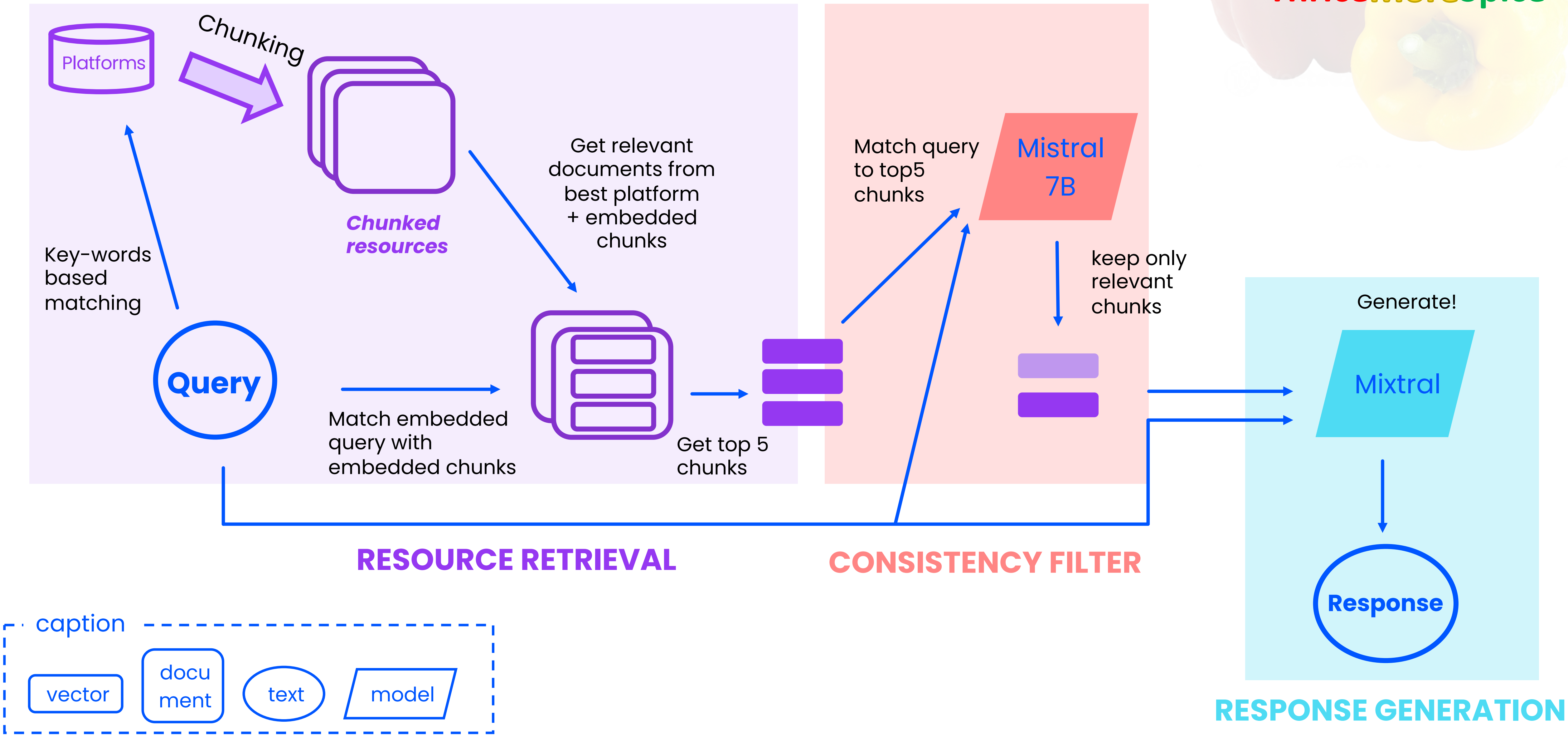
Proposed Architecture

Team Name: 3xPlusDePiment

ThriceMoreSpice



Platforms

Chunking

Chunked resources

Key-words based matching

Get relevant documents from best platform + embedded chunks

Query

Match embedded query with embedded chunks

Get top 5 chunks

Match query to top5 chunks

Mistral 7B

keep only relevant chunks

Generate!

Mixtral

Response

RESOURCE RETRIEVAL

CONSISTENCY FILTER

RESPONSE GENERATION

caption

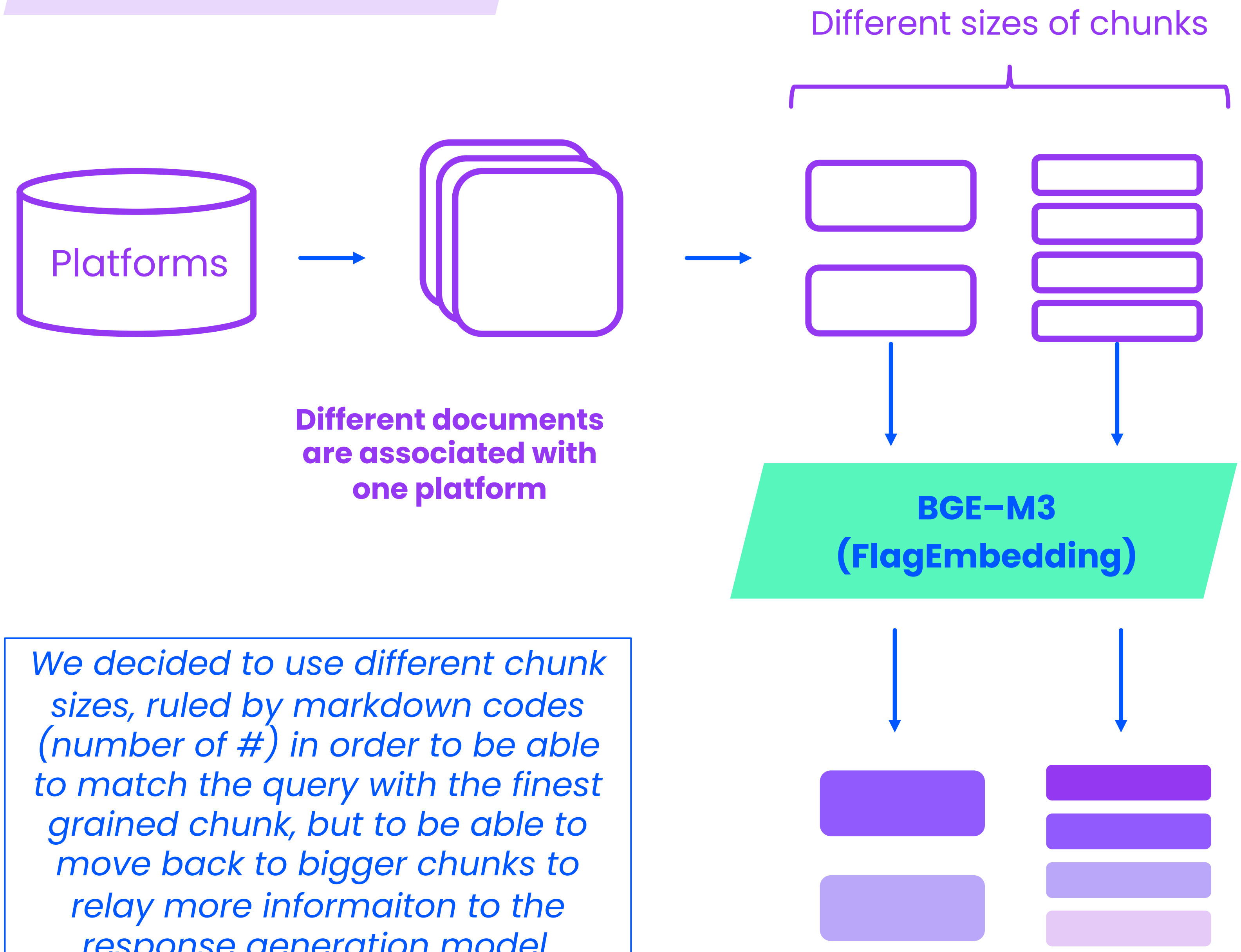
vector

docu ment

text

model

Chunking Strategy


Platforms
Different documents are associated with one platform
Different sizes of chunks
BGE–M3
(FlagEmbedding)

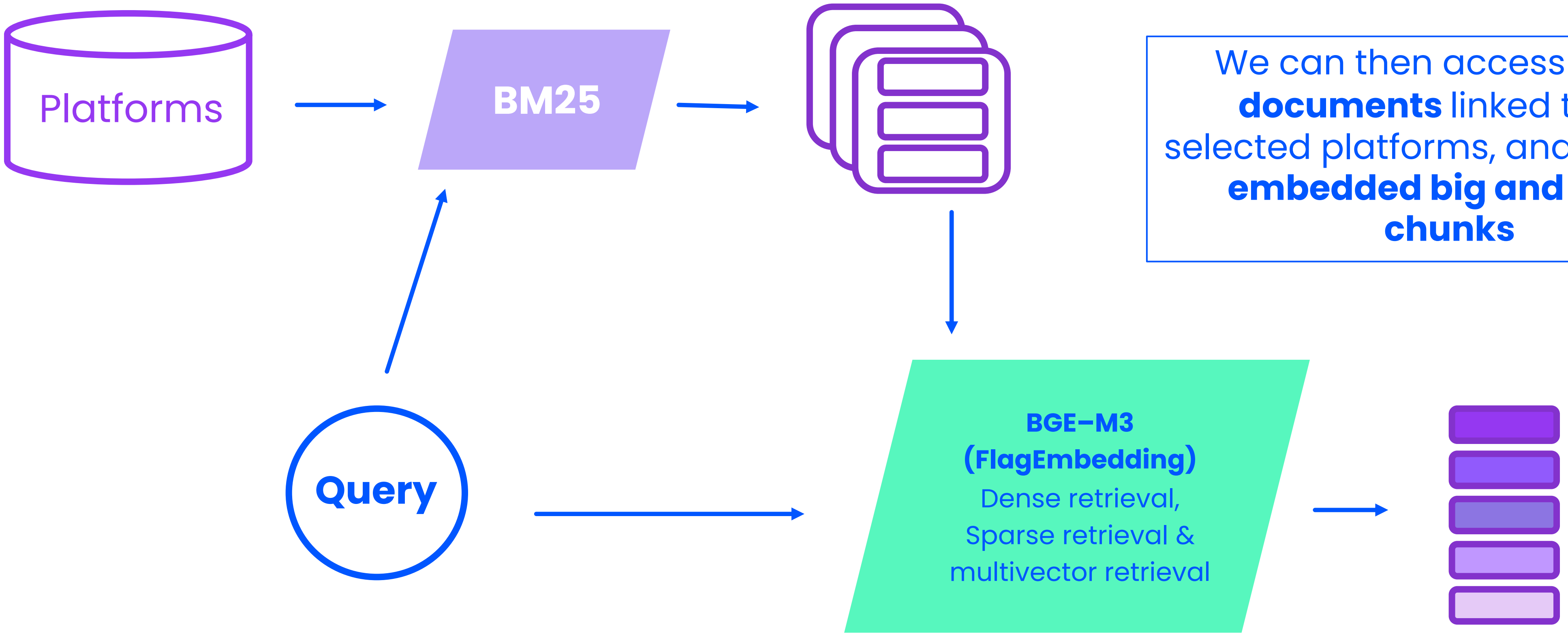
Using normalised Markdown code, chunks of different sizes are extracted from each document, for each platform.

We decided to use different chunk sizes, ruled by markdown codes (number of #) in order to be able to match the query with the finest grained chunk, but to be able to move back to bigger chunks to relay more informaiton to the response generation model.

These different-sized chunks are embedded into vectors that can be matched with an embedded query

Resource Retrieval

BM25 is a ranking function based on keywords that is going to select the most (rarely, the 2 most) **relevant platform**.



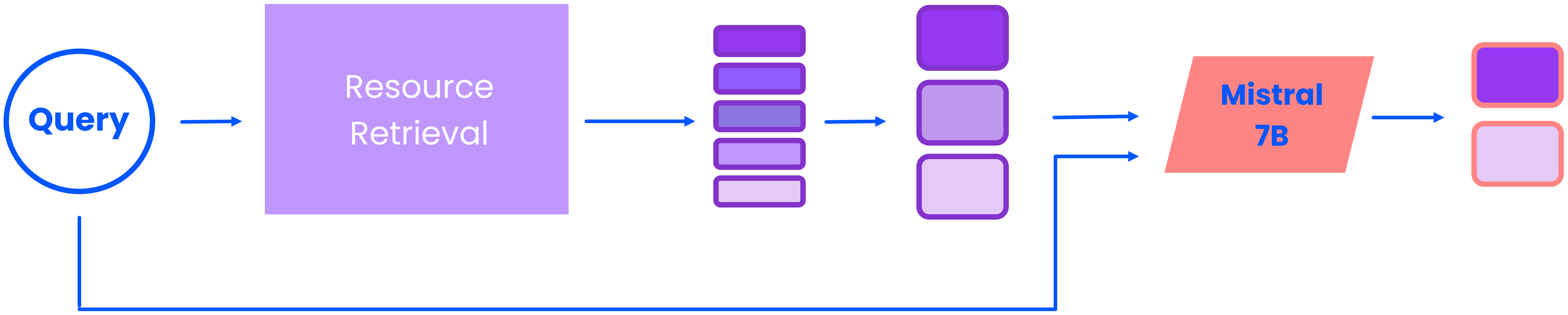
We can then access **all the documents** linked to the selected platforms, and to all **the embedded big and small chunks**

BGE-M3 is brand new (few days old!) embedding and retrieval module that **embeds the query** and matches it to embeddings of **the smaller chunks** of all retrieved documents.

BGE-M3 leverages three retrieval methods to efficiently **select the top-5 relevant chunks.**

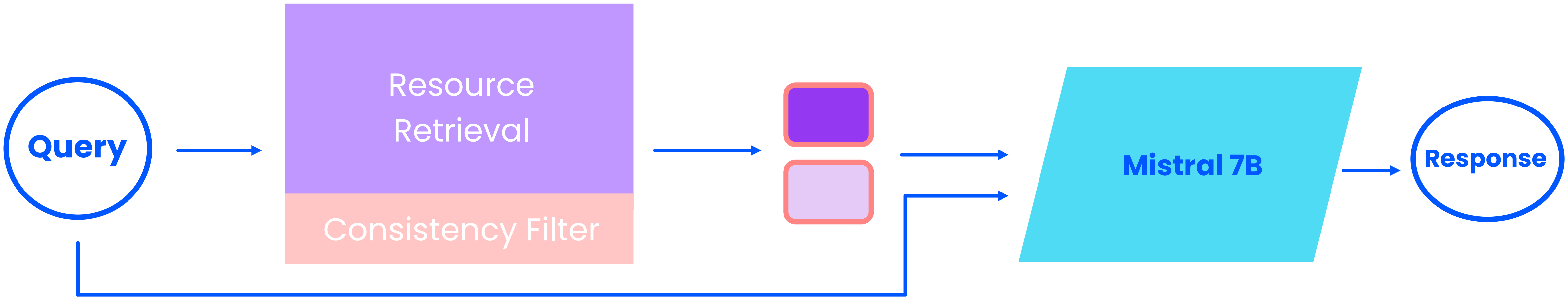
Consistency Filter

We use Mistral 7B because it's a smaller alternative to Mixtral, and we don't need as big a model for this intermediary step



- The general purpose of this step is to **verify which chunks are relevant to the initial query**.
- From the top-5 chunks, we find the **bigger, more complete chunks** (*slide 2 explains our chunking strategy*).
- Using the query, and the retrieved chunks, a Mistral 7B model is asked to return Booleans for each chunk, to **keep only the chunks judged as relevant to the query**.

Response Generation using Retrieved Resources



Now that we have the **query**, as well as the **pertinent resources**, we can generate a prompt that will use both pieces of information. This prompt is sent **to Mistral, that will output the final answer.**