How to Improve RAG with Query Transformation

Three effective techniques: query refinement, query expansion, and query decomposition

Pedram Ataee, PhD · Follow

Published in AI Advances · 6 min read · 21 hours ago

102 1

Retrieval-augmented generation (RAG) is an effective method for answering queries by leveraging contextual data not included in the model's training data. This method is often used when the resources are unfit for model training, such as rapidly changing enterprise documents requiring high privacy. This method also decreases the chance of hallucination by narrowing the response context. RAG has several steps, including **document indexing**, **information retrieval**, and **response generation**. The information retrieval step aims to extract focused context or potential responses from resources before passing them to the response generator. In this article, I highlight query transformation techniques designed for information retrieval that improve RAG performance.

User queries often lack precise descriptions or language correctness. Plus, there is no guarantee that user queries match the wording of the resources. These differences impact retrieval quality, which can’t be resolved by any LLM in the response generation step. On another note, LLMs likely make mistakes when they aim to respond to a complex query in one step. We need to break down the original query into several sub-queries before aiming to answer the original query. These challenges can be addressed by a family of techniques named “**query transformation**.”

We use various techniques daily to better understand a question (simple or complex) and respond more effectively. For example, we use techniques to expand discussion perspectives, break down complex questions, and define template responses. Similarly, we can use these techniques to let machines answer questions better through more effective information retrieval. Several techniques were previously introduced to implement these techniques in advanced RAG systems, such as (1) **query refinement**, (2) **query expansion**, and (3) **query decomposition**. You should implement these techniques with extremely large language models such as `GPT4` or `Llama3 450B` to get the best performance. Most of the usual concerns around LLMs, such as privacy or cost, don’t apply to question transformation since, for example, questions often don’t include factual information. However, if you have restrictions on using those LLMs, you should use fined-tuned instruction-following LLMs to implement question transformation techniques. In this article, I provided prompts that are tuned for `GPT4` .

The article below is a good survey article that taught me a lot. You should also read it.

A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models

Abstract. As one of the most advanced techniques in AI, Retrieval-Augmented Generation (RAG) can offer reliable and…

arxiv.org

1 — Query Refinement

> *Query refinement is a technique to improve the clarity of the original query by rewriting it to different potential forms using the context. It also corrects potential spelling or grammatical errors.*

Users interacting with natural language interfaces don't follow any rules or guidelines. They just write whatever comes into their mind, which may (and will) include spelling and grammatical errors and lack enough details. Refinement helps the query become much clearer and machine-understandable.

```
prompt = """Given the context of each text, rewrite the following text into
a well-structured question by choosing the most likely type of sentence
(declarative, interrogative, imperative, exclamatory). Feel free to use extra
words or even synonyms to enhance the text. Do this 3 times with different
words but the same structure. Don't use polite words.

Text = serial numbers Emerson valves after 2020 high-pressure"""
```

```
Query:
"serial numbrs Emerson valves after 2020 high-presure"

Transformed (Refined) Qery:
1- Which serial numbers are used for Emerson high-pressure valves produced
after 2020?
```

```
2- What are the serial numbers for Emerson valves that handle high pressure
and were made post-2020?
3- Identify the serial numbers for Emerson valves manufactured after 2020 tha
are designed for high pressure.
```

2 — Query Expansion

> *Query expansion is a technique to broaden perspectiven by adding extra details or sample responses to the original query.*

Here, I want to highlight two common query expansion techniques: *query2doc* and *HyDE*. In query2doc, the original query is concatenated with the pseudo-documents generated by LLMs in response to the original query. The pseudo-documents often contain highly relevant information to aid query disambiguation and guide retrieval. [Read More] Similarly, HyDE uses LLMs to generate hypothetical documents in response to a query with one small difference. In HyDE, we don't concatenate the document with the original query. We purely use the hypothetical document for retrieval. [Read More]

Note that pseudo or hypothetical documents refer to the same concept with different terminology. In both techniques, we first generate documents that, although they may not answer the query, serve as a valuable example of a reasonable response. You can find a sample prompt to implement the query expansion below. If you want to use HyDE, you can use the native implementation of HyDE in Llamaindex framework.

```
prompt = """Determine the context of the query, and put yourself in the shoes
of an expert in that field. Then, try to answer the question concisely in a
paragraph form. Make sure to answer the question even though you don't know
all the details. Now, respond to the following question.
```

```
Question = How can one identify a malfunctioned high-pressure Emerson valve?"
```

```
Query:
"How can one identify a malfunctioned high-pressure Emerson valve?"

Transformed (Expanded) Qery:
To identify a malfunctioning high-pressure Emerson valve, start by observing
any unusual behavior such as erratic operation, inconsistent pressure reading
or unusual noises. Check for leaks around the valve body or connections, as
these can indicate seal or fitting issues. Additionally, monitor the valve’s
response to control signals; if it fails to open or close properly or does no
maintain the set pressure, it may be malfunctioning. Conduct a visual
inspection for any signs of physical damage or wear. Finally, consult the
valve’s diagnostic tools or software if available, as these can provide
detailed error codes or performance data. If these checks indicate issues,
it may be necessary to perform further diagnostic tests or contact Emerson fc
technical support.
```

3— Query Decomposition

> *Query decomposition is a technique for breaking down complex queries into simpler sub-queries. The results of these sub-queries are then combined to generate the final response.*

In real-world use cases, RAG pipelines must effectively respond to complex queries with multiple aspects, which also need some level of reasoning. In this case, we must break down these queries into simpler ones, allowing the solution to extract answers more targeted and efficiently. Each question may need to be answered using different resources, such as an SQL database or a collection of documents. This approach leads to more accurate answers. Let me give you an example.

```
prompt = """Break down a complex question into a list of sub-questions that
help answer the main question. Break down the original question minimally as
```

```
needed. Keep answer concise and limit only to the list of sub-questions.

Question - Extract the serial numbers of Emerson valves built after 2020
designed for high-pressure scenarios and comply with API 6A."""
```

```
Query:
"Extract the serial numbers of Emerson valves built after 2020 designed for
high-pressure scenarios and comply with API 6A."

Transformed (Decomposed) Qery:
- What are the serial numbers of Emerson valves built after 2020?
- Which of these valves are designed for high-pressure scenarios?
- Which of the high-pressure valves comply with API 6A standards?
```

Last Words

You can optimize information retrieval using query transformation techniques in the pre-retrieval stage. You can also re-evaluate the retrieved documents in the post-retrieval stage using query-document relevancy techniques such as BM25 or ReRank. Enhancing a RAG solution requires more than query transformation in the real world. In this article, I highlighted the most common query transformation techniques used in the industry. However, there are more techniques. Check out the query transformation cookbook created by Llamaindex.

Query Transform Cookbook - LlamaIndex

OpenAI Agent + Query Engine Experimental Cookbook

docs.llamaindex.ai

We can borrow many other methodologies from the literature to enhance information retrieval. **What do you think?** 🍄 I would like to hear your thoughts on this. Please share your thoughts and ideas in the comments.