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

Anonymous ACL submission

Abstract

Medical question answering requires extensive access to specialized *conceptual knowledge*. The current paradigm, Retrieval-Augmented Generation (RAG), acquires expertise medical knowledge through large-scale corpus retrieval and uses this knowledge to guide a general-purpose large language model (LLM) for generating answers. However, existing retrieval approaches often overlook the importance of *factual knowledge*, which limits the relevance of retrieved conceptual knowledge and restricts its applicability in real-world scenarios, such as clinical decision-making based on Electronic Health Records (EHRs). This paper introduces RGAR, a recurrence generation-augmented retrieval framework that retrieves both relevant *factual* and *conceptual* knowledge from dual sources (i.e., EHRs and the corpus), allowing them to interact and refine each another. Through extensive evaluation across three factual-aware medical question answering benchmarks, RGAR establishes a new state-of-the-art performance among medical RAG systems. Notably, the Llama-3.1-8B-Instruct model with RGAR surpasses the considerably larger, RAG-enhanced GPT-3.5. Our findings demonstrate the benefit of extracting factual knowledge for retrieval, which consistently yields improved generation quality.

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).

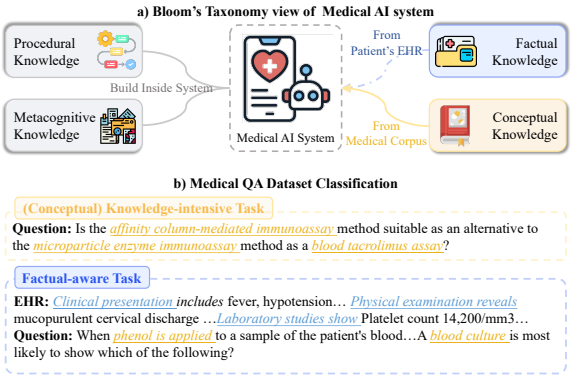


Figure 1: a) Medical AI Systems from the Perspective of Bloom's Taxonomy. b) Two Types of Medical Question Answering Tasks.

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. Recent advances, such as **Generation-Augmented Retrieval (GAR)** (Mao et al., 2021a), focus on enhancing retrieval performance by generating relevant context for query expansion.

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). Electronic Health Records (EHRs) typically contain extensive patient data, including comprehensive diagnostic test results and medical

histories (Pang et al., 2021). However, for any specific medical query, only a small subset of this information is typically relevant, and retrieval performance can be significantly degraded when queries are diluted with extraneous EHR content (Johnson et al., 2023; Lovon-Melgarejo et al., 2024).

We highlight that current *retrieval methods* often fail to adequately consider *factual information* in real-world medical scenarios. Crucially, even when applying query expansion with GAR, the persistent oversight of factual information fundamentally limits their ability to retrieve real relevant documents.

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 typically embedded within advanced RAG systems. Specifically, *Procedural Knowledge* refers to the processes and strategies required to solve problems, such as problem decomposition and retrieval (Wei et al., 2022; Zhou et al., 2023), while *Metacognitive Knowledge* pertains to an LLM's ability to assess whether it has sufficient knowledge or evidence to perform effective reasoning (Kim et al., 2023; Wang et al., 2023b).

Factual Knowledge and *Conceptual Knowledge* require retrieval from large databases containing substantial amounts of irrelevant content, corresponding to the EHRs of patients and medical corpora in answering medical questions. 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 retrieves *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-enhanced 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 (Sarthi 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 applied 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 per-

formance 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 $\mathcal{Q}$, the electronic health record (EHR) as factual knowledge $\mathcal{F}$, and a set of candidate answer $\mathcal{A} = \{a_1, ..., a_{|\mathcal{A}|}\}$. 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:

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

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 $\mathcal{C} = \{c_1, ..., c_N\}$, from a chunked corpus (knowledge base) $\mathcal{K}$. This answering process is expressed as:

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

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 $q_b = \mathcal{Q} \oplus \mathcal{F}$, where $\oplus$ denotes text concatenation.

Traditional dense retrievers, such as Dense Passage Retrieval (DPR) (Karpukhin et al., 2020), identify the top-N relevant chunks 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 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:

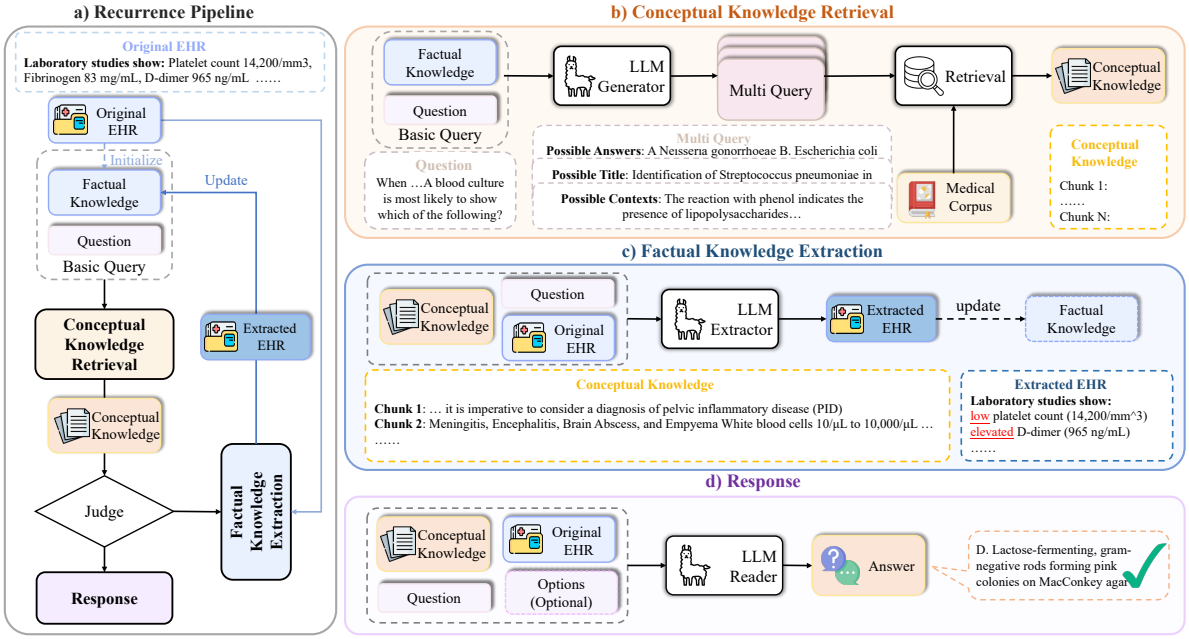


Figure 2: The Overall Framework of RGAR. a) The Recurrence Pipeline in § 3.4; b) Conceptual Knowledge Retrieval in § 3.2; c) Factual Knowledge Extraction in § 3.3; d) Response Template in § 3.4.

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

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

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

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 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:

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

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 $\mathcal{F}_e$:

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

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 Experiments

4.1 Experimental Setup

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

Table 1: Medical QA Benchmark Statistics.

Benchmarks	Max. Len	Avg. Len	Min. Len
Non-EHR QA Benchmarks			
BioASQ-Y/N	52	17	9
PubMedQA	57	23	10
EHR QA Benchmarks			
MedMCQA	207	41	11
MedQA-USMLE	872	197	50
EHRNoteQA	5782	3061	667

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.

4.1.2 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 dataset (Jin et al., 2019), a lightweight collection of 125.8k chunks derived from medical textbooks, with an average length of 182 tokens.

4.1.3 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 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

Table 2: Comparison of RGAR with Other Methods on Three Factual-Aware Datasets. Δ Indicates Improvement Over Custom, **Bold** Represents the Best, and Underline Indicates the Second-Best.

Method		MedQA-USMLE (# 1273)		MedMCQA(# 4183)		EHRNoteQA(# 962)		Average(↓)	
		Acc.	Δ	Acc.	Δ	Acc.	Δ	Acc.	Δ
w/o Retrieval	Custom	50.20	0.00	50.01	0.00	47.19	0.00	49.13	0.00
	CoT	51.45	1.25	44.53	-5.48	62.89	15.70	52.96	3.82
w/ Retrieval	RAG	53.50	3.30	50.54	0.53	61.12	13.93	55.05	5.92
	MedRAG	50.27	0.07	47.53	-2.48	70.58	23.39	56.13	6.99
	GAR	57.97	7.77	50.42	0.41	65.48	18.29	57.96	8.82
	i-MedRAG	56.24	6.04	44.94	-5.07	**74.22**	**27.03**	58.47	9.33
	RGAR	**58.83**	**8.63**	**51.02**	**1.01**	73.28	26.09	**61.04**	**11.91**

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.

4.1.4 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).

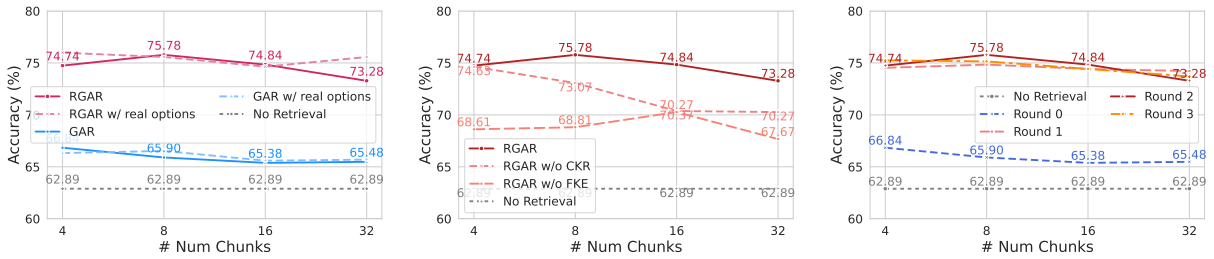
4.2 Main Results

4.2.1 Cross-Dataset Performance Improvement

We evaluate RGAR with the Llama-3.2-3B-Instruct across three factual-aware medical datasets, comparing it with several competitive baselines. Table 2 presents the results of all methods, along with their relative improvements over the Custom baseline. RGAR achieves the highest average performance across the three datasets, surpassing the second-best method, i-MedRAG, by 2%. The retrieval-based methods, even the lowest-performing RAG, consistently outperform the non-retrieval methods Custom and CoT. This highlights the importance of retrieving specialized medical knowledge when using general-purpose LLMs to answer professional medical queries. Comparing different retrieval methods, GAR outperforms vanilla RAG by approximately 3% on average, with a maximum improvement of 4.37% across datasets. This indicates that generating multiple queries for retrieval provides consistent benefits. However, while performing well on EHRNoteQA, MedRAG demonstrates a negative effect on the other two datasets compared to vanilla RAG.

Notably, the improvements achieved by our RGAR over GAR exhibit a positive correlation with the average length of the dataset's context. On EHRNoteQA, which has an average context length exceeding 3000 tokens, our approach achieved a 7.8% improvement. This validates the advantage of our *Factual knowledge Extraction* in enhancing retrieval effectiveness. Consequently, our method is particularly well-suited to real-world scenarios where complete electronic health records must be analyzed to provide medical advice. This indicates that our approach is promising for real-life applications in assisting physicians with clinical recommendations.

When analyzing performance across different datasets, we find that retrieval-based methods perform significantly better on MedQA-USMLE and EHRNoteQA, while MedMCQA showa 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. On EHRNoteQA, while



(a) Effect of Using Original Options. (b) Effect of RGAR's Two Components. (c) Effect of Rounds in RGAR.

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

RGAR's performance is slightly below that of i-MedRAG, **the latter's inference time is approximately 4 times longer, establishing RGAR as a more efficient and cost-effective alternative**.

4.2.2 Base LLMs with Different Sizes and Model Families

Table 3: Comparison of LLMs on MedQA-USMLE.

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

To further assess the versatility of RGAR, we conduct evaluations on MedQA-USMLE, 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 experimental 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.

4.3 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 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).

4.4 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 § 4.2.1, 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 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 B). The basic query shows

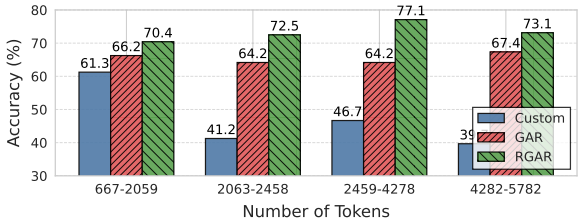


Figure 4: Fine-Grained Accuracy of EHRNoteQA After Sorting by Length and Dividing into Four Equal Parts.

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.

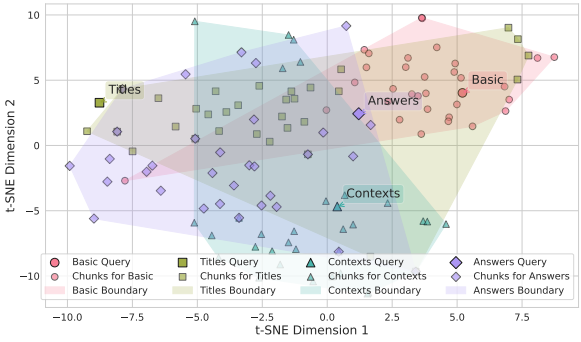


Figure 5: t-SNE Visualization of Different Queries and the Retrieved Chunks.

5 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-enhanced 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 Altenschmidt, 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*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

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

Donna M D'Alessandro, Clarence D Kreiter, and Michael W Peterson. 2004. An evaluation of information-seeking behaviors of general pediatricians. *Pediatrics*, 113(1):64–69.

Mary Forehand. 2010. Bloom's taxonomy. *Emerging perspectives on learning, teaching, and technology*, 41(4):47–56.

Jerome H Friedman, Jon Louis Bentley, and Raphael Ari Finkel. 1977. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software (TOMS)*, 3(3):209–226.

Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gianluca Moro, and Zaiqiao Meng. 2024. To generate or to retrieve? on the effectiveness of artificial contexts for medical open-domain question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9878–9919, Bangkok, Thailand. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. 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 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.

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

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

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

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

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Greg Kamradt. Llmtest_needleinahaystack: Evaluating long-context capabilities of large language models. Accessed: 2025-02-13.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 996–1009, Singapore. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. 2024. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 315–334, Bangkok, Thailand. Association for Computational Linguistics.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwang Hyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. 2024. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer.

Jesus Lovon-Melgarejo, Thouria Ben-Haddi, Jules Di Scala, Jose G. Moreno, and Lynda Tamine. 2024. Revisiting the MIMIC-IV benchmark: Experiments using language models for electronic health records. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 189–196, Torino, Italia. ELRA and ICCL.

Fengyu Lu, Jiaxin Duan, and Junfei Liu. 2023. A factual aware two-stage model for medical dialogue summarization. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2859–2866.

Kun Luo, Zheng Liu, Shitao Xiao, Tong Zhou, Yubo Chen, Jun Zhao, and Kang Liu. 2024. Landmark embedding: A chunking-free embedding method for retrieval augmented long-context large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3268–3281, Bangkok, Thailand. Association for Computational Linguistics.

Xin Lv, Yixin Cao, Lei Hou, Juanzi Li, Zhiyuan Liu, Yichi Zhang, and Zelin Dai. 2021. Is multi-hop reasoning really explainable? towards benchmarking reasoning interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8899–8911, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021a. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100, Online. Association for Computational Linguistics.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021b. Reader-guided passage reranking for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 344–350, Online. Association for Computational Linguistics.

Lynne M Markus. 2001. Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success. *Journal of management information systems*, 18(1):57–93.

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*.

Mahmoud Nazzal, Issa Khalil, Abdallah Khreishah, and NhatHai Phan. 2024. Promsec: Prompt optimization for secure generation of functional source code with large language models (llms). In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 2266–2280.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. 2021. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine Learning for Health*, pages 239–260. PMLR.

Hongjin Qian, Zheng Liu, Kelong Mao, Yujia Zhou, and Zhicheng Dou. 2024. Grounding language model with chunking-free in-context retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1311, Bangkok, Thailand. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ruiyang Ren, Yingqi Qu, Jing Liu, Xin Zhao, Qifei Wu, Yuchen Ding, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2023. A thorough examination on zero-shot dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15783–15796, Singapore. Association for Computational Linguistics.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Haitian Sun, Tania Bedrax-Weiss, and William Cohen. 2019. Pullnet: Open domain question answering with iterative retrieval on knowledge bases and text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2380–2390.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pages 601–618.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024a. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024b. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 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. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023b. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 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. *Advances in neural information processing systems*, 35:24824–24837.

Xiao Wei, Qi Xu, Hang Yu, Qian Liu, and Erik Cambria. 2024. Through the MUD: A multi-defendant charge prediction benchmark with linked crime elements. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 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. *Journal of the American Medical Informatics Association*, 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 Hardware Configuration

All experiments were conducted on an in-house workstation equipped with *dual* NVIDIA GeForce RTX 4090 GPUs, 128GB RAM, and an Intel® Core i9-13900K CPU.

Time cost across all methods on EHRNoteQA are shown in Table 4.

Table 4: Comparison of different methods in terms of execution time (hours).

Method	Custom	CoT	RAG	MedRAG	GAR	i-MedRAG	RGAR
Time (h)	0.5	0.5	1	1	2	22	6

A.2 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

B 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 5 shows the prompts template of RGAR and compared work (Using CoT ones). Table 6 shows the input of a sample, Table 7 shows the final output of RGAR.

C Framework Insight

C.1 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.

C.2 Why No Flexible Stopping Criteria

Similar multiround 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).

C.3 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 exploited, 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.

Type	Prompt Template
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 5: Prompt templates used in RGAR and Compared Methods.

Type	Texts
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/mm3 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	A. Coagulase-positive, gram-positive cocci forming mauve-colored colonies on methicillin-containing agar B. Encapsulated, gram-negative coccobacilli forming grey-colored colonies on charcoal blood agar C. Spore-forming, gram-positive bacilli forming yellow colonies on casein agar **D. Lactose-fermenting, gram-negative rods forming pink colonies on MacConkey agar**

Table 6: Case Study of a Sample Input.

Type	Texts
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	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 polysacarcharides. Option 2: Staphylococcus aureus Staphylococci can cause various infections but do not typically produce this type of capsule structure when exposed to heat. 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. 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.
Extract EHR	['Patient age: 39 years', 'Symptoms: fever, chills, left lower quadrant pain', 'Vital signs: **high** temperature (39.1°C), tachycardia (126/min), tachypnea (28/min), hypotension (80/50 mmHg)', 'Physical exam findings: mucopurulent discharge from cervix, left adnexal tenderness', 'Laboratory results: **low** platelet count (14,200/mm^3), **elevated** D-dimer (965 ng/mL)', 'Phenol test result: phosphorylated N-acetylglucosame dimmer with 6 fatty acids attached to a polysaccharide side chain']
Output Json	{"answer_choice": "D"}

Table 7: Case Study of output in RGAR.