To Enhance Graph-Based Retrieval-Augmented Generation (RAG) with Robust Retrieval Techniques

1st Maneeha Rani
DAIM
University of Hull
Hull, United Kingdom
m.rani3-2022@hull.ac.uk

2nd Bhupesh Kumar Mishra
DAIM
University of Hull
Hull, United Kingdom
bhupesh.mishra@hull.ac.uk

3rd Dhavalkumar Thakker
School of Computer Science
University of Hull
Hull, United Kingdom
D.Thakker@hull.ac.uk

4th Mohammad Nouman Khan
Pak-Austria Fachhochschule (PAF-IAST)
Haripur, Pakistan
iammuhammadnoumankhan@gmail.com

2024 18th International Conference on Open Source Systems and Technologies (ICOSST) | 979-8-3315-0869-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICOSST64562.2024.10871140

***Abstract*—Large language models have demonstrated exceptional performance in multiple domains. However, practical deployment in the healthcare sector has distinctive challenges. These challenges include hallucination, inconsistency, explainability, reasoning, authenticity, and validity of information sources. Hallucinations in LLM often emerge due to unstructured and obsolete training data and the incompetence to upgrade the model data post-training. Retrieval-augmented generation (RAG) integration with LLM decision-making helps access real-time information from external resources. However, further improvements are needed to improve accurate response generation. A knowledge Graph is a structured data comprising nodes as entities and edges as relationships. When integrated with RAG, Knowledge Graph-based retrieval offers better contextually relevant responses, traceability, and explainability of generated responses than RAG alone. This study proposes a novel knowledge graph-based RAG framework with a refined retrieval pipeline, robust chunking mechanism, and source traceability for enhanced diabetes-focused LLM. The retrieval pipeline integrates three robust retrieval strategies: keyword, graph, and vector. To ensure the authenticity of responses, a knowledge base focusing on diabetes is designed from validated sources. This verified knowledge base is preprocessed and converted to a knowledge graph to design A graph-based RAG pipeline. The empirical results demonstrate effective performance in diabetes-focused LLM, achieving a Rouge 1 score of 82.19%.**

***Index Terms*—Retrieval-augmented generation, Large language model, Knowledge graph, Diabetes, Healthcare, Graph-based Retrieval-augmented generation**

I. Introduction

Large language models (LLMs) are revolutionizing Artificial Intelligence (AI) by producing coherent text and an in-depth understanding of human language across various contexts [1]. LLMs are deployed across multiple domains, including healthcare, where factual information accuracy can significantly influence patient care[2]. Despite LLM's significant potential, practical deployment in healthcare brings considerable challenges [3]. Some general concerns include hallucination, explainability, accuracy, reliability, fetching context-specific information, and managing unstructured data [4]. Once fed to LLM, the knowledge becomes frozen, and the upgradation demands retraining[5]. Healthcare is rapidly evolving, and the latest research is emerging daily, demanding the integration of new research in already trained LLM [6]. Retaining an existing LLM to incorporate emerging research is resource-intensive and complex[7]. This has resulted in only a few disease-specific LLMs. The available disease-specific LLMs are outdated, use information from unverified sources, or are limited in scope[8]. Apart from the latest knowledge integration, existing medical literature is unstructured.

Retrieval Augmented Generation (RAG) retrieves relevant information from external resources using retrieval strategies to improve LLM's performance [9].RAG approach uses keyword search and vector sector to find contextually relevant textual information from external knowledge bases to generate coherent responses [10]. However, the lack of structured relationships between concepts is a challenge that results in a complex information retrieval process in traditional RAG approaches [11]. Traditional RAG approaches excel in specific domains, but the results obtained from keyword search and vector search can not be deployed as a standalone solution for healthcare. Further, the token-based chunking strategies used in traditional RAG may lose critical information, causing considerable risk for the healthcare sector [12]. The retrieval process could be more resource-intensive and less accurate when processing extensive unstructured data. Moreover, the accuracy in tracing the sources of information demands further improvements to ensure the explainability of generated responses[13]. Explainability refers to the capability of a model to justify the generated response, which is considered necessary in Healthcare [14] [15]. A knowledge graph can capture and organize relationships between concepts. The information in the knowledge graph is structured and capable of integrating the latest external knowledge, traceability, retrieval, and explainability. Knowledge graph-based RAG has the potential to address these challenges. This research focuses on healthcare, explicitly targeting diabetes care and control to

979-8-3315-0869-2/24/$31.00 ©2024 IEEE

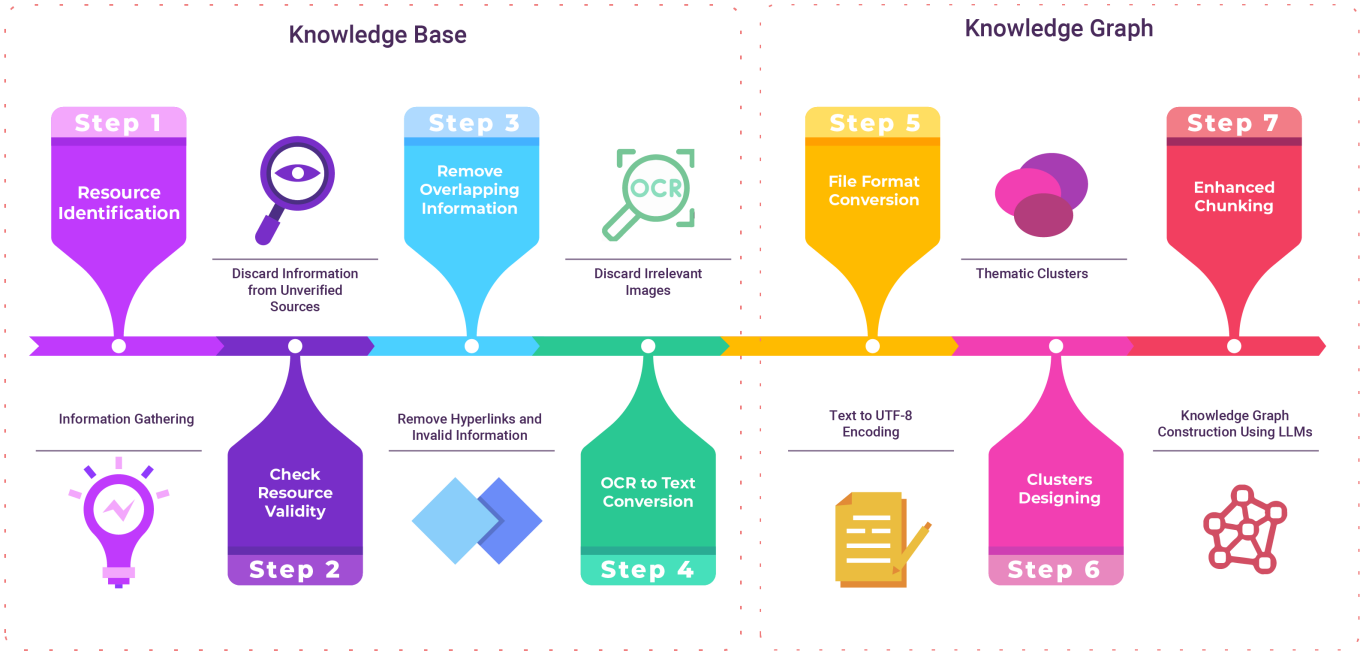

Fig. 1. Preprocessing steps for constructing the knowledge base and designing the knowledge graph

design a reliable, robust, disease-specific LLM using a Knowledge Graph. A knowledge base is created by gathering data from validated sources to construct the knowledge graph. A novel Graph RAG-based framework is proposed that integrates three retrieval strategies: keyword retrieval, vector retrieval, and graph retrieval. This integration of retrieval strategies provides an opportunity to use the strengths of all retrieval algorithms while addressing their limitations. It improves existing token-based chunking with more advanced topic-based chunking. Further, it adds a traceability mechanism to trace the source of generated answers. Overall, this combined context from three retrieval strategies helps LLM to develop more accurate answers. This paper is arranged as follows: Section 1 covers the introduction; Section 2 provides prior contributions to the field. Section 3 discusses the proposed methodology, and section 4 details experiments and evaluations. Section 4 presents the conclusion.

II. Literature Review

RAG is used to fetch contextually relevant content for an LLM. The literature in RAG is more inclined towards Graph RAG, which improves factual accuracy[16]. Graph RAG can fetch relevant subgraphs from a more extensive knowledge graph. In the Graph RAG framework, the research targets open-ended questions by improving the fetching accuracy of semantically relevant documents [17], [18]. In a recent study, KRAGEN was proposed, which uses a graph of thoughts for reasoning and response generation focused specifically on Alzheimer's disease[19]. This approach uses a similar mechanism to fetch information from a knowledge graph and generate a response. Because of its effectiveness, this research has been suggested for use in the medical domain. Further, this approach offers a graph-based data conversion from unstructured to structured format. However, the accuracy of reasoning-based questions needs improvements, resulting in 62.4%, considered borderline for the healthcare sector. Hybrid RAG[20] is another similar approach that integrates graph RAG and vector RAG to generate better responses for financial documents. The experimental results prove its applicability to the economic domain. It uses standard chunking and traceability and doesn't include keyword search, differentiating this approach from ours.

A novel algorithm named the Chain of Explorations (CoE) is proposed for a KG-based RAG framework [21]. It designs a knowledge graph to fetch relevant information and explore nodes and relationships sequentially. The empirical results and experimentation suggest reduced hallucinated content. A graph-based RAG called MedGraphRAG is proposed to produce an evidence-based information retriever[22]. This uses a static semantic chunking approach to improve the context of retrieved information. A three-tiered comprehensive graph structure is designed based on semantic similarity. It employs U-retrieval to retrieve information precisely. The results demonstrate enhanced reliability on Graph LLM for the medical domain. These two approaches use knowledge graphs but slightly different retrieval mechanisms than ours. Another approach uses tree-structured RAG [23] to represent an organization's entity hierarchy. Another approach in [24]

presents a flexible and customizable approach to integrating a knowledge graph for adjustable but contextually relevant retrieval according to variable situations. Experimental evaluations prove its applicability in various domains. G-retrieval [25] is another similar approach that uses knowledge graphs to retrieve nodes and edges. This approach uses indexes to process queries efficiently. A subgraph with all possible nodes is constructed and fetched depending on the query. Based on the subgraph, the answer is designed. Empirical results demonstrate scalability, reduced hallucinations, and applicability to multiple domains.

The above-discussed approaches use a Knowledge Graph-based RAG pipeline. However, none alone provides three different retrieval strategies, improved chunking, and robust back traceability. This distinguishes our research from existing studies. Some studies emphasize better retrieval strategies but sacrifice refined chunking. Additionally, developing a knowledge base from legitimate and validated sources for various sub-domains of diabetes coupled with knowledge graph construction is another significant achievement.

III. Methodology

This study aims to improve the LLM's capabilities in healthcare, specifically in Diabetes care and control. This section outlines the multi-step methodology employed in this paper, encompassing data collection from valid sources and preprocessing, domain-specific knowledge graph creation, a novel retrieval mechanism, and implementing a RAG framework. The knowledge graph and RAG framework are deployed to perform structured and unstructured data retrieval to enhance the relevance of retrieved answers by employing knowledge graphs, ensuring explainable, reasoning-oriented LLMs. Detailed preprocessing steps for creating a knowledge base and knowledge graphs are illustrated in Figure 1.

A. Knowledge Base

Healthcare is considered the primary domain due to the critical need for accurate information. Within healthcare, this research is particularly focused on diabetes, which demands precise knowledge. Three core domains of diabetes are identified, which include diabetes care, control, and management. These domains are further subdivided into sub-domains to ensure comprehensive coverage. These subdomains include diabetes awareness, monitoring, nutrition guidance, exercise, physical activities, medication, complications, and patient support. This hierarchical categorization of the diabetes-related information captures in-depth details of significant information, which is necessary for enhanced diabetic-focused LLM. To guarantee the credibility of the newly designed knowledge base, authentic resources are identified, and resource validity is examined. Relevant data is gathered from reputable websites and prevalidated reliable internet sources. The essential resources include Healthline, National Health Surcharge (NHS), research articles from Google Scholar, Diabetes Centre UK, the National Center for Biotechnology Information (NCBI), and the American Diabetes Association. The collected data is preprocessed by removing overlapping files and translating valuable information from images to text using optical character recognition (OCR) tools, discarding irrelevant images, and converting all information into UTF-8 encoded text. Further preprocessing steps are performed to discard unnecessary information, including hyperlinks and meaningless special characters. Considering the diverse sources of information, a systematic clustering approach is designed to streamline the information. All information files are scanned once again to extract and minimize overlapping details, if any. The extracted information is systematically organized into topic-specific clusters, each referring to an in-depth knowledge compilation related to a particular topic. This approach assures comprehensive and optimized knowledge graph construction by fetching information from well-defined clusters

```
Single-Nucleotide Polymorphisms - AFFECTS -> Diabetic Nephropathy
Capillaries - LEADS_TO -> Diabetic Nephropathy
Arterioles - LEADS_TO -> Diabetic Nephropathy
Genetic Factors - INFLUENCES -> Diabetic Nephropathy
Single-Nucleotide Polymorphisms - INFLUENCES -> Diabetic Nephropathy
Diabetes - COMPLICATION -> Diabetic Nephropathy
Diabetic Kidney Disease - CAUSE -> Diabetic Nephropathy
Contrast Media - INDUCES -> Diabetic Nephropathy
Esrd - COMPLICATION_OF -> Diabetic Nephropathy
Glomerulosclerosis - BEGINS_IN -> Diabetic Nephropathy
Diabetic Neuropathy - ASSOCIATED_WITH -> Peripheral Neuropathy
Diabetic Neuropathy - ASSOCIATED_WITH -> Diabetic Lumbosacral Plexopathy
```

Fig. 2. Graph Retriever, retrieving Entities and Relationships

B. Knowledge Graph

The scrapped knowledge base from the internet in the form of clusters is divided into chunks to structure systematically into a knowledge graph. Instead of a traditional token-based and character-based chunking strategy, this approach implements a topic-based chunking strategy to enhance accuracy. To automate the graph construction, GPT-3.5-turbo-0125 is used as a language model. LLM Graph Transformer is used to construct the knowledge graph from the clustered knowledge base. LLM Graph Transformer uses its semantic understanding to extract entities as nodes and relationships as links from text and designs a knowledge graph to be stored in a graph database. Neo4j is used as a graph database because of its capability to handle complex relationships, interactive graph visualizations for comparative analysis, and optimized graph traversals and algorithms. The newly designed knowledge graph has 6,107 nodes and 16,695 relationships. Each node is assigned an additional Entity label and source parameter to improve the indexing and query performance further. The source parameter links each node to its originating documents to generate evidence-based responses through traceability and context understanding.

C. Graph-based RAG

A graph-based retrieval augmented generation system is designed with three retrieval strategies: keyword search, vector search, and graph search. A user's query initiates the retrieval process. This query is passed to the RAG retriever, which applies three search strategies. Keyword search is performed on unstructured data, vector search is performed on vector

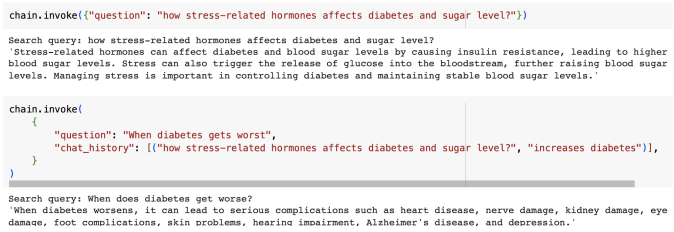

Fig. 3. Sample Questions and Generated Responses

embeddings, and graph search is performed on structured knowledge graph data. Keyword search finds semantically similar words, whereas vector search uses embeddings to find textual similarity between concepts. Graph retriever extracts relevant entities based on the full-text index. Extracted entities from the question are mapped to the knowledge graph, which returns semantically relevant neighboring nodes. The extracted nodes and relationships are flattened as entity-relationship-¿entity, demonstrated in Figure 2. These retrieval mechanisms are applied to the Neo4j database as this supports all three searches. Vector Index is used for efficient retrieval calculated by embeddings. The final retriever module will integrate the information from all retrievers to generate the final context for passing it to LLM. The Similarity Search Method is used to calculate semantic relevancy based on questions asked by the user. The final retriever module will integrate the information from all retrievers, and the combined context will be passed to LLM. LLM will generate a final response using this combined context, as illustrated in Figure 3. This unified graph-based RAG framework in Figure 4 integrates multiple retrieval techniques for structured and unstructured data to deliver high-quality, contextually relevant answers to diabetes-related healthcare questions.

IV. Evaluation and Results

To evaluate the effectiveness of our approach, a set of 100 questions covering all sub-domains of diabetes care and management is designed. This question set covers broad categories of question types from diverse topics. These question-answer categories include summarization questions, classification questions, one-word answers, generative answers, brainstorming answers, information extraction answers, and open answers. To construct a benchmark dataset for results comparison, a human annotator from the healthcare domain provided answers to each question. Answers generated by Graph RAG are evaluated against human responses using Rouge 1 and Rouge L metrics. These are well-known metrics for measuring similarity and overlap. Rouge 1 measures the common unigrams between human-generated and Graph RAG-generated answers. Rouge L measures the longest common subsequence between the generated and reference text. The ROUGE 1 and ROUGE L score for 100 entries is 82.19% and 71.34%, respectively. The same set of questions is used to generate responses using traditional RAG to make a comparative analysis for assessing performance gain. The calculated Rouge

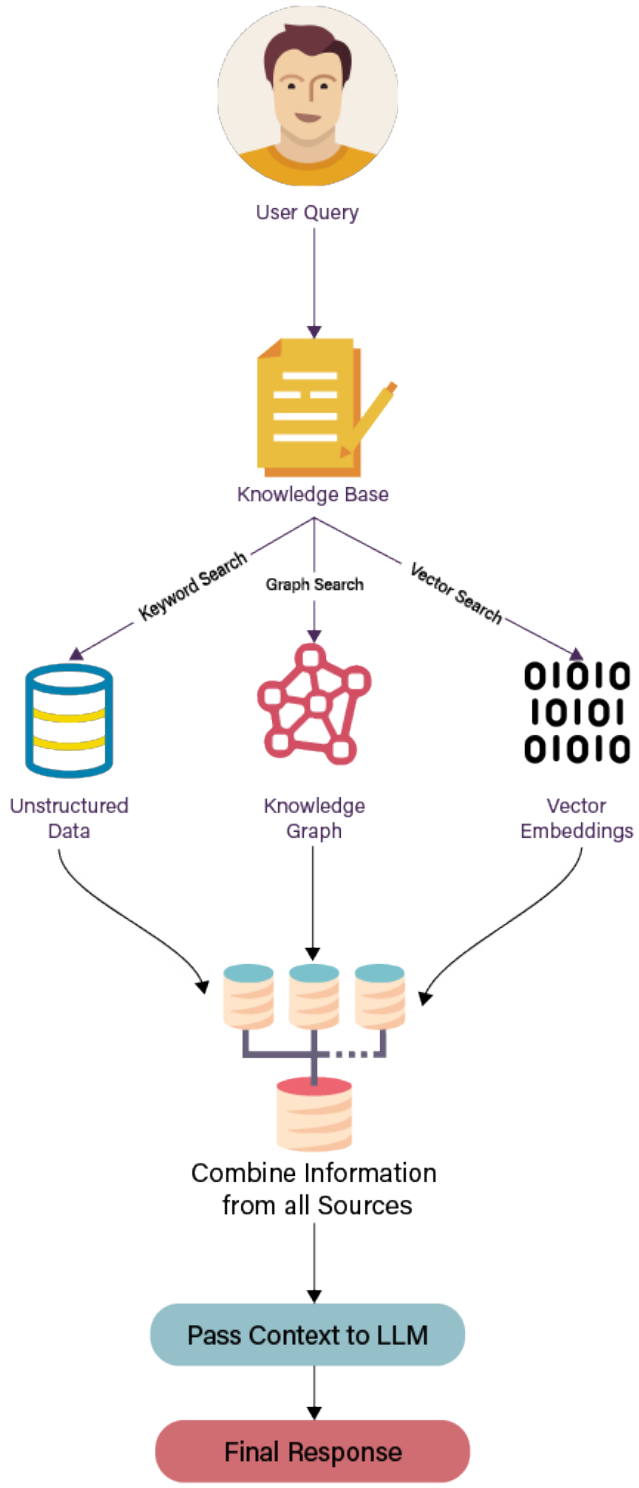

Fig. 4. Proposed Methodology

scores are 62.37% and 41.01%. Overall performance analysis is illustrated in Figure 04 and Table 03. This relative evaluation indicates the effectiveness of our approach in fetching and aligning contextually relevant information from the knowledge graph comparable to human answers. The Rouge L score is

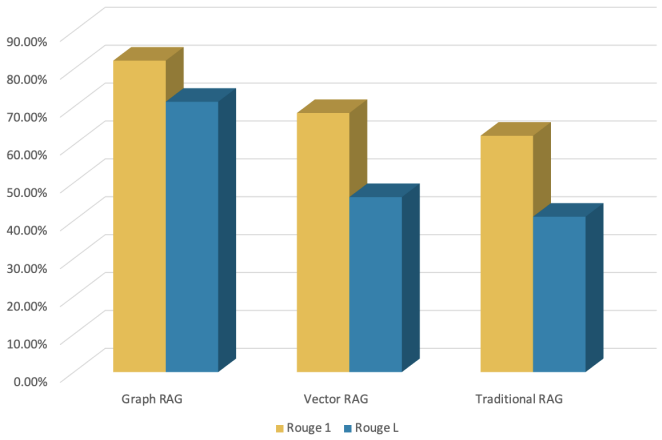

Fig. 5. Performance Analysis

slightly lower, but still, it reflects our approach's effectiveness while highlighting the areas of improvement in longer sequences considered necessary to improve results. Upon examining the results, it is found that a significant portion of questions expect generative answers that are summarized to fit into an answer width of 45 words. These generated answers during summarization challenge the rouge score. On the other hand, questions expecting one-word answers and short answers have very high scores, representing the retrieval algorithm's effectiveness. A critical area of improvement is the summarization process of long generative answers. This improvement is necessary to enhance the accuracy of obtaining comprehensive but brief and contextually relevant answers from the Graph RAG system.

TABLE I
COMPARATIVE ANALYSIS OF ROUGE 1 AND ROUGE L SCORES

Approach	**Rouge 1 %**	**Rouge L %**
Graph RAG	82.19	71.34
Vector RAG	68.37	46.18
Traditional RAG	62.37	41.01

V. CONCLUSION

The study introduces an authentic and credible knowledge base for diabetes care and controls knowledge graph construction. It implements a Graph RAG framework for integrating structured knowledge in LLMs to enhance accuracy and improve reasoning capabilities. Within the Graph RAG framework, this approach implements three different retrieval strategies for enhanced retrieval. Reasoning is improved by adding source parameters to ensure traceability for evidence-based responses. Our experiments reflect that this approach significantly improves the quality of responses for diabetes care and control by fetching highly relevant and contextually accurate information. The promising results in improved Rouge score are further proven by comparing with human-generated responses. This further opens up future directions for adding property graphs to enhance the results.

REFERENCES

[1] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," Apr. 01, 2023, Elsevier Ltd. doi: 10.1016/j.lindif.2023.102274.

[2] E. Ullah, A. Parwani, M. M. Baig, and R. Singh, "Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review," Dec. 01, 2024, BioMed Central Ltd. doi: 10.1186/s13000-024-01464-7.

[3] R. Yang et al., "Large language models in health care: Development, applications, and challenges," 2023, doi: 10.1002/hcs2.61.

[4] S. Wu et al., "STaRK: Benchmarking LLM Retrieval on Textual and Relational Knowledge Bases," Apr. 2024, [Online]. Available: http://arxiv.org/abs/2404.13207

[5] F. Stöhr, "Advancing language models through domain knowledge integration: a comprehensive approach to training, evaluation, and optimization of social scientific neural word embeddings," J Comput Soc Sci, 2024, doi: 10.1007/s42001-024-00286-3.

[6] M. Karabacak and K. Margetis, "Embracing Large Language Models for Medical Applications: Opportunities and Challenges", doi: 10.7759/cureus.39305.

[7] X. Lin, W. Wang, Y. Li, S. Yang, F. Feng, and Y. Wei, "USA-Seng Chua. 2024. Data-efficient Fine-tuning for LLM-based Recommendation", doi: 10.1145/3626772.3657807.

[8] Z. Al Nazi and W. Peng, "Large Language Models in Healthcare and Medical Domain: A Review," 2024, doi: 10.3390/informatics11030057.

[9] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," 2024. [Online]. Available: www.aaai.org

[10] K. Sawarkar, A. Mangal, and S. R. Solanki, "Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2404.07220

[11] Y. Gao, Y. Xiong, M. Wang, and H. Wang, "Modular RAG: Transforming RAG Systems into LEGO-like Reconfigurable Frameworks," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.21059

[12] P. Finardi et al., "The Chronicles of RAG: The Retriever, the Chunk and the Generator," Jan. 2024, [Online]. Available: http://arxiv.org/abs/2401.07883

[13] S. Tekkesinoglu and L. Kunze, "From Feature Importance to Natural Language Explanations Using LLMs with RAG," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.20990

[14] V. Sudhi, S. R. Bhat, M. Rudat, and R. Teucher, "RAG-Ex: A Generic Framework for Explaining Retrieval Augmented Generation," Association for Computing Machinery (ACM), Jul. 2024, pp. 2776–2780. doi: 10.1145/3626772.3657660.

[15] R. Friel, M. Belyi, and A. Sanyal, "RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems," Jun. 2024, [Online]. Available: http://arxiv.org/abs/2407.11005

[16] V. K. Kommineni, B. König-Ries, and S. Samuel, "From human experts to machines: An LLM supported approach to ontology and knowledge graph construction," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.08345

[17] T. T. Procko and O. Ochoa, "Graph Retrieval-Augmented Generation for Large Language Models: A Survey."

[18] D. Edge et al., "From Local to Global: A Graph RAG Approach to Query-Focused Summarization," Apr. 2024, [Online]. Available: http://arxiv.org/abs/2404.16130

[19] N. Matsumoto et al., "KRAGEN: a knowledge graph-enhanced RAG framework for biomedical problem solving using large language models," Bioinformatics, vol. 40, no. 6, Jun. 2024, doi: 10.1093/bioinformatics/btae353.

[20] B. Sarmah, B. Hall, R. Rao, S. Patel, S. Pasquali, and D. Mehta, "HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction," Aug. 2024, [Online]. Available: http://arxiv.org/abs/2408.04948

[21] D. Sanmartin, "KG-RAG: Bridging the Gap Between Knowledge and Creativity," May 2024, [Online]. Available: http://arxiv.org/abs/2405.12035

[22] J. Wu, J. Zhu, and Y. Qi, "Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation," Aug. 2024, [Online]. Available: http://arxiv.org/abs/2408.04187
[23] M. Fatehkia, J. K. Lucas, and S. Chawla, "T-RAG: Lessons from the LLM Trenches," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.07483
[24] S. Xu, M. Chen, and S. Chen, "Enhancing Retrieval-Augmented Generation Models with Knowledge Graphs: Innovative Practices Through a Dual-Pathway Approach," in Advanced Intelligent Computing Technology and Applications, D.-S. Huang, Z. Si, and W. Chen, Eds., Singapore: Springer Nature Singapore, 2024, pp. 398–409.
[25] X. He et al., "G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.07630