Enhancing LLM Generation with Knowledge Hypergraph for Evidence-Based Medicine

Chengfeng Dou[1,2], Ying Zhang[3], Zhi Jin[1,2]✉, Wenpin Jiao[1,2]✉, Haiyan Zhao[1,2], Yongqiang Zhao[1,2], Zhengwei Tao[1,2]

[1] School of Computer Science, Peking University;

[2] Key Laboratory of High Confidence Software Technologies(PKU), MOE, China

[3] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China

{chengfengdou,zhijin,jwp,zhhy.sei}@pku.edu.cn

{tttzw,yongqiangzhao}@stu.pku.edu.cn {19112043}@bjtu.edu.cn

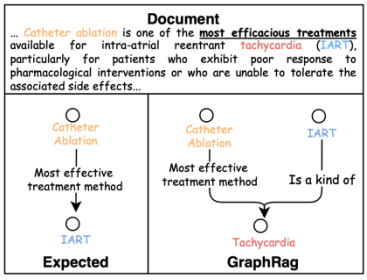

Figure 1: The phenomenon of mis-decomposition of complex relationships. LLMs omit the conditional variable "intra-atrial reentrant" for "tachycardia," leading to incorrect extraction.

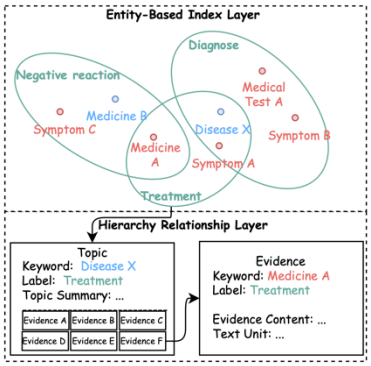

Figure 2: The Schema of EbmKG. The green ellipse denotes the hyperrelations corresponding to topic. The blue entities denote topic keywords, while red entities indicate evidence keywords. Evidence under the same topic has the same label.

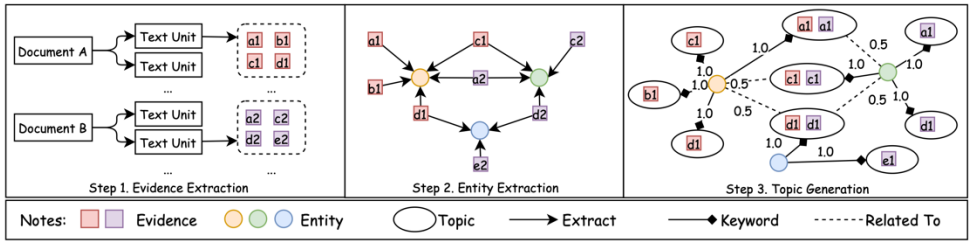

Figure 3: The construction process of EbmKG involves several key elements. Colored squares are utilized to represent evidence, with different colors distinguishing the sources of the evidence and the text within the squares indicating specific label of the evidence. Colored circles are employed to represent entities that are extracted from the evidence. White ovals represent topics, which are derived from evidence with the same aspect words.

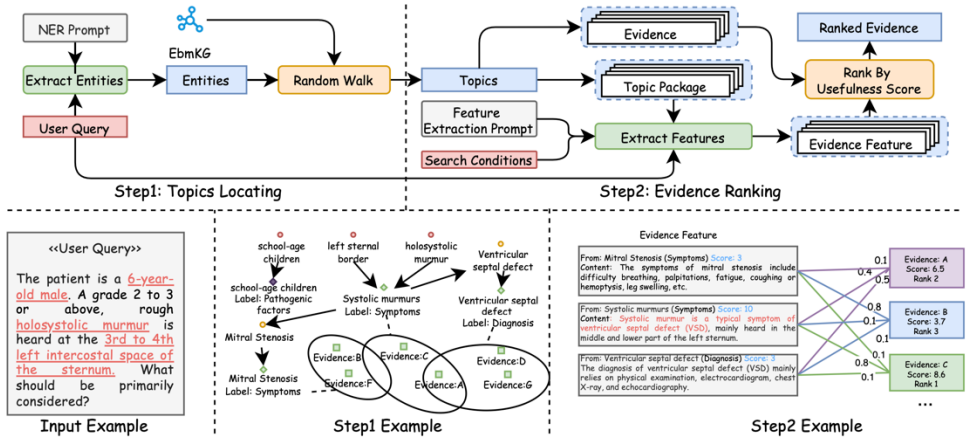

Figure 4: This figure is divided into two sections: the upper part outlines data processing, and the lower part provides an example. In 'Step 1 Example', entities in the EbmKG are shown as circles, topics as diamonds, and evidence as rectangles. Only green nodes, filtered by Personal PageRank, advance to Step 2. In 'Step 2 Example', an LLM assigns a Usefulness Score to each topic based on predefined Search Conditions, which determines the final evidence score. For NER and Feature Extraction Prompts, see Figure 8 (Search Words Extraction and Evidence Features Extraction). For Search Conditions, refer to Figure 9.

Evidence Extraction

Instruction: You are a useful medical AI assistant, please help me extract the information about {keyword}-{label} in the following document.
Document: {text_unit}

Entity Extraction

Document: {document}

Instruction: Please extract all the medical entities in the above snippet and return them according to the following template:
```json
[{{"name": ..., "type": ...,"mesh headings": [...]}}, ...]
```
Here is an explanation of the fields:
- name: The description of this entity.
- type: the type of the entity, the possible types of the entity are: [Disease, Symptom, Life Habits, Pathogen, Medical Test, Drug, Surgery].
- mesh headings: the corresponding ``Medical Subject Headings'' of the entity, one entity can correspond to multiple headings, please use List to return. If no mesh headings can be found for this entity, please rewrite the entity name according to the mesh naming convention.

Topic Summarization

Instruction: Please write a brief summary for the chapter {keyword}-{label} based on the following evidence.
Evidence: {evidence_list}

Search Words Extraction

Query: {User Query}
Task: Please extract the medical entities from the user input that will help in completing the task and try to extract more than 5 medical entities as more entities will help in the subsequent retrieval of evidence.
Output Format: [{{"name": ..., "type": ...}}, ...]

Evidence Features Extraction

Query: {User Query}
Task: To answer the user question, we looked at medical databases to find relevant information. Now after a rough ranking of documents, we have targeted the following documents, which you can consider as absolutely credible.
Documents: {Topic Chunk}
Please follow the rules below to identify the content in the document that helps answer the patient's question and score it based on relevance.
Rules: {Rules}
Output Format: {{"reference": ..., "score": ... }}

Figure 8: Prompts used by Graph Construction and Evidence Retrieval.

CMHD

Please review the reference materials and indicate whether they contain any content that demonstrates errors in the doctor's response.
These errors specifically refer to mistakes in medical knowledge, including incorrect diagnoses, medication errors, improper medical procedures, and inappropriate healthcare advice.
Please provide a score for the level of disagreement, with the score ranging from 0 to 10. A higher score indicates a stronger level of disagreement.

CMB-Clin

I would like you to identify the direct relevance between the reference materials and the doctor's questions. Please consider the following approach:

1. If the doctor is asking about diagnostic information, focus on whether the reference materials include diagnostic methods for the specific disease.
2. If the doctor is inquiring about treatment options, concentrate on the treatment plans provided in the reference materials.
3. If the doctor is seeking an overview of a disease, pay attention to the summary and review sections in the reference materials.

The scoring range is from 0 to 10, with higher scores indicating that the documents are more likely to address the doctor's questions effectively. If the reference materials do not directly provide the answer the doctor is looking for, please assign a score of 3 or lower.

MedQA and NLPEC

I would like you to evaluate the relevance of a reference to a multiple-choice question on a medical exam. Relevance in this context means that the reference provides evidence from the literature that supports or refutes one of the answer options.

Please assign a relevance score between 0 and 10, where a higher score indicates greater relevance.

DDA

I am providing you with the abstracts of reference documents. I would like you to evaluate whether these references help in identifying **unreasonable** decisions made by doctors and assign a score. Please consider the following aspects carefully:

1. If the decision is related to diagnosis, assess whether the references can help verify: cases of missed or incorrect diagnoses, and whether there was a failure to collect complete patient information.
2. If the decision is related to lifestyle advice, assess whether the references can help verify: if the advice includes content that could worsen the patient's condition or even endanger their life, such as ignoring contraindications of the disease.
3. If the decision is related to medication advice, assess whether the references can help verify: if there are errors in the knowledge provided, if there is a disregard for drug interactions, or if there is a neglect of potential side effects.
4. If the decision is related to medical care advice, assess whether the references can help verify: if the level of care recommended is appropriate, if emergency measures are suitable, if the recommended department is correct, and if the suggested tests are appropriate.

The scoring range is from [0-10] points. The higher the score, the more effectively the documents can highlight the **unreasonableness** of the doctor's decisions.

Figure 9: Search conditions for different datasets

 # Task Input
- **Input**: {input}
- **Golden Answer (Key Points)**: {key_points}
- **AI Response**: {output}

 # Task Description
You are tasked with evaluating the reliability of an AI model's response based on the provided key points in the golden answer. Your evaluation should be formatted as a JSON array, where each object represents the assessment of a specific key point.

 # Output Format
The JSON format should be as follows:
```json
[{"keypoint_number": ..., "contradict": ..., "description": ...}, ...]
```
Field Explanations:
- **keypoint_number (String):** The number or identifier of the key point. Ensure that all key points are accounted for.
- **contradict (Boolean):**
  - **True**: if the AI response misses the key point or contradicts the key point.
  - **False**: if the AI response correctly addresses the key point.
- **description (String):** A brief explanation of why the contradict field is set to True or False.

Figure 12: The prompt of key points evaluation

 # Task Input
- **User Query**: {user_query}
- **Document**: {document}

 # Task Description

Given the standard answer to the user's question: {golden_answer}, evaluate how useful the provided document is for answering the user's query.

 # Output Format

Return the evaluation in JSON format as follows:

```json
{"score": <int>, "reason": <str>}
```
Field Explanations:
- **score (Integer)**: A score between 0 and 10, indicating the usefulness of the document.
- **reason (String)**: A brief explanation justifying the score.

Figure 15: The prompt of Usefulness Evaluation

You are a powerful search engine evaluation assistant. I am using two search engines to retrieve relevant documents for a specific question. The question is as follows:
{question}

The search engines have returned their respective results:
[search_engine1]
{docs1}

[search_engine2]
{docs2}

The known correct answer to the question is: {answer}.

Please help me evaluate which search engine's results are better by considering the following dimensions:
1. **Recall**: Which search engine's results more directly provide the correct answer or are strongly related to the correct answer.
2. **Precision**: Which search engine's results contain fewer irrelevant or misleading pieces of information.

Note: Do not simply assess the quality of the documents based on the amount of information they contain. Instead, consider whether the documents provide direct assistance in answering the question.
Note: When evaluating the accuracy, only check if the content of the documents contains any incorrect information; do not consider whether the documents are relevant to the question.

Please return your results in the following well-structured format:
```json
{"recall": {"answer": ..., "exp": ..., }, "precision": {"answer": ..., "exp": ...,}}
```
Here is an explanation of the fields:
- **answer (Str)**: Choose one from ['Search Engine 1 is better', 'Search Engine 2 is better', 'Both are equally good'].
- **exp (Str)**: Please explain why you chose this value.

Figure 13: Prompt for calculating Recall and Precision.

<table>
<tr><td colspan="2">Test Sample

L4-5 central disc herniation involves ( ).
A: L4 nerve root.
B: L5 nerve root. (Graph Rag)
C: S1 nerve root.
D: Cauda equina nerve. (Native Rag, EbmKG Rag)
E: Femoral nerve.</td><td>Why Graph Rag is Wrong

An L4/L5 disc herniation does have the potential to compress a nerve root, but exactly which nerve root is compressed depends on the location and direction of the herniation. If the L4/L5 disc herniation is posterolateral, then it is more likely to compress the L5 nerve root. In the case of a centralized disc herniation, it may compress the spinal cord or cauda equina rather than a single nerve root.</td></tr>
<tr><td>Native Rag
From: Multiple radiculopathy: spinal stenosis, infectious, cancerous and inflammatory nerve root syndromes

Content: ...Cauda equina syndrome - A variant of lumbosacral polyneuropathy is cauda equina syndrome, in which the cauda equina is compressed within the spinal canal. The most common cause is a large centralized disc herniation or centralized spinal stenosis....</td><td>Graph Rag
Src: L4/L5 RADICULOPATHY
Dst: L5 NERVE ROOTS
Rel: L4/L5 RADICULOPATHY is a type of lumbar disc herniation that usually compresses the L5 nerve root.

Src: L5 RADICULOPATHY
Dst: Lumbar disc herniation
Rel: Lumbar disc herniation is a common cause of L5 radiculopathy</td><td>EbmKG Rag
From: Back pain in children and adolescents: Overview of causes
Keyword: Back pain in children and adolescents
Aspect: symptoms

Content: Low back pain, sciatica (pain radiating below the knee), limited spinal mobility, limited passive straight leg raising. Severe central disc herniation can lead to cauda equina compression, bladder dysfunction, and anesthesia in the saddle area.</td></tr>
</table>

Figure 16: Comparisons of results generated by various RAG aRandom Walkoaches.