2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Talk2Doc: A Patient Q&A system using Retrieval-Augmented Generation with Weighted Knowledge Graphs and LLMs

Asad Khan[1], Zafar Ali[1*], Irfanullah[2], Abdul Aziz[1] and Pavlos Kefalas[3]

[1] School of Computer Science and Engineering, Southeast University, Nanjing, China
{asadkhan,zafar_ali,aziz}@seu.edu.cn
[2] Department of Computer Science, Shaheed Benazir Bhutto University, Sheringal, Pakistan
irfan@sbbu.edu.pk
[3] Dashub & Department of Informatics, Aristotle University, Thessaloniki, Greece
pavloskefalas@gmail.com

Abstract. Effective retrieval in patient question-answering (Q&A) systems is essential for addressing complex medical and healthcare inquiries. Traditional retrieval-augmented generation (RAG) methods leveraging large language models (LLMs) treat historical dialogues and issue-tracking as unstructured text, overlooking critical intra- and inter-issue structures and semantic relationships. This limitation often reduces the contextual relevance and accuracy of generated responses. This paper introduces Talk2Doc, a novel Q&A system that combines RAG, weighted knowledge graphs (KGs), and LLMs to enhance response quality and contextual understanding in healthcare ap-plications. In particular, Talk2Doc constructs a weighted KG from patient questions, preserving both intra-issue structures and inter-issue relation-ships. By retrieving relevant subgraphs, the system generates precise, con-textually aware answers, effectively mitigating the drawbacks of fragmented text representations. The proposed system was rigorously evaluated using standard retrieval metrics (NDCG@K, Recall@K, MRR) and text generation metrics (BLEU, METEOR, ROUGE). Results show that Talk2Doc significantly outperforms existing approaches, improving answer accuracy and maintaining the structural integrity of patient dialogue information. By prioritizing semantic relationships among medical entities, Talk2Doc refines retrieval performance, ensuring high-quality responses. Scalable across diverse medical domains and languages, Talk2Doc represents a transformative advancement for healthcare Q&A systems.

Keywords: Weighted Knowledge Graph, Large Language Model, and Retrieval Augmented Generation, Question Answering.

* Corresponding author: zafar_ali@seu.edu.cn

1 Introduction

Effective medical support in patient question-answering (Q&A) systems is crucial for patient satisfaction and trust in the dialogue system [24, 25]. Since patient inquiries often mirror previously resolved health-related issues, efficiently retrieving relevant past cases is essential for timely and accurate resolution. Recent advancements in retrieval-augmented generation (RAG) [13, 29], embedding-based retrieval (EBR) [26, 12, 7], and large language models (LLMs) [20, 27, 30] have significantly improved question-answering and retrieval performance in this domain. This process typically involves two stages: In the first stage, historical patient queries are segmented into smaller chunks (treating them as plain text) to match the context length limitations of the embedding model and converted to embedding vectors. In the second stage, the Q&A stage, the most relevant chunks are retrieved and used as context for LLMs to answer queries. While this method is straightforward, it has several limitations, which we discuss below, along with our proposed solutions:

Limitation 1: Retrieval Accuracy Loss Due to Ignored Structures:

Issue tracking documents possess inherent semantic structures and interconnections, often with explicit references like "Issue A is related to/copied from/caused by Issue B." However, the conventional practice of segmenting such documents into smaller text chunks disregards these relationships, leading to a loss of critical contextual information. To overcome this limitation, we propose a novel approach that parses dialogues into tree structures and connects them into an integrated, weighted graph. This graph captures and preserves the intrinsic relation-ships between entities and their contextual significance, resulting in enhanced retrieval accuracy.

Limitation 2: Segmentation Reduces Answer Quality:

Dividing lengthy patient-doctor dialogues into fixed-length segments to fit the context length constraints of embedding models disrupts the logical flow of content, leading to incomplete answers. For instance, segmentation might separate a problem mentioned at the start of a dialogue from its corresponding solution at the end, resulting in the omission of crucial information. To address this, we propose a weighted graph-based approach that preserves the logical continuity and coherence of dialogue sections, ensuring comprehensive and high-quality responses.

Limitation 3: Overlooking Critical Semantic Relations:

Current models often assign equal importance to all semantic relationships, failing to account for the varying significance of different connections. However, in many cases, some relationships between entities are more critical than others. For example, a "caused by" relationship may hold greater importance than a "related to" relationship in determining the relevance of retrieved information. Our proposed solution incorporates weighted graphs to differentiate the significance of relationships, ensuring that more critical connections have a higher impact on retrieval and answer generation.

Limitation 4: Contextual Dependencies Across Segments Are Lost:

In lengthy dialogues or issue tickets, dependencies and references often span multiple segments. For example, a later segment might reference an earlier segment to provide additional clarification or to resolve ambiguities. Conventional segmentation methods fail to preserve these cross-segment dependencies, resulting in answers that

lack context and precision. Our method addresses this limitation by employing a graph structure that links related segments, allowing the Q&A system to access and integrate cross-segment dependencies for more precise and contextually accurate responses.

To address these limitations, we introduce Talk2Doc, a robust framework that integrates Retrieval-Augmented Generation (RAG), weighted Knowledge Graphs (KGs), and Large Language Models (LLMs) to advance patient question-answering in healthcare. This integration significantly enhances the quality, relevance, and contextual understanding of system responses. By combining retrieval-based methods with generative models, Talk2Doc delivers responses that are not only accurate but also contextually appropriate. The use of weighted KGs prioritizes relationships among medical entities, enabling the system to efficiently navigate complex patient inquiries. By leveraging structured medical knowledge, the framework ensures greater precision and relevance, even for multifaceted queries. Additionally, Talk2Doc offers scalability across diverse medical domains and languages, making it adaptable to a variety of healthcare settings. Ultimately, this approach enhances patient care by providing more informed and context-aware responses. The main contributions of this work are summarized below as follows:

- We introduce Talk2Doc, a novel integration of RAG, weighted KGs, and LLMs, to advance patient question-answering in healthcare.
- We enhance the accuracy and relevance of responses by combining retrieval-based methods with generative models, ensuring contextually aware answers.
- We leverage structured medical knowledge to effectively handle complex patient inquiries, improving the precision and quality of responses.
- We prioritize relationships among medical entities using weighted KGs, refining retrieval accuracy and ensuring precise, contextually relevant answers.
- We present a scalable and highly adaptable framework that seamlessly supports multiple medical domains and languages, significantly broadening its utility across diverse healthcare applications.

The remainder of the paper is structured as follows. Section 2 presents related work. Section 3 introduces the proposed model. Section 4 presents the experimental results and relevant discussion. Finally, Section 5 concludes this paper by presenting future research directions.

2 Related Works

Question-answering (Q&A) systems leveraging Knowledge Graphs (KGs) can be broadly categorized into retrieval-based, template-based, and semantic parsing-based approaches. Retrieval-based methods utilize distributed representations [6] or relation extraction [28] to retrieve answers from KGs. However, these methods struggle with questions involving multiple entities due to their limited capacity to represent complex relationships. Template-based methods rely on handcrafted tem-plates to encode intricate queries but are constrained by the finite number of available templates [23], which

limits their applicability in dynamic domains. Semantic parsing-based methods generate logical forms to extract knowledge from KGs [31, 21, 5]. While effective in specific tasks, their performance in Q&A systems remains inadequate, as highlighted in Section 1, necessitating the development of more advanced approaches that integrate LLMs and KGs.

Recent advancements have focused on combining LLMs and KGs, achieving notable improvements in reasoning and retrieval. Jin et al.[9] categorized LLMs as aligners, encoders, and predictors, highlighting their diverse roles in enhancing KG-based reasoning. Graph-based reasoning methods, such as Reasoning-on-Graph[17] and Think-on-Graph[22], utilize KGs to augment the reasoning abilities of LLMs. Yang et al.[30] enhanced factual reasoning during LLM training by incorporating KGs. Similarly, LLM-driven Q&A systems in specialized domains, such as food and medicine, have leveraged KGs to augment inference and improve accuracy [27, 20]. These advancements underscore the growing potential of integrating LLMs and KGs in Q&A tasks. Some works have further explored sophisticated pipelines to enhance Q&A systems. For example, LLeQA-RLLM[16] presents an end-to-end approach for generating long-form responses to statutory law questions using a "retrieve-then-read" framework. This system combines a bi-encoder for retrieval and an instruction-tuned LLM for answer generation, employing parameter-efficient fine-tuning for improved performance. UniGen[14] introduces a unified generative framework with a shared encoder and two decoders: one for retrieval and another for Q&A. This dual-decoder architecture enhances both retrieval and answer generation by optimizing input comprehension. RAG-CSQA [29] integrates RAG with KGs for customer service
Q&A by constructing a KG from historical issues, preserving both the internal structure of each issue and their interrelationships.

Despite these advancements, existing approaches, such as those in [16, 14, 29], have notable limitations. They fail to leverage weighted knowledge graphs, which are crucial for prioritizing significant relationships between entities. This oversight limits their ability to handle complex queries efficiently, reducing retrieval accuracy and contextual understanding while increasing response latency for large datasets. In contrast, our proposed approach addresses these limitations by integrating LLMs, RAG, and weighted knowledge graphs into a unified framework, Talk2Doc. Unlike existing methods, Talk2Doc leverages structured medical knowledge and assigns weights to relationships within KGs, enabling it to prioritize critical connections. This ensures more precise, context-aware retrieval and scalable performance across diverse medical domains and languages.
The details of our methodology are discussed in the next section.

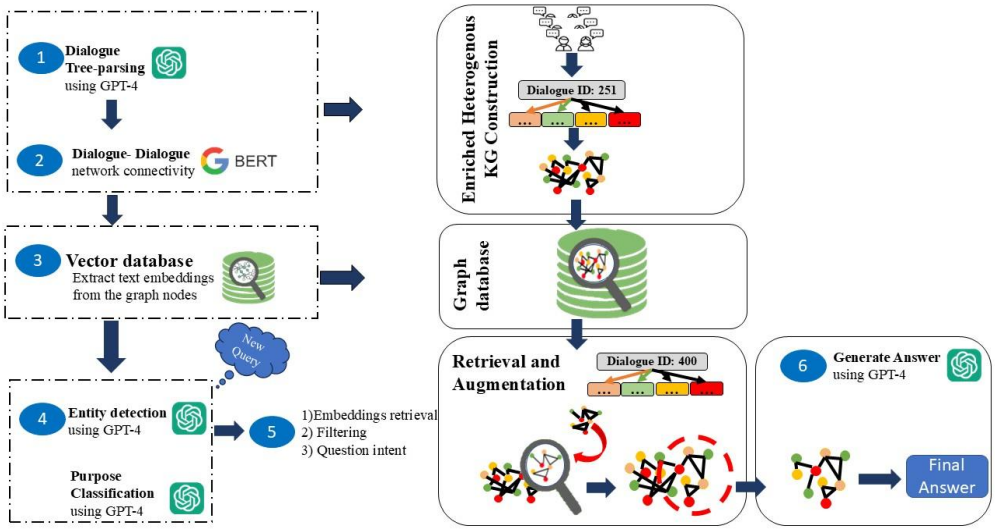

Fig. 1. The architecture of proposed Talk2Doc model.

3 The Proposed Method

We present Talk2Doc, a patient Q&A system that integrates a large language model (LLM) with retrieval-augmented generation (RAG) and a weighted knowledge graph (KG). Talk2Doc operates in two distinct phases, as illustrated in Figure 1. In the first phase, a comprehensive weighted KG is constructed from historical patient-issue dialogues. This process involves creating a tree-like representation of individual issues, establishing connections that account for their relational context, and generating embeddings for each node to enable efficient semantic searches. The second phase focuses on patient Q&A. Patient queries are parsed to extract named entities and intents. The system then navigates the weighted KG to locate relevant subgraphs and generate accurate, context-aware responses. Each phase is detailed in the following subsections.

3.1 Knowledge Graph Construction

This section provides a detailed overview of the methodology for constructing a weighted knowledge graph, designed to capture both the intricate relationships within individual issues and the broader connections between them.

Defining Graph Structure: To effectively represent historical patient issues, we propose a dual-level architecture for the KG, designed to capture both intra-issue and inter-issue relationships. This structure is depicted in Figure 1.

Defining Graph Structure: The following subsections outline the process of creating the weighted knowledge graph, focusing on intra-dialogue parsing, inter-dialogue connections, and edge weighting.

1. **Intra-Dialogue Parsing:** Each text-based dialogue (d_i) is transformed into a hierarchical tree structure (Di) in which each node represents a semantically meaningful section of the dialogue. This structure is generated using a hybrid rule-based and large language model (LLM) approach. Initially, predefined segments of the dialogue such as code sections or structured fields are extracted using deterministic rule-based methods, denoted as $d_{i,rule}$. The remaining unstructured portions, denoted $d_{i,llm}$, are processed using an instruction following LLM (e.g., GPT-4). The model is guided by a YAML template ($D_{template}$) that specifies a fixed set of questions (*e.g., summary, description, cause, importance, root cause, impact on, issue description*) representing typical information types in patient support dialogues. For each question in the template, the LLM is prompted to generate a corresponding textual answer from $d_{i,llm}$. These question–answer pairs are mapped into a tree structure, with each answer treated as a node $v_k \in D_i$. The combined tree D_i is therefore defined as:

$$d_i = d_{i,rule} \cup d_{i,llm}, \tag{1}$$

$$D_i = RuleParse(d_i, rule) + LLMParse\big(d_{i,llm}, D_{template}, prompt\big), \tag{2}$$

where RuleParse(·) and LLMParse(·) denote rule-based and LLM-based extraction modules, respectively. Once the tree Di is formed, a dense embedding vector h_k is computed for each node v_k using a pre-trained sentence embedding model. These embeddings are later used to compute semantic similarity and to weight edges in the graph construction phase.

2. **Inter-Dialogue Connection:** The individual trees (D_i) are combined into a unified graph (G) by establishing both explicit and implicit connections:

   — **Explicit Connections**($\boldsymbol{\varepsilon_{expl}}$)**:** These represent direct relationships within dialogues, such as those indicated by specific fields or references.

$$\varepsilon_{expl} = \{(D_i, D_j) | D_i \; explicitly \; connected \; to \; D_j \} \tag{3}$$

   — **Implicit Connections**($\boldsymbol{\varepsilon_{impl}}$)**:** These are inferred using textual-semantic similarities between dialogues. Advanced embedding methods and a threshold-based mechanism are employed to identify the most contextually relevant dialogue for a given query.

$$\varepsilon_{imp} = \{(D_i, D_j) | \cos\Big(embed(D_i), embed\big(D_j\big)\Big) \geq \theta\} \tag{4}$$

This dual-level connection strategy enables the graph to capture both predefined relationships and hidden semantic links, creating a robust framework for patient Q&A. After forming both explicit and implicit edges, we assign a normalized weight

we ∈ [0, 1] to each edge, capturing the strength of its structural and semantic contribution. The weighting process is detailed in the following subsection.

Edge Weighting: To quantitatively capture the strength of the relationships between nodes in the knowledge graph, we assign a normalized weight $w_e \in [0,1]$ to each edge $e \in \varepsilon = \varepsilon_{expl} \cup \varepsilon_{impl}$. Each node in the graph corresponds to a text segment generated by an LLM in response to predefined semantic prompts, such as "summary," "cause," "impact on," or "root cause." These nodes encapsulate discrete conceptual elements of a dialogue, denoted as $v_i \in D_k$, where Dk is the tree structure derived from dialogue dk. We employ a text embedding model to convert each node's textual content into a high-dimensional vector representation:

$$\boldsymbol{h}_i = embed(v_i), \tag{5}$$

where $embed(\cdot)$ denotes the embedding function (e.g., Sentence-BERT or a domain-specific LLM encoder). The weight of an edge $e = (v_i, v_j)$ is then computed using cosine similarity between the embeddings of the corresponding nodes:

$$sim(v_i, v_j) = \cos(\boldsymbol{h}_i, \boldsymbol{h}_i) = \frac{h_i . h_j}{\|h_i\| \|h_j\|}. \tag{6}$$

To ensure consistency across all edges, we normalize the similarity scores to the interval [0, 1] using the following transformation:

$$w_e = \frac{1+sim(v_i,v_j)}{2}, for\ all\ e = (v_i, v_j) \in \varepsilon \tag{7}$$

This weighting mechanism ensures that both explicit and implicit edges are evaluated not only based on their existence but also on their semantic coherence. High-weight edges indicate strong semantic alignment between concepts, while lower weights signal weaker or more tangential associations. The resulting weighted knowledge graph thus provides a rich and fine-grained structural representation of patient support dialogues, supporting downstream applications such as retrieval, reasoning, and recommendation.

Identifying Query Intent and Entities: Each user query (q) undergoes processing to extract its intents (I) and associated named entities (P) mapped as $Map(N \rightarrow U)$. Specifically, the query is parsed into key-value pairs, where each key (n) corresponds to an element in the graph template$(D_{template})$, and each value (v) represents the relevant information extracted from the query. The intents (I) capture the graph-related entities and objectives that the query seeks to address. To support this parsing, a carefully designed prompt is employed with an LLM, enabling accurate identification of intents and entities even for complex or diverse query structures. For example, given the query: *"For the past 48 hours, I had a fever, sore back, and nasal congestion"*, the extracted entities are: $P = Map(\text{'sore back'} \rightarrow \text{'fever'}, \text{'issue description'} \rightarrow \text{'The patient has the following three symptoms...'})$ and the intent set is:

$I = Set(\text{'cause of occurrence'})$. This approach leverages the LLM's advanced comprehension and interpretative capabilities, making it adaptable to a wide range of query formats. The extraction process can be summarized as follows:

$$P, I = LLM(q, D_{template}, prompt) \tag{8}$$

Retrieving Sub-Graphs Using Embeddings: We rank historical dialogues T_i using embeddings-based similarity. For each $(k, u) \in P$, we compute cosine similarity $\beta(u, n)$ between user input u and node n in section k. We then sum all relevant scores for each dialogue:

$$S_{T_i} = \sum_{(k,u)\in P} \sum_{n\in T_i} \mathbb{I}\{n.sec = k\}.\beta(u, n).\max_{e:(n\to m)} w_e \tag{9}$$

where S_{T_i} represents the relevance score computed for the historical dialogue T_i. $(k, u) \in P$ is the key-value pair from the set P, where k is a section $(e.g., symptom, description)$ and u is the corresponding user-provided text.
$n \in T_i$. Node n belongs to the set of nodes in dialogue T_i. $\mathbb{I}\{n.sec = k\}$ is the indicator function that equals 1 if node n belongs to section k, and 0 otherwise. $\beta(u, n)$ denotes the cosine similarity between the query value u and node n, computed using their respective text embeddings. $\max_{e:(n\to m)} w_e$ computes the maximum edge weight among all outgoing edges e from node n to any other node m in the graph. This equation ensures that the dialogues with the highest scores are selected as the top $K_{dialogue}$ candidates for response generation.

In the second step of sub-graph extraction, we refine the user query by incorporating the retrieved dialogue ID, which links the query to relevant historical issue dialogues. This ensures the query is contextually anchored to the $dialogue(s)$ identified during embeddings-based retrieval. The original query (q) is updated to include the dialogue ID (q'), producing a more focused version. This enhanced query is then translated into a graph query language like Cypher for Neo4j to enable precise knowledge graph queries. For example, the query *"For the past 48 hours, I've had a fever, a sore back, and nasal congestion"* is reformulated as ***"For the past 48 hours, I've had a fever, a sore back, and nasal congestion; use Dialogue ID-145 as context."*** to specify the retrieval scope. This linkage improves answer accuracy by narrowing the search to the most relevant subgraphs, as illustrated in Listing 1.1.

```
MATCH (dialogue {id: 145}) - [*] →(n)
WHERE n.section = 'symptom'
RETRUN n;
```

Listing 1.1. Cypher query to retrieve symptom nodes from a specific dialogue.

It is worth mentioning that the LLM-based query formulation can be adapted to retrieve information from subgraphs, whether these subgraphs belong to the same tree or to different trees in the KG.

Generating Answers: The data retrieved in Section 3.1 is leveraged to generate responses to the original user query. In this process, the Large Language Model (LLM) acts as a decoder, synthesizing answers based on the context provided by the retrieved information. The retrieved subgraphs are not only semantically relevant but are also prioritized based on the normalized edge weights, ensuring that the generation process is grounded in connections that reflect both structural and semantic importance. For example, in response to a user query, the *Talk2Doc* system might generate a detailed response such as: "Measure temperature every two hours for the next 2 days, drink tea and other fluids to stay hydrated, and take Panadol every 8 hours." This process ensures that the system provides contextually accurate and actionable medical advice. To maintain consistent and reliable service, a fallback mechanism is employed: if any issues arise during query execution, the system defaults to a standard textual retrieval approach, ensuring that the user receives an answer even in case of retrieval failure. This dual approach enhances system robustness and reliability, ensuring that patients always receive helpful guidance. Algorithm 1 shows the main steps of the model.

Algorithm 1 Talk2Doc: QA with RAG and Semantic KG

Require: User query q
Ensure: Answer $\hat{A}$

Phase 1: Knowledge Graph Construction

for all *dialogues* d_i **do**
  $d_{i,rule} \leftarrow RuleParse(d_i)$
  $d_{i,llm} \leftarrow LLMParse(d_i, D_{template}, prompt)$
  $D_i \leftarrow d_{i,rule} \cup d_{i,llm}$
  for all nodes $v_k \in D_i$ **do**
    $\boldsymbol{h}_k \leftarrow embed(v_k)$
  end for
end for
connect explicit edges ε_{expl}
connect implicit edges ε_{impl} using threshold θ
Compute edge weights: $w_e \leftarrow \frac{1+\cos(h_i, h_j)}{2}$

Phase 2: Knowledge Graph Construction

$(P, I) \leftarrow LLM(q, D_{template}, prompt)$
for all *dialogues* T_i **do**
  $$S_{T_i} \leftarrow \sum_{(k,u)\in P} \sum_{n\in T_i} \mathbb{I}\{n.sec = k\}\beta(u,n) \max_{e:(n\to m)} w_e$$
end for
select $top-K$ dialogues by S_{T_i}
refine query $q' \leftarrow q + dialogue\ ID$
$\hat{A} \leftarrow LLM.generate(q', subgraph)$
if generation fails **then**
  $\hat{A} \leftarrow FallbackRetrieval(q)$
end if
return $\hat{A}$

4 Experimental Evaluation

In this section, we discuss and compare the experimental results of the proposed model against other baseline models. We first describe the baseline models adopted, followed by the evaluation metrics and datasets used. Next, we present a comparative analysis of the experimental results and discuss the parameters.
It is important to note that our evaluation was conducted using datasets that included typical patient queries, doctor responses, and prescriptions. The control group relied on traditional textual embedding-based retrieval and the experimental group used Talk2Doc methodology. Both groups used the same GPT-4, LLM, and GBERT.

4.1 Baseline Models

To ensure a fair comparison, we carefully select a set of baseline models. The first baseline method is **RAG-CSQA** introduced by Xu et al. [29], which combines KG with RAG for customer service Q&A. The second baseline method, **LLeQA-RLLM**, is by Louis et al. [16], and utilizes retrieval-augmented LLMs for long-form legal Q&A. The next baseline is **UniGen**, introduced by Li et al. [14], which employs a unified generative framework for retrieval and Q&A using LLMs. Our proposed model, referred to as **Talk2Doc**, uses GPT-4 [18] in conjunction with a weighted KG and RAG to generate answers for patient queries.

4.2 Evaluation Metrics

We evaluated models retrieval effectiveness using NDCG@K, MRR, and recall@K [1,11, 7]. On the one hand, MRR assesses the performance of Q&A systems by averaging the reciprocal ranks of the first relevant answer for each query. It is computed as $(\frac{1}{Rank})$, where "$Rank$" represents the position of the first relevant answer. A higher MRR indicates that relevant answers are ranked more favorably and retrieved more effectively, reflecting the system's capability to deliver accurate and prompt responses. On the other hand, Recall@K measures the fraction of relevant answers retrieved out of all available relevant answers, indicating how well the system identifies and returns all potential correct answers to a query. A higher recall signifies better performance in capturing comprehensive and relevant information. Finally, NDCG@K evaluates ranking quality by taking into account both items' relevance and position.

For assessing Q&A performance, we compared the generated responses to the ground truth using text generation metrics such as BLEU, ROUGE, and METEOR. BLEU [19, 3] measures the overlap of n-grams (word sequences) between generated and reference answers, with higher scores indicating greater similarity and reflecting fluency and accuracy. ROUGE [15] focuses on recall, evaluating how much relevant content is captured by comparing n-grams, word sequences, and word pairs. METEOR [4] improves on BLEU by considering synonyms, stemming (matching word roots), and word order, balancing precision and recall to better capture the meaning of the

generated answer. ERTScore [32] calculates precision, recall, and F1 scores by comparing the similarity of the embeddings of the generated and reference texts. Together, these metrics help evaluate how well the generated answers align with the expected content, fluency, and relevance in question-answering systems.

4.3 Evaluation Datasets

We used the 'Diagnose Me' dataset[1], which is a Long-Form Question Answering (LFQA) dataset. It compiles dialogues between patients and doctors from icliniq.com and healthcaremagic.com. The dataset includes over 257,000 unique patient questions along with the corresponding medical advice or prescriptions provided by doctors. This extensive collection of real-world exchanges is designed to support the development and evaluation of AI systems that can comprehend and generate precise, detailed medical responses in patient-doctor interactions.

We used MedQuAD dataset, which encompasses 37 distinct question types. In this study, following the approach of [2], we focus on QA pairs categorized under "Information," which provide definitions and details about medical terms. For our experiments, we selected 4,000 QA pairs for training, 200 for validation, and 200 for testing.

4.4 Overall Comparison

Table 1 and Table 2 present the retrieval and Q&A performance, respectively. Our Talk2Doc model demonstrates consistent improvements across all metrics, significantly outperforming the baseline with an MMR score of 96.3%, recall@3 score of 100%, and nDCG@3 is 97.3% on the Diagnose me dataset. Besides, Talk2Doc gained 82.3% results on MRR, recall@3 score of 77.1 %, and nDCG@3 score of 79.1% on the MedQuAD dataset. This highlights that integrating RAG with weighted KG and LLMs substantially enhances the accuracy and relevance of patient Q&A systems. The strong performance of BLEU, METEOR, ROUGE, and BERTScore on the Diagnose Me and MedQuAD datasets, achieved through the integration of a weighted knowledge graph, LLM, and RAG, stems from their ability to evaluate both lexical and semantic alignment effectively. The weighted knowledge graph enhances relevance by prioritizing precise information, while RAG grounds the LLM's outputs in factual content, reducing hallucinations. BLEU and ROUGE focus on n-gram overlaps, METEOR accounts for semantic nuances through synonym recognition, and BERTScore captures deeper contextual similarities using embeddings. This combination ensures contextually accurate and well-aligned responses, resulting in superior metric performance.

RAG combines retrieval and generation by first extracting relevant information from a broad dataset and then using this information to produce coherent, contextually appropriate responses. The incorporation of weighted KGs adds value by structuring and prioritizing medical knowledge, emphasizing critical relationships between entities, and aiding in the navigation of complex queries. LLMs, with their advanced language processing capabilities, generate precise and contextually sensitive answers based on

[1] https://www.kaggle.com/datasets/dsxavier/diagnoise-me

the structured data from KGs and the foundational retrieval from RAG. This unified approach ensures that patient inquiries are handled with high precision and relevance, thereby improving the overall effectiveness and efficiency of the Talk2Doc system.

Table 1. Models performance in terms of retrieval accuracy.

Dataset	Model	MRR	Rec@1	Rec@3	nDCG@1	nDCG@3
Diagnose Me	RAG-CSQA	0.912	0.835	0.873	0.868	0.906
	LLeQA-RLLM	0.891	0.814	0.853	0.853	0.884
	UniGen	0.904	0.823	0.864	0.859	0.897
	Talk2Doc	**0.963**	**0.906**	**1.000**	**0.932**	**0.973**
MedQuAD	RAG-CSQA	0.781	0.712	0.732	0.742	0.756
	LLeQA-RLLM	0.793	0.721	0.731	0.754	0.759
	UniGen	0.801	0.735	0.745	0.763	0.774
	Talk2Doc	**0.823**	**0.741**	**0.771**	**0.781**	**0.791**

Table 2. Distance metric of the predicted answers to the ground truth.

Dataset	Model	BLEU	METEOR	ROUGE	Bert-Score
Diagnose Me	RAG-CSQA	0.491	0.574	0.571	0.787
	LLeQA-RLLM	0.481	0.552	0.542	0.765
	UniGen	0.492	0.571	0.561	0.774
	Talk2Doc	**0.524**	**0.651**	**0.612**	**0.853**
MedQuAD	RAG-CSQA	0.251	0.265	0.303	0.831
	LLeQA-RLLM	0.250	0.298	0.298	0.861
	UniGen	0.261	0.284	0.310	0.851
	Talk2Doc	**0.291**	**0.312**	**0.351**	**0.912**

4.5 Ablation study

In this section, we analyze the impact of integrating different modules on the experimental results. The results of the ablation study using Diagnose Me dataset are depicted in Table 3. The first model, RAG+GPT-4, employs GPT-4 and RAG without the Knowledge Graph (KG) to generate results, while the second model incorporates a simple KG with GPT-4 and RAG, demonstrating improvements over the first. On the other hand, the third variant, KGw+RAG+GPT-2, replaces the standard KG with a weighted KG (KGw) and uses GPT-2, producing inferior results compared to the first two models. Similarly, the fourth variant introduces KGw with LLAMA 3.1 and RAG, showing competitive performance but still slightly lower than GPT-4 variants.

In the fifth variant, KGw+RAG+GPT-3.5T, the system shows substantial gains by using GPT-3.5 with weighted KG, achieving higher scores across BLEU, METEOR, ROUGE, and Bert-Score. Finally, the proposed Talk2Doc model, which integrates GPT-4 with weighted KG and RAG, achieves the best performance, with a 2.3% improvement in BLEU, 2% in METEOR, approximately 3% in ROUGE, and almost 3% in Bert-Score compared to the second-best variant using the Diagnose Me dataset.

These results indicate that the inclusion of a weighted KG significantly enhances the system's performance, particularly in conjunction with GPT-4. The Talk2Doc model outperforms all other variants, demonstrating that the combination of GPT-4, RAG, and a weighted KG maximizes retrieval and question-answering performance.

Table 3. Ablation study

Model	BLEU	METEOR	ROUGE	Bert-Score
RAG+GPT 4	0.451	0.581	0.571	0.765
KG+RAG+GPT4	0.471	0.610	0.581	0.787
KG_w +RAG+GPT2	0.321	0.423	0.484	0.573
KG_w +RAG+LLAMA3.1	0.492	0.593	0.562	0.771
KG_w +RAG+GPT3.5T	0.501	0.631	0.581	0.821
Talk2Doc	**0.524**	**0.651**	**0.612**	**0.853**

4.6 Hyperparameter Settings

Optimizing Q&A systems involves several critical hyperparameters for both LLMs and RAG frameworks. For GBERT, we used the following settings: number of attention $heads = 12$, number of $layers = 12$, $hidden\ size = 768$, feedforward hidden $size = 3072$, maximum sequence length = 512, vocabulary size = 30,000, and drop-out rate = 0.1. For GPT-4, the parameters were set to: temperature = 1.0, $max\ tokens = 4096$, presence penalty = 0, frequency penalty = 0, and $top - p = 1.0$. In the RAG framework, essential parameters include retrieval batch size and retrieval top-K, which manage the number and relevance of documents retrieved per query. We set the retrieval batch size to 50 and top-K to 3. Our edge weighting scheme introduces two hyperparameters: The mixing coefficient $\lambda \in [0, 1]$, which balances structural importance ($\alpha(T)$) and semantic similarity ($\beta(u, v)$) in the convex combination $w_e = \lambda\, \alpha(T) + (1 - \lambda)\, \beta(u, v)$. We initialize $\lambda = 0.5$ and tune it in steps of 0.1 on a held-out validation set to maximize retrieval NDCG@K. The similarity threshold $\theta \in [0, 1]$, which determines when to create implicit edges via $\beta(u, v) \geq \theta$. We begin with $\theta = 0.8$ and similarly validate its optimal value. Proper adjustment of all these hyperparameters including λ and θ is crucial for achieving optimal performance in terms of both retrieval accuracy and generative answer quality.

5 Conclusion and Future Directions

In this study, we addressed significant limitations of traditional retrieval-augmented generation (RAG) methods in patient Q&A systems by integrating these techniques with a weighted knowledge graph (KG). We presented the Talk2Doc model that effectively leverages weighted KGs in conjunction with large language models (LLMs) to enhance both the accuracy and relevance of responses in healthcare settings. The key innovation of the Talk2Doc method lies in constructing a weighted KG that preserves the intricate intra- and inter-issue relationships within patient dialogues. By doing so,

the Talk2Doc approach not only captures the structural integrity of medical information but also improves contextual understanding, leading to more precise and contextually aware answers. Experimental evaluations demonstrate that Talk2Doc surpasses existing approaches, significantly improving retrieval accuracy and answer quality. Metrics such as MRR, Recall@K, NDCG@K, BLEU, ROUGE, and METEOR confirm the superiority of our approach in navigating complex medical inquiries.

The incorporation of structured medical knowledge and semantic relationships between entities proves to be instrumental in refining retrieval accuracy and mitigating the limitations imposed by text segmentation. Our approach offers a scalable solution applicable across various medical domains and languages, providing a robust framework for future advancements in patient Q&A systems. In summary, our integration of RAG with weighted KGs represents a substantial advancement in the field, addressing critical gaps in existing methods and setting a new benchmark for retrieval accuracy and answer quality in healthcare dialogue systems. Future work will focus on further optimizing the KG construction process and exploring additional applications in diverse medical contexts.

6 Limitations

The proposed patient QA system, which integrates a KG, RAG, and LLM, may underperform with limited training data. Both LLMs and retrieval systems rely on large datasets to capture complex patterns and context effectively [10]. With fewer examples, knowledge representations become less accurate, weakening both retrieval and response generation, especially for complex medical queries. Expanding the dataset or applying data augmentation can help mitigate this issue. Additionally, the system lacks a reinforcement learning mechanism to adapt through user interactions. Incorporating feedback via reinforcement learning would allow continuous refinement, improving performance over time [8]. Lastly, the model may struggle when faced with multiple dialogues describing the same issue but offering different solutions. Such inconsistencies can confuse the model and reduce its reliability when generating responses in similar future cases.

7 Ethical Considerations

The performance of our Talk2Doc method is not flawless, primarily due to limited knowledge retrieval capabilities and the possibility of errors or outdated information. Our approach has been tested on two publicly available datasets, and we acknowledge that its applicability and findings may be restricted to similar datasets or domains. Its effectiveness on other specific datasets or domains remains uncertain. Consequently, there are potential risks in applying our method to privacy-sensitive or high-stakes datasets. Caution is advised, and it is essential to verify the accuracy of the generated answers.

References

1. Ali, Z., Huang, Y., Ullah, I., Feng, J., Deng, C., Thierry, N., Khan, A., Jan, A.U., Shen, X., Rui, W., et al.: Deep learning for medication recommendation: a system- atic survey. Data Intelligence 5(2), 303–354 (2023)
2. August, T., Reinecke, K., Smith, N.A.: Generating scientific definitions with con- trollable complexity. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8298–8317 (2022)
3. Aziz, A., Ali, Z., Qi, G., Huang, Y., Kefalas, P., Aminullah, Ali, A.: Leverage diagnosis intensity in medication recommendations. In: Advanced Intelligent Com- puting Technology and Applications - 20th International Conference, ICIC 2024, Tianjin, China, August 5-8, 2024, Proceedings, Part VI (LNAI). Lecture Notes in Computer Science, vol. 14880, pp. 38–50. Springer (2024)
4. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the Association for Computational Linguistics (ACL) workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
5. Bhutani, N., Zheng, X., Qian, K., Li, Y., Jagadish, H.: Answering complex questions by combining information from curated and extracted knowledge bases. In: Proceedings of the first workshop on natural language interfaces. pp. 1–10 (2020)
6. Bordes, A., Weston, J., Usunier, N.: Open question answering with weakly supervised embedding models. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14. pp. 165–180. Springer (2014)
7. Christoforidis, G., Kefalas, P., Papadopoulos, A.N., Manolopoulos, Y.: RELINE: point-of-interest recommendations using multiple network embeddings. Knowl. Inf. Syst. 63(4), 791–817 (2021)
8. Jaques, N., Yang, Y., Kottur, S., Han, X., Goodman, N., Binns, R., Li, J.: Human-centric dialog training via offline reinforcement learning. In: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS) (2020)
9. Jin, B., Liu, G., Han, C., Jiang, M., Ji, H., Han, J.: Large language models on graphs: A comprehensive survey. arXiv preprint arXiv:2312.02783 (2023)
10. Kaplan, J., McCandlish, S., Henighan, T., Brown, T., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
11. Kefalas, P., Symeonidis, P., Manolopoulos, Y.: Recommendations based on a heterogeneous spatio-temporal social network. World Wide Web 21(2), 345–371 (2018)
12. Kung, P.P.H., Fan, Z., Zhao, T., Liu, Y., Lai, Z., Shi, J., Wu, Y., Yu, J., Shah, N., Venkataraman, G.: Improving embedding-based retrieval in friend recommendation with ann query expansion. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2930–2934 (2024)
13. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems 33, 9459–9474 (2020)
14. Li, X., Zhou, Y., Dou, Z.: Unigen: A unified generative framework for retrieval and question answering with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 8688–8696 (2024)
15. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text sum-

marization branches out. pp. 74–81 (2004)
16. Louis, A., van Dijck, G., Spanakis, G.: Interpretable long-form legal question an- swering with retrieval-augmented large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 22266–22275 (2024)
17. Luo, L., Li, Y.F., Haffari, G., Pan, S.: Reasoning on graphs: Faithful and interpretable large language model reasoning. arXiv preprint arXiv:2310.01061 (2023)
18. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
19. Post, M.: A call for clarity in reporting bleu scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers. pp. 186–191 (2018)
20. Qi, Z., Yu, Y., Tu, M., Tan, J., Huang, Y.: Foodgpt: A large language model in food testing domain with incremental pre-training and knowledge graph prompt. arXiv preprint arXiv:2308.10173 (2023)
21. Sun, H., Dhingra, B., Zaheer, M., Mazaitis, K., Salakhutdinov, R., Cohen, W.: Open domain question answering using early fusion of knowledge bases and text. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4231–4242 (2018)
22. Sun, J., Xu, C., Tang, L., Wang, S., Lin, C., Gong, Y., Ni, L.M., yeung Shum, H., Guo, J.: Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In: International Conference on Learning Representations (2023)
23. Unger, C., Bühmann, L., Lehmann, J., Ngonga Ngomo, A.C., Gerber, D., Cimiano, P.: Template-based question answering over rdf data. In: Proceedings of the 21st international conference on World Wide Web. pp. 639–648 (2012)
24. Varshney, D., Zafar, A., Behera, N.K., Ekbal, A.: Knowledge graph assisted end- to-end medical dialog generation. Artificial Intelligence in Medicine 139, 102535 (2023)
25. Varshney, D., Zafar, A., Behera, N.K., Ekbal, A.: Knowledge grounded medical dialogue generation using augmented graphs. Scientific Reports 13(1), 3310 (2023)
26. Wang, W., Guo, Y., Shen, C., Ding, S., Liao, G., Fu, H., Prabhakar, P.K.: Integrity and junkiness failure handling for embedding-based retrieval: A case study in social network search. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3250–3254 (2023)
27. Wen, Y., Wang, Z., Sun, J.: Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. arXiv preprint arXiv:2308.09729 (2023)
28. Xu, K., Feng, Y., Huang, S., Zhao, D.: Hybrid question answering over knowl- edge base and free text. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 2397–2407 (2016)
29. Xu, Z., Cruz, M.J., Guevara, M., Wang, T., Deshpande, M., Wang, X., Li, Z.: Retrieval-augmented generation with knowledge graphs for customer service ques- tion answering. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2905–2909 (2024)
30. Yang, L.F., Chen, H., Li, Z., Ding, X., Wu, X.: Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling. IEEE Transactions on Knowledge and Data Engineering 36, 3091–3110 (2023)
31. Yih, S.W.t., Chang, M.W., He, X., Gao, J.: Semantic parsing via staged query graph generation: Question answering with knowledge base. In: Proceedings of the Joint Conference of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL) and the 7th International Joint Conference on Natural Language Processing of the AFNLP (2015)
32. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)