Retrival Augmented Generatoon (RAGs)

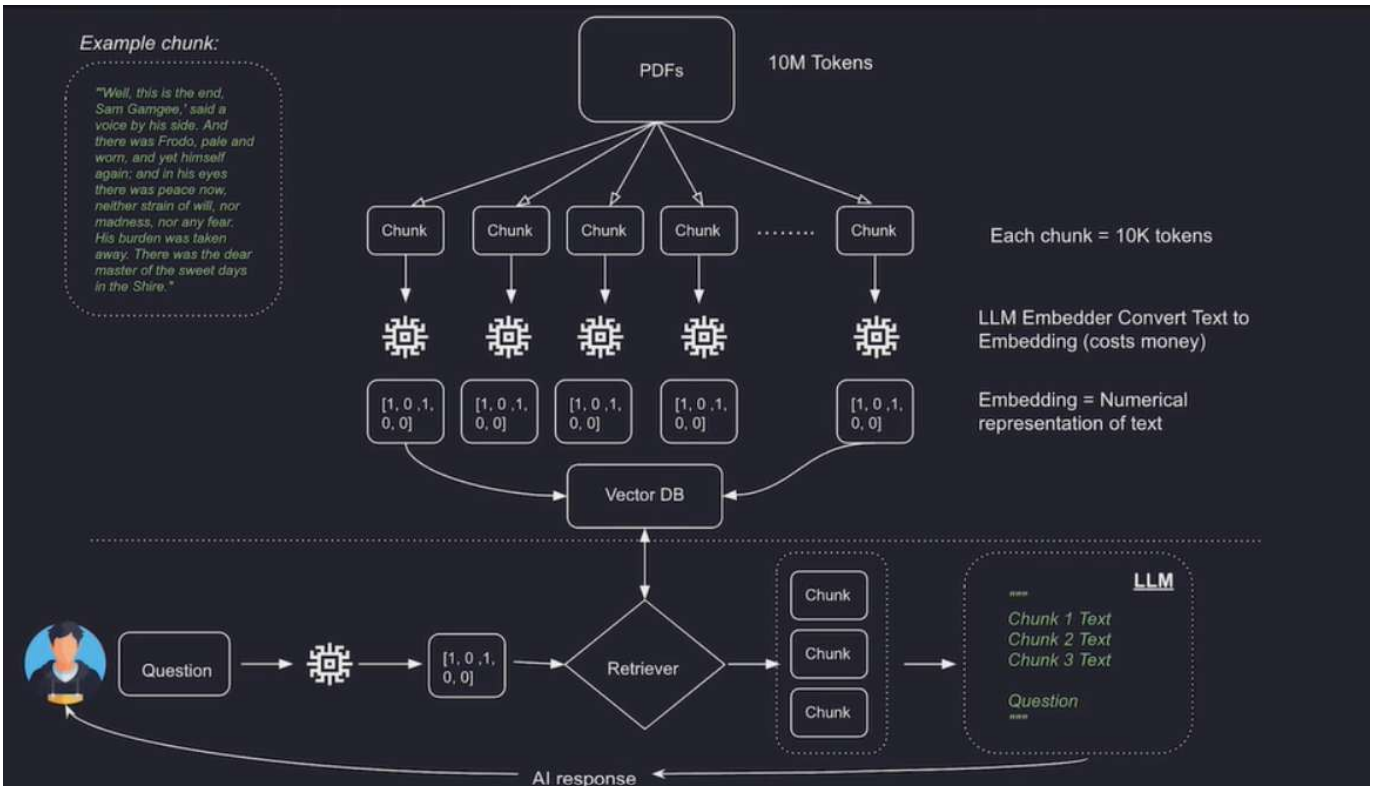
- RAGs gives LLMs additional knowledge
- In other words, we use RAGs to provide LLMs an external source of information to give better answer to our promots.

Example

We can use RAGs to provide LLMs with a list of relevant articles or books to read to answer a question. Now if you have a question , you can ask RAGs to provide you with a list of relevant articles or books to read to answer your question. This is a good way to get a better answer to your question.

Challenge

Context Window Limitation : RAGs can only consider a limited amount of context at a time, which can limit its ability to understand complex questions or provide accurate answers. This is because RAGs are trained on a fixed-size context window, which can make it difficult to capture long-range dependencies or relationships between different pieces of information.

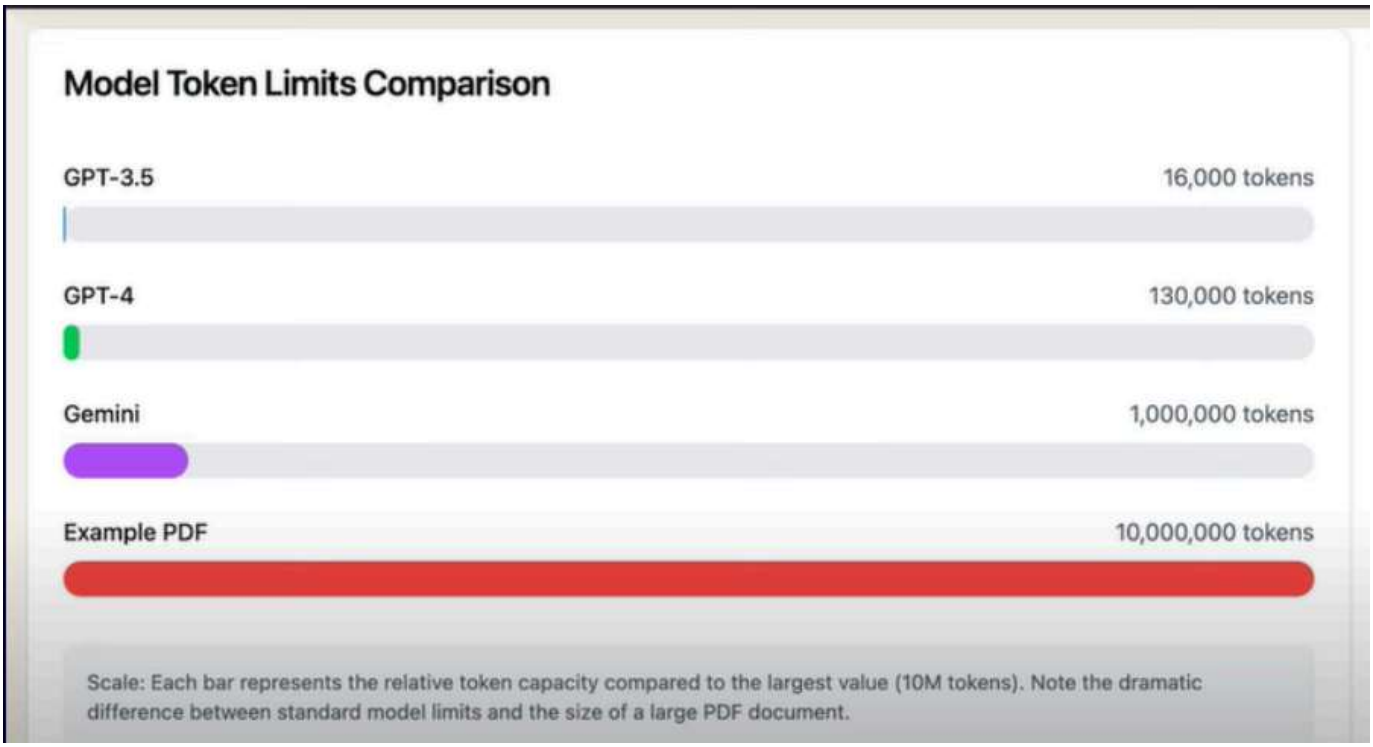


What is tokens?

In the context of language model , tokens are the smallest units of text that can be processed by the model.

- They can be words , subwords (smaller units of words), or even characters. The choice of tokenization depends on the specific model and the task at hand.

- Tokens are crucial because LLMs have a limit on how many tokens they can process in a single input. This limit is often referred to as the "sequence length " or "context length".
    - This means that if you want to input a long piece of text, you need to break it down into smaller chunks , or tokens, that can be processed individually by the model.



Token Embedding Token embedding is a technique used in NLP to convert tokens into numerical vectors that can be processed by the model .

- Each token is mapped to a unique vector in a high-dimensional space, called the *embedding space*.
- The vectors are learned during the training process and capture the semantic meaning of the tokens.

Vector DBs Vector databases are a type of database that stores and manages vectors, such as embeddings, in a way that allows for efficient querying and retrieval of similar vectors.