Medical Graph RAG: Evidence-based Medical Large Language Model via Graph Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

We introduce MedGraphRAG, a novel graph-based Retrieval-Augmented Generation (RAG) framework designed to enhance LLMs in generating evidence-based medical responses, improving safety and reliability with private medical data. We introduce Triple Graph Construction and U-Retrieval to enhance GraphRAG, enabling holistic insights and evidence-based response generation for medical applications. Specifically, we connect user documents to credible medical sources and integrate Top-down Precise Retrieval with Bottom-up Response Refinement for balanced context awareness and precise indexing. Validated on 9 medical Q&A benchmarks, 2 health fact-checking datasets, and a long-form generation test set, MedGraphRAG outperforms state-of-the-art models while ensuring credible sourcing. Our code is publicly available.

1 Introduction

The rapid advancement of large language models (LLMs), such as OpenAI's GPT-4 (OpenAI, 2023a), has accelerated research in natural language processing and driven numerous AI applications. However, these models still face significant challenges in specialized fields like medicine (Hadi et al., 2024; Williams et al., 2024; Xie et al., 2024). The first challenge is that these domains rely on vast knowledge bases -principles and notions discovered and accumulated over thousands of years; fitting such knowledge into the finite context window of current LLMs is a hopeless task. Supervised Fine-Tuning (SFT) provides an alternative to using the context window, but it is often prohibitively expensive or unfeasible due to the closed-source nature of most commercial models. Second, medicine is a specialized field that relies on a precise terminology system and numerous established truths, such as specific disease symptoms or drug side effects. In this domain, it is essential that LLMs do not distort, modify, or introduce creative elements into the data. Unfortunately, verifying the accuracy of responses in medicine is particularly challenging for non-expert users. Therefore, the ability to perform complex reasoning using large external datasets, while generating accurate and credible responses backed by verifiable sources, is crucial in medical applications of LLMs.

Retrieval-augmented generation (RAG) (Lewis et al., 2021) is a technique that answers user queries using specific and private datasets without requiring further training of the model. However, RAG struggles to synthesize new insights and underperforms in tasks requiring a holistic understanding across extensive documents. GraphRAG (Hu et al., 2024) has been recently introduced to overcome these limitations. GraphRAG constructs a knowledge graph from raw documents using an LLM, and retrieves knowledge from the graph to enhance responses. By representing clear conceptual relationships across the data, it significantly outperforms classic RAG, especially for complex reasoning (Hu et al., 2024). However, its graph construction lacks a specific design to ensure response authentication and credibility, and its hierarchical community construction process is costly, as it is designed to handle various cases for general-purpose use. We find that specific effort is required to apply it effectively in the medical domain.

In this paper, we introduce a novel graph-based RAG method for medical domain, which we refer to as Medical GraphRAG (MedGraphRAG). This technique enhances LLM performance in the medical domain by generating evidence-based responses and official medical term explanation, which not only increases their credibility but also significantly improves their overall quality. Our method builds on GraphRAG with a more sophisticated graph construction technique, called Triple Graph Construction, to generate evidence-based responses, and an efficient retrieval method, U-Retrieval, which improves response quality with few costs. In Triple Graph Construction, we design a mechanism to link user RAG data to credible medical papers and foundational medical dictionaries. This process generates triples *[RAG data, source, definition]* to construct a comprehensive graph of user documents. It enhances LLM reasoning and ensures responses are traceable to their sources and definitions, guaranteeing reliability and explainability. We also developed a unique U-Retrieval strategy to respond to user queries. Instead of building costly graph communities, we streamline the process by summarizing each graph using predefined medical tags, then iteratively clustering similar graphs to form a multi-layer hierarchical tag structure, from broad to detailed tags. The LLM generates tags for the user query and indexes the most relevant graph based on tag similarity in a top-down approach, using it to formulate the initial response. Then it refines the response by progressively integrating back the higher-level tags in a bottom-up manner until the final answer is generated. This U-Retrieval technique strikes a balance between global context awareness and the retrieval efficiency.

To evaluate our MedGraphRAG method, we implemented it on several popular open-source and commercial LLMs, including GPT (OpenAI, 2023b), Gemini(Team et al., 2023) and LLaMA (Touvron et al., 2023). The results evaluated across 9 medical Q&A benchmarks show that MedGraphRAG yields materially better results than classic RAG and GraphRAG. Our final results even surpass many specifically trained LLMs on medical corpora, setting a new state-of-the-art (SOTA) across all benchmarks. To verify its evidence-based response capability, we quantitatively tested MedGraphRAG on 2 health fact-checking benchmarks and conducted a human evaluation by experienced clinicians. Both evaluations strongly support that our responses are more source-based and reliable than previous methods.

Our contributions are as follows:

1. We are the first to propose a specialized framework for introducing graph-based RAG in the medical domain, which we named MedGraphRAG.

2. We have developed unique Triple Graph Construction and U-Retrieval methods that enable LLMs to efficiently generate evidence-based responses utilizing holistic RAG data.

3. MedGraphRAG outperforms other retrieval methods and extensively fine-tuned Medical LLMs across a wide range of medical Q&A benchmarks, establishing the new SOTAs.

4. Validated by human evaluations, MedGraphRAG is able to generate more understandable and evidence-based responses in the medical domain.

2 Method

The overall workflow of MedGraphRAG is shown in Fig. 1. We first construct the knowledge graphs from the documents by using Triple Graph Construction (Section 2.1), then tag the graphs for U-Retrieval to response the user queries (Section 2.2). We illustrate the main idea here, with detailed implementation and prompts provided in the appendix.

2.1 Triple Graph Construction

2.1.1 Preliminary: Document Chunking & Entities Extraction

Large medical documents often contain diverse content. We segment them into chunks respecting LLMs' context limits. We adopt the semantic chunking function implemented in LangChain to chunk the documents(langchain, 2024). Specifically, we isolate paragraphs P_i within the document $D = \{P_1, P_2, \ldots, P_{N_p}\}$ using a text embedding model. We then set a buffer size of 5 and enforce the token limit according to the graph construction LLM $\mathcal{L}^G$.

We then extract entities from each chunk through graph construction LLM $\mathcal{L}^G$. We prompt $\mathcal{L}^G$ to identify all relevant entities $E = \{e_1, e_2, \ldots, e_{N_e^1}\}$ in each chunk and generate a structured output with *name*, *type*, and *a description of the context*: $e = \{na, ty, cx\}$, as the examples shown in the Step2 in Fig. 1. We set *name* be the text from the document, *type* selected from the UMLS semantic types (Bodenreider, 2004), and *context* a few sentences generated by $\mathcal{L}^G$ contextualized within the document.

2.1.2 Triple Linking

Medicine relies on precise terminology and established facts, making it essential for LLMs to produce responses grounded in established facts. To achieve this, we introduced Triple Graph Construction, linking user documents to credible sources and professional definitions. Specifically, we build repository graph (RepoGraph), which is intended to be fixed across different users, providing established sources and controlled vocabulary definitions for user RAG documents. We construct Re-

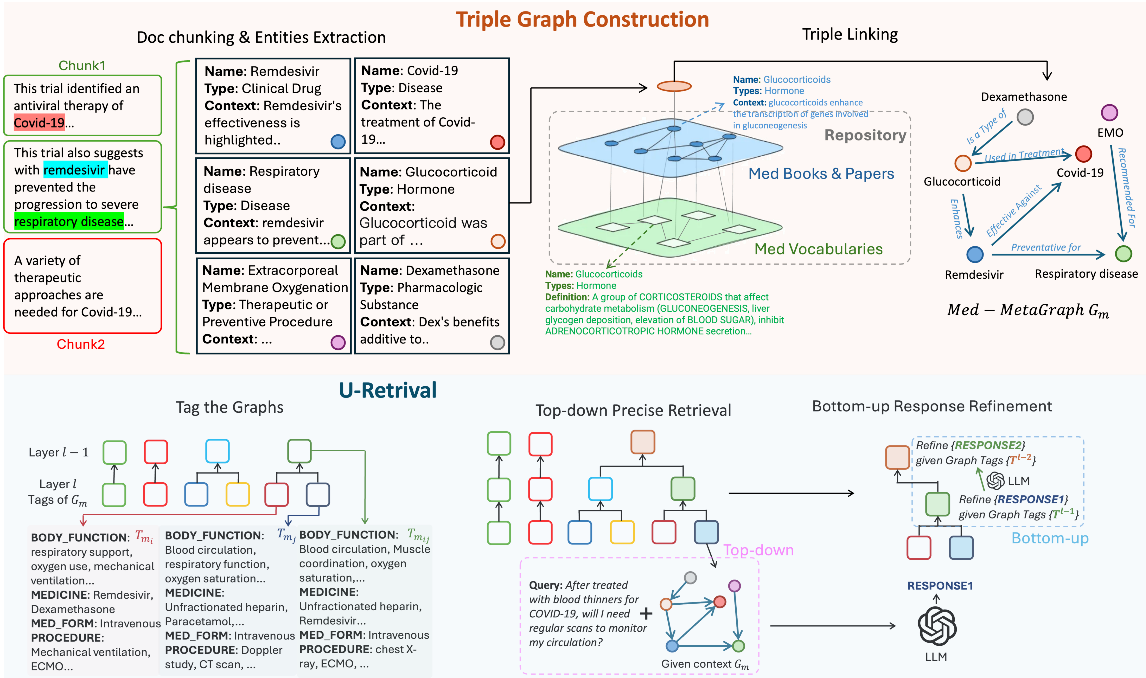


Figure 1: The overall workflow of MedGraphRAG begins with Triple Graph Construction, where documents are chunked, and entities are extracted. Triple linking then connects user entities to referenced papers and vocabulary graph layers, forming the Med-MetaGraph. In the subsequent U-Retrieval phase, graphs are tagged to enable top-down precise retrieval and bottom-up response refinement, ensuring graph-enhanced query responses.

poGraph under user RAG graph with two layers: one based on medical papers/books and another based on medical dictionaries. We build the bottom layer of RepoGraph as UMLS (Bodenreider, 2004) graph, which consist comprehensive, well-defined medical vocabularies and their relationships. The upper layer of RepoGraph is constructed from medical textbooks and scholarly articles using the same graph construction method described here.

The entities of all three tiers of graphs are hierarchically linked through semantic relationships. Let us denoted entities extracted from RAG documents as E^1. We link them to entities extracted from medical books/papers, denoted as E^2, based on their relevance, which is determined by computing the cosine similarity between their content embeddings $\phi(C_e)$. The content of an entity C_e is the concatenation of its *name*, *type*, and *context*, represented as: C_e = Text[name: na; type: ty; context: cx]. This directed linking is annotated as *the reference of*, indicating the reference relationship between entities in the two layers: $\mathrm{R}^{e^1}_{e^2} = \left\{ (e^1_i, TheReferenceOf, e^2_j) \,\middle|\, \frac{\phi(C_{e^1_i})\cdot\phi(C_{e^2_j})}{\|\phi(C_{e^1_i})\|\|\phi(C_{e^2_j})\|} \geq \delta_r \right\}$, where δ_r is the pre-defined threshold. Entities $e^2 \in E^2$ are linked to $e^3 \in E^3$ through the same way with relationships annotated as *the definition of* . Thus, RAG entities are constructed as triples *[RAG entity, source, definition]*.

We then instruct $\mathcal{L}^G$ to identify the relationships among RAG entities in each chunk, which we noted as $e^1 \in E_m$. This relationship is a concise phrase generated by $\mathcal{L}^G$ based on the content of the entity C_{e^1} and associated references $\{C_{e^2} | R^{e1}_{e2} =$ the reference of$\}$. The identified relationships specify the source and target entities, provide a description of their relationship: $R^{e^1_j}_{e^1_i} = \left\{ (e^1_i, r_{ij}, e^1_j) \,\middle|\, r_{ij} = \mathcal{L}^G_{rel}(C_{e^1_i}; C_{e^2_i}, C_{e^1_j}; C_{e^2_j}) \right\}$, where $\mathcal{L}^G_{rel}$ is $\mathcal{L}^G$ with relationship identification and generation prompt. We show an example of relationship linking in the Step4 of Fig. 1. After performing this analysis, we have generated a directed graph for each data chunk, which is referred to as Meta-MedGraphs $G_m = \{E_m, R(E_m)\}$.

2.2 U-Retrieval

2.2.1 Preliminary: Graph Tagging

Organizing and summarizing the graphs in advance is intuitive and has proven to facilitate efficient retrieval (Hu et al., 2024). However, unlike GraphRAG, we avoid constructing costly graph communities. We observe that, unlike general language content, medical text is often structured and can be summarized effectively using predefined tags. Motivated by this, we simply summarize each Meta-MedGraph G_m with several predefined tags T, and iteratively generate more abstract tag summaries for clusters of closely-related graphs. Specifically, LLM $\mathcal{L}^G$ first summarises the content of each Meta-MedGraph $\{C_e \mid e \in G_m\}$ given a set of given tags T. The tags T consist of multiple medical categories following Society for Testing and Materials (ASTM) standards for content of electronic health records, mainly including *Symptoms*, *Patient History*, *Body Functions*, and *Medication* (detailed format shown in appendix). This process generates a structured tag-summary for each G_m, denoted as T_m.

We then apply a variant agglomerative hierarchical clustering method with dynamic thresholding based on tag similarity, to group the graphs and generate synthesized tag summaries. Initially, each graph begins as its own group. At each iteration, we compute the tag similarity between all pairs of clusters and dynamically set the threshold δ_t to merge the top 20% most similar pairs. The graphs will be merged if all pairwise similarities within the group exceed δ_t. Note that we don't really link the nodes across different graphs, but generate a synthesized tag-summary for each group. Specifically, we calculate the similarity of pairs by measuring the average cosine similarity of all their tag embeddings. Let $\phi(t)$ denote the embedding of a tag $t \in T_m$. Taking two Meta-MedGraphs G_{m_i} and G_{m_j} with tag sets T_{m_i} and T_{m_j} as an example, we generate the abstract tag summery $T_{m_{ij}}$ if their cosine similarity of tag embeddings $\phi(t)$ and $\phi(t')$ higher than the threshold δ_t

$$T_{m_{ij}} = \mathcal{L}^G(T_{m_i}, T_{m_j}), \quad \text{if}$$
$$\frac{1}{|T_{m_i}| \cdot |T_{m_j}|} \sum_{t \in T_{m_i}} \sum_{t' \in T_{m_j}} \frac{\phi(t)^\top \phi(t')}{\|\phi(t)\| \, \|\phi(t')\|} \geq \delta_t;$$

These newly merged tag-summary, along with those that remain unmerged, form a new layer of tags. As tag-summaries become less detailed at higher layers, there is a trade-off between precision and efficiency. In practice, we limit the process to 12 layers, as this is sufficient for most model variants (detailed in Fig. 5).

2.2.2 Top-down Precise Retrieval

After constructing the graph, we use response LLM L^R efficiently retrieves information to respond to user queries. We begin by generating tag-summary on the user query $T_Q = \mathcal{L}^R(Q)$, and use these to identify the most relevant graph through a Top-down Precise Retrieval. Let's indicate the jth tags at layer i summarised tag T^i as $T^i[j]$, it starts from the top layer: T^0, progressively indexing down by selecting the most similar tag in each layer:

$$T^{i+1} = \underset{T^i[j] \in T^i}{\arg\max}\, sim(T_Q, T^i[j])$$

until we reach the tag for the target Meta-MedGraph G_{m_t}. We then retrieve Top N_u entities based on the embedding similarity between the query and the entity content: $E_r = \{e \mid \text{Top}N_u(sim(\phi(Q), \phi(C_e))), e \in M_t\}$, and gather all their Top k_u nearest triple neighbours $Tri^{\leq k_u}(e)$ as $E_r^{k_u} = \{e, Tri^{\leq k_u}(e), \mid e \in E_r\}$.

2.2.3 Bottom-up Response Refinement

By using all these entities and their relationships $G_r = \{E_r^{k_u}, R(E_r^{k_u})\}$, we prompt $\mathcal{L}^R$ to answer the question given the concatenated entity *names* and *relationships* in G_r: *Given QUESTION:* $\{Q\}$. *GRAPH:* $\{e_i[na]$+$R_{e_i}^{ej}$+$e_j[na]$, ...$\}$. *Answer the user question: QUESTION using the graph: GRAPH...* as $\mathcal{L}^R_{G_r}$.

In the Bottom-up Response Refinement step, we then move back to the higher-level tag retrieved in the previous step T^{i-1}, in a bottom-up manner. We provide $\mathcal{L}^R$ *QUESTION:* $\{Q\}$, *LAST RESPONSE:* ..., and *SUMMARY:* $\{T^{i-1}\}$, and ask it to *Adjust the response:RESPONSE of the question: QUESTION using the updated information: SUMMARY*. $\mathcal{L}^R$ continues refining its responses until it reaches the target layer. In practice, we retrieve 4-6 layers depends on the baseline LLM, a detailed experiment about it is shown in Fig. 5. It ultimately generate a final response after scanning all indexed graphs along the trajectory. This method enables the LLM to gain a comprehensive overview by interacting with all relevant data in the graph, while remaining efficient by accessing less relevant data in summarized form.

3 Experiment

3.1 Dataset

3.1.1 RAG data

We anticipate that users will use frequently-updated private data as RAG data, such as patient electronic medical records. Thus, we employ MIMIC-IV (Johnson et al., 2023), a publicly available electronic health record dataset, as RAG data.

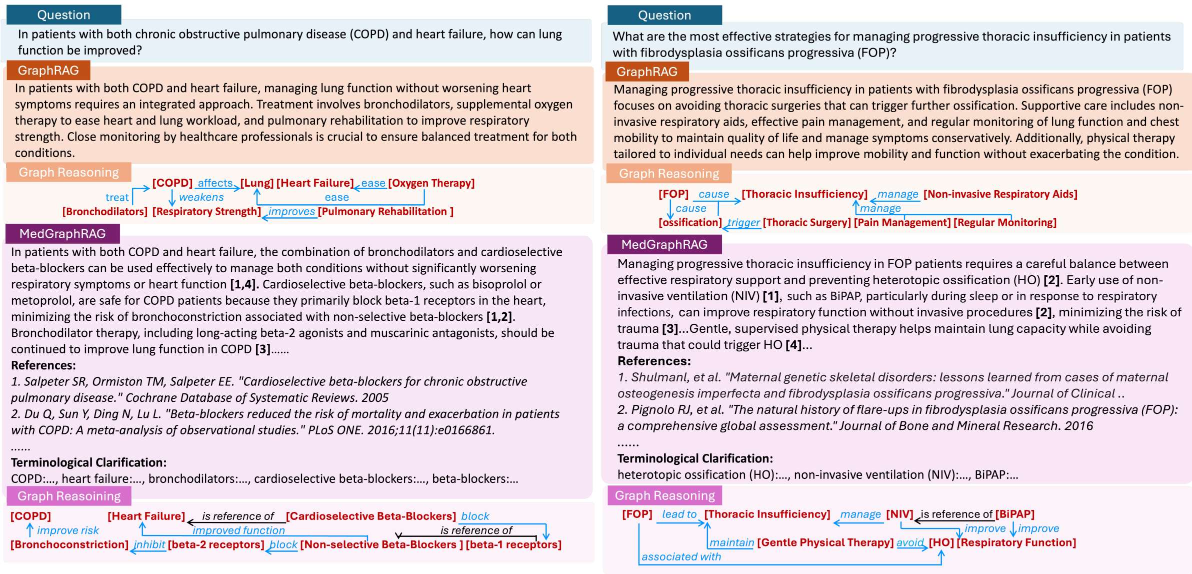


Figure 2: Example responses from GraphRAG and MedGraphRAG, with abstracted graphs. MedGraphRAG provides more detailed explanations and more complex reasoning with evidences. Full results are in the appendix.

3.1.2 Repository data

We provide repository data to support LLM responses with credible sources and authoritative vocabulary definitions. We use MedC-K (Wu et al., 2023), a corpus containing 4.8 million biomedical academic papers and 30,000 textbooks, along with all evidence publications from FakeHealth (Dai et al., 2020) and PubHealth (Kotonya and Toni, 2020), as the upper repository data, and UMLS graph, which includes authoritative medical vocabularis and semantic relationships as the bottom repository data.

3.1.3 Test Data

Our test set are the test split of 9 multiple-choice biomedical datasets from the MultiMedQA suite, 2 fact verification datasets about public health, i.e., FakeHealth (Dai et al., 2020) and PubHealth (Kotonya and Toni, 2020), and 1 test set we collected, called DiverseHealth. MultiMedQA includes MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022) PubMedQA (Jin et al., 2019) and MMLU clinic topics (Hendrycks et al., 2020). We also collected the DiverseHealth test set, focused on health equity, consisting of 50 real-world clinical questions that cover a wide range of topics, including rare diseases, minority health, comorbidities, drug use, alcohol, COVID-19, obesity, suicide, and chronic disease management. The dataset will be released alongside the paper. Details on dataset usage can be found in the appendix.

3.2 Experiment Setting

We compare different RAG methods across 6 language models as $\mathcal{L}^R$: Llama2 (13B, 70B), Llama3 (8B, 70B), Gemini-pro, and GPT-4. The Llama models were obtained from their official HuggingFace page. We used *gemini-1.0-pro* for Gemini-pro and *gpt-4-0613* for GPT-4. We primarily compare our approach with standard RAG implemented by LangChain(langchain, 2024) and GraphRAG (Edge et al., 2024a) implemented by Microsoft Azure (microsoft, 2024). All retrieval methods are compared under same RAG data and test data.

We deploy $\mathcal{L}^G$ as *Llama3-70B* to construct the graph. For text embeddings, we utilize OpenAI's *text-embedding-3-large* model. Model comparison is performed using a 5-shot response ensemble (Li et al., 2024). MedGraphRAG used U-Retrieval with 4 levels on GPT-4, and 5 levels for the others. In the retrieval, we picked top 60 entities with their 16-hop neighbors. Unless otherwise noted, all thresholds are set as 0.5. We use the same query prompt for all models to generate responses. Prompts are shown in the appendix.

3.3 Results

3.3.1 Multi-Choice Evaluation

Baselines with different retrievals First, we conducted experiments to evaluate retrieval methods on various LLM baselines, with the results shown in Table 1. We compared MedGraphRAG against baselines without retrieval, standard RAG, and

GraphRAG. Performance is measured by the accuracy of selecting the correct option. The results show that MedGraphRAG significantly enhances LLM performance on both health fact-checking and medical Q&A benchmarks. Compared to baselines without retrieval, MedGraphRAG achieves an average improvement of nearly 10% in fact-checking and 8% in medical Q&A. When compared to baselines using GraphRAG, it demonstrates an average improvement of around 8% in fact-checking and 5% in medical Q&A Notably, MedGraphRAG yields more pronounced improvements in smaller LLMs, such as $Llama2_{13B}$ and $Llama2_{8B}$. This suggests that MedGraphRAG effectively utilizes the models' own reasoning capabilities while providing them with additional knowledge beyond their parameters, serving as an external memory for information.

Comparing with SOTA Medical LLMs When applied MedGraphRAG to larger models, like $Llama_{70B}$ or GPT, it resulted in new SOTA across all 11 datasets. This result also outperforms intensively fine-tuning based medical large language models like Med-PaLM 2 (Singhal et al., 2023b) and Med-Gemini (Saab et al., 2024), establishing a new SOTA on the medical LLM leaderboard. A detailed comparison is shown in Fig. 6.

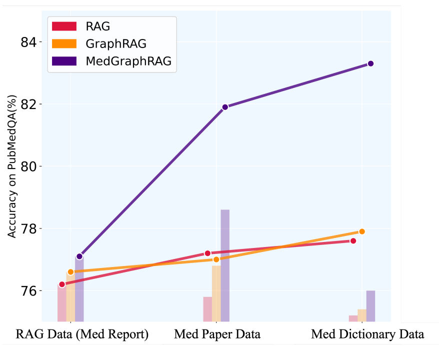


Figure 3: Impact of Repository Data on RAG, GraphRAG, and MedGraphRAG with GPT-4. Line chart: performance with incremental data inclusion; bar chart: performance with individual data inclusion.

3.3.2 Long-form Generation Evaluation

Human Evaluation We conducted human evaluations of long-form model generation on the MultiMedQA and DiverseHealth benchmarks, comparing our method to SOTA models that generate citation-backed responses, including Inline Search in (Gao et al., 2023b), ATTR-FIRST (Slobodkin et al., 2024), and MIRAGE (Qi et al., 2024). Our evaluation panel consisted of 7 certified clinicians and 5 laypersons to ensure feedback from both professional and general users. Raters completed a five-level rating survey for each model's response, assessing responses across five dimensions: *pertinence* (Pert.), *correctness* (Cor.), *citation precision* (CP), *citation recall* (CR), and *understandability* (Und.). Detailed background information on the raters and the survey questions can be found in the appendix. As shown in Table 2, MedGraphRAG consistently received higher ratings across all metrics. Notably, it showed a significant advantage in CP, CR and Und., indicating that its responses were more often backed by accurate sources and were easier to understand, even for laypersons, thanks to evidence-backed responses and clear explanations of complex medical terms.

Case Study As illustrated in Fig. 7, we compare the responses from GraphRAG and MedGraphRAG for a complex case involving patients with both chronic obstructive pulmonary disease (COPD) and heart failure (left plot). GraphRAG suggested general COPD treatments like bronchodilators and pulmonary rehabilitation but overlooked that certain bronchodilators may worsen heart failure symptoms. In contrast, MedGraphRAG provided a more comprehensive answer by recommending cardioselective beta-blockers—such as bisoprolol or metoprolol—that safely manage both conditions without adverse effects. As we can see from the graph abstracted, this superiority stems from MedGraphRAG's architecture, where entities are directly linked to key information in references, allowing retrieval of specific evidence. Conversely, GraphRAG struggles to retrieve specific information since its reference and user data are intertwined within the same layer of the graph, which leads to missing key information under the same number of nearest neighbors. And its retrieval based purely on graph summaries results in a lack of detailed insights.

3.4 Ablation and Analysis

3.4.1 Overall Ablation Study

We conducted a comprehensive ablation study to validate the effectiveness of our proposed modules, with the results presented in Table 3. Starting with GraphRAG (Hu et al., 2024) as the baseline, we incrementally incorporated our unique components, including Triple Graph Construction, and U-Retrieval. Notably, both experiments were con-

Table 1: Accuracy(%) of LLMs using different retrieval methods. Columns with a blue background represent health fact-checking benchmarks, while the others correspond to medical Q&A benchmarks. The best results are highlighted in bold.

Model	Fake Health	Pub Health	MedQA	Med MCQA	Pub MedQA	MMLU Col-Med	MMLU Col-Bio	MMLU Pro-Med	MMLU Anatomy	MMLU Gene	MMLU Clinic
Baselines without retrieval											
Llama2-13B	53.8	49.4	42.7	37.4	68.0	60.7	69.4	60.3	52.6	66.0	63.8
Llama2-70B	58.9	56.7	43.7	35.0	74.3	64.2	84.7	75.0	62.3	74.0	71.7
Llama3-8B	51.1	53.2	59.8	57.3	75.2	61.9	78.5	70.2	68.9	83.0	74.7
Llama3-70B	64.2	61.0	72.1	65.5	77.5	72.3	92.5	86.7	72.5	83.9	82.7
Gemini-pro	60.6	63.7	59.0	54.8	69.8	69.2	88.0	77.7	66.7	75.8	76.7
GPT-4	71.4	70.9	78.2	72.6	75.3	76.7	95.3	93.8	81.3	90.4	86.2
Baselines with RAG											
Llama2-13B	56.2	54.3	48.1	42.0	68.6	62.5	68.3	63.7	51.0	64.5	67.4
Llama2-70B	64.6	63.2	56.2	49.8	75.2	69.6	85.8	77.4	63.0	75.8	73.3
Llama3-8B	60.5	59.6	64.3	58.2	76.0	68.6	84.9	73.2	72.1	85.2	77.8
Llama3-70B	76.2	72.1	82.3	72.5	80.6	86.8	94.4	89.7	84.3	87.1	87.6
Gemini-pro	72.5	68.4	64.5	57.3	76.9	79.0	91.3	86.4	79.5	80.4	83.9
GPT-4	78.6	77.3	88.1	76.3	77.6	81.2	95.5	94.3	83.1	92.9	93.1
Baselines with GraphRAG											
Llama2-13B	58.7	57.5	52.3	44.6	72.8	64.1	73.0	64.6	52.1	66.2	67.9
Llama2-70B	65.7	63.8	55.1	52.4	74.6	68.0	86.4	79.2	64.6	73.9	75.8
Llama3-8B	61.7	61.0	64.8	58.7	76.6	69.2	84.3	73.9	72.8	85.5	77.4
Llama3-70B	77.7	74.5	84.1	73.2	81.2	87.4	94.8	89.8	85.2	87.9	88.5
Gemini-pro	73.8	70.6	65.1	59.1	75.2	79.8	90.8	85.8	80.7	81.5	84.7
GPT-4	78.4	77.8	88.9	77.2	77.9	82.1	95.1	94.8	82.6	92.5	94.0
Baselines with MedGraphRAG											
Llama2-13B	64.1	61.2	65.5	51.4	73.2	68.4	76.5	67.2	56.0	67.3	69.5
Llama2-70B	69.3	68.6	69.2	58.7	76.0	73.3	88.6	84.5	68.9	76.0	77.3
Llama3-8B	79.9	77.6	74.2	61.6	77.8	89.2	95.4	91.6	85.9	89.3	89.7
Llama3-70B	81.2	79.2	88.4	79.1	**83.8**	91.4	96.5	93.2	89.8	91.0	94.1
Gemini-pro	79.2	76.4	71.8	62.0	76.2	86.3	92.9	89.7	85.0	87.1	89.3
GPT-4	**86.5**	**83.4**	**91.3**	**81.5**	83.3	**91.5**	**98.1**	**95.8**	**93.2**	**98.5**	**96.4**

Table 2: Human evaluation on MedQA and Diverse-Health samples.

Data	Methods	Pert.	Cor.	CP	CR	Und.
MultiMedQA	INLINE	91	88	80	74	85
	ATTR.FIRST	93	91	86	77	93
	MIRAGE	95	90	84	75	91
	MedGrapgRAG	**97**	**94**	**92**	**86**	**95**
Diverse Health	INLINE	95	84	78	71	81
	ATTR.FIRST	96	91	81	78	85
	MIRAGE	97	89	83	76	87
	MedGrapgRAG	**97**	**96**	**89**	**84**	**93**

Table 3: An ablation study of MedGraphRAG, starting from GraphRAG, evaluated using accuracy (%) on Q&A datasets.

	MedQA	PubMedQA	MedMCQA
GraphRAG	88.9	77.9	77.2
+Triple Graph Construction	91.1	81.8	80.9
+U-Retrieval	**91.3**	**83.3**	**81.5**

ducted on the same RAG dataset, eliminating data-related improvements. The results show a gradual performance improvement as more of our modules are added, with significant gains observed when replacing GraphRAG graph construction with our Triple Graph Construction. Additionally, by replacing the summary-based retrieval(Edge et al., 2024b) in GraphRAG with our U-Retrieval method, we achieved further improvements, setting new state-of-the-art results across all three benchmarks.

3.4.2 Detailed Ablation on Triple Linking

To assess the individual effects of external RAG data and retrieval technologies, we conducted experiments comparing retrieval methods: RAG, GraphRAG, and MedGraphRAG under two settings: (1) retrieving each tier of data separately (bar chart in Fig. 3), and (2) incrementally adding all three tiers (line chart in Fig. 3). The results show that both the data and the right retrieval method must work together to unlock the full potential. When retrieving data by standard RAG, Med-Paper data individually improves performance by less than 2%, and Med-Dictionary data by less than 1%. Accumulating three tier data also leads to mediocre improvements. GraphRAG shows improvement in retrieving individual data but has minimal gains when incrementally adding more data, likely due to superficiality from linking trivial entities, as discussed in the previous case study. In contrast, MedGraphRAG efficiently handles the additional data, using its hierarchical structure to clarify relationships and show strong improvements as more data is added. With MedGraphRAG, we see significant improvements of over 6% and 8% for Med-Paper and Med-Dictionary data, respectively, highlighting the importance of the retrieval method in maximizing the impact of the data.

3.5 Detailed Ablation on U-Retrieval

In U-Retrieval, we set the retrieval depth to 4-5 layers, the number of retrieval entities to 60, and entity neighbors to 16. These settings were determined through comprehensive trials. First, we examine the impact of the retrieval range, i.e. the number of entities and neighbors, using GPT-4 with MedGraphRAG on MedQA, as shown in Fig. 4. Our findings show that retrieving more data does not necessarily lead to better performance. In fact, more data can introduce noise and exacerbate LLM performance issues with long contexts. The peak performance occurs when the retrieval size reaches approximately 120 entities with 4-hop neighbors or 60 entities with 16-hop neighbors. The 16-hop neighbors setting performed slightly better, likely due to the robustness of graph-based linking compared to vector-similarity-based retrieval.

As previously mentioned, there is also a trade-off between model accuracy and response time with retrieval layer increases. This relationship is explored in Fig. 5. The figure compares the cost time and MedQA accuracy across retrieval depths from 0 to 9 layers. We observe that both performance and response time increase as the retrieval layer increases initially. However, performance begins to degrade when retrieving more layers, as higher layers often contain less relevant information, which can interfere with refining the response. The optimal retrieval depth is 4 layers for the GPT-4 model and 5 layers for others, which we use as the default setting in our experiments.

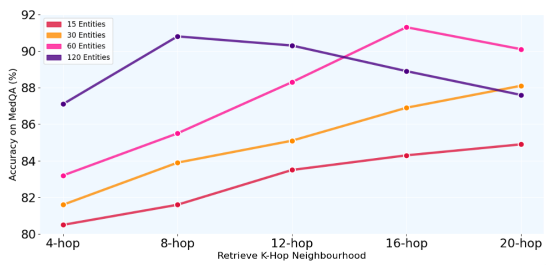


Figure 4: The effect of retrieving different number of entities and neighbourhoods. Performance evaluated by GPT-4 (MedGraphRAG) on MedQA.

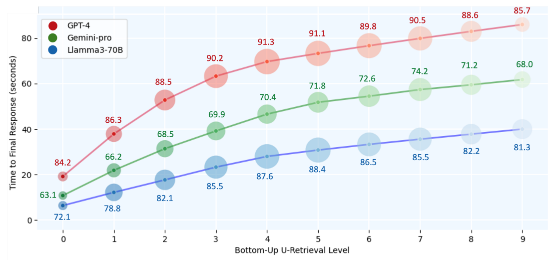


Figure 5: The relationship between U-retrieval level and time cost.

4 Related Work

Large language models (LLMs) built on Transformer architectures have advanced rapidly, leading to specialized medical LLMs such as BioGPT (Luo et al., 2022), PMC-LLaMA (Wu et al., 2023), BioMedLM (Bolton et al., 2022), and Med-PaLM 2 (Singhal et al., 2023b). While many are fine-tuned by large organizations, recent research has focused on cost-efficient, non-fine-tuned approaches, primarily using prompt engineering (Saab et al., 2024; Wang et al., 2023; Savage et al., 2024). RAG, as another non-finetuning approach, is rarely explored for medical applications (Miao et al., 2024; Xiong et al., 2024; Long et al., 2024) and lacks support for evidence-based responses and term explanations required in clinical settings.

RAG (Lewis et al., 2021) enables models to use specific datasets without additional training, improving response accuracy and reducing hallucinations (Guu et al., 2020). RAG has shown strong results across various tasks, including generating responses with citations (Gao et al., 2023b; Slobodkin et al., 2024; Qi et al., 2024; Nakano et al., 2021; Bohnet et al., 2022; Gao et al., 2023a,c; Schimanski et al., 2024; Zhang et al., 2024). GraphRAG (Hu et al., 2024) further enhances complex reasoning by constructing knowledge graphs, but lacks specific design features for generating attributed responses, and its application in medical specialization remains limited.

5 Conclusion

MedGraphRAG improves the reliability of medical response generation with its graph-based RAG framework, using Triple Graph Construction and U-Retrieval to enhance evidence-based, context-aware responses. Future work will focus on real-time data updates and validation on real-world clinical data.

6 Limitation

Despite the strong capabilities demonstrated by MedGraphRAG, the graph construction step incurs significant computational costs. In the retrieval and response stage, although the costs are lower than graph construction, they remain higher than standard large language model (LLM) calls, with each question taking around 70 seconds to process (see Figure 6 for details). Future efforts should explore methods to transfer pre-constructed graphs or accelerate the graph construction process to mitigate these computational costs.

Additionally, the scale of experimental data and the expensive nature of graph construction make it challenging to conduct comprehensive comparisons of hyper-parameter settings and technology choices. For instance, factors such as the number of paragraphs in the context window during document chunking, the use of alternative RAG datasets, and the impact of different prompts for graph construction were selected empirically based on limited data. A more rigorous and comprehensive comparison of these factors is needed in future work to identify the optimal configurations that maximize the method's potential.

Finally, regarding human evaluation, while we made efforts to ensure diversity and expertise among our raters (see Appendix for details), the evaluation may still carry biases due to the limited sample size (120 questions on MultiMedQA and 50 questions on DiverseHealth). Future research should include larger-scale and better-designed human evaluations to thoroughly assess the model's performance.

References

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

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models.

Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. 2022. Biomedlm. Stanford Center for Research on Foundation Models.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2023. Dense X Retrieval: What Retrieval Granularity Should We Use? *arXiv preprint*. ArXiv:2312.06648 [cs].

Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. *arXiv preprint arXiv:2002.00837*.

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

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024b. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint*. ArXiv:2404.16130 [cs].

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and Revising What Language Models Say, Using Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023c. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23. ArXiv:2007.15779 [cs].

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. *arXiv preprint*. ArXiv:2002.08909 [cs].

Ali Hadi, Edward Tran, Branavan Nagarajan, and Amrit Kirpalani. 2024. Evaluation of chatgpt as a diagnostic tool for medical learners and clinicians. *Plos one*, 19(7):e0307383.

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

Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2024. GRAG: Graph Retrieval-Augmented Generation. *arXiv preprint*. ArXiv:2405.16506 [cs].

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

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.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*.

langchain. 2024. Enhancing rag-based application accuracy by constructing and leveraging knowledge graphs. *https://blog.langchain.dev/enhancing-rag-based-applications-accuracy-by-constructing-and-leveraging-knowledge-graphs/*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint*. ArXiv:2005.11401 [cs].

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024. More agents is all you need. *Preprint*, arXiv:2402.05120.

Cui Long, Yongbin Liu, Chunping Ouyang, and Ying Yu. 2024. Bailicai: A domain-optimized retrieval-augmented generation framework for medical applications. *arXiv preprint arXiv:2407.21055*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Oscar A Garcia Valencia, and Wisit Cheungpasitporn. 2024. Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications. *Medicina*, 60(3):445.

microsoft. 2024. Microsoft azure graphrag. *https://github.com/Azure-Samples/graphrag-accelerator?tab=readme-ov-file*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-assisted question-answering with human feedback.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. *arXiv preprint*. ArXiv:2311.16452 [cs].

OpenAI. 2023a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2023b. Openai. introducing chatgpt. `https://openai.com/blog/chatgpt/`.

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.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. Model internals-based answer attribution for trustworthy retrieval-augmented generation. *arXiv preprint arXiv:2406.13663*.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20.

Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. 2024. Towards Faithful and Robust LLM Specialists for Evidence-Based Question-Answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1931, Bangkok, Thailand. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023a. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation. *arXiv preprint arXiv:2403.17104*.

Eric E Smith and Andrew E Beaudin. 2018. New insights into cerebral small vessel disease and vascular cognitive impairment from mri. *Current opinion in neurology*, 31(1):36–43.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint*. ArXiv:2302.13971 [cs].

Jiaqi Wang, Enze Shi, Sigang Yu, Zihao Wu, Chong Ma, Haixing Dai, Qiushi Yang, Yanqing Kang, Jinru Wu, Huawen Hu, et al. 2023. Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*.

Christopher YK Williams, Brenda Y Miao, Aaron E Kornblith, and Atul J Butte. 2024. Evaluating the use of large language models to provide clinical recommendations in the emergency department. *Nature Communications*, 15(1):8236.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-LLaMA: Towards Building Open-source Language Models for Medicine. *arXiv preprint*. ArXiv:2304.14454 [cs].

Yunfei Xie, Juncheng Wu, Haoqin Tu, Siwei Yang, Bingchen Zhao, Yongshuo Zong, Qiao Jin, Cihang Xie, and Yuyin Zhou. 2024. A preliminary study of o1 in medicine: Are we closer to an ai doctor? *arXiv preprint arXiv:2409.15277*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. 2022a. Deep Bidirectional Language-Knowledge Graph Pretraining. *arXiv preprint*. ArXiv:2210.09338 [cs].

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022b. LinkBERT: Pretraining Language Models with Document Links. *arXiv preprint*. ArXiv:2203.15827 [cs].

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, et al. 2024. Longcite: Enabling llms to generate fine-grained citations in long-context qa. *arXiv e-prints*, pages arXiv–2409.

Contents

A Details of Dataset

A.1 MIMIC-IV

The Medical Information Mart for Intensive Care (MIMIC-IV (Johnson et al., 2023)) is governed by the PhysioNet Credentialed Health Data License 1.5.0, and access is restricted to credentialed users who sign the PhysioNet Credentialed Health Data Use Agreement 1.5.0 (DUA). The dataset is a comprehensive, deidentified dataset derived from patients admitted to the emergency department or an intensive care unit (ICU) at Beth Israel Deaconess Medical Center in Boston, MA. MIMIC-IV includes data on over 65,000 ICU patients and over 200,000 emergency department patients, totaling 364,627 unique individuals. The dataset reflects 546,028 hospital admissions and 94,458 unique ICU stays. The dataset was developed through a three-step process of acquisition, preparation, and de-identification, offering modular data organization to enable the seamless integration of different data sources for research purposes.

A.2 MedC-K

The MedC-K (Wu et al., 2023) corpus is a large, medical-specific dataset consisting of 4.8 million biomedical academic papers and 30,000 textbooks. It is designed to enhance foundational language models for medical applications. To adapt a general-purpose language model for the medical domain, a data-centric knowledge injection approach is employed, which focuses on introducing the model to medical-related terminologies and definitions. This process emphasizes two key data sources: biomedical papers and textbooks, ensuring that the model is exposed to both academic research and comprehensive medical knowledge.

A.3 FakeHealth

FakeHealth (Dai et al., 2020) dataset is licensed under the Creative Commons Attribution 4.0 International (DOI: 10.5281/zenodo.3606757), which permits redistribution and reuse of the work, provided that the original creator is appropriately credited. FakeHealth consists of two datasets: HealthStory and HealthRelease. Both datasets contain news content, news reviews, social engagements, and user networks. HealthRelease consists of 606 pieces of content and reviews, while HealthStory contains 1,690. We did not use the social engagement or user network data. The models were tested on rating questions from the 'criteria' section of the news reviews, using the news content as RAG data, or repository data specifically for MedGraphRAG. The data collection process involved four steps: (1) crawling reviews of news stories and releases from HealthNewsReview.org, (2) including the source news titles and URLs in the collected files, (3) scraping the corresponding news content, and (4) gathering social engagements on Twitter (tweets, replies, and retweets) related to the source news, supplemented by user network information.

A.4 PubHealth

The PUBHEALTH (Kotonya and Toni, 2020) dataset, released under the MIT license, is designed for explainable automated fact-checking of public health claims. It consists of 11,832 claims across various health-related topics, including biomedical subjects (such as infectious diseases and stem cell research), government healthcare policies (like abortion, mental health, and women's health), and other public health issues. Each claim is labeled with a veracity tag—true, false, unproven, or mixture—and accompanied by an explanation text that justifies the assigned label. These explanations, crafted by journalists, serve as gold standard judgments to support the fact-check labels. The claims were sourced from both fact-checking websites and news or news review websites, making PUBHEALTH a comprehensive resource for understanding the veracity of health-related claims.

A.5 MultiMedQA

The MultiMedQA dataset, released under the Apache License 2.0, serves as a comprehensive benchmark for open question answering in the medical domain, combining six existing datasets that span professional medical exams, research queries, and consumer health inquiries. We used its multiple-choice question answering part including MedQA, MedMCQA, PubMedQA, and MMLU clinical topics. We did not use the included LiveQA, MedicationQA, or HealthSearchQA datasets for model testing. All datasets are in English and cover a wide range of medical and health-related topics.

A.5.1 MedQA

The MedQA (Jin et al., 2021) dataset, released under the MIT license, contains questions in the style of the US Medical License Exam, with each question offering 4 or 5 possible answer choices. These questions were sourced from the National Medical Board Examination in the USA. The dataset is divided into a development set comprising 11,450 questions and a test set containing 1,273 questions. It follows a multiple-choice, open-domain format, where each question is accompanied by a set of possible answers. MedQA serves as a valuable resource for training and evaluating models in medical question answering.

A.5.2 MedMCQA

The MedMCQA dataset, released under the MIT license, is a large-scale multiple-choice question answering dataset designed to reflect real-world medical entrance exam questions. It contains over 194,000 high-quality MCQs from AIIMS and NEET PG entrance exams, covering 2,400 healthcare topics across 21 medical subjects. The dataset is notable for its high topical diversity and an average question length of 12.77 tokens. Available in English, the dataset is divided into a development set with 187,000 questions and a test set containing 6,100 questions. The format follows a typical Q+A structure with multiple-choice, open-domain questions, making it a valuable resource for medical education and AI research.

A.5.3 PubMedQA

The PubMedQA (Pal et al., 2022) dataset, released under the MIT license, is a specialized biomedical question answering (QA) dataset designed to answer research questions using abstracts from PubMed. The task involves providing a yes/no/maybe answer based on the corresponding abstract. PubMedQA consists of 1,000 expert-annotated instances, 61.2k unlabeled instances, and 211.3k artificially generated QA instances, making it the first QA dataset that requires reasoning over biomedical research texts, particularly their quantitative aspects, to answer questions. Unlike open-domain tasks like MedQA and MedMCQA, PubMedQA is a closed-domain task, where answers must be inferred from the supporting PubMed abstract context. The dataset is structured as Q+A+context in a multiple-choice format and is divided into a development set of 500 instances and a test set of 500 instances.

A.5.4 MMLU clinic topics

The MMLU clinic topics dataset (Hendrycks et al., 2020), released under the MIT license, follows a multiple-choice question answering format (Q + A) and is designed for open-domain tasks. It includes a development set of 29 questions and a test set containing 265 questions. This structure allows for the evaluation of models in an open-domain context where a variety of potential answers are considered, making it a valuable resource for research in question answering.

A.6 DiverseHealth

The DiverseHealth test set is a carefully curated dataset we collected to advance health equity by capturing a broad and inclusive range of clinical scenarios. Sourced from real users through a private telehealth application, the dataset consists of 50 real-world clinical questions. These questions, voluntarily provided by patients who signed explicit consent to share their inquiries, reflect genuine healthcare concerns, with all personal identification removed to ensure privacy and confidentiality. The telehealth platform, used predominantly by underserved and diverse populations, offers a unique opportunity to collect a wide spectrum of medical questions that highlight the challenges faced by individuals across various socioeconomic and demographic backgrounds.

Each question in the DiverseHealth test set addresses critical aspects of healthcare that are often underrepresented in traditional datasets. Topics include rare diseases that disproportionately affect minority populations, the complexities of managing multiple comorbidities, and issues related to drug and alcohol use. Mental health questions, such as those concerning suicide prevention, offer valuable insights into the mental health struggles experienced by various communities. The dataset also includes questions related to emerging public health crises like COVID-19, focusing on its disproportionate impact on marginalized groups, as well as long-standing health concerns such as obesity and chronic disease management.

The combination of real-world clinical questions and the diversity of topics ensures that the DiverseHealth test set serves as a powerful tool for developing AI-driven healthcare solutions. By focusing on the real concerns of diverse patient populations, this dataset is crucial for improving the inclusivity of AI models, ensuring they perform effectively across different demographic groups and healthcare environments. Ultimately, the DiverseHealth test set aims to address health disparities, helping researchers build AI systems that contribute to more equitable healthcare outcomes for all.

B Detailed Implementation

In the semantic document chunking process, we apply proposition transfer (Chen et al., 2023) to each paragraph before semantic validation to extract standalone statements that are self-contained and unambiguous (e.g., transforming "It prevents respiratory disease" to "Remdesivir prevents respiratory disease"). Through proposition transfer, each paragraph is transformed into independent, clear statements. For semantic validation, we utilize an LLM to first generate a short summary and a title for the current chunk. The LLM then determines if the current paragraph belongs to this chunk based on the title and summary. If it belongs, the LLM updates the title and summary accordingly. If not, the current chunk is finalized, and the LLM generates a title and summary for the new paragraph, treating it as the start of a new chunk. (the prompts of this process are all in the Sec. D.1). If the scan reaches the end of the document, the current chunk is automatically finalized to ensure no chunk spans across multiple documents.

In the entity extraction, we include unique IDs to trace their source document. In practice, for the user privacy data, we generate a universally unique identifier (UUID) for each document as their IDs. For the medical papers and books, we use their Digital object identifier (DOI) as their IDs, and for the medical dictionaries, we use their UMLS Concept Unique Identifiers (CUI) as their IDs. This identifier is crucial for retrieving information from the source, enabling the generation of evidence-based responses later. For tag-based summary generation and merging, we insert ten tags into the prompt at a time to iteratively generate the response.

All experiments except GraphRAG related are run on Google Cloud Platform with GCP a3-megagpu-8g Accelerator Optimized: 8 NVIDIA H100 GPU, 208 vCPUs, 1872GB RAM, 16 local SSD servers. GraphRAG related experiments are run on Microsoft Azure Standard-ND96isr-H100-v5 machines by Microsoft's default implementation. We process different documents in parallel for document chunking and then parallelize the construction of the Med-MetaGraph over each data chunk. All graphs are stored and organized in Neo4j. Detailed statistic information about the constructed Med Report Graph and Med Paper Graph is shown in Tab. 4. The UMLS graph is directly deployed in its existing natural graph structure.

Table 4: Statistic information of Med Report Graph (RAG Graph) and Med Paper Graph (top tier of repository graph), including the number of Med-MetaGraphs, average nodes in each Med-MetaGraph, and average degrees of each Med-MetaGraph.

Med Report Graph			Med Paper Graph		
MetaGraph Num.	Ave. Nodes	Ave. Degrees	MetaGraph Num.	Ave. Nodes	Ave. Degrees
2,978,245	72.7	3.1	41,321,588	91.2	3.7

For testing the models on MultiMedQA, we evaluate their zero-shot performance using only the test set of each dataset, without utilizing the training data for fine-tuning or including it in the RAG data for retrieval. For evaluating accuracy on FakeHealth, we incorporate its news content into the Medical-Papers-tier graph of MedGraphRAG and into RAG data of the others, then use the criteria questions from the news content to prompt the models to respond with 'Satisfactory' or 'Not Satisfactory.' For PubHealth, we integrate its news/reviews into Medical-Papers-tier graph of MedGraphRAG and into RAG data of the others, and prompt the models to classify each claim as 'True,' 'False,' 'Unproven,' or a 'Mixture.'

C Additional Results and Analysis

C.1 Compare to SOTA Medical LLM Models

We also evaluated MedGraphRAG against a range of previous SOTA medical large language models on these benchmarks, including both intensively fine-tuned models (Gu et al., 2022)(Yasunaga et al., 2022a)(Yasunaga et al., 2022b)(Bolton et al., 2022)(Singhal et al., 2022)(Singhal et al., 2023a)(Wu et al., 2023) and non-fine-tuned models (Nori et al., 2023)(OpenAI, 2023b)(OpenAI, 2023a)(Saab et al., 2024). The results, depicted in Fig. 6, show that when combined with GPT-4, our MedGraphRAG surpasses the previous SOTA model, Medprompt (Nori et al., 2023), by a notable 1.1% on the MedQA benchmark, and also outperforms it across all 9 datasets, establishing a new SOTA on the medical LLM leaderboard. It's important to note that while Medprompt retrieves training data with similar questions and correct answers as examples for prompting, our model operates with a simple prompt containing only the original question. This improvement further demonstrates MedGraphRAG's superior capability, even when retrieving from data with a different distribution. Furthermore, when compared to intensive fine-tuning methods on these medical datasets, MedGraphRAG outperforms strong models like Med-PaLM 2 (Singhal et al., 2023b) and Med-Gemini (Saab et al., 2024), establishing a new SOTA. This superior performance highlights MedGraphRAG's ability to efficiently leverage the inherent capabilities of LLMs and enhance their performance with additional data, without the need for fine-tuning.

C.2 Case study: GPT4 with and without MedGraphRAG

As shown in Fig. 7, we compare the responses generated by vanilla GPT-4 and MedGraphRAG for a misleading case where a patient presents with symptoms commonly associated with Alzheimer's but is actually Vascular Dementia. GPT-4 was misled, returning an incorrect diagnosis. In contrast, MedGraphRAG notes the details like that the MRI showed moderate vascular changes and white matter lesions, which are indicative of chronic ischemic damage—typical of vascular dementia rather than Alzheimer's, through retrieving the findings in (Smith and Beaudin, 2018), "CBF and WMH that...causing ical impairments,". With detailed definitions of medical terms and source knowledge retrieved to assist the reasoning process, MedGraphRAG chose the correct answer and provided a detailed, easily understandable explanation with citation, enabling users to verify the response.

C.3 Case study: Long-form generation of MedGraphRAG

We provided four examples of MedGraphRAG Long-form response generation. We include the diverse cases across Comorbidity Fig. 8, 9, Rare Disease Fig. 10,11, Minority Health Fig. 12,13, and Chornic Disease Managment Fig. 14,15. We can see the unique responses provided by MedGraphRAG combining citations with clear term explanations in medical responses ensures both credibility and understanding. Citations provide a foundation of evidence, reassuring patients and professionals that recommendations are grounded in research. For example, in the hormone replacement therapy answer, the association between HRT and increased risks of cardiovascular events and thromboembolic complications is backed by "Dhejne et al., 2011," which provides long-term follow-up data on health outcomes in transgender individuals undergoing hormone therapy. This level of transparency is particularly important in healthcare, where trust is critical for patient compliance and effective care.

Clear term explanations help bridge the gap for those who might struggle with medical jargon. By explaining complex terms like cardioselective beta-blockers or hypoglycemia in simple language, patients better understand their condition and the rationale behind their treatment. This not only em-

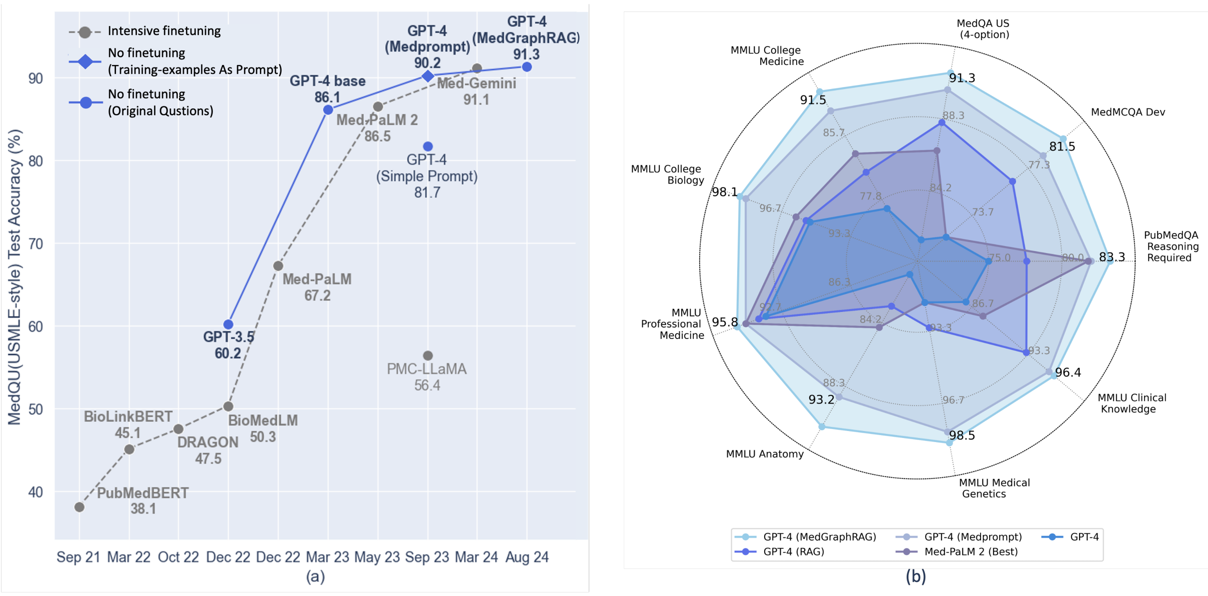


Figure 6: Compare to SOTA Medical LLM Models on MedQA benchmark.

powers them but also helps in preventing misunderstandings that could lead to improper management of their health. Altogether, using citations for evidence and plain language for explanation strikes the right balance between trust, safety, and accessibility in medical communication.

C.4 Case study: Abstracted Graph comparison between GraphRAG and MedGraphRAG

We conducted a closer examination of the abstracted graphs of GraphRAG (Fig. 16 a) and MedGraphRAG (Fig. 16 b) for the case study shown in the left plot of Fig. 7. By abstracting similar nearest neighbors of the retrieved entities (COPD and Heart Failure), we observed that MedGraphRAG accessed more detailed and specific entities, such as beta-1 receptors and Cardioselective Beta-Blockers, by linking to relevant references. While these entities are also present in the GraphRAG graph, they were not retrieved under the same number of nearest neighbors due to their indirect linkage with the retrieved entities. GraphRAG lacks a hierarchical graph that directly links these entities through an "is reference of" relationship, leading them to be overshadowed by more general neighbors at the same tier, ultimately missing retrieval.

Moreover, MedGraphRAG's approach to linking Heart Failure with Cardioselective Beta-Blockers enables further connections through beta-1 receptors in the second-tier graph, eventually linking back to Non-selective Beta-Blockers. It helps to link Heart Failure and Non-selective Beta-Blockers as neighbors in the first-tier graph relationship linking stage, which significantly enhances the LLM's ability to generate specific and accurate responses. Such an observation demonstrates the importance of including triple linking relationships when constructing the first-tier graph. MedGraphRAG leverages this unique design to build a more detailed and professional knowledge graph, resulting in better entity retrieval and richer context for accurate generation.

D Prompt used in the paper

D.1 Document Chunking

In document chunking, we used a FIFO semantic chunking method. We will first prompt LLM to generate a title and a summary of the current chunk using prompt template Fig. 17, for the current paragraph, we let LLM to determine if it should be include into this chunk based on the chunk title and summary using prompt template Fig. 18. If the paragraph is considered to be included in the chunk, then we will update the current chunk title and summary based on the prompt template Fig. 19.

D.2 Entity Extraction

We extract the entities from the chunk by prompt template Fig. 20.

D.3 Relationship Generation

We generate the relationship among entities by prompt template Fig. 21 and Fig. 22.

D.4 tag-summary generation

We use prompt template Fig. 23 and Fig. 24 to generate tag-summary for the graphs and queries. The predefined Medical Tags are in D.5.

D.5 Medical Tags

ANATOMICAL_STRUCTURE
Complex parts of the human body such as cells, organs, and systems.

SYMPTOM
Any subjective evidence of disease or physical disturbance as reported by the patient, e.g., "cough", "fatigue", "nausea".

BODY_PART
Specific part of the body, e.g., "hand", "liver", "spinal cord".

PHYSIOLOGICAL_PROCESS
The normal biological process in the body, e.g., "digestion", "circulation".

HEALTH_STATUS
General status of health or well-being, e.g., "healthy", "at risk", "critical".

MEDICAL_CONDITION
Any medical problem or condition, e.g., "hypertension", "asthma".

DISEASE_STAGE
The progression or stage of a disease, e.g., "Stage IV cancer", "early onset".

CAUSE_OF_ILLNESS
A cause or contributing factor to a condition, e.g., "infection", "genetic predisposition".

RISK_FACTOR
Any factor that increases the likelihood of developing a disease, e.g., "smoking", "family history".

PROGNOSIS
Expected outcome or forecast of the progression of a medical condition, e.g., "full recovery", "chronic".

VITAL_SIGN
Specific vital signs measured in a clinical setting, e.g., "heart rate", "blood pressure".

DIAGNOSTIC_TEST
Tests conducted for diagnosis purposes, e.g., "blood test", "CT scan".

TREATMENT_PLAN
A general plan for treating a medical condition, e.g., "chemotherapy", "physical therapy".

RECOVERY
Information regarding the recovery process, e.g., "rehabilitation needed", "recovery period of 2 weeks".

PREVENTIVE_MEASURE
Any action taken to prevent disease, e.g., "vaccination", "diet changes".

SYMPTOM_SEVERITY
The intensity of a symptom, e.g., "mild", "severe", "moderate".

CONTRAINDICATION
A condition or factor that serves as a reason to withhold a certain medical treatment, e.g., "contraindicated in pregnancy".

ALLERGY
Known allergies, e.g., "allergy to penicillin", "seasonal allergy".

BODY_FUNCTION
A function or activity carried out by the human body, e.g., "heart rate", "respiration".

BODY_FUNCTION_RESULT
The result or condition of a body function, e.g., "impaired", "normal".

BODY_MEASUREMENT
A quantifiable measurement of body function, such as vital signs obtained via basic instruments, e.g., "temperature", "blood pressure".

BODY_MEASUREMENT_RESULT
The specific value of a body measurement.

BODY_MEASUREMENT_UNIT
The unit for body measurements, e.g., "bpm", "mmHg".

LAB_TEST
A diagnostic examination performed on a bodily sample, e.g., blood test, urine test.

LAB_RESULT
A qualitative description of laboratory data, e.g., “positive”, “negative”, “elevated”.

LAB_VALUE
The numerical value obtained from lab testing.

LAB_UNIT
The unit of measurement for the lab value.

MEDICATION
Drugs or treatments prescribed for prevention or cure, e.g., “aspirin”, “insulin”.

MED_DOSE
The amount of medication prescribed at one time.

MED_FREQUENCY
The frequency at which a medication is taken, e.g., “twice a day”.

MED_ROUTE
The mode of administration of a medication, e.g., “oral”, “intravenous”.

MED_DURATION
The length of time a medication is to be taken.

MED_STATUS
The status of a medication regimen, such as “start”, “stop”, “continue”.

MED_STRENGTH
The concentration of the active ingredient in a medication.

MED_FORM
The form in which the medication is given, e.g., “tablet”, “liquid”.

MEDICAL_CONDITION
Includes diseases, findings, and symptoms.

PROCEDURE
Diagnostic or treatment procedure carried out on a patient, e.g., “MRI”, “surgery”.

PROCEDURE_RESULT
The outcome or findings of a procedure.

PROCEDURE_METHOD
Specific method or approach used during a procedure.

SEVERITY
Level of intensity or seriousness of a medical condition.

PATIENT_HISTORY
Description of past medical conditions, treatments, and procedures, e.g., “history of hypertension”.

FAMILY_HISTORY
Medical conditions and diseases present in the patient’s family, e.g., “mother had diabetes”.

LIFESTYLE_FACTORS
Relevant lifestyle details, e.g., “smoker for 10 years”, “alcohol consumption”.

ALLERGIES
Known allergies, e.g., “penicillin allergy”, “peanut allergy”.

TUMOR_DIMENSIONS
Measurements describing the size of a tumor or mass.

GENE_STUDIED
Genes linked to tumor formation, e.g., “BRCA1”.

HISTOLOGICAL_GRADE
The grade assigned to the appearance of cancerous cells.

RADIATION_DOSAGE
Amount of radiation used in cancer treatment.

CLINICAL_STATUS
Cancer status, e.g., “active”, “in remission”.

AGE
Age or life stage descriptor, e.g., “elderly”, “30 years old”.

FAMILY_STRUCTURE
Describes the patient’s family context or relationships, e.g., “married with two kids”.

LIVING_SITUATION
Details about a patient’s living arrangements, e.g., “lives alone”, “has a caregiver”.

SOCIAL_IDENTITY
Patient’s identity, including ethnicity, religion, nationality, e.g., “Hispanic”, “Catholic”.

OCCUPATION
Information regarding the patient’s employment status or history, e.g., “retired”, “construction worker”.

PERSON_NAME
Names or titles used to identify individuals.

ORGANIZATION_NAME
Names of organizations involved in patient care.

CONTACT_INFORMATION
Includes phone numbers, emails, URLs, ZIP codes.

GENERIC_ID
Identifiers such as medical record numbers, social security numbers.

D.6 Tag-Summary generation

We use prompt template Fig. 25, 26 to merge Tag summaries.

D.7 Response user query

The prompt template for generating the response with the retrieved Med-MetaGraph is shown in Fig. 27. The prompt template to revise the response based on the tag-summary is shown in Fig. 28.

E Details of Human Evaluation

E.1 Physician Answers

In the rating process, physicians were given unlimited time to return their responses and were allowed to use reference materials. They were instructed to tailor their answers to a layperson with average reading comprehension. The tasks were not tied to a specific clinical context or setting.

E.2 Physician and lay-person raters

Human evaluations were conducted by both physician and layperson raters. The physician raters included seven individuals: two from the US, three from China, and two from Singapore. Their specialties spanned ophthalmology, internal medicine, oncology, cardiology, gender dysphoria, epidemiology, and surgery. The layperson raters consisted of five individuals (three female, two male, aged 22-51) based in the US, all without a medical background. Their educational backgrounds included two with high school diplomas, two with graduate degrees, and one with postgraduate experience.

E.3 Individual evaluation of long-form answers

Raters were blinded to the source of each answer and conducted their evaluations independently, without consulting one another. The experiments are conducted on 120 questions randomly sampled from the MultiMedQA dataset and all 50 questions of DiverseHealth dataset. In all new rating experiments, each response was evaluated by all 12 raters. Inter-rater reliability analysis showed very good agreement ($\kappa > 0.8$) among raters for the MultiMedQA answers, and good agreement ($\kappa > 0.6$) for the DiverseHealth dataset, including assessments on whether answers lacked important citations or contained unrelated citation information.

E.4 Survey Questions

For each response generated by the models being evaluated, please rate the response based on the following dimensions. Provide your rating on a scale of 1 to 5, where 1 is the lowest and 5 is the highest. Additionally, provide comments if necessary.

1. **Pertinence (Pert.)**
   - How relevant is the response to the given medical query?
   - Rating Scale: 1 (Not Relevant) to 5 (Highly Relevant)
   - **Optional Comment**: What aspects of the response made it relevant or irrelevant?

2. **Correctness (Cor.)**
   - How accurate is the information provided in the response?
   - Rating Scale: 1 (Incorrect) to 5 (Completely Accurate)
   - **Optional Comment**: Identify any specific inaccuracies or issues in the response.

3. **Citation Precision (CP)**
   - How well does the provided citation support the statements made in the response?
   - Rating Scale: 1 (Not at All Supported) to 5 (Fully Supported)
   - **Optional Comment**: Are there any instances where the citation does not adequately support the statement?

4. **Citation Recall (CR)**
   - Does every important claim or medical statement have a corresponding citation to support it?

- Rating Scale: 1 (Many Claims Lack Citations) to 5 (All Claims are Supported by Citations)
- **Optional Comment**: Are there any claims made without proper citations that should be addressed?

5. **Understandability (Und.)**
   - How easy is it to understand the response, given the medical content?
   - Rating Scale: 1 (Difficult to Understand) to 5 (Very Easy to Understand)
   - **Optional Comment**: If any part of the response was unclear, what made it difficult to understand?

Rating Scale Definitions (1-5)

1. **1**: Very poor – lacks relevance, accuracy, proper citations, or clarity.
2. **2**: Poor – has significant shortcomings in one or more areas.
3. **3**: Fair – some issues are present, but the response is generally acceptable.
4. **4**: Good – response meets most expectations, with minor issues.
5. **5**: Excellent – fully meets expectations, with no significant issues.

F Boarder Impact

Our MedGraphRAG enhances LLMs by providing intrinsic source citations, significantly improving transparency, interpretability, and verifiability. This tool ensures that each response generated by the LLM is accompanied by provenance or source grounding information, clearly demonstrating that answers are rooted in the dataset. The availability of cited sources for each assertion allows users, especially in the critical field of medicine, to swiftly and accurately audit the LLM's output against the original source material. This feature is crucial where (a) a trustworthy relationship -based on transparency of reasoning, needs to exist between human experts and LLM agents and (b) where the security/privacy of the information shared with the agent is mission critical. By employing this method, we have developed an Evidence-based Medical LLM. Clinicians can easily verify the source of the reasoning and adjust the model's responses, ensuring the safe use of LLMs in clinical scenarios.

Question

Question:
Mr. John Smith, a 68-year-old retired engineer, presents with increasing forgetfulness and occasional episodes of sudden confusion and speech difficulty. His medical history includes type 2 diabetes, hypertension, and a transient ischemic attack two years ago. Neuroimaging reveals moderate vascular changes and white matter lesions, while his PET scan shows patchy decreased glucose metabolism. Given these symptoms and findings, which of the following is the most likely diagnosis?
Options:
A. Alzheimer's Disease
B. Vascular Dementia
C. Frontotemporal Dementia
D. Lewy Body Dementia
Answer:
B. Vascular Dementia

GPT-4

Answer:
A. Alzheimer's Disease
Explanation:
1.Cognitive Testing: His moderate cognitive impairment on the Montreal Cognitive Assessment (MoCA) might also be consistent with Alzheimer's, considering the general decline in cognitive functions.
2.Neuroimaging: Although the MRI shows vascular changes, the patchy areas of decreased glucose metabolism observed in the PET scan can sometimes overlap with early Alzheimer's patterns, reflecting disrupted brain activity.
3.Symptom Pattern: While his symptoms show vascular influences, the general forgetfulness and worsening cognitive decline over the past year could also align with the progressive nature of Alzheimer's disease.

MedGraphRAG

Answer:
B. Vascular Dementia
Explanation:
Medical History and Symptoms: His history of vascular risk factors and the nature of his episodic symptoms, which include sudden confusion and difficulties with speech, are indicative of neurological disturbances typically associated with vascular incidents like micro-strokes or TIAs [1].
1.Neuroimaging Results: The MRI showing moderate vascular changes and white matter lesions is characteristic of chronic ischemic damage prevalent in vascular dementia, rather than the neurodegenerative patterns typically observed in Alzheimer's or other dementias [2].
2.Additional Findings: His episodic cognitive symptoms also support a vascular origin rather than a degenerative neurological disease like Alzheimer's, which would generally present a gradual, consistent cognitive decline rather than episodic [3].
These factors collectively support the diagnosis of vascular dementia, aligning more with the implications of his vascular medical history and the episodic nature of his cognitive disturbances.

Vascular Dementia: A form of dementia caused by an impaired supply of blood to the brain, often resulting from strokes or other vascular injuries, which leads to cognitive decline.
Episodic symptoms: Symptoms that occur in discrete episodes, rather than being continuous. These can vary in severity and nature and are transient, typically related to neurological events.
Micro-strokes or TIAs: Micro-strokes are small, unnoticed strokes that can cause temporary, minor neurological symptoms. TIAs are brief episodes of neurological dysfunction caused by loss of blood flow to the brain, typically lasting less than 24 hours without causing permanent damage.
White matter lesions: These are indicators of vascular pathology in the brain, including changes in blood vessel structure and areas of damage in the white matter, often due to chronic ischemic (lack of blood flow) conditions.
Chronic ischemic damage: Long-standing damage to brain tissue caused by repeated or persistent reduction in blood supply, typically resulting in loss of neuronal function and health.
Neurodegenerative patterns: Refers to the structural and functional changes in the brain associated with progressive neurological diseases like Alzheimer's, where there is gradual degradation and death of neurons.
Episodic cognitive symptoms: Fluctuating cognitive impairments that appear in episodes, contrasted with steady cognitive decline seen in degenerative diseases.

[1]: O'Brien, J. T., & Thomas, A. (2015). Vascular dementia. *Lancet*, 386(10004), 1698-1706.
[2]: Smith, E. E., & Beaudin, A. E. (2018). New insights into cerebral small vessel disease and vascular cognitive impairment from MRI. *Current Opinion in Neurology*, 31(1), 36-43.
[3]: Gorelick, P. B., Scuteri, A., Black, S. E., et al. (2011). Vascular contributions to cognitive impairment and dementia: A statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*, 42(9), 2672-2713.

Figure 7: Example case shows MedGraphRAG generating evidence-based responses with grounded citations and terminology explanations.

Comorbidity

Question

In patients with both chronic obstructive pulmonary disease (COPD) and heart failure, how can lung function be improved?

MedGraphRAG

In patients with both **COPD** and **heart failure**, the combination of **bronchodilators** and **cardioselective beta-blockers**can be used effectively to manage both conditions without significantly worsening respiratory symptoms or heart function **[1,4]**. **Cardioselective beta-blockers**, such as bisoprolol or metoprolol, are safe for **COPD** patients because they primarily block **beta-1 receptors** in the heart, minimizing the risk of **bronchoconstriction** associated with non-selective beta-blockers **[1,2]**. **Bronchodilator** therapy, including long-acting beta-2 agonists and **muscarinic antagonists**, should be continued to improve lung function in **COPD [3]**.

Bronchodilators, such as long-acting beta-2 agonists (LABAs**)**, help relax the airway muscles, which reduces **bronchoconstriction** and increases airflow, thereby improving lung function **[3,5]**. **Muscarinic antagonists**, like tiotropium, work by blocking the action of acetylcholine on airway muscles, further aiding in bronchodilation and easing the symptoms of **COPD [5]**. These medications should be continued as part of the treatment regimen to enhance respiratory efficiency.

Cardioselective beta-blockers, such as bisoprolol and metoprolol, are particularly beneficial for **COPD** patients as they predominantly target **beta-1 receptors** in the heart, reducing cardiovascular strain without significantly affecting the **beta-2 receptors** that are crucial for lung function **[1,2]**. This selective targeting helps avoid the **bronchoconstriction** risk often seen with non-selective **beta-blockers**, which can negatively impact patients with respiratory conditions **[2,6]**.

For patients managing both **COPD** and **heart failure**, the use of **cardioselective beta-blockers** in combination with **bronchodilators** can be a balanced and effective strategy to optimize both cardiovascular and respiratory health. Studies have shown that the careful use of **cardioselective beta-blockers** reduces mortality risk without exacerbating respiratory symptoms **[2]**. This combination treatment helps maintain heart function while avoiding respiratory compromise, ultimately contributing to improved lung function and better overall outcomes in patients with these comorbid conditions.

References:

1.Salpeter SR, Ormiston TM, Salpeter EE. "Cardioselective beta-blockers for chronic obstructive pulmonary disease." *Cochrane Database of Systematic Reviews*. 2005.

2.Du Q, Sun Y, Ding N, Lu L. "Beta-blockers reduced the risk of mortality and exacerbation in patients with COPD: A meta-analysis of observational studies." *PLoS ONE*. 2016; 11(11): e0166861.

3.O'Donnell DE, Parker CM. "COPD exacerbations . 3: Pathophysiology." *Thorax*. 2006.

4.Hawkins NM, Petrie MC, Jhund PS, Chalmers GW, Dunn FG, McMurray JJ. "Heart failure and chronic obstructive pulmonary disease: diagnostic pitfalls and epidemiology." *European Journal of Heart Failure*. 2009; 11(2): 130-139.

5.Tashkin DP, Celli B, Senn S, et al. "A 4-year trial of tiotropium in chronic obstructive pulmonary disease." *New England Journal of Medicine*. 2008; 359(15): 1543-1554.

6.Rutten FH, Cramer MJ, Lammers JW, Grobbee DE, Hoes AW. "Heart failure and chronic obstructive pulmonary disease: An ignored combination?" *European Journal of Heart Failure*. 2006; 8(7): 706-711.

Figure 8: MedGraphRAG long-form generation response on Comorbidity example-part1.

Terminological Clarification:
•**COPD (Chronic Obstructive Pulmonary Disease)**: A chronic inflammatory lung disease that obstructs airflow, causing breathing difficulties, cough, and mucus production.
•**Heart Failure**: A condition in which the heart cannot pump blood effectively, leading to symptoms such as fatigue, shortness of breath, and fluid retention.
•**Bronchodilators**: Medications that relax the muscles surrounding the airways, leading to widened air passages and improved airflow to ease breathing in **COPD** patients.
•**Cardioselective Beta-Blockers**: A type of **beta-blocker** that primarily targets **beta-1 receptors** in the heart, which reduces the risk of respiratory complications. They are used to manage heart conditions without significantly affecting lung function.
•**Beta-Blockers**: A class of medications used to manage cardiovascular conditions like hypertension and **heart failure**. They work by blocking **beta receptors** to decrease heart rate and blood pressure. Non-selective **beta-blockers** can also affect **beta-2 receptors**, leading to **bronchoconstriction**.
•**Beta-1 Receptors**: Receptors primarily found in the heart that, when blocked, lead to reduced heart rate and force of contraction, making them a key target for treating cardiovascular diseases.
•**Beta-2 Receptors**: Receptors found in the smooth muscles of the airways. Blocking these receptors can lead to **bronchoconstriction**, which is why cardioselective **beta-blockers** that do not significantly impact these receptors are preferred for **COPD** patients.
•**Bronchoconstriction**: The narrowing of the airways in the lungs due to the tightening of surrounding muscles, leading to difficulty in breathing. This is a concern for **COPD** patients and can be exacerbated by non-selective **beta-blockers**.
•**Beta-2 Agonists**: Medications that stimulate **beta-2 receptors** to relax airway muscles and promote bronchodilation, improving airflow in **COPD** patients.
•**Muscarinic Antagonists**: Medications that block the action of acetylcholine on muscarinic receptors, which helps to relax the airways and improve airflow, thus reducing symptoms in **COPD**.

Figure 9: MedGraphRAG long-form generation response on Comorbidity example-part2.

Rare Disease

Question

What are the most effective strategies for managing progressive thoracic insufficiency in patients with fibrodysplasia ossificans progressiva (FOP)?

MedGraphRAG

Managing progressive **thoracic insufficiency** in **FOP** patients requires a careful balance between effective respiratory support and preventing **heterotopic ossification (HO) [2]**. Early use of **non-invasive ventilation (NIV) [1]**, such as **BiPAP**, particularly during sleep or in response to **respiratory infections**, can improve respiratory function without invasive procedures **[2]**, minimizing the risk of trauma **[3]**. Gentle, supervised **physical therapy** helps maintain lung capacity while avoiding trauma that could trigger **HO [4]**.

Non-invasive ventilation (NIV), like **BiPAP**, can be used to provide respiratory support, especially during periods of compromised breathing due to **respiratory infections** or sleep. This avoids the need for invasive procedures, which can provoke **HO** in **FOP** patients **[1,2]**. **BiPAP** specifically offers different levels of air pressure during inhalation and exhalation, making it a comfortable and effective choice for improving respiratory function **[3,5]**.

Supervised **physical therapy** plays a crucial role in maintaining lung capacity while minimizing trauma. It is important that such exercises are gentle and performed under supervision to avoid any injury that might lead to the development of **heterotopic ossification [4,6]**. The main goal is to preserve thoracic movement without stimulating the abnormal bone growth characteristic of **FOP**.

Nutritional support is also essential to optimize respiratory function in **FOP** patients. Maintaining adequate nutrition helps to improve overall health and supports the muscles needed for breathing **[7]**.

References:

1.Shulmanl, et al. "Maternal genetic skeletal disorders: lessons learned from cases of maternal osteogenesis imperfecta and fibrodysplasia ossificans progressiva." *Journal of Clinical*.

2.Pignolo RJ, et al. "The natural history of flare-ups in fibrodysplasia ossificans progressiva (FOP): a comprehensive global assessment." *Journal of Bone and Mineral Research*. 2016.

3.Kaplan FS, et al. "Fibrodysplasia ossificans progressiva: Clinical and genetic aspects." *Orphanet Journal of Rare Diseases*. 2008.

4.Additional source not explicitly listed (general reference to supervised physical therapy for maintaining lung capacity in **FOP** patients).

5.Rocke DM, et al. "Pulmonary function and its management in FOP." *Respiratory Care*. 2017.

6.Kaplan FS, Glaser DL, Shore EM, et al. "The medical management of fibrodysplasia ossificans progressiva: current treatment considerations." *Clinical Reviews in Bone and Mineral Metabolism*. 2005.

7.Al Mukaddam M, et al. "Nutrition in Fibrodysplasia Ossificans Progressiva: Strategies and Management." *Journal of Clinical Endocrinology & Metabolism*. 2020.

Figure 10: MedGraphRAG long-form generation response on Minority Health example-part1.

Terminological Clarification:
•**Heterotopic Ossification (HO)**: The formation of bone tissue in areas outside of the skeletal system, usually in soft tissues like muscles. It is common in **FOP** and can be triggered by trauma or injury, resulting in limited mobility and pain.
•**Non-Invasive Ventilation (NIV)**: A respiratory support method that uses a mask to deliver air to the lungs, assisting breathing without requiring surgical procedures. It is commonly used to avoid complications in **FOP**patients where trauma can lead to **HO**.
•**BiPAP (Bilevel Positive Airway Pressure)**: A type of **NIV** that provides two levels of air pressure—higher pressure when inhaling and lower pressure when exhaling. It helps maintain effective ventilation and supports breathing during sleep or when respiratory function is compromised.
•**Fibrodysplasia Ossificans Progressiva (FOP)**: A rare genetic disorder characterized by the formation of bone in muscles, tendons, and other connective tissues, leading to progressive immobility. Even minor trauma can trigger flare-ups, resulting in abnormal bone growth.
•**Thoracic Insufficiency**: A condition where the chest wall cannot adequately support normal breathing or lung growth, leading to reduced lung function. It is a common complication in **FOP** patients due to **HO** affecting the ribcage and spine.
•**Respiratory Infections**: Infections affecting the respiratory tract, which can significantly worsen breathing in **FOP**patients. Prevention and effective management are crucial to avoid triggering **HO**.
•**Physical Therapy**: Supervised exercises designed to maintain or improve mobility. For **FOP** patients, it must be done cautiously to avoid any physical stress or trauma that could lead to **HO**.
•**Nutritional Support**: Adequate nutrition is crucial for maintaining general health and supporting respiratory muscle function, especially in **FOP** patients where compromised lung capacity can lead to difficulties in meeting nutritional needs.

Figure 11: MedGraphRAG long-form generation response on Minority Health example-part2.

Minority Health

Question

Will long-term hormone replacement therapy (HRT) injections affect life expectancy in a male-to-female transgender individual?

MedGraphRAG

Yes, long-term **hormone replacement therapy (HRT)** injections in **male-to-female transgender** individuals can affect life expectancy by potentially increasing certain health risks **[1,4]**, particularly **cardiovascular events [2]** and **thromboembolic complications [3]**. However, with proper medical supervision **[5]**, **risk assessment**, and **management strategies**, these risks can be mitigated, leading to a life expectancy similar to the general population.
Proper medical supervision is crucial for monitoring hormone levels, screening for potential complications, and managing any risks associated with long-term **HRT [5]**. **Cardiovascular events**, such as heart attacks or strokes, are potential risks that should be closely monitored in individuals undergoing long-term **HRT**, especially when risk factors like smoking or pre-existing heart conditions are present **[2]**.
Thromboembolic complications, including deep vein thrombosis (DVT) or pulmonary embolism, are also possible but can be managed through careful risk stratification and preventive measures **[3,6]**. With regular medical check-ups and the use of personalized **management strategies**, individuals undergoing **HRT** can mitigate these risks and achieve a life expectancy comparable to that of the general population.

References:

1.Dhejne C, et al. "Long-Term Follow-Up of Transsexual Persons Undergoing Sex Reassignment Surgery: Cohort Study in Sweden." *PLoS ONE*. 2011.
2.Nota NM, et al. "Occurrence of Acute Cardiovascular Events in Transgender Individuals Receiving Hormone Therapy." *European Journal of Endocrinology*. 2019.
3.Asscheman H, et al. "Venous Thrombosis and Long-Term Use of Estrogens in Transgender Women." *Journal of Sexual Medicine*. 2014.
4.Wierckx K, et al. "Long-term evaluation of cross-sex hormone treatment in transsexual persons." *Journal of Clinical Endocrinology & Metabolism*. 2012.
5.Hembree WC, et al. "Endocrine Treatment of Gender-Dysphoric/Gender-Incongruent Persons: An Endocrine Society Clinical Practice Guideline." *Journal of Clinical Endocrinology & Metabolism*. 2017.
6.Seal LJ, et al. "Health Implications of Gender Affirmation in Transgender Individuals." *Lancet Diabetes Endocrinology*. 2016.

Figure 12: MedGraphRAG long-form generation response on Rare Disease example1.

Terminological Clarification:
•**Hormone Replacement Therapy (HRT)**: A treatment involving the administration of hormones, such as estrogen or testosterone, to align an individual's physical characteristics with their gender identity. In **male-to-female transgender** individuals, **HRT** typically includes estrogen and anti-androgens.
•**Male-to-Female Transgender**: Refers to individuals assigned male at birth who identify and transition to female. **HRT** is a part of the transition process that helps in developing feminine physical traits.
•**Cardiovascular Events**: Medical conditions affecting the heart or blood vessels, including heart attacks, strokes, and other complications related to the cardiovascular system. These events can be a risk factor for individuals on long-term **HRT**.
•**Thromboembolic Complications**: Conditions that involve blood clots forming in the veins, which may lead to **deep vein thrombosis (DVT)** or **pulmonary embolism**. These are possible risks associated with **HRT** in transgender individuals.
•**Proper Medical Supervision**: Regular healthcare support that includes monitoring hormone levels, evaluating risk factors, and adjusting treatments as necessary. This helps mitigate potential risks associated with **HRT**.
•**Risk Assessment**: The process of identifying potential health risks, such as cardiovascular issues or thromboembolic complications, in individuals undergoing **HRT**.
•**Management Strategies**: Techniques or treatments designed to minimize health risks. For individuals on **HRT**, this may include lifestyle modifications, regular monitoring, and tailored medical care to address potential complications.
•**Deep Vein Thrombosis (DVT)**: A condition where a blood clot forms in a deep vein, usually in the legs. **DVT** is one of the possible **thromboembolic complications** associated with **HRT**.
•**Pulmonary Embolism**: A blockage in one of the pulmonary arteries in the lungs, usually caused by a blood clot that has traveled from the legs (**DVT**). It is a serious complication that can occur with long-term **HRT**.

Figure 13: MedGraphRAG long-form generation response on Rare Disease example-part2.

Question

Chornic Disease Managment

I have Type 1 Diabetes and I'm planning to exercise for about 30 minutes after work. How should I adjust my insulin and eating?

MedGraphRAG

To avoid **low blood sugar** during and after your 30-minute exercise session, start by checking your **blood glucose** before you begin; if it's below your target range, consume a small **carbohydrate** snack (about 15–30 grams) **[2,3,6]**. Consider reducing your **rapid-acting insulin** dose before exercise by 25–50% if you've taken it within the last few hours **[2,3,7]**, as exercise increases **insulin sensitivity** and can lower **blood sugar levels [1,2,3,4]**. Keep **fast-acting carbohydrates**readily available during your workout in case you experience symptoms of **hypoglycemia [1,3,5]**.

Monitoring **blood glucose** levels before, during, and after exercise helps maintain stable **blood sugar levels**. Consuming **carbohydrates** before exercise if **blood glucose** is low will help prevent **hypoglycemia** and allow you to exercise more comfortably **[2,3,6]**. **Rapid-acting insulin** adjustments are important, as exercise can significantly increase **insulin sensitivity**, leading to a faster decline in **blood sugar levels [2,3,7]**.

Fast-acting carbohydrates (such as glucose tablets or sugary drinks) should be easily accessible during exercise in case of **hypoglycemia**. These types of carbohydrates act quickly to raise **blood sugar levels** and can help stabilize levels if they drop too low during activity **[1,2,3,5]**.

References:

1.Bakris, G., et al. "American Diabetes Association (ADA) standards of medical care in diabetes 2016 cardiovascular disease and risk management." *Diabetes Care*. 2016.

2.Riddell MC, et al. "Exercise management in type 1 diabetes: a consensus statement." *The Lancet Diabetes & Endocrinology*. 2017.

3.Colberg SR, et al. "Physical Activity/Exercise and Diabetes: A Position Statement of the American Diabetes Association." *Diabetes Care*. 2016.

4.Yardley JE, et al. "Effects of Exercise on Glucose Levels in Type 1 Diabetes: A Continuous Glucose Monitoring Study." *Diabetes Technology & Therapeutics*. 2013.

5.Adolfsson P, et al. "Improvements in Diabetes Control During Physical Activity Intervention for Adolescents with Type 1 Diabetes Mellitus." *Journal of Clinical Endocrinology & Metabolism*. 2017.

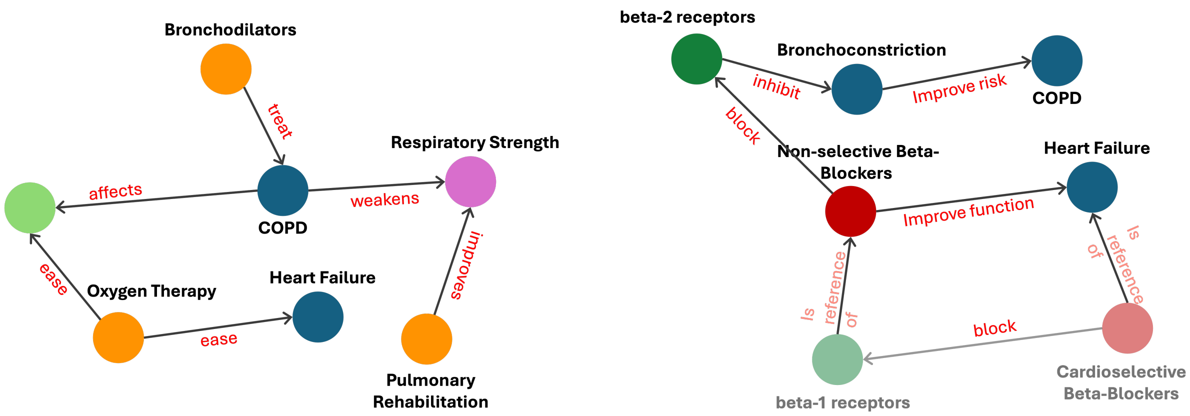
6.Rabasa-Lhoret R, et al. "Exercise in Type 1 Diabetes: A practical review of its benefits and challenges." *Journal of Diabetes and its Complications*. 2009.

7.Heinemann L, et al. "Adjustment of insulin therapy for physical activity in type 1 diabetes mellitus." *Diabetes Obesity and Metabolism*. 2014.

Figure 14: MedGraphRAG long-form generation response on Chornic Disease Managment example-part1.

Terminological Clarification:
•**Low Blood Sugar (Hypoglycemia)**: A condition where **blood glucose** levels fall below the normal range, leading to symptoms like shakiness, sweating, confusion, or even loss of consciousness if untreated.
•**Blood Glucose**: The concentration of glucose (sugar) in the blood, commonly referred to as **blood sugar levels**. Monitoring **blood glucose** is crucial for individuals with **Type 1 Diabetes** to manage their health.
•**Carbohydrate**: A macronutrient found in foods such as bread, fruits, and sweets that provides energy. Consuming **carbohydrates** before exercise can help maintain **blood glucose** levels, especially in individuals with diabetes.
•**Rapid-Acting Insulin**: A type of insulin that starts to work quickly to reduce **blood glucose** levels, typically within 15 minutes of injection. It helps manage the spikes in **blood sugar** that occur after meals.
•**Insulin Sensitivity**: The body's responsiveness to insulin, meaning how effectively insulin helps cells absorb glucose. Exercise increases **insulin sensitivity**, which means the body requires less insulin to lower **blood glucose**levels.
•**Blood Sugar Levels**: The amount of glucose present in the blood at any given time. Managing **blood sugar levels**is essential for individuals with diabetes to prevent both **hypoglycemia** and hyperglycemia.
•**Fast-Acting Carbohydrates**: Carbohydrates that are rapidly absorbed into the bloodstream, quickly raising **blood glucose** levels. Examples include glucose tablets, sugary drinks, and candies. These are used to treat **hypoglycemia**.
•**Hypoglycemia**: A condition characterized by abnormally low **blood glucose** levels, which can be caused by too much insulin, insufficient food intake, or increased physical activity without proper adjustments.

Figure 15: MedGraphRAG long-form generation response on Chornic Disease Managment example-part2.



a. Abstracted graph of GraphRAG

b. Abstracted graph of MedGraphRAG

Figure 16: The comparison of abstracted graph between GraphRAG and MedGraphRAG.

System Prompt:

You are the steward of a chunk, which represents a collection of paragraphs discussing a similar medical topic.
Task:
•You should generate both a **concise chunk title** and a **1-sentence chunk summary** that clearly indicate what the chunk is about.
•The **chunk title** should be a few words that clearly describe the topic of the chunk.
•The **chunk summary** should provide a brief overview of the content, explaining what is covered.
Requirements for Titles and Summaries:
•The **title** should be **brief** but fully capture the essence of the chunk.
•The **summary** should describe what the chunk is about and provide any clarifying instructions for additions.
•Your titles and summaries should **anticipate generalization** where appropriate. For example:

- If the chunk is about **high blood pressure**, generalize to **cardiovascular conditions**.
- If the content is about **a specific treatment**, generalize to **therapeutic approaches**.

Examples:
1.Input: Chunk Content: This chunk contains multiple paragraphs about the use of basal insulin in managing Type 1 diabetes, the benefits of continuous glucose monitoring (CGM) devices, and the role of patient education in understanding insulin dosing adjustments.
Output:
Chunk Title: Basal Insulin and Monitoring in Type 1 Diabetes
Chunk Summary: Detailed exploration of basal insulin use, benefits of continuous glucose monitoring, and the role of patient education in optimizing insulin dosing for Type 1 diabetes management.
2.Input:
Chunk Content: This chunk includes paragraphs on the use of oseltamivir as an antiviral treatment for influenza, the importance of annual influenza vaccination for high-risk populations, and the benefits of early initiation of antiviral drugs to reduce complications.
Output:
Chunk Title: Oseltamivir and Vaccination for Influenza
Chunk Summary: Focused discussion on oseltamivir as an antiviral, the significance of annual flu vaccination, and the role of early antiviral treatment in managing influenza and reducing complications.

User Prompt:

Determine the title and summary of the chunk based on its content:
Chunk Content: {chunk_content}

Figure 17: Prompt to generate chunk outline.

System Prompt:

Task:
Determine whether the given paragraph should be included in the current chunk.
Criteria:
•A paragraph should be included in a chunk if its meaning, direction, or intention is similar to that of the chunk.
Goal:
•The objective is to group similar paragraphs into cohesive chunks.
Context Provided:
•You will be given the outline of the current chunk and the previous four paragraphs as context.
Instructions:
•If you think the paragraph should be joined with the current chunk, return **"Yes"**.
•If you think the paragraph should not be joined with the current chunk, return **"No"**.
Examples:
1.Current Chunk Outline:
1. **Chunk Name:** Whole Grain Intake for Glycemic Control in Type 2 Diabetes
2. **Chunk Summary:** Specific analysis of how increasing whole grain consumption can contribute to better glycemic control in Type 2 diabetes patients.

Given Paragraph:
"A balanced diet low in processed sugars and rich in whole grains is crucial for managing Type 2 diabetes effectively."
Return: "Yes"
Reason: The paragraph directly discusses increasing whole grain intake, which aligns well with the chunk's specific focus on glycemic control through whole grains.
2.Current Chunk Outline:
1. **Chunk Name:** ACE Inhibitors for Hypertension Management
2. **Chunk Summary:** Detailed discussion of the role of ACE inhibitors in lowering blood pressure and their mechanisms of action for effective hypertension management.

Given Paragraph:
"Regular physical activity and reducing sodium intake are essential lifestyle changes for managing high blood pressure."
Return: "No"
Reason: The paragraph is about non-pharmacological interventions for hypertension management, whereas the chunk specifically focuses on the pharmacological role and mechanisms of ACE inhibitors.

User Prompt:

Current Chunks:
--Start of current chunks—
{current_chunk_outline}
--End of current chunks—

Previous Paragraphs:
--Start of previous paragraphs—
{the four previous paragraphs}
--End of previous paragraphs—

Determine if the following paragraph should belong to the current chunk: {current paragraph}

Figure 18: Prompt to determine whether the given paragraph should be included in the chunk.

System Prompt:

You are the steward of a chunk, which represents a collection of paragraphs discussing a similar medical topic.
A new paragraph has been added to the chunk, and your task is to **update both the chunk title and summary** to ensure they accurately reflect the current content.
Requirements for Title and Summary Updates:
•The **title** should be **brief** but fully capture the updated essence of the chunk.
•The **summary** should provide an accurate overview of the updated chunk content.
•Both the title and summary should **anticipate generalization** where appropriate. For example:

- If the content is about **high blood pressure**, generalize to **cardiovascular conditions**.
- If the content is about **a specific treatment**, generalize to **therapeutic approaches**.

Example:
Input:
New Paragraph: This paragraph discusses the benefits of basal insulin in maintaining consistent blood sugar levels for patients with Type 1 diabetes.
Current Chunk Title: Insulin Therapy for Type 1 Diabetes
Current Chunk Summary: The use of continuous glucose monitoring (CGM) for real-time tracking of blood glucose levels and basal and bolus insulin therapy for optimal glycemic control.
Output:
Updated Chunk Title: Insulin Therapy for Type 1 Diabetes
Updated Chunk Summary: The use of continuous glucose monitoring (CGM) for real-time tracking of blood glucose levels, basal insulin therapy for maintaining consistent blood sugar levels, and bolus insulin therapy for optimal glycemic control in patients with Type 1 diabetes.

User Prompt:

A new paragraph will be added to the chunk. Update the chunk title and summary based on the current ones:
New paragraph :
{paragraph_content}
Current Chunk Title:
{current_title}
Current Chunk Summary:
{current_summary}

Figure 19: Prompt to update the chunk outline based on the new paragraph.

System Prompt:

You are tasked with extracting key entities from a given paragraph and structuring them into Entity objects.
Your Task:
•Analyze the input paragraph to identify important entities.
•Entities can be noun phrases or key concepts mentioned in the content.
Entity Structure:
For each identified entity, create an Entity object that includes:
•**Name**: Extracted directly from the text or inferred from context.
•**Type**: Chosen from **UMLS semantic types**
•**Context**: A few sentences explaining the entity's relevance within the paragraph.
Instructions:
•Carefully read the provided content.
•Identify distinct entities and assign an appropriate UMLS semantic type.
•Provide contextual information that explains the entity's role within the paragraph.
Example:
Input Paragraph: "Metformin is widely used for treating Type 2 diabetes. It helps lower blood sugar levels and improve insulin sensitivity. The medication is recommended by most endocrinologists."
Expected Output:
Entities:
1.Entity:
1. **Name**: Metformin
2. **Type**: Pharmacologic Substance
3. **Context**: "Metformin is widely used for treating Type 2 diabetes."

2.Entity:
1. **Name**: Type 2 diabetes
2. **Type**: Disease or Syndrome
3. **Context**: "Metformin is widely used for treating Type 2 diabetes."

3.Entity:
1. **Name**: Endocrinologists
2. **Type**: Biomedical Occupation or Discipline
3. **Context**: "The medication is recommended by most endocrinologists."

User Prompt:

Extract entities from the given paragraph and structure them into Entity objects with the specified properties:
Paragraph:
{paragraph}

Figure 20: Entity extraction prompt.

System Prompt:

You are tasked with generating a concise description of the relationship between two given medical entities. The two entities and their reference entities are provided, and your job is to identify and describe the relationship between them based on established medical knowledge and context.

Instructions:

Review the Given Entities:

Examine the provided **primary entities** that need to be related.
Each entity includes a **context**, which is a description of the entity within the paragraph.
You will also receive **reference entities** to help understand the context.

Identify Relevant Relationships:

Use the primary entities and their context, along with reference entities, to determine the relationship.
The relationship should be based on established medical associations, such as:

- **Treats**: A drug treats a disease.
- **Causes**: A condition or substance causes a symptom.
- **Associated With**: Two entities are linked or commonly occur together.
- **Improves**: A medication or intervention improves a function.
- **Diagnoses**: A test or procedure diagnoses a condition.

Create a Relationship Object:

For each identified relationship, provide:

- **Subject (subj)**: The first entity Name.
- **Object (obj)**: The second entity Name.
- **Relationship Description**: A concise phrase describing the relationship (e.g., "treats," "associated with").

Output Formatting:

Present the relationship as a structured object as shown in the example.
Ensure clarity and consistency for easy understanding.

Example:

Given Entities:

Primary Entities:

Entity 1:

- **Name: Metformin**
- **Type:** Pharmacologic Substance
- **Context:** "Metformin is widely used for treating Type 2 diabetes."

Entity 2:

- **Name: Type 2 Diabetes**
- **Type:** Disease or Syndrome
- **Context:** "Metformin is widely used for treating Type 2 diabetes."

Reference Entities:

Entity 1:

- **Name: Metformin**
- **Parent**: Metformin
- **Type:** Pharmacologic Substance
- **Context:** "Metformin helps to lower blood sugar levels."

Entity 2:

- **Name: Type 2 Diabetes**
- **Parent**: Type 2 Diabetes
- **Type:** Disease or Syndrome
- **Context:** "Endocrinologists often recommend Metformin for managing Type 2 diabetes."

Figure 21: Relationship generation prompt-part1.

Generated Relationship:
Relationship:
Subject (subj):
Metformin
Object (obj):
Type 2 Diabetes
Relationship Description: "Treats"
Note:
Ensure that the relationship description is **concise and accurate**.
Use only verified medical relationships based on the context provided by the primary and reference entities.
The relationship description should be a **concise phrase** (e.g., "treats," "causes," "associated with").

User Prompt:

Given the following two entities, generate a **concise relationship description** between them using the specified format.
Primary Entities:
•**Entity 1:**
- **Name:** {entity_1_name}
- **Type:** {entity_1_type}
- **Context:** {entity_1_context}

•**Entity 2:**
- **Name:** {entity_2_name}
- **Type:** {entity_2_type}
- **Context:** {entity_2_context}

Reference Entities:
{The list of reference entities.}

Figure 22: Relationship generation prompt-part2.

System Prompt:

You are tasked with summarizing medical content from provided medical graph with structure [entity content][relationship][entity content]. Each summary must strictly adhere to a set of predefined categories, ensuring that the extracted information is well-structured and presented in a concise manner. For each category, provide the information using the format: 'CATEGORY_NAME: Key details'. Only include information relevant to each category, and avoid unnecessary elaboration: {Medical Tags and Description provided.}
Each category should be addressed only if relevant to the source content. The summary should be concise, precise, and well-structured, allowing for rapid comprehension and use.
The summary should include the reference DOI at the end if the content from the reference is used in constructing the summary. Additionally, include the CUI (Concept Unique Identifier) number after specific terms if the term may be obscure and needs clarification for individuals without a clinical background.

User Prompt:

Generate a structured summary from the following graph {a set of [entity content][relationship][entity content]}

Figure 23: tag-summary generation prompt for the graphs.

System Prompt:

You are tasked with summarizing medical content from provided user query. Each summary must strictly adhere to a set of predefined categories, ensuring that the extracted information is well-structured and presented in a concise manner. For each category, provide the information using the format: 'CATEGORY_NAME: Key details'. Only include information relevant to each category, and avoid unnecessary elaboration: {Medical Tags and Description provided.}
Each category should be addressed only if relevant to the source content. The summary should be concise, precise, and well-structured, allowing for rapid comprehension and use.

User Prompt:

Generate a structured summary from the following query {user query}

Figure 24: tag-summary generation prompt for the queries.

System Prompt:

You are tasked with merging two summaries of medical content, each structured by a set of predefined medical categories. Your goal is to combine these summaries into a single cohesive summary while preserving all relevant information and maintaining clarity.

Instructions:

1.Combine Information for Each Category:

1. For each category, **compare** the content from both summaries.
2. If both summaries mention different instances within a category, identify **shared characteristics** and refer to them using a **function word** or a more **general term (parent category)**.
    1. **Examples**:
        1. "disease: glaucoma" and "disease: diabetic retinopathy" should be merged as "disease: eye disease."
        2. "ANATOMICAL_STRUCTURE: retina" and "ANATOMICAL_STRUCTURE: cornea" should be generalized as "ocular structure."
        3. "SYMPTOM: blurred vision, eye pain, vision loss, redness" should be generalized as "visual disturbance."
3. If there are **common elements** between both summaries, include them only once in the merged output.

2.Avoid Redundancy:

1. Remove repeated information while retaining key details in a concise form.

3.Category Integrity:

1. Maintain clear labels for each category, and include all relevant categories in the final summary.
2. If a category is irrelevant to both summaries, it can be omitted.

4.Well-Structured Format:

1. The final output should be concise, well-structured, and easy to understand, ensuring that the original meaning is retained.

Output Format: Provide the merged information for each category in the following format:
CATEGORY_NAME: Key details.

Figure 25: tag-summary merge prompt, part-1.

Merge Example:
•**Input Summary A**:
- **ANATOMICAL_STRUCTURE**: retina
- **SYMPTOM**: blurred vision, eye pain
- **DISEASE**: glaucoma

•**Input Summary B**:
- **ANATOMICAL_STRUCTURE**: cornea, retina
- **SYMPTOM**: vision loss, redness
- **DISEASE**: diabetic retinopathy

Expected Output:
•**ANATOMICAL_STRUCTURE**: ocular structure
- **Explanation**: "Retina" and "cornea" are generalized as "ocular structure."

•**SYMPTOM**: visual disturbance
- **Explanation**: Symptoms such as "blurred vision," "eye pain," "vision loss," and "redness" are summarized as "visual disturbance."

•**DISEASE**: eye disease
- **Explanation**: Both "glaucoma" and "diabetic retinopathy" are classified as "eye disease."

User Prompt:

Now, merge the provided summaries accordingly by finding similarities, referring to shared characteristics, using common parent categories where appropriate, and ensuring all information is presented clearly and concisely under each category.

Tag Summary 1:
{Tag Summary 1}
Tag Summary 2:
{Tag Summary 2}

Figure 26: tag-summary merge prompt, part-2.

System Prompt:

You are tasked with answering a medical question using the provided retrieved medical graph. The graph has a structured format: [entity content] [relationship] [entity content]. Your response should be well-reasoned and backed by information from the medical graph, ensuring clarity and accuracy.

Instructions:

1.Graph-Based Reasoning: Use the information from the provided graph structure to reason and generate an accurate response to the medical question. Ensure that all relevant entities and relationships are used effectively to construct a well-supported answer.

2.Reference Inclusion:

1. If the response uses information from a reference, include the **reference DOI** in the appropriate position within the response to indicate sourcing or backing of the statement.

3.Clarification of Clinical Terms:

1. For specific terms that may be **obscure** or require additional context, include the **CUI (Concept Unique Identifier)** after the term to assist individuals without a clinical background in understanding.

4.Formatting:

1. Ensure that the response is **clear**, **well-structured**, and **concise**.
2. Cite DOIs and CUIs at relevant points to enhance understanding and credibility.

Output Example:

•Response should be in a natural conversational tone, ensuring readability.

•Example of reference inclusion: "The treatment has been shown effective for hypertension [DOI:10.1234/example]."

•Example of CUI inclusion: "Patients may experience tachycardia (CUI: C0004238)."

User Prompt:

Answer the user question: {QUESTION} using the graph: {GRAPH}.

Figure 27: Prompt to response based on retrieved graph.

System Prompt:

You are tasked with adjusting a previous response to a medical question using an updated, structured summary based on predefined categories. The updated summary provides additional or clarified information, and your goal is to integrate these changes while preserving the structure and quality of the original response.

Instructions:

1.Integration of Updated Information:
1. Carefully integrate relevant details from the structured summary into the previous response. The summary is organized by categories, and the adjustments should use this categorized information to enhance or update specific aspects of the answer.
2. Ensure consistency between the original response and the new information provided.

2.Reference Inclusion:
1. Include the **reference DOI** at the appropriate point if the updated or newly added content uses information from a reference.
2. The DOI should be positioned where the reference supports reasoning, backing, or sourcing within the response.

3.Term Clarification:
1. For any **obscure terms** that may need clarification for a non-clinical audience, include the **CUI (Concept Unique Identifier)** after the term. This will help improve understanding for readers with less clinical background.

4.Formatting:
1. Ensure the adjusted response remains **clear**, **well-structured**, and **concise**.
2. Maintain the conversational tone of the original response, incorporating new references and CUIs where needed.

User Prompt:

Adjust the response: {model last response} of question: {user question} using the updated information: {tag summary}

Figure 28: Prompt to revise the response based on tag-summary.