

Chat Documents Using Retrieval Augmented Generation (RAG)

...

Context Stuffing

A mango is an edible fruit produced by the tropical tree *Mangifera indica*.

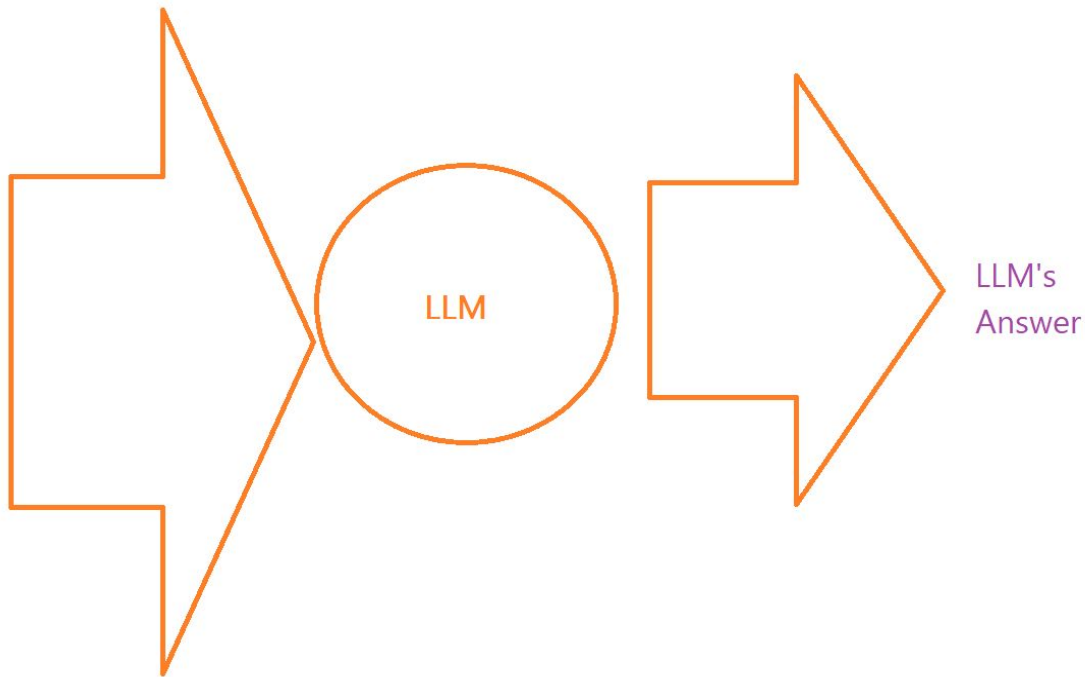
Ripe intact mangoes give off a distinctive resinous, sweet smell.

Mango trees grow to 30–40 metres tall, with a crown radius of 10–15 m.

context =

query =

What is MANGO? Is it safe to eat? It tastes bad?



Steps RAG

1. Create embedding of document and store it in vector DB



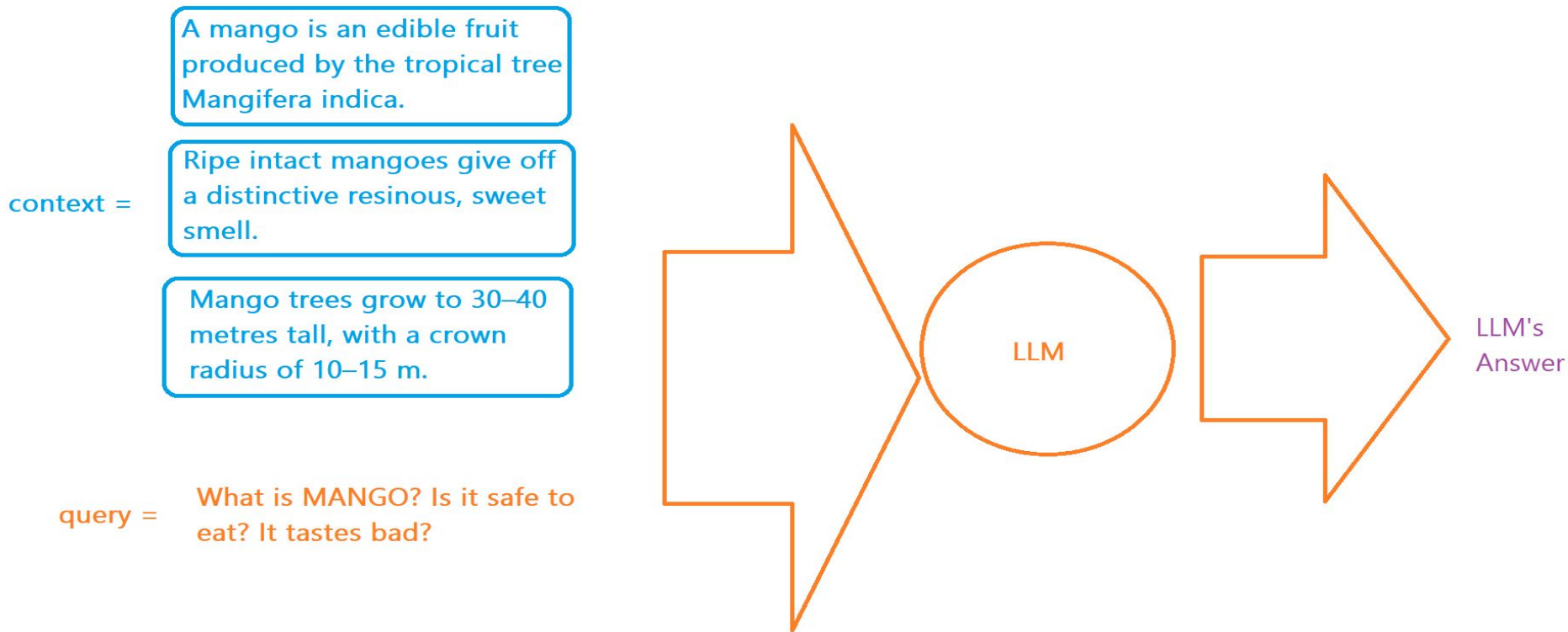
Steps RAG

2. For given query, create vector and perform similarity search in VectorDB

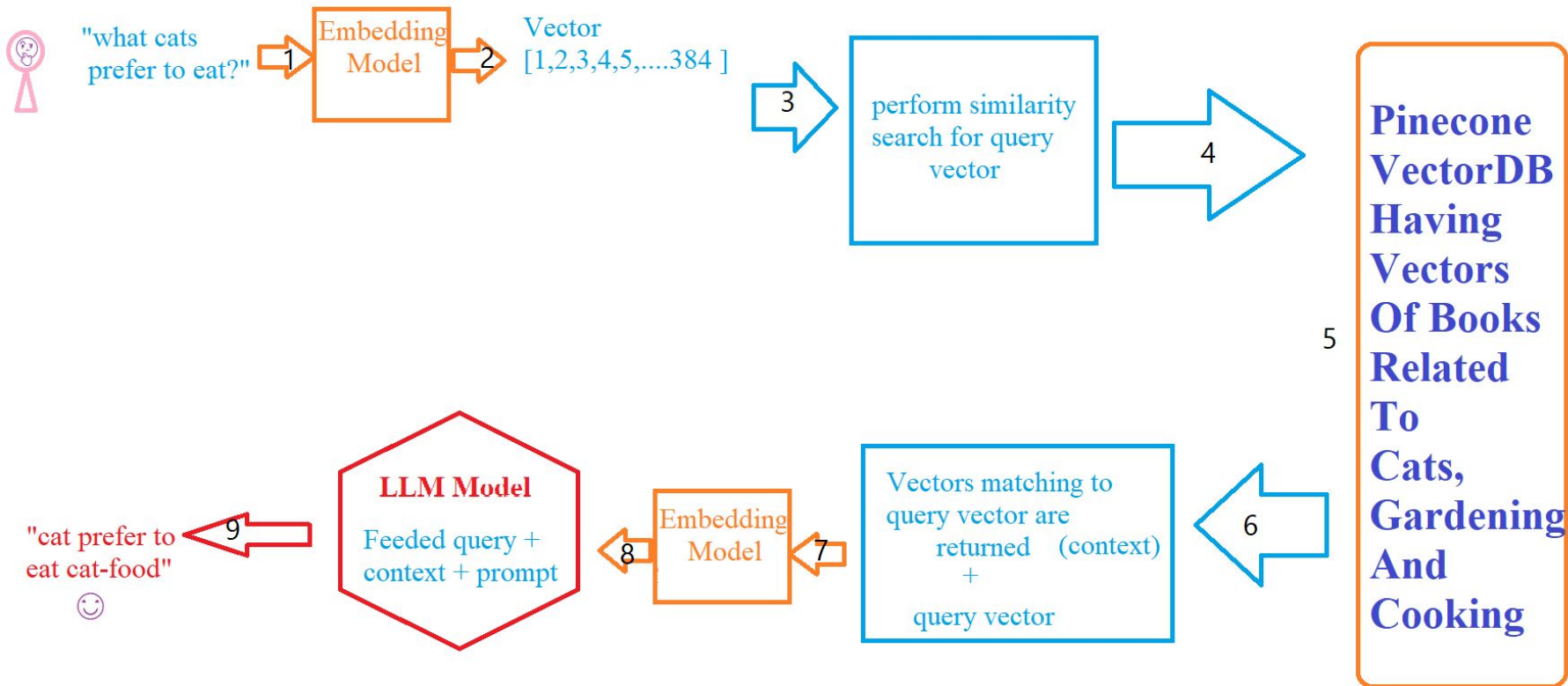


Steps RAG

3. Context stuffing:- create prompt using query and documents received in similarity search



Retrieval Augmented Generation



RAG - Links

[Retrieval Augmented Generation \(RAG\): Reducing Hallucinations in GenAI Applications | Pinecone](#)

https://python.langchain.com/docs/modules/data_connection/

<https://arxiv.org/abs/2307.01155>