidence-based medical references. For each document, we preserve source metadata, document titles, section headings, and paragraph-level content, and convert the corpus into a unified markdown format. To support clinical-chain retrieval, we organize the corpus into a hierarchical topic tree, where documents are assigned to terminal nodes with LLM-assisted classification. This structure enables Med-Chain to retrieve evidence through explicit clinical paths, i.e., domain $\rightarrow$ subtopic $\rightarrow$ document $\rightarrow$ section, while keeping the corpus construction procedure modular and extensible. Details are provided in Appendix.

3.3 Clinical-Chain Evidence Retrieval

Given a medical question q , Clinical-Chain Evidence Retrieval constructs multiple candidate paths over the hierarchical guideline corpus. At each level, lexical relevance is used as a lightweight prior to shortlist plausible child nodes, while the clinical routing decision is mainly made by an LLM router. Specifically, for a current path $p_{\ell-1}$, we first retain the top- B child nodes according to lexical relevance:

$$\mathcal{C}_{\ell}(p_{\ell-1}) = \text{TopB}_c s_{\text{lex}}(q, c), \quad c \in \text{Child}(p_{\ell-1}).$$

The LLM router then selects the most appropriate child $\hat{c}_{\ell}$ from $\mathcal{C}_{\ell}(p_{\ell-1})$. We represent this decision

as a sparse routing score:

$$r_{\text{LLM}}(q, p_{\ell-1}, c) = \mathbb{I}[c = \hat{c}_{\ell}].$$

Each candidate child c expands the current path as $p_{\ell} = p_{\ell-1} \oplus c$, and the cumulative path score is updated by

$$S(p_{\ell}) = S(p_{\ell-1}) + r_{\text{LLM}}(q, p_{\ell-1}, c) + s_{\text{lex}}(q, c).$$

After each expansion round, all candidate paths are ranked by their cumulative scores, and only the top- K_p paths are retained for the next level. This beam-style routing allows the retriever to preserve alternative clinical paths instead of greedily committing to a single early decision.

When terminal documents are reached, the highest-ranked path p^* is selected for evidence extraction, while other high-ranked paths are retained as alternative retrieval candidates. Let d^* be the terminal document of p^* . We split d^* into heading-based evidence chunks, preserving section titles and local context, and rank each chunk by its local relevance to the question:

$$\mathcal{E} = \text{TopK}_{e \in \text{Chunk}(d^*)} S_{\text{ev}}(q, e).$$

The resulting retrieval state, including the selected path p^* , ranked candidate paths, and top evidence

chunks $\mathcal{E}$, is passed to Adaptive Retrieval Refinement for evidence verification and possible refinement.

The same retrieval procedure is used in both initial retrieval and refinement. In the initial round, the history is empty. In later refinement rounds, the retriever can be conditioned on the controller action and accumulated history H_t , which provide feedback about previously diagnosed retrieval failures and guide the retriever to revise the clinical path rather than perform a new flat retrieval.

3.4 Adaptive Retrieval Refinement

Adaptive Retrieval Refinement verifies and repairs the retrieved clinical evidence path by identifying fixable mismatches in the clinical reasoning chain, rather than simply retrieving more passages. Given the current retrieval state $(\mathcal{P}_\tau, \mathcal{E}_\tau)$, the verifier produces a structured diagnosis consisting of evidence quality scores, an error check, and a refinement decision. The quality scores assess relevance, sufficiency, specificity, and discriminativeness, indicating whether the evidence addresses the clinical question, contains enough information, matches the patient context and decision point, and helps distinguish among answer options.

The error check determines whether the retrieval chain contains a repairable mismatch and localizes it to one of four levels: *coarse topic*, *subtopic*, *document*, or *none*. Based on this diagnosis, a deterministic controller either stops refinement when the evidence is adequate or no reliable error is detected, or repairs the retrieval path at the diagnosed level. Coarse-topic errors trigger rollback to a broader clinical domain, subtopic errors preserve the reliable prefix and re-route the remaining branch, and document-level errors search for a more suitable evidence source within the current branch. The repaired document is then segmented and re-ranked to update the evidence state.

Each refinement round records the verifier diagnosis, controller action, and explanation as history, which is passed to subsequent rounds to avoid repeating previous errors. The loop terminates when no further refinement is needed or the refinement budget is reached, after which the final evidence is summarized into the compact context used by the answerer.

4 Experiments

This section presents the experimental setup, main results, and analysis. We answer the following research questions (RQs): **RQ1**: How effective is Med-Chain for medical question answering? **RQ2**: Do the main components of Med-Chain work? **RQ3**: Does retrieving more evidence always improve performance? **RQ4**: Does clinical-chain retrieval improve retrieval quality? **RQ5**: How does model scale affect retrieval and answer generation? **RQ6**: What are the inference-time and token costs of Med-Chain?

4.1 Experimental Setup

4.1.1 Datasets

We evaluate Med-Chain on seven text-based medical question answering benchmarks, covering general medical knowledge, board-style clinical reasoning, and expert-level medical QA. Specifically, we use HLE-Med (Center for AI Safety et al., 2026), MedXpertQA-Text (Zuo et al., 2025), MMLU-Pro-Med (Wang et al., 2024), Medbullets (Chen et al., 2025), SGPQA (Du et al., 2025), MedQA-US (Jin et al., 2021), and MMLU-Med (Hendrycks et al., 2020). These benchmarks vary in difficulty, required medical knowledge, and clinical reasoning depth, allowing us to evaluate Med-Chain across diverse medical QA scenarios.

4.1.2 Baselines

We compare Med-Chain with a broad set of baselines, including frontier models (GPT (OpenAI, 2025), Gemini (Comanici et al., 2025), DeepSeek (DeepSeek-AI et al., 2026)), general-purpose LLMs (Qwen3.5 (Qwen Team, 2026), Llama-3.1 (Grattafiori et al., 2024), InternVL3 (Zhu et al., 2025)), medical LLMs (HuluMed (Jiang et al., 2025), HealthGPT (Lin et al., 2025), Lingshu (LASA Team et al., 2025), MedDr (He et al., 2024)), and RAG-style medical QA methods (MedRAG (Zhao et al., 2025), i-MedRAG (Xiong et al., 2024b), GENREAD (Yu et al., 2022), MedGENIE (Frisoni et al., 2024), GRG (Abdallah and Jatowt, 2023), CGAP (Su et al., 2023), MedRGAG (Li et al., 2026)). These baselines cover closed-source references, direct prompting with general and medical LLMs, and existing retrieval-augmented approaches. We further instantiate Med-Chain with different models to assess its generality across model families and

Table 1: Overall experimental results of Med-Chain and baselines on seven medical QA benchmarks. Within each model group, the best results are highlighted in bold and the second-best results are underlined.

Methods	HLE-Med	MedXpert-Text	MMLU-Pro-Med	Medbullets	SGPQA	MedQA-US	MMLU-Med	Avg.
Frontier Models								
GPT-5.2	12.1	40.3	<u>79.8</u>	81.0	51.2	90.0	90.5	<u>63.6</u>
Gemini-2.5-Pro	8.9	40.1	81.0	76.0	63.0	92.9	91.8	64.8
Gemini-2.5-Flash	7.1	35.6	70.0	<u>77.6</u>	<u>53.3</u>	<u>91.2</u>	84.2	59.9
DeepSeek-V4-Pro	8.5	25.3	71.8	64.5	49.3	81.2	<u>91.0</u>	55.9
General-purpose LLMs								
Qwen3.5-4B	11.0	15.1	53.2	41.9	27.3	57.6	72.8	39.8
Llama-3.1-8B	<u>10.6</u>	13.1	51.3	56.5	28.6	60.6	73.4	42.0
Qwen3.5-9B	<u>10.6</u>	18.2	61.7	56.5	33.1	67.7	83.4	47.3
InternVL3-38B	7.5	16.0	<u>72.1</u>	54.6	42.5	73.5	83.8	50.0
Qwen3.5-27B	7.5	27.5	75.9	64.5	51.1	85.3	91.3	57.6
Qwen3.5-35B-A3B	8.5	<u>20.6</u>	69.3	<u>58.1</u>	<u>43.4</u>	<u>73.9</u>	<u>86.4</u>	<u>51.5</u>
Medical LLMs								
Hulu-Med-4B	11.1	20.3	58.6	62.7	31.8	73.1	77.5	47.9
Hulu-Med-30B-A3B	<u>10.0</u>	31.3	69.3	71.4	42.2	82.8	85.8	56.1
HealthGPT-14B	6.8	11.3	63.4	39.8	25.7	66.2	80.2	41.9
Lingshu-32B	8.6	<u>22.7</u>	70.2	<u>65.4</u>	<u>41.1</u>	<u>74.7</u>	<u>84.7</u>	<u>52.5</u>
MedDr-40B	8.9	12.0	55.6	44.3	24.0	59.2	65.2	38.5
RAG-Style LLMs								
MedRAG	9.6	19.2	67.0	62.1	38.3	75.0	85.5	51.0
i-MedRAG	9.6	20.1	67.0	60.5	38.3	77.5	86.4	51.3
GENREAD	10.7	21.9	68.5	62.1	39.5	81.2	87.3	53.0
MedGENIE	<u>10.0</u>	<u>24.5</u>	72.8	72.6	<u>44.6</u>	<u>85.3</u>	<u>88.8</u>	<u>56.9</u>
GRG	9.6	20.4	67.8	62.9	39.2	77.9	86.5	52.0
CGAP	9.6	17.8	65.3	54.8	38.1	71.1	83.8	48.7
MedRGAG	8.6	27.4	<u>72.2</u>	<u>71.0</u>	45.1	86.0	89.1	57.0
Med-Chain								
Ours (Qwen3.5-4B)	9.2	30.9	77.4	74.2	53.5	87.8	89.7	60.4
Ours (Llama-3.1-8B)	9.6	26.0	67.6	73.4	45.7	84.1	87.3	56.2
Ours (Qwen3.5-9B)	<u>10.6</u>	<u>32.9</u>	78.1	74.2	54.4	89.5	91.1	61.6
Ours (Qwen3.5-27B)	9.2	34.7	78.7	77.4	56.6	<u>90.3</u>	93.2	62.9
Ours (Qwen3.5-35B-A3B)	11.0	32.1	<u>78.6</u>	<u>75.8</u>	<u>55.5</u>	89.5	92.1	<u>62.1</u>
Ours (DeepSeek-V4-Pro)	7.5	31.8	<u>78.6</u>	<u>75.8</u>	55.3	90.7	<u>93.0</u>	61.8

scales. Configuration details for the RAG-style baselines are provided in Appendix.

4.1.3 Implementation Details

By default, we use Qwen3.5-27B as the clinical-chain retriever. We use Qwen3.5 models (4B, 9B, 27B, 35B-A3B), Llama-3.1-8B, and DeepSeek-V4-Pro as answerers. The branch factor, beam width, and number of candidate retrieval paths are set to 3, 5, and 5, respectively. The top 3 reranked evidence chunks are used for verification and answer generation. ARR is capped at 5 refinement steps to avoid unbounded retrieval. Performance is measured by multiple-choice accuracy. To prevent data leakage, the retrieval corpus excludes all benchmark questions and answer labels.

4.2 Main Results (RQ1)

As shown in Table 1, we report the performance of Med-Chain and baselines.

Med-Chain demonstrates superior performance by retrieving evidence along clinically structured paths and refining misaligned re-

trieval results. Compared with standard LLM prompting, Med-Chain brings substantial average improvements, with Qwen3.5-27B increasing from 57.6 to 62.9 and Qwen3.5-35B-A3B increasing from 51.5 to 62.1. Compared with existing medical RAG methods, Med-Chain also achieves stronger overall performance, demonstrating the effectiveness of clinical-chain retrieval and verifier-guided evidence refinement. These results suggest that flat retrieval or direct parameter-based medical reasoning is insufficient for challenging medical QA, while structured evidence acquisition provides a more reliable way to support answer generation.

Med-Chain exhibits strong generalization capabilities across model families and model sizes. Med-Chain consistently improves the corresponding backbone models on average. Notably, Qwen3.5-9B improves from 47.3 to 61.6, and Llama-3.1-8B improves from 42.0 to 56.2, showing that structured retrieval can substantially enhance smaller or weaker base models. Stronger backbones such as Qwen3.5-35B-A3B

and DeepSeek-V4-Pro also benefit from the framework, indicating that Med-Chain is a general inference-time framework rather than a backbone-specific optimization.

The improvements are particularly evident on knowledge-intensive and expert-level clinical reasoning benchmarks. On MedXpert-Text, Med-Chain improves Qwen3.5-27B from 27.5 to 34.7 and Qwen3.5-35B-A3B from 20.6 to 32.1, showing clear gains on expert-level medical QA. On MMLU-Pro-Med, Med-Chain further improves Qwen3.5-35B-A3B from 69.3 to 78.6 and achieves 78.7 with Qwen3.5-27B. Similar gains are observed on Medbullets, SGPQA, MedQA-US, and MMLU-Med. These results indicate that Med-Chain is especially effective when questions require specialized medical knowledge, multi-step clinical reasoning, and precise evidence grounding.

4.3 Ablation Study (RQ2, RQ3)

Figure 3 reports the ablation results averaged over seven medical QA benchmarks. All ablation experiments use Qwen3.5-35B-A3B for both the retriever and answerer.

Impact of Clinical-Chain Evidence Retrieval (C²R). For this comparison, ARR is disabled, and the retrieved evidence chunks are passed directly to the answerer. Compared with the No-RAG baseline, flat retrieval methods (Robertson and Zaragoza, 2009; Jin et al., 2023; Wang et al., 2022) bring only marginal or unstable gains: BM25 improves the average score from 51.5 to 52.0, while MedCPT and e5 achieve 51.6 and 50.7, respectively. In contrast, C²R substantially improves the average score to 57.8. This shows that the key to medical RAG is not merely retrieving external evidence, but retrieving evidence through clinically structured paths.

Impact of Adaptive Retrieval Refinement (ARR). ARR further improves the average score from 57.8 to 60.3. This indicates that, after the initial clinical-chain retrieval, verifier-guided refinement can correct part of the clinically misaligned evidence and improve downstream answering. Therefore, ARR acts as a repair mechanism rather than a standalone retrieval module: it strengthens Med-Chain by refining retrieval paths that are initially incomplete or misaligned.

Impact of Retrieval Breadth and Evidence Depth. The results show that more retrieval is not always better. For child-node expansion, using

top-3 child nodes achieves the best average score of 57.9, outperforming both top-1 and top-5. This suggests that moderate branching balances exploration and noise. For retrieval paths, keeping top-5 paths performs best, indicating that path-level diversity helps improve retrieval robustness. For evidence chunks, using top-3 blocks achieves the best score, while increasing to top-5 slightly decreases performance. This suggests that excessive evidence may introduce distracting or less relevant information, and that Med-Chain benefits from selective evidence use rather than simply adding more context.

4.4 Analysis of Med-Chain Retrieval Quality (RQ4)

To further evaluate retrieval quality, we sample 100 representative examples from MedXpertQA-Text. Two licensed physicians score the retrieved evidence of each system across five dimensions: relevance, sufficiency, specificity, discriminativeness, and verifiability. Scores are averaged across annotators and examples.

As shown in Figure 4, Med-Chain achieves the best overall retrieval quality across all five dimensions. Compared with flat retrievers and existing medical RAG baselines, our method retrieves evidence that is more clinically relevant, more sufficient for answering the question, and more discriminative for distinguishing among answer options. Notably, Med-Chain shows a clear advantage in verifiability. Unlike flat retrievers that return isolated passages, Med-Chain exposes the retrieval trajectory, including the selected topic, subtopic, document, and evidence chunks. This makes it easier to audit whether the retrieved evidence follows a clinically plausible route and whether it supports the final answer.

4.5 Analysis of Model Scale and Stage Allocation (RQ5)

Figure 5 shows the effect of model allocation between retriever and answer generation.

Larger models improve performance, but balanced stage allocation matters more. We vary the model used for retrieval routing and evidence summarization, and independently vary the model used for final answer generation. The results show that model scale has a clear impact on Med-Chain, but the effect is not strictly monotonic. The best average performance is achieved when Qwen3.5-27B is used for both retriever and an-

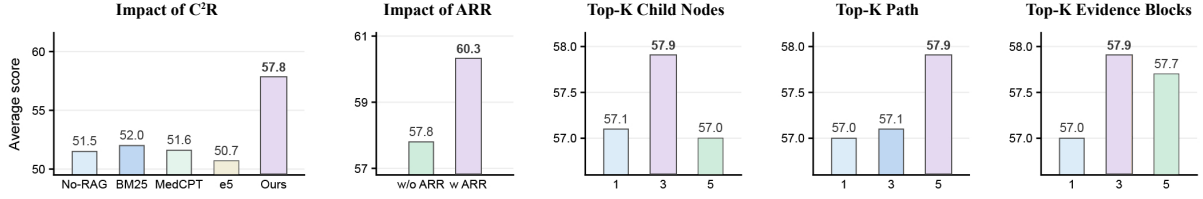


Figure 3: Ablation results of Med-Chain. We report the average accuracy for clinical-chain evidence retrieval (C^2R), adaptive retrieval refinement (ARR), and different settings of top- k child nodes, retrieval paths, and evidence chunks. The results show that both C^2R and ARR improve performance, while moderate retrieval breadth and evidence depth provide the best trade-off between evidence coverage and noise.

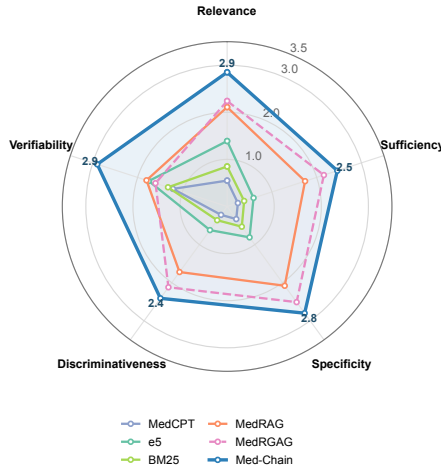


Figure 4: Retrieval Quality Evaluation.

swer generation, reaching an average score of 62.9. This indicates that stronger models can improve both evidence acquisition and final answering, but assigning the largest model to every stage does not necessarily yield the best result.

A smaller retriever can still benefit a larger answerer. The results show that the retrieval stage does not always require the largest model to be effective. Even when a smaller model is used for retrieval, it can still provide clinically relevant evidence that improves a larger answer model. This suggests that retrieval routing and answer generation play complementary roles: the retriever determines whether the answerer receives useful evidence, while the answerer determines how effectively such evidence is used. Therefore, Med-Chain benefits from balanced model allocation across stages rather than scaling only a single component.

4.6 Efficiency Analysis of Med-Chain (RQ6)

Table 2 reports the efficiency of Med-Chain in terms of average time per query (TPQ) and token

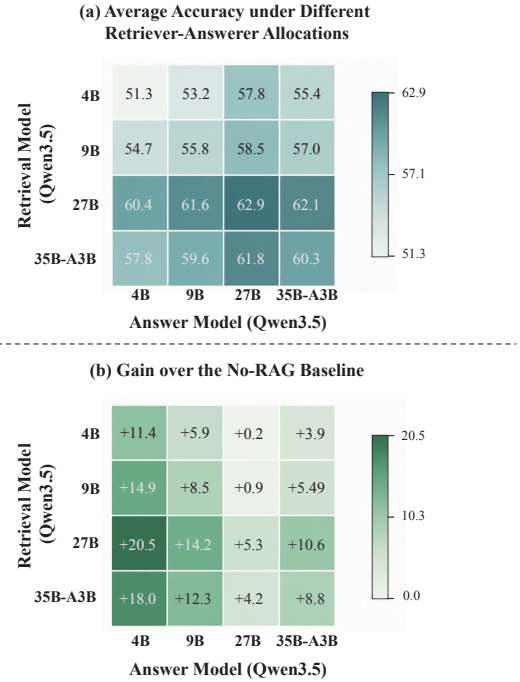


Figure 5: Effect of model scale and stage allocation in Med-Chain. Rows indicate the retriever model and columns indicate the answerer model. The upper heatmap shows average accuracy across seven medical QA benchmarks, while the lower heatmap shows the gain over the same answerer without retrieval.

consumption per thousand queries (Token/KQ). Med-Chain requires 0.38 seconds per query and 6.92M tokens per thousand queries. Although it is more costly than lightweight baselines such as GENREAD and MedRAG, the additional cost comes from clinical-chain routing and verifier-guided refinement. Compared with stronger iterative RAG methods, Med-Chain remains efficient and uses substantially fewer tokens than MedRGAG while achieving better overall performance.

Table 2: Efficiency Comparison.

Method	TPQ	Token/KQ
CGAP	0.20s	2.36M
GENREAD	0.10s	1.52M
GRG	0.89s	7.14M
i-MedRAG	0.33s	6.45M
MedGENIE	0.26s	2.97M
MedRAG	0.11s	2.08M
MedRGAG	0.74s	10.59M
Ours	0.38s	6.92M

5 Conclusion

In this work, we present Med-Chain, a training-free medical evidence retrieval framework that retrieves evidence through inspectable and repairable clinical-chain paths. By combining Clinical-Chain Evidence Retrieval with Adaptive Retrieval Refinement, Med-Chain reduces clinical misalignment in flat retrieval, where lexically relevant passages may fail to support the intended medical decision. Experiments on seven medical QA benchmarks demonstrate its effectiveness and generalizability. These results highlight the importance of clinically structured and level-aware evidence retrieval for reliable medical QA.

Limitations

Med-Chain has several limitations. Its effectiveness depends on the coverage, quality, freshness, and organization of the underlying medical corpus, and missing or outdated sources may affect both routing and retrieval. Since the evaluated benchmarks provide answer labels but no gold supporting guidelines or clinical paths, our results demonstrate the downstream utility of the corpus rather than directly validating its completeness or hierarchical organization. Moreover, although the retrieved sources are traceable, the summarizer and answerer may still draw on parametric knowledge. Med-Chain therefore supports retrieval-level grounding but does not guarantee sentence-level provenance for every generated statement. Future work will focus on corpus validation, finer-grained provenance tracking, verifier robustness, and inference efficiency.

Acknowledgments

This work is supported by the Pioneer and Leading Goose R&D Program of Zhejiang (Grant No. 2025C01128), the National Natural Science Foundation of China (Grant No. 62476241), and the

ZJU-Angelalign R&D Research Center.

References

- Abdelrahman Abdallah and Adam Jatowt. 2023. Generator-retriever-generator approach for open-domain question answering. *arXiv preprint arXiv:2307.11278*.
- Elie A. Akl, Joerg J. Meerpohl, Julian Elliott, Lara A. Kahale, Holger J. Schünemann, and Living Systematic Review Network. 2017. Living systematic reviews: 4. living guideline recommendations. *Journal of clinical epidemiology*, 91:47–53.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. [Self-rag: Learning to retrieve, generate, and critique through self-reflection](#). In *International Conference on Learning Representations*, volume 2024, pages 9112–9141.
- Center for AI Safety, Scale AI, and HLE Contributors Consortium. 2026. [A benchmark of expert-level academic questions to assess AI capabilities](#). *Nature*, 649:1139–1146.
- Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599.
- Natalie LY Chow, Natalie Tateishi, Alexa Goldhar, Rabia Zaheer, Donald A Redelmeier, Amy H Cheung, Ayal Schaffer, and Mark Sinyor. 2023. Does knowledge have a half-life? an observational study analyzing the use of older citations in medical and scientific publications. *Bmj Open*, 13(5):e072374.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, and 300 others. 2026. [Deepseek-v4: Towards highly efficient million-token context intelligence](#). *Preprint*, arXiv:2606.19348.
- Peter Densen. 2011. Challenges and opportunities facing medical education. *Transactions of the American clinical and climatological association*, 122:48–58.
- Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Zhu, Minghao Liu, Yiming Liang,

- Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shuyue Guo, Shian Jia, Sichao Jiang, Yiyao Liao, Rui Li, Qinrui Li, Sirun Li, and 76 others. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *Advances in Neural Information Processing Systems*, 38.
- Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gianluca Moro, and Zaiqiao Meng. 2024. To generate or to retrieve? on the effectiveness of artificial contexts for medical open-domain question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9878–9919.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Sunan He, Yuxiang Nie, Hongmei Wang, Shu Yang, Yihui Wang, Zhiyuan Cai, Zhixuan Chen, Yingxue Xu, Luyang Luo, Huiling Xiang, et al. 2024. Gsco: Towards generalizable ai in medicine via generalist-specialist collaboration. *arXiv preprint arXiv:2404.15127*.
- 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*.
- Songtao Jiang, Yuan Wang, Sibao Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. 2025. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*.
- 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, 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.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781.
- LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. 2025. *Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning*. Preprint, arXiv:2506.07044.
- 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.
- Lei Li, Xiao Zhou, Yingying Zhang, and Xian Wu. 2026. From retrieval to generation: Unifying external and parametric knowledge for medical question answering. In *Proceedings of the ACM Web Conference 2026*, pages 2419–2430.
- Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. 2025. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*.
- Jiaxiang Liu, Tianxiang Hu, Jiawei Du, Ruiyuan Zhang, Joey Tianyi Zhou, and Zuozhu Liu. 2025. Kpl: Training-free medical knowledge mining of vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18852–18860.
- Haoran Luo, Haihong E, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, and Anh Tuan Luu. 2025. Hypergraphrag: Retrieval-augmented generation via hypergraph-structured knowledge representation. *Advances in Neural Information Processing Systems*, 38:152206–152234.
- OpenAI. 2025. *Introducing GPT-5.2*. Published December 11, 2025. Accessed August 28, 2026.
- Qwen Team. 2026. *Qwen3.5: Towards native multimodal agents*.
- Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. *Foundations and Trends in Information Retrieval*, 3(4):333–389.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950.
- Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2025. Rationale-guided retrieval augmented generation for medical question answering. In *Proceedings of the 2025 Conference of*

- the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12739–12753.
- Dan Su, Mostofa Patwary, Shrimai Prabhumoye, Peng Xu, Ryan Prenger, Mohammad Shoeybi, Pascale Fung, Animashree Anandkumar, and Bryan Catanzaro. 2023. Context generation improves open domain question answering. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 793–808.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.
- Yuan Wang, Shujian Gao, Jiayang Liu, Songtao Jiang, Haoxiang Xia, Xiaotian Zhang, Zhaolu Kang, Yemin Wang, and Zuozhu Liu. 2025. [Beyond n-grams: A hierarchical reward learning framework for clinically-aware medical report generation](#). Preprint, arXiv:2512.02710.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.
- Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*.
- Qinxin Wu, Fucheng Niu, Hengchuan Zhu, Yifan Sun, Ye Shen, Xu Li, Han Wu, Leqi Liu, Zhiwen Pan, Zuozhu Liu, et al. 2026. Bridging the skill gap in clinical cbct interpretation with cbctreptd. *arXiv preprint arXiv:2603.10933*.
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251.
- Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024b. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 199–214. World Scientific.
- Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.
- Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.
- Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. 2023. HuatuoGPT, towards taming language model to be a doctor. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 10859–10885.
- Weike Zhao, Chaoyi Wu, Yanjie Fan, Pengcheng Qiu, Xiaoman Zhang, Yuze Sun, Xiao Zhou, Shuju Zhang, Yu Peng, Yanfeng Wang, et al. 2026. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, 651(8106):775–784.
- Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pages 4442–4457.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.
- Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A Appendix

A.1 Med-Chain Details

A.1.1 Guideline Corpus Construction

We construct a hierarchical guideline corpus as the external evidence source for Med-Chain. We collect 12.1K publicly available clinical guideline documents, medical reference documents, and disease-specific reports, and convert all documents into a unified markdown format. During conversion, we preserve the original document structure as much as possible, including titles, section headings, subheadings, recommendation statements, diagnostic criteria, and clinically meaningful paragraph boundaries.

The corpus is organized into a three-level hierarchy under the guidance of clinical experts. Level 1 corresponds to broad clinical domains or body systems, such as cardiology, respiratory medicine, infectious diseases, endocrinology, neurology, and oncology. Level 2 further groups documents by disease category, clinical problem, or management scenario, such as heart failure, arrhythmia, pulmonary embolism, tuberculosis, diabetes, or stroke. Level 3 corresponds to the concrete guideline or report document. This design follows the common process of clinical evidence seeking: identifying the relevant specialty, narrowing it to a disease or clinical problem, and then consulting the corresponding guideline. It also enables retrieval errors to be localized to the coarse domain, subtopic, or document level for layer-aware refinement.

We use DeepSeek-V3 to perform model-assisted document classification. For each markdown document, the model is given the title, source metadata, major headings, and a short content preview, and then assigns the document to the most appropriate Level-1 and Level-2 categories. Ambiguous, low-confidence, or overlapping cases are reviewed and adjusted by clinical experts. Finally, each markdown document is segmented into evidence chunks. We split documents according to original section boundaries whenever possible, and further divide long sections into chunks of approximately 2,500 characters. Each chunk is stored with its document title, section heading, source path, and hierarchical breadcrumb for traceable retrieval.

Algorithm 1 Med-Chain Inference Procedure

Require: Question q , options O , hierarchical corpus $\mathcal{D}$, maximum rounds T

Ensure: Final answer $\hat{y}$

```
1:  $H_0 \leftarrow \emptyset$ 
2:  $(\mathcal{P}_0, \mathcal{E}_0) \leftarrow R_{\text{chain}}(q, \mathcal{D}; H_0)$ 
3:  $\tau \leftarrow 0$ 
4: for  $t = 1$  to  $T$  do
5:    $v_t \leftarrow V(q, O, \mathcal{P}_\tau, \mathcal{E}_\tau)$ 
6:    $a_t \leftarrow \pi(v_t)$ 
7:    $H_t \leftarrow H_{t-1} \cup \{(v_t, a_t)\}$ 
8:   if  $a_t = \text{STOP}$  then
9:     break
10:  else
11:     $(\mathcal{P}_t, \mathcal{E}_t) \leftarrow R_{\text{chain}}(q, \mathcal{D}; \mathcal{P}_\tau, a_t, H_t)$ 
12:     $\tau \leftarrow t$ 
13:  end if
14: end for
15:  $\mathcal{C}_\tau \leftarrow \text{Summarize}(q, O, \mathcal{E}_\tau)$ 
16:  $\hat{y} \leftarrow G_{\text{final}}(q, O, \mathcal{C}_\tau)$ 
17: return  $\hat{y}$ 
```

A.1.2 Med-Chain Pipeline

Algorithm 1 summarizes the inference procedure of Med-Chain. Given a medical question and a hierarchical guideline corpus, Med-Chain first performs clinical-chain retrieval to select structured evidence paths and retrieve relevant evidence chunks. At each refinement round, a verifier evaluates the current retrieval state, and a deterministic controller selects either STOP or a level-aware refinement action. If refinement is required, the retriever is re-invoked using the current path, controller action, and retrieval history. The process terminates when the controller chooses STOP or the maximum number of rounds is reached. Finally, the evidence from the last completed retrieval state is summarized into a compact context for final answer generation.

A.2 Experiment Details

All experiments are implemented under the same evaluation pipeline. Experiments are conducted on 8 NVIDIA H20 GPUs with bfloat16 inference. For multi-GPU inference, we use tensor/model parallelism according to model size and GPU memory capacity. The guideline corpus is preprocessed offline, including markdown conversion, hierarchical classification, and evidence chunking, and is reused across all retrieval-based methods.

For RAG-style baselines, we follow their offi-

cial configurations and implementation settings as closely as possible, including their retrieval format, prompt template, number of retrieved documents, and evidence construction strategy. To ensure a controlled comparison, every LLM-based module is instantiated with Qwen3.5-35B-A3B while keeping the remaining pipeline unchanged. All baselines are evaluated using the same scripts, answer extraction rules, and metric computation as our method. We use the original test split of each benchmark and do not modify the data partition.

For each dataset, we report accuracy based on the final parsed option letter. If the model output cannot be parsed into a valid option, it is counted as incorrect. This protocol is applied consistently to all methods. For methods involving retrieval, the retrieved evidence is provided only at inference time and no model parameters are updated using the evaluation data.

The ablation studies focus on the retrieval component of Med-Chain. We use Qwen3.5-35B-A3B as the backbone model for retrieval routing, evidence-based answer generation, verification, and final answer generation. To isolate the effect of retrieval design, ARR is disabled for the retrieval-strategy comparison, while the top-k sensitivity analyses use a maximum refinement budget of one step. Specifically, the retrieval-strategy variants perform only the initial retrieval, whereas the top-k variants allow at most one refinement step. This setting allows us to separately evaluate retrieval design and retrieval hyperparameter sensitivity under controlled refinement budgets. We compare variants such as flat retrieval, clinical-chain retrieval without refinement, and alternative evidence selection strategies under the same backbone model, prompt format, evidence budget, and evaluation script.

A.2.1 Efficiency

We report the efficiency of different methods using two metrics: time per query (TPQ) and tokens per thousand queries (Token/KQ). TPQ measures the average wall-clock inference time required to answer one query, including retrieval, intermediate reasoning or refinement steps, and final answer generation. A lower TPQ indicates faster end-to-end inference. Token/KQ measures the total number of input and output tokens consumed per 1,000 queries, which reflects the computational cost of LLM calls. This metric is especially important for methods that involve multiple retrieval, verifi-

cation, or generation stages.

All efficiency measurements were obtained on 8 NVIDIA H20 GPUs using bfloat16 inference. Models were served with vLLM using batched inference and 256 asynchronous workers. For N completed queries, TPQ is calculated as T/N , where T is the total wall-clock time spent on on-line inference. This interval covers retrieval, routing, verification, refinement, summarization, and final answer generation, while one-time corpus preprocessing and index construction are excluded. Retrieval indices are reused across queries, and no answer-generation or refinement outputs are cached. Token/KQ aggregates the input and output tokens from all LLM calls.

A.3 Human Evaluation Rubric

To further analyze retrieval quality, we conduct a human evaluation on a subset of examples. We sample 100 representative cases from MedExpertQA and ask two licensed clinicians to score the retrieved evidence. Each retrieved evidence set is evaluated along five dimensions: relevance, sufficiency, specificity, discriminativeness, and verifiability. Each dimension is scored from 0 to 5, where 0 indicates that the retrieved evidence is completely unhelpful and 5 indicates that the evidence is highly useful for answering the question. The dimensions are:

Relevance: Whether the retrieved evidence is medically related to the question, including the same disease, symptom, mechanism, test, treatment, or clinical scenario.

Sufficiency: Whether the evidence provides enough information to support answering the question, rather than only mentioning a related topic.

Specificity: Whether the evidence is specific to the target condition, population, diagnosis, treatment, or management context, instead of being overly general.

Discriminativeness: Whether the evidence helps distinguish the correct option from plausible distractors.

Verifiability: Whether the retrieved evidence is traceable, explicit, and easy for users to inspect or verify through the retrieved clinical path and document source.

A.4 Path-Level Diagnostics

Because the datasets provide answer labels but no gold clinical paths, we evaluate clinically acceptable rather than exact paths. On 700 cases (100

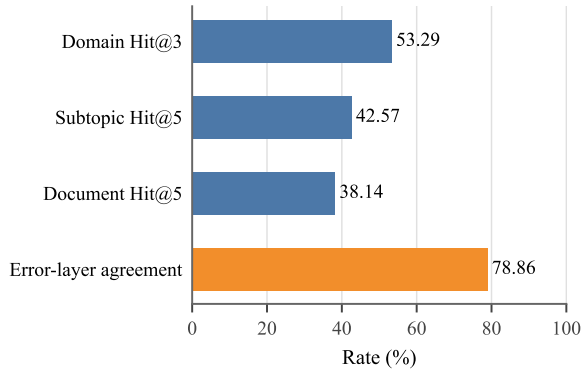


Figure 6: Path-level diagnostics on 700 cases. Hit@ k measures the percentage of cases in which a clinically acceptable target appears in the top- k predictions. Error-layer agreement measures agreement between the controller and the automatic judge on the earliest erroneous layer.

per dataset), DeepSeek-V4-Flash judges the top-3 domains, top-5 subtopics, top-5 terminal documents, and earliest erroneous layer. This diagnostic uses Qwen3.5-35B-A3B for both retrieval and answering. Med-Chain achieves 53.29% Domain Hit@3, 42.57% Subtopic Hit@5, 38.14% Document Hit@5, and 78.86% error-layer agreement. Manual review of 150 judgments yields 92% human-judge agreement. Without gold paths, these values are descriptive diagnostics rather than standard accuracy.

A.5 Benchmark Details

This section introduces the medical benchmarks used in our evaluation. These benchmarks cover a wide range of difficulty levels, from standard medical knowledge tests to expert-level clinical reasoning and frontier academic evaluation. Together, they assess LLMs’ capabilities in medical knowledge recall, clinical decision-making, multi-step reasoning, and robustness under challenging question formats.

- HLE-Med.** HLE-Med is the health and medicine subset of Humanity’s Last Exam (HLE), an expert-level closed-ended benchmark for evaluating LLMs on difficult academic medical questions. Each question has a clear and verifiable answer, but typically requires specialized biomedical or clinical reasoning rather than simple retrieval. It is useful for assessing advanced medical reasoning beyond conventional QA benchmarks.
- MedXpert-Text.** MedXpert-Text is the text-
- only subset of MedXpertQA, a challenging expert-level medical benchmark covering 17 specialties and 11 body systems. It contains carefully reviewed questions designed through filtering, augmentation, synthesis, and expert validation to reduce ambiguity and data leakage. Compared with standard benchmarks such as MedQA, it emphasizes specialty-level clinical reasoning and comprehensive medical understanding.
- MMLU-Pro-Med.** MMLU-Pro-Med is the medical subset of MMLU-Pro, an enhanced version of MMLU designed for more difficult and reasoning-oriented evaluation. MMLU-Pro improves upon the original MMLU by removing noisy or overly simple questions, increasing answer choices from four to ten, and introducing more complex reasoning tasks. As a result, MMLU-Pro-Med provides a more discriminative evaluation of medical knowledge and reasoning while reducing the effect of random guessing.
- Medbullets.** Medbullets is a recent medical multiple-choice QA benchmark consisting of 308 USMLE Step 2/3-style clinical questions collected from open-access posts on X after April 2022. Each question includes a clinical vignette, five answer options, and expert-written explanations for both correct and incorrect choices. Its recent construction makes it less likely to overlap with LLM pre-training data and suitable for evaluating both answer accuracy and explanation quality.
- SGPQA.** SGPQA is the medical subset of SuperGPQA, a graduate-level benchmark spanning 285 disciplines. SuperGPQA uses a Human-LLM collaborative filtering process to remove trivial, ambiguous, or unreliable questions through iterative refinement with model responses and expert feedback. In the medical domain, SGPQA evaluates specialized biomedical knowledge and clinical reasoning beyond standard medical QA tasks.
- MedQA-US.** MedQA-US is the English subset of MedQA, a multilingual medical multiple-choice QA benchmark derived from professional medical board exams. The full MedQA dataset includes English, simplified

Chinese, and traditional Chinese questions, while MedQA-US is widely used to assess English-language clinical knowledge, diagnostic reasoning, and board-exam-style medical question answering.

- **MMLU-Med.** MMLU-Med is the medical subset of the Massive Multitask Language Understanding benchmark. It covers subjects such as anatomy, clinical knowledge, college medicine, medical genetics, and professional medicine. Through multiple-choice questions, it provides a standard evaluation of broad medical knowledge and general medical reasoning ability in LLMs.

A.6 Baseline Details

We provide additional details on the general-purpose and medical-domain LLMs, and retrieval-augmented baselines used in this study:

- **GPT-5.2** is OpenAI’s advanced frontier model series designed for professional knowledge work and long-running agentic tasks. It shows strong capabilities in coding, tool use, long-context reasoning, image understanding, and artifact generation. These abilities make it a representative general-purpose baseline for professional, technical, and multimodal tasks.
- **Gemini-2.5-Pro/Flash** is a multimodal reasoning model developed for coding, long-context understanding, and agentic workflows. It can process complex multimodal inputs, including long video content, and performs well on frontier reasoning and coding benchmarks. Its combination of reasoning, multimodality, and long-context capability makes it suitable for complex problem-solving tasks.
- **DeepSeek-V4-Pro** is a large Mixture-of-Experts language model with efficient activation of a subset of its parameters during inference. It supports very long context lengths and is designed for reasoning, coding, and long-horizon tasks. Its architecture improves efficiency in long-context scenarios while maintaining strong model capability.
- **Qwen3.5** is a family of open-weight vision-language models designed for reasoning, coding, agentic tasks, and multimodal understanding. The series adopts an efficient hybrid architecture that combines linear attention with sparse Mixture-of-Experts design. In this study, we use the 4B, 9B, 27B, and 35B-A3B variants to evaluate different model scales and efficiency settings.
- **Llama-3.1** is a family of open foundation language models supporting multilingual understanding, coding, reasoning, and tool use. Its largest version supports long-context processing and achieves competitive performance with leading proprietary models on a variety of tasks. It also explores extensions to multimodal capabilities through image, video, and speech components.
- **InternVL3** is an open-source multimodal language model that jointly learns from multimodal data and pure-text corpora. Unlike models that add visual capabilities after text-only pretraining, InternVL3 adopts a native multimodal pre-training paradigm. This design improves visual-language alignment and supports strong performance across diverse multimodal reasoning tasks.
- **Hulu-Med** is a transparent generalist medical vision-language model designed to integrate medical text, 2D images, 3D volumes, and videos in a unified architecture. It is trained on large-scale public or synthetic medical data across multiple anatomical systems and imaging modalities. Hulu-Med also introduces a token-reduction strategy to improve efficiency for 3D and video inputs.
- **HealthGPT** is a medical vision-language model that unifies medical visual understanding and generation within an autoregressive framework. It progressively adapts medical comprehension and generation knowledge to pretrained LLMs through specialized training strategies. HealthGPT is designed for medical visual comprehension and generation tasks.
- **Lingshu** is a medical-specialized multimodal large language model for medical visual question answering, text-based QA, and report generation. It uses a comprehensive data curation pipeline that combines medical imaging, medical text, and general-domain data.

Through multi-stage training and reasoning-oriented data, Lingshu aims to improve medical knowledge coverage, reduce hallucinations, and enhance reasoning ability.

- **MedDr** is a generalist medical vision-language foundation model designed for diverse healthcare modalities, including radiology, pathology, dermatology, retinography, and endoscopy. It uses a diagnosis-guided bootstrapping strategy to construct medical image-text data from image and label information. During inference, it further applies retrieval-augmented diagnosis to improve generalization across medical tasks.
- **MedRAG** is a medical retrieval-augmented generation framework enhanced by knowledge graph-guided reasoning. It retrieves relevant electronic health records based on patient manifestations and combines them with structured diagnostic knowledge. This allows the model to provide more specific diagnostic and treatment support, especially for diseases with similar clinical presentations.
- **i-MedRAG** is an iterative retrieval-augmented generation framework for medical question answering. Instead of retrieving evidence only once, i-MedRAG allows LLMs to ask follow-up queries across multiple rounds. This iterative process helps gather missing information and form reasoning chains for complex medical questions.
- **GENREAD** is a generate-then-read framework for knowledge-intensive question answering. It replaces or complements document retrieval by prompting an LLM to generate contextual documents relevant to the question. These generated contexts are then used to support final answer generation.
- **MedGENIE** is a medical generate-then-read framework for multiple-choice medical question answering. It uses domain-specific models to generate relevant artificial contexts before answering questions. By providing additional medical background information, MedGENIE improves reasoning and answer accuracy on medical QA benchmarks.
- **GRG** is a Generator-Retriever-Generator framework for open-domain question answering. It first generates contextual documents using an LLM and retrieves relevant documents from an external corpus in parallel. The generated and retrieved evidence are then combined to support final answer generation.
- **CGAP** is a two-stage closed-book question answering framework that uses the knowledge stored in pretrained language models. It first prompts the model to generate a relevant context for the question, and then uses this context to produce the answer. This coarse-to-fine process helps activate parametric knowledge without relying on external retrieval.
- **MEDRGAG** is a unified retrieval-generation augmented framework for medical question answering. It combines retrieved external knowledge with generated background knowledge from the model. By selecting useful evidence from both sources, MEDRGAG aims to reduce incomplete retrieval and unreliable generation, improving the robustness of medical reasoning.

Table 3: Ablation results for C²R.

Method	Retrieval	HLE-Med	MedXpert-Text	MMLU-Pro-Med	Medbullets	SGPQA	MedQA-US	MMLU-Med	Avg
Ours(Qwen3.5-35B-A3B)	No-Retrieval	8.5	20.6	69.3	58.1	43.4	73.9	86.4	51.5
	BM25	9.2	23.6	62.5	71.0	42.5	75.6	79.9	52.0
	MedCPT	8.9	23.9	65.6	60.5	42.5	77.2	82.5	51.6
	e5	9.9	25.0	65.9	57.3	41.5	75.8	79.6	50.7
	C ² R	9.2	27.7	73.3	72.6	50.5	84.1	87.1	57.8

Table 4: Ablation results for ARR.

Method	ARR	HLE-Med	MedXpert-Text	MMLU-Pro-Med	Medbullets	SGPQA	MedQA-US	MMLU-Med	Avg
Ours(Qwen3.5-35B-A3B)	w/o	9.2	27.7	73.3	72.6	50.5	84.1	87.1	57.8
	w	10.6	30.0	76.6	72.6	51.9	88.8	91.6	60.3

Table 5: Ablation results for top-*k* paths.

Method	Top- <i>k</i> Path	HLE-Med	MedXpert-Text	MMLU-Pro-Med	Medbullets	SGPQA	MedQA-US	MMLU-Med	Avg
Ours(Qwen3.5-35B-A3B)	1	9.2	27.2	72.5	67.7	49.6	84.5	87.9	57.0
	3	8.5	28.1	72.3	70.2	49.7	84.2	87.0	57.1
	5	10.3	28.5	72.3	72.6	51.1	83.9	86.6	57.9

Table 6: Ablation results for top-*k* evidence blocks.

Method	Top- <i>k</i> Evidence Block	HLE-Med	MedXpert-Text	MMLU-Pro-Med	Medbullets	SGPQA	MedQA-US	MMLU-Med	Avg
Ours(Qwen3.5-35B-A3B)	1	9.9	26.9	75.2	66.9	48.8	84.3	86.6	57.0
	3	10.3	28.5	72.3	72.6	51.1	83.9	86.6	57.9
	5	9.9	27.1	71.9	73.4	48.6	85.2	87.8	57.7

Table 7: Ablation results for top-*k* child nodes.

Method	Top- <i>k</i> Child Nodes	HLE-Med	MedXpert-Text	MMLU-Pro-Med	Medbullets	SGPQA	MedQA-US	MMLU-Med	Avg
Ours(Qwen3.5-35B-A3B)	1	9.2	27.5	74.5	66.1	49.9	84.7	87.5	57.1
	3	10.3	28.5	72.3	72.6	51.1	83.9	86.6	57.9
	5	9.9	27.3	73.5	71.0	48.4	83.5	85.6	57.0

Table 8: Performance under different retriever and answerer combinations.

Retriever	Answerer	HLE-Med	MedXpert-Text	MMLU-Pro-Med	Medbullets	SGPQA	MedQA-US	MMLU-Med	Avg
—	Qwen3.5-4B	11.0	15.1	53.2	41.9	27.3	57.6	72.8	39.8
—	Qwen3.5-9B	10.6	18.2	61.7	56.5	33.1	67.7	83.4	47.3
—	Qwen3.5-27B	7.5	27.5	75.9	64.5	51.1	85.3	91.3	57.6
—	Qwen3.5-35B-A3B	8.5	20.6	69.3	58.1	43.4	73.9	86.4	51.5
Qwen3.5-4B	Qwen3.5-4B	11.7	20.2	66.1	59.7	36.8	79.0	85.4	51.3
Qwen3.5-4B	Qwen3.5-9B	12.1	21.6	69.3	62.9	39.1	80.8	86.7	53.2
Qwen3.5-4B	Qwen3.5-27B	10.3	24.7	74.5	74.2	44.2	86.4	90.0	57.8
Qwen3.5-4B	Qwen3.5-35B-A3B	11.0	22.8	71.5	67.7	42.2	84.1	88.4	55.4
Qwen3.5-9B	Qwen3.5-4B	11.0	22.6	69.3	70.2	40.4	82.0	87.6	54.7
Qwen3.5-9B	Qwen3.5-9B	11.7	24.1	71.5	68.6	42.7	83.4	88.7	55.8
Qwen3.5-9B	Qwen3.5-27B	8.2	28.2	75.5	71.8	48.4	86.8	90.8	58.5
Qwen3.5-9B	Qwen3.5-35B-A3B	8.9	25.8	72.9	69.4	46.2	85.5	90.0	57.0
Qwen3.5-27B	Qwen3.5-4B	9.2	30.9	77.4	74.2	53.5	87.8	89.7	60.4
Qwen3.5-27B	Qwen3.5-9B	10.6	32.9	78.1	74.2	54.4	89.5	91.1	61.6
Qwen3.5-27B	Qwen3.5-27B	9.2	34.7	78.7	77.4	56.6	90.3	93.2	62.9
Qwen3.5-27B	Qwen3.5-35B-A3B	11.0	32.1	78.6	75.8	55.5	89.5	92.1	62.1
Qwen3.5-35B-A3B	Qwen3.5-4B	9.2	28.2	74.1	70.2	47.6	86.1	89.5	57.8
Qwen3.5-35B-A3B	Qwen3.5-9B	12.4	30.0	74.5	71.0	50.2	88.1	91.2	59.6
Qwen3.5-35B-A3B	Qwen3.5-27B	9.2	33.1	77.4	75.8	53.8	90.6	92.6	61.8
Qwen3.5-35B-A3B	Qwen3.5-35B-A3B	10.6	30.0	76.6	72.6	51.9	88.8	91.6	60.3

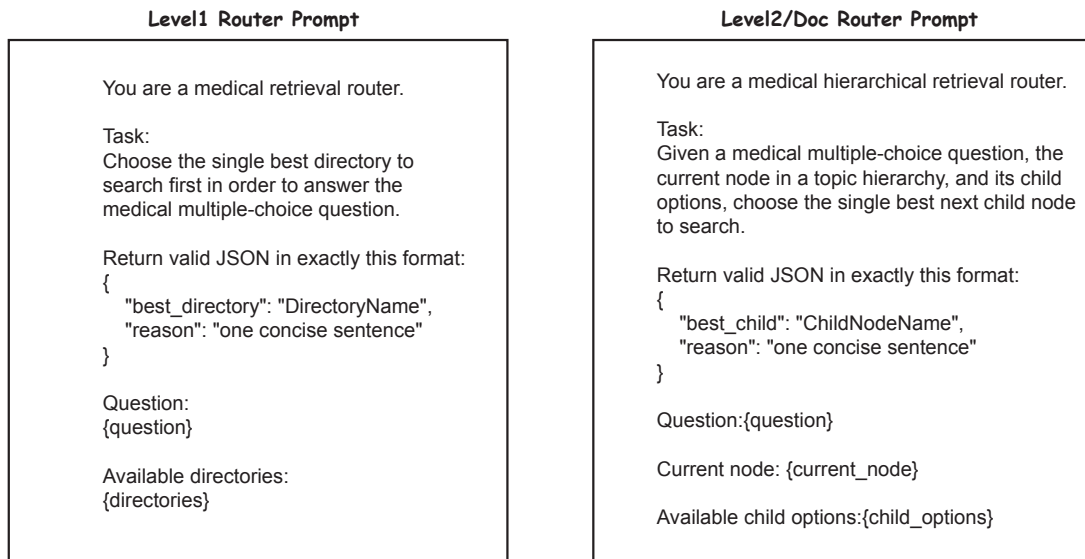


Figure 7: Prompts Used by Med-Chain Routers

Table 9: Institutes/platforms and corresponding links for guideline document collection.

Country/Region	Institute/platform	Website
United States	U.S. Centers for Disease Control and Prevention	www.cdc.gov
United States	Emergency Care Research Institute	home.ecri.org
United Kingdom	National Institute for Health and Care Excellence	www.nice.org.uk
United States	U.S. Preventive Services Task Force	www.uspreventiveservicestaskforce.org/uspstf/home
International	World Health Organization	www.who.int
International	PubMed	pubmed.ncbi.nlm.nih.gov/

Verifier Prompt

You are a conservative verifier for a medical RAG retrieval pipeline.

Your job has two separate parts:

1. Judge evidence quality for answering the exact multiple-choice question.
2. Decide whether another retrieval attempt is likely to improve the evidence.

Important distinction:

- Evidence can be insufficient or non-discriminative even when retrieval is structurally acceptable.
- Do NOT request another retrieval just because the evidence is incomplete.
- Continue retrieval only when the selected path/document is clearly off-topic, wrong population/anatomy/disease/task, or a nearby document/branch is likely to fix the mismatch.
- If the evidence is broadly relevant but not enough to decide the answer, set `has_retrieval_error=false` and `continue_retrieval=false`.
- If several previous attempts would likely keep cycling among unrelated documents, set `continue_retrieval=false`.

Question: {question}

Options: {options}

Retrieval path context: {retrieval_path_context}

Evidence blocks: {evidence_blocks}

Return valid JSON only, exactly one object, with this schema:

```
{
  "evidence_quality": {
    "relevance": 0.0,
    "sufficiency": 0.0,
    "specificity": 0.0,
    "discriminative": 0.0,
    "reason": "short explanation"
  },
  "retrieval_error": {
    "has_retrieval_error": false,
    "error_layer": "coarse_topic|subtopic|doc|none",
    "reason": "short explanation"
  },
  "continue_retrieval": false,
  "continue_reason": "short explanation",
  "confidence": 0.0,
}
```

Layer definitions:

- `coarse_topic`: the selected top-level medical area is broadly wrong.
- `subtopic`: the broad area is right but the narrower branch is wrong.
- `doc`: the branch is plausible but the chosen document is wrong and another document in the branch is likely better.
- `none`: retrieval is acceptable or another retrieval is unlikely to help.

Figure 8: Prompts Used by Med-Chain Verifier

Summarizer Prompt

You are a medical evidence/context summarizer.

Given a medical question, answer options, and retrieved guideline evidence, produce the final context that is useful for answering the question.

The final output should be context only.

Core rules:

- Do not answer the question.
- Do not predict the final answer.
- Do not mention answer letters.
- Do not add unsupported claims as if they came from retrieved evidence.
- Retrieved evidence is authoritative.
- Generated background knowledge is auxiliary and should only be used when critical knowledge is missing from the retrieved evidence.
- Do not include irrelevant evidence.
- If retrieved evidence contains no useful information, you may include only concise background knowledge that is necessary to understand the medical concept, diagnostic criterion, treatment threshold, or differentiating concept needed for the question.
- Do not fabricate citations.
- Do not include reasoning steps, selection rationale, JSON, headings like "useful facts", or metadata.
- Keep the context concise and clinically focused.

Internal procedure:

1. Extract only the retrieved evidence facts that are directly useful for answering the question.
2. If the retrieved evidence is incomplete, identify the critical missing medical knowledge needed to answer the question.
3. Generate concise auxiliary background notes only for those missing knowledge points.
4. Select the most useful retrieved facts and, if needed, auxiliary notes.
5. Prefer retrieved evidence whenever it directly supports the answer.
6. Exclude generated notes that conflict with retrieved evidence.
7. Output only the final integrated context.

Question:

{question}

Options, if provided:

{options}

Retrieved evidence:

{retrieved_evidence}

Return JSON only:

```
{  
  "evidence": "concise context statement",  
}
```

Figure 9: Prompts Used by Med-Chain Summarizer