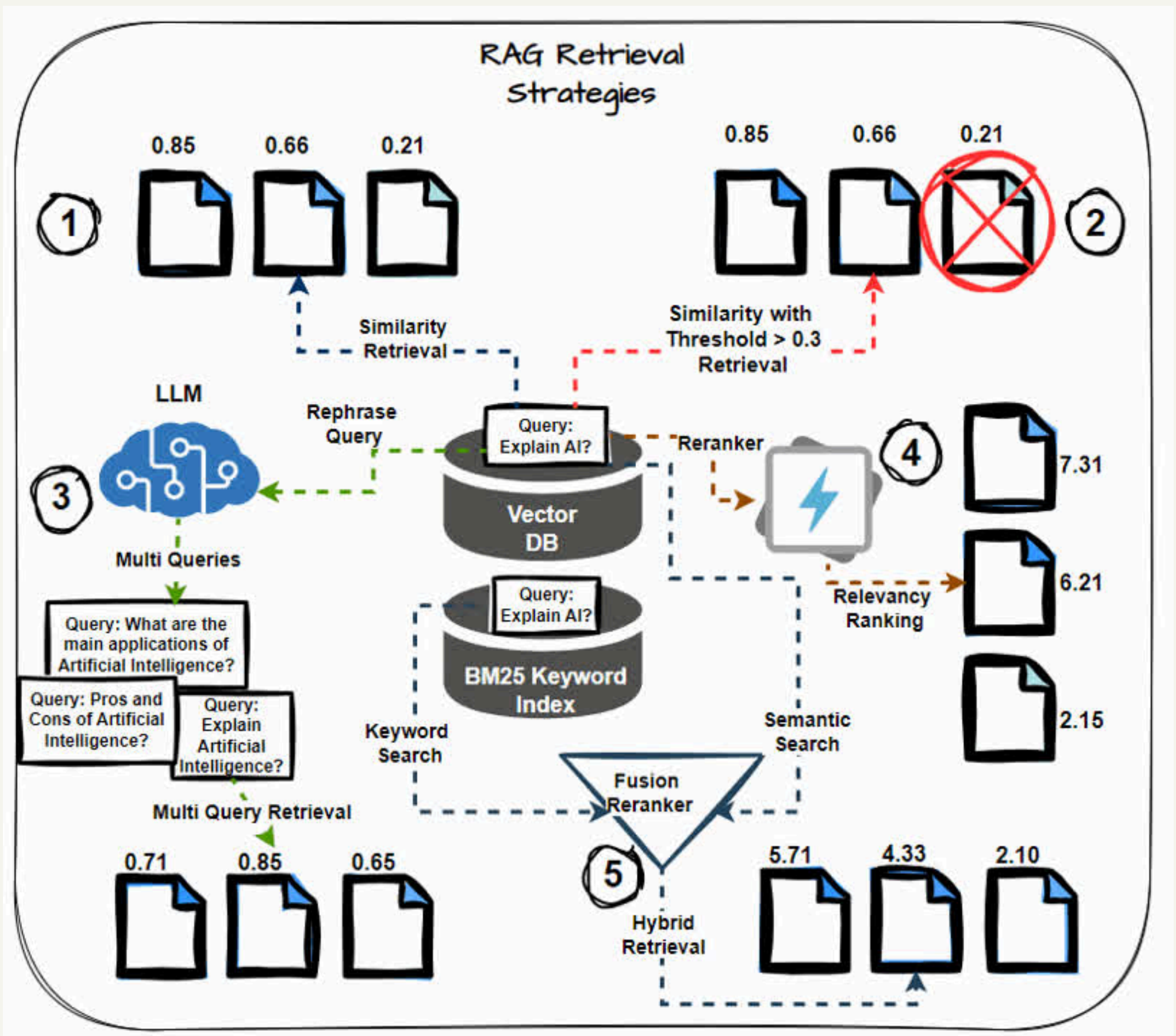
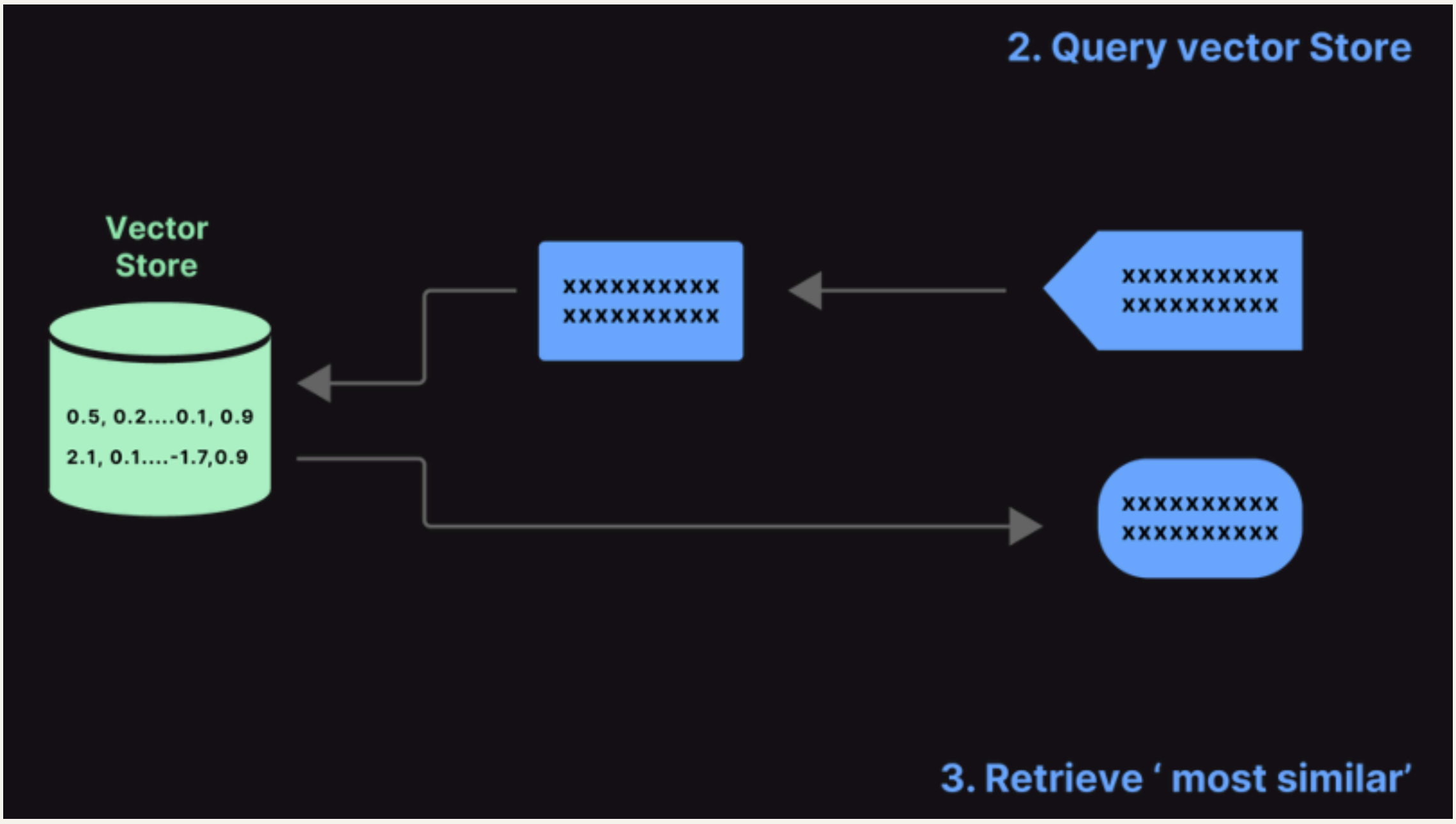
Popular RAG Retrieval Strategies



Dipanjan (DJ)

Purpose of Retrieval



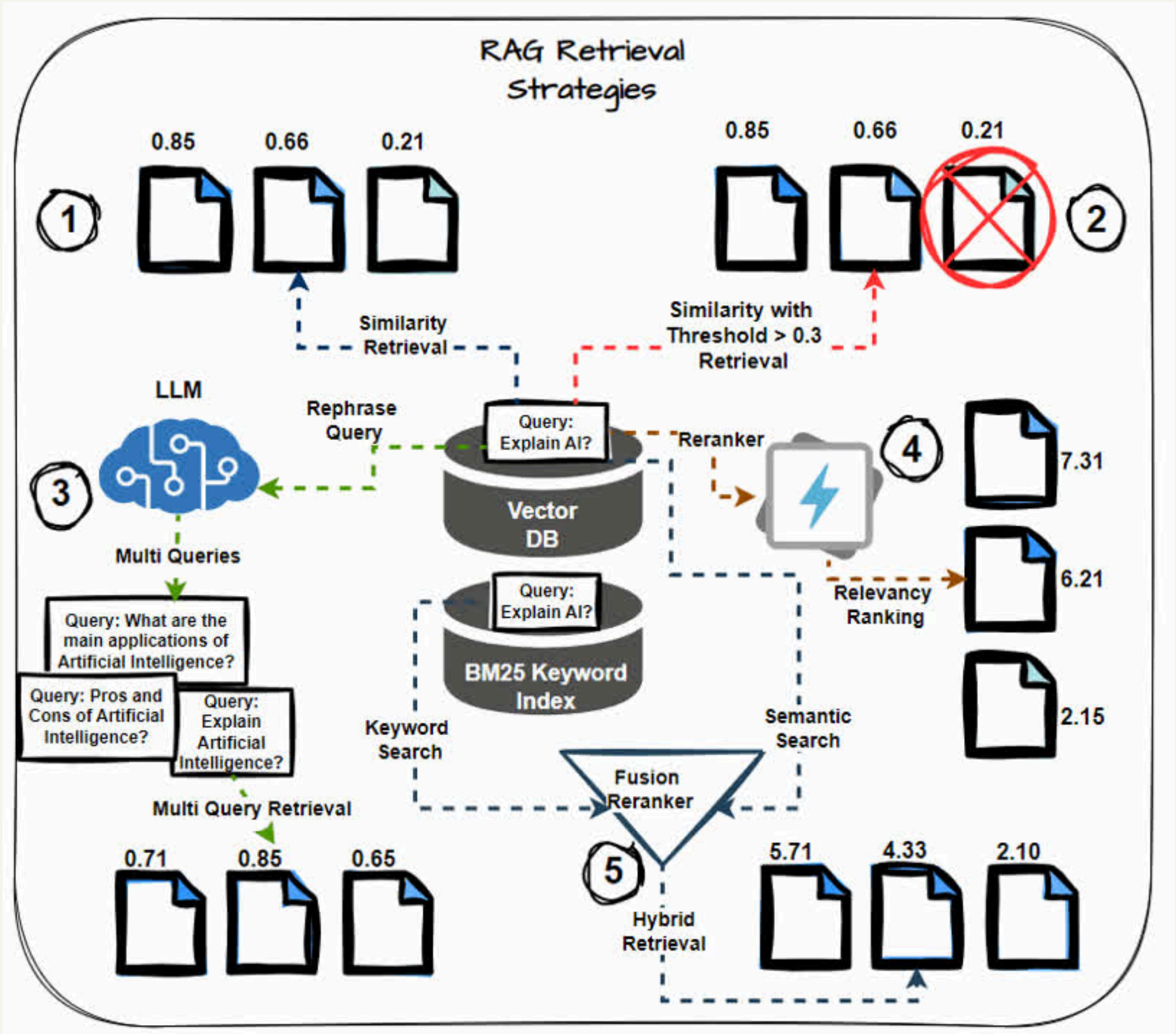
- **RAG connects an external knowledge base to augment the existing knowledge of a LLM**

- **RAG leverages a vector database to first retrieve relevant context for a query and makes the LLM use this context to answer queries**

- **Retrieval is the process of using a proper strategy to find out the most similar context documents to the given user query**

Source: Improving Real-World RAG Systems: Key Challenges & Practical Solutions

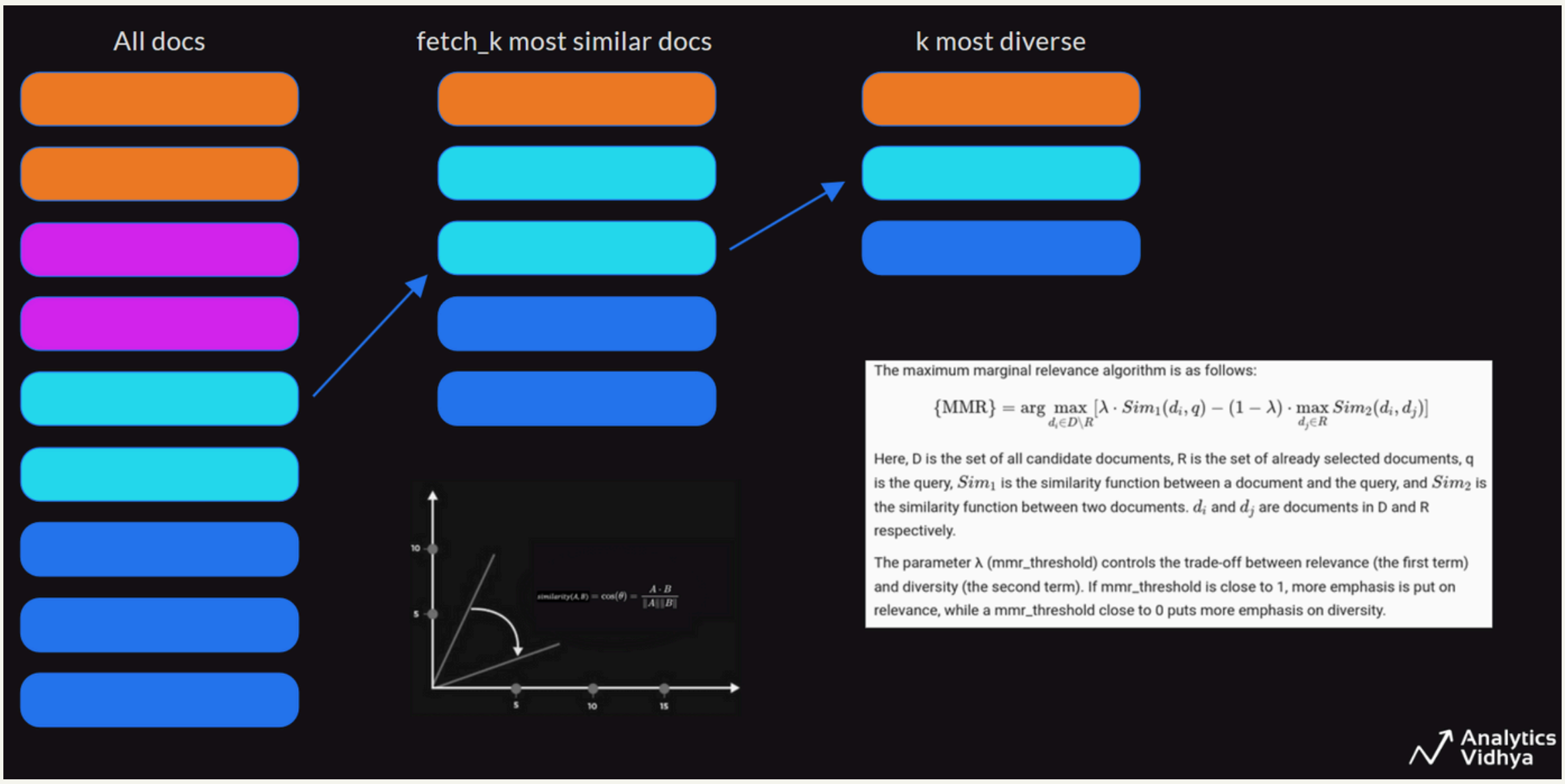
Popular Retrieval Strategies



- **Semantic Similarity**
- **Similarity with Threshold**
- **Multi-Query Retrieval**
- **Rerankers**
- **Hybrid Search**
- **Ensemble Retrieval**
- **Contextual Compression**
- **Self-Query and more . . .**

Source: Improving Real-World RAG Systems: Key Challenges & Practical Solutions

Semantic Similarity



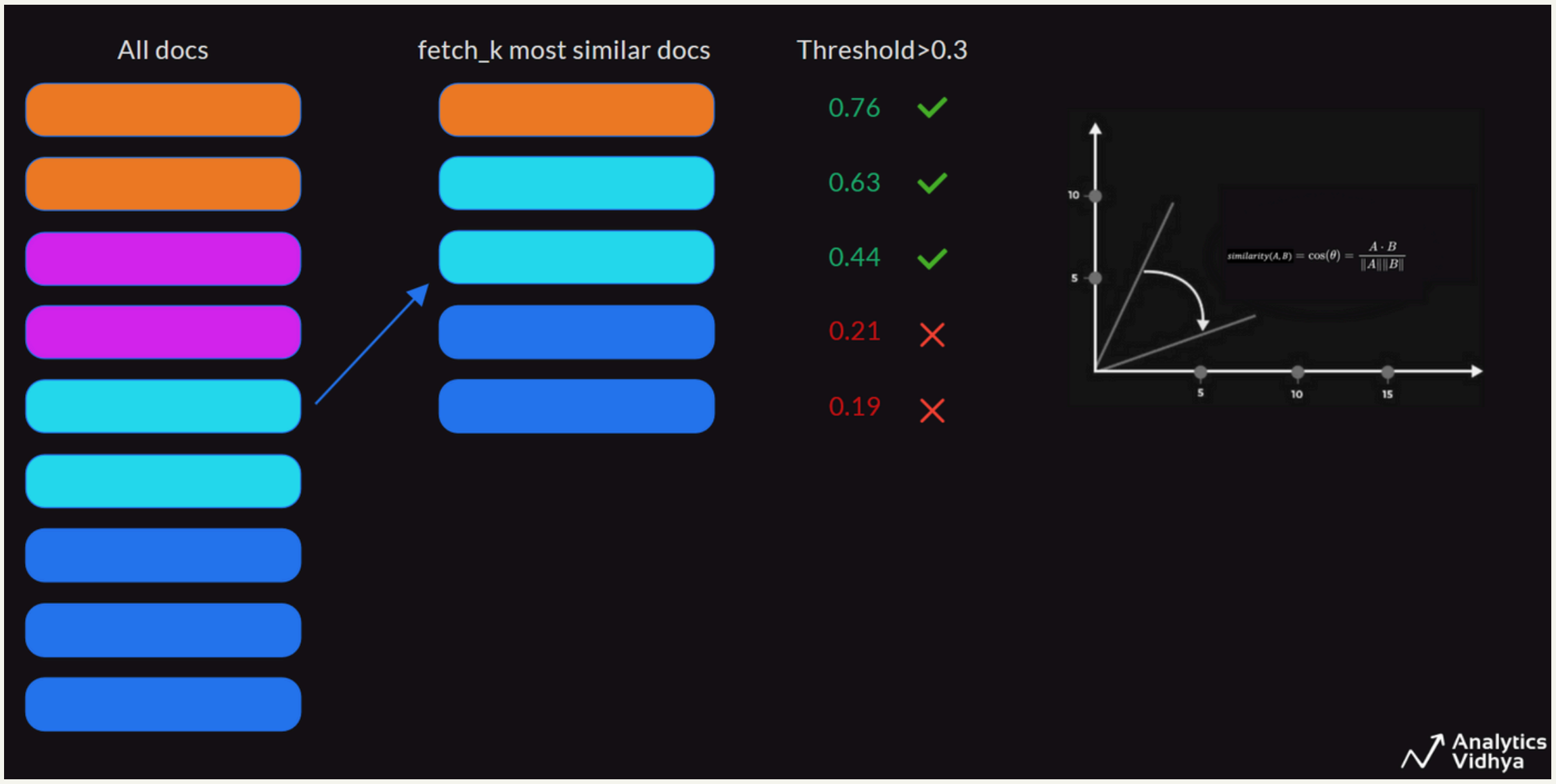
- **Semantic similarity usually measures the cosine similarity between query and context document embedding vectors**

- **Other similarity ranking methods can also be used like Maximum Marginal Relevance**

- **One of the most common methods for retrieving similar context documents**

Source: Improving Real-World RAG Systems: Key Challenges & Practical Solutions

Similarity with Threshold



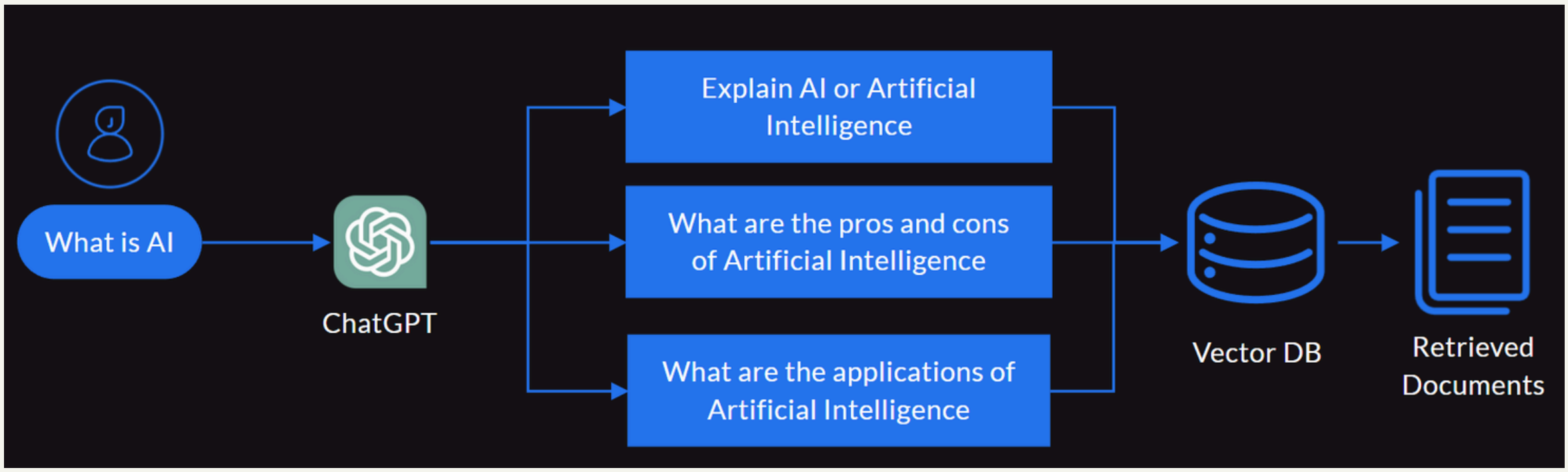
- **Very similar to semantic similarity, except it puts a cap on the minimum similarity which MUST exist between the query and context document embedding vectors**

- **All context documents returned are greater than the similarity threshold**

- **Very useful to prevent returning irrelvant context documents with low similarity to user queries**

Source: Improving Real-World RAG Systems: Key Challenges & Practical Solutions

Multi-Query Retrieval



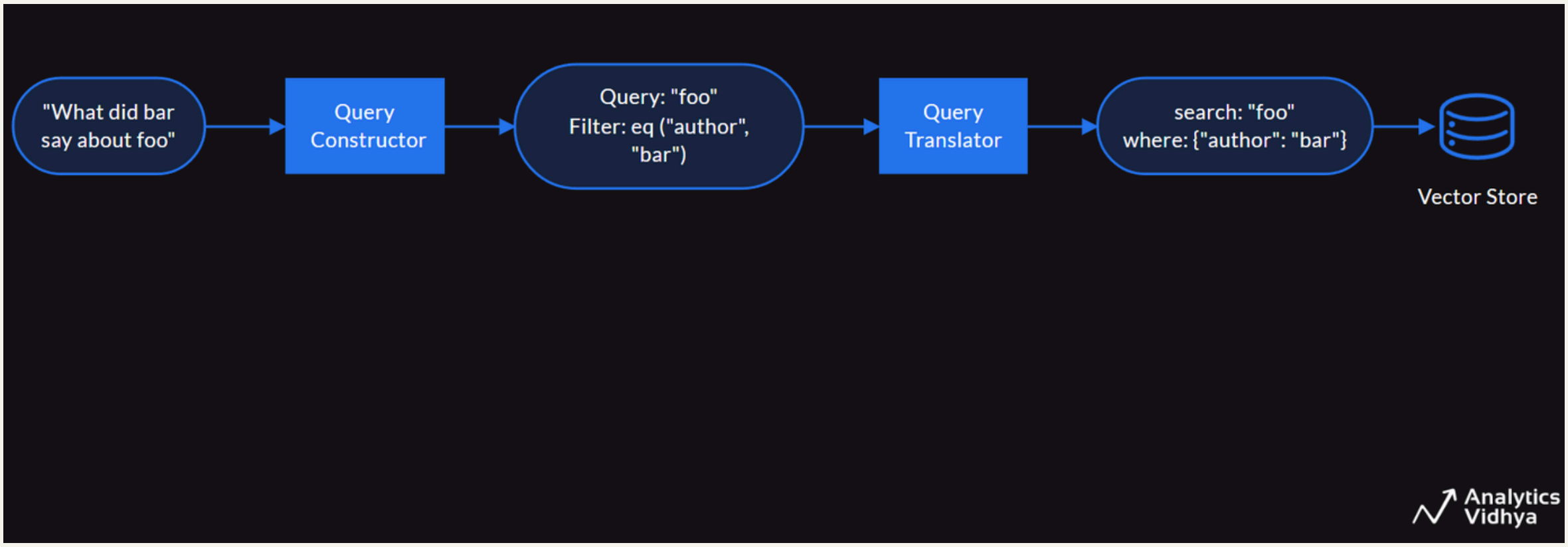
- **Multi-Query retrieval usually leverages an LLM to create multiple variants of the user query and then retrieves context documents similar to each of these user queries**

- **Finally it takes a union of the retrieved context documents and returns a unique list of context documents similar to all the variant user queries**

- **The idea is to create better versions of the user query so that we get more coverage in terms of context documents retrieved which can handle things like abbreviations, synonyms and more.**

Source: Improving Real-World RAG Systems: Key Challenges & Practical Solutions

Self-Query Retrieval



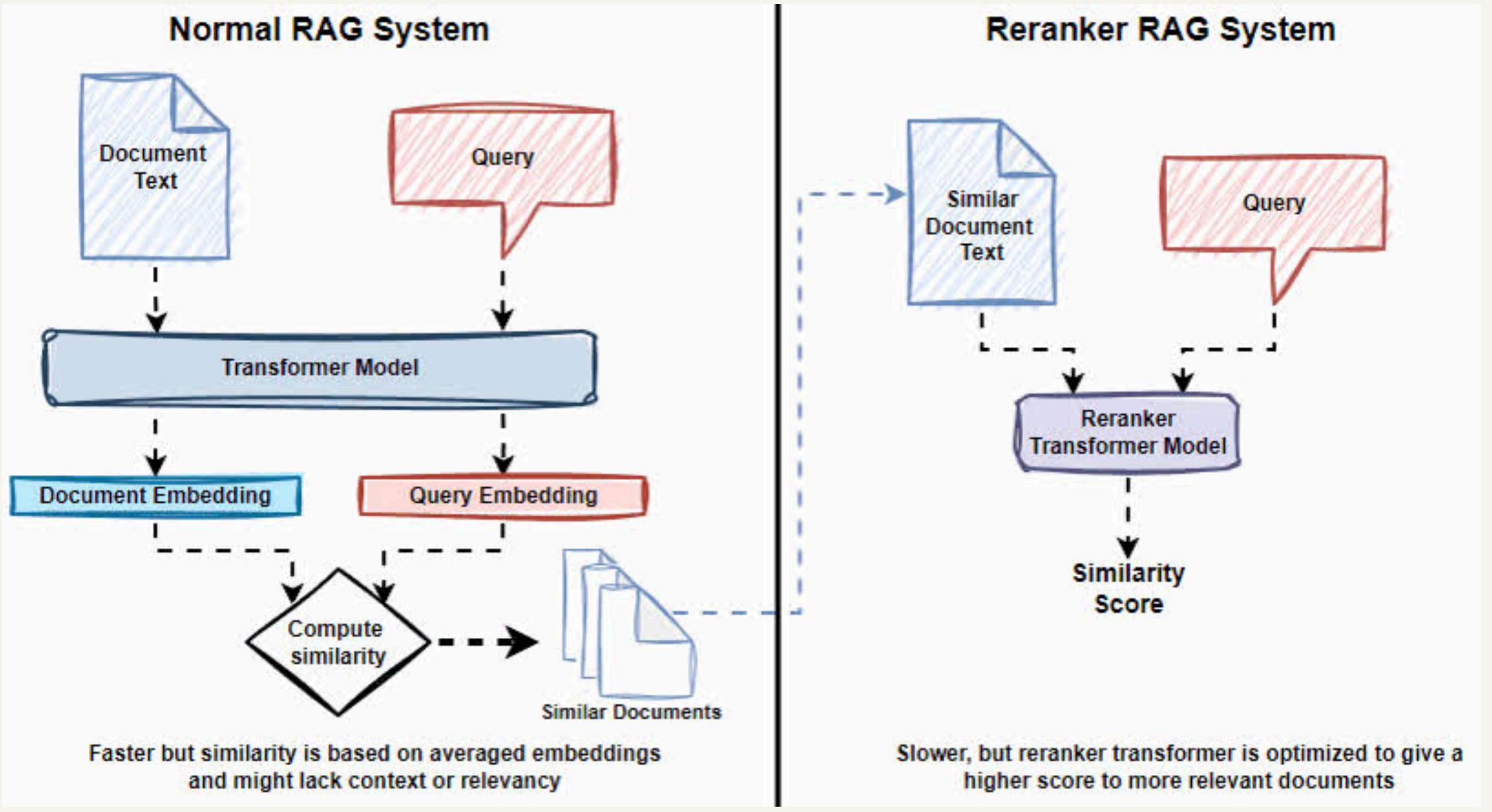
- **Self-Query Retrieval tries to first extract specific elements like structured field elements and unstructured elements from a natural language text query**

- **The retriever uses a query-constructing LLM chain to extract structured fields (like author, rating, product category etc.) which acts as filters**

- **Allows us to use standard semantic search using embeddings on natural language text and also filter retrieved documents further based on aspects mentioned around structured fields which are usually stored in the metadata**

Source: Improving Real-World RAG Systems: Key Challenges & Practical Solutions

Reranker Retrieval



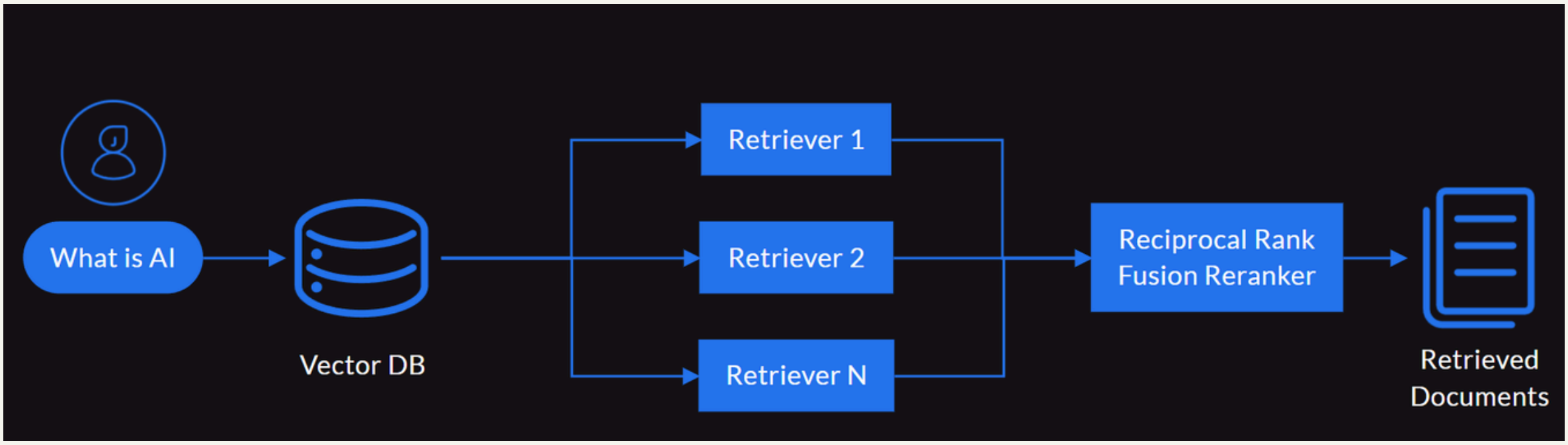
- **Rerankers are trained and fine-tuned cross-encoder transformers which have been trained on labeled data**

- **These models have been trained on (Query, Context) pairs and have learnt to predict relevancy scores in terms of how relevant is the Context to the given Query**

- **Very useful in reranking already retrieved context documents based on the user query by focusing more on relevancy rather than just embedding similarity**

Source: Improving Real-World RAG Systems: Key Challenges & Practical Solutions

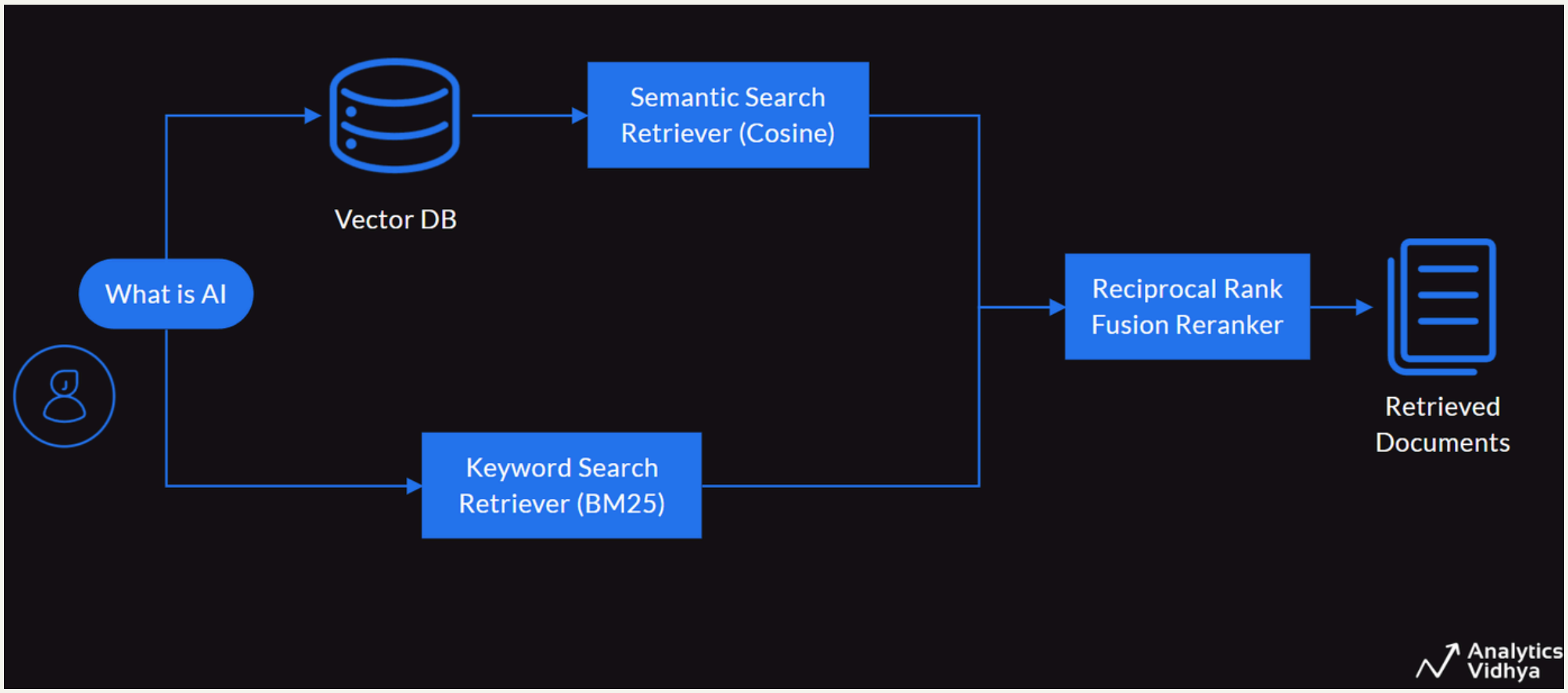
Ensemble Retrieval



- **Ensemble Retrieval is all about combining different retrieval strategies together**
- **Each retrieval strategy retrieves context documents based on user query independently**
- **Finally uses reciprocal rank fusion to combine and rerank the context documents**

Source: Improving Real-World RAG Systems: Key Challenges & Practical Solutions

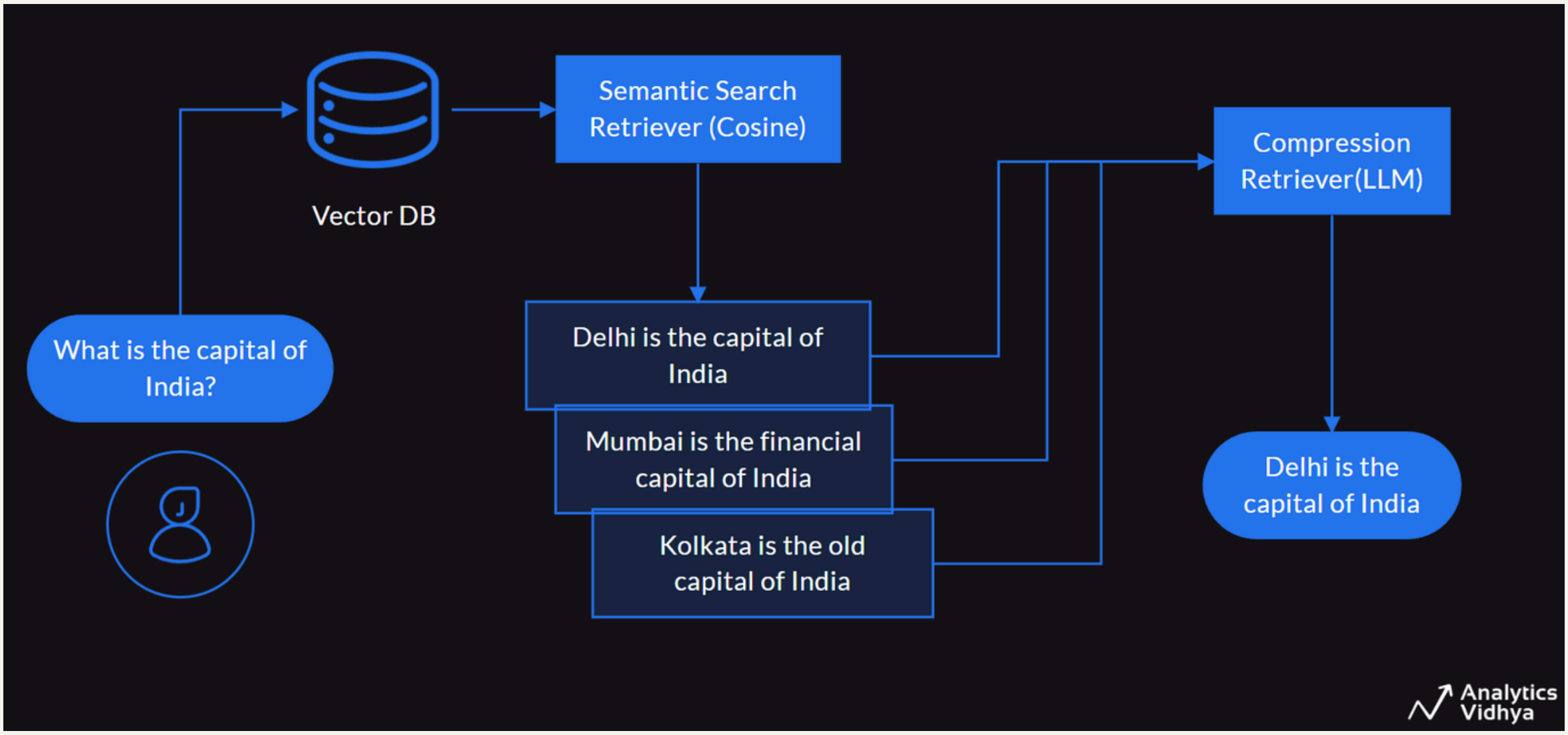
Hybrid Search



- **Hybrid Search combines semantic similarity (like cosine similarity) along with keyword search**
- **Keyword search can use any common keyword similarity techniques like bag of words, TF-IDF or BM25**
- **Very useful especially in cases when syntax and semantics both matter**

Source: Improving Real-World RAG Systems: Key Challenges & Practical Solutions

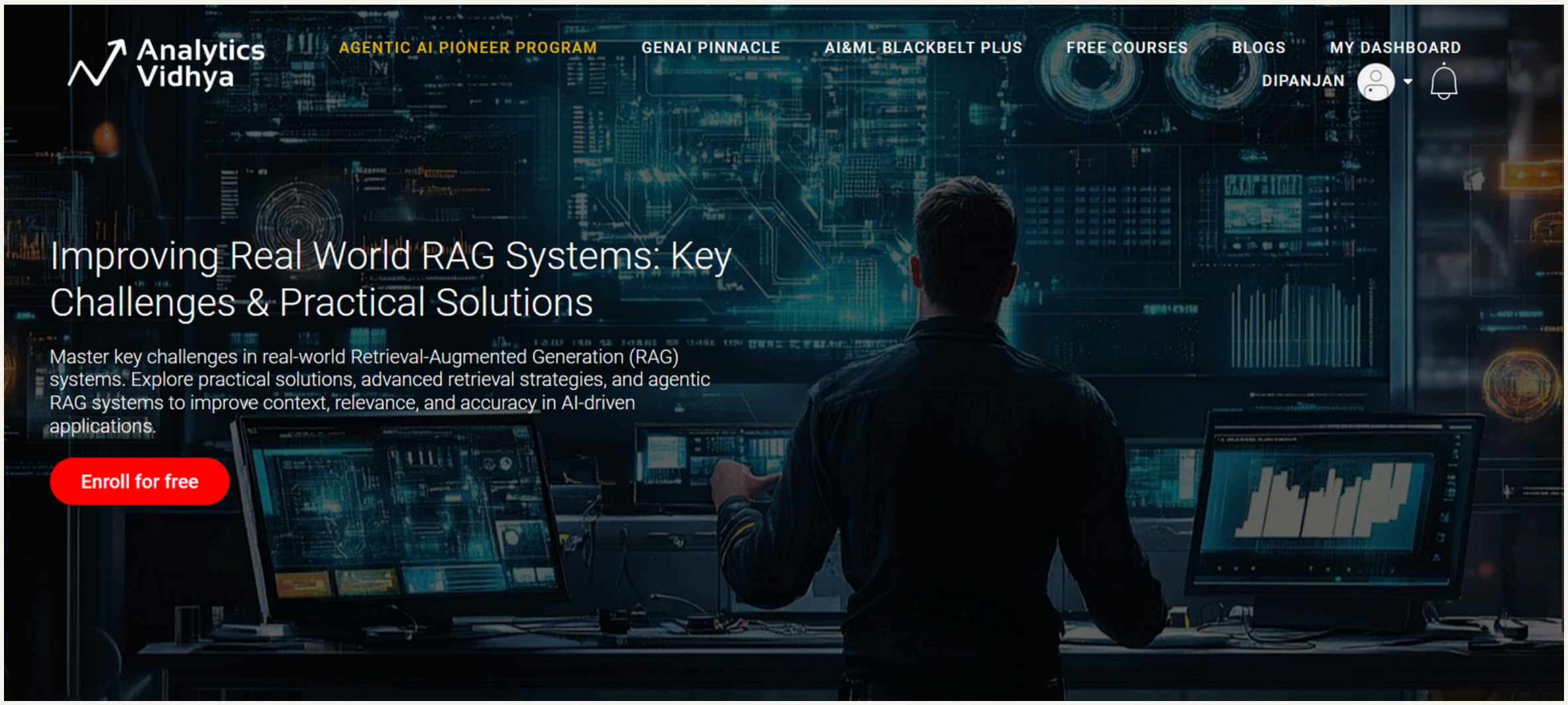
Contextual Compression Retrieval



- **Uses standard semantic (or any other strategy) retrieval as the first step**
- **Compresses retrieved context documents by removing irrelevant content from the documents**
- **One method is context extraction which just extracts content from the context documents relevant to user query**
- **Another method is context filtering which filters out complete context documents which are not relevant to the user query**

Source: Improving Real-World RAG Systems: Key Challenges & Practical Solutions

Free Course on Improving RAG Systems



CHECK OUT THE FREE COURSE HERE