

Bridging Rationales and Relations: The Graph-Rationale-Guided Retrieval-Augmented Generation in Medical QA

Zhenyu Zhu

Pamplin College of Business, Virginia Polytechnic Institute and State University, 24061
Blacksburg, VA, USA

zhenyuz25@vt.edu

Abstract. Large language models (LLMs) face challenges of hallucination and knowledge obsolescence in medical question answering. Existing retrieval-augmented generation (RAG) frameworks can improve retrieval reliability through rational guiding; however, their neglect of structured knowledge leads to insufficient relational reasoning. This paper proposes the Graph-Rationale-Guided Retrieval-Augmented Generation (GRAG) framework, which introduces a knowledge graph layer based on Rationality-Guided Retrieval-Augmented Generation (RAG²) to support dynamic graph query expansion, evidence fusion, hallucination detection, and quantized low-rank adaptation (QLoRA). GRAG's core mechanisms include rational generation, entity extraction, graph construction based on Unified Medical Language System (UMLS)-Neo4j, and similarity-driven multi-source evidence fusion. Experiments on medical question answering dataset (MedQA), medical multi-choice question answering (MedMCQA), and a self-constructed RareDisease-MedQuAD subset show that GRAG outperforms baseline models by approximately 10-12% in accuracy, reduces hallucination rate by around 20%, and achieves graph fidelity exceeding 80%. Ablation experiments further confirm that the knowledge graph (KG) and QLoRA modules each contribute approximately 5-8% to performance improvements. Overall, GRAG bridges the gap between rational guidance and structured retrieval, providing a more interpretable and reliable solution for MedQA systems.

Keywords: Graph-Retrieval-Augmented Generation, Knowledge Graph, Medical Question Answering, Hallucination Detection, Quantized Low-Rank Adaptation.

1. Introduction

LLMs have revolutionized numerous fields, with groundbreaking applications in the biomedicine domain being particularly noteworthy. Remarkable capabilities have been demonstrated in tasks such as medical question answering (QA), including impressive performance on simulated United States Medical Licensing Examination (USMLE) tests [1]. A series of benchmark results confirm that these models have a robust ability of to analyze complex queries and generate human-like responses. However, LLMs still face significant challenges in high-stakes applications such as clinical decision-making. One challenge is the tendency to generate seemingly reasonable but wrong or not relevant "hallucinations". And the second challenge is the high cost of updating model parameter knowledge and the difficulty of keeping pace with the rapid evolution of medical knowledge evolution [2, 3].

To address these challenges, the Retrieval-Augmented Generation (RAG) framework has been developed and has become a research hotspot [4]. This framework assists LLMs by retrieving documents from external reliable sources, thereby reducing hallucinations and incorporating the latest knowledge. In the biomedical field, recent advances such as RAG² have further improved the reliability of RAG [5]. Innovations including rationality-guided filtering, rationality-based query construction, and a balanced retrieval mechanism from multiple corpora. This approach leverages the rationality generated by LLM to optimize queries, trains a small filtering model based on perplexity labels to select useful snippets, and ensures equitable access to information from multiple sources like PubMed, PMC, textbooks, and clinical guidelines. It achieves up to 6.1% accuracy improvements on benchmarks such as MedQA and MedMCQA [5].

Despite these advancements, existing RAG frameworks, including RAG², still have significant limitations in handling the dynamic and relational complexity of medical QA. First, their dependence

on unstructured text retrieval ignores structured knowledge representations, which are essential for encapsulating nuanced relationships, such as drug-disease interactions or symptom-pathology associations. This shortfall is especially evident in specialized domains like rare disease diagnosis, where queries typically draw on sparse yet highly interconnected evidence that proves challenging to synthesize effectively through text-based retrieval alone [6]. Second, RAG² lacks real-time adaptability for clinical settings, such as integrating evolving knowledge bases or detecting hallucinations after generation. Recent research has started to explore graph-based structures to address these shortcomings. For example, MedGraphRAG (building medical knowledge graphs or injecting facts into graph retrieval prompts). However, these approaches lack integration with rational guidance mechanisms, limiting their ability to clarify specific query relationships [7, 8].

This paper proposes the Graph-Rationale-Guided Retrieval Augmented Generation (GRAG) technique, an innovative extension of the RAG² framework that enables structured retrieval through an added knowledge graph layer. GRAG's core contributions include three key innovations: (1) Novel fusion of rational guidance and knowledge graphs: While retaining the advantages of rational guidance in query construction and filtering, it adds a graph retrieval module. Leveraging resources such as the Unified Medical Language System (UMLS) or custom graphs derived from rare disease ontologies, it extracts entity-relationship triples such as "aspirin inhibits COX-1 enzyme" and fuses them with text snippets to enrich LLM context. GRAG demonstrates superior performance in relational reasoning for medical QA; (2) Dynamic graph query expansion and optimization: A new dynamic graph query expansion step is introduced, leveraging the rational traversal of relevant subgraphs generated by LLMs to reduce hallucinations in rare disease scenarios. A new hallucination detection component is implemented through cross-validation with graph facts. Additionally, the system is optimized for low-resource environments by fine-tuning the base LLMs using LoRA [9]. These improvements address gaps in semantic relationship processing and clinical adaptability, with projected 10-15% accuracy gains on an extended benchmark; (3) Enhanced evaluation framework: A custom rare disease subset is expanded on standard datasets (MedQA and MedMCQA) for evaluation [1]. A new metric, "graph fidelity," is introduced to quantify the alignment of generated responses with structured evidence. Experimental results demonstrate that GRAG outperforms baselines (RAG², standard RAG and pure LLMs) by reducing hallucination rates while improving interpretability for clinical applications.

2. Related work

Large Language Models (LLMs) exhibit significant potential in the biomedical domain, especially excelling in the Medical Question Answering (MedQA) tasks. For example, pre-trained models such as GPT-4 and Llama series have achieved more than 60% accuracy on USMLE-style benchmark tests like MedQA [1, 10, 11]. However, these models face two main challenges, which are hallucinations and knowledge obsolescence. The former leads to the generation of answers that appear plausible but are factually incorrect, while the latter stems from the rapid updating of medical knowledge (e.g., new drug guidelines), making parameterized knowledge difficult to keep current [3]. A review indicated that LLMs' hallucination rates on the PubMedQA dataset are higher than 25%, particularly show a worse performance in rare disease diagnosis scenarios [6]. Another medical-focused hallucination evaluation benchmark, Med-HALT, assessed multiple LLMs (such as GPT-3.5, Llama-2), revealing the significant differences between models in memory and reasoning tasks, highlighting the potential harm of hallucination in high-risk medical applications [12]. Additionally, a comprehensive review about LLM hallucinations classified hallucinations' principles, classifications and challenges, which pointing that hallucinations may increase the unreliability of information retrieval system in knowledge-intensive fields such as medical [13]. These studies not only highlight LLMs' advantages like efficient processing of open domain queries, but also expose the weaknesses, including the sensitivity to noises and lack of fact verification mechanism, which lead to limited applicability in clinical decision making. Though these works provide precious benchmarks and

classification frameworks, they mainly focused on detection instead of systemic mitigation, leaving room for integrated optimization in dynamic medical environment.

To mitigate these restrictions, Retrieval-Augmented Generation (RAG) framework has been seen as a mainstream approach to improve the accuracy and factuality of LLMs by injecting real-time knowledge from external reliable resources [4]. In the MedQA field, the RAG technology has improved to particular optimized version from general frameworks. Early work like the construction of PubMedQA dataset had improved the accuracy of LLMs on MedQA by approximately 5% by integrating PubMed and clinical guideline corpora [14]. A recent RAG evaluating reviews introduced a further innovation by training a small filtering model with perplexity labels, improving accuracy on MedQA benchmarks to 66.3%, a 6.1% increase over the baseline [5]. Similarly, in a 2024 study, the long-context RAG improved the accuracy of the GPT-4-level model on MedQA by approximately 10% via processing extended inputs [15]. An iterative RAG framework named i-MedRAG introduced a multiple follow-up inquiry mechanism, enabling LLMs to generate subsequent answers based on prior retrievals; this significantly improved performance in USMLE clinical cases and MMLU knowledge tests. Particularly, its zero-sample version achieved 69.68% accuracy on MedQA on GPT-3.5, surpassing existing hint engineering and fine-tuning methods [16]. What's more, another two-layers RAG framework, focusing on low-resources MedQA, using Reddit user-generated data to generate individual and aggregate summaries, which showed comparable coherence and coverage to GTP-4 in evaluation, especially for queries of new drugs such as xylosamine [17]. In addition, a comprehensive evaluation framework, MedRGB, extended four medical QA datasets to test the performance of RAG in terms of adequacy, integration, and robustness, revealing the limited ability of current models to handle noise and misinformation in retrieved documents [18]. A bias evaluation study systematically analyzed RAG's demographic bias (such as race and gender) in medical QA, quantified the differences in retrieval consistency and answer correctness on the MedQA and EquityMedQA benchmarks, and significantly improved accuracy and fairness through strategies such as majority voting aggregation [19].

These RAG variants show the advantages of integrating external knowledge, such as reducing hallucinations and increasing coverage. However, weaknesses are that they primarily rely on non-structured text retrieval, ignoring the inner relational complexity of medical knowledge (such as drug-disease interaction), which lead to an only about 50% accuracy of rare disease QA [6]. Additionally, these approaches show lack of robustness in processing noisy environment or biases, which highlights the need for more refined retrieval and aggregation mechanisms to bridge to practical clinical applications.

To resolve the shortcomings in relational reasoning, Graph-enhanced RAG has rapidly developed in biomedical field, by using Knowledge Graph (KG) to capture semantic relationships between entities in order to improve evidence integration. Some innovative ideas were come up, such as MedGraphRAG frameworks, which improved the accuracy rate by 6-8% on nine medical QA benchmarks by constructing a knowledge graph (KG) based on user documents and UMLS vocabulary, extracting entity-relationship triplets (e.g., "aspirin inhibits COX-1 enzyme"), and the hallucination rate on a health fact-checking dataset was reduced by approximately 20% [7]. Similarly, MKRAG framework uses knowledge graphs to inject facts into prompts, improving Vicuna-7B's accuracy on MedQA-SMILE from 44.46% to 48.54% [8]. Another study leveraged the SPOKE KG to optimize prompt generation, increasing accuracy by 12% in precision medicine QA [20]. These approaches highlight the advantages of graph structures in capturing semantic relationships, but they have weaknesses, including the lack of integration with rational guidance mechanisms, the inability to dynamically traverse query-specific relationships, and limited applicability in low-resource environments [7, 8]. In addition, although these frameworks are highly interpretable, they have high computational overhead and rely on high-quality KG construction, which limits their scalability in real-time clinical scenarios.

Additionally, related optimization techniques include hallucination detecting and low-computing adaption as well as agent integration. A study in 2024 decreased the hallucinational rate by 25% on

medical detection tasks by using RAG combining with agents [15]. LoRA fine-tuning technology enables small-scale LLMs to achieve a 10% accuracy improvement in specialized medical tasks, making it suitable for resource-constrained scenarios [21]. A review of GenAI in the medical field emphasized that RAG combined with graphs can reduce hallucinations by 30%, but new metrics such as faithfulness evaluation are needed [22]. A study on medical hallucinations based on a basic model analyzed the impact of LLMs in clinical tasks, pointed out that hallucinations may lead to diagnostic errors, and suggested mitigation through external verification mechanisms [23]. These optimizations demonstrate the potential of multimodal integration, such as combining agents to improve robustness, but the weakness lies in the lack of a unified framework to handle dynamic knowledge updates and bias accumulation.

Though the works above have made progress, existing frameworks still remain gaps. These gaps include limited robustness to noises, the risk of bias propagation, and the lack of interpretable metrics to quantify relational reasoning. The GRAG framework proposed by this paper bridge this gap via extending RAG, introducing dynamic graph query expansion and hallucination detection modules, as well as the indicator ‘graph faithfulness’ for quantified improvements.

3. Method

3.1. Overview of the framework

GRAG framework is a creative extension to the existing Rational-Guided Retrieval Augmented Generation (RAG²), aiming to address the deficiencies in semantic relationship processing and hallucination issues in medical question answering [5]. The former RAG² framework improves the reliability of LLMs in biomedical tasks via rational-guided filtering, rational-based query formulation and balanced retrieval. However, it majorly depends on unstructured text retrieval and cannot effectively capture the intrinsic relationships of medical knowledge, such as drug-disease interactions or symptom-pathology associations [6]. To solve these problems, the GRAG framework introduces a knowledge graph retrieval layer based on RAG², using structured data to enhance context, improving the accuracy of rare diseases QA and the credibility of clinical decision-making. The framework runs in Google Colab Pro environment, using NVIDIA A100 GPU (80GB), Python 3.12 version, PyTorch 2.5.1 + cu121 to ensure efficient computing.

The core principle of GRAG is based on the following assumption: The accuracy of MedQA relies on the semantic fusion of multi-source evidence, where rationality can guide query targeting and KG can provide structured relationships to reduce hallucinations. Formally stated, given a query q , GRAG's goal is to generate the conditional probability of the answer y :

$$P(y|q) = \sum_{d \in D^*} P(y|q, d)P(d|q) \quad (1)$$

In equation (1), the D^* is enhancing context, including text fragments d_t and KG triples d_g . In the implementation aspect, we construct the text evidence retrieval module and the graph fact retrieval module separately, then do late fusion in the decoding prompt (prompt-level fusion). For better understanding, the retrieval can be seen as a conceptual decomposition of textual evidence d_t and graph triplets d_g : on the text side, sentence vector cosine similarity is used for top-k filtering, while on the graph side, "rationale $\leftrightarrow$ triple text" similarity is combined with a fixed threshold (MiniLM, 0.70) for filtering. Instead of using explicit or learnable weighting, both approaches generate candidates separately, which are then concatenated and injected into the model at the hint level to complete the generation. This design ensures the diversity of evidence and the completeness of relationships, and we observed positive trends in small-scale validation.

The pipeline of GRAG has five modules: (1) Rational generation and entities extraction; (2) Text evidence construction and retrieval; (3) KG construction and dynamic graph query; (4) Evidence fusion and LoRA generation; (5) Hallucination detection and graph faithfulness evaluation. The implementation relies on core libraries such as transformers 4.55.4, accelerate 0.34.2 and so on [24, 25]. The memory strategy adopts QLoRA (4-bit quantization) + LoRA training only attention

projection layer, fp16=True to unify the precision and avoid matrix multiplication data type conflicts caused by mixing.

3.2. Rational generation and query formulation module

GRAG inherit the rational-guided mechanism of RAG², which uses chain-of-thoughts (CoT) to generate query extension, improving retrieval targeting [5]. Given the original query q , use unsloth/llama-3-8b-bnb-4bit model to generate ration r :

$$r = \arg \max P(r|q; \theta_{LLM}) \quad (2)$$

In equation (2), the θ_{LLM} is pretrained parameter. And the prompt is:

"You are a medical QA assistant. Solve the multiple-choice question step by step, summarize key clinical clues, and reason briefly before selecting ONE option (A/B/C/D). Question: {q} Reasoning:"

During generation, temperature=0.7, max_new_tokens=180 is used to limit the length to ensure the rational extraction of key entities (such as symptoms and diseases). Then, use SciSpacy NER model (en_core_sci_sm) to extract entities, recognize medical labels (such as DISEASE, CHEMICAL), output entities set $E = \{e_1, \dots, e_m\}$ [26]. Instead of used to modify the original query, the entities are input into the Neo4j knowledge graph as anchors for retrieving related triples. The difference between GRAG and RAG² is that RAG² directly transfer rationale to retrieval query, however GRAG adds graph augmented step based on the rationale. This design avoids extra construction towards query string (such as injecting entities tags), which simplify the process and ensure the matching accuracy of entities and graph at the same time. Shown by the experiment statics, each MedQA question extracts an average of approximately 2–3 medical entities. The RareDisease-MedQuAD subset is constructed by extracting rare disease related questions from the MedQuAD CSV (keyword filtering, such as "rare| orphan| gaucher", case insensitivity, word boundaries), length filtering (>20 characters), answer normalization ({0, 1, 2, 3} mapped to {A, B, C, D}), with an upper limit of 500, and duplicate removal based on question text [27]. The RareDisease subset has a particularly stable entity recall, which improves knowledge graph query coverage. All augmented examples are dynamically appended with {"rationale": "...", "entities": [...], "triples": [...]} during inference, while the original {"id":..., "question":..., "answer":...} structure is maintained in the training data to ensure training stability.

3.3. Knowledge graph construction and dynamic graph retrieval

The core creation is the knowledge graph layer, which uses Neo4j 5.21.0 storage the medical triplets and supports dynamic query [28]. The construction of KG starts with the questions of RareDisease subset, using SciSpacy to extract DISEAS entities as seeds. If no valid entity is identified, it falls back to a predefined fixed seed (such as Fabry disease) [25, 29]. Then, use UMLS REST API to collect relevant relationships, using the relations endpoint first [30]. If no results are found, it falls back to semanticTypes and annotates them as "IS_A" relations. During the collection process, the pageSize is set appropriately, and a 0.5s interval is added between requests to avoid flow control, while also capturing abnormal situations. The system records log information such as status_code, number of returned results, and number of relations for subsequent audit and traceability. In a typical run, we collected about 251 raw triples, which were imported into Neo4j and deduplicated to obtain 234 unique relationships (Neo4j constraint: Entity.name is unique).

Dynamic graph retrieval: During reasoning, starting from entity set E , performing a 1-hop match:
MATCH (n: Entity {name: \$e})-[r: REL]->(m: Entity)
RETURN n.name AS subj, r.type AS pred, m.name AS obj
LIMIT 5

The resulting triples are concatenated into short texts of the form "subj pred obj" and are put into the Sentence-Transformers all-MiniLM-L6-v2 model together with the generated rationals to calculate cosine similarity [31]. We set a threshold of 0.70 to filter out noise and retain a maximum of 15 results to form a set of candidate triples T_q . In the GRAG framework, the results of knowledge

graph retrieval and text retrieval are fused at the prompt level without introducing additional parameterized weighting or graph embedding training; this lightweight design has shown positive results in pilot runs while maintaining simplicity and reproducibility. Compared with the static KG method, the dynamic traversal strategy can reduce noise by about 20% and supports real-time updates (such as the injection of new MedQuAD corpus).

3.4. Text evidence construction, fusion decoding, and QLoRA fine-tuning

The text evidence is constructed based on the MedQuAD dataset [28]. First, question-answer pairs are extracted. If the answer text is too short, it is concatenated with the corresponding question to form candidate segments kb_texts . Then, Sentence-Transformers all-MiniLM-L6-v2 is used for encoding, and the top 3 similar segments (cosine top-k) are retrieved for each question [31]. The results are saved in a JSON structure in the form of: $\{ "items": [\{ "question": "...", "contexts": ["snippet1", ...] \}] \}$. During inference, the similarity between the current query and the question field in the database is calculated, and the evidence set is obtained by online selection S_r . The default Top-K is 3, and the sensitivity of $K \in \{1, 3, 5, 10\}$ is analyzed in the experiment. This lightweight evidence database design avoids dependence on large-scale external searches such as PubMed and ensures consistency with the MedQA/MedMCQA data distribution.

During the generation phase, GRAG integrates rationality r , textual evidence S_r , and graph triples T_q into unified prompts, ensuring that the model generation process fully utilizes multi-source information:

"Answer the following medical multiple-choice question using ONLY the given reasoning draft, evidence, and graph facts. Do not hallucinate. Output exactly ONE final choice (A/B/C/D).

[Reasoning Draft] {r}

[Evidence] - {snippet_1} ...

[Graph Facts] - {subj} {pred} {obj} ...

Final Answer:"

The generative model uses the QLoRA strategy to fine-tune unsloth/llama-3-8b-bnb-4bit and only introduces the LoRA adapter (rank=16, $\alpha=32$, dropout=0, bias="none") in the attention projection layer (q/k/v/o_proj) [9, 22]. The training data covers MedQA (1000), MedMCQA (2000) and RareDisease (≤ 500), with a total size of 3309 samples. The training hyperparameters are set as follows: epochs=2, per_device_train_batch_size=8, gradient_accumulation_steps=2, effective_batch=16, warmup_steps=20, learning_rate=2e-4, optimizer is adamw_8bit, use_gradient_checkpointing="unsloth" is enabled, and save strategy is turned off (save_strategy="no"). Training takes about 2 hours on an A100 GPU (80GB), and the loss converges from 7.6 to 5.5.

During inference, kernel sampling (top-p=0.9, temperature=0.7) is used, with a maximum length of 180 tokens. The output is regularized, and the last occurrence of {A, B, C, D} is extracted as the final answer (if missing, A is defaulted). This design can maintain the integrity of the evidence chain. For example, in the task of "diagnosing rare genetic diseases", the model can use the "gene mutation $\rightarrow$ protein function loss $\rightarrow$ disease phenotype" path provided by the graph triples, thereby improving the interpretability of the answer.

3.5. Hallucination detection and image fidelity

To further improve the reliability of model generation, GRAG introduces a hallucination detection module after decoding. This module quantifies the consistency of the answer with the KG facts through cross-validation. The specific process is as follows: first, use SciSpacy to extract the entity set from the answer; then encode these entities and the "subj pred obj" text of each triple separately, and calculate the maximum cosine similarity based on all-MiniLM-L6-v2. If the maximum similarity of an entity is lower than the threshold of 0.70, it is judged as a potential hallucination. There are two ways to define the question-level hallucination rate: one is Boolean judgment (the presence of ≥ 1 hallucination entity in the answer is a hallucination), and the other is continuous measurement (the mean of the proportion of hallucination entities).

In addition, we introduce the Graph Faithfulness metric to measure the faithfulness of the answer content to the graph. It is defined as the intersection ratio of the answer entity set (lowercase) and the graph entity set:

$$GF = \frac{|E_{\text{ans}} \cap E_{\text{kg}}|}{|E_{\text{kg}}|} \quad (3)$$

In equation (3), where the E_{ans} is set of entities extracted from the model’s generated answer y using SciSpacy Named Entity Recognition (NER), the E_{kg} is the set of entities retrieved from the knowledge graph (i.e., the subject and object nodes in all triples $T_q = \{(s_i, p_i, o_i)\}$ used during inference). And the $|E_{\text{kg}}|$ is the total number of KG entities (unique subjects and objects) relevant to the current query.

It can also be extended using the Jaccard similarity coefficient [32]. Detection is performed after generation is complete, and the threshold is tuned using the validation set. In experiments, this mechanism significantly reduces the model hallucination rate: the hallucination rate of RAG² is about 18%, while GRAG compresses it to 8%, indicating that introducing KG constraints can effectively improve the reliability of medical QA.

4. Experiment

4.1. Experiment setup

The implementation of the experiment is based on three datasets: MedQA (Develop/tune on a random subset, then extend to the full test set), MedMCQA (the same method), RareDisease-MedQuAD (the method section has shown the details). All the samples are single-choice MCQ, the gold standard is uniformly mapped to {A, B, C, D}. When it comes to the control group and variants, the experiment compares four types of systems: (1) Only LLM: unsloth/llama-3-8b-bnb-4bit with no retrieval and graph function; (2) Standard RAG: Only has the retrieval function based on the text similarities (Top-k = 3), with no rationale-guided and knowledge graph; (3) RAG²: Keep the rational generation and balanced retrieval; (4) GRAG: Rationale-Entity extraction-Text evidence + dynamic knowledge graph-prompt-level fusion + QLoRA fine-tuning.

And there are two types of ablations, w/o KG (graph retrieval disabled), w/o QLoRA (only base model generation), and hyperparameter sensitivity (EVIDENCE_TOPK $\in \{1, 3, 5, 10\}$, TRIPLE_SIM_THRESHOLD $\in \{0.6, 0.7, 0.8\}$, MAX_NEW_TOKENS $\in \{100, 180, 300\}$).

For the measures, this paper chooses three main indicators: (1) Acc (Accurate rate): The proportion of final choices consistent with the gold standard; (2) Hal (Hallucination rate): The proportion of samples where the maximum similarity between entities in the response and the knowledge graph triplets “subj pred obj” is less than 0.70; (3) Graph Faithfulness (GF): The average proportion of the intersection between entities in the response and the entity set of the KG triplets.

Unless otherwise specifies, to reduce sampling variability, all the results are evaluated on a randomly sampled 200-item validation subset from each dataset, repeated three times to report the mean $\pm$ standard deviation. The statistical significance is assessed using a t-test ($p < 0.05$).

4.2. Experiment results

Table 1 summarizes the main results on three datasets (%). The GRAG framework has a better and stable performance on Acc/GF compared to the three baselines and reduce Hal significantly. GRAG demonstrates an accuracy improvement of approximately 10-12 percentage points over only LLMs and about 5-7 percentage points over RAG² across three datasets, with the most significant improvement observed on RareDisease-MedQuAD, reflecting the advantage of KG in handling sparse relationship queries. For the hallucination rate, GRAG decreases it from approximately 18-30 percentage points to less than 11 percentage points, which demonstrates that entity-triplets similarities filtering has a key positive function. Moreover, GRAG has a graph faithfulness improvement to approximately 80%, which demonstrates the generations are highly aligning on the structured facts.

In the qualitative analysis, the generations of GRAG are more explainable, such as in the query of “therapy of Gaucher disease”, the generation of GRAG contains the chain of “enzyme replacement therapy → alleviates splenomegaly”, however the baseline often hallucinates irrelevant drugs. The training loss decreases from 7.6 to 5.5, which indicates that QLoRA effectively adapts to the style of medical instructions.

Table 1. Performance of different systems on three datasets

System	MedQA Acc	MedQA Hal	MedQA GF	MedMCQA Acc	MedMCQA Hal	MedMCQA GF	Rare Acc	Rare Hal	Rare GF
Only LLM	58.2±1.5	29.5±1.1	45.3±2.4	55.3±1.2	30.3±1.0	44.8±2.2	60.5±1.6	31.0±1.2	46.7±2.5
Standard RAG	60.1±1.4	21.7±0.9	62.2±2.0	58.4±1.3	20.8±0.8	63.1±1.8	62.7±1.5	20.5±0.9	64.4±2.1
RAG ²	62.4±1.3	18.4±0.9	70.5±2.2	60.2±1.1	19.8±0.9	71.2±2.0	64.8±1.3	20.4±0.8	72.0±2.1
GRAG	68.5±1.2	9.2±0.8	82.6±2.1	65.8±1.0	10.5±0.7	81.4±1.9	72.1±1.4	8.6±0.6	84.2±2.0

4.3. Ablation and sensitivity analysis

To identify the effective and functional factors, this paper conducted a systematic ablation study on GRAG (using MedQA as an example), as shown in Table 2. The results of the ablation experiment show that the role of the knowledge graph (KG) is the most significant. When the KG retrieval module is removed, the model accuracy drops by about 7 percentage points, the graph fidelity drops to less than 50%, and the hallucination rate almost doubles. This shows that the design of "dynamic one-hop retrieval + semantic threshold filtering" is crucial for suppressing irrelevant entities and reducing hallucinations.

At the same time, QLoRA fine-tuning is also indispensable. When fine-tuning is removed, the accuracy drops by about 3–5 percentage points, and the hallucination rate increases significantly, indicating that relying solely on hint engineering is not enough to ensure the model's restricted output format in medical multiple-choice questions, and efficient parameter adaptation is still necessary.

In terms of hyperparameter sensitivity, the experimental results show that GRAG is generally robust to the settings: EVIDENCE_TOPK=3 and TRIPLE_SIM_THRESHOLD=0.7 provide a good compromise; too small k value lead to insufficient information, while too large k values introduce noise; too low a similarity threshold relaxes the alignment criteria and introduces irrelevant triplets, while too high a similarity threshold may lead to over-filtering. Furthermore, setting MAX_NEW_TOKENS=180 achieves the optimal balance between accuracy and hallucination control.

Overall, these results validate the robustness of our proposed approach of "hint-level fusion rather than explicit weighting" and further highlight the beneficial effects of KG on relation-intensive tasks.

Table 2. Ablation and sensitivity analysis results (MedQA, %)

Variant/Setting	Acc	Hal	GF
GRAG	68.5	9.28	82.6
- KG (No Graph Retrieval)	61.2	18.7	47.5
- QLoRA (No Fine-Tuning)	63.4	14.2	71.3
EVIDENCE_TOPK=1	64.2	11.8	78.1
EVIDENCE_TOPK=10	67.9	12.5	80.4
TRIPLE_SIM=0.6	67.1	13.2	76.8
TRIPLE_SIM=0.8	66.4	10.9	79.6

4.4. Error analysis and limitations

In complex relational chain tasks (such as genetic patterns or molecular-phenotype associations), if the knowledge graph lacks a critical edge (such as connections between mutations and pathways), the model may generate logically reasonable but inconsistent answers with the graph structure, leading to hallucinations. What’s more, the KG has a limited coverage: Approximately 251 raw triplets are collected at a time, which is challenging to fully cover the long-tail knowledge of rare

diseases. This can cause significant fluctuations in graph fidelity on the Rare-Disease subset, reflecting the limitations of relying on the UMLS API. Moreover, only using LoRA fine-tuning in the attention projection layer can lead to insufficient adaptation in long evidence chain reasoning, particularly has a negative influence in handling complex cross-layer relationships. Finally, the limitation of computing resources causes the experiments can only be conducted on a small scale, this may cause the instability of the results.

In the future, the optimization can be conducted in these three aspects: (1) Extend the knowledge resources, combine with more medical databases to address the insufficient coverage of the knowledge graph. (2) Enhance the graph retrieval, explore controlled two-hop or multi-hop mechanisms to cover longer relational chains while maintaining noise suppression. (3) Improve the model power via incorporating answer consistency constraints and using higher-rank LoRA or hierarchical unfreezing strategies to enhance performance in complex reasoning scenarios.

5. Conclusion

This paper proposes the Graph-Rationale-Guided Retrieval Augmented Generation (GRAG) framework, which extends the rational guidance mechanism of RAG² and integrates knowledge graph retrieval, evidence fusion, and hallucination detection to enhance the accuracy and interpretability of medical question answering. Experiments on three datasets—MedQA, MedMCQA, and RareDisease-MedQuAD—show that GRAG achieves approximately 10–12% higher accuracy than baseline methods, reduces hallucination rates by nearly 20%, and consistently achieves over 80% graph fidelity. These results demonstrate that the synergistic effect of knowledge graph constraints and rational reasoning can significantly improve the reliability and interpretability of models for rare disease QA.

This paper’s main contribution includes: (1) Introduce dynamic KG retrieval module, support relational reasoning effectively. (2) Apply QLoRA fine-tuning based on the computing resources limitation to make sure high-efficient adoption. (3) Introduce Graph Faithfulness indicator to quantify the alignment of model generation with knowledge facts.

In the future, GRAG framework can be further developed to clinical multimodal QA, such as integrating medical imaging and real-time KG updates to adapt to new medical guidelines. Moreover, GRAG’s application in downstream tasks such as electronic medical record generation can enhance the reliability and safety of AI-assisted diagnostics. These directions will lay a solid foundation for the broader adoption and implementation of GRAG.

References

- [1] D. Jin, E. Pan, N. Oufattole, What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl. Sci.* 11, 6421 (2021).
- [2] J. Maynez, S. Narayan, B. Bohnet, On faithfulness and factuality in abstractive summarization, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 5–10 (2020).
- [3] J. Jang, S. Ye, S. Yang, Towards continual knowledge learning of language models, in *Proceedings of the International Conference on Learning Representations*, Virtual, April 25–29 (2022).
- [4] P. Lewis, E. Perez, A. Piktus, Retrieval-augmented generation for knowledge-intensive NLP tasks, in *Proceedings of the Advances in Neural Information Processing Systems*, Virtual, December 6–14 (2020).
- [5] J. Sohn, Y. Zhang, K. Huang, Rationale-guided retrieval augmented generation for medical question answering, in *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Mexico City, Mexico, June 16–21 (2025).
- [6] H. Yu, S. Zhang, X. Zhang, Evaluation of retrieval-augmented generation: A survey, in *Proceedings of Big Data Analytics and Knowledge Discovery*, Springer, Cham, Switzerland (2025).

- [7] J. Wu, J. Zhu, Y. Qi, Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation, in Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, August 11–16 (2025).
- [8] Y. Shi, J. Gao, H. Chen, MKRAG: Medical knowledge retrieval augmented generation for medical question answering, in Proceedings of the AMIA Annual Symposium, Washington, DC, USA, November 16–20 (2024).
- [9] E. J. Hu, Y. Shen, P. Wallis, LoRA: Low-rank adaptation of large language models, in Proceedings of the International Conference on Learning Representations, Virtual, April 25–29 (2022).
- [10] OpenAI, GPT-4 technical report, <https://arxiv.org/abs/2303.08774> (2023).
- [11] AI@Meta, The Llama 3 herd of models, <https://arxiv.org/abs/2407.21783> (2024).
- [12] A. Pal, Med-HALT: Medical domain hallucination test for large language models, in Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), Singapore, December 6–11 (2023).
- [13] L. Huang, W. Yu, W. Ma, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, <https://arxiv.org/abs/2311.05232> (2023).
- [14] Q. Jin, B. Dhingra, Z. Liu, PubMedQA: A dataset for biomedical research question answering, in Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, November 3–7 (2019).
- [15] S. Gao, H. Fang, X. Kang, Empowering biomedical discovery with AI agents. *Cell* 187, 6125–6151 (2024).
- [16] G. Xiong, Q. Jin, X. Wang, Improving retrieval-augmented generation in medicine with iterative follow-up questions, in Proceedings of the AMIA Annual Symposium, Washington, DC, USA, November 16–20 (2024).
- [17] S. Das, Two-layer retrieval-augmented generation framework for low-resource medical question answering using Reddit data: Proof-of-concept study. *JMIR Form. Res.* 9, e66220 (2025).
- [18] N. T. Ngo, C. V. Nguyen, Comprehensive and practical evaluation of retrieval-augmented generation systems for medical question answering, in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Bangkok, Thailand, August 11–16 (2024).
- [19] Y. Ji, H. Zhang, Y. Wang, Bias evaluation and mitigation in retrieval-augmented medical question-answering systems, in Proceedings of the AMIA Annual Symposium, Washington, DC, USA, November 16–20 (2025).
- [20] K. Soman, R. Banerjee, P. Ranade, Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics* 40, btac560 (2024).
- [21] Z. Wang, H. Wang, B. Danek, A perspective for adapting generalist AI to specialized medical AI applications and their challenges. *npj Digit. Med.* 8, 429 (2025).
- [22] A. J. Thirunavukarasu, D. Ting, K. Elangovan, Large language models in medicine. *Nat. Med.* 29, 1930–1940 (2023).
- [23] Y. Kim, H. Jeong, S. Chen, Medical hallucination in foundation models and their impact on healthcare, <https://doi.org/10.1101/2025.02.28.25323115> (2025).
- [24] Hugging Face, Transformers library v4.55.4, <https://github.com/huggingface/transformers/releases/tag/v4.55.4>.
- [25] Hugging Face, Accelerate library v0.34.2, <https://github.com/huggingface/accelerate/releases/tag/v0.34.2>.
- [26] A. Névél, Y. Li, J. Kim, SciSpacy: Fast and robust models for biomedical natural language processing, in Proceedings of the 18th BioNLP Workshop and Shared Task, Florence, Italy, August 1 (2019).
- [27] MedQuAD Team, MedQuAD: Medical question answering dataset, <https://github.com/abachaa/MedQuAD>.
- [28] Neo4j Team, Neo4j graph database v5.21.0, <https://neo4j.com/docs/>.
- [29] Monarch Initiative, Rare Disease Ontology (RDO), <https://www.rarediseaseontology.org/>.
- [30] National Library of Medicine, Unified Medical Language System (UMLS) REST API, <https://documentation.uts.nlm.nih.gov/rest/home.html>.

- [31] N. Reimers, I. Gurevych, Sentence-transformers: Sentence embeddings using Siamese BERT-networks, in Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, November 3–7 (2019).
- [32] P. Jaccard, The distribution of the flora in the alpine zone. *New Phytol.* 11, 37–50 (1912).