RAG for Enhanced Medical Question Answering

Author:Su Yang

Student ID: 223040097

Email: 223040097@link.cuhk.edu.cn

Introduction

In this poster, I present a Retrieval-Augmented Generation (RAG) pipeline aimed at improving large language model (LLM) performance in medical question answering.

At the heart of this approach is a fact corpus—a curated set of concise, precise medical knowledge. While LLMs are pretrained on vast general data and excel at natural language tasks, their effectiveness in specialized fields like medicine often suffers from limited domain-specific knowledge.

A common solution is to fine-tune the LLM on domain-specific data, but this method faces two key challenges: difficulty in balancing generalization and specialization, and high computational and energy costs.

RAG addresses these issues by dynamically retrieving relevant facts and incorporating them into the model's prompts. This allows the LLM to apply its existing language skills while being guided by targeted knowledge.

The pipeline is implemented using the DeepSeek-R1 chat model and evaluated by comparing answer indices against correct references. Results show that this RAG approach significantly boosts task performance.

Methodology

An example for RAG prompting

Prompt:

System: You are an expert in doing QA exams. You need to answer some questions. You can refer to any examples, hints, and facts if provided.

Question:

Six days after undergoing surgical repair of a hip fracture, a previously healthy 79-year-old woman becomes agitated and confused. She is unarousable during the day but awake and impulsive at night, requiring frequent reorientation. Her husband reports that she usually drinks one to two glasses of wine per week. Her only current medication is oxycodone for pain. Vital signs are within normal limits. She appears distressed and is oriented to person but not to place or time. Neurologic examination reveals inattentiveness but no focal deficits. A urine dipstick test is normal.Which of the following is the most likely cause of her current condition?

A. Dementia
B. Opioid intoxication
C. Delirium
D. Urinary tract infection

Hint:

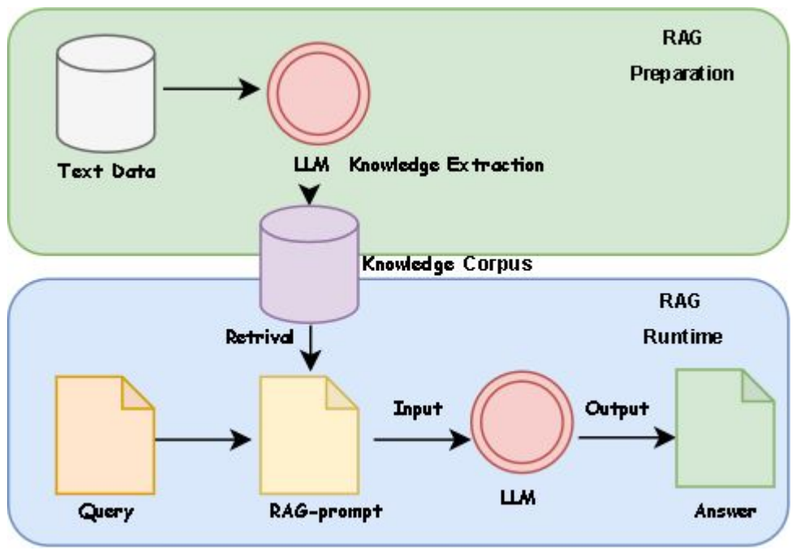
1. Dementia: Characterized by chronic, progressive cognitive decline. It typically does not present with sudden onset or fluctuating symptoms.
2. Opioid intoxication: Can cause sedation and confusion, but the symptoms are usually persistent rather than fluctuating.
3. Delirium: Presents with acute onset, fluctuating levels of consciousness, inattention, and disorientation. It is common in elderly patients following surgery.
4. Urinary tract infection: Can lead to delirium in older adults; however, a normal urine dipstick makes this diagnosis less likely.

Answer:

Expected Output:

C. Delirium.

Methodology



I. Fact corpus Building

- *One-shot similarity matching*
  The most straightforward way of RAG for questioning-answering task is embedding all collected training QA samples, and then retrieve the most similar one for the given test question to form one-shot prompting.
- *Option Hint*
  Instead of retrieving the most similar samples, option hint focus on building option hint for each option appear in all training QA samples. The option hint is a piece of illustrative text for the option given a question context, and it is generated by LLM with preciseness and conciseness.

II. Retrieval-Augmented Generation on MedQA

I use MedQA as the major dataset for the experiment. For option hint generation, I use Deepseek-R1 chat as generator. For one-shot similarity RAG, I use MiniLM-L6-V2 from SentenceTransformers as the encoder, and Faiss as vector database for embeddings' storage and retrieval. It should be noted that, option-hint's retrieval is implemented through direct key hashing in hashmap.

Results

RAG-approach	Accuracy	Total Questions
None	0.85	100
One-shot	0.83	100
Fact-corpus	0.91	100

The result shows that option-hint RAG greatly boost model performance with no need for training, which implies the importance of entity extraction and precise knowledge retrieval. One-shot similarity RAG performs worse than baseline, which is because it creates longer context, making it more difficult for semantic embedding, retrieval and LLM inference.