31-12-23

DAY-11

Introducing Open source LLM models;

Open source LLM;

(i) not hosted anywhere; need to download the model & load the model manually

(ii) Advantage — no charge

(iii) Disadvantage — You need good configuration system
→ i3, i5 processor
→ 8GB of RAM
→ GPU would be an add on

Some very popular & Powerful open source LLMs

- Meta Llama 2
- Google PaLM 2 { Backend of Google Bard }
- Falcon

Google Search open llms github
↳ this repo has details of all open source LLM

Llama2 is better than gpt 3.5 turbo → meta claims

Search for => llama2.ai

Quantized Model -

→ For a simple neural network there are multiples of parameters which are in float that gets trained/adjusted during BackPropagation (BP).

Data Type	32 bit	64 bit
short	1 byte	1 byte
string	2 byte	2 byte
int	4 byte	4 byte
long	4 byte	4 byte
float	8 byte	8 byte

} out of all datatype float consumes lot of memory/space during BP.

Therefore the process of converting float to integer is called QUANTIZATION.

$w_1 = 1.23$ → After quantization → 1

$w_2 = 2.45$ → After quantization → 4

} offcourse there is a data loss but this is done to quantize the model

Initial Model A (30GB) → After quantization → Quantized model (5GB)

} The performance of quantized model is less compared to raw model

Input text → Pipeline → Response

{ ⓐ Apply preprocessing – Auto Tokenizer

ⓑ Text to number conversion

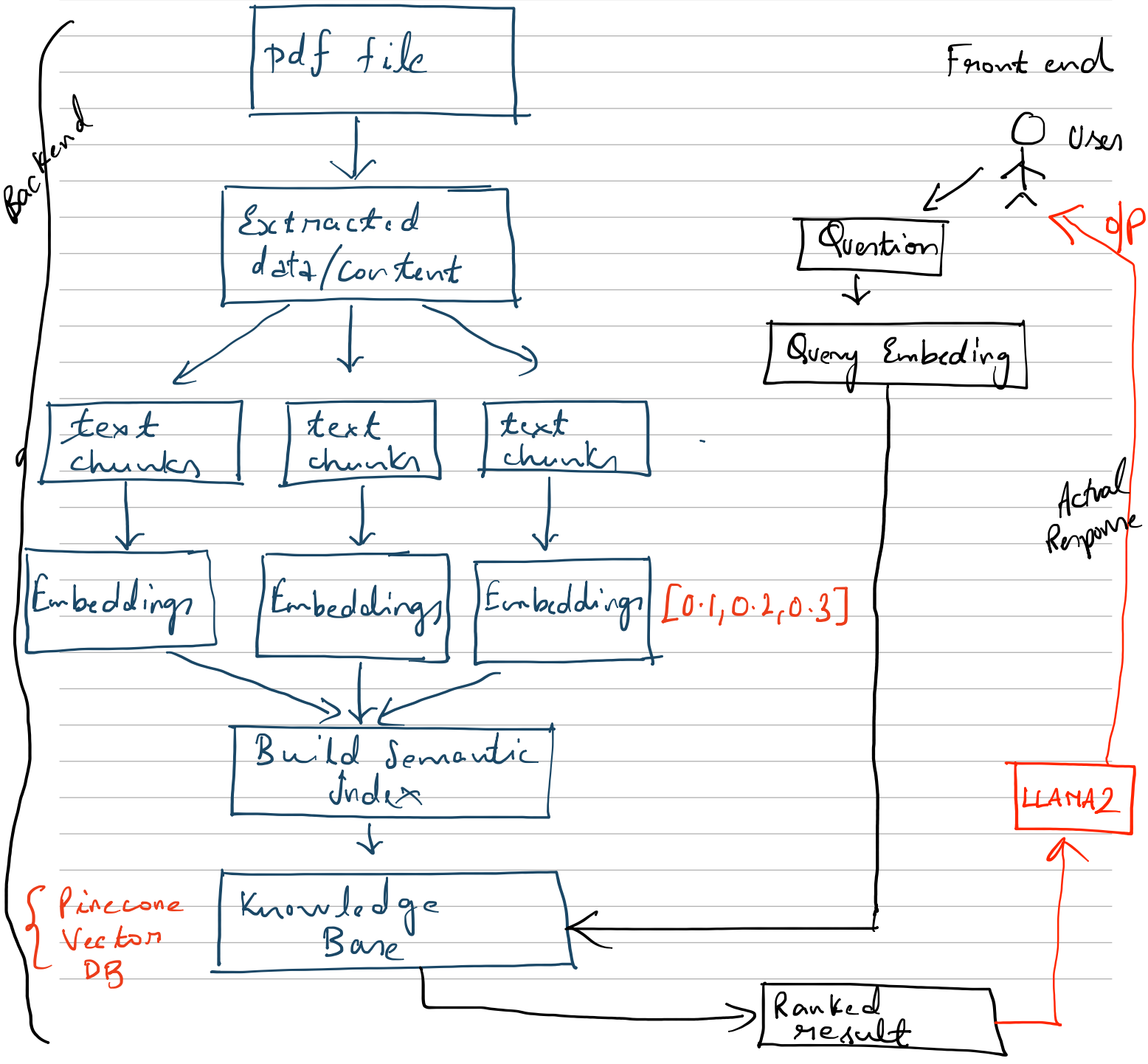
ⓒ Model → prediction

ⓓ Response }

02/01/2024

DAY-12

Implementing a Medical chat bot with custom data ;



Tech Stack used:

1. Python
2. LangChain / Llama Index → Gen AI framework
3. Frontend / web app → flask
4. LLM → Meta Llama 2
5. Vector DB → Pinecone