Raphaël MANSUY

https://www.linkedin.com/in/raphaelmansuy/

AI

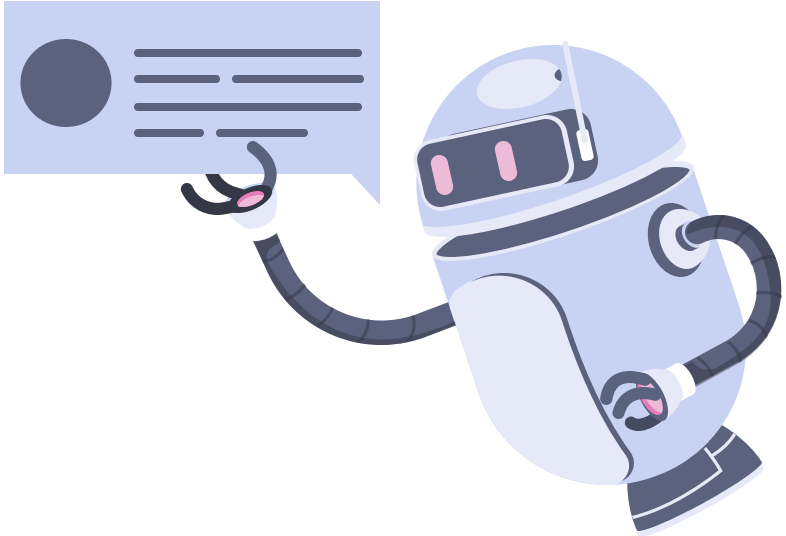
A gentle introduction to Generative AI

Hong Kong - 27/08/2023

01 →

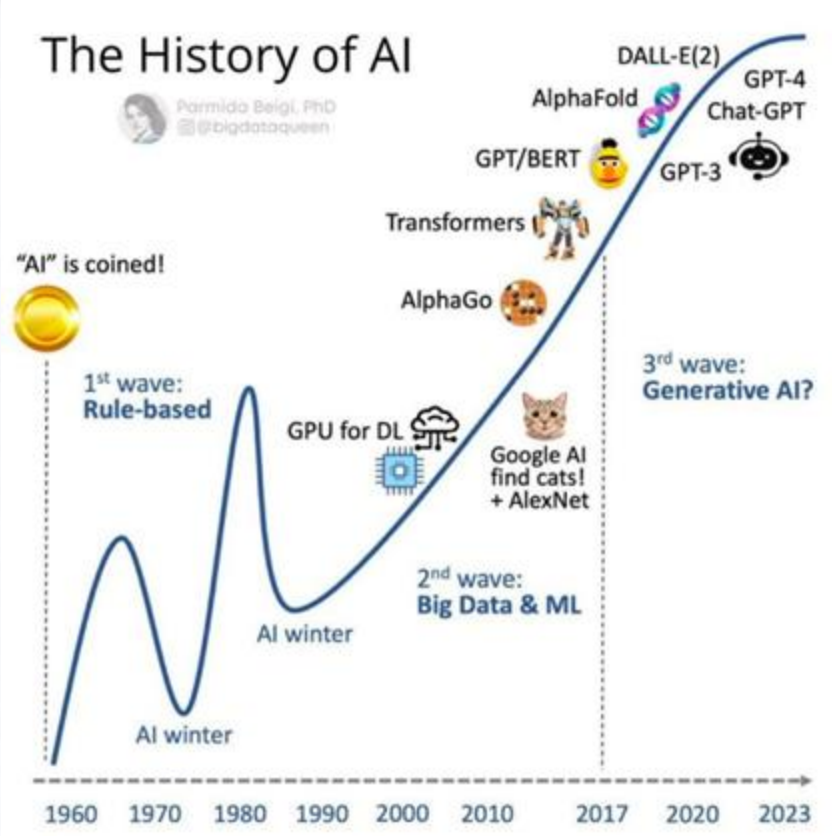
What is artificial intelligence?

(AI) = Artificial intelligence



Artificial intelligence

The **simulation** of human intelligence in machines that are programmed to **think** and **learn** like humans, including tasks such as visual perception, speech recognition, decision-making, image creation, and language translation.



02 →

What is generative AI ?

Generative AI

Generative AI is a branch of AI that focuses on creating new content.

- Text generation (LLM / transformers)
- Music generation (transformers)
- Image generation (Diffusion models)

ChatGPT

Midjourney

Aiva

Generative AI 🧠

Generative AI is a branch of AI that focuses on creating new content.

- Text generation (LLM / transformers)
- Music generation (transformers)
- Image generation (Diffusion models)

ChatGPT

Midjourney

Aiva

What is an LLM ?

Large

Language

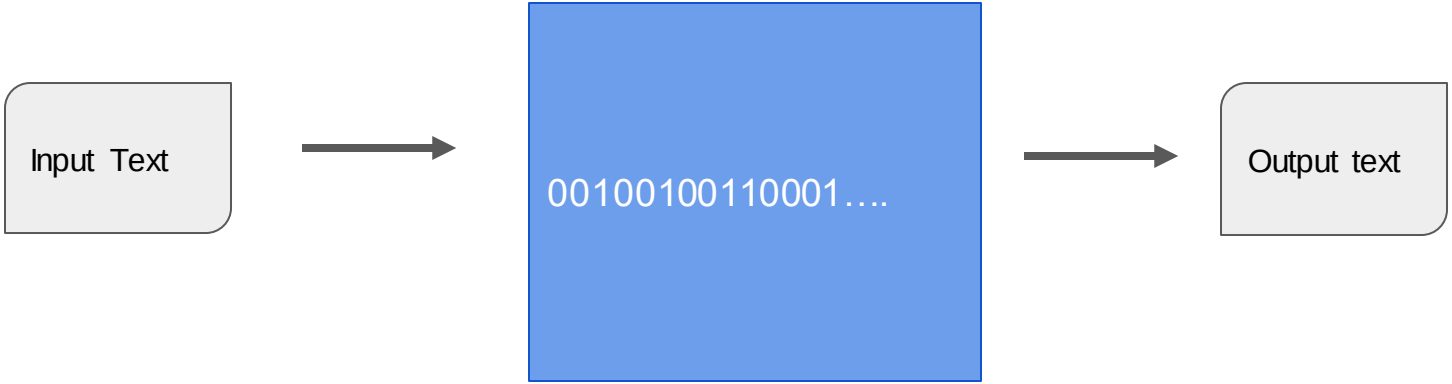
Model

“LLM Are Alien Technology”

Simon Willison

A LLM is just a Function …

LLM: A **F**unction **F**ormed from **N**umbers



model

An AI 🧠 Model … is just a set of binary files …

main · longchat-7b-v1.5-32k

lmzheng Update config.json 16deb63

.gitattributes		1.52 kB
config.json		720 Bytes
generation_config.json		174 Bytes
pytorch_model-00001-of-00002.bin	pickle	9.98 GB LFS
pytorch_model-00002-of-00002.bin	pickle	3.5 GB LFS
pytorch_model.bin.index.json		26.8 kB
special_tokens_map.json		435 Bytes
tokenizer.model		500 kB LFS
tokenizer_config.json		747 Bytes

🤗 Hugging Face

Search models, datasets, users...

lmsys/longchat-7b-v1.5-32k

like 34

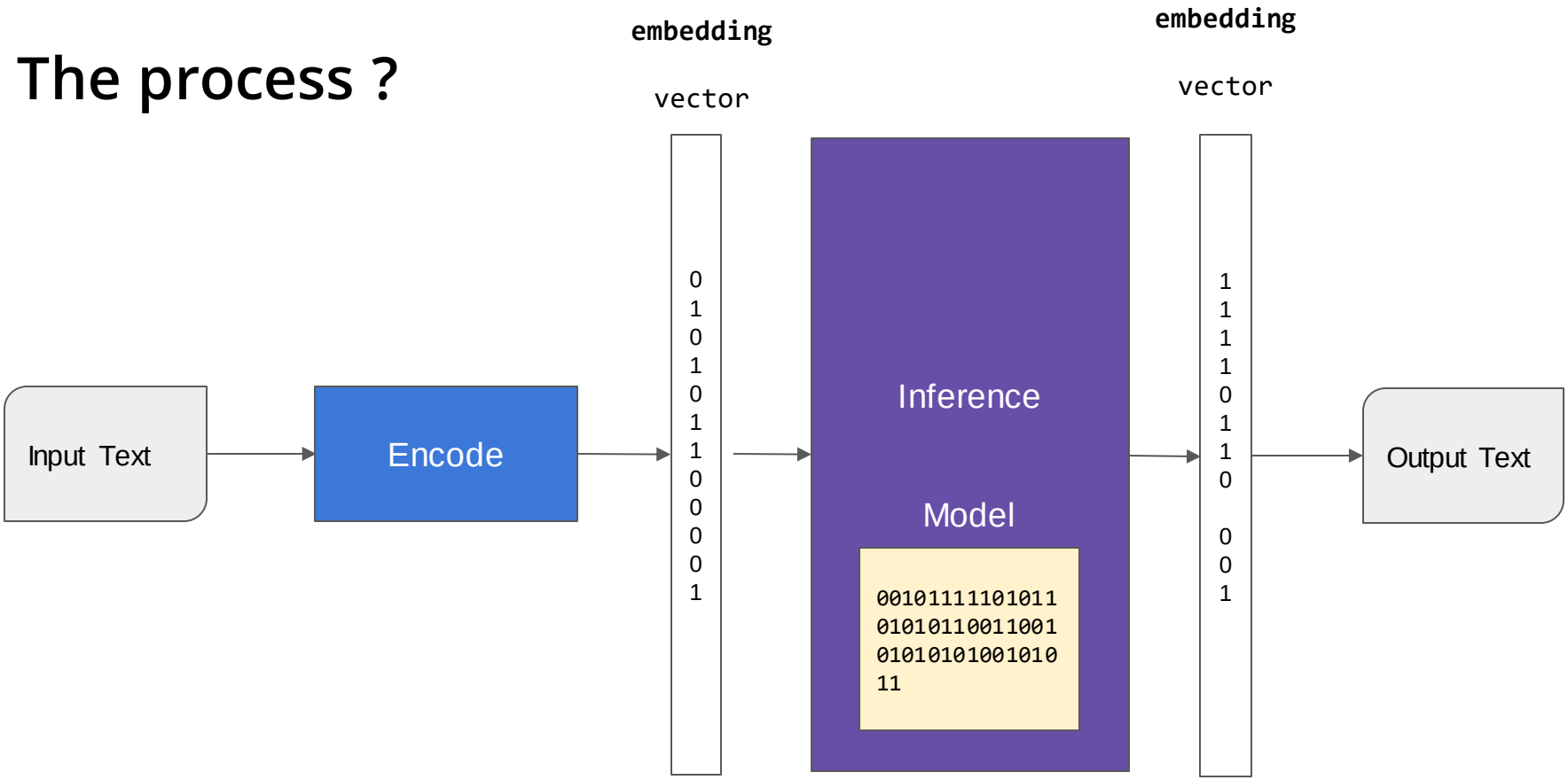
Text Generation · PyTorch · Transformers · llama · text-generation-inference

How to use it ?

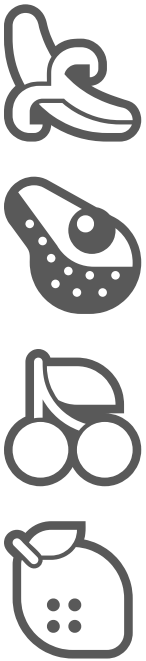
1. Load the model
2. Use an Inference function that use the weights of the model
3. Create message
4. Encode the message
5. Submit the encoded message function to the inference layer
6. Get the output from the inference layer
7. Decode the output to generate text

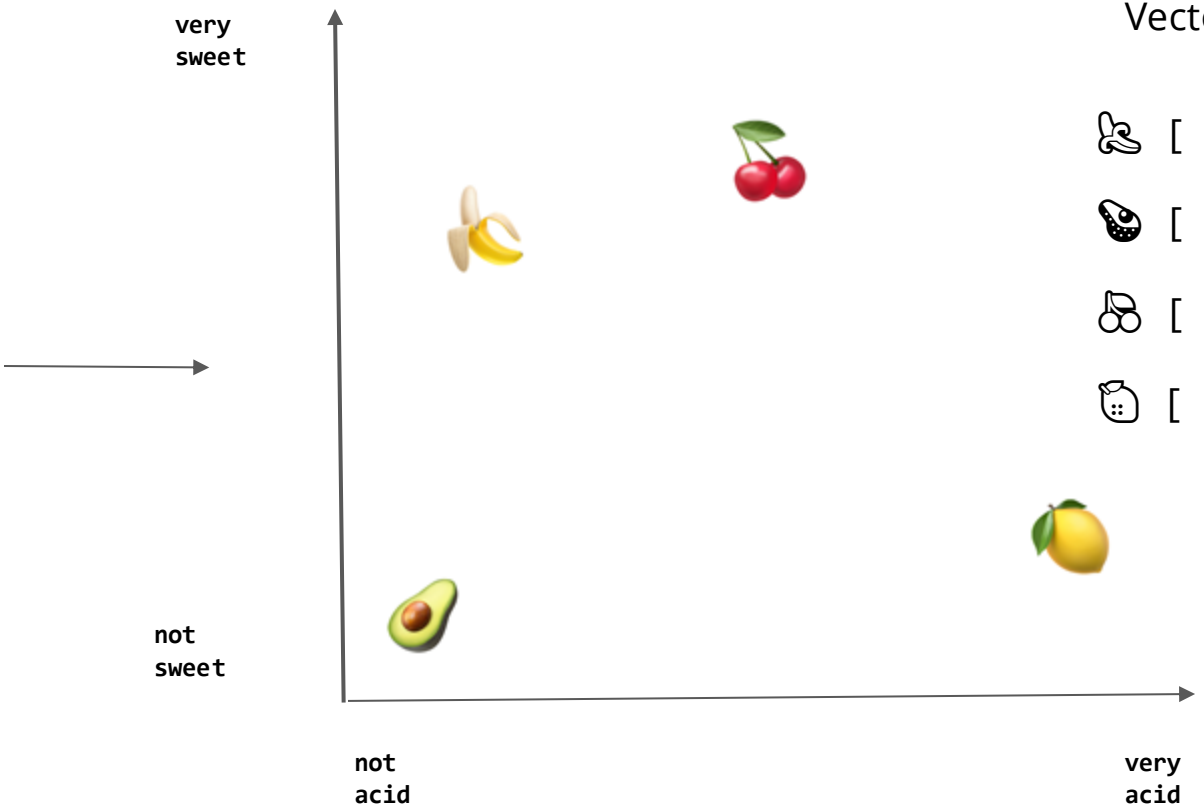
```
import torch
from transformers import AutoModelForCausalLM, AutoTokenizer, pipeline

tokenizer = AutoTokenizer.from_pretrained(
    "stabilityai/StableBeluga-7B", use_fast=False)
model = AutoModelForCausalLM.from_pretrained(
    "stabilityai/StableBeluga-7B", torch_dtype=torch.float16, low_cpu_mem_usage=T
system_prompt = "### System:\nYou are StableBeluga, an AI that follows instructio

message = "Write me a poem please"
prompt = f"{system_prompt}### User: {message}\n\n### Assistant:\n"
inputs = tokenizer(prompt, return_tensors="pt").to("cuda")
output = model.generate(**inputs, do_sample=True,
                        top_p=0.95, top_k=0, max_new_tokens=256)

print(tokenizer.decode(output[0], skip_special_tokens=True))
```

The process ?


embedding
vector
embedding
vector
Input Text
Encode
0
1
0
1
0
1
1
0
0
0
0
1
Inference
Model
00101111101011
01010110011001
01010101001010
11
1
1
1
1
0
1
1
0
0
0
1
Output Text

A fictive 2D Embedding / embeddings = “meaning”





Vectors

[0.2, 0.7]

[0.1, 0.1]

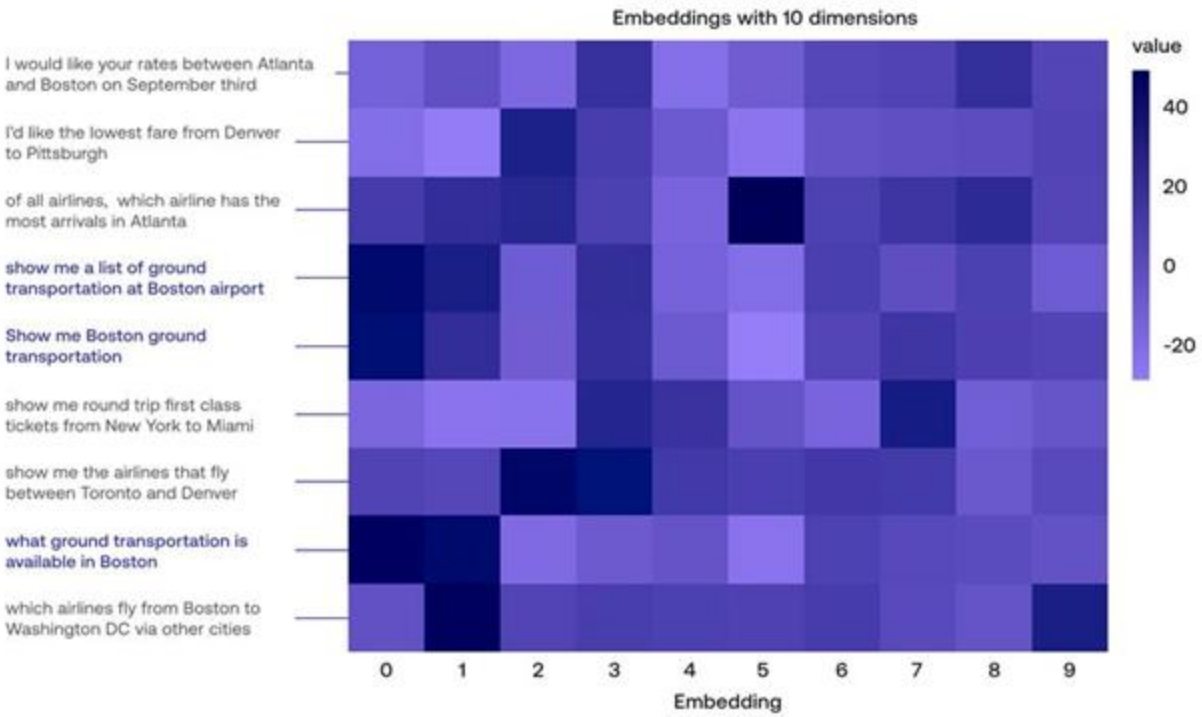
[0.6, 07]

[0.9, 0.2]

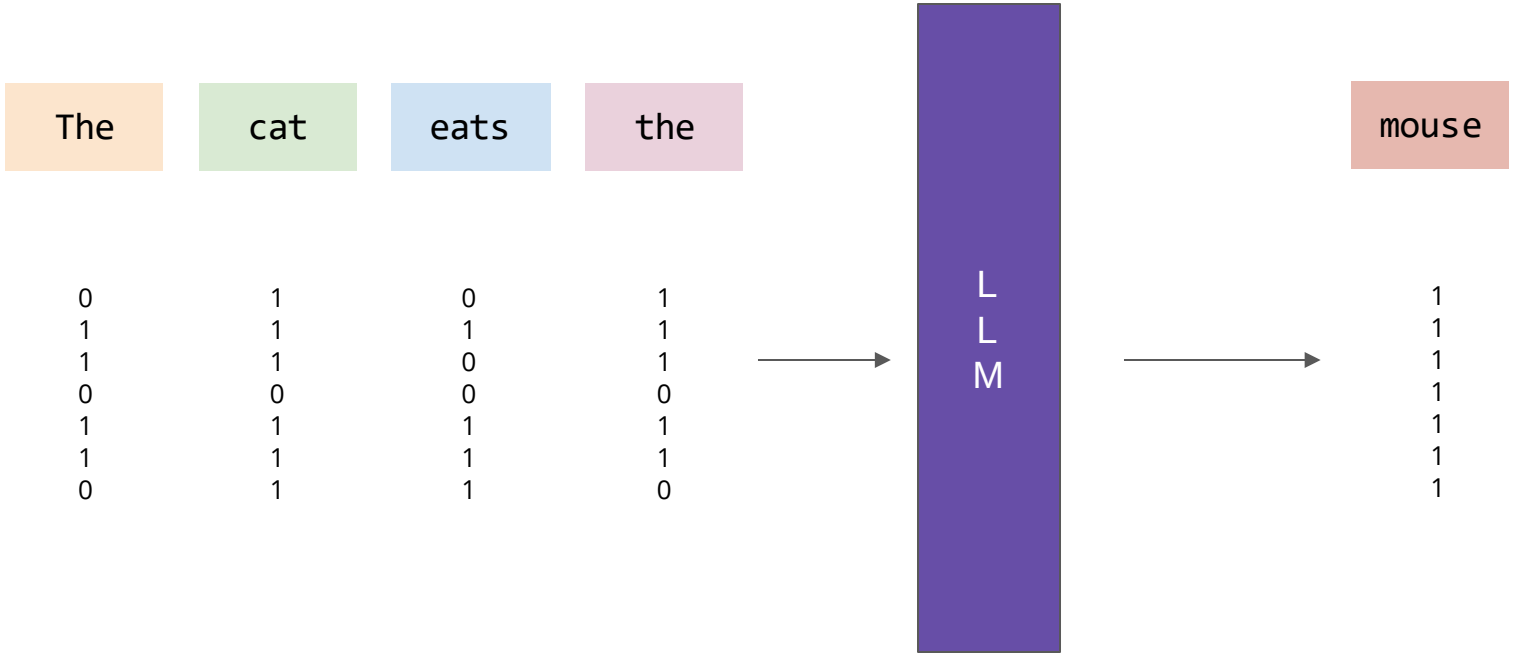
With a LLM we encode in N dimensions

Example

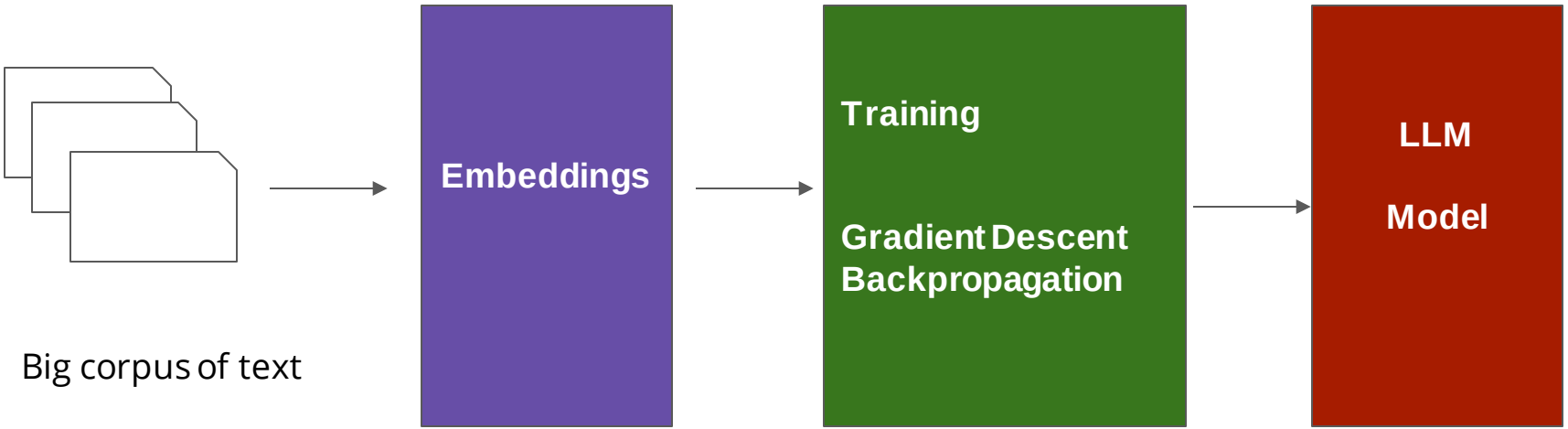
from cohere.com



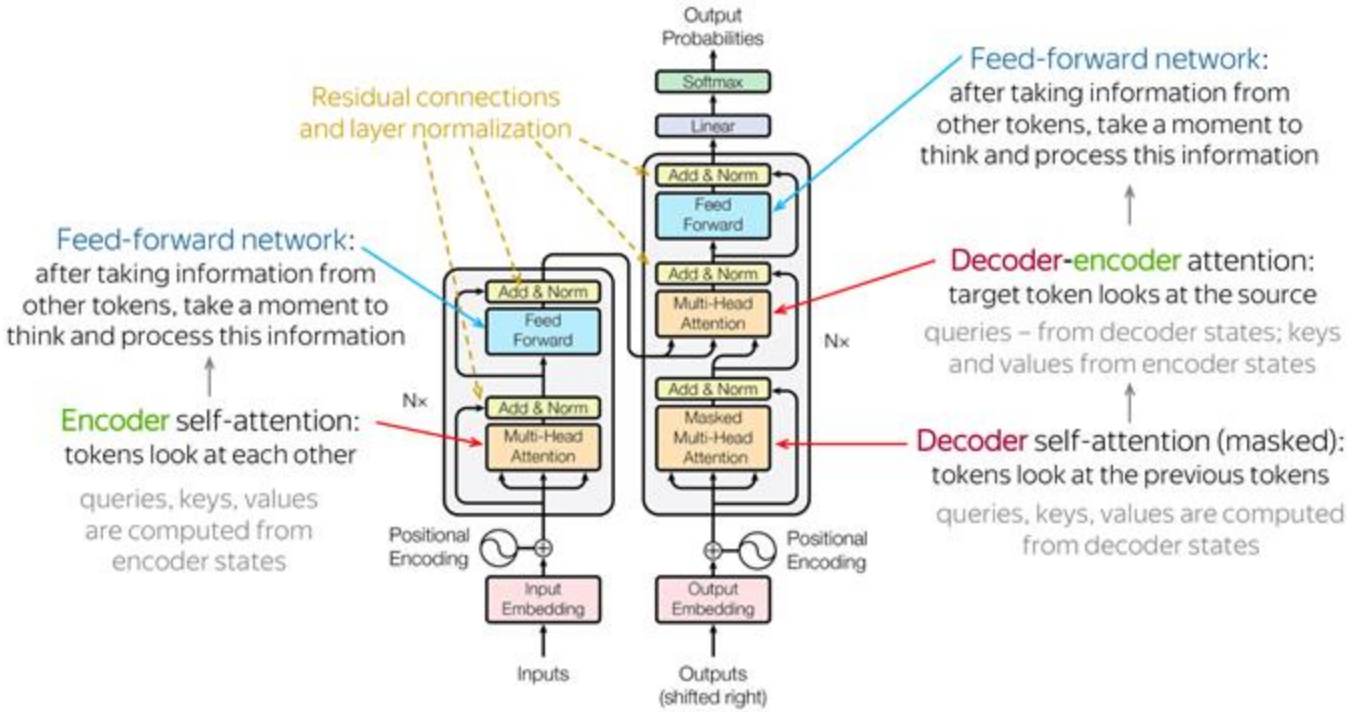
LLM are functions that guess the next words !



How LLM are trained ?



The transformer architecture



03 →

Question ?

Who I am ?

Raphaël MANSUY

CTO ELITIZON ltd

https://www.linkedin.com/in/raphaelmansuy/

TECHNOLOGY VENTURE STUDIO

We launch 🚀 innovative services.

As a technology venture studio we transform **ideas** into **services** and services into **companies**.