MC-RAG System: A Structure-Driven RAG System for Multi-Constraint Queries

Xiao Zhang[1] , **Yang Wan**[1] , **Yi Li**[2] , **Miao Xie**[1] , **Chunli Lv**[1,3]
[1]College of Information and Electrical Engineering, China Agricultural University, China
[2]College of Computing and Data Science, Nanyang Technological University, Singapore
[3]Key Laboratory of Agricultural Informatization Standardization, Ministry of Agriculture and Rural Affairs, China
B20243080802@cau.edu.cn, SY20253082208@cau.edu.cn, liyi0067@e.ntu.edu.sg, 0520shui@163.com, lvcl@cau.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) systems are widely adopted in question answering, yet they often fail to satisfy complex multi-constraint queries, leading to constraint violations, factual inconsistencies, or hallucinations. We present STRUCTURE-DRIVEN RAG SYSTEM FOR MULTI-CONSTRAINT QUERIES(MC-RAG), a structure-driven RAG system that reformulates retrieval as a subgraph matching problem over a knowledge graph. By integrating semantic and structural embeddings with path-level indexing, MC-RAG performs interpretable, structure-aware, and constraint-consistent retrieval and generation. During the demonstration, participants can input medical or encyclopedic multi-constraint queries, visualize how the system parses constraints, performs structural matching, and generates answers, thereby experiencing an end-to-end, interactive, and explainable RAG pipeline. A demo video is available at https://youtu.be/J8kahzmAnu0.

1 Introduction

Retrieval-Augmented Generation (RAG) [Fan *et al.*, 2024] has become a core framework in knowledge-driven question answering systems [Izacard and Grave, 2021; Borgeaud *et al.*, 2022], where external documents are retrieved to ground the generation of LLMs. While effective for simple, factual queries, real-world user queries often impose multiple constraints to be satisfied simultaneously, rather than matched by loosely related texts to provide rough knowledge. For example, in the nutritional scenario shown in Fig.1, a user may ask: *"Which food provides abundant protein, contains vitamin D, offers calcium beneficial for bone health, and helps with weight management?"*. The answer must jointly satisfy five constraints (i.e. containing protein, calcium, and vitamin D). However, existing RAG systems struggle to address such multi-constraint requirements. Mainstream RAG systems (i.e., chunk-based RAG systems[Lewis *et al.*, 2020]) retrieve documents solely by semantic similarity between the query and text chunks, ignoring the constraints in queries [Karpukhin *et al.*, 2020; Xu *et al.*, 2024; Khattab and Zaharia, 2020]. Recent advancements in graph-based RAG systems[Guo *et al.*, 2025; Liang *et al.*, 2025; Edge *et al.*, 2024; He *et al.*, 2024] aim to mitigate the problem by organizing external knowledge as a knowledge graph and retrieves entities or subgraphs most similar to the query. Yet, such methods still rely on similarity ranking over local graph structures, and thus cannot ensure precise satisfaction of single or multiple constraints.

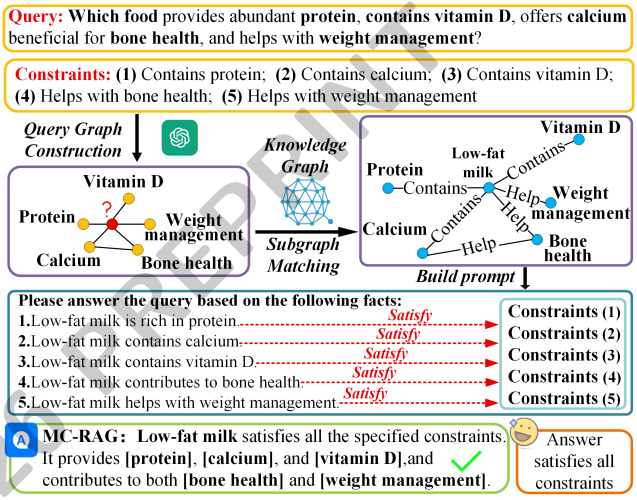

Figure 1: Multi-constraint query example.

To address this limitation, we propose MC-RAG, a demonstration system built upon our structure guided RAG framework [Xie *et al.*, 2026]. It integrates constraint parsing, structural subgraph retrieval, constraint verification, and generation grounded in retrieved evidence into an interactive RAG workflow. Our core idea is to replace similarity ranking with *constraint-satisfying subgraph matching* over a knowledge graph [Han *et al.*, 2013; Kim *et al.*, 2018; Xie *et al.*, 2018; Xie *et al.*, 2017; Zhang *et al.*, 2024], and use the matched subgraph as verifiable evidence to guide generation. Specifically, the system first parses the natural language query into a query graph, where the constraints are represented as nodes or edges. Unlike prior path-dominance indexing for fully labeled query graphs [Ye *et al.*, 2024], MC-RAG handles unknown nodes by completing missing labels during online matching. Then, it retrieves subgraphs in the knowl-

edge graph that are isomorphic or approximately isomorphic to the query graph, which satisfies all constraints. Finally, the retrieved structured evidence is used to guide the generation.

In addition, MC-RAG introduce an R*-Tree index to organize external knowledge represented by embeddings, which avoids the NP-complete bottleneck[Cordella *et al.*, 2004] of subgraph isomorphism on large-scale knowledge graphs and achieves near-linear retrieval efficiency. In this demonstration, we present an interactive prototype of MC-RAG. Unlike conventional "black-box" QA systems, users can visually inspect the full pipeline, including query parsing, subgraph matching, and answer generation. The demonstration highlights three key capabilities: (1) **Exact structural retrieval**: subgraph matching for multi-constraint queries, enabled by path-level embeddings and R*-Tree indexing; (2) **Constraint-consistent generation**: generating answers only when all query constraints are satisfied; and (3) **Interactive interpretability**: visualizing the complete process from query graph construction to final generation. As shown in Fig 1, MC-RAG identifies an isomorphic subgraph centered on "low-fat milk" that satisfies all constraints, thereby producing a logically complete answer.

In summary, MC-RAG reformulates multi-constraint queries as subgraph matching and enables efficient retrieval via indexing with embeddings. The demo lets users visually inspect how constraints are parsed, matched, and verified before generation, turning RAG from a black-box process into an transparent workflow and providing a reusable basis for future RAG systems for constraint handling.

2 System Architecture

The overall architecture of the MC-RAG is shown in Fig.2.

Data Layer. To better represent the semantic and structural information of external knowledge and lay the foundation for precise subgraph matching, we design the Data Layer to extract entity and relationship from raw knowledge, construct a knowledge graph, and build efficient path-level indexes.

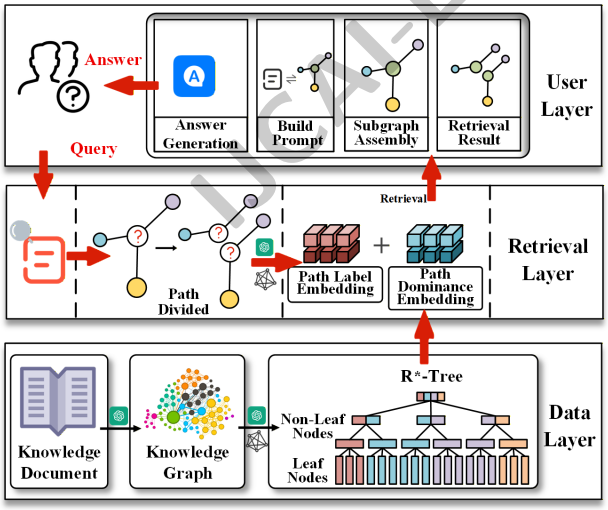

Figure 2: Overall architecture of the MC-RAG system.

MC-RAG adopts a dual-embedding and indexing mechanism, where each path is encoded into two embeddings: (1) a label embedding capturing the semantic meaning of nodes in the path, and (2) a dominant embedding learned by a graph neural network (GNN) to represent the structural pattern of the path. All path embeddings are organized in an R*-Tree.

Retrieval Layer. Retrieval Layer serves as the core component of MC-RAG, responsible for parsing user queries and performing subgraph matching-based retrieval on the knowledge graph. When a natural language question is submitted, the system first converts it into a query graph. The query graph is then decomposed into several paths, and the system executes path-level matching based on the R*-Tree index. Finally, the matched paths are aggregated and validated, producing a candidate subgraph that satisfies the full query conditions. This structured subgraph is then passed to the generation module to produce the answer.

User Layer. The User Layer is an interactive visualization interface where users can input natural language queries and observe the entire process of query parsing, subgraph retrieval, and answer generation in real time. This interaction mode allows users to intuitively understand how MC-RAG conducts interpretable and structure-driven retrieval-augmented generation.

3 Index Construction and Subgraph Matching-Based RAG

This section introduces the core process of MC-RAG.

Offline: Index Construction. We build a reusable path index over a knowledge graph to support efficient multi-constraint structural retrieval. Given raw documents, the system (i) extracts entities and relations to form a KG, and encodes each node label with an LLM to obtain semantic label embeddings; (ii) trains a lightweight GNN encoder to obtain structure-aware *dominant* embeddings that are comparable across local and global neighborhoods; and (iii) enumerates KG paths of a fixed length and stores each path with a pair of **(semantic, structural)** embeddings in an R*-Tree. The R*-Tree organizes embedding regions for hierarchical pruning, enabling fast candidate path retrieval for downstream query matching.

Online: Subgraph Matching-Based RAG. Given a user queries, the system uses an LLM to parse constraints and construct a query graph, and normalizes query mentions to KG entities via embedding-based retrieval. The query graph is decomposed into fixed-length paths, where missing labels are handled by wildcard completion using the index to propose candidate types/entities. We then perform path-level retrieval on the R*-Tree with joint semantic-and-structural filtering, aggregate matched paths, and validate them to form a candidate subgraph that satisfies all constraints, which is finally fed to the LLM as structured evidence for answer generation: If exact isomorphic matches exist, MC-RAG uses all matched subgraphs as evidence. Otherwise, it selects the approximate match with the minimum edit distance; if no reliable approximate match exists, it falls back to the core entity and its one hop neighbors.

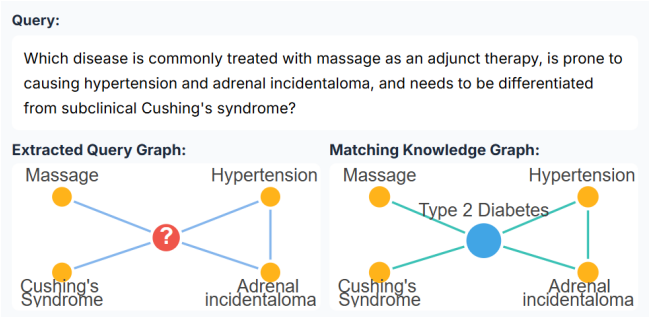

Figure 3: Visualization of query graph and matched subgraphs.

4 Demonstration Interface and Functions

System Interface Overview. The main interface consists of three functional modules: *(1) Knowledge Graph Construction*, *(2) Knowledge Graph Display*, and *(3) Interactive QA*.

In *Knowledge Graph Construction* module, users can upload raw knowledge documents (PDF, DOCX, TXT, etc.), and the system automatically extracts entities and relations to construct a knowledge graph while displaying the processing progress in real time. Once construction is complete, the system generates both label and dominant embeddings for all paths in the background and builds an R*-Tree index to support subsequent structure-aware retrieval. In *Knowledge Graph Display* module, users can view the global topology and type distribution of the constructed graph. The interface supports filtering by node type, zooming, and drag-and-drop interaction. Different colors represent different entity types. The system dynamically displays the number of nodes and edges, helping users inspect extraction quality and graph connectivity before running queries.

In *Interactive QA* module, as shown in Fig.3, users can enter natural language queries such as: *"Which disease commonly uses massage as an adjuvant therapy, is prone to have adrenal incidentaloma and hypertension, and requires differential diagnosis from subclinical Cushing's syndrome, adrenal incidentaloma often lead to hypertension?"* The system automatically parses the question using an LLM-based query parser to extract both semantic and structural constraints, and presents them as a structured list, as shown in Fig.4. These constraints, together with the target entity type, form the query graph q. Based on the parsed query graph, the system performs path-level matching against the indexed knowledge graph. Interactive QA module then visualizes the

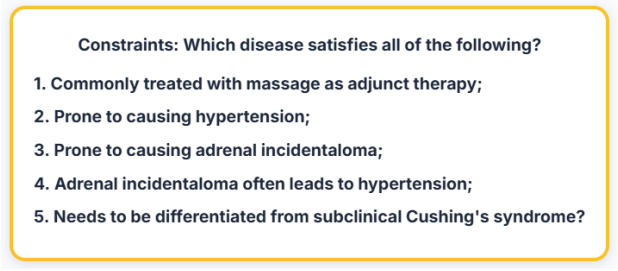

Figure 4: Query parsing and constraint extraction

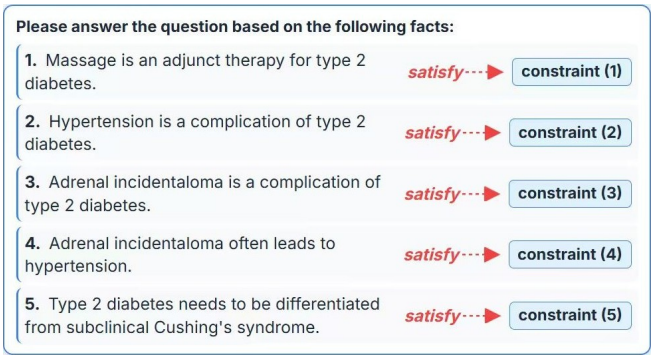

Figure 5: Constraint verification before answer generation.

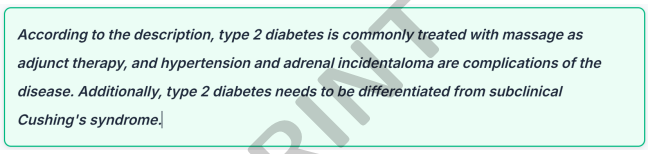

Figure 6: Final answer and reasoning chain visualization.

query graph and the retrieved candidate subgraphs that satisfy all constraints.

As shown in Fig.3, users can clearly observe the structural evidence underlying the generated answer. As shown in Fig. 5, once all constraints are verified, the generation module converts the final matched subgraph into a structured prompt and invokes the language model to produce a natural language explanation. As shown in Fig. 6, the interface simultaneously displays the generated answer and its logical reasoning chain.

Experimental Evaluation. Before the demo, we systematically evaluated MC-RAG using the multi-constraint dataset *ERQA*[1] and the single-constraint dataset *Natural Questions*[Kwiatkowski *et al.*, 2019]. *ERQA* covers 20 domains and contains approximately $206,000$ queries, with each query involving an average of 4.6 constraints. Experimental results show that MC-RAG significantly outperforms representative state-of-the-art RAG systems and LLMs. Compared with NaiveRAG[Lewis *et al.*, 2020], GraphRAG[Edge *et al.*, 2024], LightRAG[Guo *et al.*, 2025], KAG[Liang *et al.*, 2025], GPT-5, GLM-4V, Gemini-2.5 and Qwen-3, MC-RAG achieves consistent absolute gains on precision, recall, F1 and Hit@1 (up to $+37.04$ points in Hit@1, $+42.45$ points in Recall, and $+0.40$ points in F1). In terms of efficiency, MC-RAG achieves an average end-to-end generation time of 10.8 seconds per query; the subgraph matching pipeline takes about 6–7 seconds, comparable to KAG. During the offline index construction phase, MC-RAG requires approximately 11 hours in total, slightly longer than LightRAG(around 10 hours). The index can be reused across datasets, and the GNN encoder is trained only once offline.

[1]ERQA:(https://github.com/CAU-X-AI-Lab/ERQA)

Acknowledgements

This work was supported by the China Agricultural University "Young Researcher" Start-up Fund No. QNYJY2024144 and the Visiting Scholar Program of the China Scholarship Council (CSC) No. 202506350123.

Contribution Statement

Xiao Zhang and Yang Wan contributed equally to this work. Miao Xie is the corresponding author.

References

[Borgeaud *et al.*, 2022] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.

[Cordella *et al.*, 2004] Luigi P. Cordella, Pasquale Foggia, Carlo Sansone, and Mario Vento. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(10):1367–1372, 2004.

[Edge *et al.*, 2024] 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. *CoRR*, abs/2404.16130, 2024.

[Fan *et al.*, 2024] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on RAG meeting llms: Towards retrieval-augmented large language models. In Ricardo Baeza-Yates and Francesco Bonchi, editors, *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6491–6501. ACM, 2024.

[Guo *et al.*, 2025] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 10746–10761. Association for Computational Linguistics, 2025.

[Han *et al.*, 2013] Wook-Shin Han, Jinsoo Lee, and Jeong-Hoon Lee. Turboiso: towards ultrafast and robust subgraph isomorphism search in large graph databases. In *Proceedings of the 2013 ACM SIGMOD international conference on management of data*, pages 337–348, 2013.

[He *et al.*, 2024] Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Advances in Neural Information Processing Systems*, 37:132876–132907, 2024.

[Izacard and Grave, 2021] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880, 2021.

[Karpukhin *et al.*, 2020] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP*, pages 6769–6781, 2020.

[Khattab and Zaharia, 2020] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020.

[Kim *et al.*, 2018] Kyoungmin Kim, In Seo, Wook-Shin Han, Jeong-Hoon Lee, Sungpack Hong, Hassan Chafi, Hyungyu Shin, and Geonhwa Jeong. Turboflux: A fast continuous subgraph matching system for streaming graph data. In *Proceedings of the 2018 international conference on management of data*, pages 411–426, 2018.

[Kwiatkowski *et al.*, 2019] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

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

[Liang *et al.*, 2025] Lei Liang, Zhongpu Bo, Zhengke Gui, Zhongshu Zhu, Ling Zhong, Peilong Zhao, Mengshu Sun, Zhiqiang Zhang, Jun Zhou, Wenguang Chen, Wen Zhang, and Huajun Chen. KAG: boosting llms in professional domains via knowledge augmented generation. In Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov, editors, *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025 - 2 May 2025*, pages 334–343. ACM, 2025.

[Xie *et al.*, 2017] Miao Xie, Sourav S. Bhowmick, Gao Cong, and Qing Wang. Panda: toward partial topology-based search on large networks in a single machine. *The VLDB Journal*, 26(2):203–228, April 2017.

[Xie *et al.*, 2018] Miao Xie, Sourav S Bhowmick, Hao Su, Gao Cong, and Wook-Shin Han. Panda: a system for partial topology-based search on large networks. *Proc. VLDB Endow.*, 11(12):1966–1969, August 2018.

[Xie *et al.*, 2026] Miao Xie, Xiao Zhang, Yi Li, and Chunli Lv. Structure guided retrieval-augmented generation for factual queries, 2026.

[Xu *et al.*, 2024] Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. Unsupervised information refinement training of large language models for retrieval-augmented generation. *arXiv preprint arXiv:2402.18150*, 2024.

[Ye *et al.*, 2024] Yutong Ye, Xiang Lian, and Mingsong Chen. Efficient exact subgraph matching via gnn-based path dominance embedding. *Proc. VLDB Endow.*, 17(7):1628–1641, 2024.

[Zhang *et al.*, 2024] Zhijie Zhang, Yujie Lu, Weiguo Zheng, and Xuemin Lin. A comprehensive survey and experimental study of subgraph matching: trends, unbiasedness, and interaction. *Proceedings of the ACM on Management of Data*, 2(1):1–29, 2024.