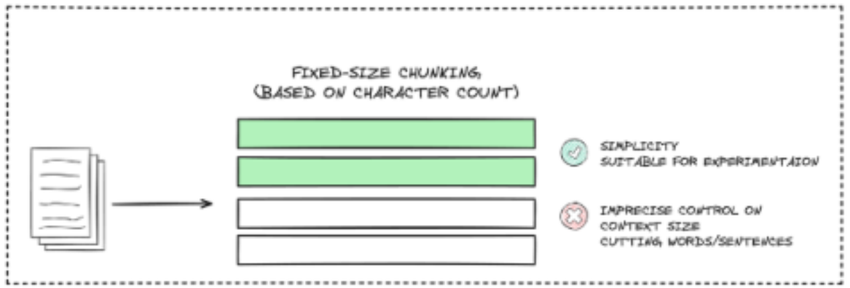
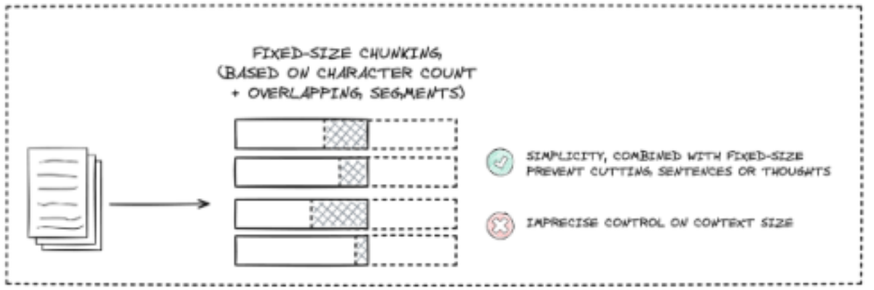
ADVANCED RAG

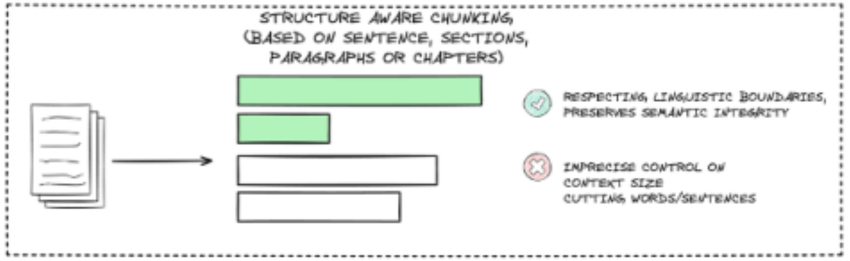
Eduardo Ordax
Go to Market Lead
Generative AI
AWS

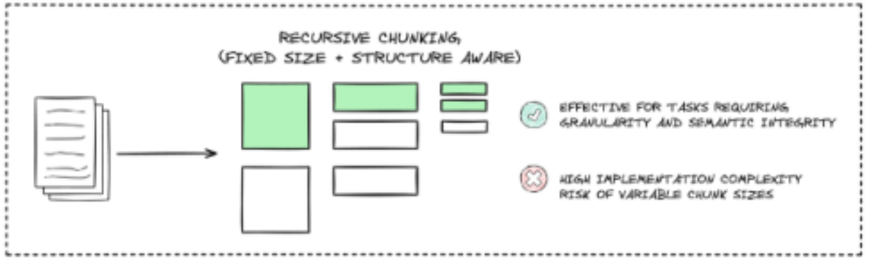
Wednesday, July 17, 2024
Madrid, ES

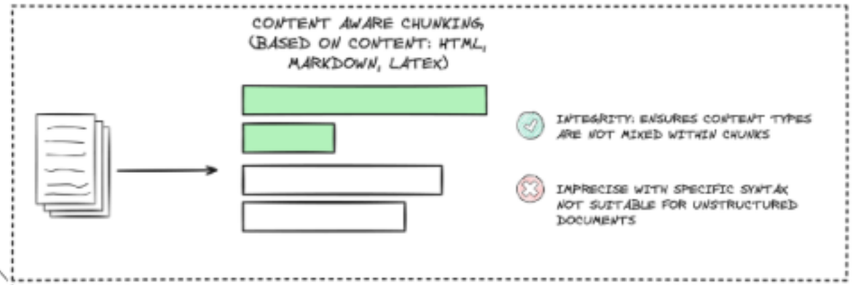
CHUNKING STRATEGIES

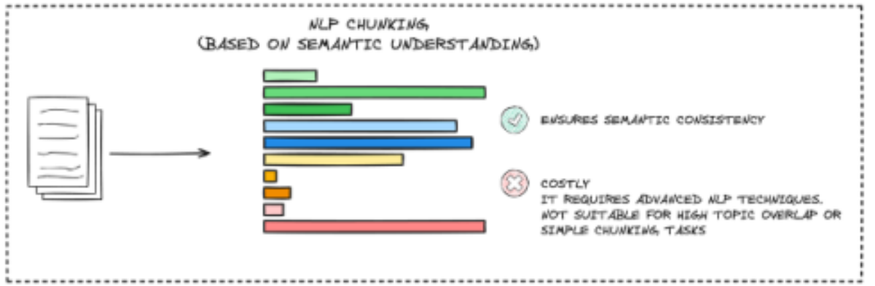






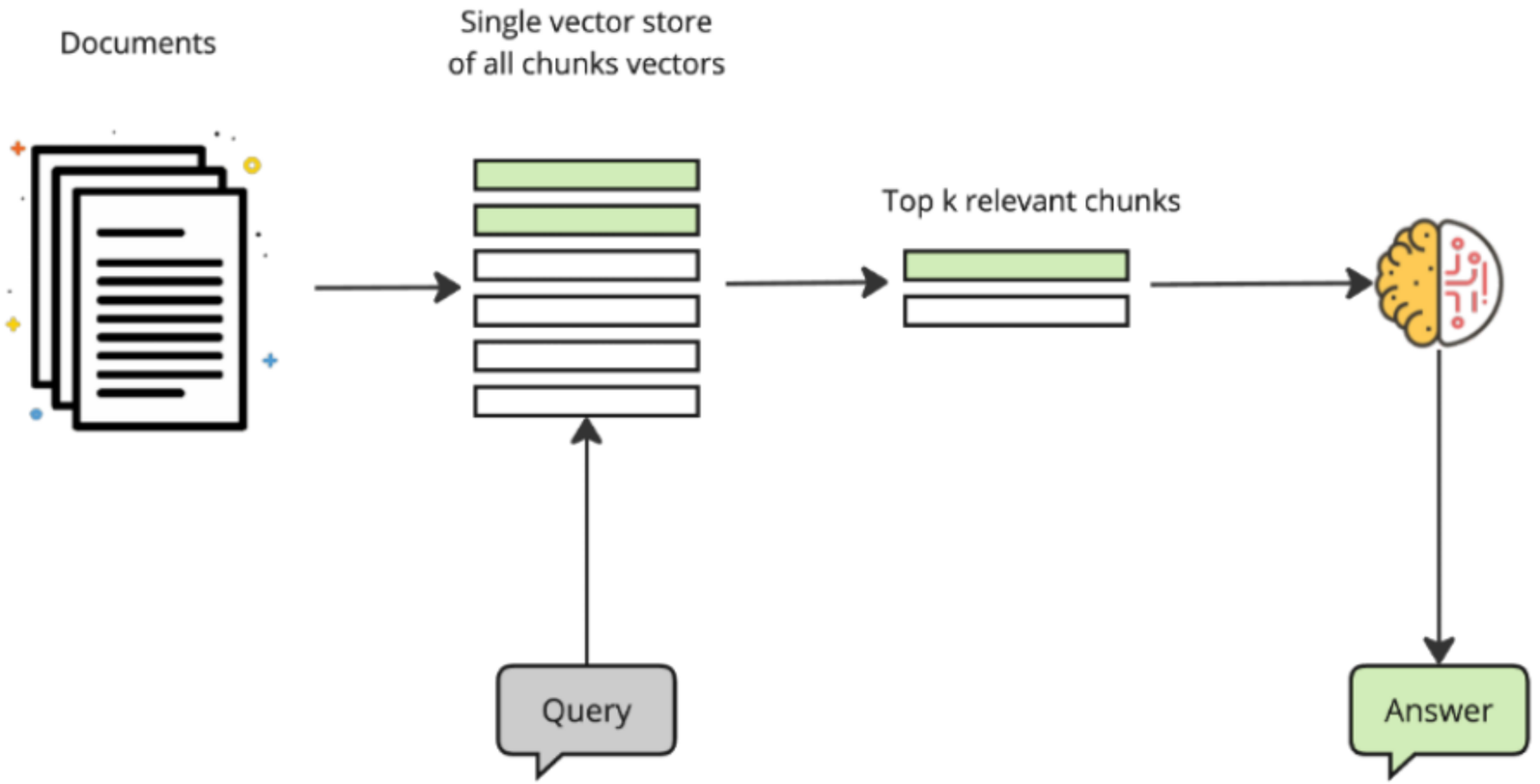






EDUARDO ORDAX
WWW.LINKEDIN.COM/IN/EORDAX

BASIC INDEX RETRIEVAL

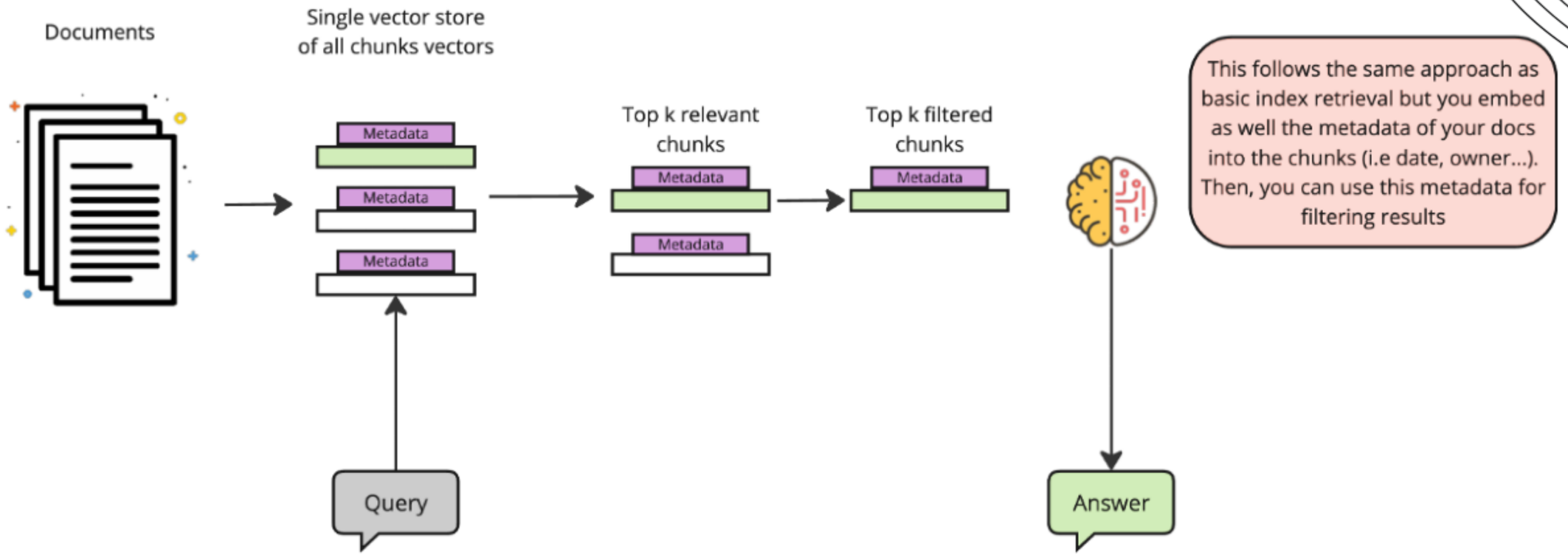


The most naive implementation uses a flat index — a brute force distance calculation between the query vector and all the chunks' vectors.

A proper search index, optimised for efficient retrieval on 10000+ elements scales is a vector index using some Approximate Nearest Neighbours implementation.

EDUARDO ORDAX
WWW.LINKEDIN.COM/IN/EORDAX

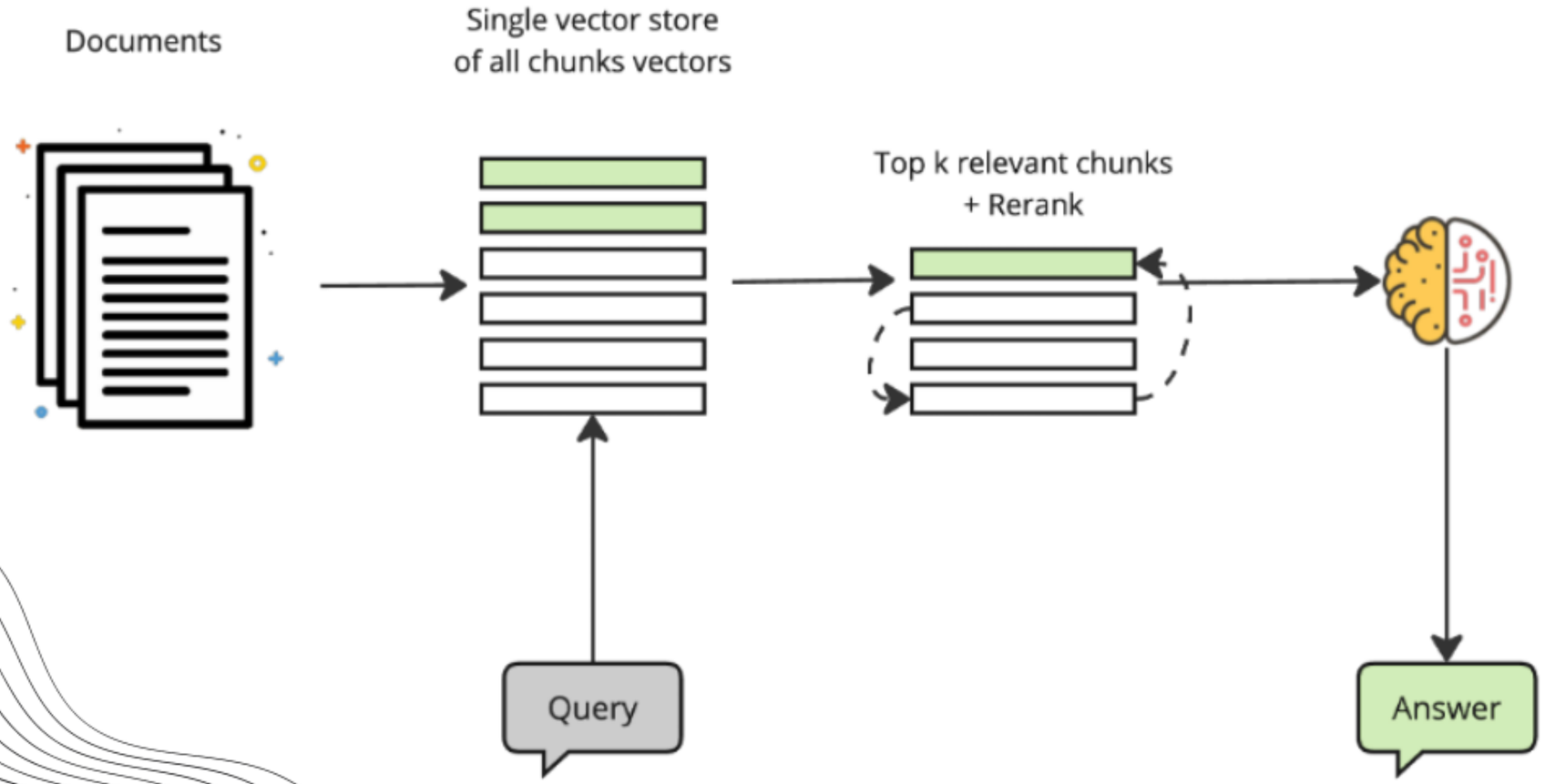
BASIC INDEX RETRIEVAL (+METADATA)



ABNASIA.ORG

EDUARDO ORDAX
WWW.LINKEDIN.COM/IN/EORDAX

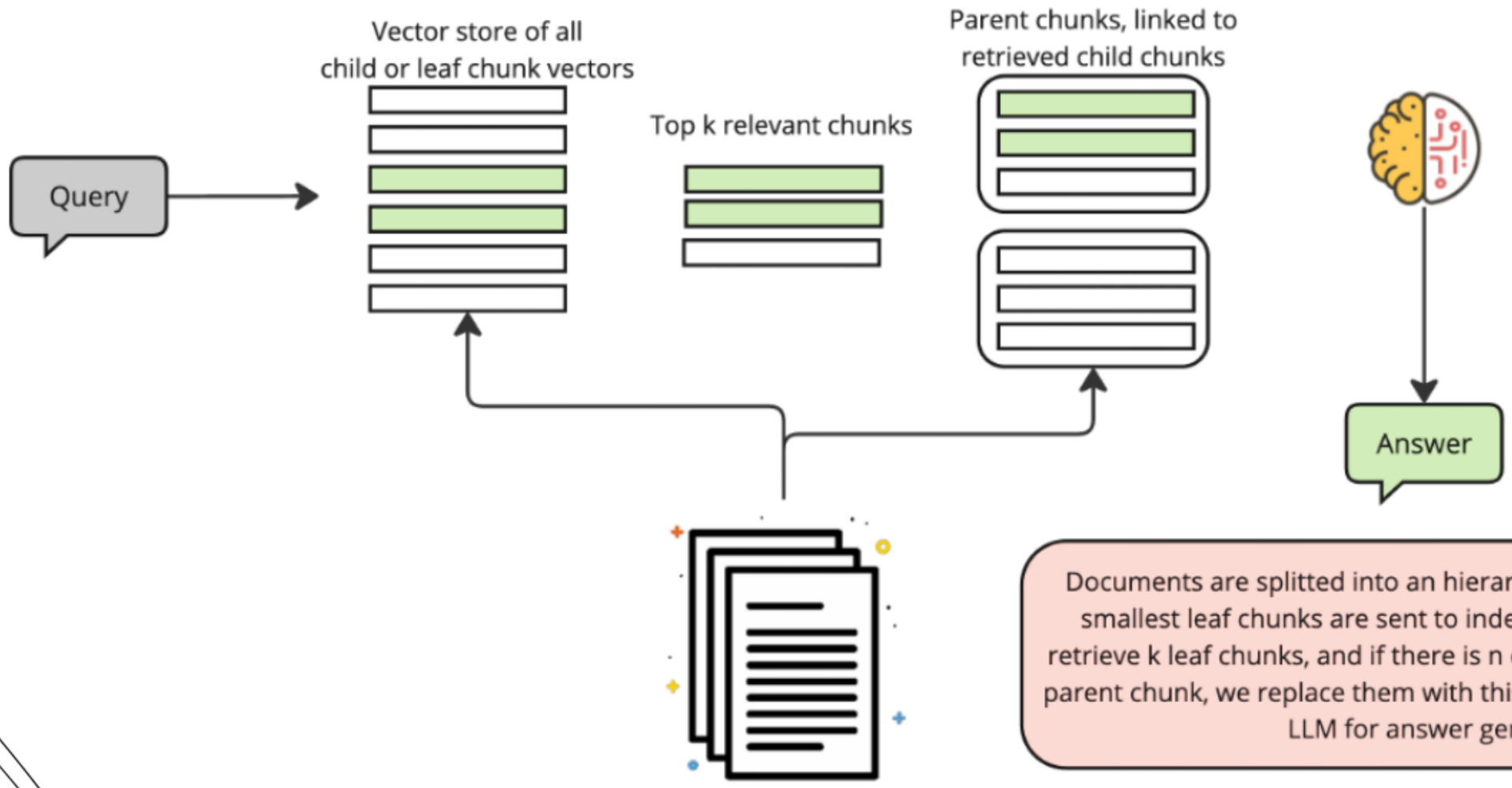
RERANKING



The most naive implementation uses a flat index — a brute force distance calculation between the query vector and all the chunks' vectors. After this, we use a reranker to rank the most relevant chunks for answer generation

EDUARDO ORDAX
WWW.LINKEDIN.COM/IN/EORDAX

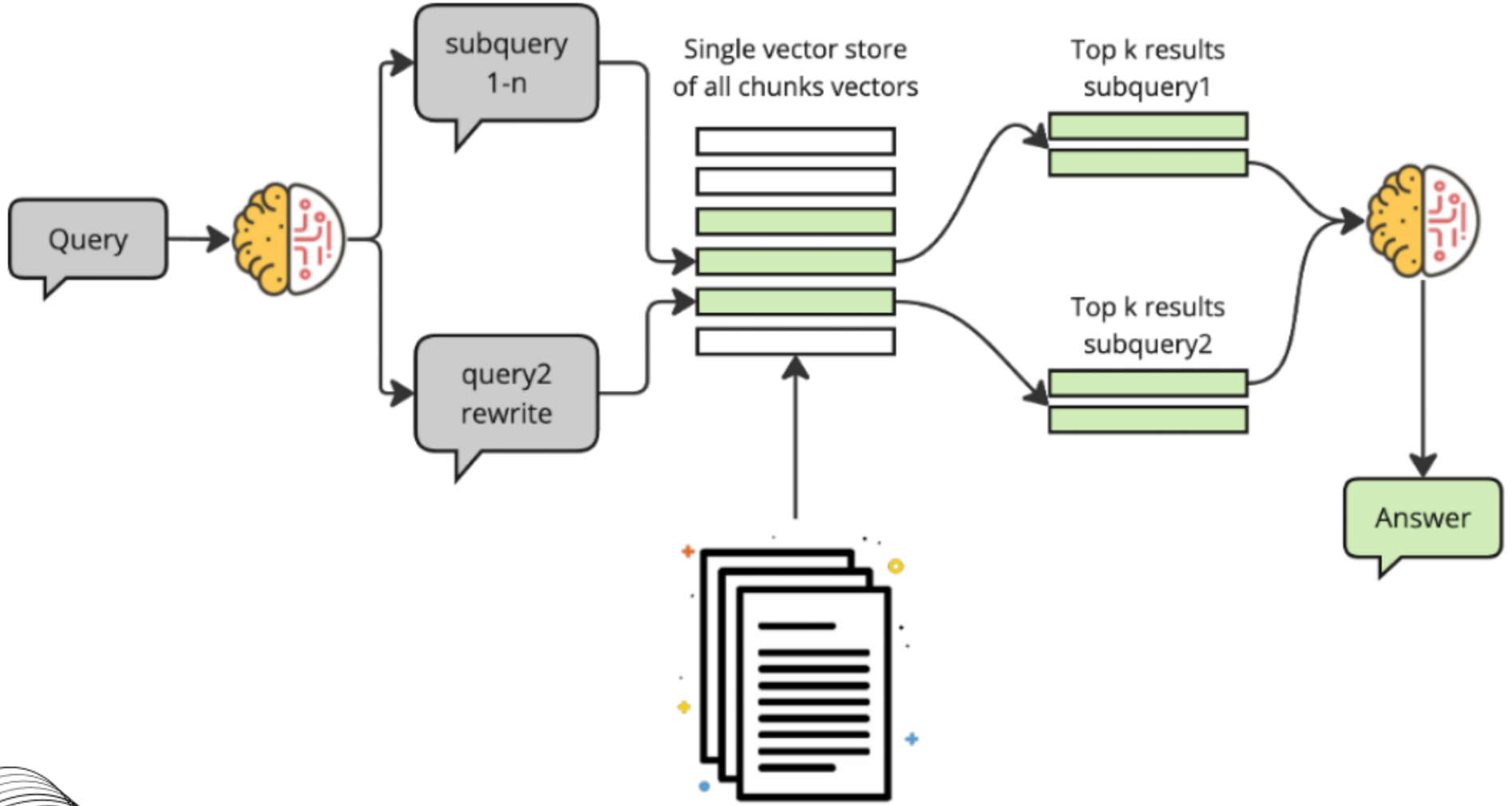
PARENT-CHILD CHUNKS RETRIEVAL



Documents are splitted into an hierarchy of chunks and then the smallest leaf chunks are sent to index. At the retrieval time we retrieve k leaf chunks, and if there is n chunks referring to the same parent chunk, we replace them with this parent chunk and send it to LLM for answer generation.

EDUARDO ORDAX
WWW.LINKEDIN.COM/IN/EORDAX

QUERY TRANSFORMATION



Query transformations are a family of techniques using an LLM as a reasoning engine to modify user input in order to improve retrieval quality. If the query is complex, LLM can decompose it into several sub queries.

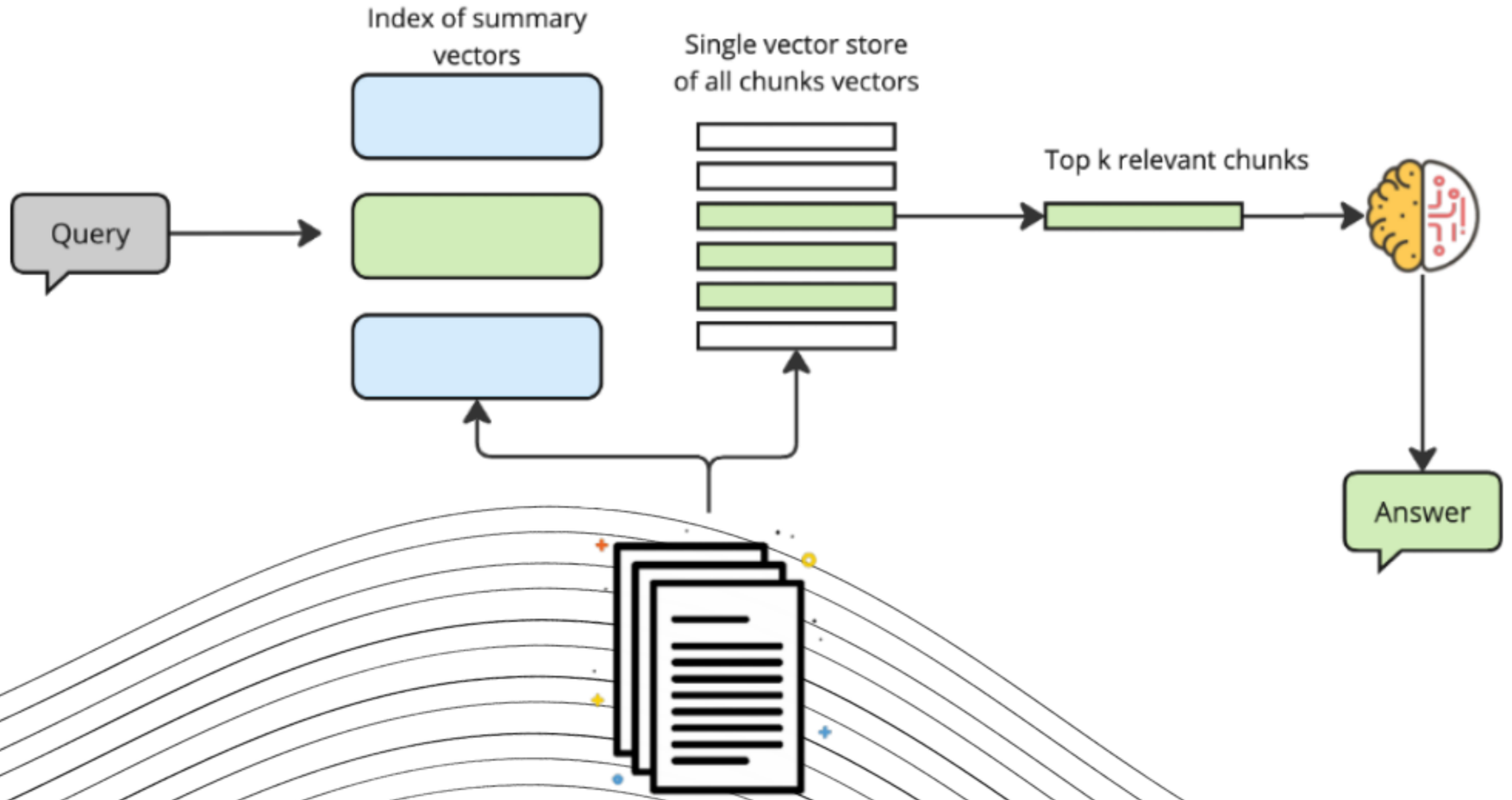
1/ Step-back prompting generates a more general query

2/ Query re-writing uses LLM to reformulate initial query in order to improve retrieval

EDUARDO ORDAX

WWW.LINKEDIN.COM/IN/EORDAX

HIERARCHICAL INDEX RETRIEVAL

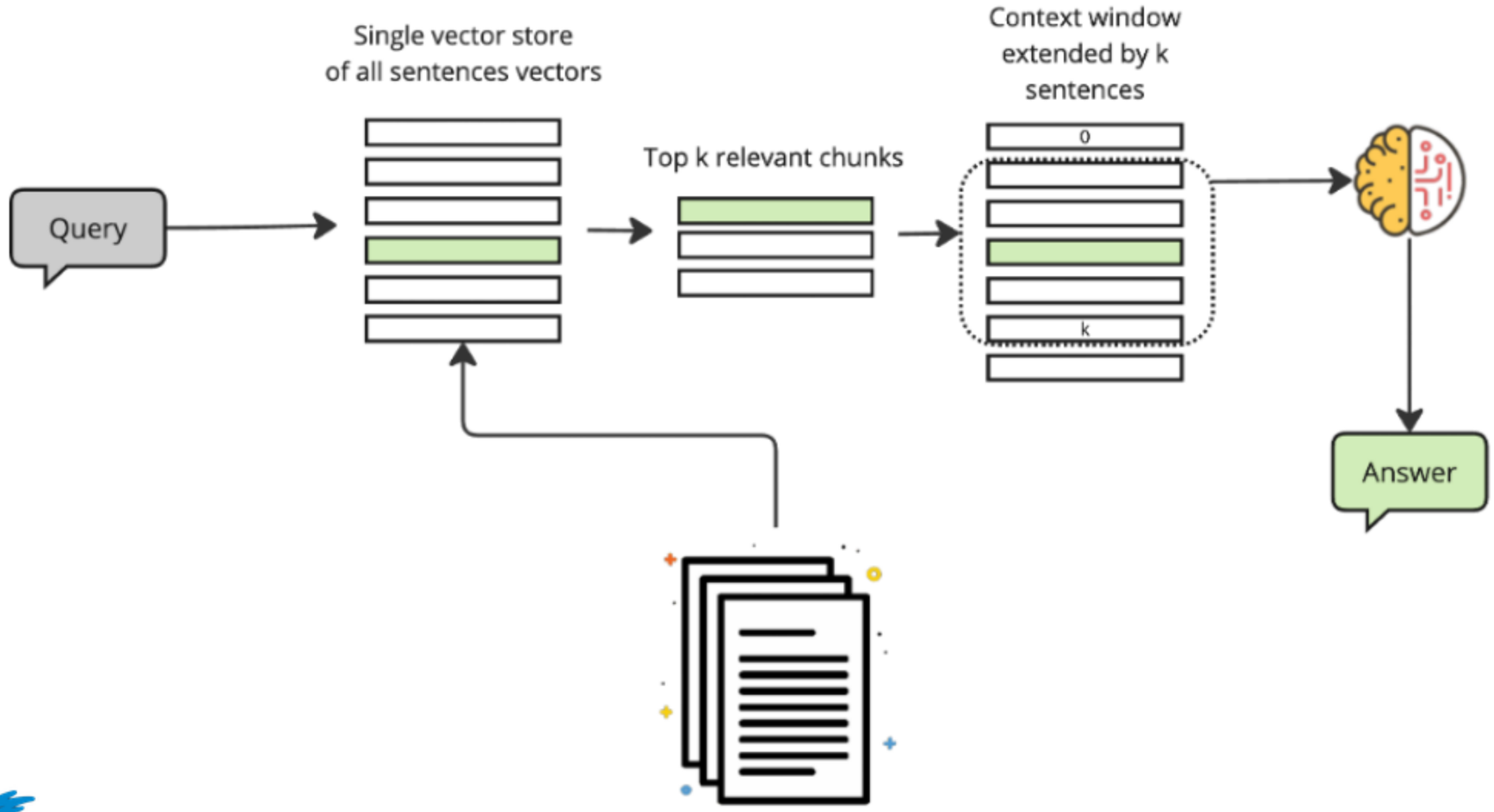


In case you have many documents to retrieve from, you need to be able to efficiently search inside them, find relevant information and synthesise it in a single answer with references to the sources. An efficient way to do that is to create two indices — one composed of summaries and the other one composed of document chunks, and to search in two steps

ABNASIA.ORG

EDUARDO ORDAX
WWW.LINKEDIN.COM/IN/EORDAX

SENTENCE WINDOW RETRIEVAL



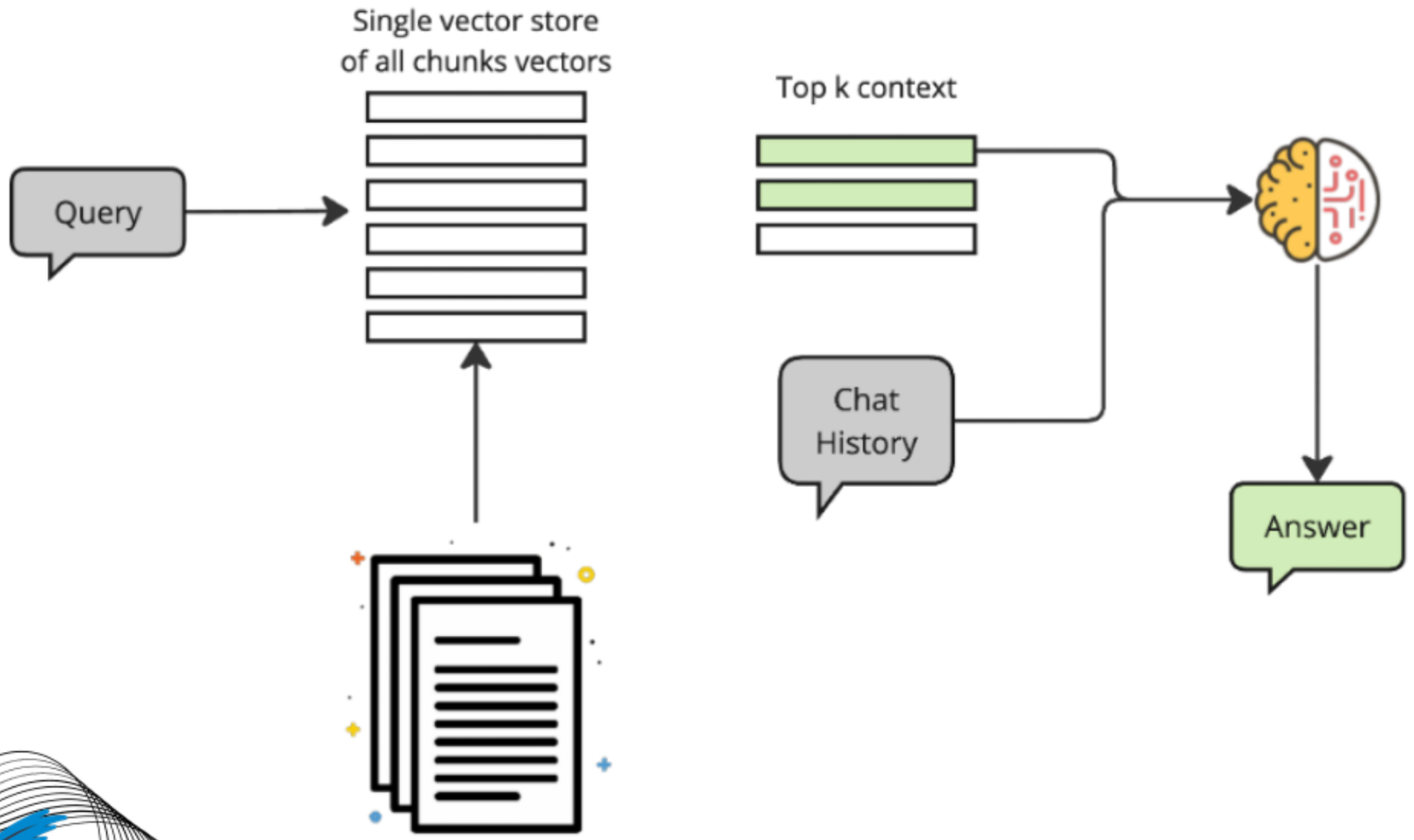
Each sentence in a document is embedded separately which provides great accuracy of the query to context cosine distance search.
In order to better reason upon the found context after fetching the most relevant single sentence we extend the context window by *k* sentences before and after the retrieved sentence and then send this extended context to LLM

ABNASIA.ORG

EDUARDO ORDAX

WWW.LINKEDIN.COM/IN/EORDAX

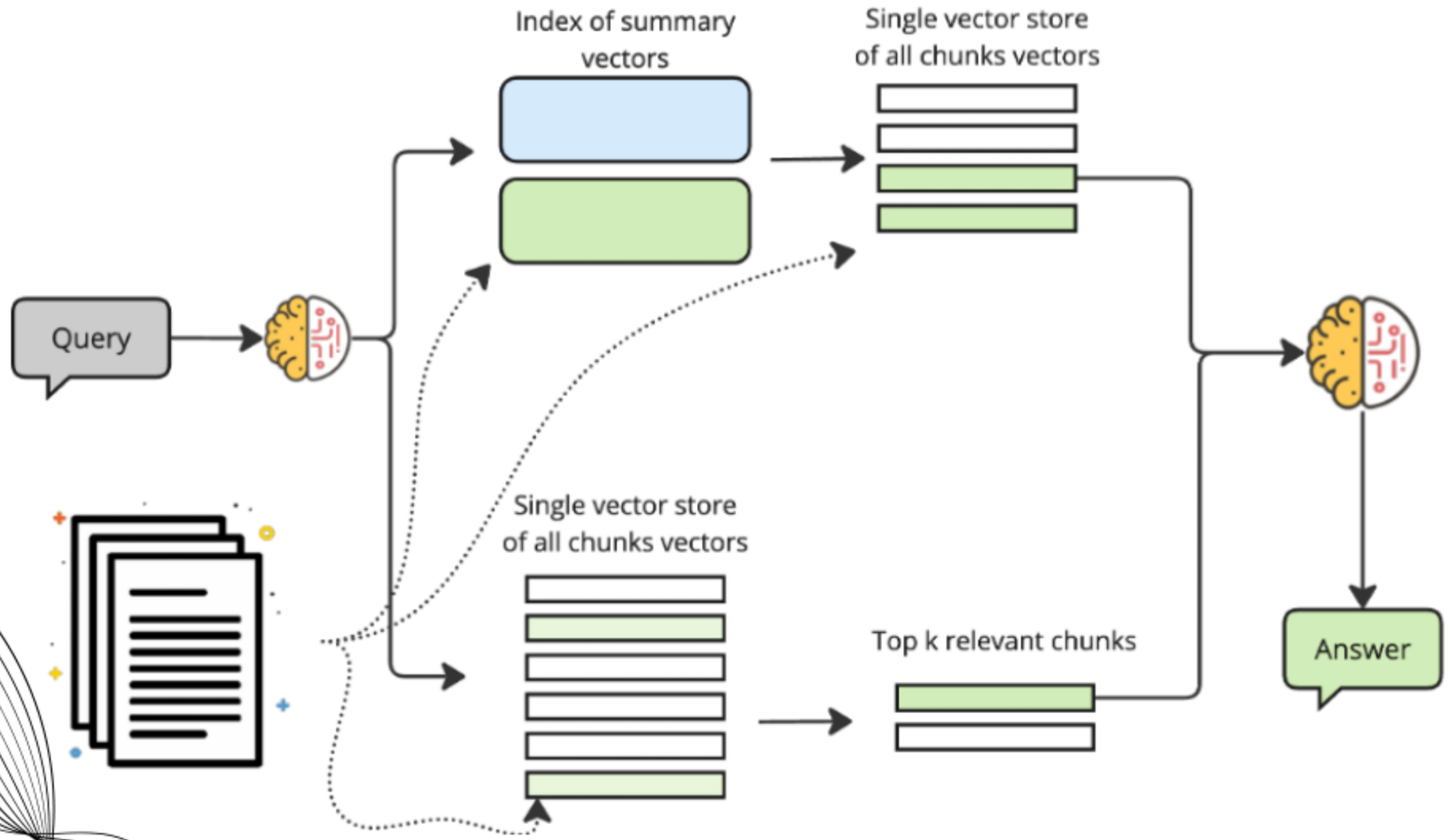
CHAT ENGINE CONTEXT



This technique takes into account the dialogue context.
It retrieves top k context relevant to user's query and then sending it to LLM along with chat history from the *memory* buffer for LLM to be aware of the previous context while generating the next answer.

EDUARDO ORDAX
WWW.LINKEDIN.COM/IN/EORDAX

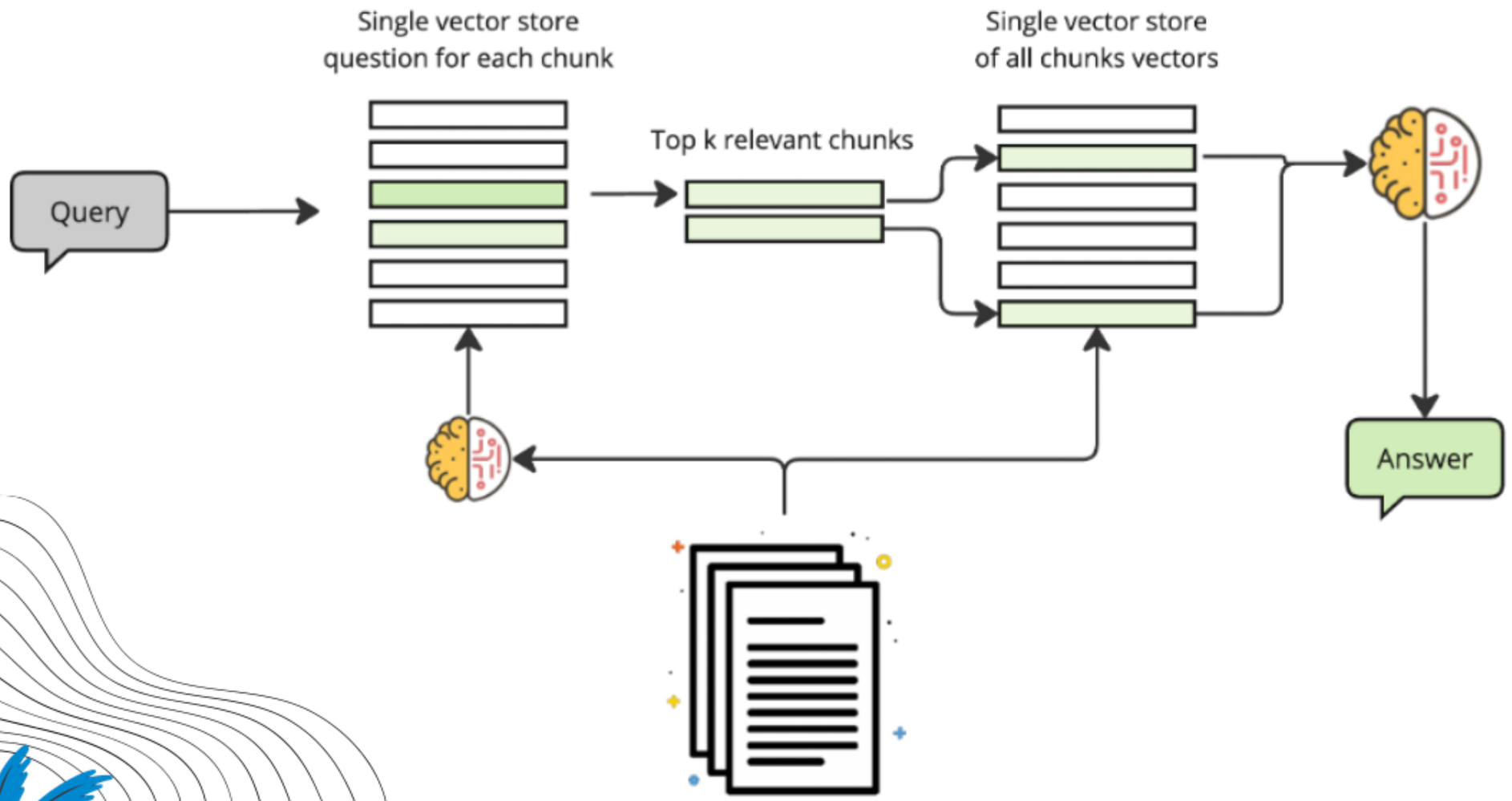
QUERY ROUTING



Query routing is the step of LLM-powered decision making upon what to do next given the user query — the options usually are to summarise, to perform search against some data index or to try a number of different routes and then to synthesise their output in a single answer.
The selection of a routing option is performed with an LLM call, returning its result in a predefined format, used to route the query to the given index

EDUARDO ORDAX
WWW.LINKEDIN.COM/IN/EORDAX

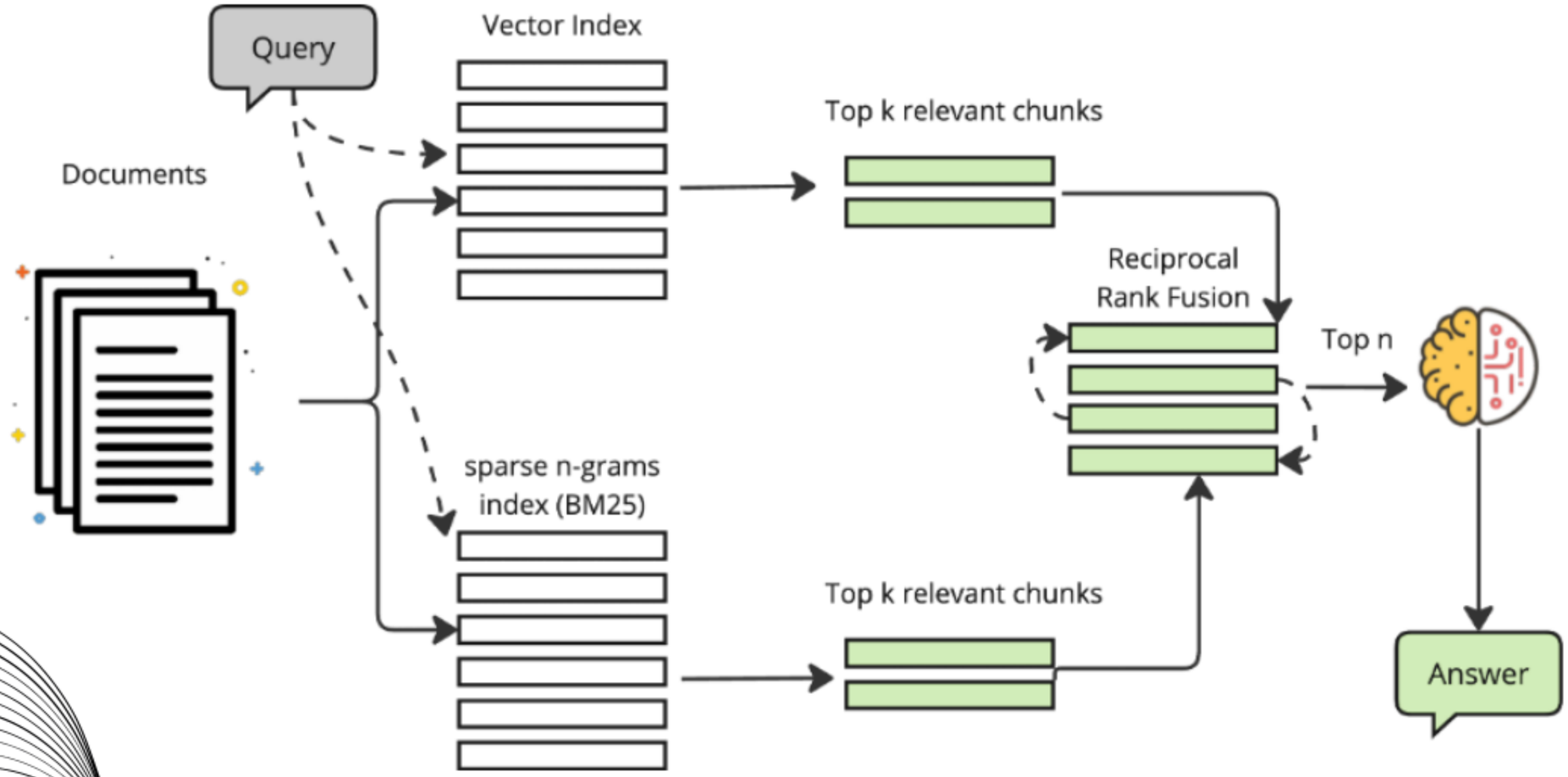
HYDE RETRIEVAL



Another approach is to ask an LLM to generate a question for each chunk and embed these questions in vectors. At runtime, performing query search against this index of question vectors (replacing chunks vectors with questions vectors in our index) and then after retrieval route to original text chunks and send them as the context for the LLM to get an answer. This approach improves search quality due to a higher semantic similarity between query and hypothetical question

EDUARDO ORDAX
WWW.LINKEDIN.COM/IN/EORDAX

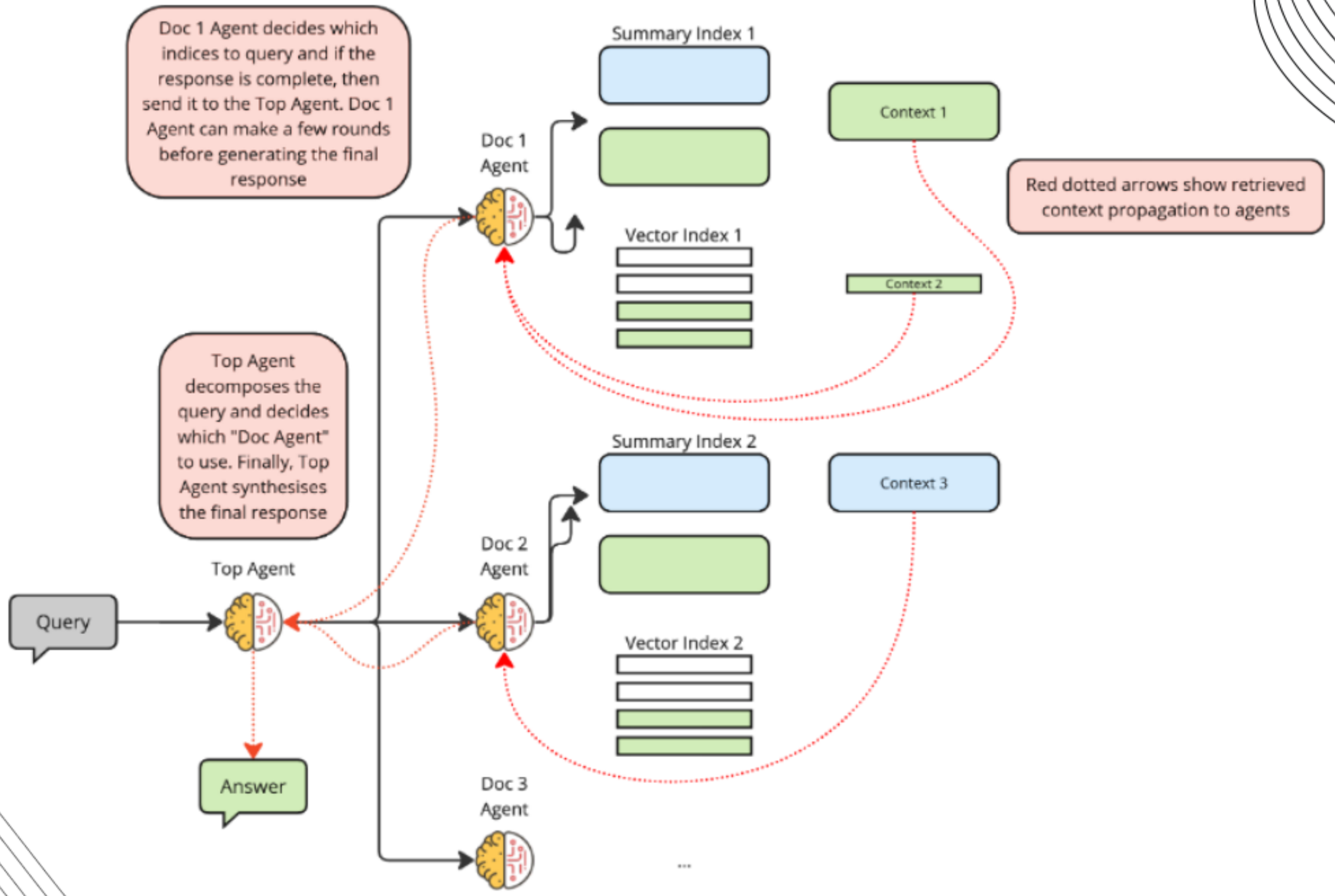
HYBRID SEARCH



Hybrid search combines two distinct search methodologies: keyword-based search and vector-based search. Reciprocal Rank Fusion algorithm is used for reranking the retrieved results with different similarity scores for the final output

EDUARDO ORDAX
WWW.LINKEDIN.COM/IN/EORDAX

RAG WITH AGENTS



THANKS!

Visualisation by Eduardo Ordax (https://www.linkedin.com/in/eordax/)
Original Idea by Eduardo Ordax

ABNASIA.ORG