Chunk + document hybrid retriever

**Chunk-Based Retrieval:** This involves breaking down documents into smaller chunks and retrieving these chunks based on their relevance to the query.

**Document-Based Retrieval:** This involves retrieving entire documents that are relevant to the query.

By combining these approaches and using long-context embeddings, the system aims to improve the precision and relevance of retrieved information, especially for queries requiring detailed, context-rich answers..

**Components and Mechanism**

1. **Document Segmentation:**
   * **Chunking:** The documents are divided into smaller, manageable chunks. Each chunk can be a paragraph, a section, or a pre-defined length of text.
   * **Document Retention:** The original documents are retained as full entities for potential retrieval.
2. **Embedding Generation:**
   * The system uses advanced language models (e.g., Transformer-based models) to generate embeddings that capture the contextual meaning of both chunks and entire documents. These embeddings consider long-range dependencies and contextual relationships within the text, providing a deep understanding of the content.
3. **Indexing:**
   * **Chunk Indexing:** Each chunk is indexed separately with its corresponding embedding.
   * **Document Indexing:** Full documents are also indexed with their embeddings.
4. **Query Processing:**
   * **Query Embedding:** When a query is received, it is converted into an embedding using the same language model to capture its contextual meaning.
5. **Retrieval:**
   * **Chunk Retrieval:** The query embedding is used to search the chunk index to retrieve the most relevant chunks.
   * **Document Retrieval:** Simultaneously, the query embedding is used to search the document index to retrieve the most relevant documents.
6. **Hybrid Scoring and Ranking:**
   * **Relevance Scoring:** Retrieved chunks and documents are scored based on their relevance to the query embedding.
   * **Hybrid Ranking:** A combined ranking is generated, prioritizing the most relevant chunks and documents. This hybrid approach ensures that specific relevant information (from chunks) and broader context (from documents) are both considered.
7. **Answer Synthesis:**
   * **Chunk Synthesis:** Relevant chunks are synthesized to form coherent, contextually relevant responses.
   * **Document Context:** Retrieved documents provide additional context to ensure the synthesized answer is comprehensive and accurate.

**Example Workflow**

1. **Query:** "What are the benefits of using renewable energy sources?"
2. **Query Embedding:** The query is converted into a long-context embedding.
3. **Chunk Retrieval:** Relevant chunks from various documents discussing the benefits of renewable energy are retrieved.
4. **Document Retrieval:** Entire documents that provide in-depth discussions on renewable energy are also retrieved.
5. **Hybrid Ranking:** Chunks and documents are scored and ranked based on their relevance.
6. **Answer Synthesis:** The most relevant chunks are synthesized into a coherent answer, with additional context from the retrieved documents