MedLTRAG: Co-Augmentation of Knowledge Graphs and Large Language Models for Long-tail Medical Question Answering

1st Mian Wu
School of Software
Beihang University
Beijing, China
wumian@buaa.edu.cn

2nd Xu Wang
School of Software
Beihang University
Beijing, China
xuwang@buaa.edu.cn

3rd Jie Sun
Zhongguancun Laboratory
Beijing, China
sunjie@zgclab.edu.cn

4th Bing Wei
School of Cyberspace Security
Hainan University
Haikou, China
weibing@hainanu.edu.cn

***Abstract*—Despite the remarkable success of large language models (LLMs), their ability to effectively incorporate long-tail domain knowledge remains a fundamental challenge. In specialized domains such as medicine, reliable question answering requires structured and precise knowledge that is often poorly captured by purely parametric representations. We observe that the key limitation of existing LLM-based approaches to long-tail medical QA lies in a misalignment between how specialized knowledge is represented and how it is accessed during inference. To address this limitation, we propose MedLTRAG, a framework that tightly integrates LLMs with knowledge graphs (KGs) to enable grounded medical reasoning. MedLTRAG follows an offline-to-online paradigm: domain-specific KGs are first constructed from medical literature to encode specialized expertise, and are then selectively retrieved at inference time to support reasoning through a multi-stage beam search mechanism. By externalizing long-tail knowledge into structured representations and explicitly incorporating them into the inference process, MedLTRAG reduces the brittleness of purely parametric reasoning. Extensive experiments on multiple disease-centric medical QA benchmarks demonstrate that MedLTRAG consistently outperforms state-of-the-art methods. Beyond empirical improvements, this work outlines a generalizable paradigm for integrating structured knowledge with LLMs, offering a principled pathway toward robust reasoning in knowledge-intensive, long-tail domains.**

***Index Terms*—LLM-KG Co-Augmentation, Retrieval-Augmented Generation, Large Language Models, Long-tail Knowledge, Medical Question Answering**

I. Introduction

Large Language Models (LLMs) have achieved strong performance in natural language understanding and reasoning tasks [1]. Recent advances in the biomedical domain show that LLMs assist clinicians and researchers in complex diagnostic QA. However, reliably answering questions involving long-tail medical knowledge remains a fundamental challenge [2].

In medical QA, long-tail knowledge refers to specialized, low-frequency information that is critical for reasoning about specific diseases, symptoms, or treatments, yet is sparsely represented in large-scale corpora. Such knowledge is often scattered across heterogeneous medical literature, expressed through diverse terminologies, and connected by implicit semantic relations. As a result, long-tail medical knowledge is difficult to organize, retrieve, and reason over using conventional data curation pipelines [3], and remains poorly captured by purely parametric language models.

In this work, we identify the core bottleneck of long-tail medical QA as the disconnect between specialized knowledge representation and inference-time utilization. To address this issue, we propose MedLTRAG, a co-augmentation framework that tightly integrates LLMs and knowledge graphs for grounded medical reasoning. MedLTRAG is designed as a plug-and-play framework that can be seamlessly integrated with existing LLMs without additional training. Our contributions are summarized as follows:

- We propose MedLTRAG, a novel framework that tightly integrates LLMs with domain-specific KGs through an offline-to-online paradigm, enabling grounded reasoning over long-tail medical knowledge.
- We introduce a fine-grained entity sampling strategy together with a novel multi-stage beam search reranking mechanism, which progressively selects question-relevant subgraphs and aligns structured knowledge retrieval with the model's reasoning process.
- We construct three disease-centric long-tail medical QA benchmarks to fill the gap in specialized evaluation. Extensive experiments demonstrate that MedLTRAG consistently outperforms state-of-the-art biomedical LLMs and existing RAG baselines across diverse backbones.

II. Related Work

KG Construction and Completion with LLMs. Leveraging the generative power of LLMs, researchers have explored various methods for KG construction and completion. For instance, SF-GPT [4] utilizes an entity extraction filter and LLMs for subgraph fusion , while EP-RSR [5] introduces the "Entity Pair-Relation-Fact" paradigm for relation extraction. Our work distinguishes itself by proposing an LLM-KG co-augmentation framework specifically for long-tail medical question answering. **Injecting Knowledge Graphs into**

Corresponding author: sunjie@zgclab.edu.cn

LLM Reasoning. Advancing LLM capabilities have led researchers to inject structured graph information via prompt engineering to support biomedical reasoning. For example, MindMap [6] employs a dual-path strategy to create chain-of-thought evidence subgraphs , while DALK [7] utilizes a self-aware retrieval module to filter redundant data. However, these methods often rely on general-purpose KGs, limiting their effectiveness when the KG coverage is insufficient or semantically mismatched with the query. **Multi-Agent Systems in Medical Reasoning.** Multi-agent systems have recently entered the medical domain, with MAIN-RAG [8] utilizing multiple agents and adaptive filtering to reduce document noise. Similarly, ToR [9] employs tree-structured reasoning paths and cross-verification to ensure decision consistency. However, these multi-round agent discussions are prone to topic drift and information loss, and often lack process interpretability by focusing only on final diagnostic outputs.

In summary, although existing studies in the medical domain have separately explored the applications of LLMs and KGs, to the best of our knowledge, no prior work has investigated the potential of mutual co-augmentation between the two in long-tail medical knowledge scenarios. Given the high information density of domain-specific KGs, capturing question-relevant key information is essential [10]. Therefore, we propose MedLTRAG to leverage LLM-KG co-augmentation, enhancing knowledge representation and reasoning for long-tail medical question answering.

III. Methodology

This section elaborates on the proposed MedLTRAG co-augmentation framework. Fig. 1 illustrates the overall pipeline of MedLTRAG.

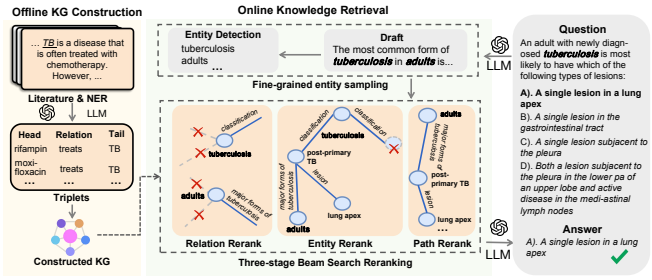


Fig. 1. **The overview pipeline of MedLTRAG.** X denotes the filtered-out knowledge during the reranking process. We first extract structured knowledge from unstructured corpora and construct a domain-specific knowledge graph (Section III-B). Then, we utilize a fine-grained entity sampling strategy combined with a novel three-stage beam-search-based reranking mechanism to select appropriate knowledge from the knowledge graph (Section III-C).

A. Background

High-quality and large-scale training corpora are crucial for LLMs to comprehensively capture disease-related details. However, real-world medical data exhibit imbalanced sample distributions. Taking leprosy, measles, and tuberculosis as examples, these diseases have relatively low incidence rates in developed countries and regions [11], resulting in their limited presence in mainstream public corpora and general medical texts. Nevertheless, in many resource-limited regions and developing countries, they continue to pose persistent challenges to public health systems and primary healthcare practice. Therefore, we select leprosy, measles, and tuberculosis as three representative long-tail medical diseases to systematically evaluate the applicability and effectiveness of the proposed approach in long-tail medical question answering.

B. Offline KG construction

1) Corpus Collection: To construct an accurate and comprehensive knowledge graph (KG), we systematically retrieved PubMed articles that primarily focus on leprosy, measles, and tuberculosis. To capture recent advances in the field and avoid the interference of outdated knowledge in earlier publications on KG quality, we restricted our search to articles published after 2005. Subsequently, we compiled keyword lists for leprosy, measles, and tuberculosis based on authoritative medical resources, specifically drawing from two main sources: (1) the World Health Organization (WHO); and (2) the Centers for Disease Control and Prevention (CDC). We then used these keyword lists to systematically filter the retrieved literature.

2) KG Construction: To construct an accurate and high-quality KG, we systematically categorize LLM-based KG construction methods into two major types for comprehensive comparison: (a) LLM-based relation extraction [7]; and (b) LLM-based generative construction [12], as illustrated in Fig. 2. We denote the KGs constructed by these two types of methods as KG_{re} and KG_{cons}, respectively.

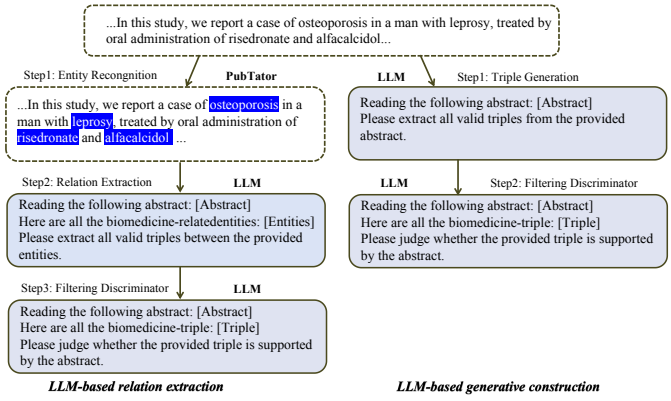


Fig. 2. **The detailed process of disease-specific KG construction.** The left panel illustrates the LLM-based relation extraction pipeline. The right panel depicts the LLM-based generative construction, with both methods employing a filtering discriminator to ensure quality.

The LLM-based relation extraction method, we utilize PubTator Central (PTC) [13], developed and continuously maintained by NCBI, to extract relevant entities from the corpus. PTC is a widely used biomedical entity recognition tool for PubMed literature, and supports six types of biomedical concepts, including genes, diseases, chemicals, mutations, species, and cell lines. The extracted entities are treated as nodes in the KG. Then, we design structured prompt templates to guide the LLM to take the document and its corresponding

entity set as input and identify semantic relations among the entities, which are then treated as edges in the knowledge graph.

The LLM-based generative construction method employs prompt templates to guide the LLM to directly identify all triples from the document. To reduce irrelevant or noisy triples, we further append a filtering discriminator after both construction methods. The document D together with a triple (e_h, r, e_t) is provided to the model as input. If the document contains the candidate triple relation, the model outputs "YES"; otherwise, it outputs "NO". The corresponding formulation is defined as $y_{\text{test}} = \text{LLM}(P_{\text{prompt}}, D, (e_h, r, e_t))$. Table I presents detailed statistics of the KG_{re} and KG_{cons}, respectively, including the number of corpora used and the numbers of nodes and triples.

TABLE I
DETAILED STATISTICS ABOUT KG_{RE} AND KG_{CONS}

	KG_{Leprosy}	KG_{Measles}	$KG_{\text{Tuberculosis}}$
#Corpus	6,319	5,832	9,332
#KG_{re}			
Nodes	7,389	5,081	9,174
Triples	16,173	10,129	20,388
#KG_{cons}			
Nodes	21,625	18,216	28,236
Triples	33,567	31,786	55,130

C. Online knowledge retrieval

1) Fine-grained entity sampling: To more effectively identify entities relevant to queries and enable accurate information retrieval from the KG, we introduce a fine-grained entity sampling strategy. Specifically, for a given query Q, a preliminary response draft is first generated using an LLM. Assertions requiring further verification are then extracted from this draft. Subsequently, the domain-specific entity set E is extracted from these assertions. All these steps are performed by the LLM without manual intervention.

We then follow MindMap [6] to perform similarity-based entity linking, aligning entities in the entity set E with those in the target knowledge graph G. Specifically, we use Sentence-BERT [14] to encode all entities in G and E as dense vector representations, denoted H_G and H_E, respectively. Then using cosine similarity, we link each entity in E to its nearest neighbor in G. This process yields the initial aligned entity set $E_G = \{e_1, \ldots, e_N\}$ for subsequent processing.

2) Three-stage beam search reranking: We observe that noise is a common challenge in automatically constructed KGs. To address this issue, we draw inspiration from beam search [15] and reranking techniques in the medical domain [16], and propose a three-stage beam-search-based reranking method. Specifically, we decompose the retrieval process into three stages, namely relation reranking, path reranking, and subgraph reranking, and adopt the medical pre-trained model MedCPT [17] as the ranking model to highlight key information relevant to the question and filter out noisy information. Furthermore, considering that the constructed graph exhibits strong topical relevance, we introduce a reranking mechanism at each retrieval step to improve search efficiency and reduce computational cost.

In each iteration of the beam search algorithm, the top-K reasoning paths are expanded, defined as $P_K^D = \{(e_{h,k}, r_k, e_{t,k})\}$, where $e_{h,k}$, r_k, and $e_{t,k}$ denote the head entity, relation, and tail entity, respectively, and D denotes the current iteration of the beam search process.

Relation rerank. This stage performs a beam search with depth 1 and beam width N from the current entity e_n to identify relevant relations r. For e_n, we collect all connected out-degree and in-degree relations to form a candidate set $R_{\text{cand},n}$. Then, we rerank the candidate relations in $R_{\text{cand},n}$ based on the semantic context of the input question and select the top N relevant relations. The corresponding formulation is defined as $R^D = \text{Rerank}\,(Q, \{(e_n, R_{\text{cand},n}), (R_{\text{cand},n}, e_n)\}, N)$.

Path rerank. First, we identify a new entity set E^D based on the selected relation r_n. Specifically, for entity e_n, we construct the candidate entity set, where the corresponding formulation is defined as $E_{\text{cand},n} = \{e \mid (e_n, r_n, e) \in G\} \cup \{e \mid (e, r_n, e_n) \in G\}$. To further optimize retrieval, we rerank the candidate paths $P_{\text{cand},n}$ obtained from entity e_n, retaining at most N paths with the highest semantic relevance, where the corresponding formulation is defined as $P^D = \text{Rerank}\,(Q, P_{\text{cand},n}, N)$.

Subgraph rerank. In the D-th iteration, all candidate paths $P_{\text{cand},n}$ are aggregated into a global candidate set P_{cand}. We rerank the paths based on their semantic relevance and select the top K paths, denoted as P_K^D, for subsequent iterations or as final evidence subgraphs, i.e., $P_K^D = \text{Rerank}\,(Q, P_{\text{cand}}, K)$.

Universality. When the maximum search depth is reached and the LLM evaluates that the current reasoning paths are insufficient to generate an answer, MedLTRAG generates an answer based solely on the inherent knowledge in the LLM. The pseudo-code is shown in Algorithm 1.

IV. EXPERIMENTS

This section systematically evaluates the effectiveness of MedLTRAG on long-tail medical question answering. Our experiments are designed to answer the following research questions. **Q1**: Can MedLTRAG consistently improve long-tail medical QA performance across different diseases and model backbones? **Q2**: How does MedLTRAG compare with existing biomedical LLMs and retrieval-augmented approaches? **Q3**: How do individual components of MedLTRAG contribute to its overall performance?

To answer these questions, we conduct comprehensive evaluations on disease-centric QA benchmarks, compare MedLTRAG with strong biomedical and retrieval-based baselines, and perform extensive ablation and sensitivity analyses.

A. QA Benchmark

Evaluating long-tail medical question answering requires benchmarks that emphasize specialized, low-frequency disease knowledge rather than common clinical facts. To this end, we construct three disease-centric QA benchmarks focusing on

Algorithm 1 MedLTRAG

Require: Questions $\mathcal{Q}$, Knowledge Graph G, max depth L
Ensure: Answer generated by MedLTRAG

```
for all q ∈ Q do
  E ← EXTRACT_ENTITIES(q)
  topic ← ENTITY_MATCHING(E)
  if topic = ∅ then
    return baseline LLM answer
  end if
  for d = 1 to L do
    R ← ⋃_{e∈topic} RELATION_SEARCH_PRUNE(e, G, q, N)
    P ← ⋃_{r∈R} PATH_SEARCH_PRUNE(e, r, G, q, N)
    if P = ∅ then
      return baseline LLM answer
    end if
    paths ← SUBGRAPH_PRUNE(P, q, K)
    if paths = ∅ then
      return baseline LLM answer
    end if
    (isRes, ans) ← REASONING(q, paths)
    if isRes then
      return ans
    else
      topic ← UPDATE_TOPIC(paths)
    end if
  end for
end for
```

leprosy, measles, and tuberculosis, which are representative long-tail diseases that remain clinically important yet underrepresented in general medical corpora.

We derive the benchmarks from two large-scale multiple-choice medical QA datasets, MedQA [18] and MedMCQA [19], comprising 205,878 questions of diverse biomedical topics. To extract long-tail disease-related questions, we adopt a two-stage filtering strategy. We first perform keyword-based matching using curated disease-specific vocabularies (Section III-B1). Then, we employ an LLM to semantically assess disease relevance and eliminate spurious matches. This process yields 435 leprosy-related, 443 measles-related, and 1,105 tuberculosis-related questions, forming three challenging long-tail QA benchmarks that emphasize structured medical reasoning over specialized knowledge.

B. Experiment Settings

Biomedical LLMs: We compare MedLTRAG with several representative biomedical language models, including BioMedGPT-LM-7B [20], Bio-Medical-Llama-3-8B [21], and Llama3-OpenBioLLM-70B [22]. These models are explicitly fine-tuned on biomedical corpora and serve as strong domain-specific baselines.

Retrieval-Augmented LLMs: We further include state-of-the-art retrieval-augmented approaches in the biomedical domain, including Almanac [23], Clinfo.ai [24], MindMap [6], DALK [7], and ToR [9]. These methods represent diverse RAG paradigms, ranging from text-based document retrieval to KG-guided reasoning.

Implementation Details: Unless otherwise specified, all experiments use GPT-5.1 as the backbone LLM. For text-based RAG baselines, documents are chunked into segments of up to 128 tokens, and the top-8 most relevant chunks are retrieved for inference. For KG-based methods, the same disease-specific KG is used to ensure fair comparison. In MedLTRAG, we use the precision-oriented KG constructed via LLM-based relation extraction approach KG_{re} by default. The beam width is set to N=8 and the top-K retained reasoning paths to K=15, with further analysis provided in Section IV-F.

C. Experimental Results

We first report the overall performance of MedLTRAG to answer **Q1** and **Q2**. Table II summarizes the results across the three disease-centric long-tail QA benchmarks.

As shown in Table II, MedLTRAG consistently achieves the best performance across all three datasets, outperforming both biomedical LLMs and existing retrieval-augmented methods. These results demonstrate that explicitly integrating structured domain knowledge into the inference process substantially improves long-tail medical QA performance.

TABLE II
RESULTS ON THE QA BENCHMARK.

Method	QA_{leprosy}	QA_{measles}	QA_{TB}
Biomedical LLMs			
MedFound-7B [25]	25.7	30.0	22.5
BioMedGPT-LM-7B [20]	31.5	34.3	37.6
Asclepius-Llama2-7B [26]	29.0	27.3	30.0
Meditron-7B [27]	25.3	23.9	27.8
Bio-Medical-Llama-3-8B [21]	63.7	66.4	72.0
Llama3-OpenBioLLM-70B [22]	75.6	68.8	75.1
Retrieval-Augmented LLMs			
Clinfo.ai [24]	81.1	82.6	82.7
Almanac [23]	78.2	80.1	79.9
MindMap [6]	81.9	82.6	83.2
DALK [7]	82.1	83.7	85.3
ToR [9]	<u>82.8</u>	<u>84.9</u>	<u>86.0</u>
GPT-5.1	80.0	81.9	81.9
MedLTRAG	**85.5**	**86.2**	**87.8**

Compared with text-based RAG approaches, MedLTRAG yields larger performance gains, indicating that structured subgraph retrieval provides more precise and less noisy evidence than unstructured document retrieval. Moreover, the improvement over vanilla GPT-5.1 highlights the limitation of purely parametric reasoning in long-tail settings and validates the effectiveness of the proposed co-augmentation paradigm.

We further observe that biomedical LLMs generally lag behind general-purpose LLMs, and we attribute this gap to their relatively smaller parameter sizes and insufficient task-specific fine-tuning. Although they may perform adequately in general medical scenarios, they fall short in long-tail medical settings that require stronger domain-specific knowledge, thereby limiting their generalization ability in specialized QA tasks. Overall, these experimental results demonstrate the effectiveness of the proposed method on long-tail medical

question answering tasks and further suggest that combining general-purpose LLMs with structured domain knowledge holds strong potential for biomedical question answering.

D. Ablation Study on the MedLTRAG framework

To answer **Q3**, we conduct an ablation study to quantify the contribution of each key component in MedLTRAG, including fine-grained entity sampling, adaptive iterative retrieval, and the multi-stage beam search reranking mechanism.

As shown in Table III, removing the multi-stage reranking mechanism leads to the most significant performance degradation, underscoring its importance in suppressing noisy knowledge and progressively aligning retrieved subgraphs with the reasoning objective. Removing adaptive iterative retrieval or fine-grained entity sampling also results in noticeable performance drops, indicating that accurate entity grounding and progressive retrieval are both essential for effective long-tail reasoning. Importantly, these results confirm that MedLTRAG's gains do not stem from naive knowledge injection, but from carefully structured and iteratively refined knowledge utilization.

TABLE III
ABLATION STUDY OF THE MEDLTRAG FRAMEWORK

Method	QA_{leprosy}	QA_{measles}	QA_{TB}
MedLTRAG w/o reranking in beam search	82.1	80.7	82.6
MedLTRAG w/o adaptive iterative	83.4	84.4	84.5
MedLTRAG w/o Fine-grained entity sampling	85.3	84.2	84.7
GPT-5.1	80.0	81.9	81.9
MedLTRAG	**85.5**	**86.2**	**87.8**

E. Ablation Study on KG Construction

We further investigate how different KG construction strategies affect downstream QA performance. Specifically, we compare the precision-oriented KG_{re} (LLM-based relation extraction) with the scale-oriented KG_{cons} (LLM-based generative construction).As shown in Table IV, MedLTRAG with KG_{re} consistently outperforms its KG_{cons} counterpart. Although KG_{cons} contains more entities and triples, its increased noise negatively impacts retrieval quality and reasoning effectiveness. These results highlight that, for long-tail medical QA, knowledge precision is more valuable than sheer graph size, and careful KG curation is essential.

TABLE IV
ABLATION STUDY RESULTS WITH KG_{re} AND KG_{cons}.

Method	QA_{leprosy}	QA_{measles}	QA_{TB}
GPT-5.1	80.0	81.9	81.9
MedLTRAG w/ KG_{cons}	79.7	80.5	80.3
MedLTRAG w/ KG_{re}	85.5	86.2	87.8

F. Hyperparameter Analysis

We analyze the sensitivity of MedLTRAG to key hyperparameters, including the beam width N and the Top-K candidate size. Fig. 3 reports results for $K \in \{5, 10, 15, 20\}$ and $N \in \{4, 6, 8, 10\}$. The results indicate that the performance initially improves as these parameters increase, but declines beyond a certain threshold due to the introduction of redundant and noisy subgraphs. This trade-off further motivates the design of the multi-stage reranking mechanism, which balances retrieval diversity and precision during inference.

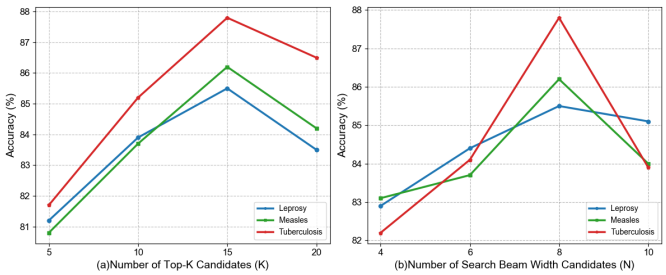


Fig. 3. **Hyperparameter Analysis.** (a) Effect of Top-K candidates on accuracy (K). (b) Effect of search beam width on accuracy (N).

G. Case Study

We present qualitative case studies in Fig. 4 and Fig. 5 to illustrate how MedLTRAG integrates structured knowledge into the reasoning process. In representative examples, MedLTRAG retrieves compact, question-relevant subgraphs that explicitly encode disease–symptom–treatment relations, enabling the LLM to arrive at correct answers that are difficult to obtain through unstructured retrieval alone. These examples qualitatively demonstrate how MedLTRAG grounds LLM reasoning in structured medical knowledge, reducing hallucinations and improving interpretability.

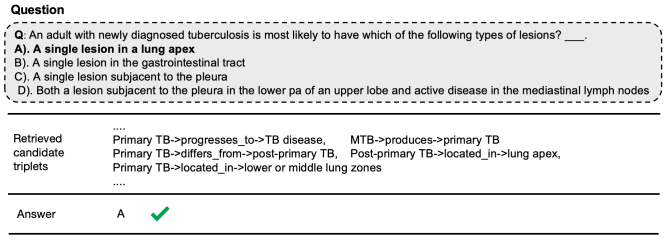
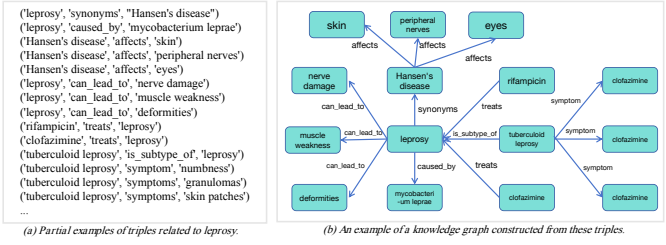


Fig. 4. **An Example for the Case Study.** MedLTRAG retrieves question-relevant subgraphs from the knowledge graph to guide the reasoning process, thereby enabling the large language model to produce the correct answer.



(a) Partial examples of triples related to leprosy. *(b) An example of a knowledge graph constructed from these triples.*

Fig. 5. **An illustrative example of knowledge extraction and visualization.**

H. Experiment Results with GPT-5.2

To evaluate robustness and generalizability, we further apply MedLTRAG to GPT-5.2, a stronger and more recent LLM. As shown in Table V, MedLTRAG continues to yield consistent performance improvements over the vanilla GPT-5.2 model. This result suggests that MedLTRAG is model-agnostic and complements advances in foundation models, making it well suited for future LLM generations.

TABLE V
PERFORMANCE ON GPT-5.2

Method	QA_{leprosy}	QA_{measles}	QA_{TB}
Vanilla GPT-5.2	81.4	79.5	84.0
MedLTRAG + GPT-5.2	86.4	86.9	89.7

V. CONCLUSION

We propose MedLTRAG, a framework integrating offline KG construction with online retrieval via fine-grained entity sampling and a three-stage beam search reranking strategy. Benchmarks demonstrate the effectiveness of this approach in enhancing medical QA reasoning.

VI. ACKNOWLEDGEMENT

This work was supported by the National Key Research and Development Program of China (Grant No. 2024YFC3308500) and by the National Natural Science Foundation of China (No. 62072017).

REFERENCES

[1] X. Wang, M. Zhang, X. Meng, J. Zhang, Y. Liu, and C. Hu, "Element-based automated dnn repair with fine-tuned masked language model," *Proceedings of the ACM on Software Engineering*, vol. 2, no. FSE, pp. 106–129, 2025.

[2] B. Lv, C. Tang, N. Liu, X. Liu, R. Zhang, and Y. Yu, "Mire: A medical information enhanced framework for long-tail medical dialogue synthesis," *Expert Systems with Applications*, p. 130905, 2025.

[3] Z. Wu, K. Guo, E. Luo, T. Wang, S. Wang, Y. Yang, X. Zhu, and R. Ding, "Medical long-tailed learning for imbalanced data: Bibliometric analysis," *Computer Methods and Programs in Biomedicine*, vol. 247, p. 108106, 2024.

[4] L. Sun, P. Zhang, F. Gao, Y. An, Z. Li, and Y. Zhao, "Sf-gpt: A training-free method to enhance capabilities for knowledge graph construction in llms," *Neurocomputing*, vol. 613, p. 128726, 2025.

[5] F. Zhang, H. Yu, J. Cheng, and H. Xu, "Entity pair-guided relation summarization and retrieval in LLMs for document-level relation extraction," in *Findings of the Association for Computational Linguistics: NAACL 2025*. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 4022–4037.

[6] Y. Wen, Z. Wang, and J. Sun, "MindMap: Knowledge graph prompting sparks graph of thoughts in large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 10 370–10 388.

[7] D. Li, S. Yang, Z. Tan, J. Y. Baik, S. Yun, J. Lee, A. Chacko, B. Hou, D. Duong-Tran, Y. Ding *et al.*, "Dalk: Dynamic co-augmentation of llms and kg to answer alzheimer's disease questions with scientific literature," in *EMNLP (Findings)*, 2024.

[8] C.-Y. Chang, Z. Jiang, V. Rakesh, M. Pan, C.-C. M. Yeh, G. Wang, M. Hu, Z. Xu, Y. Zheng, M. Das, and N. Zou, "MAIN-RAG: Multi-agent filtering retrieval-augmented generation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 2607–2622.

[9] Q. Peng, J. Cui, J. Xie, Y. Cai, and Q. Li, "Tree-of-reasoning: Towards complex medical diagnosis via multi-agent reasoning with evidence tree," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 1744–1753.

[10] P. Jiang, C. Xiao, A. Cross, and J. Sun, "Graphcare: Enhancing healthcare predictions with personalized knowledge graphs," in *ICLR*, 2024.

[11] "National organization for rare disorders (nord)," https://rarediseases.org/, 2026, accessed: 2026-01-25.

[12] B. Jimenez Gutierrez, Y. Shu, Y. Gu, M. Yasunaga, and Y. Su, "Hipporag: Neurobiologically inspired long-term memory for large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 59 532–59 569, 2024.

[13] C.-H. Wei, H.-Y. Kao, and Z. Lu, "Pubtator: a web-based text mining tool for assisting biocuration," *Nucleic acids research*, vol. 41, no. W1, pp. W518–W522, 2013.

[14] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[15] J. Wang and J. Han, "PropRAG: Guiding retrieval with beam search over proposition paths," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 6212–6227.

[16] M. Jia, J. Duan, Y. Song, and J. Wang, "medIKAL: Integrating knowledge graphs as assistants of LLMs for enhanced clinical diagnosis on EMRs," in *Proceedings of the 31st International Conference on Computational Linguistics*. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 9278–9298.

[17] Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, and Z. Lu, "Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval," *Bioinformatics*, vol. 39, no. 11, p. btad651, 2023.

[18] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.

[19] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering," in *Conference on health, inference, and learning*. PMLR, 2022, pp. 248–260.

[20] Y. Luo, J. Zhang, S. Fan, K. Yang, M. Hong, Y. Wu, M. Qiao, and Z. Nie, "Biomedgpt: An open multimodal large language model for biomedicine," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[21] "Contactdoctor-bio-medical: A high-performance biomedical language model," https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B, 2024.

[22] M. S. Ankit Pal, "Openbiollms: Advancing open-source large language models for healthcare and life sciences," 2024.

[23] C. Zakka, R. Shad, A. Chaurasia, A. R. Dalal, J. L. Kim, M. Moor, R. Fong, C. Phillips, K. Alexander, E. Ashley *et al.*, "Almanac—retrieval-augmented language models for clinical medicine," *Nejm ai*, vol. 1, no. 2, p. AIoa2300068, 2024.

[24] A. Lozano, S. L. Fleming, C.-C. Chiang, and N. Shah, "Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature," in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*. World Scientific, 2023, pp. 8–23.

[25] X. Liu, H. Liu, G. Yang, Z. Jiang, S. Cui, Z. Zhang, H. Wang, L. Tao, Y. Sun, Z. Song *et al.*, "A generalist medical language model for disease diagnosis assistance," *Nature Medicine*, pp. 1–11, 2025.

[26] S. Kweon, J. Kim, J. Kim, S. Im, E. Cho, S. Bae, J. Oh, G. Lee, J. H. Moon, S. C. You, S. Baek, C. H. Han, Y. B. Jung, Y. Jo, and E. Choi, "Publicly shareable clinical large language model built on synthetic clinical notes," 2023.

[27] Z. Chen, A. Hernández-Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, A. Sallinen, A. Sakhaeirad, V. Swamy, I. Krawczuk, D. Bayazit, A. Marmet, S. Montariol, M.-A. Hartley, M. Jaggi, and A. Bosselut, "Meditron-70b: Scaling medical pretraining for large language models," 2023.