**Contextual Question vs Non Contextual Question for Document QnA**

***Contextual Questions***

Contextual questions are those that have semantic similarity with the data present in the uploaded documents. These questions are directly related to the content of the document and require information from it to generate accurate responses.

**Example:**

* If a document on oncology is uploaded, and the user asks:
  + "What are the common treatments for lung cancer?"
  + "Explain the stages of cancer mentioned in the document."
  + "What is the survival rate for breast cancer?"

Since these questions refer to the information present in the document, they are considered contextual.

**Implementation :**

* For contextual questions , all parameters required for querying Azure Cognitive Search, including the index name, are passed into a single GPT function.
* This GPT function is specifically designed to interact with Azure Cognitive Search, fetch relevant results, and generate the final response.
* The function ensures that the query maintains full context, leveraging all necessary parameters and conversation history to retrieve and return the most accurate and relevant information.
* Since the GPT function directly handles the search query, the response remains tightly coupled to the indexed document data, providing precise answers based on the stored content.

**Limitation** :

* **No Answer Generation Without Relevant Matches**
  + If the search does not return relevant documents, the function does not attempt to fabricate an answer.
  + It ensures that all responses are grounded in the indexed content, which can sometimes result in "No relevant information found" messages.
* **Limited to Predefined Parameters**
  + The search is strictly controlled by the predefined parameters, making it less adaptable to ambiguous or vague queries.
  + Users must phrase their questions within the expected structure to get meaningful results.

***Non-Contextual Questions***

Non-contextual questions do not have any semantic similarity with the uploaded document's data. These questions are more generic or request metadata about the document rather than its content.

**Example:**

* "Tell me something about the document."
* "Summarize the document."
* "List all the documents uploaded."

These questions are not derived from the content of the document but rather about the document itself or unrelated general knowledge.

**Implementation :**

* Before using the non-contextual approach, the system first attempts the contextual approach. If the output states that the response is not available, then the non-contextual approach is used as a fallback mechanism.
* In the non-contextual approach, the system first queries Azure Cognitive Search independently to fetch the top 15 most relevant chunks of data related to the question.
* After retrieving these top 15 chunks, a static prompt is added to structure the response generation process.
* This entire constructed prompt—containing the 15 chunks, static instructions, and conversation history—is then passed into the GPT function for processing.
* The GPT function generates the final response based on the provided context without directly interacting with Azure Cognitive Search during this stage.
* This approach provides flexibility and ensures that responses are still relevant even if the search query was broad or vague.

**Limitations :**

* **Improper Citation Generation:** The function does not generate citations correctly and may not always reference the exact document source.
* **Slight Accuracy Deviation:** While responses remain relevant, they might not be as precise as the contextual approach, leading to slight variations in accuracy.
* **Potential Hallucinations:** In cases where relevant document data is insufficient, the model might generate responses that contain fabricated or misleading information.

**ChatDSI Input Context Handling Information**

**1. General QnA(Non-Contextual Questions)**

For general question-and-answer interactions, we capture the entire history of the conversation until the token limit of **3000 tokens** is reached.

**2. Document QnA (Contextual Questions):**

For questions related to uploaded documents, the history capture mechanism is handled based on the type of query:

* **For PDF/PPT/DOCX/TXT:** We capture the last **two** questions from the history.
* **For** **CSV/Excel and Image-Based QnA :** For this approach history is not captured.