Q&A

1.什么是RAG?

In the field of artificial intelligence, especially natural language processing (NLP), RAG refers to a method that combines information retrieval and text generation techniques. Retrieval-augmented generation(RAG) — models which combine pre-trained parametric and non-parametric memory for language generation.Parameter memory refers to the memory composed of the parameters of a model. In traditional neural network models, such as Transformer architecture, all knowledge is encoded into the weights (parameters) of the model. In RAG systems, non parametric memory typically refers to an external knowledge base or collection of documents from which relevant information can be retrieved to assist in the generation process. This knowledge base can be dynamically updated at runtime without the need to retrain the model.RAG can be ﬁne-tuned on any seq2seq task, whereby both the generator and retriever are jointly learned.

1. RAG是哪一年提出的概念，在什么背景下诞生了RAG？

The Retrieval Augmented Generation (RAG) model was proposed by researchers from Facebook AI Research (now Meta AI) in 2020. The background of this model is mainly to address the limitations of traditional generative and retrieval models.

Traditional generative models excel at generating smooth and context relevant text, but their knowledge is limited by training data and cannot be easily updated to include the latest information. On the other hand, retrieval models answer questions by retrieving relevant information from a pre built knowledge base. Although they can provide more accurate information, their expressive power is limited, making it difficult to generate complex and coherent long text answers.

To overcome these limitations, the RAG model combines the advantages of generative and retrieval models. It first finds relevant information fragments from a large number of documents through a retrieval component, and then uses this information to guide a powerful pre trained language model for text generation. This approach not only makes the generated text more accurate and fact based, but also maintains the flexibility and expressiveness of generative models.

1. 什么是KG和LLM？

KG (Knowledge Graph, Knowledge Graph）：

A knowledge graph is a graphical structure used to represent knowledge, consisting of entities and relationships between them. Each entity represents a specific object or concept, such as a person, place, organization, etc., while relationships describe the connections between these entities, such as "place of birth," "workplace," "founder," etc. Knowledge graphs organize information in a structured manner, enabling machines to understand and process complex relational networks.

LLM (Large Language Model), Large Language Model）：

Large scale language models refer to deep learning models with a large number of parameters, typically based on Transformer architecture, capable of handling natural language processing tasks such as text generation, translation, question answering, etc. This type of model learns the statistical rules of language by training on a large amount of textual data and can generate coherent and context relevant text. As the model size grows and the amount of training data increases, the ability of LLM is also constantly enhancing, enabling it to perform more complex language understanding and generation tasks. Famous examples include OpenAI's GPT series.

1. RAG的工作原理是什么？

The basic workflow of RAG:

1. Input processing: When receiving a query or request from a user, the input is first processed for use in subsequent steps.

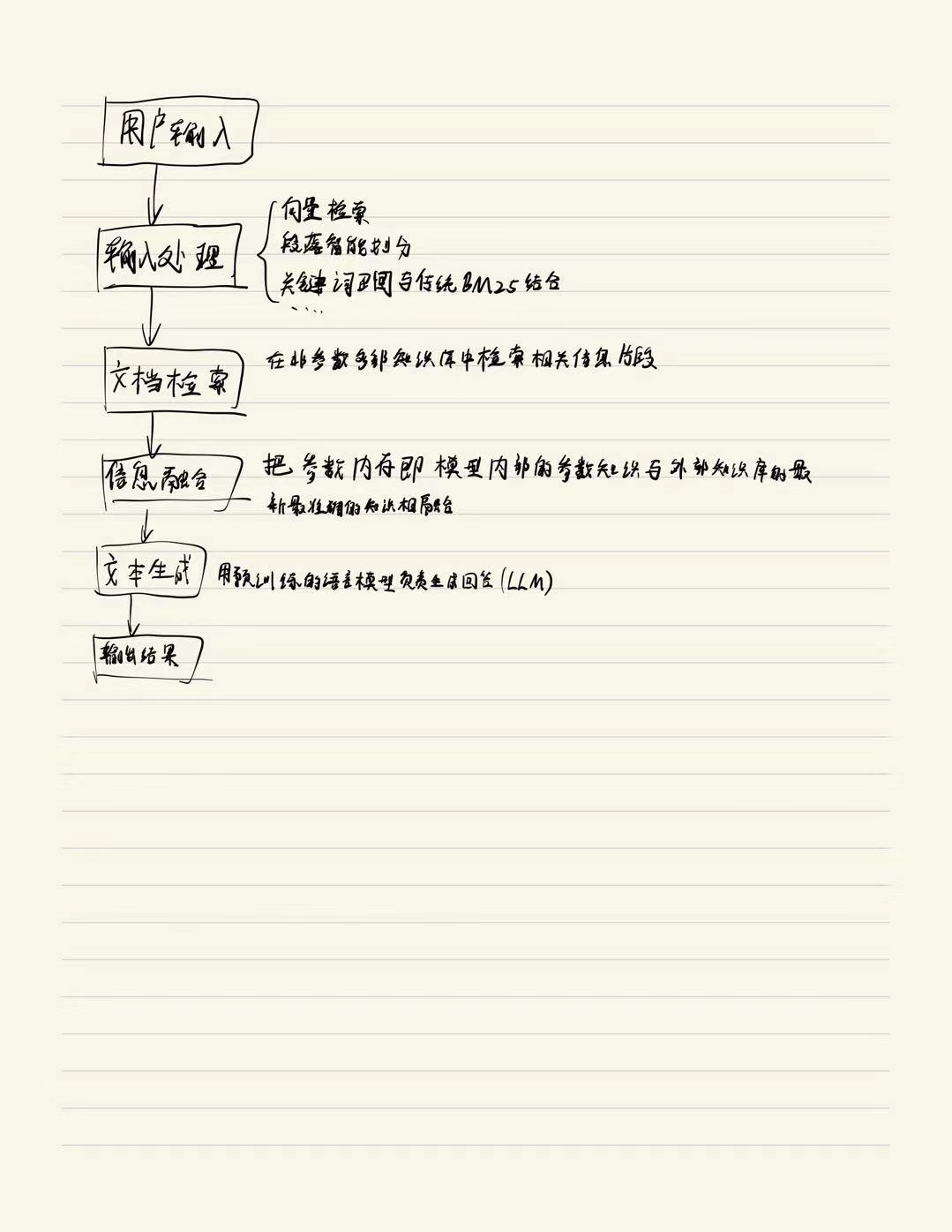
2. Retrieval Component: The RAG system retrieves the most relevant information fragments for user queries from a large-scale unstructured text database, such as the Wikipedia article collection or other types of document libraries. This step typically relies on an efficient retrieval mechanism that can quickly locate document paragraphs that may contain the desired information. The retrieval process may involve calculating similarity scores between the query and the document, and selecting the highest scoring documents as candidates.

3. Information fusion: Relevant document paragraphs retrieved will be integrated to form a comprehensive contextual environment. This context contains the latest and most accurate information obtained from external knowledge sources to supplement or update the knowledge stored in the internal parameters of the model.

4. Generation Component: Based on the context environment constructed above, pre trained language models (usually large language models based on Transformer architecture) are responsible for generating answers. Due to the support of specific background information, the generated answers are not only more accurate, but also better adapted to specific query needs.

5. Output result: Ultimately, the answer processed by the generating component will be returned to the user. This answer not only utilizes specific facts from external knowledge bases, but also leverages the advantages of language models in terms of fluency and coherence.

1. 自己在草稿纸上画一张RAG工作流程图



1. 现在的RAG已经用了什么技术，解决了什么场景的问题？

Main technical applications：

Vector retrieval: a method that uses semantic similarity calculation, such as converting text into vector representations through pre trained language models, and then efficiently retrieving based on these vectors.

Paragraph intelligent partitioning: Improved the slicing method of documents to ensure that each chunk is neither too large nor too small, in order to improve query accuracy and contextual integrity.

Dynamic Vector Generation: Adjust the vector generation strategy dynamically according to different QA scenarios, such as using the Instructor tool to optimize vector quality.

HyDE intermediate layer: first create virtual or hypothetical documents, and then conduct recalls to improve recall rates.

Combining keyword recall with traditional BM25: When vector recall is insufficient, traditional keyword matching methods are used to supplement information retrieval.

LLMLingua compression: deals with information overload caused by too many recalls.

LongLoRA Extended Window: allows a larger range of content to be input into large language models to solve the problem of long document summarization.

Scenario problems solved:

Scenarios with frequent updates of private data: Due to RAG's ability to quickly retrieve the latest data and integrate it into answers, it is highly suitable for enterprises or services that require real-time updates to their knowledge base.

Scenarios requiring citation of the original text: RAG can provide specific sources of literature, which is particularly important for fields such as academic research and legal consulting that require precise citation.

In the case of limited hardware resources: Compared to fine-tuning a model for each enterprise's private library, RAG only needs one fine-tuning to adapt to multiple needs, reducing costs and resource consumption.

Not good at answering small-scale descriptive questions: By improving content slicing and intelligent parsing techniques, RAG can now better handle such problems.

1. 现在的RAG面临什么挑战，在哪方面做的有欠缺？

Main challenges and existing problems:

Challenges in the retrieval process

**Accuracy and recall:** During the retrieval phase, there may be issues with low accuracy, where the retrieved information is not always highly relevant to the query; At the same time, there may also be issues with low recall rates, which means that some relevant document blocks may have been missed.

**Outdated data:** If the knowledge base is not updated in a timely manner, it may contain a large amount of outdated information, resulting in generated answers based on inaccurate data.

Challenges during the generation process

**Insufficient information:** When the retrieved information is insufficient to answer the user's question, the model may fabricate information or provide incomplete and incorrect answers.

**Unrelated answer:** Sometimes the model may generate answers that are unrelated to the user's query, usually because the model cannot correctly understand the relationship between the retrieved knowledge content and the user's query.

Challenges in the process of constructing Prompt

**Prompt construction:** How to construct a suitable Prompt guided model based on the retrieved relevant knowledge content is another challenge. Even with high-quality knowledge fragments, if the Prompt is not designed properly, it can still affect the quality of the final answer.

**Redundancy and duplication:** When multiple retrieved knowledge contents contain similar information, it is easy to encounter redundancy and duplication issues, which may lead to lengthy and unnecessary duplication of generated content.

Challenges in knowledge base maintenance

**Resource consumption:** The implementation of RAG systems typically requires high computational resources, including resources for supporting retrieval mechanisms and database maintenance.

**Real time requirements:** In order to ensure the accuracy of retrieval, the knowledge base needs to be regularly updated and cleaned, which poses a requirement for automated processes.

1. RAG技术演进历程，现在最新的技术？

### **Advanced RAG**

### **Naive RAG**

### **Modular RAG**

9.RAG在政策问答上的研究论文列举出来，已经实现的应用的GitHub链接

snapshot/README\_2024.md

papers/2024年04月/README.md

HyPA-RAG: A Hybrid Parameter Adaptive Retrieval-Augmented Generation System for AI Legal and Policy Applications <https://arxiv.org/abs/2409.09046>

Responsible Retrieval Augmented Generation for Climate Decision Making from Documents <https://arxiv.org/abs/2410.23902>