(1)

Main components:

- **Document Preprocessing:**
  - Split document into chunk, and pre-compute the TF-IDF vector or dense vector of the document or chunk. That is for query matching.
  - Extract key entity and its key information, like the city name and its temperature. And output them as structured data. For example,

    ```
    {
    'city name': 'HONG KONG',
    'temperature': '22',
    'date': xxx,
    }
    ```

- **Query matching:**
  - Match the relative chunk or document, and feed it into prompt as the context, so that we can answer the question based on context via llm.
  - Match the relative structured data and feed it into prompt
- **Prompt engineering:**
- **Answer:**
  - we can apply llm model with causal chain thinking and without causal chain to answer the question.

(2)

- **Question 1:**
  - **system prompt:**

    > You are a helpful assistant. Please answer the question based on the content within 2 brackets like <\context> <\context>

  - **user prompt:**

    > <\context> <\context>
    >
    > <\context> <\context>
    >
    > What is the temperature in HONG KONG, provide me a short answer with only several keywords.

- **Question 2:**
  - Must use causal thinking llm model, like `deepseek-r1` or distilled thinking models.
  - **system prompt** is same as above
  - **user prompt**:

    > <\context> <\context>
    >
    > <\context> <\context>

> What is the temperature variation in HONG KONG. Please answer after a CoT thinking, and provide the final answer within 2 brackets like <\box> <\box>

(3)

- **increase number of retrieval documents:** add more documents as the context.
- **Further inference:** Apply reasoning model like `deepseek-r1` to extract the information and inference.
- **Voting:** We ask the same question and get 10 answers, then we choose one answer with highest frequency as our final answer.
- **Validation:** After we get the answers from llm, we can input the answer with context to a validation llm models like `deepseek-r1`. We ask it `Is the answer right or not?` And we need to set a lower `temperature` and lower `top p` to get a reliable answer. If the result is `The answer is not right`, we will repeat the second step to regenerate another answer.