Vector store-backed retriever

A vector store retriever is a retriever that uses a vector store to retrieve documents. It is a lightweight wrapper around the vector store class to make it conform to the retriever interface. It uses the search methods implemented by a vector store, like similarity search and MMR, to query the texts in the vector store.

Maximum Marginal Relevance (MMR) Retrieval:

- **Focuses on diversity:** MMR aims to retrieve documents that are not only similar to the query but also cover different aspects of the topic. This can be beneficial for tasks like summarizing a topic or providing different viewpoints.
- **Configurable parameters:** You might be able to control the balance between relevance and diversity by adjusting parameters like the weighting of similarity scores and novelty scores.
- **Computational cost:** MMR retrieval can be computationally more expensive than simple similarity search, especially for large datasets.

Similarity Score Threshold Retrieval:

- **Precision control:** This method allows you to control the precision of retrieved documents by setting a minimum similarity score threshold. This ensures only highly relevant documents are returned.
- **Potential for missed results:** Setting a high threshold might exclude relevant documents with slightly lower similarity scores.
- **Tuning the threshold:** Finding the optimal threshold can be crucial for achieving a good balance between precision and recall (finding all relevant documents).

Specifying Top K:

- **Control result size:** This parameter allows you to limit the number of retrieved documents (top K), making the retrieval process faster and reducing the amount of data to process further.
- **Relevance ranking:** The retrieved documents are still ranked by their similarity score within the top K results. This allows you to prioritize the most relevant documents.
- **Choosing K:** The optimal value for K depends on your specific use case. For tasks like question answering, a smaller K (top 1-3) might be sufficient, while for browsing or exploration, a larger K (top 10-20) could be useful.