2nd Session:
NLP meets GenAI :LLMs & RAG

27/12/2024

Google Developer Groups
On Campus • SUP'COM

EPI Educational Group

By Mohammed Arbi Nsibi

MOHAMED ARBI NSIBI

- Final year ICT engineering student@ SUP'COM
- GDG Carthage member
- Mentor of GDGoC SUP'COM & ISAMM
- Former GDSC Lead 23/24

mohammedarbinsibi@gmail.com

https://huggingface.co/Goodnight7

https://www.linkedin.com/in/mohammed-arbi-nsibi-584a43241/

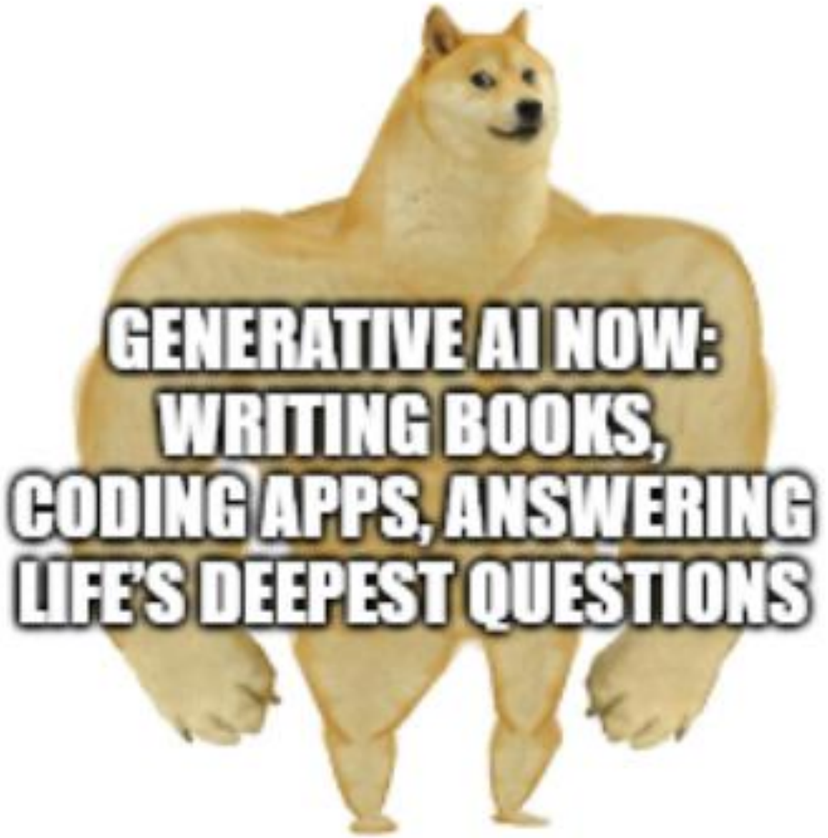
Content

- NLP
  - Pipeline
- LLMs
  - Transformers
  - Let’s code

- Why we need RAG
- Speaking on Your Behalf : Building a ChatBot
- QUIZ

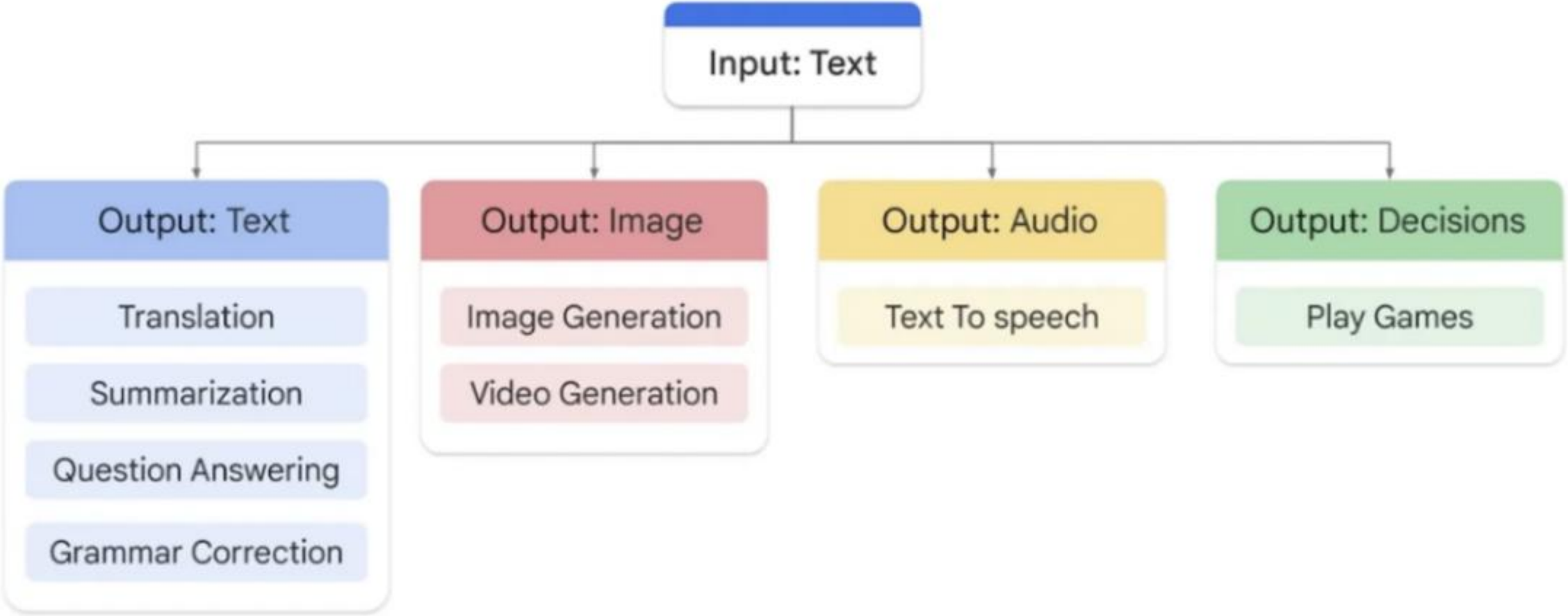
Google Developer Groups
On Campus • SUP'COM

By Mohammed Arbi Nsibi

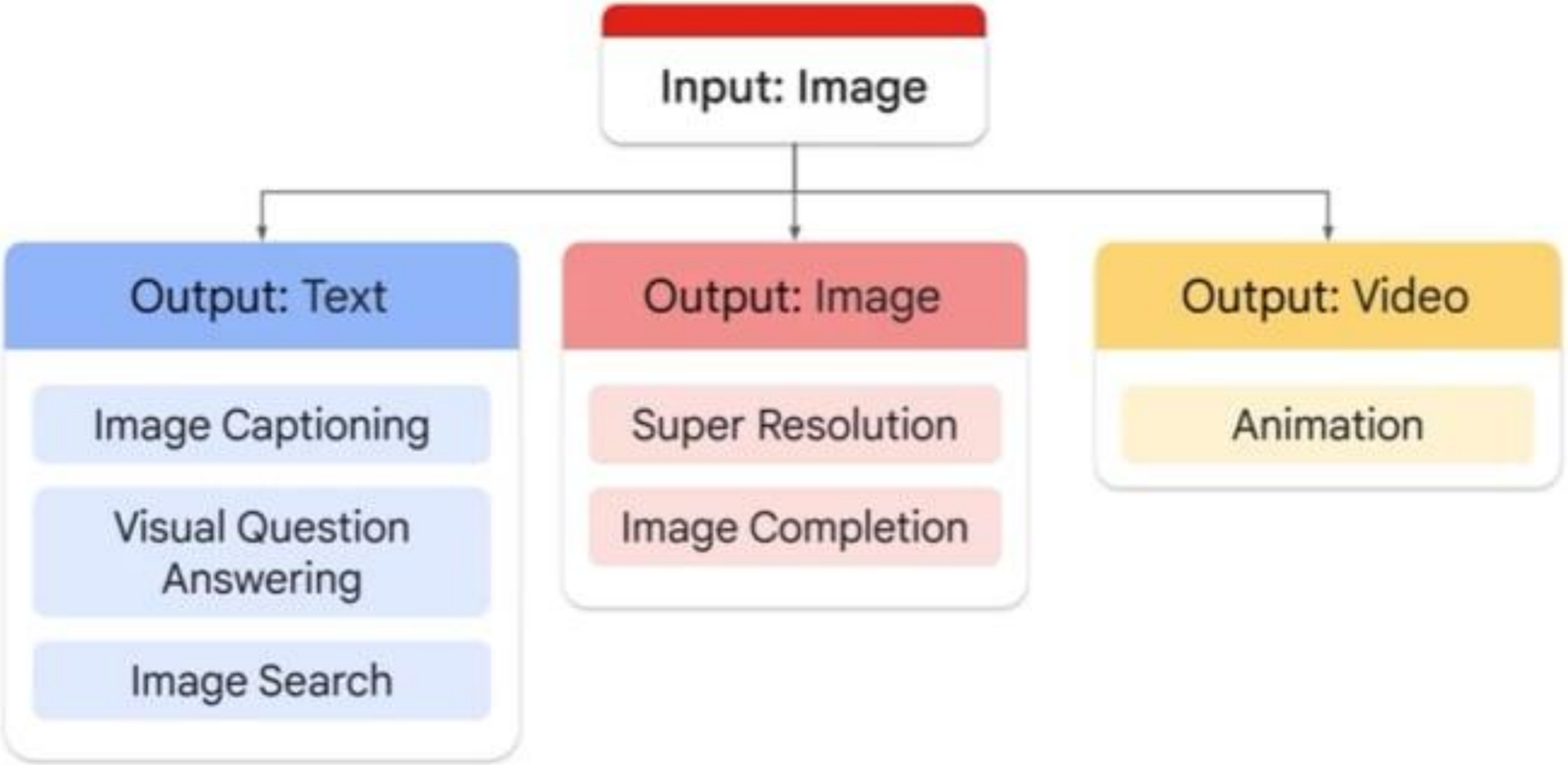


By Mohammed Arbi Nsibi

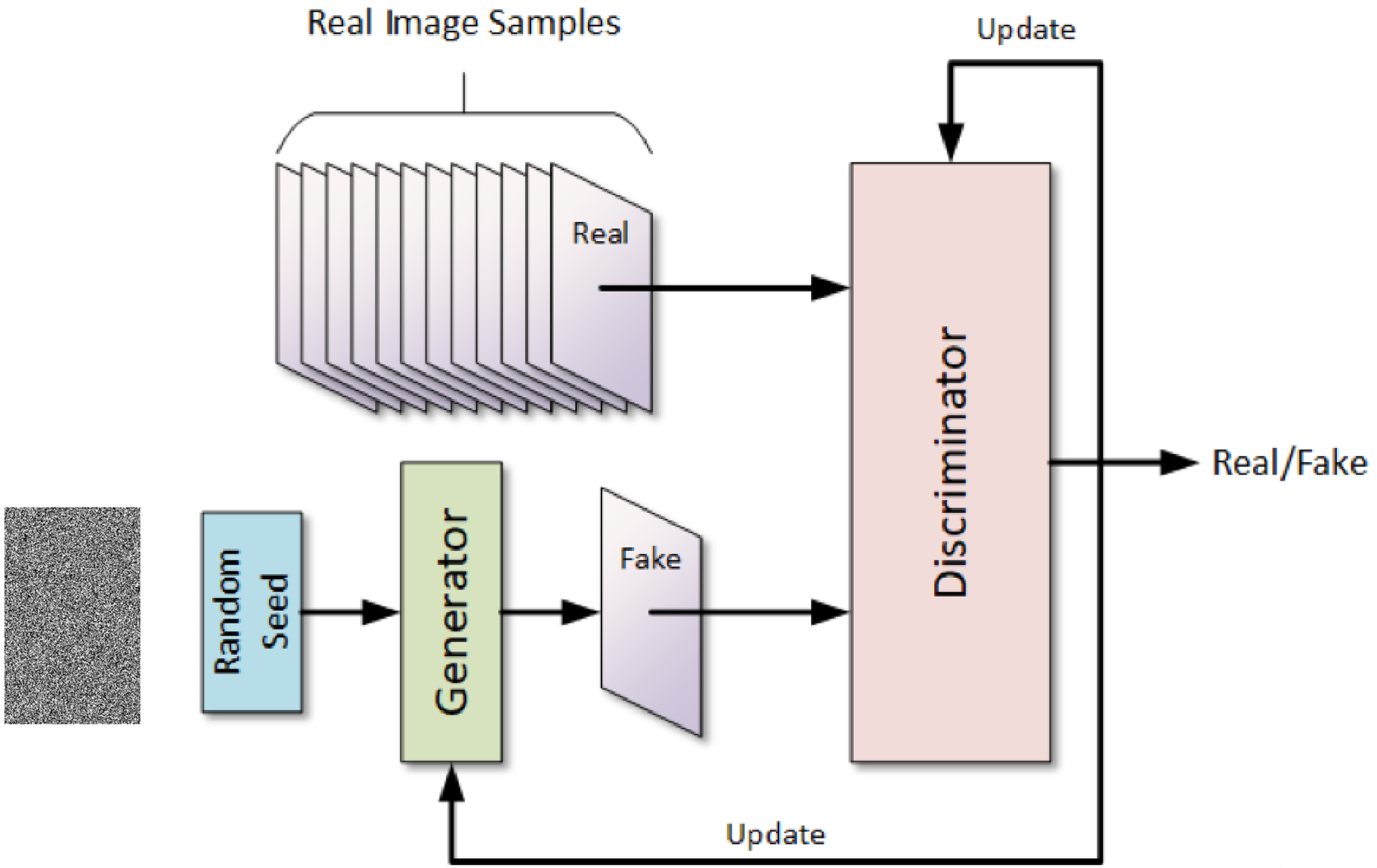
Types of Generative AI Based on Data



Types of Generative AI Based on Data



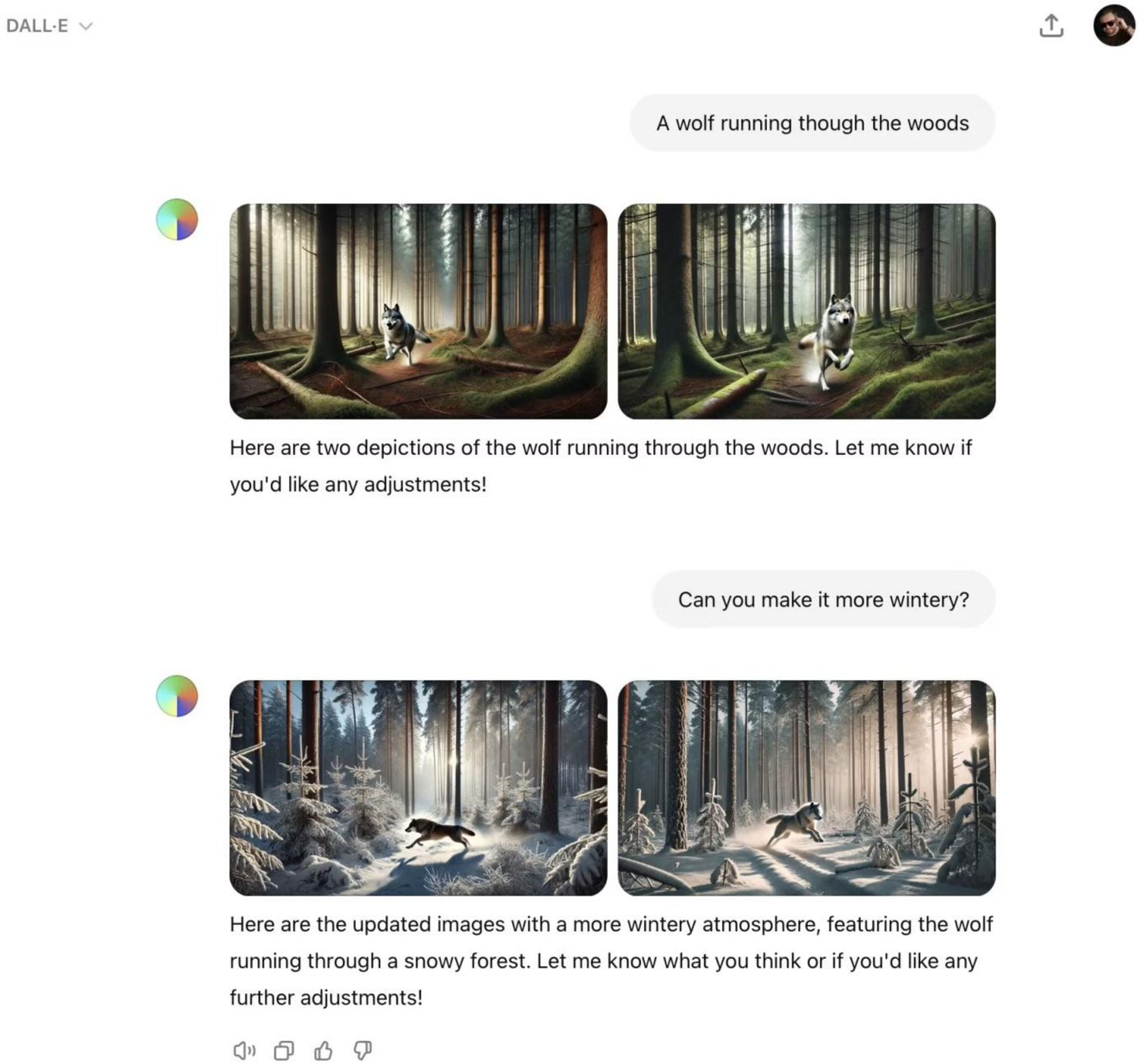
GANs: Generative adversarial network



https://arxiv.org/abs/1406.2661

Diffusion models :

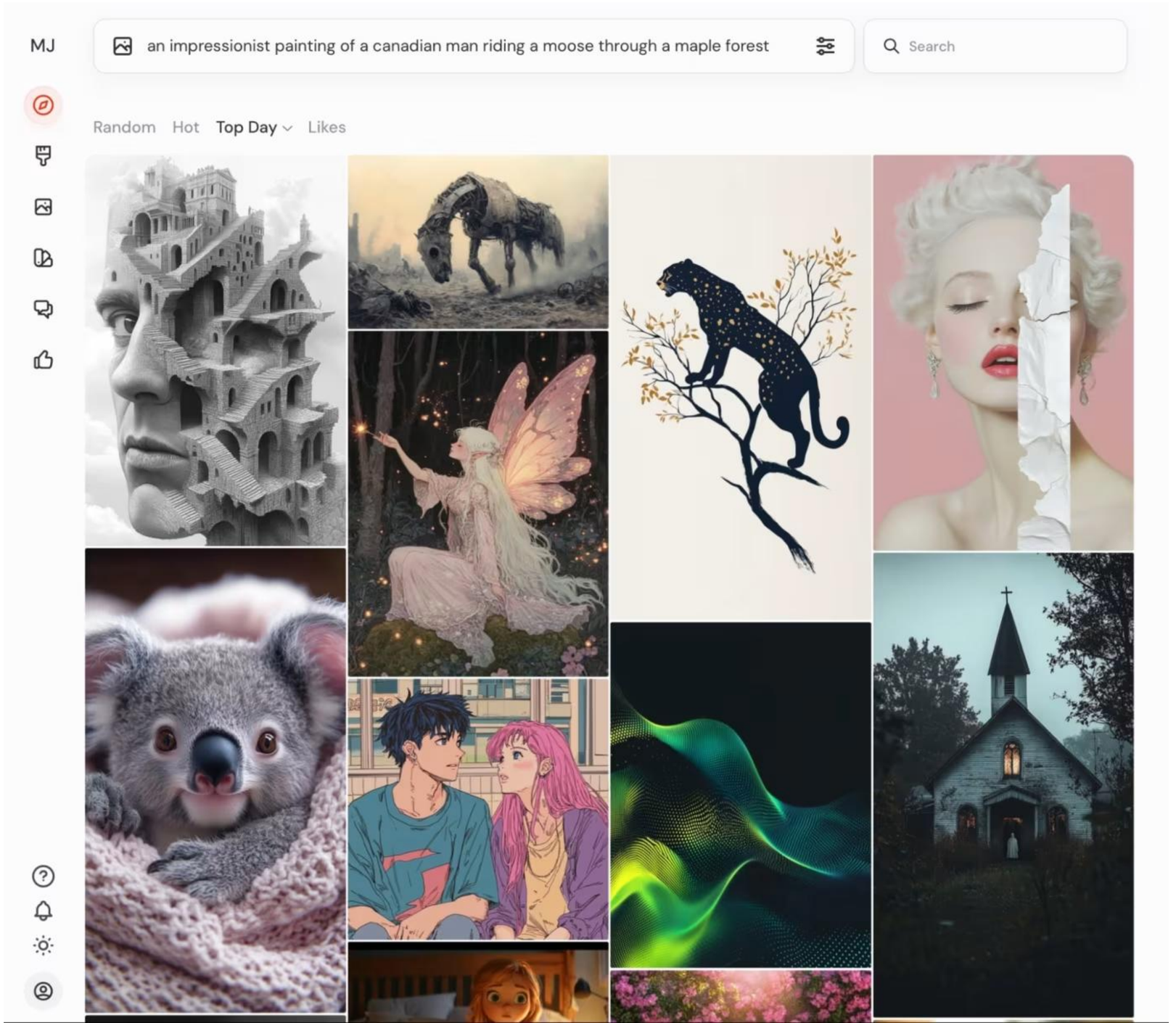
DALL-E



https://openai.com/index/dall-e-2/

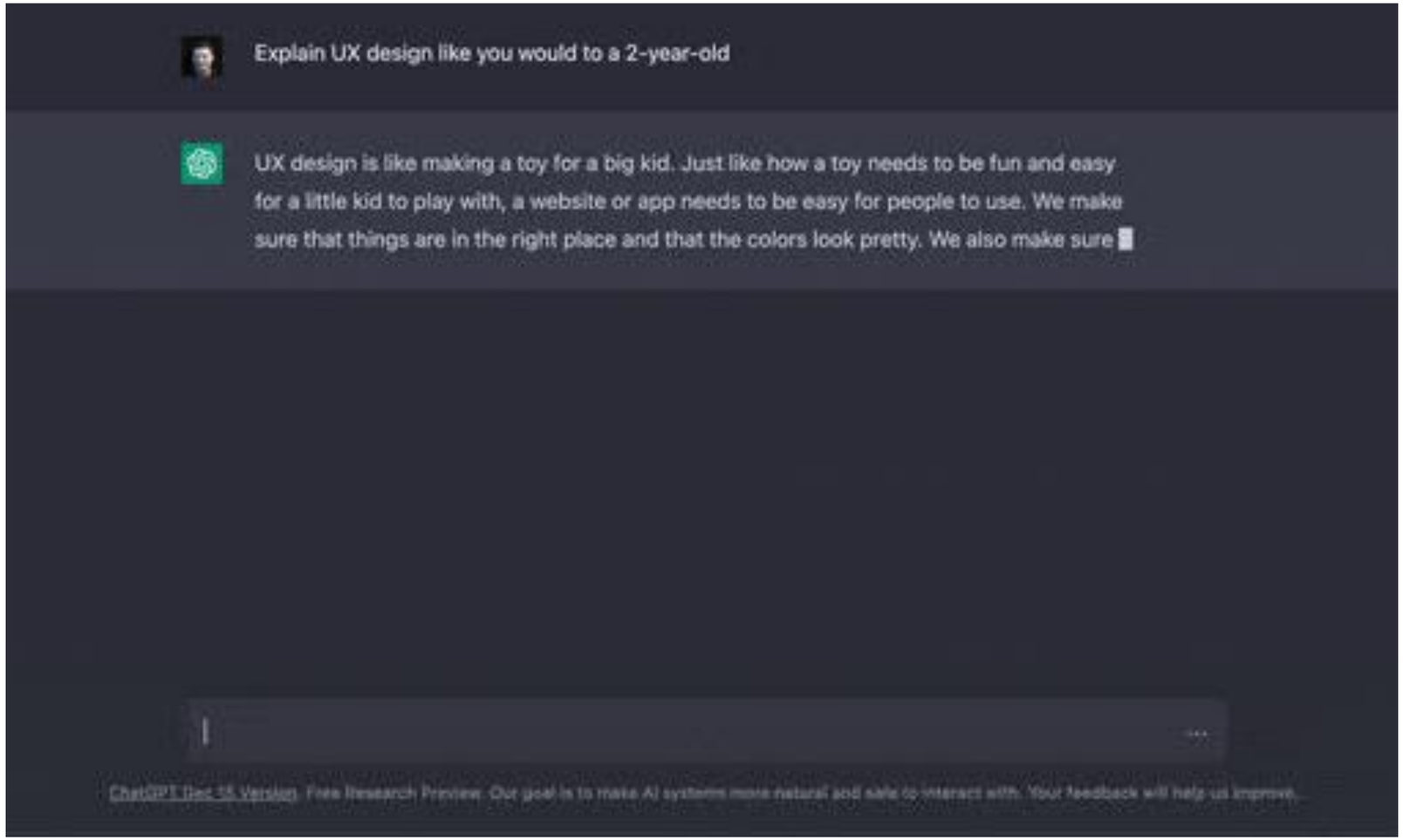
Diffusion models :

Midjourney



https://www.midjourney.com/home

ChatGPT


Explain UX design like you would to a 2-year-old
UX design is like making a toy for a big kid. Just like how a toy needs to be fun and easy for a little kid to play with, a website or app needs to be easy for people to use. We make sure that things are in the right place and that the colors look pretty. We also make sure
ChatGPT Dec 15 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.


ME THINKING ABOUT THE DAYS
WHEN I HAD TO GOOGLE EVERYTHING
GPT-4,
MY NEW
BEST FRIEND

WHAT IS NLP ?

Google Developer Groups
On Campus • SUP'COM

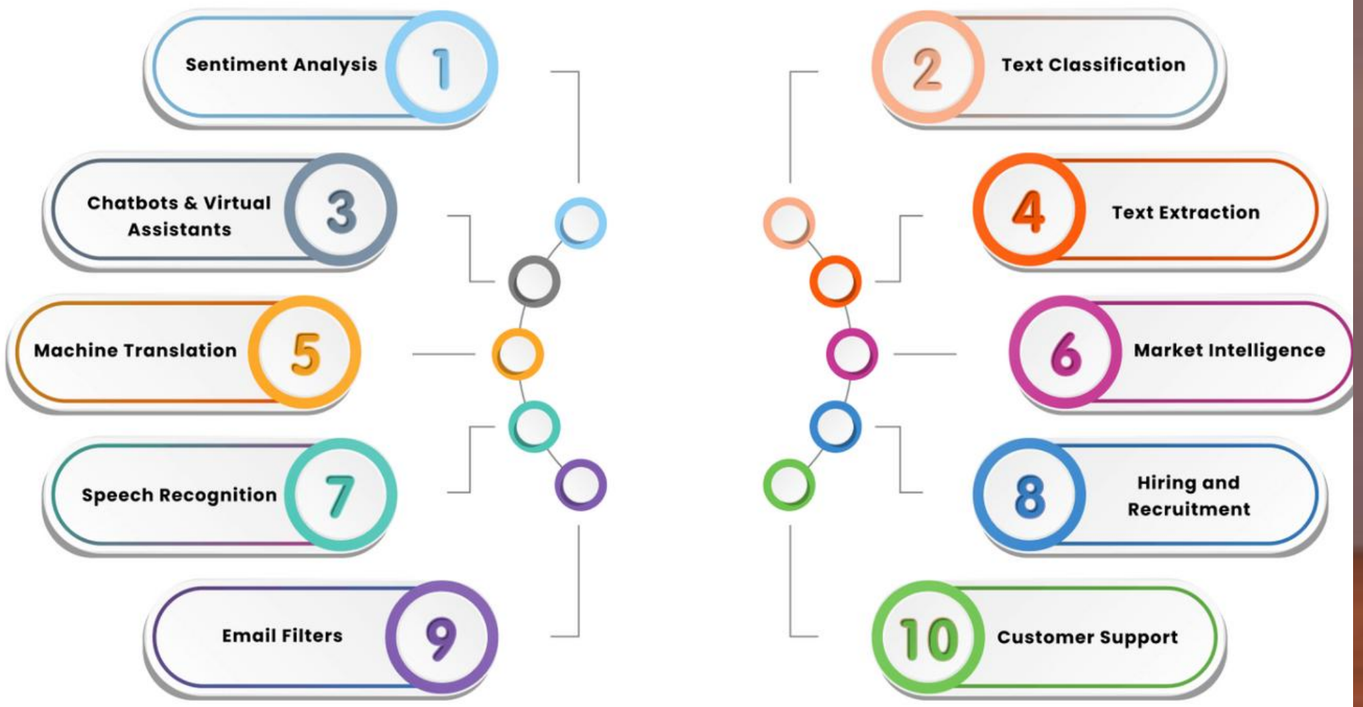
EPI Educational Group

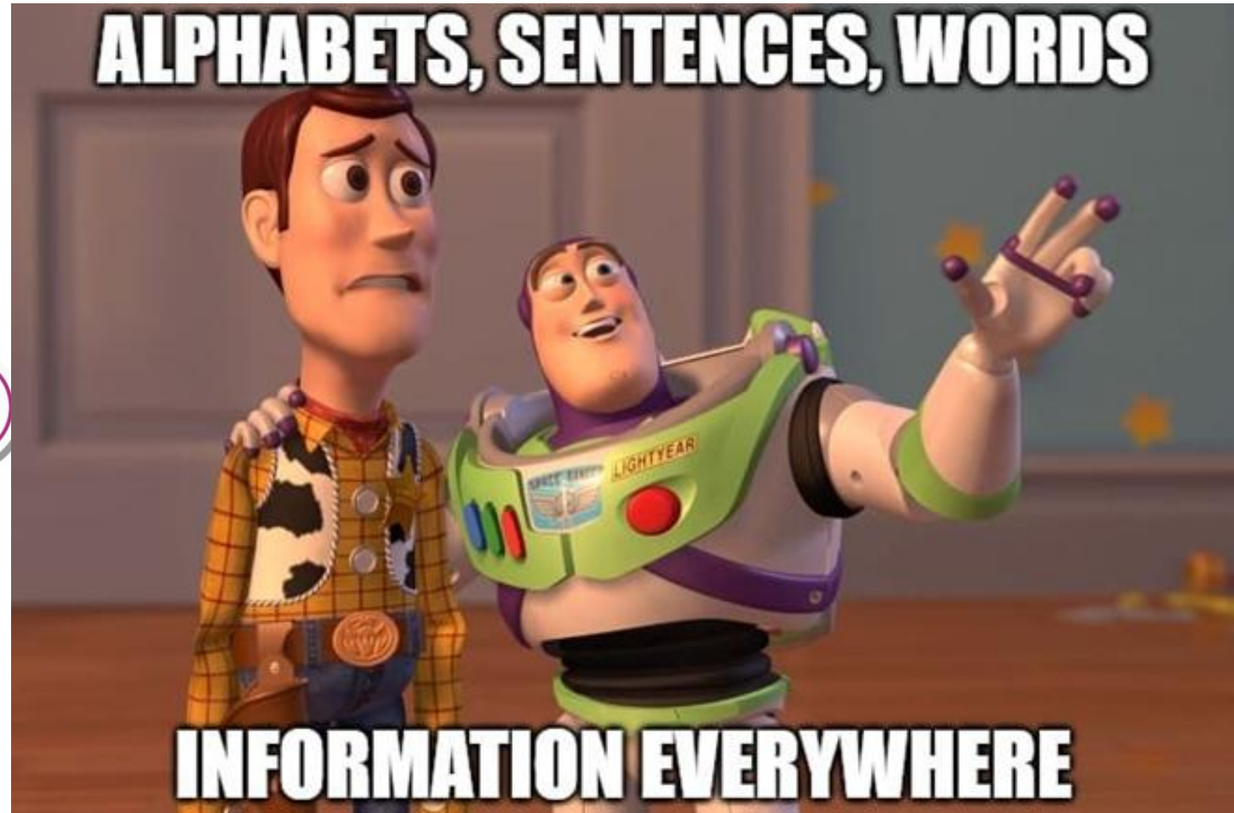
By Mohammed Arbi Nsibi

WHAT IS NLP ?

is referred to as NLP. It is a **subset** of AI that enables machines to comprehend and analyze **human languages**. Text or audio can be used to represent human languages.

NLP applications





NLP Pipeline

Data Acquisition → Text Cleaning → Text Preprocessing → Feature Engineering → Model Building → Evaluation → Deployment

GDG Carthage

By Mohammed Arbi Nsibi

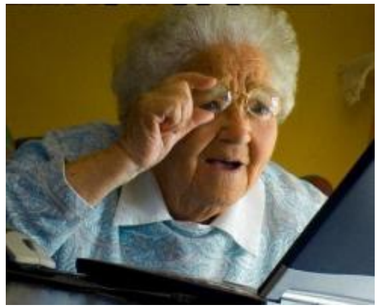
Text Cleaning

Regex or Regular Expression

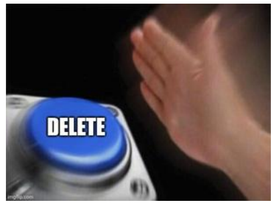
Spelling corrections

used for searching the string of specific patterns. Suppose our data contain phone number, email-Id, and URL. we can find such text using the regular expression. After that either we can keep or remove such text patterns as per requirements.

Text Preprocessing



- Lowercasing
- Stop word removal
- Stemming or lemmatization



- Removing digit/punctuation
- POS tagging (adj, noun, adv)
- Named Entity Recognition (NER) (name, location,..)

GDG Carthage

By Mohammed Arbi Nsibi

"Elon Musk founded SpaceX in 2002."

Named Entity Recognition (NER)

[Elon Musk: PERSON],
[SpaceX: ORG],
[2002: DATE].

POS tagging

Elon/NNP (Proper noun)
Musk/NNP
founded/VBD
SpaceX/NNP
in/IN
2002/CD (Cardinal number)

Feature Engineering = Text Representation = Text Vectorization.

Our main agenda is to represent the text in the numeric vector in such a way that the ML algorithm can understand the text attribute.

Traditional Approach

- One Hot Encoding

Traditional Approach

- One Hot Encoding

id	color
1	red
2	blue
3	green
4	blue

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Traditional Approach

- TF-IDF (Term Frequency – Inverse Document Frequency) 1972

  Give more weight to rare words and less to common terms

$$TF(t,d) = \frac{(Number\ of\ occurrences\ of\ term\ t\ in\ document\ d)}{(Total\ number\ of\ terms\ in\ the\ document\ d)}$$

$$IDF(t,D) = \log_e \frac{(Total\ number\ of\ documents\ in\ the\ corpus)}{(Number\ of\ documents\ with\ term\ t\ in\ them)}$$

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

Traditional Approach

- Neural Approach (Word embedding)

car =[0.8, 0.9, 0.9, 0.01, 0.75]
bike =[0.8, 0.7, 0.2, 0.01, 0.5]

not interpretable for humans

try to incorporate the contextual meaning of the words.

Word	car	bike
Road	0.8	0.8
Speed	0.9	0.7
Fuel	0.9	0.2
Animal	0.01	0.01
Price	0.75	0.5

How can we get these word embedding vectors?

How can we get these word embedding vectors?

Train our own embedding layer:

- **CBOW (Continuous Bag of Words)**
- **SkipGram**

Pre-Trained Word Embeddings

- **Word2vec by Google 2013**
- **GloVe by Stanford**
- **fasttext by Facebook**

Visualizing High-Dimensional Space:
https://www.youtube.com/watch?v=wvsE8jm1GzE&ab_channel=GoogleforDevelopers

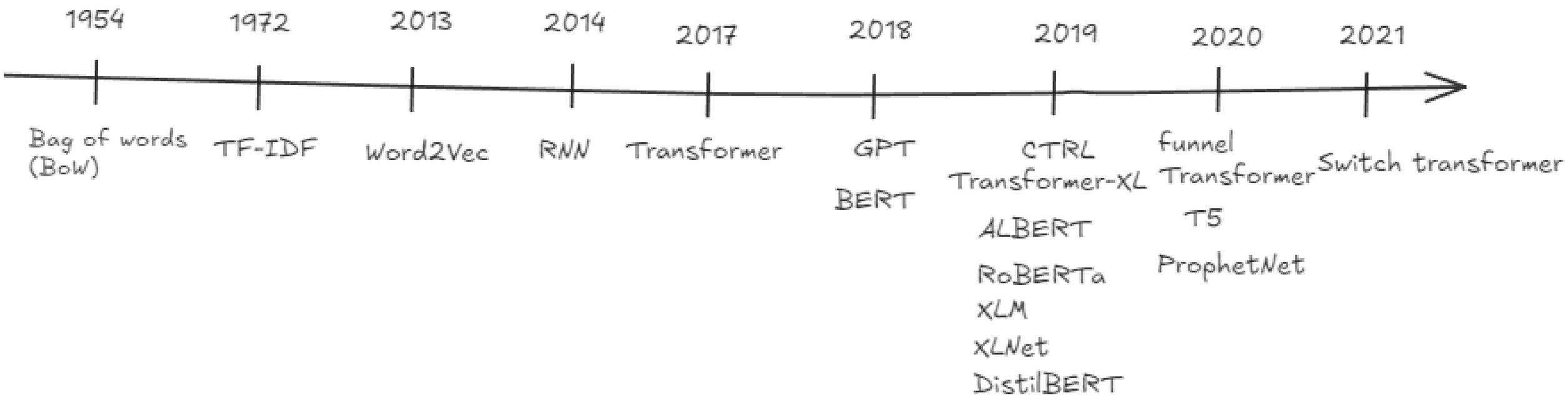
By Mohammed Arbi Nsibi

LET'S CODE

NLP vs LLM

GDG Carthage

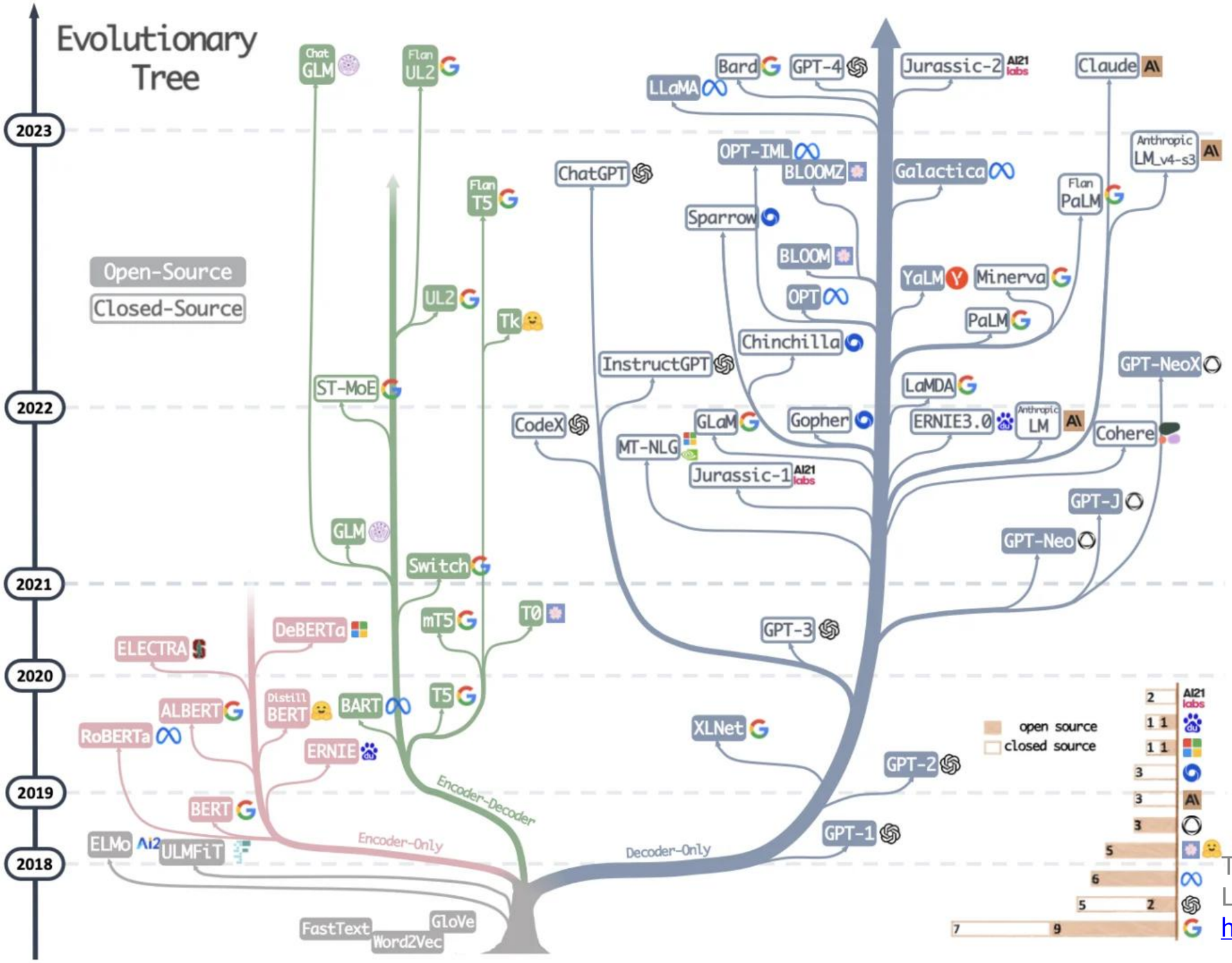
By Mohammed Arbi Nsibi



Google Developer Groups
On Campus • SUP'COM

EPI Educational Group

By Mohammed Arbi Nsibi



The evolutionary tree of modern LLMs via https://arxiv.org/abs/2304.13712.

GPT-1

(June 2018)

6 years: What has changed?

Llama **3.2**

(September 2024)

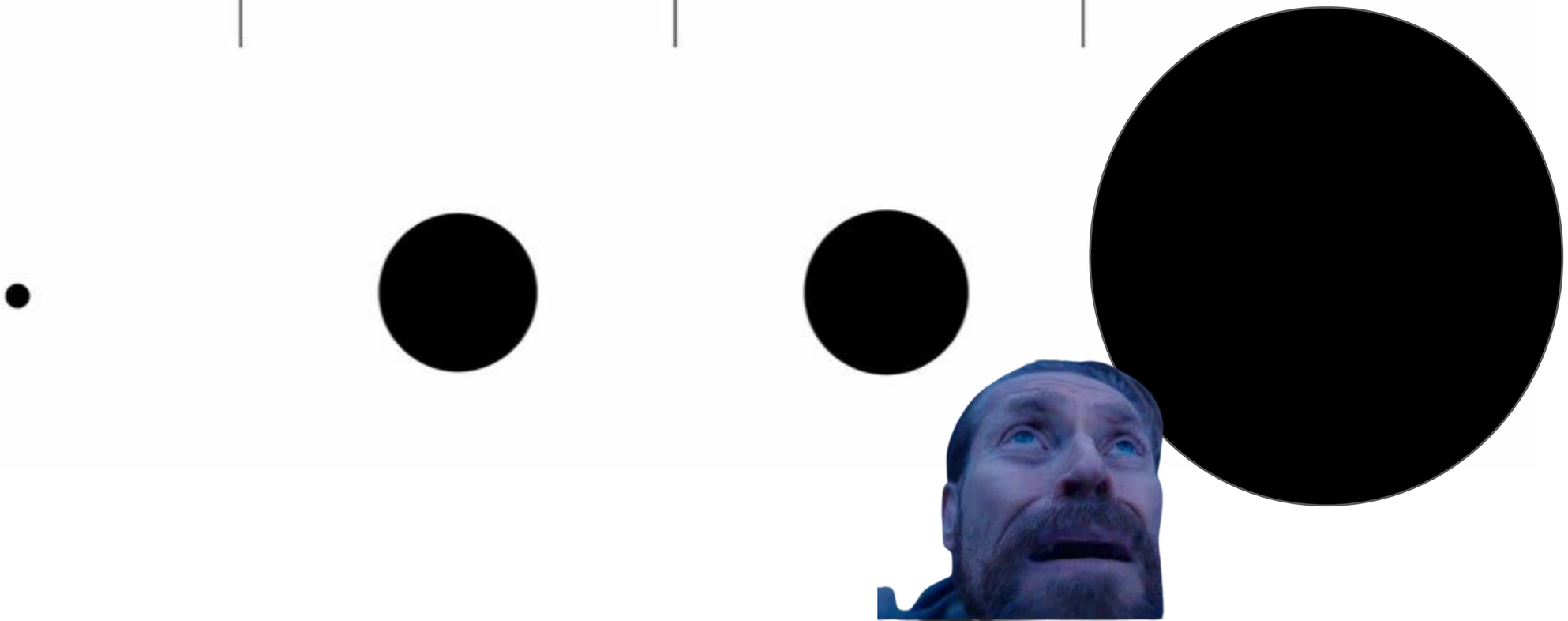
GDG Carthage

Model size

2019	2023	2023
GPT-2	Llama-1	Llama-2
124M to 1.5B	7B to 65B	7B to 70B

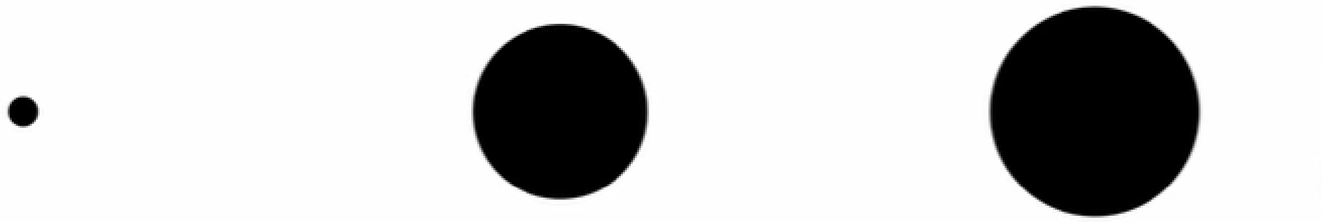
Model size

2019	2023	2023	2024
GPT-2	Llama-1	Llama-2	Llama-3
124M to 1.5B	7B to 65B	7B to 70B	8B to 405B



Dataset

2019	2023	2023
GPT-2	Llama-1	Llama-2
40B tokens	1.4T tokens	2T tokens



Dataset

2019	2023	2023	2024
GPT-2	Llama-1	Llama-2	Llama-3
40B tokens	1.4T tokens	2T tokens	15T tokens

GDG Carthage

"Attention is All You Need"

Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*] [†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

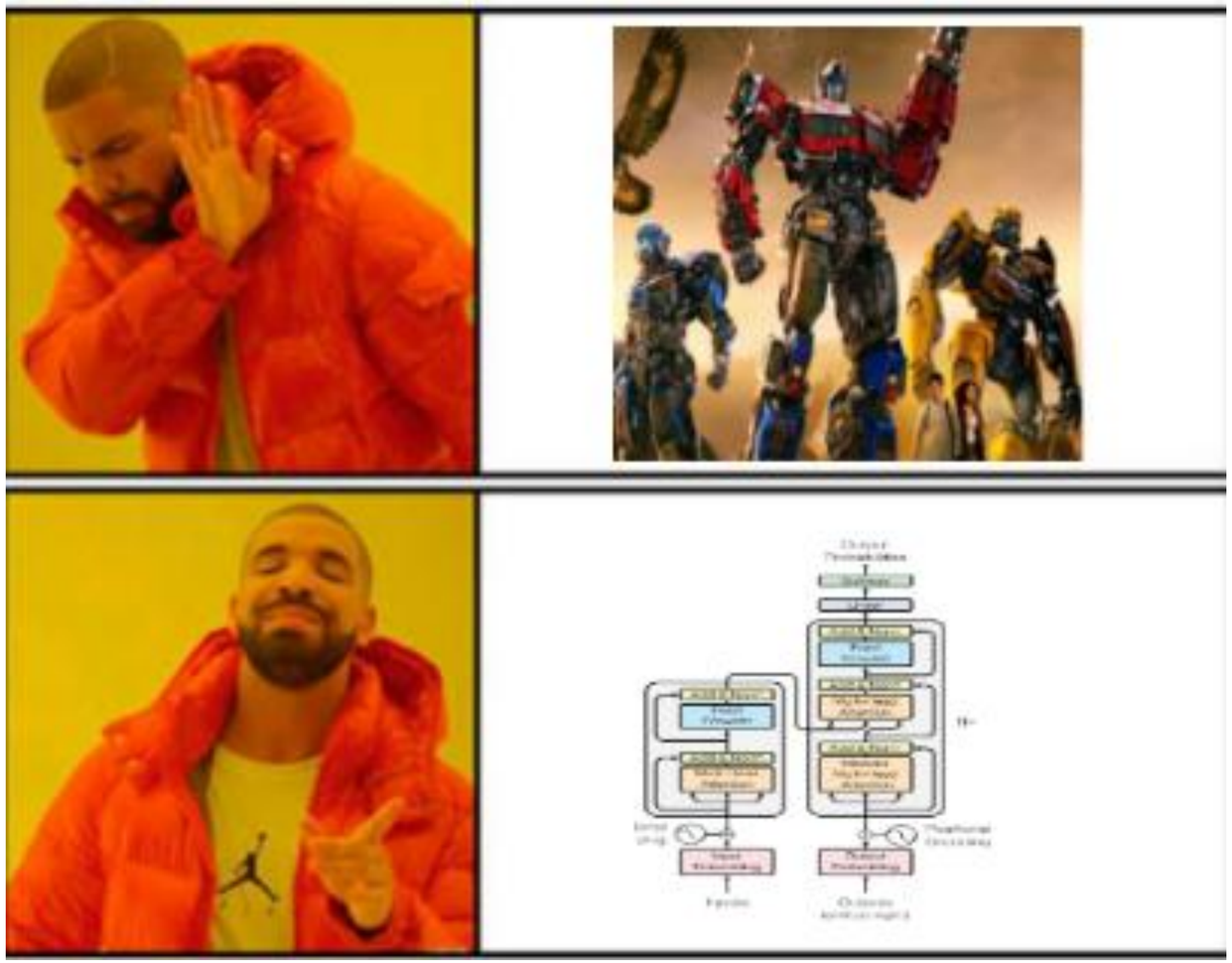
Abstract

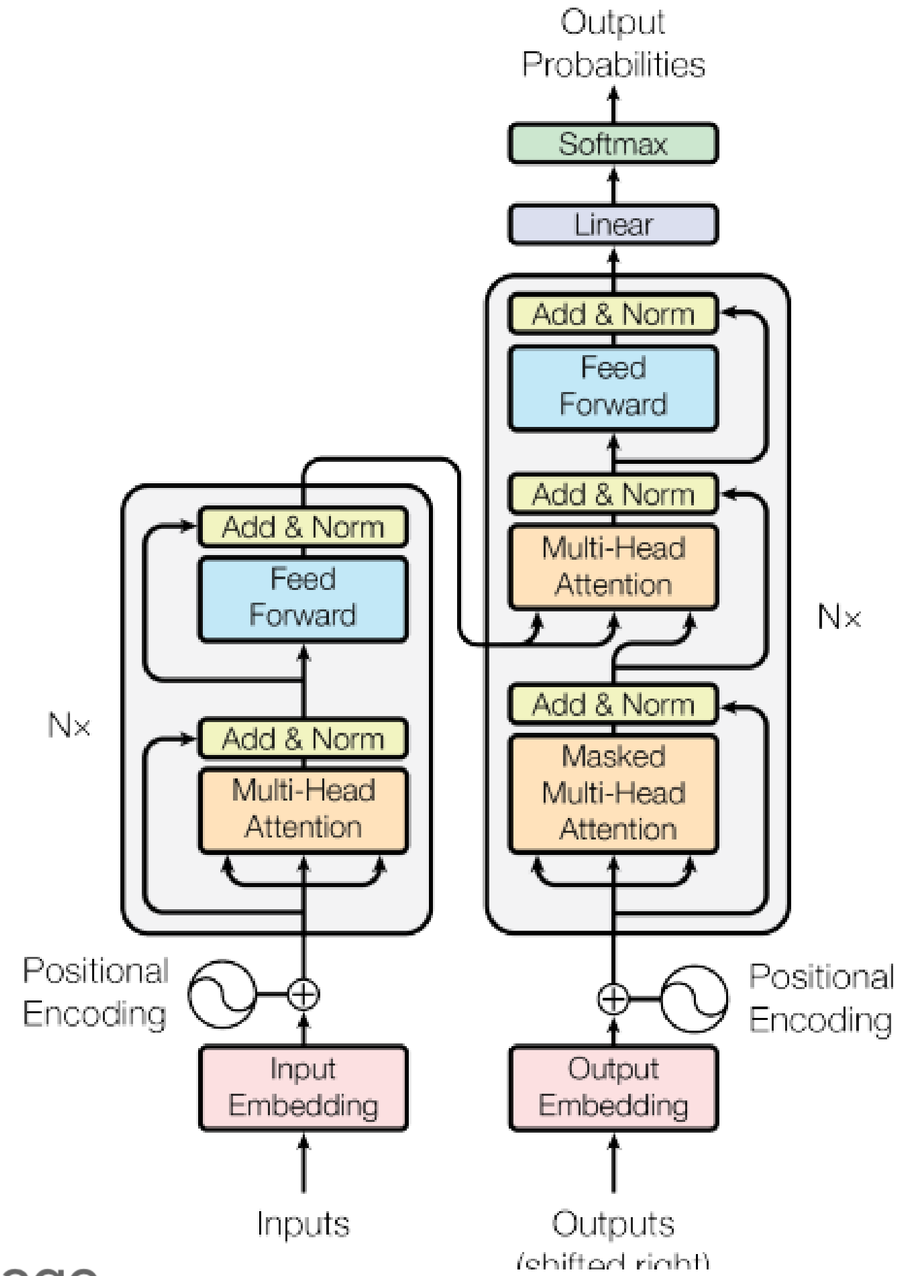
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

GDG Carthage

By Mohammed Arbi Nsibi

Transformers





GDG Carthage

Source: https://arxiv.org/abs/1706.03762

X_1 for "I"

X_2 for "love"

X_3 for "NLP"

$$PE_{(pos\ 2i)} = sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

$$PE_{(pos\ 2i+1)} = cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \end{bmatrix}$$

Position	PE (dim 0)	PE (dim 1)	PE (dim 2)	PE (dim 3)
0	sin(0) = 0	cos(0) = 1	sin(0) = 0	cos(0) = 1
1	sin(1/10000^0) ≈ 0.8415	cos(1/10000^0) ≈ 0.5403	sin(1/10000^0) ≈ 0.8415	cos(1/10000^0) ≈ 0.5403
2	sin(2/10000^0) ≈ 0.9093	cos(2/10000^0) ≈ -0.4161	sin(2/10000^0) ≈ 0.9093	cos(2/10000^0) ≈ -0.4161

$$E = X + PE$$

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

الوضع ريلاكس بلا صياح

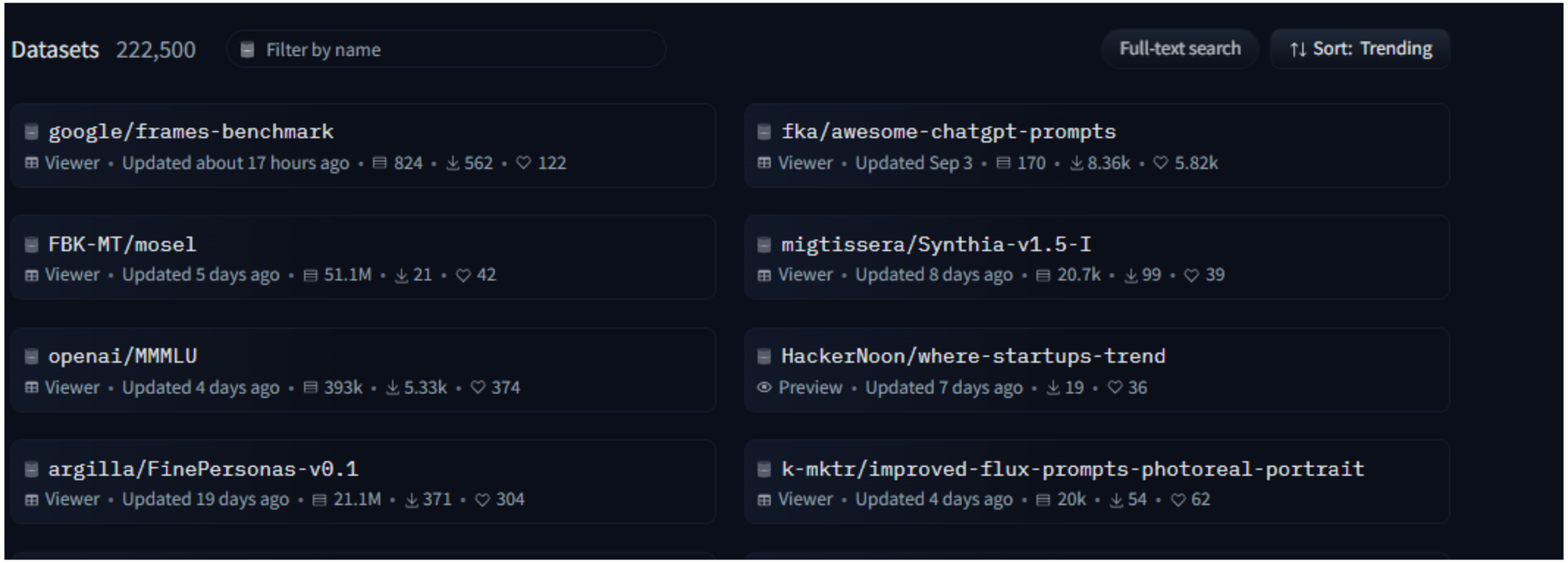
Where to find pretrained LLMs ?

Hugging Face

GDG Carthage

By Mohammed Arbi Nsibi

Where to find pretrained LLMs ?



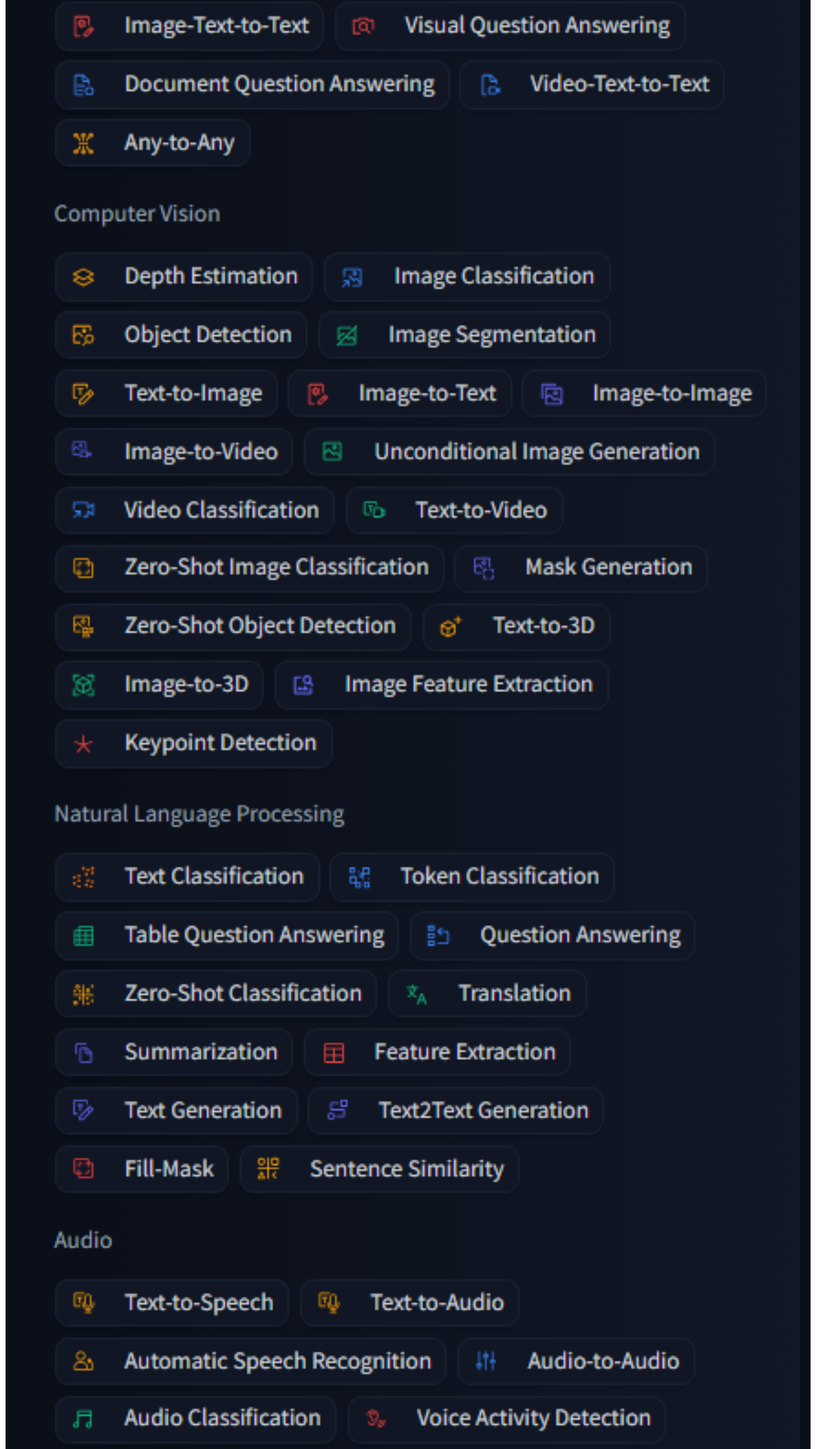
Model types

Text-to-text

Text-to-image

Text-to-video

Sentence-similarity



How to create my own application

```
import gradio as gr
from transformers import pipeline

pipe = pipeline("sentiment-analysis")

def predict(new_input):
  out = pipe(new_input)
  out = out[0]["label"]
  return out

gr.Interface(predict,inputs=["text"],outputs=["text"]).launch()
```

https://huggingface.co/learn/nlp-course/chapter1/3#working-with-pipelines

LET'S CODE

By Mohammed Arbi Nsibi

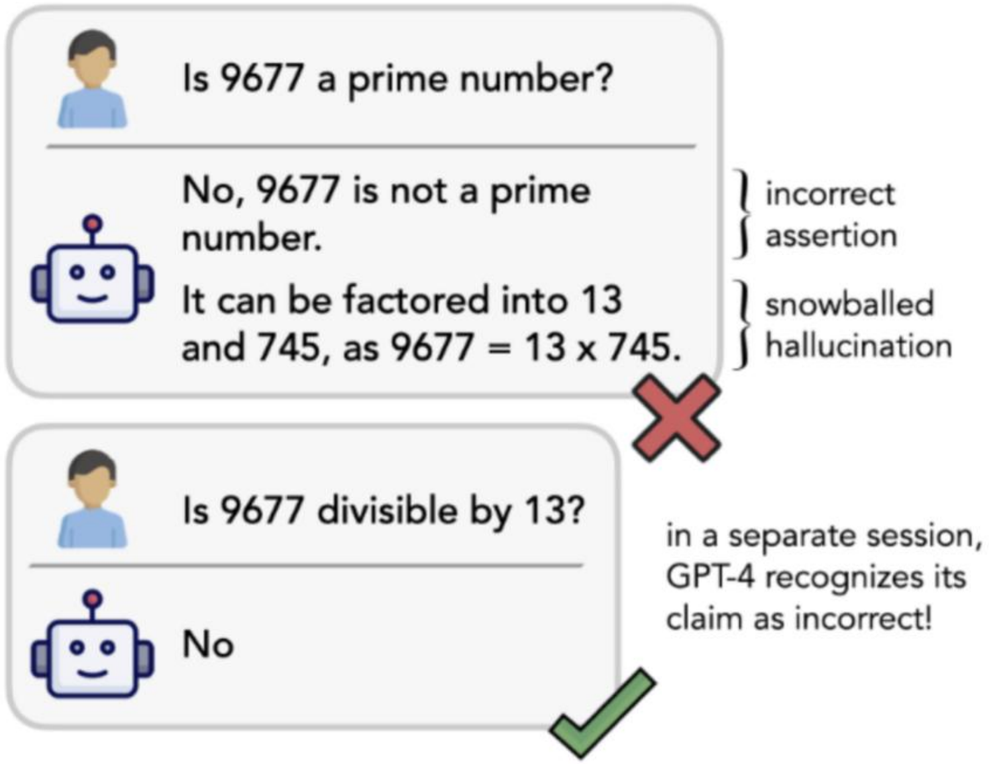
Hallucinations

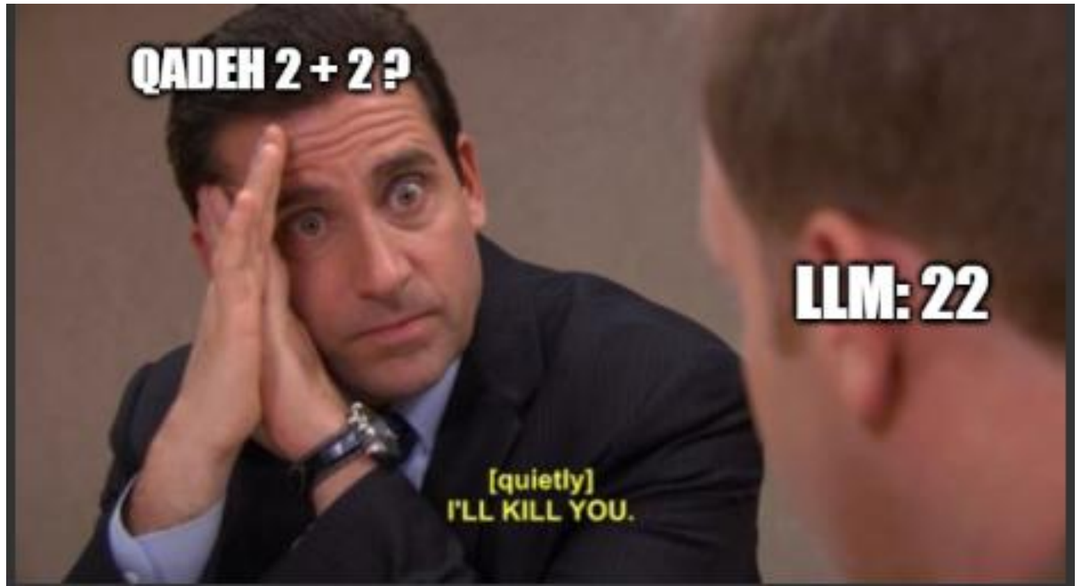
- The model is not trained on enough data.
- The model is trained on noisy or dirty data.
- The model is not given enough context .
- The model is not given enough constraints.

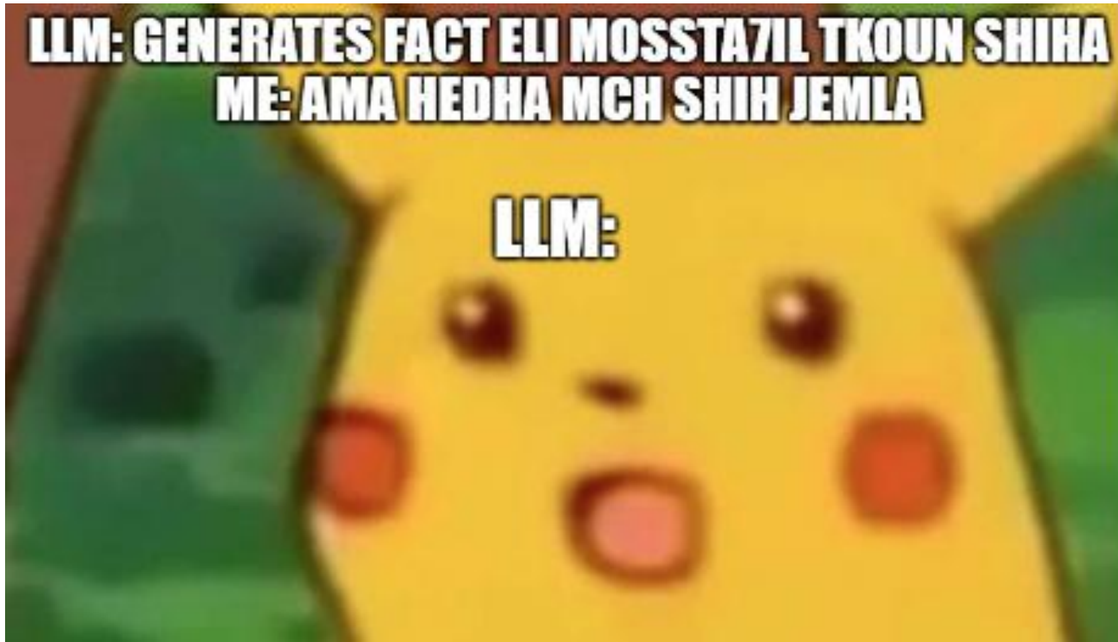
GDG Carthage

By Mohammed Arbi Nsibi

Hallucinations





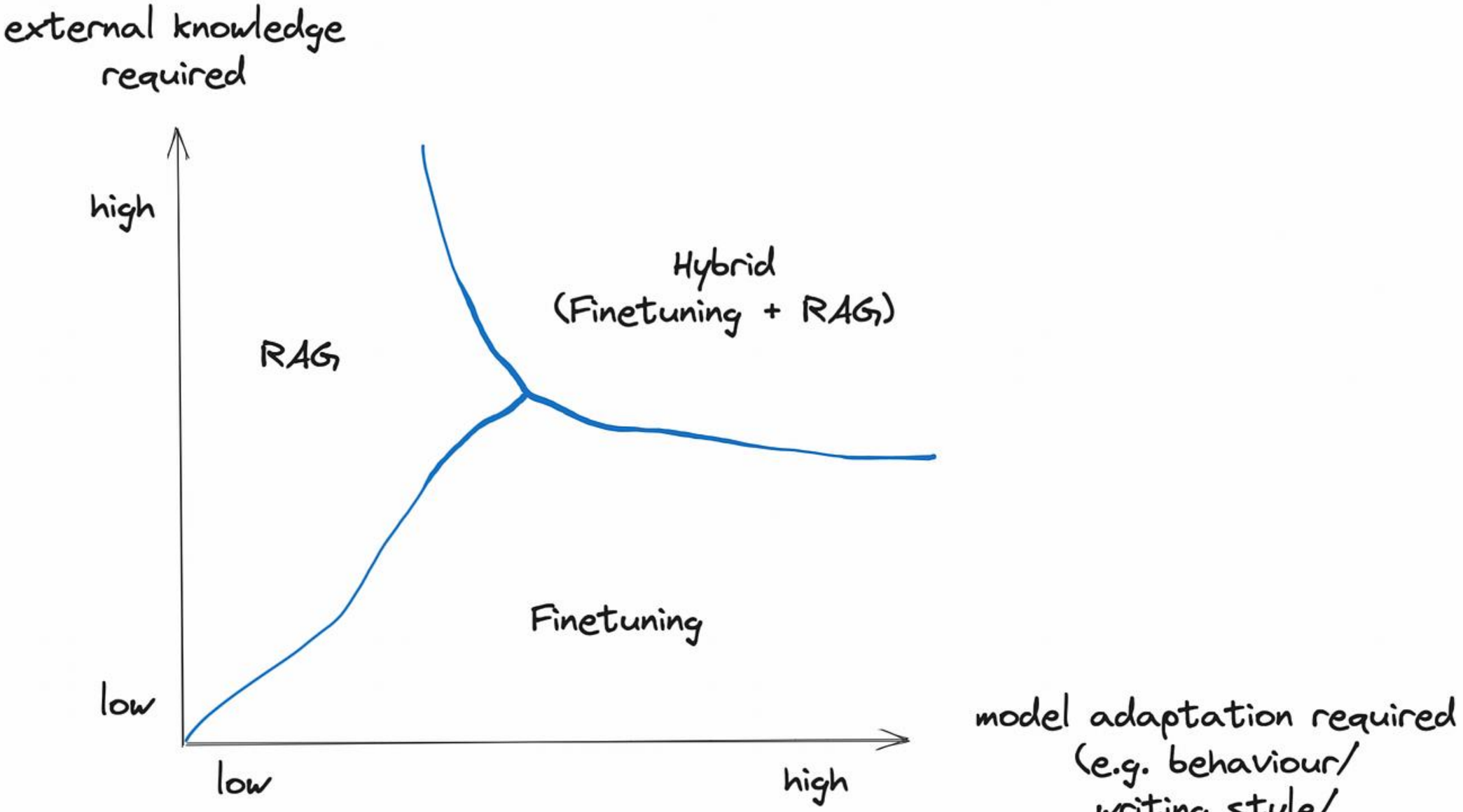


Solutions ?

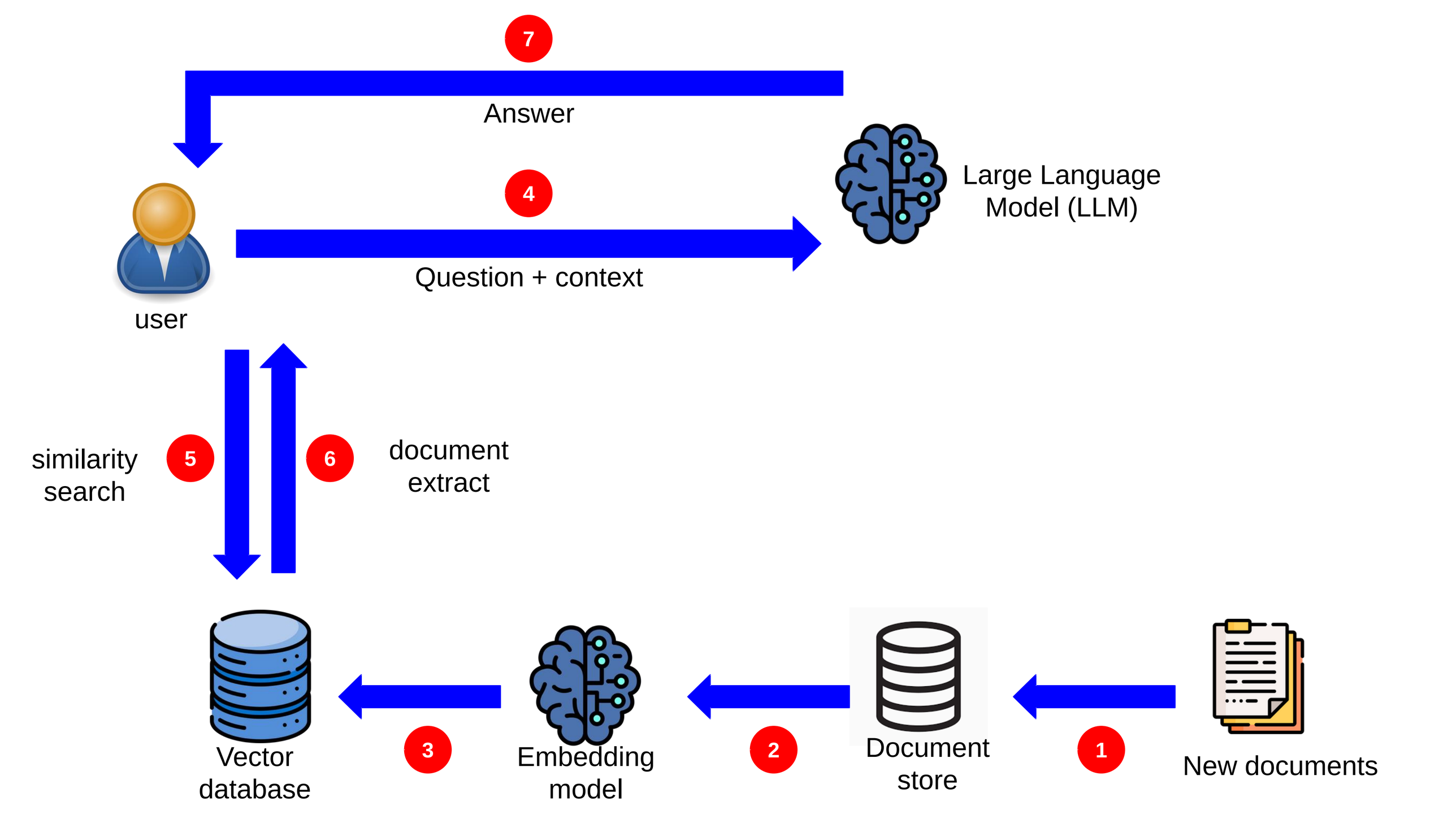
GDG Carthage

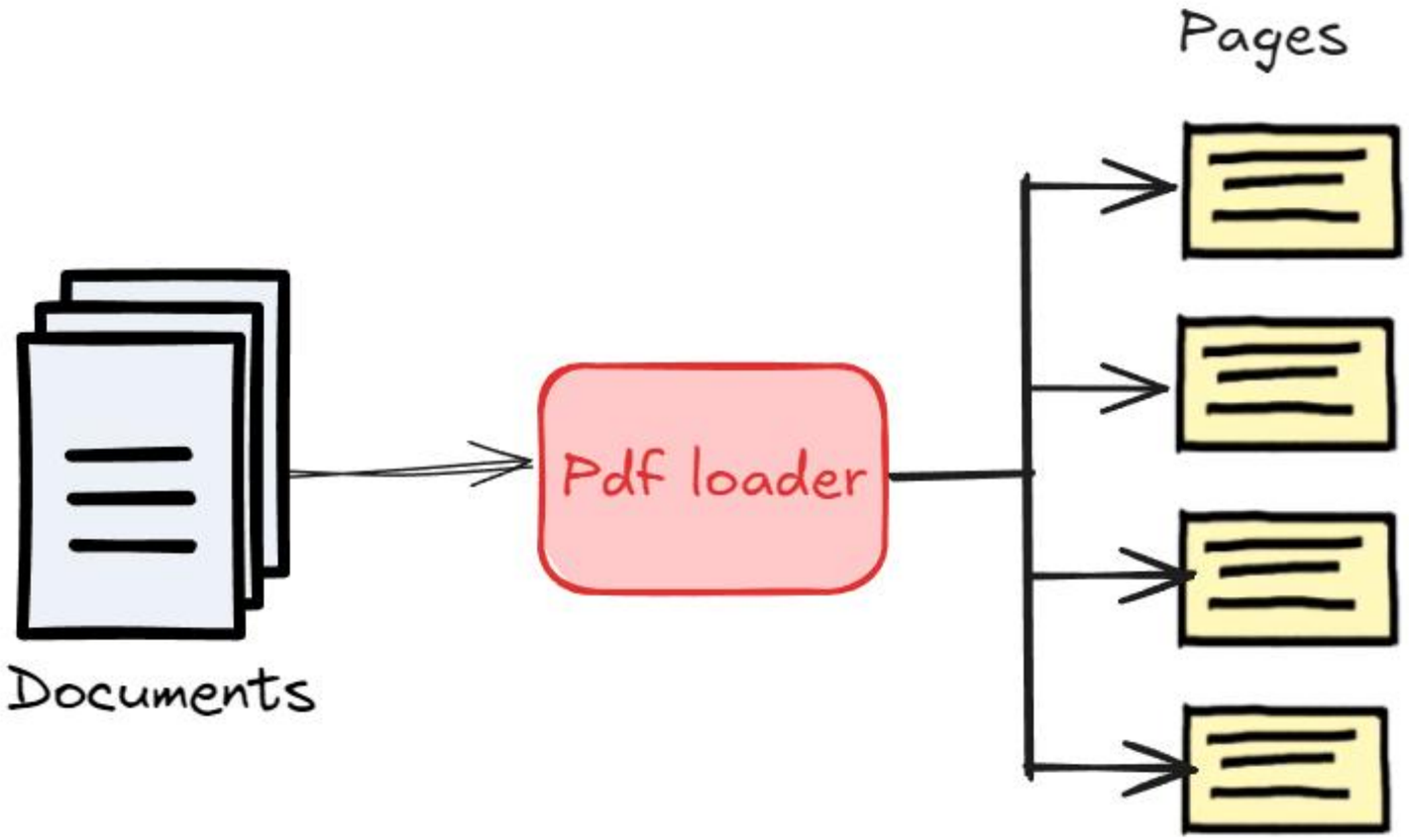
By Mohammed Arbi Nsibi

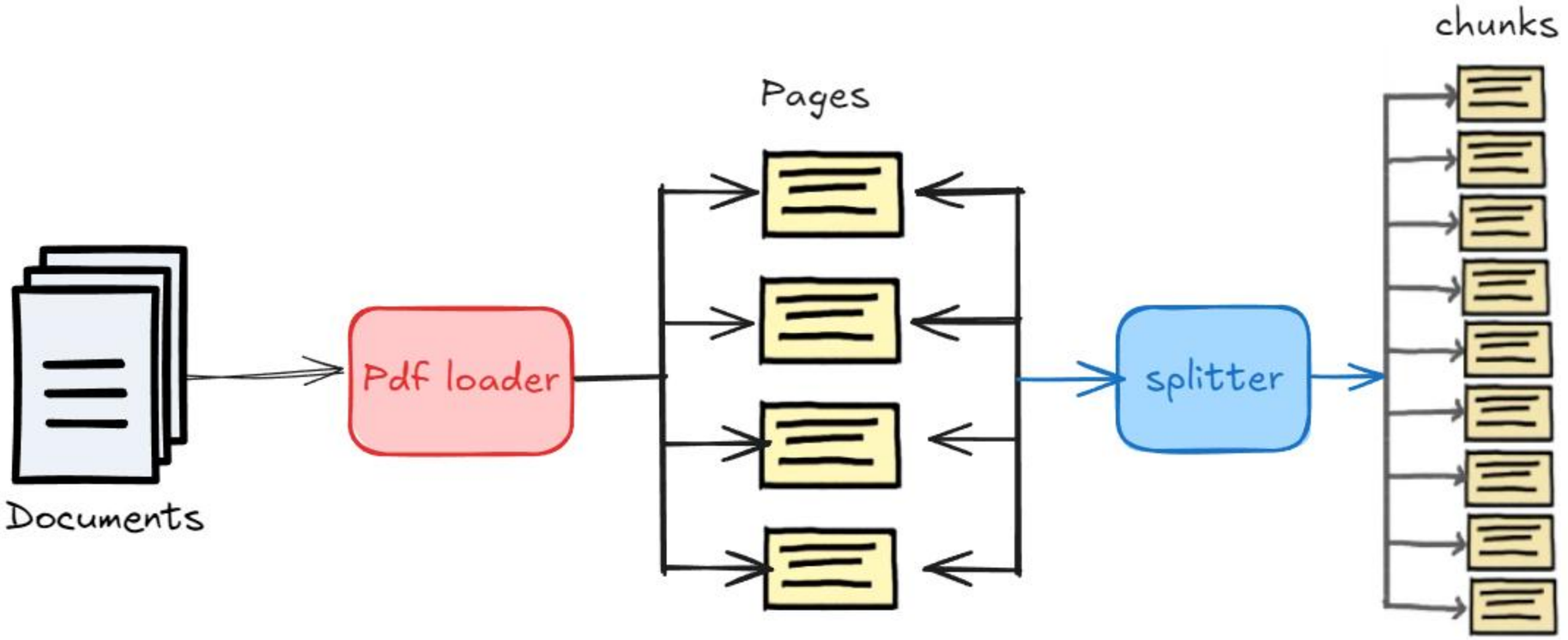
RAG / Fine-tuning



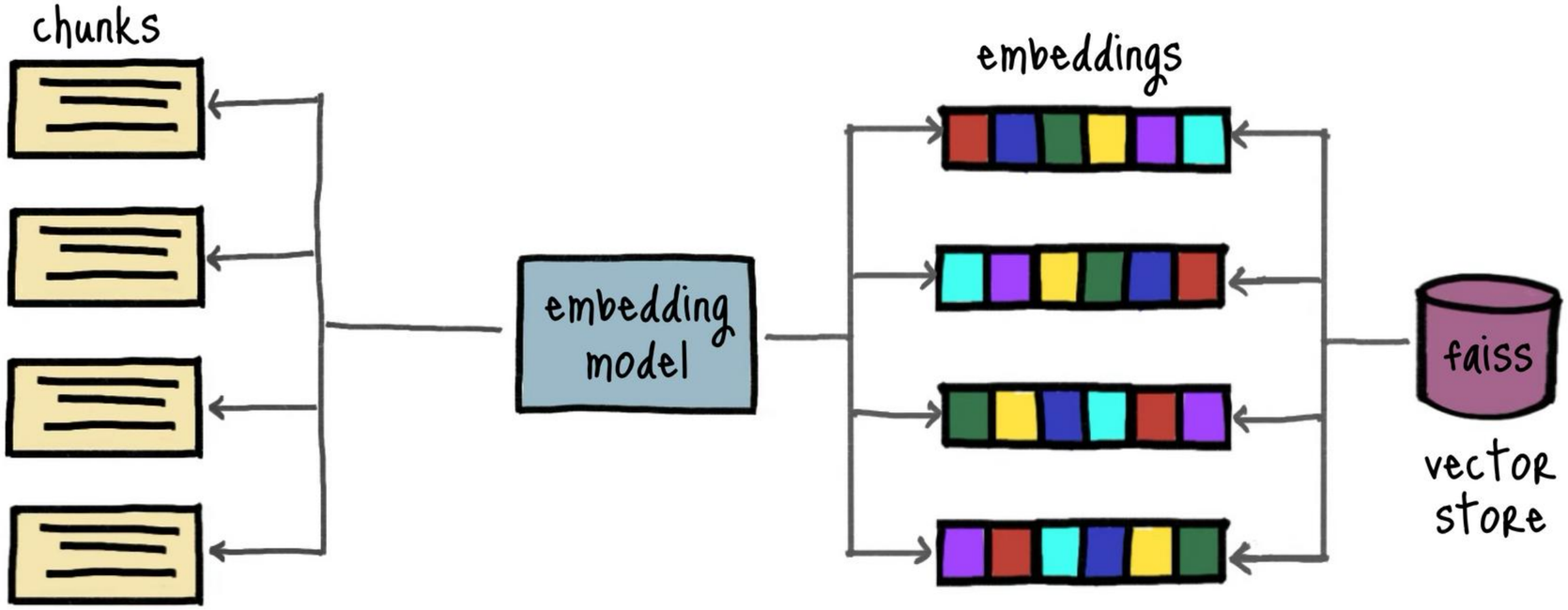
But wait..
WTF is RAG ?

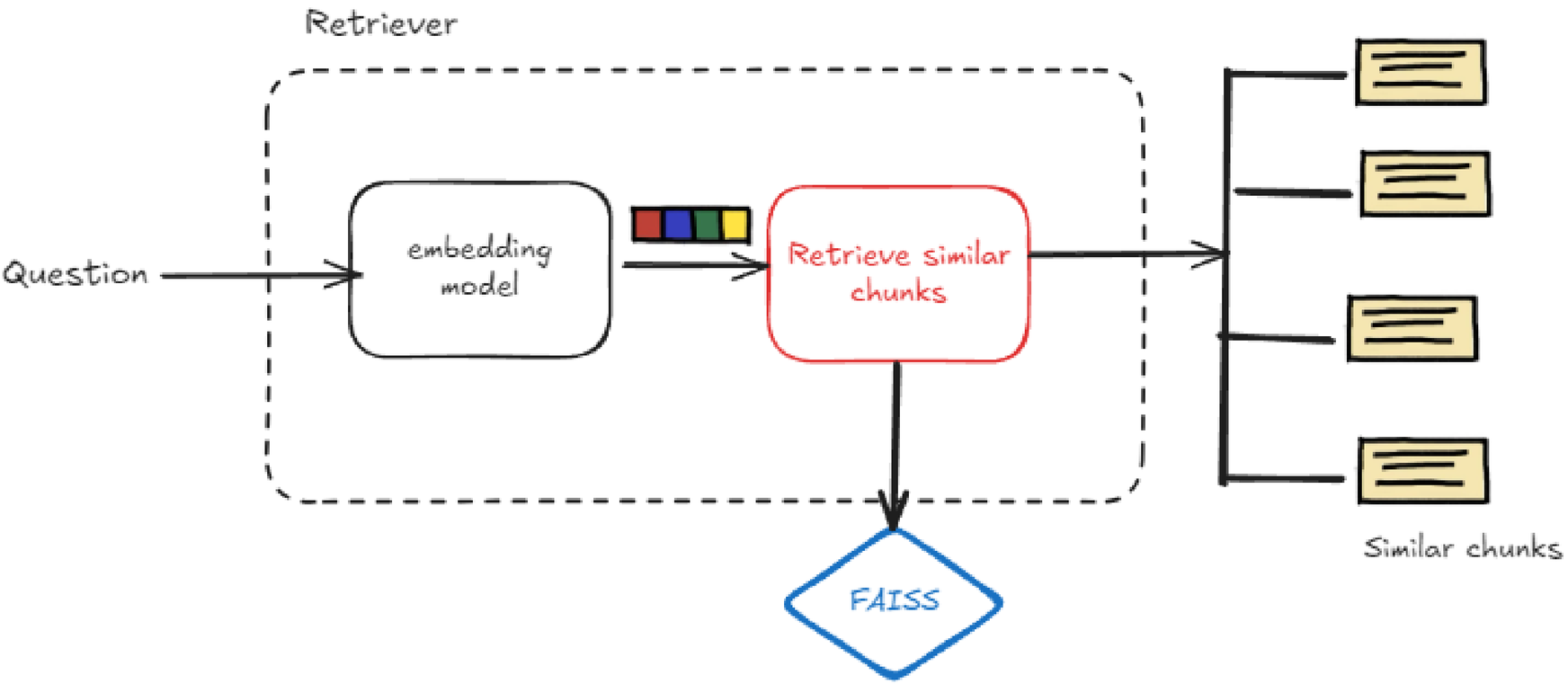

7
Answer
4
Question + context
Large Language Model (LLM)
user
similarity search
5
6
document extract
Vector database
3
Embedding model
2
Document store
1
New documents


Pages
Pdf loader
Documents

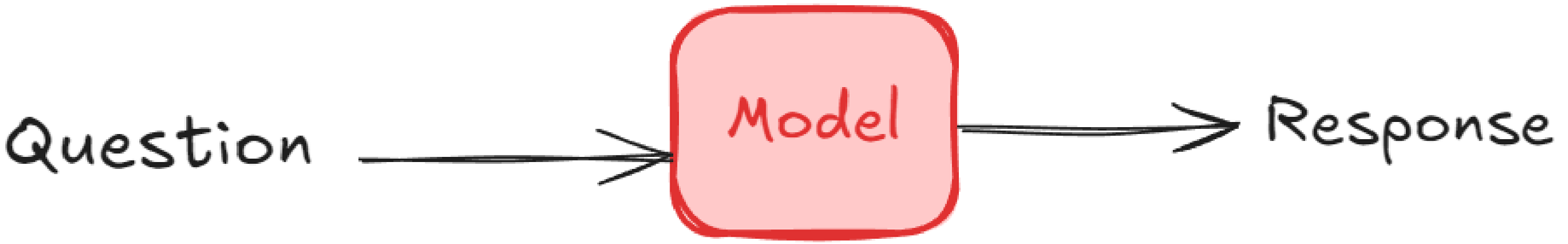


By Mohammed Arbi Nsibi

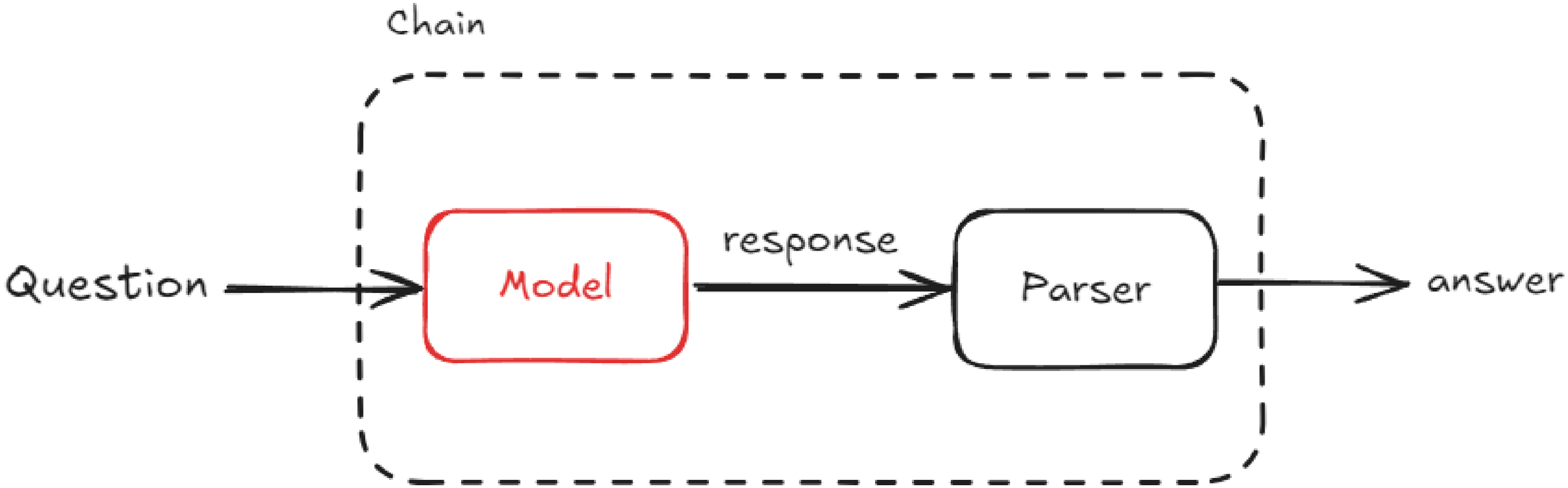

chunks
embedding model
embeddings
faiss
vector store



By Mohammed Arbi Nsibi

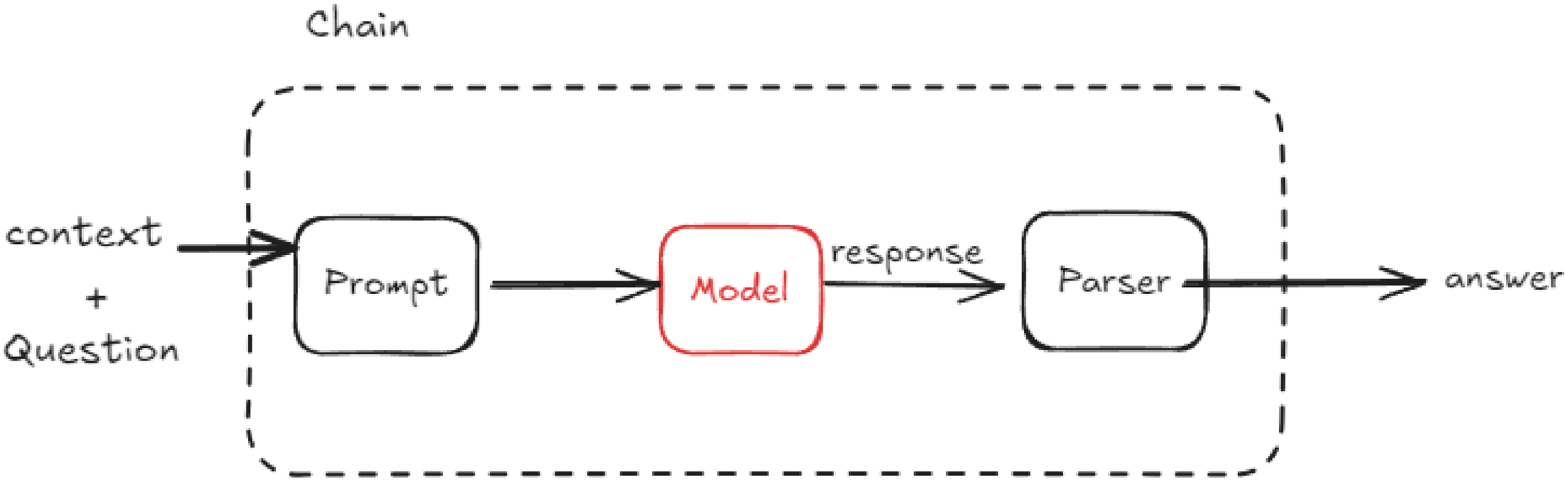

Question
Model
Response

```
AIMessage(content='As of my last update in April 2023, Joe Biden is the President of the United States. He took office on January 20, 2021,
```

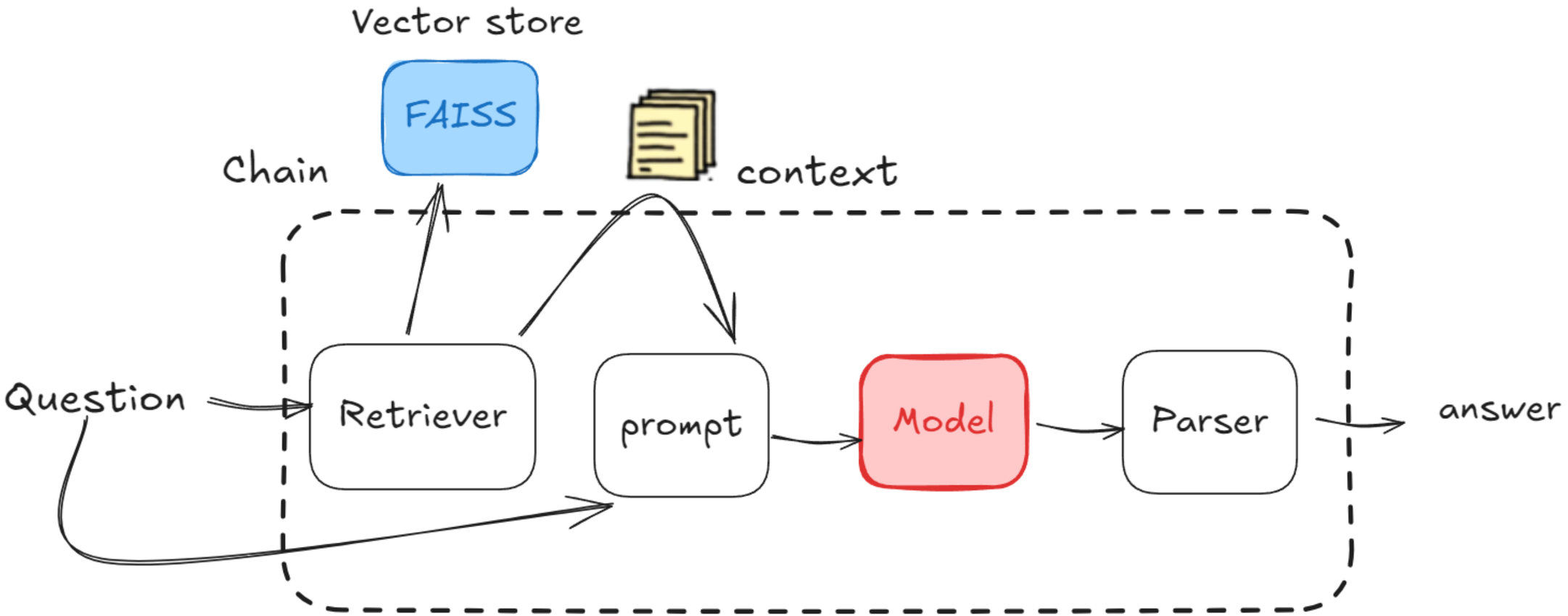



```
'As of my last update in April 2023, Joe Biden is the President of the United States. He took office on January 20, 2021,
```


context
+
Question
Prompt template
Prompt



By Mohammed Arbi Nsibi



Google Developer Groups
On Campus • SUP'COM

EPI Educational Group

By Mohammed Arbi Nsibi

RAG (Retrieval-augmented generation)

By Mohammed Arbi Nsibi


always has been
wait so creating chatbots is that easy ?


THEY SAID U CREATED A CHATBOT?
I DID
WHAT DID IT COST ?
NOTHING, IT'S FREE

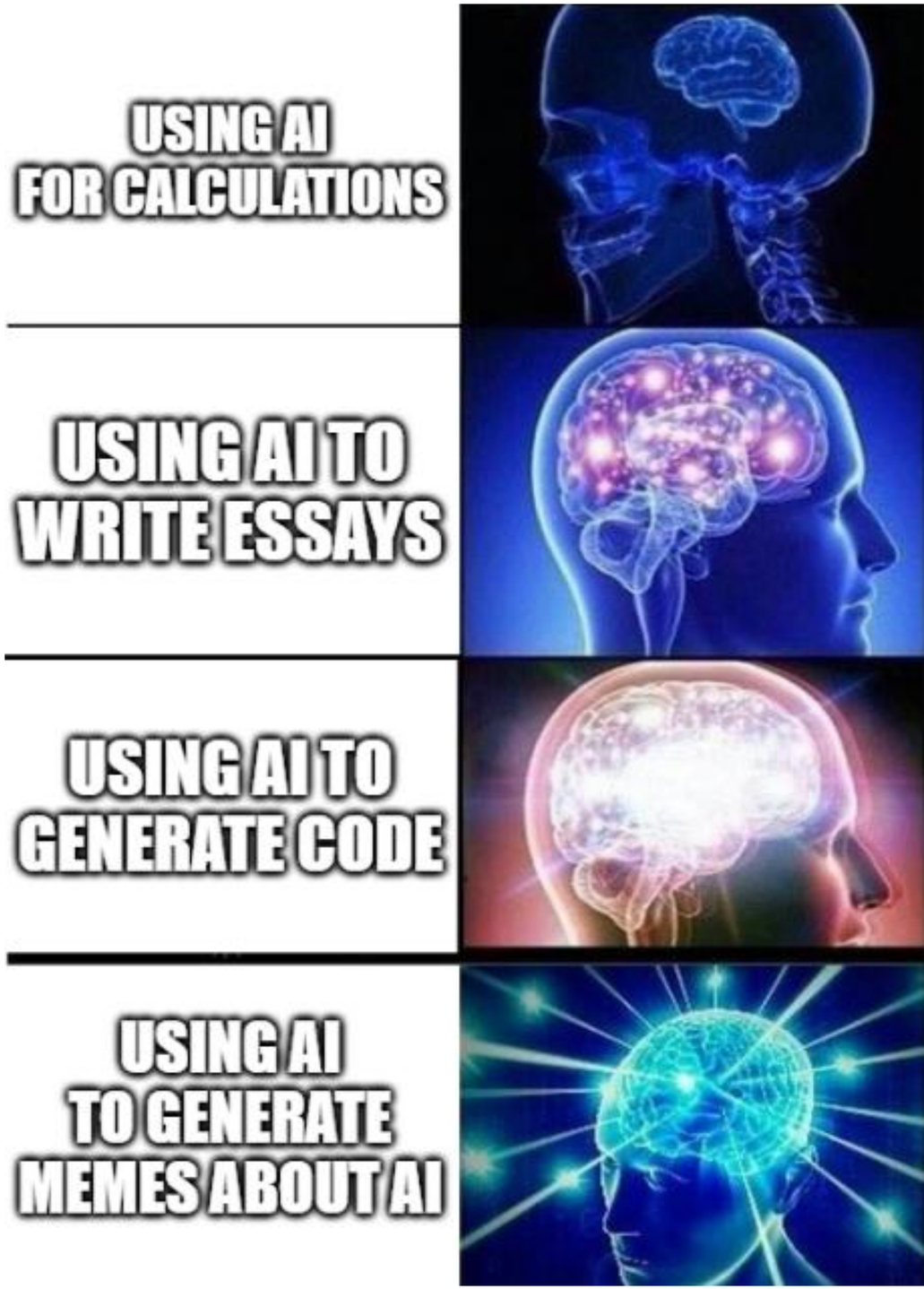
Q&A

GDG Carthage

By Mohammed Arbi Nsibi

GDG Carthage

THANK YOU for you attention!!



https://www.linkedin.com/in/mohammed-arbi-nsibi-584a43241/

QUIZ TIME

GDG Carthage

By Mohammed Arbi Nsibi

Cost to pretrain a 685B parameter LLM
(not including failed runs, hyperparameter tuning, or personnel costs)

Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours	2664K	119K	5K	2788K
in USD	$5.328M	$0.238M	$0.01M	$5.576M

Table 1 | Training costs of DeepSeek-V3, assuming the rental price of H800 is $2 per GPU hour.

$5 million!

https://github.com/deepseek-ai/DeepSeek-V3/blob/main/DeepSeek_V3.pdf

It makes me appreciate all these openly available models!

Math:
- The total number of GPU hours needed is 184,320 hours.
- The cost of running one A100 instance per hour is approximately $33.
- Each instance has 8 A100 GPUs.

That's 184320 / 8 * 33 = $760,000