

# RGAR: Recurrence Generation-augmented Retrieval for Factual-aware Medical Question Answering

Anonymous EMNLP submission

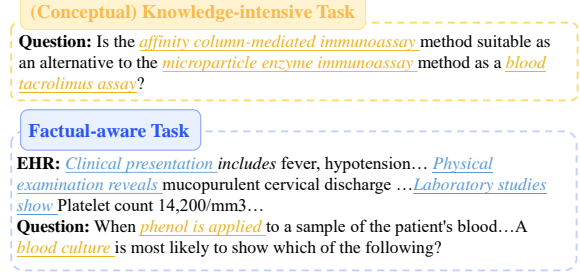
## Abstract

Medical question answering fundamentally relies on accurate clinical knowledge. The dominant paradigm, Retrieval-Augmented Generation (RAG), acquires expertise *conceptual* knowledge from large-scale medical corpus to guide general-purpose large language models (LLMs) in generating trustworthy answers. However, existing retrieval approaches often overlook the patient-specific *factual knowledge* embedded in Electronic Health Records (EHRs), which limits the contextual relevance of retrieved *conceptual knowledge* and hinders its effectiveness in vital clinical decision-making. This paper introduces RGAR, a recurrence generation-augmented retrieval framework that synergistically retrieves both *factual* and *conceptual* knowledge from dual sources (i.e., EHRs and the corpus), allowing mutual refinement through iterative interaction. Across three factual-aware medical QA benchmarks, RGAR establishes new state-of-the-art performance among medical RAG systems. Notably, RGAR enables the Llama-3.1-8B-Instruct model to surpass the considerably larger GPT-3.5 augmented with traditional RAG. Our findings demonstrate the benefit of explicitly mining patient-specific factual knowledge during retrieval, consistently improving generation quality and clinical relevance.

## 1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in general question answering (QA) tasks, achieving impressive performance across diverse scenarios (Achiam et al., 2023). However, when facing domain-specific questions that require specialized expertise, from medical diagnosis (Jin et al., 2021) to legal charge prediction (Wei et al., 2024), these models face significant challenges, often generating unreliable conclusions due to both hallucinations (Ji et al., 2023) and potentially stale knowledge embedded in their parameters (Wang et al., 2024a).

a) Types of Medical QA Datasets



b) Bloom's Taxonomy view of Medical AI system

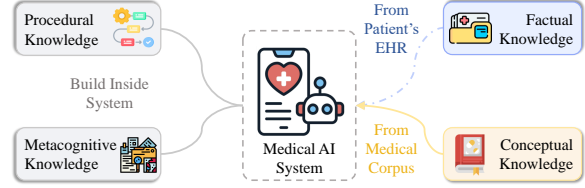


Figure 1: a) Two Types of Medical Question Answering Tasks. b) Medical AI Systems from the Perspective of Bloom's Taxonomy.

**Retrieval-Augmented Generation (RAG)** (Lewis et al., 2020) has emerged as a promising approach to address these challenges by leveraging extensive, trustworthy knowledge bases to support LLM reasoning. The effectiveness of this approach, however, heavily depends on the relevance of retrieved documents.

In the medical domain, current RAG approaches concatenate all available contextual information from a given example into a single basic query for retrieval, aiming to provide comprehensive context for model reasoning (Xiong et al., 2024a). While this method has demonstrated substantial improvements on early *knowledge-intensive* medical QA datasets such as PubMedQA (Jin et al., 2019), its limitations have become increasingly apparent with the emergence of EHR-integrated datasets that better reflect real-world clinical practices (Kweon et al., 2024).

As shown in Figure 1 a), Electronic Health

Records (EHRs) typically contain extensive patient data, including diagnostic test results, medical histories, and other longitudinal information (Pang et al., 2021; Johnson et al., 2023; Lovon-Melgarejo et al., 2024). However, for any specific medical query, only a small subset of this information is typically relevant (Sackett, 1997; D’Alessandro et al., 2004). Incorporating all available EHRs into retrieval queries often introduces substantial irrelevant information, which degrades the performance of LLM- and RAG-based QA systems (Fang et al., 2024; Shi et al., 2023). Despite ongoing efforts to improve retrieval through query expansion and generation, such as Generation-Augmented Retrieval (GAR) (Mao et al., 2021a), how to effectively extract and utilize query-relevant factual knowledge from noisy and large-scale EHRs remains an open problem.

As shown in Figure 1 b), inspired by Bloom’s taxonomy (Forehand, 2010; Markus, 2001), we categorize the knowledge required to address real-world medical QA problems into four types: *Factual Knowledge*, *Conceptual Knowledge*, *Procedural Knowledge*, and *Metacognitive Knowledge*. The latter two represent higher-order knowledge commonly integrated into advanced RAG systems. *Procedural* and *Metacognitive Knowledge* capture reasoning strategies and self-assessment capabilities, respectively, and have been explored in recent studies (Wei et al., 2022; Zhou et al., 2023; Kim et al., 2023; Wang et al., 2023b).

*Factual Knowledge*, such as patient-specific information from EHRs, and *Conceptual Knowledge*, such as general medical understanding from corpora, together form the complete context inputs required for answering medical questions. Processing both types of knowledge requires navigating long contexts filled with irrelevant information. Unfortunately, current RAG systems do not differentiate between these types of *retrieval targets*, overlooking the necessity of retrieval from EHRs.

To overcome this limitation, we propose **RGAR**, a system designed to simultaneously retrieve *Factual Knowledge* and *Conceptual Knowledge* through a recurrent query generation and interaction mechanism. This approach iteratively refines queries to enhance the relevance of retrieved professional and factual knowledge, thereby improving performance on *knowledge-intensive* and *factual-aware* medical QA tasks.

Our key contributions are listed as follows:

- We are the first to analyze RAG systems through the lens of Bloom’s taxonomy, addressing the current underrepresentation of *Factual Knowledge* in existing frameworks.
- We introduce RGAR, a dual-end retrieval system that facilitates recurrent interactions between *Factual* and *Conceptual Knowledge*, bridging the gap between LLMs and real-world clinical applications.
- Through extensive experiments on three medical QA datasets involving *Factual Knowledge*, we demonstrate that RGAR achieves superior average performance compared to state-of-the-art (SOTA) methods, enabling Llama-3.1-8B-Instruct model to outperform the considerably larger RAG-based GPT-3.5-turbo.

## 2 Related Work

**RAG Systems.** RAG systems are characterized as a "Retrieve-then-Read" framework (Gao et al., 2023). The development of Naive RAG has primarily focused on retriever optimization, evolving from discrete retrievers such as BM25 (Friedman et al., 1977) to more sophisticated and domain-specific dense retrievers, including DPR (Karpukhin et al., 2020) and MedCPT (Jin et al., 2023), which demonstrate superior performance.

In recent years, numerous advanced RAG systems have emerged. Advanced RAG systems focus on designing multi-round retrieval structures, including iterative retrieval (Sun et al., 2019), recursive retrieval (Sarathi et al., 2024), and adaptive retrieval (Jeong et al., 2024). A notable work in medical QA is MedRAG (Xiong et al., 2024a), which analyzes retrievers, corpora, and LLMs, offering practical guidelines. Follow-up work, *i*-MedRAG (Xiong et al., 2024b), improved performance through multi-round decomposition and iteration, albeit with significant computational costs.

These approaches focus solely on optimizing the retrieval process, overlooking the retrievability of *factual knowledge*. In contrast, RGAR introduces a recurrent structure, enabling continuous query optimization through dual-end retrieval and extraction from EHRs and professional knowledge corpora, thereby enhancing access to both knowledge types.

**Query Optimization.** As the core interface in human-AI interaction, query optimization (also known as prompt optimization) is the key to improving AI system performance. It is widely ap-

plied in tasks such as text-to-image generation (Liu et al., 2022; Wu et al., 2024b) and code generation (Nazzal et al., 2024).

In the era of large language models, query optimization for retrieval tasks has gained increasing attention. Representative work includes GAR (Mao et al., 2021a), which improves retrieval performance through query expansion using fine-tuned BERT models (Devlin et al., 2019). GENREAD (Yu et al., 2023) further explored whether LLM-generated contexts could replace retrieved professional documents as reasoning evidence. MedGENIE (Frisoni et al., 2024) extended this approach to medical QA.

Another line of work focuses on query transformation and decomposition, breaking down original queries into multiple sub-queries tailored to specific tasks, enhancing retrieval alignment with model needs (Dhuliawala et al., 2023). Subsequent work has reinforced the effectiveness of query decomposition through fine-tuning (Ma et al., 2023).

Using expanded queries directly as reasoning evidence lacks the transparency of RAG, as RAG relies on retrievable documents that provide traceable and trustworthy reasoning, which is crucial in the medical field. Besides, the effectiveness of query expansion and query decomposition approaches is heavily dependent on fine-tuning LLMs, which limits scalability.

In contrast, our work focuses on query optimization without fine-tuning LLMs. Specifically, retrieval from EHRs can be seen as query filtering that eliminates irrelevant information, thereby obtaining pertinent *factual knowledge*. Extracting factual knowledge enhances the effectiveness of retrieval from the corpus.

### 3 Methodology

In this section, we introduce RGAR framework, as illustrated in Figure 2. It begins by prompting a general-purpose LLM to generate multiple queries from an initial basic query. These multiple queries are then used to **retrieve conceptual knowledge** from the corpus (§ 3.2). Then retrieved conceptual knowledge is subsequently used to **extract factual knowledge** from the electronic health records (EHRs) and transform it into retrieval-optimized representations (§ 3.3). The **recurrence pipeline** continuously updates the basic query and iteratively executes the two aforementioned components. This process optimizes the retrieved results, ultimately

improving the quality of responses (§ 3.4).

#### 3.1 Task Formulation

In *factual-aware* medical QA, each data sample comprises the following elements: a patient’s natural language query  $Q$ , the electronic health record (EHR) as factual knowledge  $\mathcal{F}$ , and a set of candidate answers  $\mathcal{A} = \{a_1, \dots, a_{|\mathcal{A}|}\}$ . The overall goal is to identify the correct answer  $\hat{a}$  from  $\mathcal{A}$ .

A *non-retrieval* approach directly prompts an LLM to act as a **reader**, processing the entire context and generating an answer, formulated as:

$$\hat{a} = \text{LLM}(\mathcal{F}, Q, \mathcal{A} | \mathcal{T}_r) \quad (1)$$

where  $\mathcal{T}_r$  is the prompts. However, this approach relies exclusively on the conceptual knowledge encoded within LLM, without leveraging external, trustworthy medical knowledge sources.

To overcome this limitation, recent studies have explored *retrieval-based* approaches, which enhance the model’s knowledge by retrieving a specified number  $N$  of chunks, denoted as  $\mathcal{C} = \{c_1, \dots, c_N\}$ , from a chunked corpus (knowledge base)  $\mathcal{K}$ . This answering process is expressed as:

$$\hat{a} = \text{LLM}(\mathcal{F}, Q, \mathcal{A}, \mathcal{C} | \mathcal{T}_r). \quad (2)$$

#### 3.2 Conceptual Knowledge Retrieval (CKR)

To maintain consistency with the *option-free retrieval approach* proposed by (Xiong et al., 2024a), we do not incorporate the answer options  $\mathcal{A}$  during retrieval. This design is in line with real-world medical quality assurance scenarios, where answer choices are typically not available in advance.

Following their method, we construct the **basic query** by concatenating the EHR and the patient’s query, formally defined as  $q_b = Q \oplus \mathcal{F}$ , where  $\oplus$  denotes text concatenation.

Traditional dense retrievers, such as Dense Passage Retrieval (DPR) (Karpukhin et al., 2020), identify the top- $N$  relevant chunks  $\mathcal{C}$  from the knowledge base  $\mathcal{K}$  by computing similarity scores using an encoder  $E$ :

$$\begin{aligned} \text{sim}(q_b, c_i) &= E(q_b)^\top E(c_i), \\ \mathcal{C} &= \text{top-}N(\{\text{sim}(q_b, c_i)\}). \end{aligned} \quad (3)$$

Vanilla GAR (Mao et al., 2021a) expands  $q_b$  using a fine-tuned BERT (Devlin et al., 2019) to produce three types of content that enhance retrieval: potential answers  $q_e^a$ , contexts  $q_e^c$ , and titles

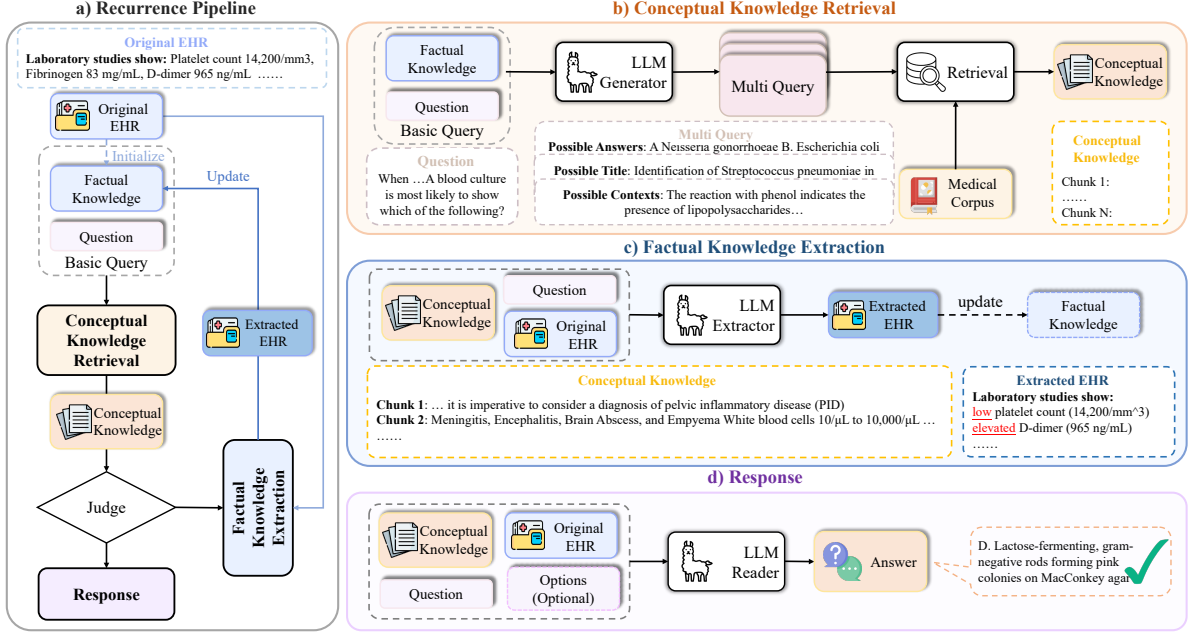


Figure 2: The Overall Framework of RGAR. a) The Recurrence Pipeline in § 3.4; b) Conceptual Knowledge Retrieval in § 3.2; c) Factual Knowledge Extraction in § 3.3; d) Response Template in § 3.4.

$q_e^t$ . With the growing zero-shot generation capabilities of LLMs (Kojima et al., 2022), a common practice is to prompt LLMs to serve as train-free query **generators**, producing expanded content  $\tilde{q}_e$  using prompt templates  $\mathcal{T}_g$  (Frisoni et al., 2024). The three types of content generation process can be formulated as:

$$\begin{aligned}\tilde{q}_e^a &= \text{LLM}(q_b | \mathcal{T}_g^a), \\ \tilde{q}_e^c &= \text{LLM}(q_b | \mathcal{T}_g^c), \\ \tilde{q}_e^t &= \text{LLM}(q_b | \mathcal{T}_g^t).\end{aligned}\quad (4)$$

The final score  $Sc$  for retrieving  $\mathcal{C}$  is then computed by normalizing and averaging the similarities of these expanded queries:

$$Sc(c_i) = \sum_{\tilde{q}_e \in \{\tilde{q}_e^a, \tilde{q}_e^c, \tilde{q}_e^t\}} \frac{\exp(\text{sim}(\tilde{q}_e, c_i))}{\sum_{c_j} \exp(\text{sim}(\tilde{q}_e, c_j))}. \quad (5)$$

### 3.3 Factual Knowledge Extraction (FKE)

In EHR, only a small portion of necessary information constitutes problem-relevant factual knowledge (D’Alessandro et al., 2004). Direct input of lengthy EHR content containing substantial irrelevant information into dense retrievers can degrade retrieval performance (Ren et al., 2023). While a straightforward approach would be to retrieve EHR content based on question  $\mathcal{Q}$  (Lu et al., 2023), this fails to fully utilize conceptual knowledge obtained

from the previous Conceptual Knowledge Retrieval Stage. Furthermore, the necessary chunking of EHR for retrieval introduces content discontinuity (Luo et al., 2024).

Given that EHRs more closely resemble long passages from the Needle in a Haystack task (Kamradt) rather than necessarily chunked corpus, and inspired by large language models’ capability to precisely locate answer spans in reading comprehension tasks (Cheng et al., 2024), we propose leveraging LLMs for text span tasks (Rajpurkar et al., 2016) on EHR to filter relevant factual knowledge efficiently and effectively using conceptual knowledge. We define this filtered factual knowledge as  $\mathcal{F}_s$ , with prompts  $\mathcal{T}_s$ , expressed as:

$$\mathcal{F}_s = \text{LLM}(\mathcal{F}, \mathcal{Q}, \mathcal{C} | \mathcal{T}_s). \quad (6)$$

In addition, EHRs often contain numerical report results (Lovon-Melgarejo et al., 2024) that require conceptual knowledge to interpret their significance. Furthermore, medical QA involves multi-hop questions (Pal et al., 2022), where retrieved conceptual knowledge can generate explainable new factual knowledge conducive to reasoning. Drawing from LLM zero-shot summarization prompting strategies (Wu et al., 2025), we analyze and summarize the filtered EHR  $\mathcal{F}_s$  with prompts  $\mathcal{T}_e$ , yielding an enriched representation  $\mathcal{F}_e$ :

$$\mathcal{F}_e = \text{LLM}(\mathcal{F}_s, \mathcal{Q}, \mathcal{C} | \mathcal{T}_e). \quad (7)$$

This process, which we refer to as the LLM **Extractor**, completes the extraction of original EHR information. In practice, RGAR implements these two phases using single-stage prompting to reduce time overhead.

### 3.4 The Recurrence Pipeline and Response

Building on the  $\mathcal{F}_e$ , we **update** the basic query for Conceptual Knowledge Retrieval as  $q_b = \mathcal{Q} \oplus \mathcal{F}_e$ . This establishes a **recurrence interaction** between factual and conceptual knowledge, guiding next retrieval toward more relevant content. Iterative execution enhances the stability of both retrieval and extraction. The entire pipeline recurs for a predefined number of iterations, ultimately yielding the final retrieved conceptual knowledge  $\mathcal{C}^*$ .

During the response phase, we follow the approach in Equation 2 to generate answers. Notably, the  $\mathcal{F}_e$  are restricted to the retrieval phase and are not used in the response phase. The sole difference lies in the retrieved chunks, highlighting the impact of retrieval quality on the responses.

## 4 Experimental Setup

### 4.1 Benchmark Datasets

We evaluated RGAR on three *factual-aware* medical QA benchmarks featuring multiple-choice questions that require human-level reading comprehension and expert reasoning to analyze patients’ clinical conditions.

**MedQA-US** (Jin et al., 2021) and **MedMCQA** (Pal et al., 2022) consist of questions derived from professional medical exams, evaluating specialized expertise such as disease symptom diagnosis and medication dosage requirements. The problems frequently involve patient histories, vital signs (e.g., blood pressure, temperature), and final diagnostic evaluations (e.g., CT scans), making it necessary to retrieve relevant medical knowledge tailored to the patient’s specific circumstances. However, due to their exam-oriented format, the provided information has already been filtered, reducing the difficulty of extracting factual knowledge from EHR.

**EHRNoteQA** (Kweon et al., 2024) is a recently introduced benchmark that provides authentic, complex EHR data derived from MIMIC-IV (Johnson et al., 2023). This dataset encompasses a wide range of topics and demands that models emulate genuine clinical consultations, ultimately generating accurate discharge recommendations. Consequently, EHRNoteQA challenges models to

Table 1: Medical QA Benchmark Statistics.

Benchmarks	Max. Len	Avg. Len	Min. Len
Non-EHR QA Benchmarks			
BioASQ-Y/N (Tsatsaronis et al., 2015)	52	17	9
PubMedQA (Jin et al., 2019)	57	23	10
MMLU-Med (Hendrycks et al., 2021)	961	87	17
EHR QA Benchmarks			
MedMCQA (Pal et al., 2022)	207	41	11
MedQA-US (Jin et al., 2021)	872	197	50
EHRNoteQA (Kweon et al., 2024).	5782	3061	667

identify which *factual details* within the EHR are relevant to the questions at hand and apply domain-specific knowledge to address them.

Table 1 highlights that the chosen datasets, which include EHR information, tend to have significantly **longer** content compared to datasets without EHRs. Notably, the EHRNoteQA dataset has a maximum length exceeding 4,000 tokens. This raises concerns about the reasonableness of directly employing these EHRs for retrieval. While the MMLU-Med dataset contains relatively long questions, it is still categorized as a Non-EHR QA Benchmark, as its content does not derive from factual information. Representative question samples are provided in the Appendix E.2.

### 4.2 Retriever and Corpus

To ensure a fair comparison, we adopt the same retriever, corpus, and parameter settings as previous work (Xiong et al., 2024a). We use MedCPT (Jin et al., 2023), a dense retriever specialized for the biomedical domain, configured to retrieve 32 chunks by default. For the corpus, we employ the Textbooks corpus (Jin et al., 2019), a lightweight collection of 125.8k chunks. Results on a much larger-scale corpus are presented in Appendix B.4.

### 4.3 LLMs and Baselines

We focus on the effect of RGAR on general-purpose LLMs without domain-specific knowledge. Therefore, we exclude LLMs fine-tuned on the medical domain, such as PMC-Llama (Wu et al., 2024a). Our primary experiments utilize Llama-3.2-3B-Instruct, while ablation studies include a range of models from the Llama-3.1/3.2 (Dubey et al., 2024) and Qwen-2.5 (Yang et al., 2024a) families, ranging from 1.5B to 8B parameters. All selected models feature a context length of approximately 128K tokens. Temperatures are set to zero to ensure reproducibility through greedy decoding.

For *non-retrieval methods*, we consider a zero-shot approach Custom (Kojima et al., 2022) as a

Table 2: Comparison of RGAR with SOTA Methods on Three Factual-Aware Datasets and MMLU-Med.  $\Delta$  Indicates Improvement Over Custom, **Bold** Represents the Best, and Underline Indicates the Second-Best.

Method		MedQA-US (# 1273)		MedMCQA(# 4183)		EHRNoteQA(# 962)		Average( $\downarrow$ )			MMLU-Med(# 1089)	
		Acc.	$\Delta$	Acc.	$\Delta$	Acc.	$\Delta$	Acc.	$\Delta$	Avg.Rank	Acc.	$\Delta$
w/o Retrieval	Custom	50.20	0.00	50.01	0.00	47.19	0.00	49.13	0.00	6.00	64.46	0.00
	CoT	51.45	1.25	44.53	-5.48	62.89	15.70	52.96	3.82	5.67	62.99	-1.47
w/ Retrieval	RAG	53.50	3.30	<u>50.54</u>	<u>0.53</u>	61.12	13.93	55.05	5.92	4.00	65.47	1.01
	MedRAG	50.27	0.07	47.53	-2.48	70.58	23.39	56.13	6.99	4.67	63.91	-0.55
	GAR	<u>57.97</u>	<u>7.77</u>	50.42	0.41	65.48	18.29	57.96	8.82	<u>3.00</u>	<u>66.12</u>	<u>1.66</u>
	<i>i</i> -MedRAG	56.24	6.04	44.94	-5.07	<b>74.22</b>	<b>27.03</b>	<u>58.47</u>	<u>9.33</u>	3.33	64.74	0.28
	RGAR	<b>58.83</b>	<b>8.63</b>	<b>51.02</b>	<b>1.01</b>	<u>73.28</u>	<u>26.09</u>	<b>61.04</b>	<b>11.91</b>	<b>1.33</b>	<b>66.48</b>	<b>2.02</b>

baseline and evaluate improvements relative to it. To fully exploit the reasoning capabilities of the LLMs, we incorporate chain-of-thought (CoT) reasoning (Wei et al., 2022). For *retrieval-based methods*, we evaluate the classic RAG model (Lewis et al., 2020), the domain-adapted MedRAG (Xiong et al., 2024a), and *i*-MedRAG (Xiong et al., 2024b), a medical-domain RAG system designed to decompose questions and iteratively provide answers.

We adopt GAR (Mao et al., 2021a) as a representative *query-optimized RAG method*, implemented train-free in accordance with § 3.2. RGAR defaults to 2 rounds of recurrence.

#### 4.4 Evaluation Settings

Following MIRAGE (Xiong et al., 2024a), we adopt the following evaluation framework. In **Option-Free Retrieval**, no answer options are provided for retrieval (§3.2), ensuring a more realistic medical QA scenario. In **Zero-Shot Learning**, RAG systems are evaluated without in-context few-shot learning, reflecting the lack of similar exemplars in real-world medical questions. For **Metrics**, we employ Accuracy, defined as the proportion of correctly answered questions, and we extract model outputs by applying regular expression matching to the entire generated responses (Wang et al., 2024b).

### 5 Experimental Analysis

#### 5.1 Cross-Dataset Performance Improvement

We evaluate RGAR using the LLaMA-3.2-3B-Instruct model on three factual-aware medical QA datasets, comparing it against several competitive baselines. The results, presented in Table 2, include the absolute performance of each method as well as their relative improvements over the Custom baseline. **RGAR achieves the highest average performance across all three datasets, outperforming the second-best method, *i*-MedRAG, by 2%. Retrieval-based methods—despite variability**

in quality—consistently surpass non-retrieval baselines (Custom and CoT), underscoring the importance of incorporating specialized medical knowledge when leveraging general-purpose LLMs to answer professional medical queries.

Among the retrieval-based approaches, GAR outperforms vanilla RAG by approximately 3% on average, with a maximum gain of 4.37% across datasets. This demonstrates the effectiveness of multi-query generation in improving retrieval quality. However, MedRAG, while performing well on EHRNoteQA, exhibits degraded performance on the other two datasets compared to vanilla RAG, highlighting its limited robustness.

**A key advantage of our proposed RGAR framework lies in its stable and consistent performance improvements**—an essential property for medical applications. As shown in Table 2, RGAR ranks among the top two methods across all datasets, delivering reliable gains over both RAG and GAR. In contrast, *i*-MedRAG, despite incurring substantial time overhead, performs poorly on MedMCQA and ranks near the bottom, which significantly undermines its suitability for real-world deployment.

Notably, the performance improvements of RGAR over GAR exhibit a positive correlation with the average context length in each dataset. For example, in EHRNoteQA, which has an average context length exceeding 3000 tokens, RGAR achieves a 7.8% improvement, validating the benefit of our Factual Knowledge Extraction module in enhancing retrieval effectiveness. This suggests that RGAR is particularly well-suited to practical clinical scenarios where complete electronic health records must be analyzed to generate accurate medical recommendations.

To further assess generalizability, we evaluated the models on Conceptual Knowledge-Intensive Tasks, where factual knowledge extraction is ex-

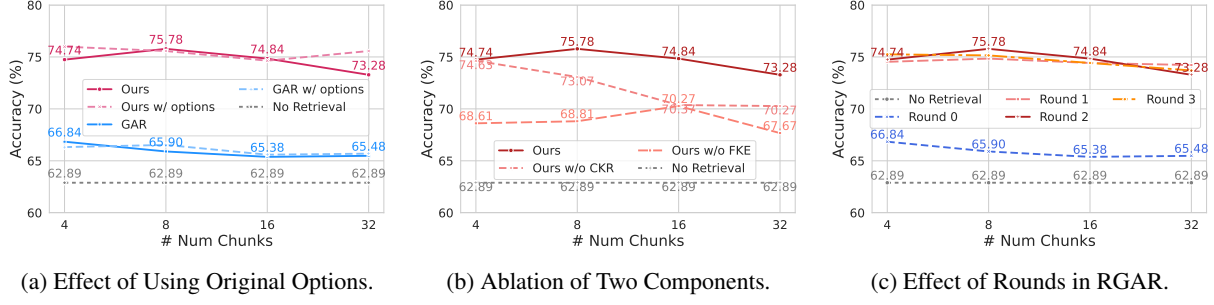


Figure 3: Accuracy with Different Numbers of Retrieved Chunks on EHRNoteQA Dataset.

Table 3: Comparison of LLMs on MedQA-US.

Model	Custom	RAG	GAR	RGAR
Llama-3.2-1B-Instruct	38.96	29.30	30.79	29.85
Llama-3.2-3B-Instruct	50.20	53.50	57.97	58.83
Llama-3.1-8B-Instruct	60.80	62.14	67.39	69.52
Qwen2.5-1.5B-Instruct	43.99	41.48	43.42	42.58
Qwen2.5-3B-Instruct	48.23	49.96	53.50	54.28
Qwen2.5-7B-Instruct	59.46	58.83	63.39	63.86
Average	50.27	49.20	52.74	53.15

pected to have less impact. On the MMLU-Med dataset, RGAR continues to outperform GAR and also surpasses *i*-MedRAG, demonstrating its robustness across diverse task scenarios.

## 5.2 Base LLMs with Different Sizes and Model Families

To further assess the versatility of RGAR, we conduct evaluations on MedQA-US, a widely used medical dataset, by utilizing base LLMs of various sizes and model families, specifically from Llama and Qwen. The results in Table 3 show that RGAR consistently achieves the best average performance.

When considering model size, we find that retrieval-based approaches fall short of the non-retrieval Custom baseline for smaller models, such as Llama-3.2-1B-Instruct and Qwen2.5-1.5B-Instruct. These smaller models, constrained by their weaker performance, are not well-suited to leverage retrieval-enhanced information. As the model size increases, however, all retrieval-enhanced approaches exhibit notable performance gains, with RGAR yielding the most significant improvements. This trend becomes particularly pronounced for larger models. For example, RGAR achieves a 7.38% improvement over RAG on Llama-8B, 5.33% on Llama-3B, 5.03% on Qwen-8B, and 4.32% on Qwen-3B.

Moreover, we find that under the same ex-

perimental conditions, **Llama-3.1-8B-Instruct achieves a performance of 69.52% with RGAR, surpassing the 66.22% reported by MedRAG for GPT-3.5-16k-0613 (Achiam et al., 2023).** This significant improvement underscores the practicality of using well-optimized retrieval methods with smaller models, enabling performance rivals those of proprietary large-scale foundational models in real-world medical recommendation tasks.

## 5.3 Ablation Study

Due to the absence of ground-truth retrieval chunks, we evaluate retrieval effectiveness through QA performance, systematically varying the number of retrieved chunks  $N$  from 4 to 32. A reduced retrieval number serves as a more stringent assessment of retrieval quality. We investigate three primary factors in Figure 3: the effect of options generated by GAR versus those originally provided by the dataset, the contributions of CKR and FKE components, and the impact of RGAR’s recurrence rounds.

We first compare the retrieval performance between LLM-generated options and original dataset options. Figure 3a shows how RGAR and GAR perform across different values of  $N$ . Both approaches maintain stable performance across different  $N$ , indicating reliable retrieval quality. While using original options shows slightly higher average Accuracy, the difference is minimal. This suggests that even when GAR generates options that differ from the originals, it achieves similar retrieval results as long as the core topics align.

We then examine the impact of RGAR’s two main components—CKR and FKE—as shown in Figure 3b. When we remove the conceptual knowledge interaction from the FKE phase, the system shows only moderate improvements when extracting factual knowledge from EHR without conceptual knowledge, demonstrating the importance of integrating both types of knowledge. Removing the

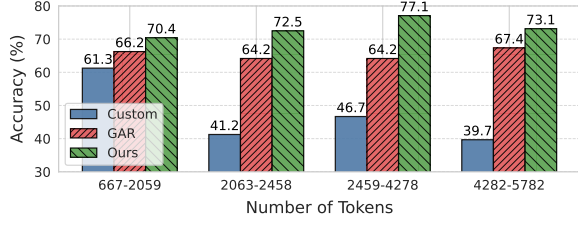


Figure 4: Fine-Grained Accuracy of EHRNoteQA After Sorting by Length and Dividing into Four Equal Parts.

multi-query generation step from CKR causes performance to degrade as  $N$  increases, indicating that multiple queries are necessary to maintain stable retrieval.

Finally, we analyze the effect of rounds in RGAR (Round 0 means GAR), as illustrated in Figure 3c. Our results show that even a single iteration significantly improves performance by enabling interaction between factual and conceptual knowledge. Multiple rounds work similarly to a reranking mechanism (Mao et al., 2021b), improving the ranking of important chunks and showing substantial gains even with relatively small  $N$ . With  $N = 8$ , the default two-round setup achieves a performance of 75.78%, almost 1% better than using a single round. However, adding more rounds shows no clear benefits, as they tend to generate multi-hop factual knowledge during the FKE phase, leading CKR to retrieve multi-hop conceptual knowledge, which may cause LLMs to over-infer (Yang et al., 2024b). Given that each round involves one reasoning step from both the LLM extractor and LLM query generator, two rounds sufficiently support multi-hop reasoning needs (Lv et al., 2021).

#### 5.4 Fine-Grained Performance Analysis

While the previous sections examined overall dataset performance and established preliminary findings, this section provides a detailed analysis of specific aspects of our results. In § 5.1, we showed that RGAR performs better on real-world medical recommendation tasks involving comprehensive EHRs. To verify this finding, we conduct a detailed analysis of EHRNoteQA by grouping questions based on context length and dividing them into four bins. Within each bin, we compare the performance of RGAR, GAR, and Custom. As shown in Figure 4, Custom shows decreasing accuracy with increasing context length. GAR improves accuracy across all bins, with RGAR achieving further performance gains. Notably, the improvements are

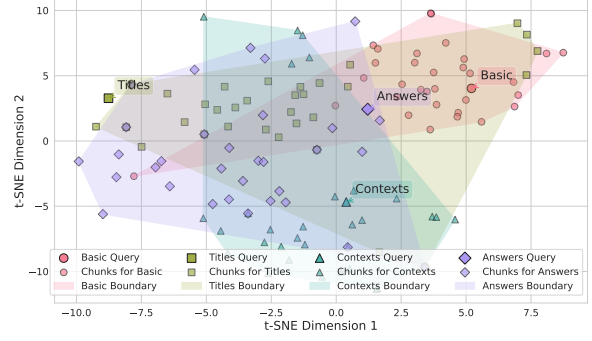


Figure 5: t-SNE Visualization of Different Queries and the Retrieved Chunks.

more significant in the three bins with longer contexts compared to the first bin. The results show that RGAR maintains consistent average performance across different context length.

It is also important to note that generating multiple queries from different aspects within RGAR helps stabilize retrieval. Figure 5 presents a t-SNE visualization of different queries and their individually retrieved chunks for a sample question (details provided in Appendix C). The basic query shows limited suitability for retrieval, as its coverage area differs from that of the three queries generated by RGAR. RGAR clearly introduces some variation in retrieval content. Although the regions corresponding to the three generated queries overlap, the specific chunks retrieved do not overlap significantly. This underscores the need to average the retrieval similarities of these three queries to achieve more stable retrieval results.

## 6 Conclusion

In this work, we propose RGAR, a novel RAG system that distinguishes two types of retrievable knowledge. Through comprehensive evaluation across three factual-aware medical benchmarks, RGAR demonstrates substantial improvements over existing methods, emphasizing the significant impact of in-depth factual knowledge extraction and its interaction with conceptual knowledge on enhancing retrieval performance. Notably, our RGAR enables the Llama-3.1-8B-Instruct model to outperform the considerably larger, RAG-based proprietary GPT-3.5. From a broader perspective, RGAR offers a promising approach for enhancing general-purpose LLMs in clinical diagnostic scenarios where extensive factual knowledge is crucial, with potential for extension to other professional domains demanding precise factual awareness.

## Limitations

Despite RGAR achieving superior average performance, several limitations warrant discussion. Our RGAR requires corpus retrieval, and its time complexity scales proportionally with the size of the corpus, which is an inherent issue within the RAG paradigm. Approaches that generate reasoning evidence directly through domain-specific LLMs (Yu et al., 2023; Frisoni et al., 2024) avoid the computational challenges at inference time. However, they face difficulties in updating LLMs to incorporate new medical knowledge, which results in frequent updates and training costs.

Comparative approaches such as MedRAG (Xiong et al., 2024a) and *i*-MedRAG (Xiong et al., 2024b) explore integration possibilities with prompting techniques like Chain-of-Thought (Wei et al., 2022) and Self-Consistency (Wang et al., 2023a) to enhance reasoning capabilities. Our investigation focused specifically on validating how additional factual knowledge processing improves retrieval performance, without examining the impact of these prompting strategies. Furthermore, unlike multi-round methods such as *i*-MedRAG (Xiong et al., 2024b) that implement LLM-based early stopping to reduce computational costs, our system operates with fixed time complexity. However, it is noteworthy that, because *i*-MedRAG requires multiple rounds of query decomposition, retrieval, and answer aggregation, the actual time overhead of RGAR is significantly smaller than that of *i*-MedRAG.

Our EHR extraction approach assumes LLMs can process complete EHR contextual input, justified by current mainstream LLMs exceeding 128K context windows with anticipated growth. However, in extreme cases where EHR content exceeds LLM context limits, integration with chunk-free approaches may be necessary (Luo et al., 2024; Qian et al., 2024). Finally, as RGAR operates in a zero-shot setting without instruction fine-tuning, its effectiveness is partially contingent on the model’s instruction-following capabilities—which we cannot fully mitigate.

## Ethical Statement

This research adheres to the ACL Code of Ethics. All medical datasets utilized in this study are either open access or obtained through credentialed access protocols. To ensure patient privacy protection, all datasets have undergone comprehensive

anonymization procedures. While Large Language Models (LLMs) present considerable societal benefits, particularly in healthcare applications, they also introduce potential risks that warrant careful consideration. Although our work advances the relevance of retrieved content for medical queries, we acknowledge that LLM-generated responses based on retrieved information may still be susceptible to errors or perpetuate existing biases. Given the critical nature of medical information and its potential impact on healthcare decisions, we strongly advocate for a conservative implementation approach. Specifically, we recommend that all system outputs undergo rigorous validation by qualified medical professionals before any practical application. This stringent verification process is essential to maintain the integrity of clinical and scientific discourse and prevent the propagation of inaccurate or potentially harmful information in healthcare settings. These ethical safeguards reflect our commitment to responsible AI development in the medical domain, where the stakes of misinformation are particularly high and the need for reliability is paramount.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. 2002. *Finite-time analysis of the multiarmed bandit problem*. *Mach. Learn.*, 47(2–3):235–256.
- Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, et al. Stealing part of a production language model. In *Forty-first International Conference on Machine Learning*.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.
- Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. *Adapting large language models via reading comprehension*. In *The Twelfth International Conference on Learning Representations*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of*

720	<i>the North American Chapter of the Association for</i>	Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi	776
721	<i>Computational Linguistics: Human Language Tech-</i>	Rungta, Krithika Iyer, Yuning Mao, Michael	777
722	<i>nologies, Volume 1 (Long and Short Papers)</i> , pages	Tontchev, Qing Hu, Brian Fuller, Davide Testuggine,	778
723	4171–4186, Minneapolis, Minnesota. Association for	et al. 2023. Llama guard: Llm-based input-output	779
724	Computational Linguistics.	safeguard for human-ai conversations. <i>arXiv preprint</i>	780
		<i>arXiv:2312.06674</i> .	781
725	Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu,	Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju	782
726	Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Ja-	Hwang, and Jong Park. 2024. <a href="#">Adaptive-RAG: Learn-</a>	783
727	son Weston. 2023. Chain-of-verification reduces hal-	<a href="#">ing to adapt retrieval-augmented large language mod-</a>	784
728	lucination in large language models. <i>arXiv preprint</i>	<a href="#">els through question complexity</a> . In <i>Proceedings of</i>	785
729	<i>arXiv:2309.11495</i> .	<i>the 2024 Conference of the North American Chap-</i>	786
730	Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey,	<i>ter of the Association for Computational Linguistics:</i>	787
731	Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman,	<i>Human Language Technologies (Volume 1: Long</i>	788
732	Akhil Mathur, Alan Schelten, Amy Yang, Angela	<i>Papers)</i> , pages 7036–7050, Mexico City, Mexico. As-	789
733	Fan, et al. 2024. The llama 3 herd of models. <i>arXiv</i>	sociation for Computational Linguistics.	790
734	<i>preprint arXiv:2407.21783</i> .		
735	Donna M D’Alessandro, Clarence D Kreiter, and	Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan	791
736	Michael W Peterson. 2004. An evaluation of	Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea	792
737	information-seeking behaviors of general pediatri-	Madotto, and Pascale Fung. 2023. Survey of halluci-	793
738	cians. <i>Pediatrics</i> , 113(1):64–69.	nation in natural language generation. <i>ACM Comput-</i>	794
		<i>ing Surveys</i> , 55(12):1–38.	795
739	Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiao-	Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng,	796
740	jun Chen, and Ruifeng Xu. 2024. <a href="#">Enhancing noise</a>	Hanyi Fang, and Peter Szolovits. 2021. What disease	797
741	<a href="#">robustness of retrieval-augmented language models</a>	does this patient have? a large-scale open domain	798
742	<a href="#">with adaptive adversarial training</a> . In <i>Proceedings</i>	question answering dataset from medical exams. <i>Ap-</i>	799
743	<i>of the 62nd Annual Meeting of the Association for</i>	<i>plied Sciences</i> , 11(14):6421.	800
744	<i>Computational Linguistics (Volume 1: Long Papers)</i> ,		
745	pages 10028–10039, Bangkok, Thailand. Association	Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William	801
746	for Computational Linguistics.	Cohen, and Xinghua Lu. 2019. <a href="#">PubMedQA: A</a>	802
747	Mary Forehand. 2010. Bloom’s taxonomy. <i>Emerging</i>	<a href="#">dataset for biomedical research question answering</a> .	803
748	<i>perspectives on learning, teaching, and technology</i> ,	In <i>Proceedings of the 2019 Conference on Empirical</i>	804
749	41(4):47–56.	<i>Methods in Natural Language Processing and the</i>	805
750	Jerome H Friedman, Jon Louis Bentley, and Raphael Ari	<i>9th International Joint Conference on Natural Lan-</i>	806
751	Finkel. 1977. An algorithm for finding best matches	<i>guage Processing (EMNLP-IJCNLP)</i> , pages 2567–	807
752	in logarithmic expected time. <i>ACM Transactions on</i>	2577, Hong Kong, China. Association for Computa-	808
753	<i>Mathematical Software (TOMS)</i> , 3(3):209–226.	tional Linguistics.	809
754	Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gi-	Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau,	810
755	anluca Moro, and Zaiqiao Meng. 2024. <a href="#">To generate</a>	Lana Yeganova, W John Wilbur, and Zhiyong Lu.	811
756	<a href="#">or to retrieve? on the effectiveness of artificial con-</a>	2023. Medcpt: Contrastive pre-trained transformers	812
757	<a href="#">texts for medical open-domain question answering</a> .	with large-scale pubmed search logs for zero-shot	813
758	In <i>Proceedings of the 62nd Annual Meeting of the</i>	biomedical information retrieval. <i>Bioinformatics</i> ,	814
759	<i>Association for Computational Linguistics (Volume 1:</i>	39(11):btad651.	815
760	<i>Long Papers)</i> , pages 9878–9919, Bangkok, Thailand.		
761	Association for Computational Linguistics.	Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin	816
762	Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia,	Gayles, Ayad Shammout, Steven Horng, Tom J Pol-	817
763	Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen	lard, Sicheng Hao, Benjamin Moody, Brian Gow,	818
764	Wang. 2023. Retrieval-augmented generation for	et al. 2023. MIMIC-IV, a freely accessible electronic	819
765	large language models: A survey. <i>arXiv preprint</i>	health record dataset. <i>Scientific data</i> , 10(1):1.	820
766	<i>arXiv:2312.10997</i> .		
767	Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt,	Greg Kamradt. <a href="#">Llmtest_needleinahaystack: Evaluating</a>	821
768	and Demet Yuksel. 2025. Large language models	<a href="#">long-context capabilities of large language models</a> .	822
769	lack essential metacognition for reliable medical rea-	Accessed: 2025-02-13.	823
770	soning. <i>Nature communications</i> , 16(1):642.		
771	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou,	Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick	824
772	Mantas Mazeika, Dawn Song, and Jacob Steinhardt.	Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and	825
773	2021. <a href="#">Measuring massive multitask language under-</a>	Wen-tau Yih. 2020. <a href="#">Dense passage retrieval for open-</a>	826
774	<a href="#">standing</a> . In <i>International Conference on Learning</i>	<a href="#">domain question answering</a> . In <i>Proceedings of the</i>	827
775	<i>Representations</i> .	<i>2020 Conference on Empirical Methods in Natural</i>	828
		<i>Language Processing (EMNLP)</i> , pages 6769–6781,	829
		Online. Association for Computational Linguistics.	830

831	Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joon-	<i>the Association for Computational Linguistics (Vol-</i>	887
832	suk Park, and Jaewoo Kang. 2023. <a href="#">Tree of clarifica-</a>	<i>ume 1: Long Papers</i> ), pages 3268–3281, Bangkok,	888
833	<a href="#">tions: Answering ambiguous questions with retrieval-</a>	Thailand. Association for Computational Linguistics.	889
834	<a href="#">augmented large language models</a> . In <i>Proceedings</i>		
835	<i>of the 2023 Conference on Empirical Methods in</i>	Xin Lv, Yixin Cao, Lei Hou, Juanzi Li, Zhiyuan Liu,	890
836	<i>Natural Language Processing</i> , pages 996–1009, Sin-	Yichi Zhang, and Zelin Dai. 2021. <a href="#">Is multi-hop rea-</a>	891
837	gapore. Association for Computational Linguistics.	<a href="#">soning really explainable? towards benchmarking</a>	892
		<a href="#">reasoning interpretability</a> . In <i>Proceedings of the</i>	893
838	Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yu-	<i>2021 Conference on Empirical Methods in Natural</i>	894
839	taka Matsuo, and Yusuke Iwasawa. 2022. Large lan-	<i>Language Processing</i> , pages 8899–8911, Online and	895
840	guage models are zero-shot reasoners. <i>Advances in</i>	Punta Cana, Dominican Republic. Association for	896
841	<i>neural information processing systems</i> , 35:22199–	Computational Linguistics.	897
842	22213.		
843	Abhishek Kumar, Robert Morabito, Sanzhar Umbet,	Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao,	898
844	Jad Kabbara, and Ali Emami. 2024. <a href="#">Confidence</a>	and Nan Duan. 2023. Query rewriting in retrieval-	899
845	<a href="#">under the hood: An investigation into the confidence-</a>	augmented large language models. In <i>Proceedings</i>	900
846	<a href="#">probability alignment in large language models</a> . In	<i>of the 2023 Conference on Empirical Methods in</i>	901
847	<i>Proceedings of the 62nd Annual Meeting of the As-</i>	<i>Natural Language Processing</i> , pages 5303–5315.	902
848	<i>sociation for Computational Linguistics (Volume 1:</i>		
849	<i>Long Papers</i> ), pages 315–334, Bangkok, Thailand.	Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong	903
850	Association for Computational Linguistics.	Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen.	904
		2021a. <a href="#">Generation-augmented retrieval for open-</a>	905
851	Sunjun Kweon, Jiyouon Kim, Heeyoung Kwak,	<a href="#">domain question answering</a> . In <i>Proceedings of the</i>	906
852	Dongchul Cha, Hangyul Yoon, Kwang Hyun Kim,	<i>59th Annual Meeting of the Association for Compu-</i>	907
853	Jeewon Yang, Seunghyun Won, and Edward Choi.	<i>tational Linguistics and the 11th International Joint</i>	908
854	2024. Ehrnoteqa: An llm benchmark for real-world	<i>Conference on Natural Language Processing (Vol-</i>	909
855	clinical practice using discharge summaries. In <i>The</i>	<i>ume 1: Long Papers</i> ), pages 4089–4100, Online. As-	910
856	<i>Thirty-eight Conference on Neural Information Pro-</i>	sociation for Computational Linguistics.	911
857	<i>cessing Systems Datasets and Benchmarks Track</i> .		
858	Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio	Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong	912
859	Petroni, Vladimir Karpukhin, Naman Goyal, Hein-	Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen.	913
860	rich Küttler, Mike Lewis, Wen-tau Yih, Tim Rock-	2021b. <a href="#">Reader-guided passage reranking for open-</a>	914
861	täschel, et al. 2020. Retrieval-augmented generation	<a href="#">domain question answering</a> . In <i>Findings of the Asso-</i>	915
862	for knowledge-intensive nlp tasks. <i>Advances in Neu-</i>	<i>ciation for Computational Linguistics: ACL-IJCNLP</i>	916
863	<i>ral Information Processing Systems</i> , 33:9459–9474.	2021, pages 344–350, Online. Association for Com-	917
		putational Linguistics.	918
864	Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and	Lynne M Markus. 2001. Toward a theory of knowl-	919
865	Joshua B Tenenbaum. 2022. Compositional visual	edge reuse: Types of knowledge reuse situations and	920
866	generation with composable diffusion models. In	factors in reuse success. <i>Journal of management</i>	921
867	<i>European Conference on Computer Vision</i> , pages	<i>information systems</i> , 18(1):57–93.	922
868	423–439. Springer.		
869	Jesus Lovon-Melgarejo, Thouria Ben-Haddi, Jules	Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman	923
870	Di Scala, Jose G. Moreno, and Lynda Tamine.	Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad	924
871	2024. <a href="#">Revisiting the MIMIC-IV benchmark: Exper-</a>	Farajtabar. 2025. <a href="#">GSM-symbolic: Understanding</a>	925
872	<a href="#">iments using language models for electronic health</a>	<a href="#">the limitations of mathematical reasoning in large</a>	926
873	<a href="#">records</a> . In <i>Proceedings of the First Workshop on</i>	<a href="#">language models</a> . In <i>The Thirteenth International</i>	927
874	<i>Patient-Oriented Language Processing (CL4Health)</i>	<i>Conference on Learning Representations</i> .	928
875	@ <i>LREC-COLING 2024</i> , pages 189–196, Torino,		
876	Italia. ELRA and ICCL.	Mahmoud Nazzal, Issa Khalil, Abdallah Khreishah, and	929
877	Fengyu Lu, Jiaxin Duan, and Junfei Liu. 2023. <a href="#">A factual</a>	NhatHai Phan. 2024. Promsec: Prompt optimization	930
878	<a href="#">aware two-stage model for medical dialogue summa-</a>	for secure generation of functional source code with	931
879	<a href="#">rization</a> . In <i>2023 IEEE International Conference</i>	large language models (llms). In <i>Proceedings of the</i>	932
880	<i>on Bioinformatics and Biomedicine (BIBM)</i> , pages	<i>2024 on ACM SIGSAC Conference on Computer and</i>	933
881	2859–2866.	<i>Communications Security</i> , pages 2266–2280.	934
882	Kun Luo, Zheng Liu, Shitao Xiao, Tong Zhou, Yubo	Ankit Pal, Logesh Kumar Umapathi, and Malaikan-	935
883	Chen, Jun Zhao, and Kang Liu. 2024. <a href="#">Landmark</a>	nan Sankarasubbu. 2022. Medmcqa: A large-scale	936
884	<a href="#">embedding: A chunking-free embedding method for</a>	multi-subject multi-choice dataset for medical do-	937
885	<a href="#">retrieval augmented long-context large language mod-</a>	main question answering. In <i>Conference on health,</i>	938
886	<a href="#">els</a> . In <i>Proceedings of the 62nd Annual Meeting of</i>	<i>inference, and learning</i> , pages 248–260. PMLR.	939
		Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew	940
		Spotnitz, RuiJun Chen, Adler Perotte, and Karthik	941
		Natarajan. 2021. Cehr-bert: Incorporating temporal	942

943	information from structured ehr data to improve prediction tasks. In <i>Machine Learning for Health</i> , pages 239–260. PMLR.	999
944		1000
945		1001
946	Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. 2023. <a href="#">Efficiently scaling transformer inference</a> . In <i>Proceedings of Machine Learning and Systems</i> , volume 5, pages 606–624. Curan.	1002
947		
948		1003
949		1004
950		1005
951		1006
952		1007
953	Hongjin Qian, Zheng Liu, Kelong Mao, Yujia Zhou, and Zhicheng Dou. 2024. <a href="#">Grounding language model with chunking-free in-context retrieval</a> . In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 1298–1311, Bangkok, Thailand. Association for Computational Linguistics.	1008
954		1009
955		1010
956		1011
957		1012
958		1013
959	Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J Topol. 2024. Llm-based agentic systems in medicine and healthcare. <i>Nature Machine Intelligence</i> , 6(12):1418–1420.	1014
960		1015
961		1016
962		1017
963		1018
964	Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. <a href="#">SQuAD: 100,000+ questions for machine comprehension of text</a> . In <i>Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing</i> , pages 2383–2392, Austin, Texas. Association for Computational Linguistics.	1019
965		1020
966		1021
967		1022
968		1023
969		1024
970	Ruiyang Ren, Yingqi Qu, Jing Liu, Xin Zhao, Qifei Wu, Yuchen Ding, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2023. <a href="#">A thorough examination on zero-shot dense retrieval</a> . In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 15783–15796, Singapore. Association for Computational Linguistics.	1025
971		1026
972		1027
973		1028
974		1029
975		1030
976		1031
977	David L. Sackett. 1997. <a href="#">Evidence-based medicine</a> . <i>Seminars in Perinatology</i> , 21(1):3–5. Fatal and Neonatal Hematology for the 21st Century.	1032
978		1033
979		1034
980	Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. <a href="#">RAPTOR: Recursive abstractive processing for tree-organized retrieval</a> . In <i>The Twelfth International Conference on Learning Representations</i> .	1035
981		1036
982		1037
983		1038
984		1039
985	Max Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. <i>Scientific Data</i> , 7(1):322.	1040
986		1041
987		1042
988		1043
989	Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. <a href="#">Large language models can be easily distracted by irrelevant context</a> . In <i>Proceedings of the 40th International Conference on Machine Learning</i> , volume 202 of <i>Proceedings of Machine Learning Research</i> , pages 31210–31227. PMLR.	1044
990		1045
991		1046
992		1047
993		1048
994		1049
995		1050
996	Haitian Sun, Tania Bedrax-Weiss, and William Cohen. 2019. Pullnet: Open domain question answering with iterative retrieval on knowledge bases and text.	1051
997		1052
998		1053
		1054
		1055
	In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 2380–2390.	
	Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction {APIs}. In <i>25th USENIX security symposium (USENIX Security 16)</i> , pages 601–618.	
	George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. <i>BMC bioinformatics</i> , 16:1–28.	
	Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024a. Knowledge editing for large language models: A survey. <i>ACM Computing Surveys</i> , 57(3):1–37.	
	Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024b. <a href="#">“my answer is C”: First-token probabilities do not match text answers in instruction-tuned language models</a> . In <i>Findings of the Association for Computational Linguistics: ACL 2024</i> , pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.	
	Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. <a href="#">Self-consistency improves chain of thought reasoning in language models</a> . In <i>The Eleventh International Conference on Learning Representations</i> .	
	Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023b. <a href="#">Self-knowledge guided retrieval augmentation for large language models</a> . In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 10303–10315, Singapore. Association for Computational Linguistics.	
	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. <i>Advances in neural information processing systems</i> , 35:24824–24837.	
	Xiao Wei, Qi Xu, Hang Yu, Qian Liu, and Erik Cambria. 2024. <a href="#">Through the MUD: A multi-defendant charge prediction benchmark with linked crime elements</a> . In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 2864–2878, Bangkok, Thailand. Association for Computational Linguistics.	
	Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024a. Pmc-llama: toward building open-source language models for medicine. <i>Journal of the American Medical Informatics Association</i> , page ocae045.	

Chaoyi Wu, Pengcheng Qiu, Jinxin Liu, Hongfei Gu, Na Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Towards evaluating and building versatile large language models for medicine. *npj Digital Medicine*, 8(1):58.

Zongyu Wu, Hongcheng Gao, Yueze Wang, Xiang Zhang, and Suhang Wang. 2024b. [Universal prompt optimizer for safe text-to-image generation](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6340–6354, Mexico City, Mexico. Association for Computational Linguistics.

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, Bangkok, Thailand. Association for Computational Linguistics.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024b. [Do large language models latently perform multi-hop reasoning?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. [Generate rather than retrieve: Large language models are strong context generators](#). In *The Eleventh International Conference on Learning Representations*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. [Least-to-most prompting enables complex reasoning in large language models](#). In *The Eleventh International Conference on Learning Representations*.

Hongyu Zhu, Sichu Liang, Wentao Hu, Fang-Qi Li, Yali Yuan, Shi-Lin Wang, and Guang Cheng. 2024. [Improve deep forest with learnable layerwise augmentation policy schedules](#). In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6660–6664.

## A Implementation Details

### A.1 Code and Results

The core implementation of the RGAR framework and the output json files can be accessed via the **Anonymous Repository**: <https://anonymous.4open.science/r/RGAR-C613>

### A.2 Hardware Configuration

All experiments were conducted on an in-house workstation equipped with *dual* NVIDIA GeForce RTX 4090 GPUs and an Intel® Core i9-13900K CPU.

## B Additional Experimental Results

### B.1 Additional Evaluation Metric

This work and all compared methods rely solely on accuracy as the evaluation metric. This is because the corpora used for retrieval lack pre-annotated ground truth chunks specific to each question (if such ground truth were available, metrics like recall and precision for retrieved chunks could be calculated). The absence of pre-annotated ground truth more accurately reflects real-world medical advisory scenarios.

Inspired by a recent study from 2025 (Griot et al., 2025), we introduce an additional metric—model confidence in its answers—to provide a finer-grained assessment of whether the results are sufficiently “reliable”. We extract the logits for the four answer options at the position of the generated answer in the format { “answer\_choice”: “X” } and use the probability of the most likely option as the model’s confidence in its answer. Notably, Methods using CoT cannot be evaluated in this manner because their output options are not position-specific, making logit extraction infeasible. The changes in accuracy and average confidence on MedQA-US are presented in Table 4:

Table 4: Accuracy and Model Confidence of Different Methods.

	Custom	RAG	GAR	RGAR
Accuracy	50.20%	53.50%	57.97%	58.83%
Average Confidence	0.6632	0.6524	0.6811	0.7116

It is evident that average confidence generally correlates positively with actual accuracy, with the exception of non-retrieval-based methods, which tend to produce overly confident probability estimates.

First, we categorize results based on actual confidence values. The accuracy and corresponding sample counts for different confidence thresholds are in Table 5:

Table 5: Accuracy of Methods Across Confidence Ranges.

Confidence	Custom	RAG	GAR	RGAR
>0.85	61.60%	76.09%	81.93%	83.60%
<0.85	46.10%	45.84%	47.70%	46.07%
>0.9	65.46%	80.60%	83.50%	86.78%
<0.9	46.48%	47.45%	50.61%	47.69%
Nums	Custom	RAG	GAR	RGAR
>0.85	336	322	382	433
<0.85	937	951	891	840
>0.9	249	232	285	363
<0.9	1024	1041	988	910

These results show that RGAR has a higher number of samples with higher confidence (e.g., >0.85 and >0.9), and these samples exhibit significantly higher accuracy. This indicates that RGAR’s results are sufficiently reliable, allowing users to decide whether further verification of the answer’s correctness is needed based on confidence levels.

Furthermore, to calibrate for potential biases in confidence estimates (e.g., some methods being overly confident or conservative), we rank samples within each method by confidence and compute their respective accuracies. The results are in Table 6:

Table 6: Accuracy by Relative Model Confidence Levels.

Confidence Ranking	Custom	RAG	GAR	RGAR
top 25%	74.21%	76.42%	82.70%	87.11%
25%-50%	55.35%	59.43%	66.67%	63.52%
50%-75%	40.88%	44.03%	49.37%	49.06%
75%-100%	30.41%	34.17%	33.23%	35.74%

Notably, in the top 25% confidence ranking, RGAR achieves a substantial improvement of approximately 4% over the second-best method, GAR. In contrast, GAR’s correctly answered samples are more likely to fall into lower confidence regions, suggesting a ”lack of essential metacognition” in its responses.

## B.2 Additional Analysis of Time Cost

Time cost across all methods on EHRNoteQA are shown in Table 7.

Table 7: Comparison of different methods in terms of execution time (hours).

Method	Custom	CoT	RAG	MedRAG	GAR	<i>i</i> -MedRAG	RGAR
Time/h	0.13	0.96	0.47	1.26	1.52	19.03	4.49

Balancing time overhead and performance is crucial, and our approach achieves this balance. As shown in Tables 2 and 7, RGAR’s time overhead is less than  $0.3 \times$  that of *i*-MedRAG while maintaining comparable or superior performance. On average, RGAR requires about 20 seconds per sample, whereas *i*-MedRAG exceeds 60 seconds, making its overhead impractical. Although RGAR’s per-round overhead is  $1.5 \times$  that of GAR, Figure 3 shows a clear performance gain. For real-time applications, a single-round RGAR offers an optimal trade-off. Other methods lag significantly behind both *i*-MedRAG and RGAR, making them unsuitable for medical applications where reliability is critical.

We further analyze the overhead of different pipeline components in all methods:

(1) Corpus retrieval: Since embedded vectors are pre-saved, retrieval overhead is in the second range, making multiple retrievals negligible. Custom and RAG methods have similar costs.

(2) LLM generation: The CoT method has unstable token lengths (110–4096, avg. 2,433), making its overhead only  $0.6 \times$  to GAR’s. GAR involves three generations, each under 1,000 tokens. GAR’s three queries share input except for prompts (see Equation 4), and existing methods (Pope et al., 2023) suggest that sharing KV cache could potentially make it more efficient.

(3) *i*-MedRAG: Its LLM generation’s overhead in each round includes query decomposition, CoT-based answering of each query, and summarization, leading to a  $4.2 \times$  higher cost than RGAR, even with early stopping.

In summary, RGAR significantly improves upon GAR in just one round, enabling flexible time-performance trade-offs. GAR-like methods may further reduce overhead via shared KV cache techniques.

## B.3 Advantages Over *i*-MedRAG

(1) The average performance improvement of RGAR compared to *i*-MedRAG is relatively modest, largely because *i*-MedRAG is an extremely complex approach, with a time overhead three times that of RGAR. The focus of RGAR is to

demonstrate the importance of extracting factual knowledge from EHRs and the interaction between factual and conceptual knowledge. This is convincingly supported by the comparisons with RAG and GAR in Section 5.1 and the ablation study in Section 5.3.

(2) A key advantage of RGAR is its stable and consistent performance improvement, which is critical for the requirements in medical applications. As shown in Table 2, RGAR ranks among the top two across all three datasets, demonstrating a stable enhancement over both RAG and GAR. In contrast, *i*-MedRAG, despite its substantial time overhead, performs poorly on MedmcQA, ranking near the bottom. This significantly limits its potential for real-world deployment.

(3) An additional advantage of RGAR is its flexibility. Its two main components—factual knowledge extraction and conceptual knowledge retrieval—can be easily integrated into various existing RAG frameworks. For instance, we experimented by adding the factual knowledge extraction module to the initial cycle of *i*-MedRAG. On the MedQA-US dataset, this improved its performance from 56.24% to 58.13%, surpassing GAR’s 57.97% and coming close to RGAR’s 58.83%. This highlights the extensibility and effectiveness of the factual knowledge extraction module. However, due to the prohibitive time overhead—*i*-MedRAG generates  $m$  queries, and combining these with the  $n$  queries from conceptual knowledge retrieval would result in  $m * n$  queries—we did not pursue further combinations. The focus of this paper is to validate the effectiveness of the RGAR approach. Future work will aim to integrate RGAR’s methodology with existing RAG techniques, reduce time overhead, and develop systems that offer a better trade-off between performance and efficiency.

#### B.4 Additional Corpus

While our main experiments are conducted using corpus Textbooks, we acknowledge that corpus size and coverage may influence absolute performance. However, our objective is not to optimize the corpus itself, but rather to investigate how explicit factual knowledge extraction can enhance the architecture of RAG systems. As demonstrated in our main results, the proposed method consistently outperforms strong baselines—including GAR—under the same corpus conditions. This validates the effectiveness of our approach independent of corpus scale.

It is important to note that prior work, such as the MIRAGE benchmark (Xiong et al., 2024a), has shown that while a larger corpus may improve overall accuracy, it does not fundamentally alter the relative advantages among RAG architectures.

In our study, we focus on the practical scenario of deploying a personal health assistant (Qiu et al., 2024) on a consumer-grade GPU (e.g., 24GB VRAM) and a standard desktop system. From this perspective, extremely large corpora such as MedCorp present significant resource challenges. Specifically, the complete storage requirement for MedCorp—including the raw documents and MedCPT embeddings—amounts to 336 GB, and its deployment requires a minimum of 256 GB RAM to load the retrieval index, which poses substantial overhead for individual users or lightweight healthcare applications.

To further support the generality of our findings, we include additional experiments on a mid-sized corpus (StatPearls) and a large-scale evaluation on MedCorp in Table 8. These results reaffirm the effectiveness of our architecture, demonstrating that it remains beneficial across different corpus scales, without relying on massive storage or compute resources.

Table 8: Performance of Different Methods with Varying Corpus Sizes on MedQA-US.

Corpora	Custom	RAG	GAR	RGAR
TextBooks(#125.8k)	50.20%	53.50%	56.24%	58.83%
StatPearls(#301.2k)	50.20%	54.83%	56.48%	58.99%
MedCorp(#65.3M)	50.20%	55.77%	58.20%	60.64%

Specifically, all retrieval-based methods benefit from a larger corpus on the MedQA-US dataset. However, even when using the much larger MedCorp corpus, GAR does not outperform RGAR evaluated on the smaller textbook corpus. This indicates that the performance gains of RGAR over GAR remain stable across corpora of different sizes.

In terms of runtime, as shown in Figure 6, all multi-stage retrieval methods experience a significant increase in latency on MedCorp, primarily due to the cost of retrieval rather than generation—each retrieval step incurs an average delay of approximately 10 seconds. This further highlights RGAR’s suitability for deployment on consumer-grade systems, where both memory and latency are limited.

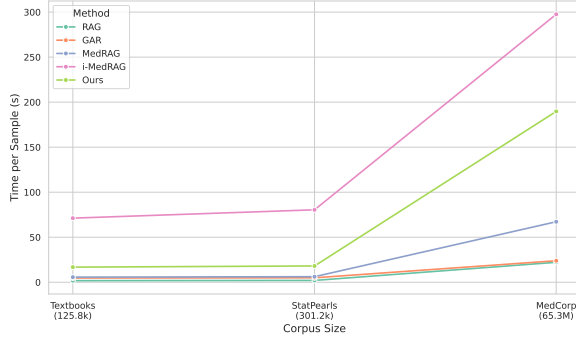


Figure 6: Time Overhead of Different Methods with Varying Corpus Sizes.

## B.5 Fine-grained Analysis of Performance on Different Datasets

When analyzing performance across different datasets, we find that retrieval-based methods perform significantly better on MedQA-US and EHRNoteQA, while MedMCQA shows a negative effect—consistent with results reported by MedRAG (Xiong et al., 2024a). A closer analysis reveals that MedMCQA incorporates arithmetic reasoning questions (roughly 7% of the total), and the addition of extensive retrieved contexts diminishes the model’s numerical reasoning capabilities, which could potentially be fixed with larger base LLMs (Mirzadeh et al., 2025). Nonetheless, among retrieval-based methods, our RGAR stands out as the only approach that outperforms vanilla RAG on this dataset, delivering an improvement of more than 1% over Custom.

## C Prompt Template and Case Study

For simplicity, we merged EHR and question in the prompt words of the answer and treated them as question in the prompt words. Table 9 shows the prompts template of RGAR and compared work (Using CoT ones). Table 10 shows the input of a sample, Table 11 shows the final output of RGAR.

### C.1 Simple Case Study

Given that our method operates with a retrieval budget of only 32 documents—and that medical question answering inherently requires domain-specific reasoning—we include a simplified case study to illustrate why traditional approaches may fall short under such constraints. This example highlights the challenges faced by earlier methods in capturing and integrating critical patient-specific risk factors with external medical knowledge, and contrasts

them with the advantages of our proposed framework.

**Case.** A 60-year-old male patient presents with persistent cough, fever, and mild dyspnea. The hospital’s EHR includes not only symptom descriptions but also chest X-ray results, blood oxygen levels, prior diagnoses (e.g., diabetes, COPD), allergy history, and lab findings. In addition, external medical literature provides authoritative guidelines on pulmonary infections, comorbidity considerations, and evidence-based treatment strategies.

**Limitations of Traditional Methods.** Basic retrieval-based methods tend to issue dispersed queries over all surface-level mentions in the EHR, retrieving fragmented conceptual knowledge loosely related to individual symptoms. This makes it difficult to focus on high-risk factors specific to the current condition—such as comorbid diabetes or COPD. Query decomposition further fragments retrieval results, lacking coherence or clinical focus. These methods heavily depend on a sufficiently strong language model to accurately identify critical information from a large and often noisy textual input after retrieval—whether it be high-risk factors embedded in the original EHR or appropriate treatment strategies extracted from retrieved documents. This reliance becomes particularly problematic when deploying LLMs on resource-constrained environments, such as consumer-grade GPUs, where inference capabilities and context handling are limited.

Generative retrieval approaches like GAR allow the model to generate an intermediate answer and use it to retrieve supporting content. While this enables partial recognition of high-risk factors and relevant treatments, it heavily relies on the model’s internal conceptual knowledge. As a result, its effectiveness declines in complex cases requiring deeper medical understanding.

**Advantages of RGAR.** Our method explicitly encourages the integration of conceptual and factual knowledge. In the first retrieval stage, external conceptual knowledge (e.g., clinical guidelines on comorbidities) is retrieved and provided to the model. This serves as guidance for iterative reasoning, allowing the model to focus on identifying patient-specific high-risk factors—such as the interaction between diabetes, COPD, and current symptoms.

Through multiple rounds of reasoning, the model captures critical factual elements from the EHR (e.g., allergy history, abnormal blood oxygen lev-

els) and aligns them with relevant conceptual insights (e.g., recommended antibiotic choices for diabetic patients). This integrated process results in more accurate and interpretable treatment suggestions, grounded in both structured medical evidence and patient-specific context.

## D Framework Insight

### D.1 the Rationality of Bloom’s Taxonomy

We recognize that there may be concerns regarding the use of Bloom’s Taxonomy in our framework, particularly the potential implication that it imposes a rigid dichotomy between factual and conceptual knowledge. However, both the original taxonomy and our application through the RGAR framework emphasize the integration—rather than the separation—of these two forms of knowledge. Our work does not advocate for treating factual and conceptual knowledge as disjoint entities; rather, it highlights their complementary roles in effective problem-solving, a view that is explicitly articulated in our abstract and substantiated through empirical analyses, including targeted ablation studies.

In Section D.2, we further clarified that factual and conceptual knowledge originate from fundamentally different sources, and this distinction naturally aligns with the two types of knowledge defined in Bloom’s Taxonomy. Traditional RAG setups often fail to recognize the distinction between factual and conceptual knowledge, which leads to a lack of dedicated extraction for factual knowledge and makes it impossible to facilitate interaction between the two types of knowledge. In contrast, our method is specifically designed to handle these two forms of knowledge separately and enables meaningful interaction between them.

If the concern is that using separate modules enforces an artificial separation, then by this logic all RAG systems are inherently “bifurcated”, since they embed a query including factual knowledge to retrieve conceptual knowledge. The key distinction of our work lies in foregrounding factual knowledge extraction and promoting its interaction with conceptual retrieval, which stands in contrast to the rigid separation in existing systems. **Our approach, RGAR, does not divide knowledge more but integrates it more effectively.**

### D.2 Organization of early datasets

Early RAG-based methods were shaped by the structure of existing QA datasets. For instance,

early benchmarks like PubMedQA provided only the question as input for retrieval. Subsequent datasets, such as MedQA-US, introduced associated factual knowledge but presented it concatenated with the question, effectively treating the two as a single input. It was not until the introduction of EHRNoteQA that EHRs and questions were explicitly provided as separate components. As a result, existing retrieval methods were naturally designed to operate on unified question–context inputs, without explicitly distinguishing factual knowledge from the query itself.

### D.3 Another View of the Recurrence Pipeline

We conceptualize the Recurrence Pipeline as an exploration-exploitation process within the reinforcement learning framework (Auer et al., 2002). In GAR, even when generated content is only partially accurate (or potentially inaccurate), it remains valuable for retrieval if it correlates with passages containing correct information (e.g., co-occurrence with correct answers), thus representing an exploratory phase. Conversely, EHR extraction serves as an exploitation phase, thoroughly utilizing explored knowledge by selecting relevant components and synthesizing new evidence (factual knowledge). Based on this newly derived evidence, subsequent iterations can initiate fresh exploration-exploitation cycles, creating a continuous knowledge transmission process (Zhu et al., 2024).

In scenarios where additional factual knowledge is not required, the retrieved content tends to remain relatively constant, and utilizing this content under identical prompting conditions would likely yield similar factual knowledge through extraction and summarization. However, when conceptual knowledge is needed to derive new factual knowledge through reasoning from existing factual information, the updated basic query facilitates easier retrieval of conceptual knowledge supporting current reasoned factual knowledge, thereby maintaining the integrity of reasoning chains. Furthermore, leveraging current factual knowledge for retrieval enables the exploration and discovery of novel knowledge domains.

### D.4 Why No Flexible Stopping Criteria

Similar multi-round RAG systems have adopted more flexible stopping criteria. For instance, Adaptive RAG (Jeong et al., 2024) determines whether to retrieve further by consulting the model itself. *i*-MedRAG (Xiong et al., 2024b), while setting a

maximum number of retrieval iterations, also supports early stopping.

In our RGAR framework, we do not adopt such settings. On the one hand, we focus on evaluating how additional processing of *factual knowledge* enhances retrieval performance, raising awareness of this often-overlooked type of knowledge in previous RAG systems, while flexible stopping criteria mainly showcase procedural knowledge and metacognitive knowledge. On the other hand, the metacognitive capabilities of current LLMs remain under question, as a model’s self-evaluation of the need for additional retrieval information often does not match actual requirements (Kumar et al., 2024).

## D.5 Generalizability of the Framework

Since RGAR maintains the same input-output structure as standard RAG systems, it is well-suited for any retrieval scenario, regardless of the presence of Electronic Health Records (EHRs). Its advantages become particularly evident when handling long EHR texts. The framework accepts a string input, which undergoes additional partitioning to extract EHRs and questions. In scenarios where EHRs are unavailable, the factual knowledge extraction module is not executed; instead, the question is rewritten with retrieved conceptual knowledge. The output is formatted as a JSON object, facilitating the inclusion of intermediate system outputs.

From the perspective of future scalability, the evolution of LLM agents suggests that private LLM health assistants will gain access to more extensive historical health data from owners (patients), including EHRs, enabling more comprehensive question answering (Qiu et al., 2024). This anticipated expansion emphasizes the importance of distinguishing inputs beyond the question, particularly factual information, thereby validating the rationale behind our framework.

To demonstrate the framework’s generalizability, we evaluated its performance on the MMLU-Med dataset (Hendrycks et al., 2021) in Table 2, which lacks EHRs. Our experimental results, indicate that RGAR consistently outperformed GAR, albeit with a relatively modest improvement compared to datasets containing EHRs.

## D.6 Future Work

Our RGAR framework leverages retrieved medical domain knowledge to deliver exceptional answer quality. However, we are concerned that such powerful generative capabilities, if maliciously ex-

ploited, could pose security risks. For instance, when the retrieved corpus contains private or copyrighted information, malicious users could exploit the LLM’s responses to extract and disclose sensitive data from the corpus (Carlini et al., 2021). Additionally, malicious users might attempt to replicate our base LLM (Tramèr et al., 2016) by collecting large volumes of question-answer pairs or infer internal details of our retrieval-based generation framework (Carlini et al.). We will make every effort to mitigate these risks, such as verifying the legitimacy of queries (Inan et al., 2023), ensuring that RGAR is used responsibly and legally.

## E Dataset Description and Analysis

### E.1 Dataset Coverage Overview

The datasets used in our study collectively span a broad range of medical domains:

- MedQA-US focuses on general clinical medicine within the scope of the USMLE examination.
- MedMCQA encompasses 21 medical specialties, including cardiology, oncology, dermatology, and more.
- EHRNoteQA covers real-world scenarios such as inpatient management, emergency medicine, and intensive care.
- MMLU-Med targets basic medical sciences and related fields, including anatomy, genetics, medical ethics, and public health.

The Textbooks corpus utilized in our study comprises content from 18 widely recognized medical textbooks, extensively used by medical students and USMLE candidates. This corpus encompasses a broad spectrum of medical disciplines, including internal medicine, pediatrics, surgery, obstetrics and gynecology, psychiatry, pharmacology, pathology, and foundational sciences such as anatomy, physiology, and biochemistry. Given this extensive coverage, the Textbooks corpus aligns well with the domains addressed in our evaluated datasets—MedQA-US, MedMCQA, EHRNoteQA, and MMLU-Med—thereby serving as a representative and appropriate retrieval corpus for our experiments.

Regarding question types, multiple-choice QA is the most commonly used format and is the type

employed in all comparative analyses in this paper. Open-ended (generative) QA datasets, which primarily evaluate the quality of generated text summaries rather than the ability to solve medical problems (Savery et al., 2020), are beyond the scope of this study. However, addressing such datasets is indeed a necessary step toward real-world applications.

## E.2 Representative Examples of different datasets

Figure 1 and Table 1 in the main text illustrate the distinctions among datasets with respect to the involvement of factual knowledge, specifically electronic health records (EHRs) in the case of medical questions. In this section, we present representative samples from all the datasets referenced throughout the paper.

From the example of MMLU-Med in Table 15, it can be observed that the length of the input primarily stems from the inclusion of extensive references to literature viewpoints and empirical findings, which are used to support complex reasoning. This characteristic contributes to its status as a representative and challenging medical QA dataset. However, in comparison to the three datasets discussed above, MMLU-Med still contains little to no factual knowledge specific to individual patients; that is, it lacks detailed depictions of patient-specific information. As shown in Table 2, RGAR continues to exhibit strong performance on this type of dataset. To some extent, this highlights the generalizability of our approach: the FKE module remains effective in scenarios involving lengthy inputs that require distillation and extraction of key information.

## E.3 Comparative Analysis of Dataset Length Distributions

In this section, we present additional visualizations comparing the two categories of datasets we described, and explain our rationale for excluding the MMLU-Med dataset (Hendrycks et al., 2021). We plotted smoothed Kernel Density Estimation (KDE) curves for these datasets, as shown in Figure 7. Our analysis confirms that datasets containing Electronic Health Records (EHR) consistently demonstrate greater length compared to those without EHR content. However, certain datasets exhibit complex question sources and types. For instance, while the MMLU-Med dataset exhibits a considerable mean length of 84 tokens and a maximum

length of up to 961 tokens, the primary source of this length is not factual knowledge such as EHRs. Moreover, its length distribution is highly skewed: the majority of samples are relatively short, with only a small fraction being significantly longer. This distribution differs substantially from that of medical QA datasets involving EHRs, where longer inputs are more consistently present. As a result, we exclude MMLU-Med from our main experimental evaluation. Nevertheless, we still report results on this dataset, given its prominence and representativeness in the current landscape of medical QA benchmarks.

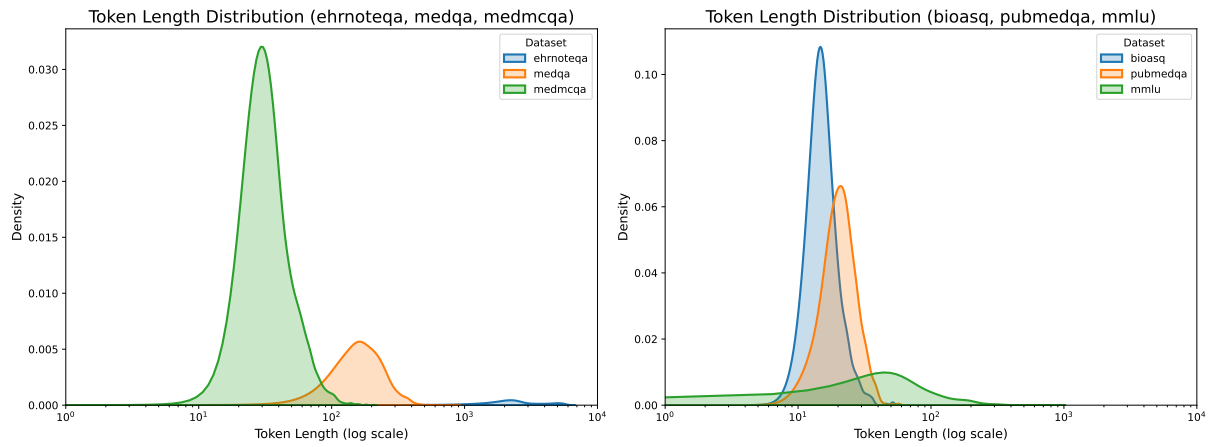


Figure 7: Length Distribution Analysis of Medical QA Datasets with and without EHR.

System prompts for Non-CoT
You are a helpful medical expert, and your task is to answer a multi-choice medical question using the relevant documents. Organize your output in a json formatted as Dict {"answer_choice": Str{A/B/C/...}}. Your responses will be used for research purposes only, so please have a definite answer. Please just give me the json of the answer.
System prompts for using CoT
You are a helpful medical expert, and your task is to answer a multi-choice medical question. Please first think step-by-step and then choose the answer from the provided options. Organize your output in a json formatted as Dict{"step_by_step_thinking": Str(explanation), "answer_choice": Str{A/B/C/...}}. Your responses will be used for research purposes only, so please have a definite answer. Please just give me the json of the answer.
Answer prompts for Non-CoT
Here are the relevant documents: {{context}}
Here is the question: {{question}}
Here are the potential choices: {{options}}
Please just give me the json of the answer. Generate your output in json:
Answer prompts for Using CoT
Here are the relevant documents: {{context}}
Here is the question: {{question}}
Here are the potential choices: {{options}}
Please think step-by-step and generate your output in one json:
Extracting EHR prompts
Here are the relevant knowledge sources: {{context}}
Here are the electronic health records: {{ehr}}
Here is the question: {{question}}
Please analyze and extract the key factual information in the electronic health records relevant to solving this question and present it as a Python list. Use concise descriptions for each item, formatted as ["key detail 1", ..., "key detail N"]. Please only give me the list. Here is the list:
Generating Possible Answer prompts
Please give 4 options for the question. Each option should be a concise description of a key detail, formatted as: A. "key detail 1" B. "key detail 2" C. "key detail 3" D. "key detail 4"
Generating Possible Title prompts
Please generate some titles of references that might address the above question. Please give me only the titles, formatted as: ["title 1", "title 2", ..., "title N"]. Please be careful not to give specific content and analysis, just the title.
Generating Possible Contexts prompts
Please generate some knowledge that might address the above question. please give me only the knowledge.

Table 9: Prompt templates used in RGAR and Compared Methods.

EHR
A 39-year-old woman is brought to the emergency department because of fevers, chills, and left lower quadrant pain. Her temperature is 39.1°C (102.3°F), pulse is 126/min, respirations are 28/min, and blood pressure is 80/50 mm Hg. There is blood oozing around the site of a peripheral intravenous line. Pelvic examination shows mucopurulent discharge from the cervical os and left adnexal tenderness. Laboratory studies show: Platelet count 14,200/mm <sup>3</sup> Fibrinogen 83 mg/mL (N = 200–430 mg/dL) D-dimer 965 ng/mL (N < 500 ng/mL)
Questions
When phenol is applied to a sample of the patient's blood at 90°C, a phosphorylated N-acetylglucosamine dimer with 6 fatty acids attached to a polysaccharide side chain is identified. A blood culture is most likely to show which of the following?
Options
<p>A. Coagulase-positive, gram-positive cocci forming mauve-colored colonies on methicillin-containing agar</p> <p>B. Encapsulated, gram-negative coccobacilli forming grey-colored colonies on charcoal blood agar</p> <p>C. Spore-forming, gram-positive bacilli forming yellow colonies on casein agar</p> <p><b>D. Lactose-fermenting, gram-negative rods forming pink colonies on MacConkey agar</b></p>

Table 10: Case Study of a Sample Input.

Generated Title
"Identification of Streptococcus pneumoniae in Clinical Specimens", "Phosphorylcholine-Binding Protein A of Staphylococcus aureus", "D-Dimer Assay for Diagnosing Deep Vein Thrombosis", "The Role of Fibrin Degradation Products in Hemostasis", "Infection Control Measures for Hospital-Acquired Bloodstream Infections"
Generate Context
The description suggests Neisseria gonorrhoeae infection. Neisseria gonorrhoeae produces lipopolysaccharides in its outer membrane, specifically LPS or endotoxin. When heated, these lipopolysaccharides undergo cross-linking reactions resulting in the formation of phosphorylcholine-attached oligomers such as those described in the question stem. These structures can be detected by ELISA using antibodies against specific epitopes on the surface of N. gonorrhoeae cells. The presence of this structure would suggest an active infection due to N. gonorrhoeae. Blood cultures for N. gonorrhoeae typically require specialized media containing antibiotics like vancomycin, colistin, and nystatin to inhibit competing bacteria
Generate Answer
<p>Option 1: Streptococcus pneumoniae The presence of a high fever, tachycardia, hypotension, and low platelets suggests sepsis or severe infection. The laboratory findings also indicate disseminated intravascular coagulation (DIC). The identification of a phosphorylated N-acetylglucosane dimer with 6 fatty acid chains attached to a polysaccharide side chain after applying phenol at 90°C is consistent with pneumococcal capsular polysaccharides.</p> <p>Option 2: Staphylococcus aureus Staphylococci can cause various infections but do not typically produce this type of capsule structure when exposed to heat.</p> <p>Option 3: Escherichia coli E. coli does have capsules that may be detected by certain methods such as latex agglutination tests; however, they don't match the description given in the prompt.</p> <p>Option 4: Neisseria meningitidis Neisseria species like N. meningitidis would more commonly present with symptoms related to central nervous system involvement rather than pelvic signs described in the scenario.</p>
Extract EHR
<p>Patient age: 39 years</p> <p>Symptoms: fever, chills, left lower quadrant pain</p> <p>Vital signs: <b>high</b> temperature (39.1°C), tachycardia (126/min), tachypnea (28/min), hypotension (80/50 mmHg)</p> <p>Physical exam findings: mucopurulent discharge from cervix, left adnexal tenderness</p> <p>Laboratory results: <b>low</b> platelet count (14,200/mm<sup>3</sup>), <b>elevated</b> D-dimer (965 ng/mL)</p> <p>Phenol test result: phosphorylated N-acetylglucosamine dimer with 6 fatty acids attached to a polysaccharide side chain</p>
Output Json
{"answer_choice": "D"}

Table 11: Case Study of output in RGAR.

---

Representative sample from the **MedQA-US** dataset

---

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

---

Table 12: Case study of a representative sample from the **MedQA-US** dataset.

---

Representative sample from the **MedMCQA** dataset

---

29 yrs old woman with a pregnancy of 17 week has a 10 years old boy with down syndrome. She does not want another down syndrome kid; best advice to her is

- A. No test is required now as her age is below 35 years
- B. Ultra sound at this point of time will definitely tell her that next baby will be down syndromic or not
- C. Amniotic fluid samples plus chromosomal analysis will definitely tell her that next baby will be down syndromic or not
- D. blood screening at this point of time will clear the exact picture

---

Table 13: Case study of a representative sample from the **MedMCQA** dataset.

---

Representative sample from the **EHRNoteQA** dataset

---

Patient ID: 15455707\nAdmission ID: 24016271\nChartdate: 2172-06-17\nName: \_\_\_\_ Unit No: \_\_\_\_\nAdmission Date:\_\_\_\_ Discharge Date:\_\_\_\_\nDate of Birth:\_\_\_\_ Sex: M\nService: PLASTIC\nAllergies:\nNo Known Allergies / Adverse Drug Reactions\nAttending:\_\_\_\_\nChief Complaint:\nCrush injury to bilateral index fingers consistent with a flexor\ntendon laceration\nMajor Surgical or Invasive Procedure:\n\_\_\_\_ Bilateral IF flexor tendon repairs\nHistory of Present Illness:\n\_\_\_\_ otherwise healthy male s/p work accident on \_\_\_\_ when his\nhands were pulled into conveyor belt. He is here today for\nrepair of bilateral index finger crush injuries.\nPast Medical History:\nNone\nSocial History:\n\_\_\_\_\nFamily History:\nNon-contributory\nPhysical Exam:\nPre-procedure physical exam as documented in Dr.\_\_\_\_: He is well appearing.\nCARDIAC: He has palpable pulses without arrhythmia.\nLUNGS: He is breathing room air without shortness breath or\ncough.\nMUSCULOSKELETAL: Focused of upper extremity examination, hands\nare well perfused bilaterally with palpable radial artery with\ngood cap refill in all five digits including lacerated digits\nwith volar lacerations overlying the P2 of the left index finger\nand as well as the P2 and P3 of the right index finger with\nsegmental lacerations transversely. He denies paresthesias in\nthe radial and ulnar border of the index, middle, ring, small or\nthumb bilaterally. He is unable to make a composite fist with\nno active motion demonstrated at the PIP of either index finger\nnor DIP of either index finger.\nBrief Hospital Course:\nThe patient was admitted to the plastic surgery service on\n\_\_\_\_ and had operative repair of bilateral index finger\ncrush injuries. Please see operative note for further details\nof procedure. The patient tolerated the procedure well.\n\nNeuro: Post-operatively, the patient received IV pain medication\nwith good effect and adequate pain control. When tolerating oral\nintake, the patient was transitioned to oral pain medications.\n\nCV: The patient was stable from a cardiovascular standpoint;\nvital signs were routinely monitored.\n\nPulmonary: The patient was stable from a pulmonary standpoint;\nvital signs were routinely monitored.\n\nGI/GU: Post-operatively, the patient was given IV fluids until\ntolerating oral intake. His diet was advanced when appropriate,\nwhich was tolerated well. He was also started on a bowel regimen\nto encourage bowel movement. Intake and output were closely\nmonitored.\n\nAt the time of discharge on POD#1, the patient was doing well,\nafebrile with stable vital signs, tolerating a regular diet,\nambulating, voiding without assistance, and pain was well\ncontrolled. Patient had bilateral splints in place.\nMedications on Admission:\nNone\nDischarge Medications:\n1. Acetaminophen 650 mg PO Q6H:PRN pain\n2. OxycodONE (Immediate Release) \_\_\_\_ mg PO Q4H:PRN pain\nDischarge Disposition:\nHome\nDischarge Diagnosis:\nbilateral index fingers crush injury consistent with bilateral\nflexor tendon lacerations\nDischarge Condition:\nMental Status: Clear and coherent. \_\_\_\_Speaking)\nLevel of Consciousness: Alert and interactive.\nActivity Status: Ambulatory - Independent.\nDischarge Instructions:\nFollowup Instructions: \_\_\_\_\nQuestion: What was the patient's condition like at the time of discharge, particularly focused on his vital signs, pain management and mobility?

- A. The patient was fairly stabilized, with pain under control, consuming a regular diet, and able to walk and relieve himself without assistance
  - B. The patient was on a repetitive intake of IV fluids and required IV painkillers.
  - C. Patient was experiencing altered states of consciousness, still in distress due to pain, and not able to ambulate
  - D. The patient was responding well to the oral pain medications and was capable of consistent motion at the PIP of index fingers
  - E. Patient still required high-dependency care with heart rate and blood pressure under constant monitoring
- 

Table 14: Case study of a representative sample from the **EHRNoteQA** dataset.

Sauna use, sometimes referred to as *sauna bathing*, is characterized by short-term passive exposure to extreme heat. This exposure elicits mild hyperthermia and an increase in the body's core temperature that induces a thermoregulatory response involving neuroendocrine, cardiovascular, and cytoprotective mechanisms that work together to restore homeostasis and condition the body for future heat stressors. In recent decades, sauna bathing has emerged as a means to increase lifespan and improve overall health, based on compelling data from observational, interventional, and mechanistic studies. Of particular interest are the findings from studies of participants in the Kuopio Ischemic Heart Disease Risk Factor (KIHD) Study, an ongoing prospective population-based cohort study of health outcomes in more than 2,300 middle-aged men from eastern Finland, which identified strong links between sauna use and reduced death and disease. The KIHD findings showed that men who used the sauna two to three times per week were 27 percent less likely to die from cardiovascular-related causes than men who didn't use the sauna. [2] Furthermore, the benefits they experienced were found to be dose-dependent: Men who used the sauna roughly twice as often, about four to seven times per week, experienced roughly twice the benefits and were 50 percent less likely to die from cardiovascular-related causes. [2] In addition, frequent sauna users were found to be 40 percent less likely to die from all causes of premature death. These findings held true even when considering age, activity levels, and lifestyle factors that might have influenced the men's health. [2]... The KIHD also revealed that frequent sauna use reduced the risk of developing dementia and Alzheimer's disease in a dose-dependent manner. Men who used the sauna two to three times per week had a 66 percent lower risk of developing dementia and a 65 percent lower risk of developing Alzheimer's disease, compared to men who used the sauna only one time per week. The health benefits associated with sauna use extended to other aspects of mental health, as well. Men participating in the KIHD study who used the sauna four to seven times per week were 77 percent less likely to develop psychotic disorders, regardless of the men's dietary habits, socioeconomic status, physical activity, and inflammatory status (as measured by C-reactive protein). Exposure to high temperature stresses the body, eliciting a rapid, robust response. The skin and core body temperatures increase markedly, and sweating ensues. The skin heats first, rising to 40°C (104°F), and then changes in core body temperature occur, rising slowly from 37°C (98.6°F, or normal) to 38°C (100.4°F) and then rapidly increasing to 39°C (102.2°F). Cardiac output, a measure of the amount of work the heart performs in response to the body's need for oxygen, increases by 60 to 70 percent, while the heart rate (the number of beats per minute) increases and the stroke volume (the amount of blood pumped) remains unchanged. [5] During this time, approximately 50 to 70 percent of the body's blood flow is redistributed from the core to the skin to facilitate sweating. The average person loses approximately 0.5 kg of sweat while sauna bathing. [11] Acute heat exposure also induces a transient increase in overall plasma volume to mitigate the decrease in core blood volume. This increase in plasma volume not only provides a reserve source of fluid for sweating, but it also acts like the water in a car's radiator, cooling the body to prevent rapid increases in core body temperature and promoting heat tolerance. Repeated sauna use acclimates the body to heat and optimizes the body's response to future exposures, likely due to a biological phenomenon known as hormesis, a compensatory defense response following exposure to a mild stressor that is disproportionate to the magnitude of the stressor. Hormesis triggers a vast array of protective mechanisms that not only repair cell damage but also provide protection from subsequent exposures to more devastating stressors. The physiological responses to sauna use are remarkably similar to those experienced during moderate- to vigorous-intensity exercise. In fact, sauna use has been proposed as an alternative to exercise for people who are unable to engage in physical activity due to chronic disease or physical limitations. [13]

Based on the article, which of the following statements is the author likely to agree with?

- A. Heart surgery patients who cannot run on treadmills may benefit from sauna use.
- B. Patients on a diet would benefit from sauna use.
- C. Salt restriction would be equal to sauna use for hypertensive patients.
- D. Patients with skin conditions may be cured with sauna use.

---

Table 15: Case study of a representative sample from the **MMLU-Med** dataset.

Representative sample from the <b>BioASQ-Y/N</b> dataset
Can losartan reduce brain atrophy in Alzheimer’s disease?
A. Yes
B. No

Table 16: Case study of a representative sample from the **BioASQ-Y/N** dataset.

Representative sample from the <b>PubMedQA</b> dataset
Is anorectal endosonography valuable in dyschesia?
A. Yes
B. No
C. Maybe

Table 17: Case study of a representative sample from the **PubMedQA** dataset.