Search This step is responsible for retrieving relevant documents based on a given query. Queries are submitted using the respective API to retrieve relevant documents for web search engine retrieval. For local resources, the query component is responsible for formatting the query in the format required by different sparse or dense retrieval methods. Then, the query is submitted to the retrieval system, which returns a set of relevant documents along with their scores.

In both local and web-based scenarios, an optional reranker can be employed to refine the ranking of retrieved documents further. The reranker usually comprises a more complex and larger model that considers additional features of the documents and the given query. These additional features often include the semantic relationship between the query and the document content, document importance or popularity, and other custom measures specific to the information need at hand.

A.2 Generation Component

The evaluable output for the generation component is the *response* of LLMs and the *structured or formatted output* from the phrased response.

Prompting The generation process critically hinges on prompting, where a query, retrieval outcomes, and instructions converge into a single input for the language model. Research showcases various strategic prompting tactics such as the Chain of Thought (CoT) [60], Tree of Thought (ToT) [3], and Self-Note [31], each significantly shaping the model's output. These methods, especially the step-by-step approach, are pivotal in augmenting LLMs for intricate tasks.

Prompting innovations have introduced methods like Rephrase and Respond (RaR) [8], enhancing LLMs by refining queries within prompts for better comprehension and response. This technique has proven to boost performance across diverse tasks. The latest RAG benchmarks [61,62] in the specific domains start to evaluate the robustness of various prompting engineering skills, including CoT, RaR, etc.

Inference The final input string prepared in the prompting step is then passed on to the LLMs as input, which generates the output. The inference stage is where the LLM operates on the input derived from the retrieval and the prompting stages in the pipeline to generate the final output. This is usually the answer to the initial query and is used for downstream tasks.

Depending on the specifics of the task or expected output structure, a post-processing step may be implemented here to format the generated output suitably or extract specific information from the response. For example, the classification problems (multi-choice questions) or if the task requires the extraction of specific information from the generated text, this step could involve additional named entity recognition or parsing operations.