2024-06-19

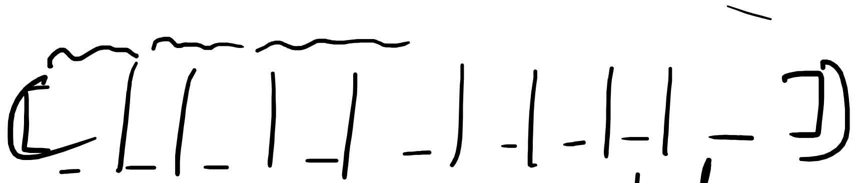
19 June 2024 10:22

Problems w/ LSTM & RNN

-> Slow

-> Requires a lot of data

-> Accuracy (Vanishing Gradient Problem)

-> 

Tokenize then Embed

-> Transformer: Picks up entire sequence.

-> Query : What is the text (Systematic Syntax
Key : Where is the text & Derivable
Value : Here is the text Semantics)

-> Encoder & Decoder

q, k, V Tries to understand q, k, V.

-> GAN ~ Generative Adversarial Networks.

-> Masking What is ◯ name ?
your

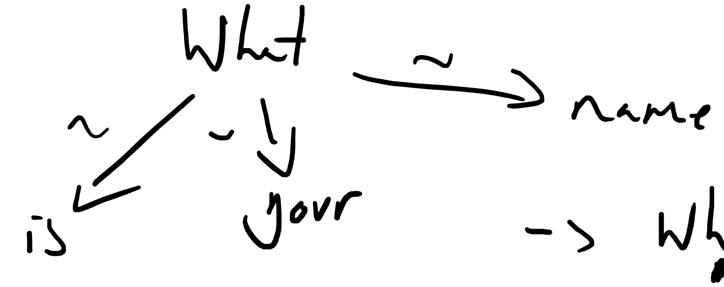
[illegible]

your

-> Multi-Headed Attention

[MSK]

[CLS]



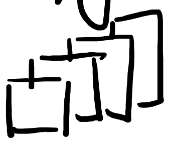
-> What is your name?

-> Cuál es [tu] nombre?

-> Como se llama?

-> Separate Encoders & Decoders

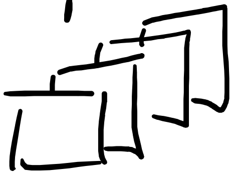
Google



BERT

Bidirectional Encoder Representation Transformer

OpenAI



GPT

Generative Pretrained Transformer

-> Machine Translation & Text-Generation. (GPT)

-> Text Classification & Text Summarization (BERT)

150M 7B

-> GPT-1, GPT-2, GPT-3, GPT-4 ~ Not o source.

-> BERT, RoBERTa, XLNet, DistilBERT ~ Hugging Face

-> BERT, RoBERTa, XLNet, DistilBERT ~ Hugging face
110M 135M 1024 GPU ~ V100

-> GPT-3.5 & GPT-4 ~ FAIR ~ Yann LeCun

-> Meta - LLaMa - 3.5 B parameters
Wikipedia & News ~ 80B tokens

V100 & T100

-> LLM ~ Large Language Model (Hightech transformer)
~ Manifests - Use & calibrate this model .

[Agri / Automobiles / Electronics]

[7000010000]

~ Tell me about Lamb.

-> Prompt Engineering

Refining the query

- GPT ~ OpenAI - Paid
- Claude ~ Anthropic ~ Paid
- Llama ~ Meta ~ Open Source
- Gemma ~ Google ~ OS
- Vistra AI ~ Paid

-> Llama Index

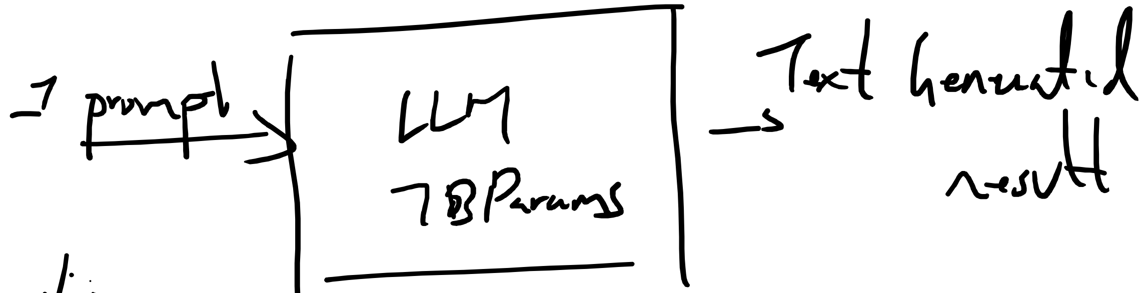
-> Langchain ✓

-> Colab
↳ Local ~ Ollama

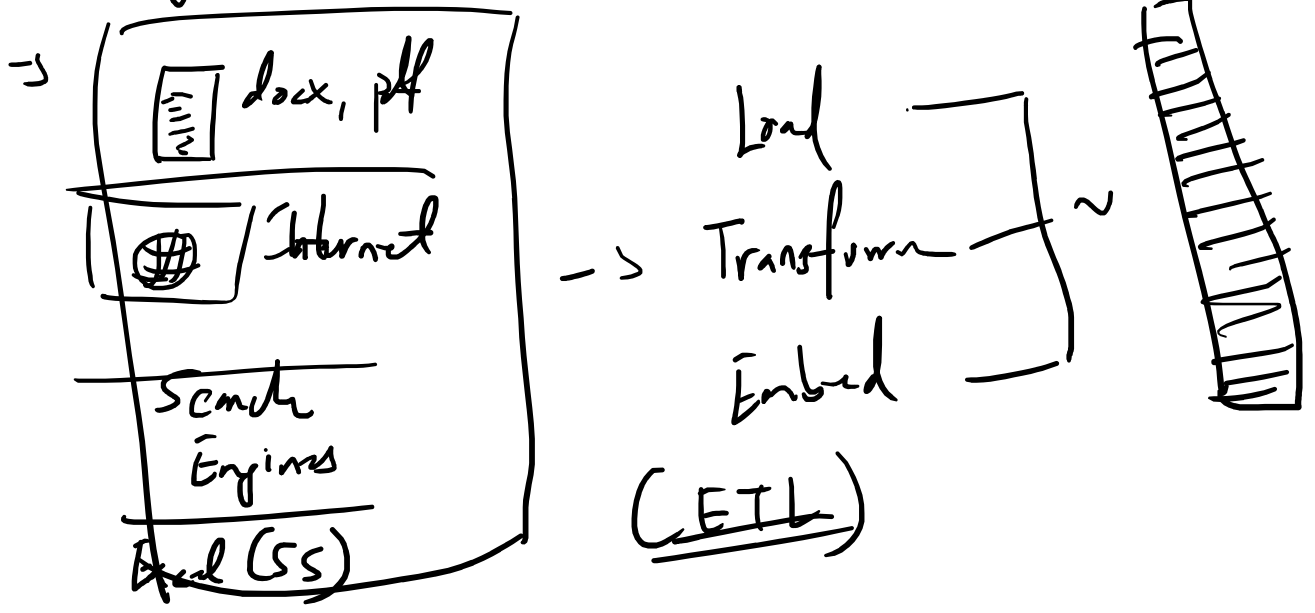
Vistra AI ~ Paid   ↳ Local ~ Ollama

-> **RAG**: Retrival Augmented Generation

-> Information is sparsed

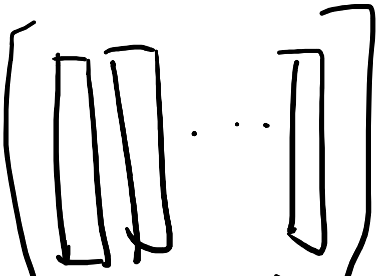


Data Ingestion



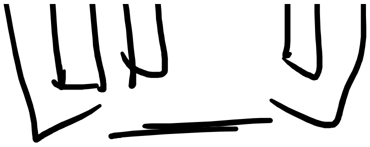
-> Load, Transform & Embed converts data into vectors.

-> Chroma DB
-> FAISS
-> ObjectBox



~ Vector Store or Vector Database

}-> FAISS
L-> ObjectBox
-> Cassandra
(ASTRA)



Vector Database

Map & Reduce

-> Similarity Search

Task: Create a query engine for paper.