MEDICAL GRAPH RAG: TOWARDS SAFE MEDICAL LARGE LANGUAGE MODEL VIA GRAPH RETRIEVAL-AUGMENTED GENERATION

Junde Wu
University of Oxford
jundewu@ieee.org

Jiayuan Zhu
University of Oxford

Yunli Qi
University of Oxford

ABSTRACT

We introduce a novel graph-based Retrieval-Augmented Generation (RAG) framework specifically designed for the medical domain, called **MedGraphRAG**, aimed at enhancing Large Language Model (LLM) capabilities and generating evidence-based results, thereby improving safety and reliability when handling private medical data. Our comprehensive pipeline begins with a hybrid static-semantic approach to document chunking, significantly improving context capture over traditional methods. Extracted entities are used to create a three-tier hierarchical graph structure, linking entities to foundational medical knowledge sourced from medical papers and dictionaries. These entities are then interconnected to form meta-graphs, which are merged based on semantic similarities to develop a comprehensive global graph. This structure supports precise information retrieval and response generation. The retrieval process employs a U-retrieve method to balance global awareness and indexing efficiency of the LLM. Our approach is validated through a comprehensive ablation study comparing various methods for document chunking, graph construction, and information retrieval. The results not only demonstrate that our hierarchical graph construction method consistently outperforms state-of-the-art models on multiple medical Q&A benchmarks, but also confirms that the responses generated include source documentation, significantly enhancing the reliability of medical LLMs in practical applications.

1 INTRODUCTION

The rapid advancement of large language models (LLMs), such as OpenAI's ChatGPT OpenAI (2023a) and GPT-4 OpenAI (2023b), has significantly transformed research in natural language processing and sparked numerous AI applications in everyday scenarios. However, these models still face limitations when applied to fields requiring specialized knowledge, such as finance, law, and medicine. There are two primary challenges: First, deploying trained LLMs for specific uses is complex due to their struggles with extremely long contexts and the high costs or impracticality of fine-tuning large models on specialized datasets. Second, in domains like medicine where precision is crucial, LLMs may produce hallucinations—outputs that seem accurate but lead to incorrect conclusions, which can be dangerous. Additionally, they sometimes provide overly simplistic answers without offering new insights or discoveries, which falls short in fields that demand high-level reasoning to derive correct answers.

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. Originally designed for situations where the necessary answers are found within specific text regions, RAG sometimes struggles to synthesize new insights from disparate pieces of information linked by shared attributes. Additionally, it underperforms in tasks requiring a holistic understanding of summarized semantic concepts across large datasets or extensive documents. To address these limitations, the graph RAG Hu et al. (2024) method has been introduced. This approach leverages LLMs to create a knowledge graph from the private dataset, which, in conjunction with graph machine learning, enhances prompt augmentation during query processing. GraphRAG has demonstrated significant improvements, outperforming previous methods applied to private datasets by offering greater intelligence and mastery in information synthesis.

In this paper, we introduce a novel graph RAG method for applying LLMs to the medical domain, which we refer to as Medical Graph RAG (MedRAG). This technique improves LLM performance in the medical domain by response queries with grounded source citations and clear interpretations of medical terminology, boosting the transparency and interpretability of the results. This approach involves a three-tier hierarchical graph construction method. Initially, we use documents provided by users as our top-level source to extract entities. These entities are then linked to a second level consisting of more basic entities previously abstracted from credible medical books and papers. Subsequently, these entities are connected to a third level—the fundamental medical dictionary graph—that provides detailed explanations of each medical term and their semantic relationships. We then construct a comprehensive graph at the highest level by linking entities based on their content and hierarchical connections. This method ensures that the knowledge can be traced back to its sources and the results are factually accurate.

To respond to user queries, we implement a U-retrieve strategy that combines top-down retrieval with bottom-up response generation. The process begins by structuring the query using predefined medical tags and indexing them through the graphs in a top-down manner. The system then generates responses based on these queries, pulling from meta-graphs—nodes retrieved along with their TopK related nodes and relationships—and summarizing the information into a detailed response. This technique maintains a balance between global context awareness and the contextual limitations inherent in LLMs.

Our medical graph RAG provides Intrinsic source citation can enhance LLM transparency, interpretability, and verifiability. The results provides the provenance, or source grounding information, as it generates each response, and demonstrates that an answer is grounded in the dataset. Having the cited source for each assertion readily available also enables a human user to quickly and accurately audit the LLM's output directly against the original source material. It is super useful in the field of medicine that security is very important, and each of the reasoning should be evidence-based. By using such a method, we construct an evidence-based Medical LLM that the clinician could easiely check the source of the reasoning and calibrate the model response to ensure the safty usage of llm in the clinical senarios.

To evaluate our medical graph RAG, we implemented the method on several popular open and closed-source LLMs, including ChatGPT OpenAI (2023a) and LLaMA Touvron et al. (2023), testing them across mainstream medical Q&A benchmarks such as PubMedQA Jin et al. (2019), MedMCQA Pal et al. (2022), and USMLE Kung et al. (2023). For the RAG process, we supplied a comprehensive medical dictionary as the foundational knowledge layer, the UMLS medical knowledge graph Lindberg et al. (1993) as the foundamental layer detailing semantic relationships, and a curated MedC-K dataset Wu et al. (2023) —comprising the latest medical papers and books—as the intermediate level of data to simulate user-provided private data. Our experiments demonstrate that our model significantly enhances the performance of general-purpose LLMs on medical questions. Remarkably, it even surpasses many fine-tuned or specially trained LLMs on medical corpora, solely using the RAG approach without additional training.

Our contributions are as follows:

1. We are pioneers in proposing a comprehensive pipeline for applying graph RAG specifically in the medical field.

2. We have developed unique graph construction and data retrieval methods that enable LLMs to generate evidence-based responses utilizing holistic private data.

3. We conducted validation experiments across mainstream benchmarks, achieving state-of-the-art (SOTA) performance with various model variants.

2 METHOD

MedGraphRAG enhances Large Language Models (LLMs) with a medical graph RAG tailored for handling private medical data. It involves segmenting medical documents into chunks, extracting entities, and organizing them into a hierarchical graph structure across three levels—from user-provided documents to foundational medical information. These entities form meta-graphs, which are then merged based on content similarity into a comprehensive global graph. For user queries, the LLM retrieves and synthesizes information efficiently from the graph, enabling precise and contextually relevant medical responses.

2.1 Medical Graph Construction

Semantic Document Segmentation Large medical documents often contain multiple themes or diverse content. To process these effectively, we first segment the document into data chunks that conform to the context limitations of Large Language Models (LLMs). Traditional methods such as chunking based on token size or fixed characters typically fail to detect subtle shifts in topics accurately. Consequently, these chunks may not fully capture the intended context, leading to a loss in the richness of meaning.

To enhance accuracy, we adopt a mixed method of character separation coupled with topic-based segmentation. Specifically, we utilize static characters (line break symbols) to isolate individual paragraphs within the document. Following this, we apply a derived form of the text for semantic chunking. Our approach includes the use of proposition transfer, which extracts standalone statements from a raw text Chen et al. (2023). Through proposition transfer, each paragraph is transformed into self-sustaining statements. We then conduct a sequential analysis of the document to assess each proposition, deciding whether it should merge with an existing chunk or initiate a new one. This decision is made via a zero-shot approach by an LLM. To reduce noise generated by sequential processing, we implement a sliding window technique, managing five paragraphs at a time. We continuously adjust the window by removing the first paragraph and adding the next, maintaining focus on topic consistency. We set a hard threshold that the longest chunk cannot excess the context length limitation of LLM. After chunking the document, we construct graph on each individual of the data chunk.

Element Extraction We then identify and extract instances of graph nodes from each chunk of source text. This is accomplished using a LLM prompt designed to recognize all relevant entities within the text. For each entity, the LLM is prompted to output the name, type, and a description. The name may either be the exact text from the document or a derivative term commonly used in medical contexts, carefully chosen to reflect professional medical terminology suitable for subsequent processing. The type is selected from a predefined table by the LLM, and the description is an LLM-generated explanation of the entity, contextualized within the document. To ensure the model's effectiveness, we provide a few examples to guide the LLM in generating outputs as desired.

For each entity data structure, we include a unique ID to trace its source document and paragraph. This identifier is crucial for retrieving information from the source, enabling the generation of evidence-based responses later.

To enhance the quality of extraction and reduce noise and variance, we repeat the extraction process multiple times. This iterative approach encourages the LLM to detect any entities it may have initially overlooked. The decision to continue or halt the repetitive process is also determined by the LLM itself.

Hierarchy Linking Medicine is a specialized field characterized by its consistent use of a precise terminology system and its foundation on numerous established truths, such as specific symptoms of diseases or side effects of drugs. In this domain, it is crucial that LLMs do not distort, modify, or add creative or random elements to the data, unlike their applications in other, less constrained contexts.

Recognizing this, we have developed a unique structure within the medical domain to link each entity to grounded medical knowledge and terms. This approach aims to provide credible source and profound definitions for each entity concept, thereby enhancing the authenticity of the responses and reducing the occurrence of hallucinations, a significant challenge when applying LLMs to medicine.

Specifically, we construct a three-tiered RAG data structure to develop a comprehensive medical graph. The first level consists of user-provided documents, such as highly confidential medical reports from a specific hospital. After extracting entities from these documents as previously described, we link them to a more foundational level of commonly accepted information. The second level is constructed using medical textbooks and scholarly articles. We pre-construct a graph from these medical sources using the same methods outlined earlier, prior to receiving real user documents. Entities from the first level are linked to corresponding entities in the second level, based on relevance detected by LLMs.

The entities of the second-level graph are then connected to a third level, which includes several well-defined medical terms and their knowledge relationships. This grounded information is sourced from reliable resources such as the Unified Medical Language System (UMLS), which integrates various health and biomedical vocabularies and their semantic relationships. For each entity, we compare the text embedding of its name with those of medical vocabularies in UMLS, selecting vocabularies where the cosine similarity falls below a specified threshold. Each linked vocabulary is further associated with its professional definitions and relationships in UMLS, and these relationships are translated into plain text, as demonstrated in Wu et al. (2023).

Relationship Linking We then instruct the LLM to identify all relationships between clearly-related entities. This decision is based on comprehensive information about each entity, including its name, description, definition, and associated lower-level medical foundation knowledge. The identified relationships specify the source and target entities, provide a description of their relationship, and include a score indicating the closeness of this relationship. To maintain order and precision in assessing relationship distance, we prompt the LLM to choose from a predefined list of descriptors—very related, related, medium, unrelated, very unrelated. After performing this analysis, we generate a weighted directed graph for each data chunk. These graphs serve as the fundamental building blocks in our system and are referred to as meta-graphs.

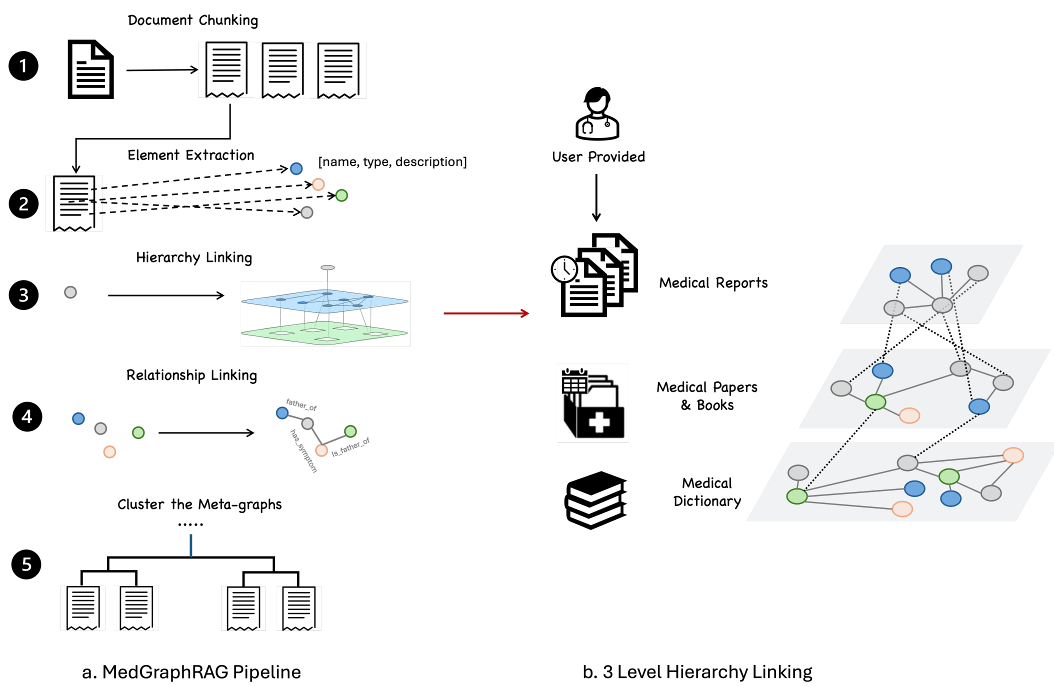

Figure 1: MedGraphRAG framework.

Tags generation and merge the graphs After constructing the meta-graphs, our next step is to scan the data across each chunk to develop a global graph that links all meta-graphs together. The nodes in merged meta-graphs would link together based on the linking rule we used in the last paragraph. To achieve this, we calculate the distance between each pair of meta-graphs and sequentially merge the closest ones into larger entities. For efficient merging, we use the LLM to summarize the content of each meta-graph based on predefined medical categories, such as symptoms, patient history, body functions, and medications. The LLM generates a summary for each category derived from the meta-graph's content, resulting in a list of tags that succinctly describe its main themes.

Using these tags, the LLM calculates the similarity between two meta-graphs. Those with the highest similarity are considered for merging. The merged graph becomes a new graph, but retains its original meta-graphs and tags for easier indexing later. Subsequently, new summarized tag information is generated for the new graph, and its similarity to others is recalculated for potential further merging. This process can be repeated until a single global graph remains. However, as the summarized tag information accumulates, it loses detail, presenting a trade-off between merging effectiveness and efficiency. In practice, we limit the process to 24 iterations to prevent excessive loss of detail.

2.2 Retrieve from the graph

After constructing the graph, the LLM efficiently retrieves information to respond to user queries through a strategy we called U-retrieve. We begin by generating summarized tag descriptions, similar to the previous step, and use these to identify the most relevant graph through a top-down matching process. This starts with one of the larger graphs, progressively indexing down to the smaller graphs it contains. This matching process is repeated until we reach the meta-graph layer and retrieve multiple relevant entities. Subsequently, we gather all pertinent content related to these activated entities and their TopK related entities. This includes the content of the entities themselves, their associated foundational medical knowledge, their relevance and relationships to other entities, and the content of any linked entities.

Once the relevant content is identified, the LLM is prompted to generate an intermediate response using these information, presented in text form. This intermediate response is preserved and combined with the higher-level graph's summarized tag information to formulate a more detailed or refined response. The LLM repeats this response generation process in a bottom-up manner until it reaches the highest level, generating a final response after scanning all the indexed graphs along the trajectory. This method allows the LLM to have a comprehensive overview, as it interacts with all the data in the graph, while also remaining efficient by accessing less relevant data in a summarized form.

3 Experiment

3.1 Dataset

3.1.1 RAG data

In our RAG data structure, we design three distinct levels of data, each serving different roles in practice. The top-level data comprises private user information, such as medical reports in a hospital, which are confidential and not to be shared or exposed. This data is user-specific and subject to the highest frequency of updates or changes when using the LLM in a practical setting. The middle level consists of up-to-date, peer-reviewed, and credible medical books and papers. This layer provides users with the latest medical advancements and knowledge, ensuring they do not miss any cutting-edge discoveries. While these resources can be set as default data for different users, they can also be updated regularly by users or administrators to maintain currency. This data is updated at a medium frequency, typically annually. The bottom level includes data that define medical terms and their semantic relationships, primarily sourced from established vocabularies. This data is the most authoritative and serious and should be set as the default for every user intending to use the medical LLM. It is updated with the lowest frequency, approximately every five years or more.

Top-level We employ MIMIC-IV, a publicly available electronic health record dataset, as our primary dataset. This dataset originates from Beth Israel Deaconess Medical Center and encompasses patient admissions spanning from 2008 to 2019. MIMIC-IV is designed to facilitate research and educational pursuits, encompassing a wide range of data including patient measurements, diagnoses, procedures, treatments, and anonymized clinical notes. This dataset is the product of a collaborative effort between the hospital and MIT, and is meticulously gathered, processed, and deidentified to comply with privacy standards. It is structured into three distinct modules—hospital, intensive care unit, and clinical notes—each specifically designed to meet various research requirements.

Medium-level We utilized MedC-K, a substantial medical-specific corpus, as our medium-level data source. This corpus comprises 4.8 million biomedical academic papers and 30,000 textbooks. It includes the S2ORC dataset by Lo et al. (2020), which contains 81.1 million English-language academic papers. From this extensive collection, we have extracted 4.8 million papers related to biomedical studies from PubMed Central, amounting to over 75 billion tokens that encapsulate advanced medical knowledge. Additionally, we curated a collection of 30,000 medical textbooks from various libraries and publishers. After thorough cleaning and de-duplication processes, this collection provides approximately 4 billion tokens of essential medical knowledge.

Bottom-level We utilize the UMLS dataset as our foundational bottom-level data. The Unified Medical Language System (UMLS), developed by the U.S. National Library of Medicine, is an extensive dataset that unifies various medical vocabularies to enhance the interoperability of health information systems. It consists of three main components: the Metathesaurus, which amalgamates over 200 medical vocabularies including SNOMED CT and ICD-10; the Semantic Network, which organizes medical concepts and delineates their interrelationships; and the SPECIALIST Lexicon, which aids in natural language processing by providing detailed linguistic insights. UMLS is crucial for facilitating tasks such as electronic health record integration and clinical decision support, thereby improving the management and comprehension of medical data.

3.1.2 TEST DATA

PubMedQA Developed by Jin et al. in 2019, PubMedQA is a biomedical question-answering dataset derived from PubMed abstracts. This dataset primarily focuses on addressing research questions through a multiple-choice format with options like yes, no, or maybe. It comprises three distinct parts: the PQA-L, which includes 1,000 manually labeled pairs used for testing; PQA-U, consisting of 61.2k unlabeled pairs which are not used; and PQA-A, featuring 211.3k artificially generated pairs.

MedMCQA Introduced by Pal, Umapathi et al. in 2022, MedMCQA is a dataset of multiple-choice questions formulated from practice and previous examinations for Indian medical school entrance tests, specifically the AIIMS and NEET-PG. The dataset splits into a training set with 182,822 questions and a testing set containing 4,183 questions, each question offering four possible answers. This dataset serves as a significant resource for testing knowledge of medical school candidates.

USMLE Created by Jin, Pan et al. in 2021, the USMLE dataset consists of multiple-choice questions from the United States Medical Licensing Exams, tailored to assess medical professionals' readiness for board certification. This dataset is unique in its multilingual coverage, providing questions in English, Simplified Chinese, and Traditional Chinese. For the purpose of this description, only the English portion is considered, which includes 10,178 + 1,273 + 1,273 pieces of data.

3.2 LLM MODELS

LLAMA2 Building upon the original LLAMA dataset, LLAMA2 extends the evaluation framework by including more diverse and complex language tasks, potentially addressing the limitations and gaps identified in the initial version. Although specific details on LLAMA2 might be hypothetical or speculative in nature, one can expect that it would continue the focus on robust, comprehensive linguistic analysis, refining the tools and methods to better measure nuances in language understanding and generation.

LLAMA3 LLAMA3 is the latest iteration in the LLAMA series of large language models, developed to advance the capabilities of natural language understanding and generation. Building on the successes of its predecessors, LLAMA and LLAMA2, LLAMA3 incorporates even more sophisticated algorithms and a broader dataset to enhance its performance across a wide array of linguistic tasks.

GPT-4 Developed by OpenAI, ChatGPT-4 is an iteration of the generative pre-trained transformer models that has been trained on a diverse range of internet text. As a more advanced version, ChatGPT-4 features improvements over its predecessors in terms of its ability to understand and generate human-like text, making it capable of engaging in more coherent and contextually relevant conversations. This model is designed to perform a wide range of tasks including but not limited to translation, question-answering, and content generation, showcasing significant advancements in handling complex dialogue scenarios and nuanced language subtleties.

Gemini Google's Gemini is a cutting-edge language model designed to enhance the capabilities of conversational AI systems. Developed as part of Google's ongoing efforts in natural language processing, Gemini aims to provide more nuanced and context-aware interactions than previous models. This model leverages deep learning techniques to understand and generate human-like responses, making it suitable for a wide range of applications including virtual assistants, customer support, and interactive applications.

3.3 Results

3.3.1 Medical Graph RAG Effect

First, we conducted experiments to assess the impact of our Medical Graph RAG on various large language models, with the results presented in Table 1. The data reveals that our MedGraphRAG significantly enhances the performance of LLMs on medical benchmarks. This improvement is attributed to the implementation of zero-shot RAG, which is more cost-effective, faster, and more convenient than fine-tuning or using adapters. Notably, MedGraphRAG yields more substantial improvements in smaller LLMs, such as LLaMA2-13B and LLaMA3-8B, which typically underperform on these benchmarks, thus broadening its applicability across a wider user base. MedGraphRAG also significantly boosts the performance of more powerful, closed-source LLMs like GPT and LLaMA3-70B, helping them achieve state-of-the-art (SOTA) results on multiple benchmarks. These outcomes surpass the accuracy of human experts, demonstrating the strong potential of AI to enhance clinical workflows.

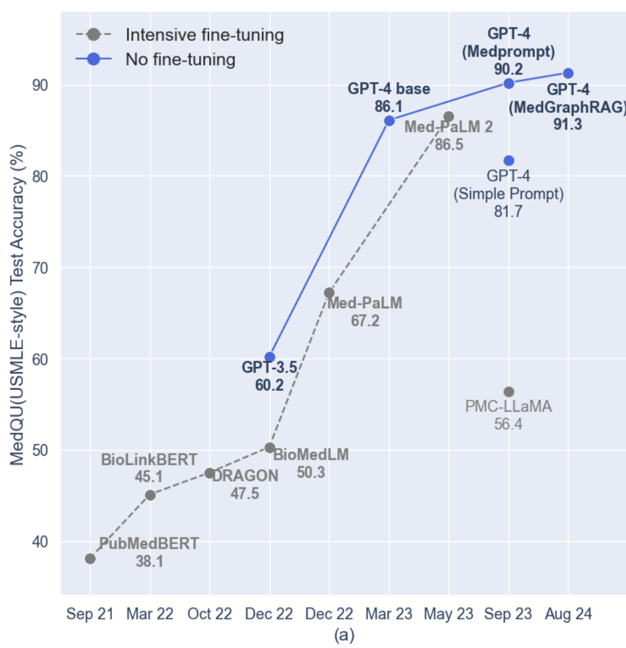

Figure 2: Compare to SOTA Medical LLM Models on MedQA benchmark.

3.3.2 Evidence-based response

Thanks to the graph linking mechanism in our MedGraphRAG, we can prompt LLMs to generate evidence-based responses to complex medical questions, enhancing both safety and explainability. As illustrated in Figure 3, we compare the responses generated by GPT-4 alone and those enhanced by MedGraphRAG for a challenging medical diagnosis question. In this case, the patient exhibits

Table 1: The improvement of MedGraphRAG on various LLMs.

Model	Size	Open-sourced	MedQA	MedMCQA	PubMedQA
LLaMA2	13B	yes	42.7	37.4	68.0
LLaMA2-MedGraphRAG	13B	yes	65.5	51.4	73.2
LLaMA2	70B	yes	43.7	35.0	74.3
LLaMA2-MedGraphRAG	70B	yes	69.2	58.7	76.0
LLaMA3	8B	yes	59.8	57.3	75.2
LLaMA3-MedGraphRAG	8B	yes	74.2	61.6	77.8
LLaMA3	70B	yes	72.1	65.5	77.5
LLaMA3-MedGraphRAG	70B	yes	88.4	79.1	**83.8**
Gemini-pro	-	no	59.0	54.8	69.8
Gemini-MedGraphRAG	-	no	72.6	62.0	76.2
GPT-4	-	no	81.7	72.4	75.2
GPT-4 MedGraphRAG	-	no	**91.3**	**81.5**	83.3
Human (expert)	-	-	87.0	90.0	78.0

symptoms commonly associated with Alzheimer's—increasing forgetfulness and occasional episodes of sudden confusion and speech difficulty. However, a careful analysis by an experienced human expert would identify the condition as Vascular Dementia. The MedGraphRAG-enhanced response not only accurately identifies Vascular Dementia over Alzheimer's but also provides detailed explanations supported by authentic citations. This ensures that each claim is verifiable, making the information trustworthy for clinicians. Additionally, the response includes simplified explanations of medical terms, making it accessible to users without a medical background. This evidence-based, user-friendly approach is crucial in clinical practice where safety is paramount.

3.3.3 Compare to SOTA Medical LLM Models

We also evaluated MedGraphRAG against a range of previous state-of-the-art (SOTA) 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. (2023)Wu et al. (2023) and non-fine-tuned models Nori et al. (2023)OpenAI (2023a)OpenAI (2023b) on the MedQA benchmark. The results, depicted in Figure 2, show that when applied to a powerful GPT-4 LLM, our MedGraphRAG surpasses the previous SOTA prompted model, Medprompt Nori et al. (2023), by a significant 1.1%. Even when compared with intensive fine-tuning methods on these medical datasets, MedGraphRAG outperforms all and achieves the SOTA. This superior performance stems from leveraging the inherent capabilities of the robust GPT-4 model. This further underscores the advantages of our non-fine-tuned MedGraphRAG approach: it inherits the strong capabilities of a closed-source model and outperforms many models that require costly and exhaustive fine-tuning.

3.3.4 Ablation Study

We conducted a comprehensive ablation study to validate the effectiveness of our proposed modules, the results of which are presented in Table 2. This study compares various methods for document chunking, hierarchy graph construction, and information retrieval. Specifically, for document chunking, we evaluated our hybrid static-semantic method against a purely static approach. For hierarchy graph construction, we contrasted our method with the basic construction approach used in LangChain. For information retrieval, we compared the summarized-based retrieval Edge et al. (2024) with our U-retrieve method. These methods were assessed across the three Q&A benchmarks previously mentioned.

The results, as shown in the table, indicate that our hybrid semantic method significantly enhances performance over the vanilla model, underscoring the importance of sophisticated data chunking in all RAG pipelines. When comparing the base graph construction method with our proposed hierarchical approach, it's clear that constructing the graph enhances RAG performance. Furthermore, our hierarchical graph construction technique yields the most significant improvements, surpassing the performance of most state-of-the-art (SOTA) methods. Additionally, replacing the summarized retrieval with our U-retrieve method further boosts performance, demonstrating the effectiveness of U-retrieve in enhancing retrieval accuracy and relevance.

Table 2: An ablation study on MedGraphRAG.

Doc. Chunking		Graph Construction		Retrieve		MedQA	MedMCQA	PubMedQA
Static	Hyb-Semantic	Base	Hierarchy	SumR	UR	(%)	(%)	(%)
✓				✓		83.6	74.2	75.8
	✓			✓		87.4	77.2	77.9
	✓	✓		✓		88.8	78.7	80.6
	✓		✓	✓		90.7	80.8	82.5
	✓	✓	✓		✓	**91.3**	**81.5**	**83.3**

4 CONCLUSION

In conclusion, this paper introduces MedGraphRAG, a novel graph-based Retrieval-Augmented Generation (RAG) framework for the medical domain, enhancing the capabilities of Large Language Models (LLMs). Our method combines advanced document chunking with a hierarchical graph structure, significantly improving data organization and retrieval accuracy. Our ablation studies confirm superior performance over state-of-the-art models on medical Q&A benchmarks and provide credible, source-linked responses essential for medical applications. Moving forward, we aim to expand this framework to include more diverse datasets and explore its potential in real-time clinical settings.

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

REFERENCES

Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. Biomedlm, 2022. URL https://crfm.stanford.edu/2022/12/15/biomedlm.html. Stanford Center for Research on Foundation Models.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. Dense X Retrieval: What Retrieval Granularity Should We Use?, December 2023. URL http://arxiv.org/abs/2312.06648. arXiv:2312.06648 [cs].

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From Local to Global: A Graph RAG Approach to Query-Focused Summarization, April 2024. URL http://arxiv.org/abs/2404.16130. arXiv:2404.16130 [cs].

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, January 2022. ISSN 2691-1957, 2637-8051. doi: 10.1145/3458754. URL http://arxiv.org/abs/2007.15779. arXiv:2007.15779 [cs].

Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. GRAG: Graph Retrieval-Augmented Generation, May 2024. URL http://arxiv.org/abs/2405.16506. arXiv:2405.16506 [cs].

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A Dataset for Biomedical Research Question Answering, September 2019. URL http://arxiv.org/abs/1909.06146. arXiv:1909.06146 [cs, q-bio].

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198, February 2023. ISSN 2767-3170. doi: 10.1371/journal.pdig.0000198. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9931230/.

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. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, April 2021. URL http://arxiv.org/abs/2005.11401. arXiv:2005.11401 [cs].

D. A. Lindberg, B. L. Humphreys, and A. T. McCray. The Unified Medical Language System. *Methods of Information in Medicine*, 32(4):281–291, August 1993. ISSN 0026-1270. doi: 10.1055/s-0038-1634945.

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. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine, November 2023. URL http://arxiv.org/abs/2311.16452. arXiv:2311.16452 [cs].

OpenAI. Openai. introducing chatgpt. https://openai.com/blog/chatgpt/, 2023a.

OpenAI. Gpt-4 technical report, 2023b.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering, March 2022. URL http://arxiv.org/abs/2203.14371. arXiv:2203.14371 [cs].

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. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

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. Towards expert-level medical question answering with large language models, 2023.

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. LLaMA: Open and Efficient Foundation Language Models, February 2023. URL http://arxiv.org/abs/2302.13971. arXiv:2302.13971 [cs].

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-LLaMA: Towards Building Open-source Language Models for Medicine, August 2023. URL http://arxiv.org/abs/2304.14454. arXiv:2304.14454 [cs].

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. Deep Bidirectional Language-Knowledge Graph Pretraining, October 2022a. URL http://arxiv.org/abs/2210.09338. arXiv:2210.09338 [cs].

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. LinkBERT: Pretraining Language Models with Document Links, March 2022b. URL http://arxiv.org/abs/2203.15827. arXiv:2203.15827 [cs].