From Words to Wonders: An Introduction to NLP and LLMs

16/10/2024

GDG Carthage

By Mohammed Arbi Nsibi

GDG Carthage

MOHAMED ARBI NSIBI

- Final year ICT engineering student@ SUP'COM
- GDG Carthage member
- Former lead of GDSG SUP'COM 23/24

mohammedarbinsibi@gmail.com

https://huggingface.co/Goodnight7

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

Content

- Intro to AI & Gen AI
- What is NLP
  - Pipeline
  - Let’s code
- LLMs
  - Transformers
  - Let’s code
- QUIZ

GDG Carthage

By Mohammed Arbi Nsibi

WHAT IS Generative AI ?

GDG Carthage

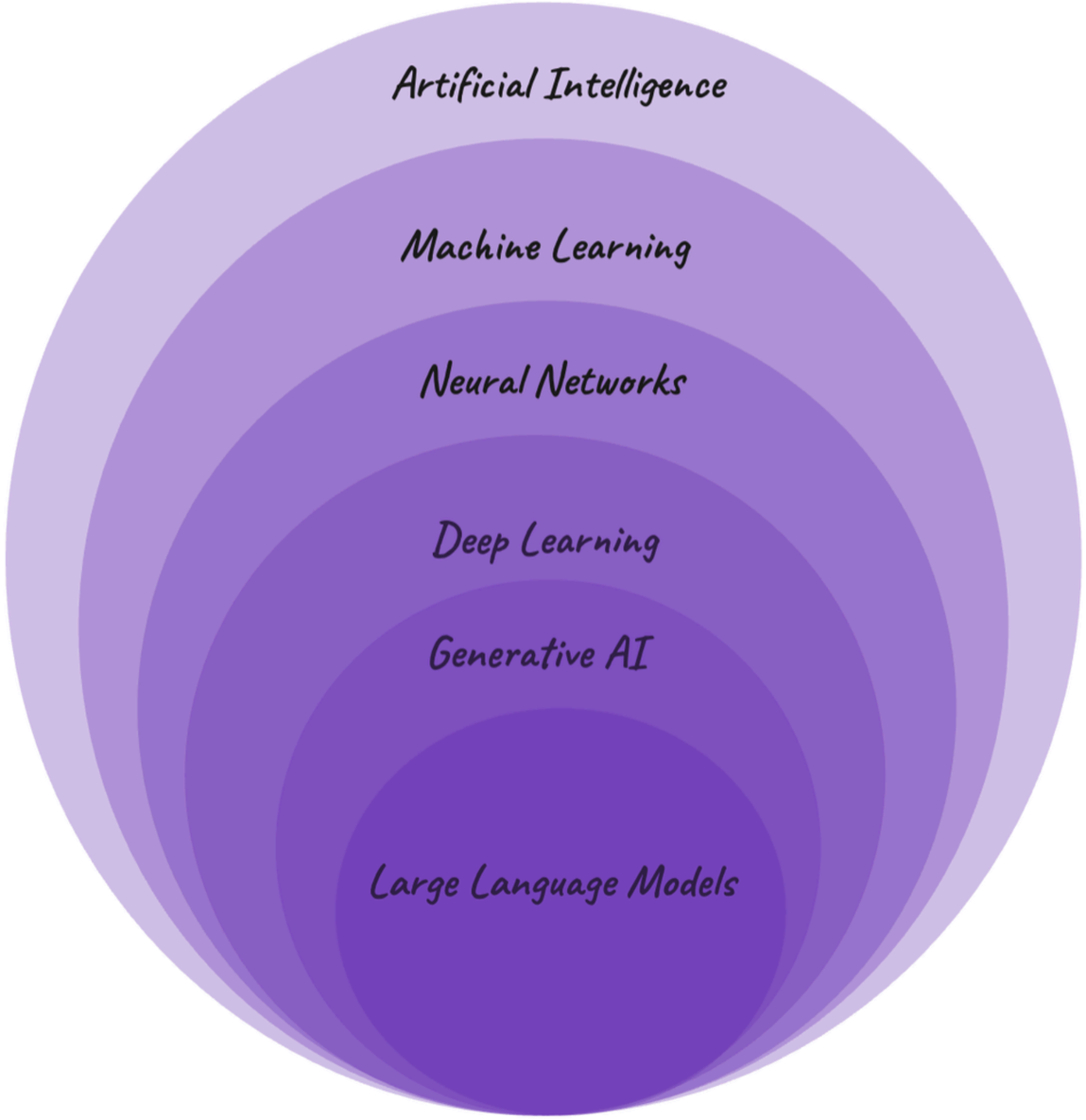
What is AI ?

AI is a **branch of computer science** that deals with the creation of intelligent **agents** which are systems that can **reason** and learn (is the development of computer systems able to perform tasks normally requiring **human intelligence**)

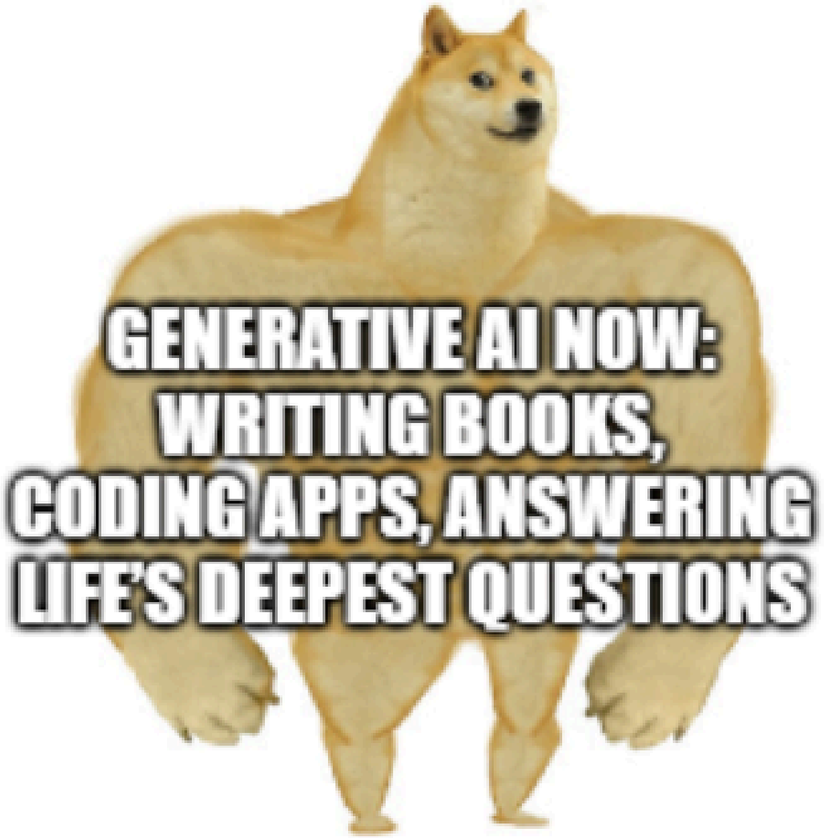
What is GenAI ?

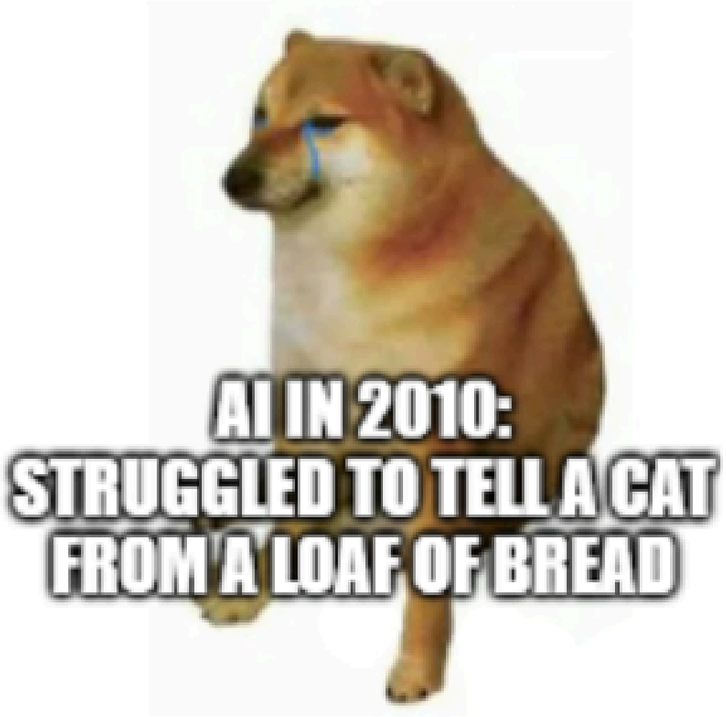
Gen AI is a type of AI technology that can **produce** various types of content , including text, imagery, audio and synthetic data .

GDG Carthage



GDG Carthage

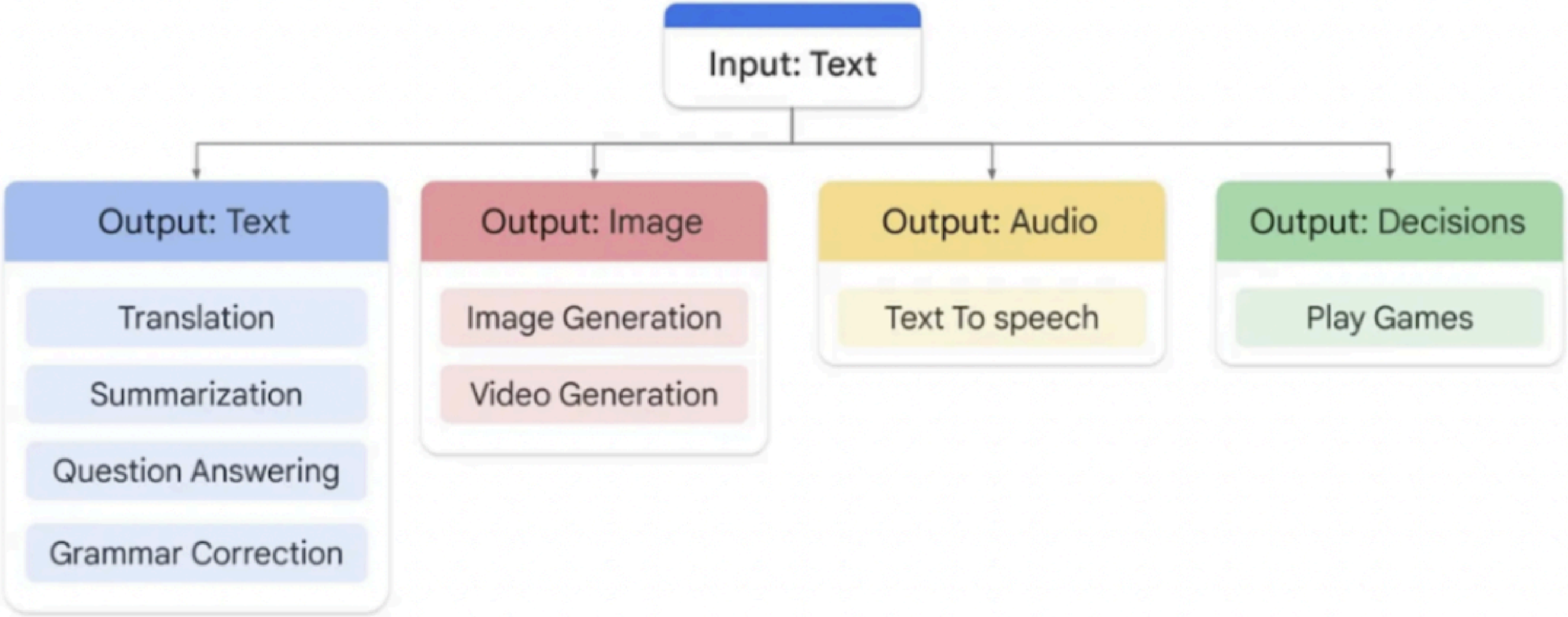




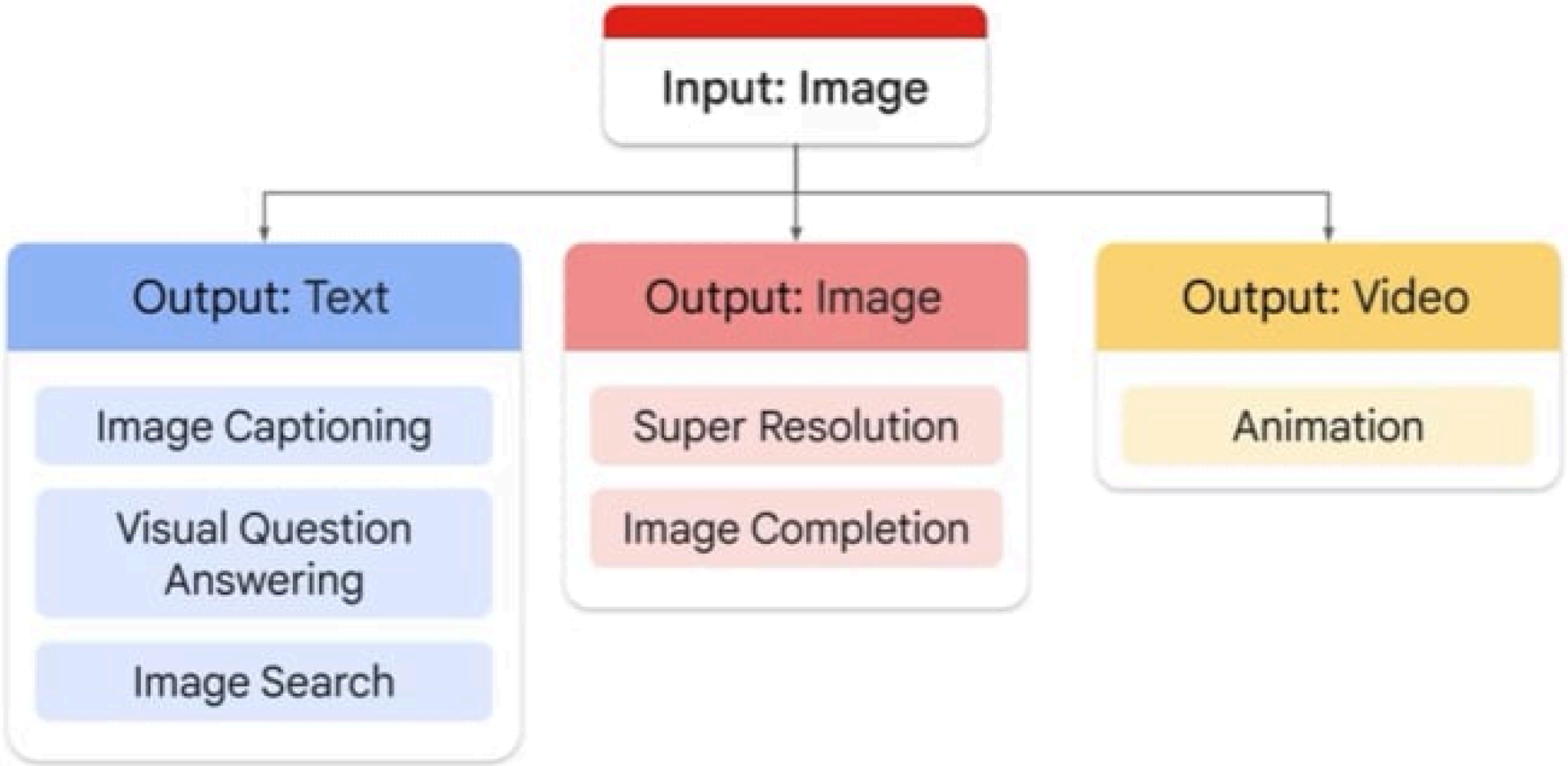
GDG Carthage

By Mohammed Arbi Nsibi

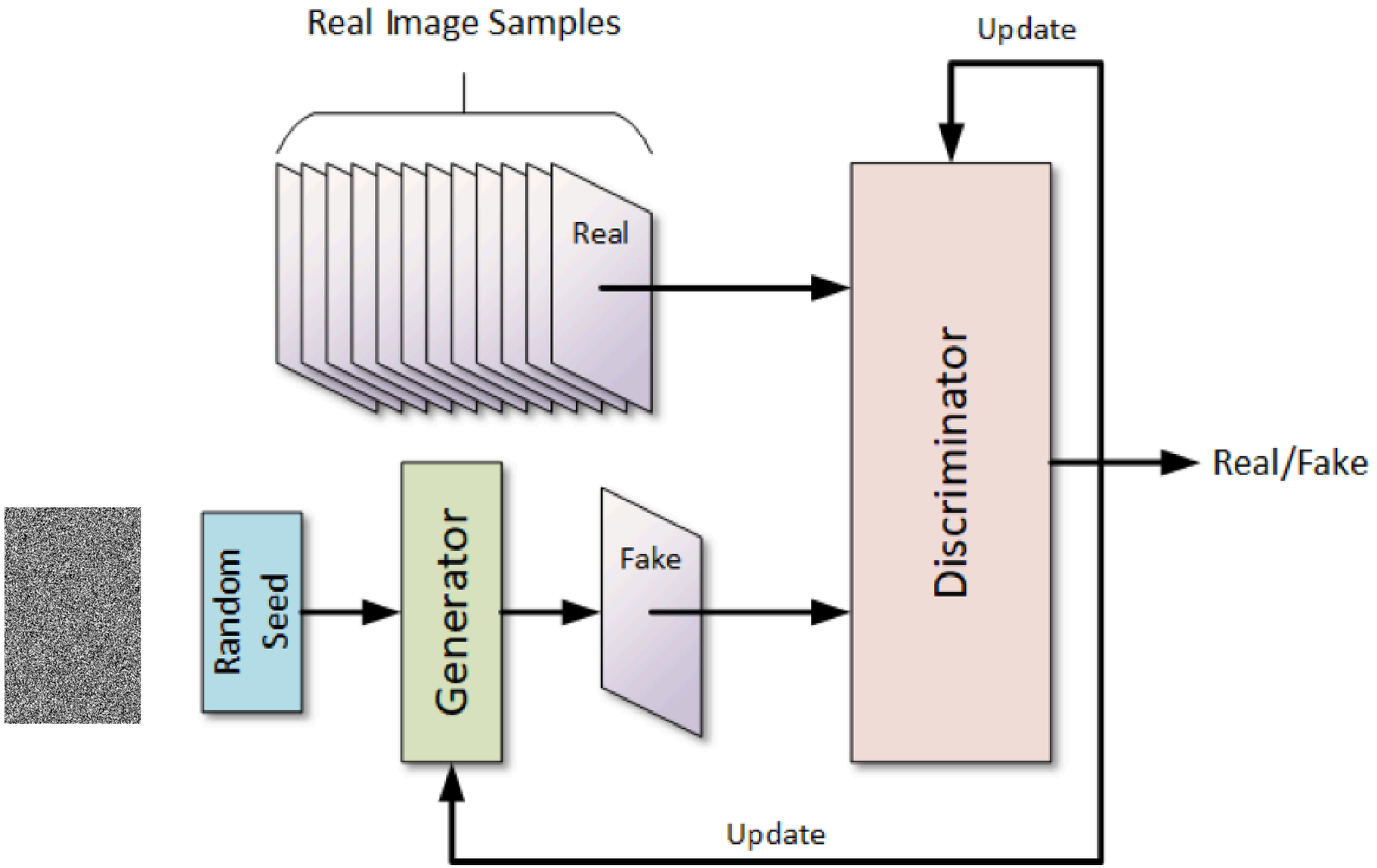
Types of Generative AI Based on Data



Types of Generative AI Based on Data



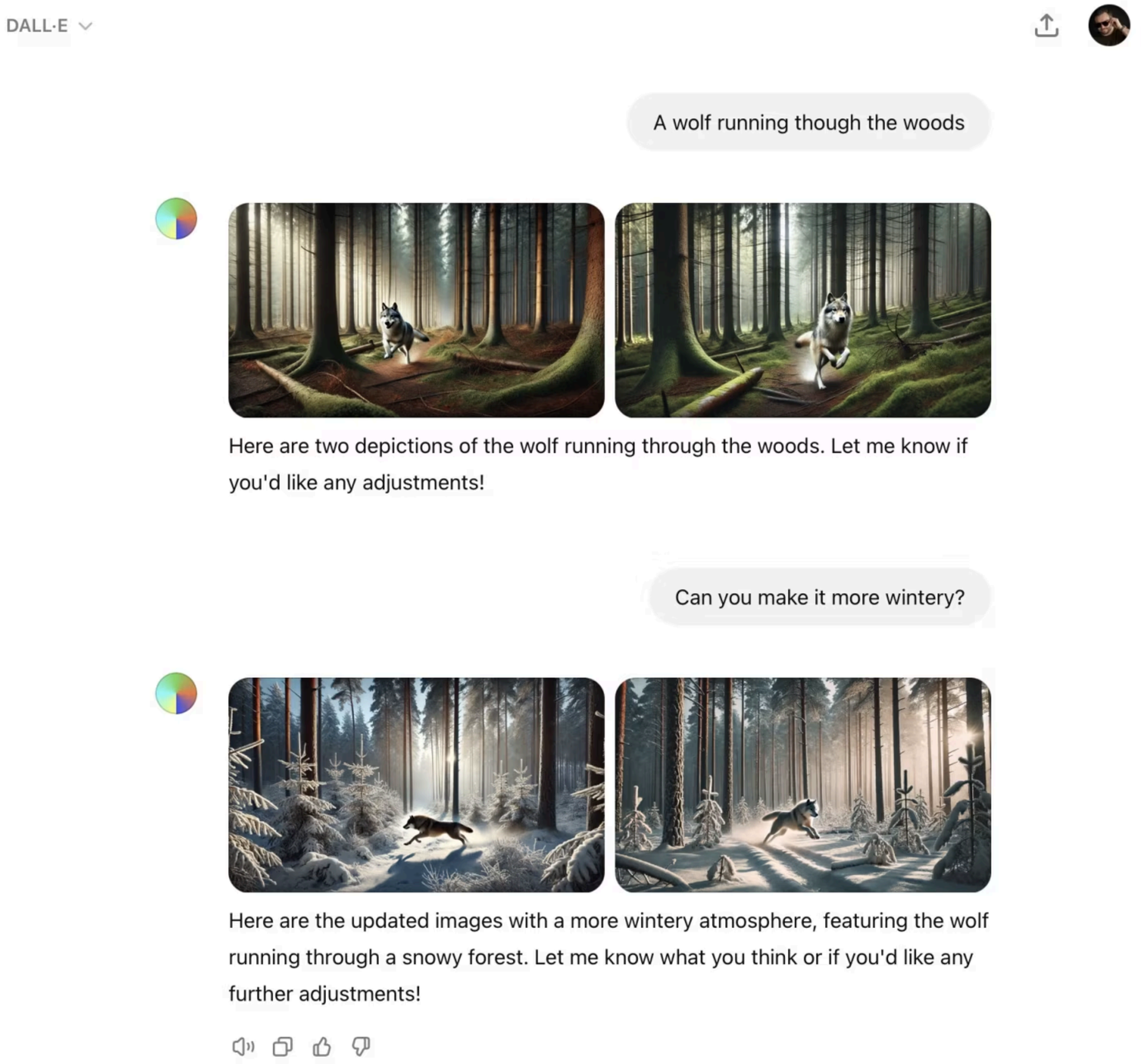
GANs: Generative adversarial network



https://arxiv.org/abs/1406.2661

Diffusion models :

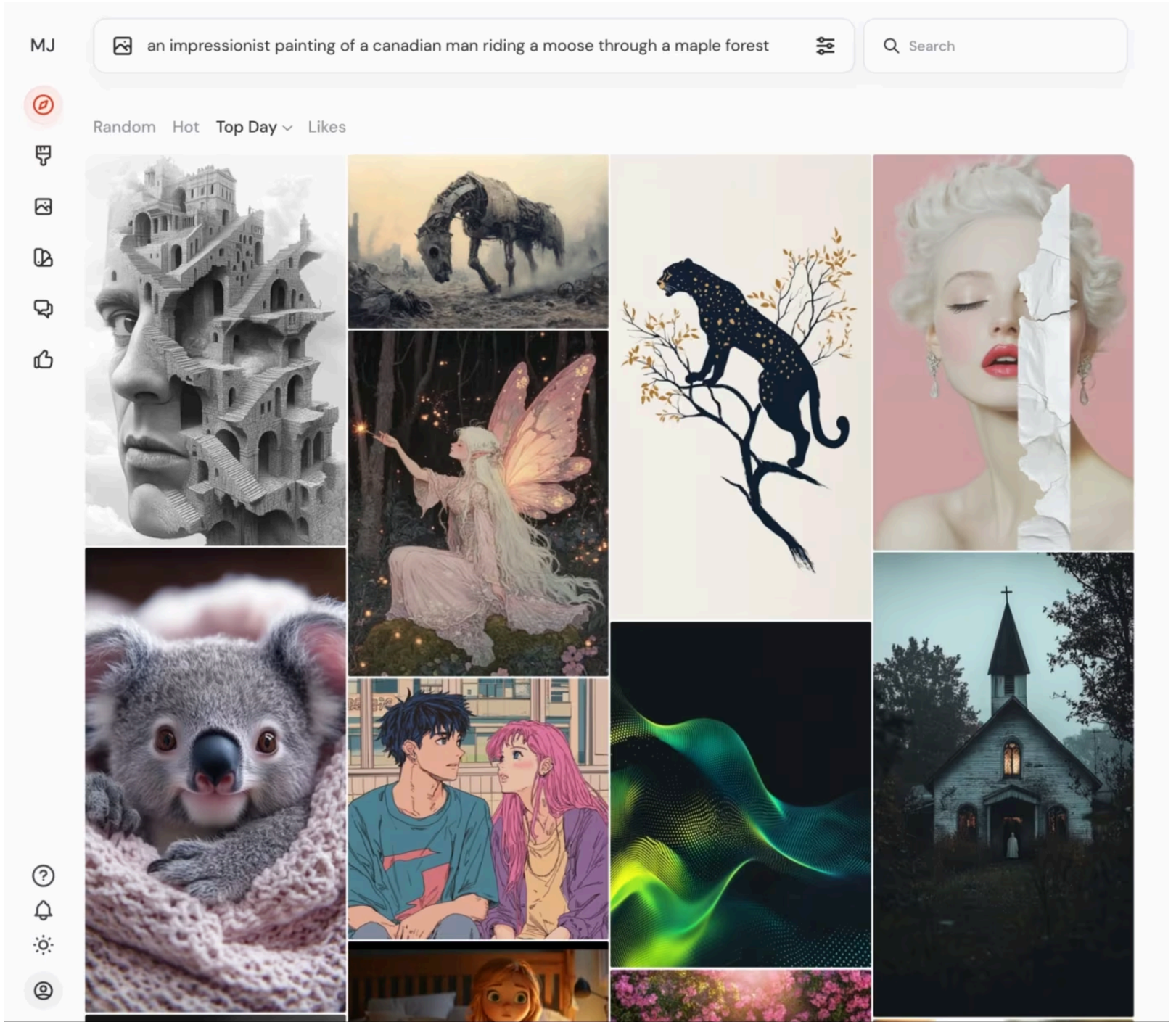
DALL-E



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

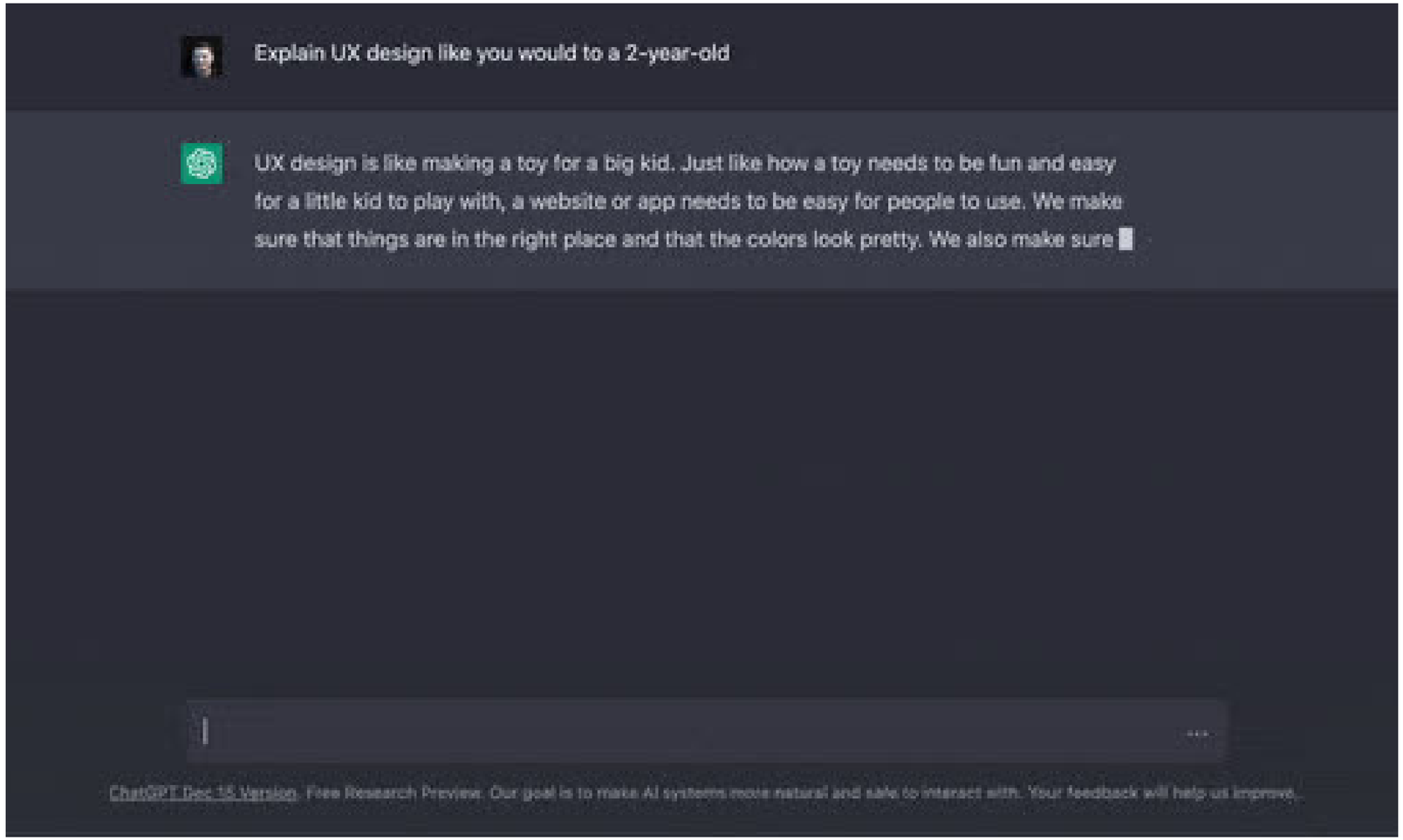
Diffusion models :

Midjourney



https://www.midjourney.com/home

ChatGPT



GDG Carthage

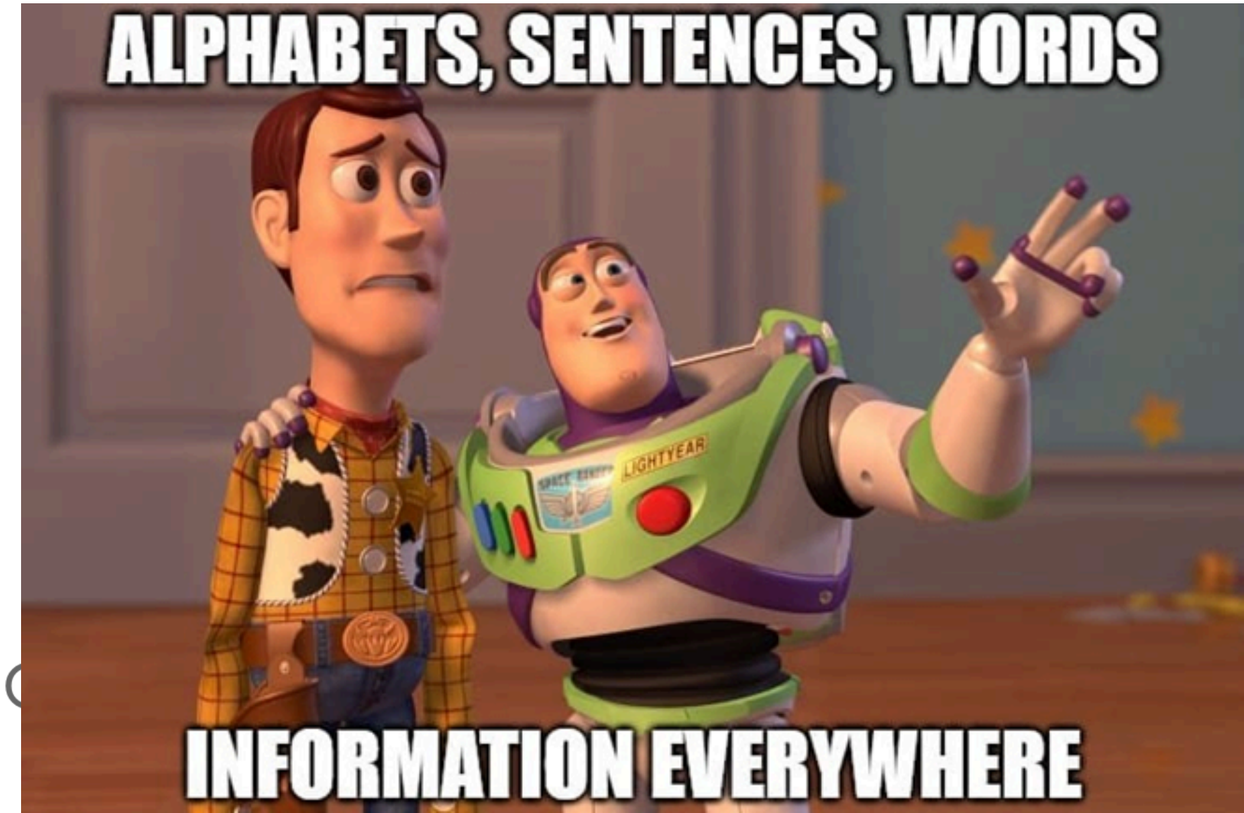
GPT-4 VS HUMAN TESTS (MAY/2023)

Selected highlights only. Percentiles; 50 refers to the 50th percentile as average, and may not be the testing average for some tests. Alan D. Thompson. April 2023. https://lifearchitect.ai/iq-testing-ai/

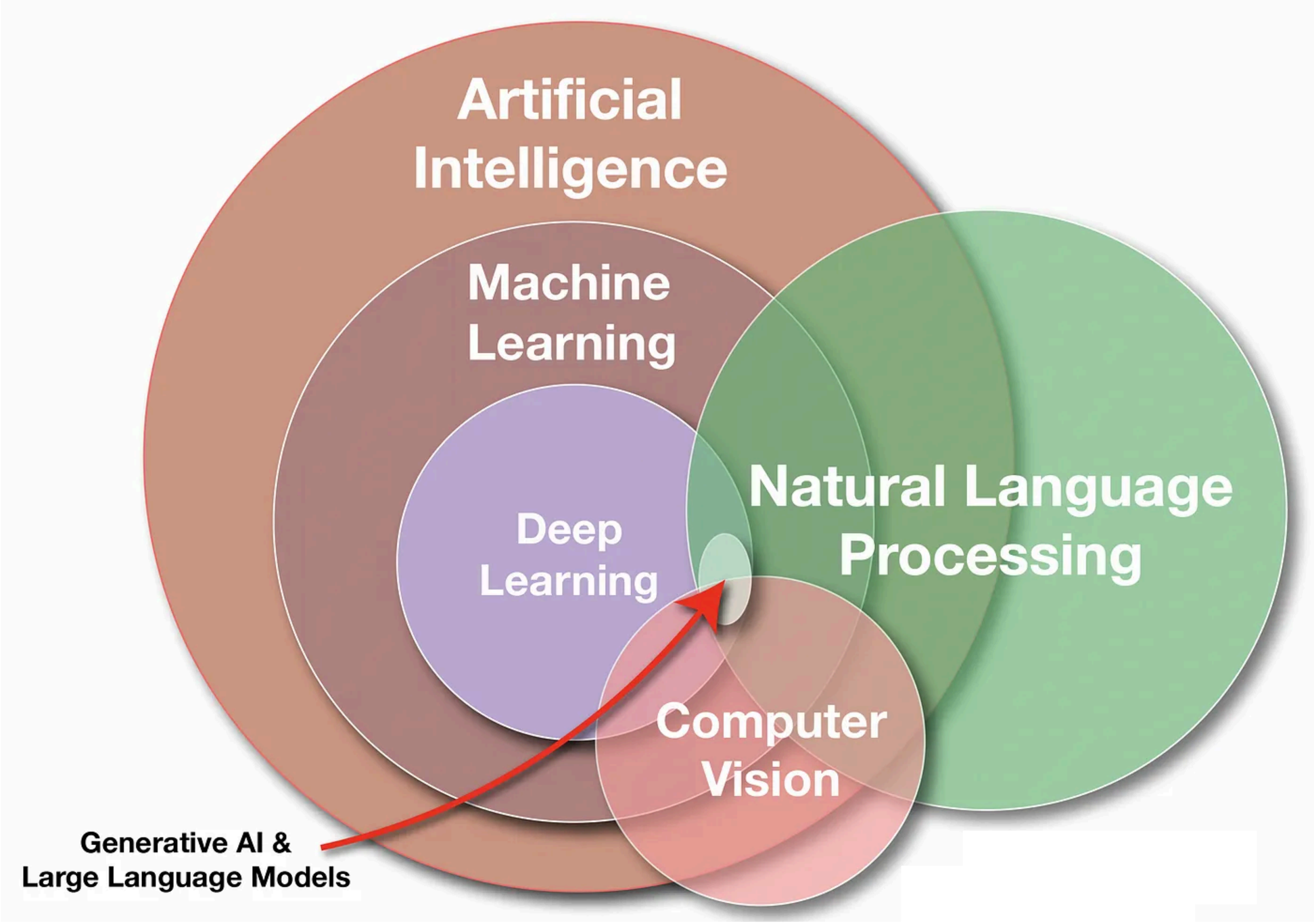
WHAT IS NLP ?

WHAT IS NLP ?

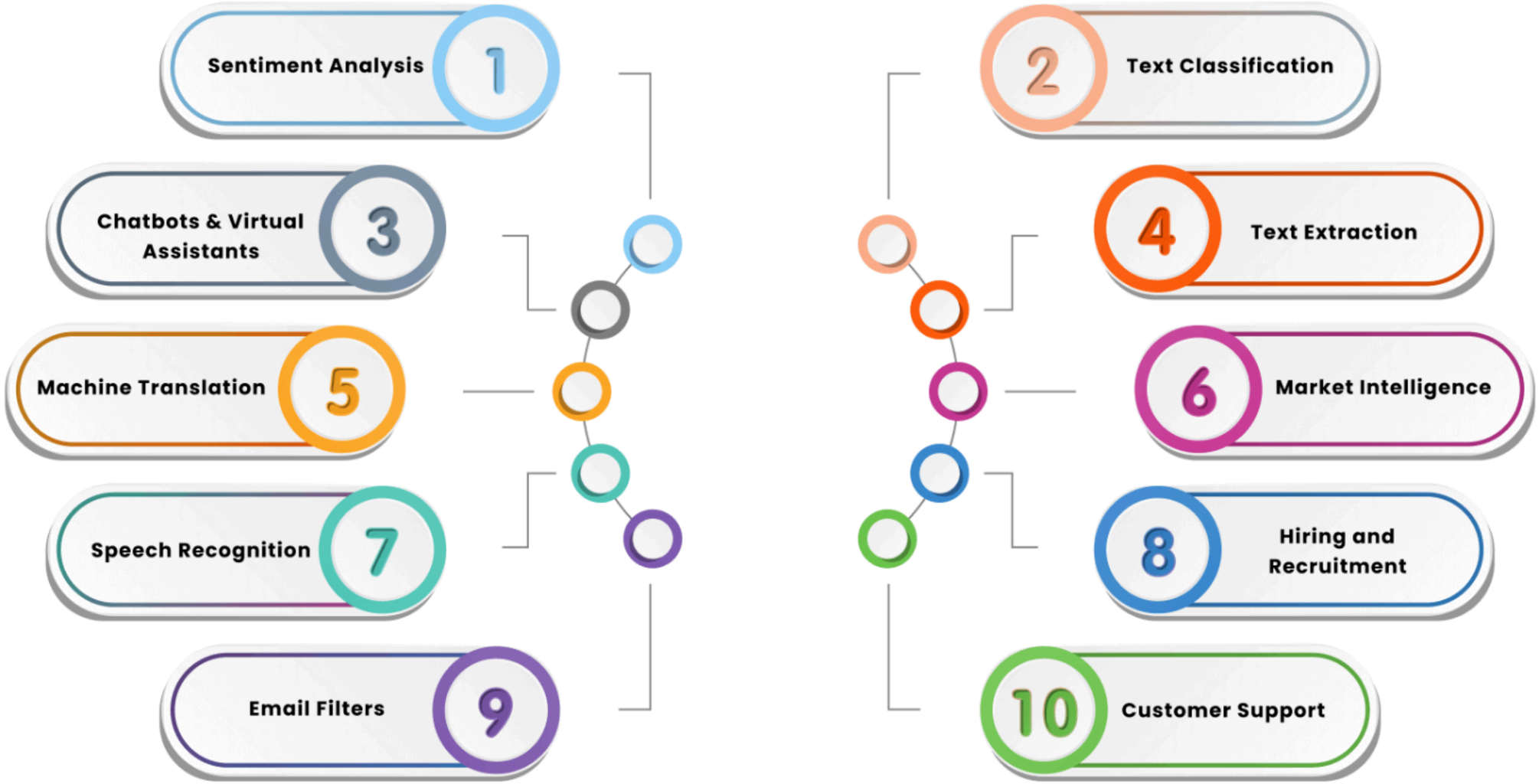
is referred to as NLP. It is a **subset** of AI that enables machines to comprehend and analyze **human languages**. Text or audio can be used to represent human languages.



GDG C

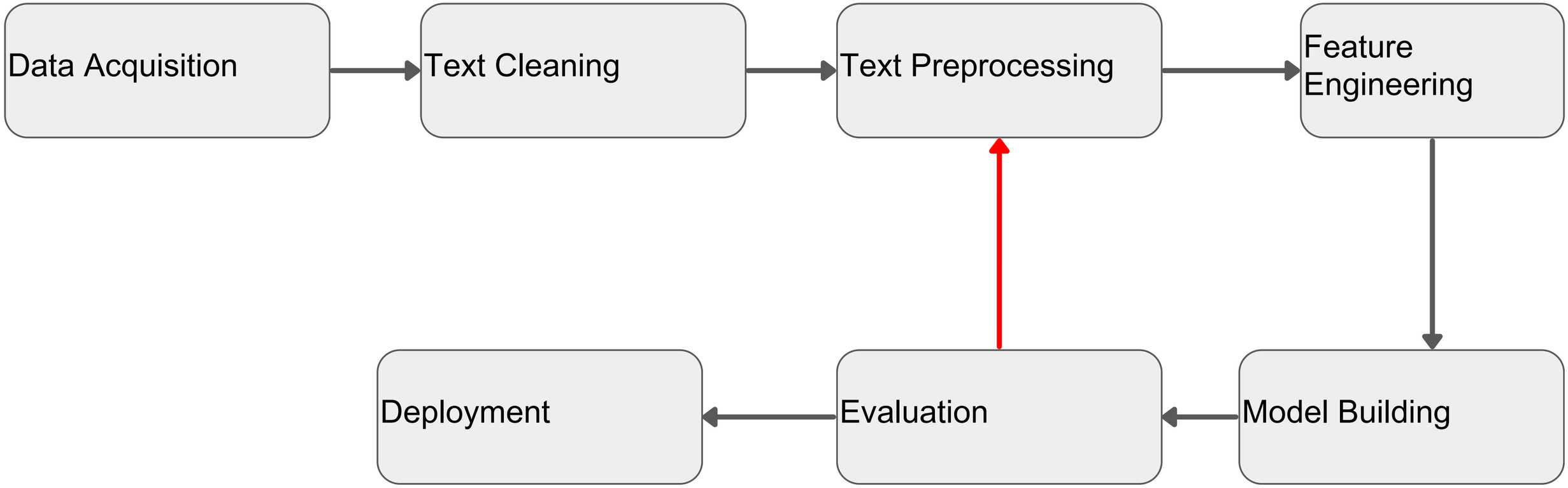

Artificial Intelligence
Machine Learning
Deep Learning
Natural Language Processing
Computer Vision
Generative AI & Large Language Models

NLP applications



GDG Carthage

NLP Pipeline



GDG Carthage

By Mohammed Arbi Nsibi

Text Cleaning

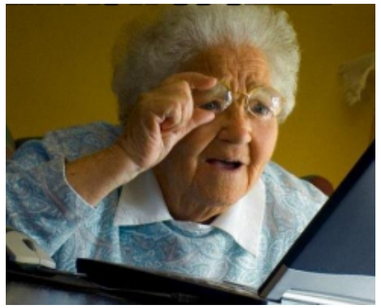
Regex or Regular Expression

Spelling corrections

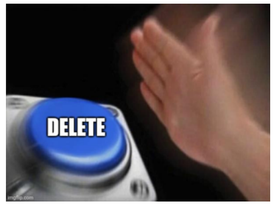
used for searching the string of specific patterns. Suppose our data contain phone number, email-Id, and URL. we can find such text using the regular expression. After that either we can keep or remove such text patterns as per requirements.

GDG Carthage

Text Preprocessing



- Lowercasing
- Stop word removal
- Stemming or lemmatization
- Removing digit/punctuation
- POS tagging
- Named Entity Recognition (NER)



GDG Carthage

By Mohammed Arbi Nsibi

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

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

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

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

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?

GDG Carthage

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**

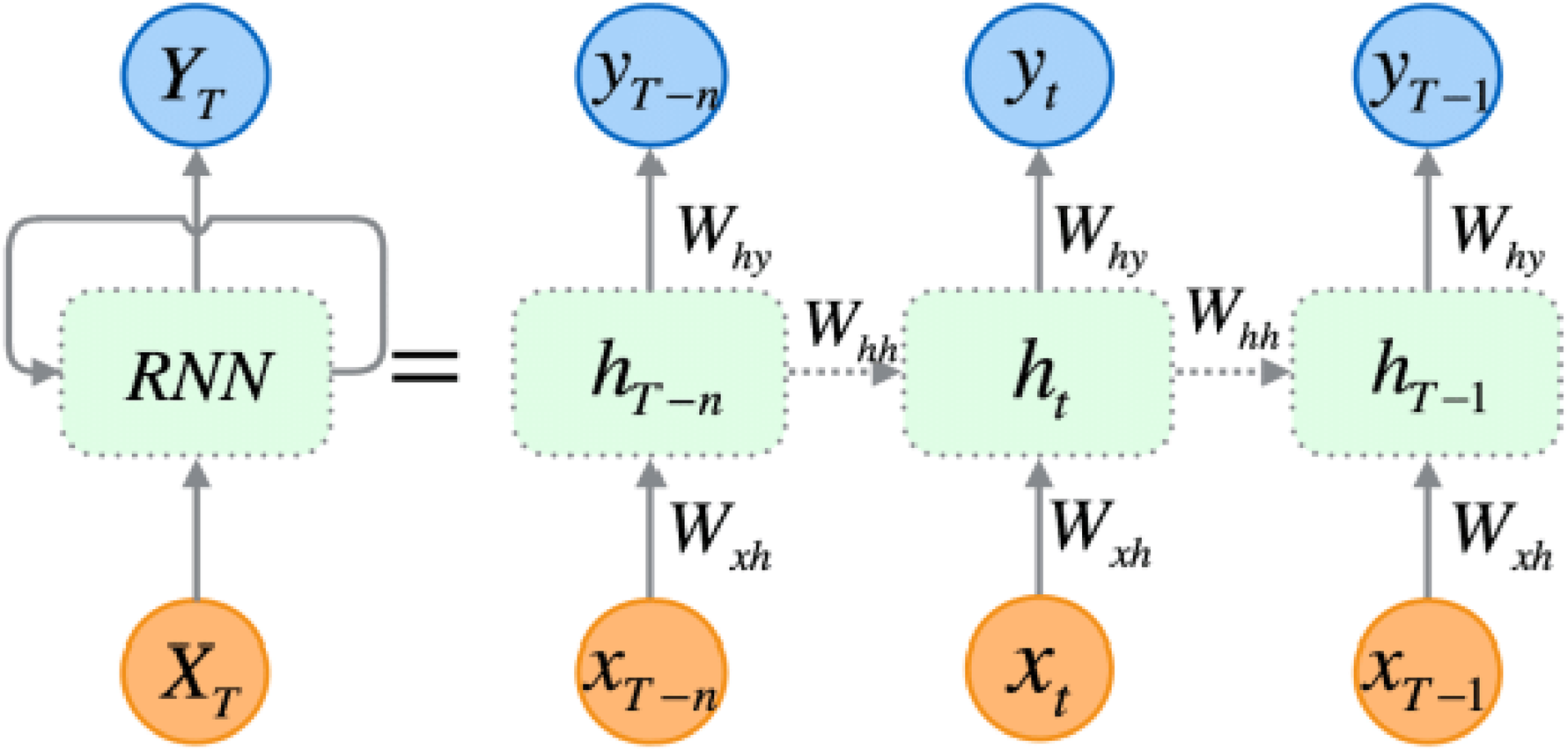
Visualizing High-Dimensional Space: https://www.youtube.com/watch?v=wvsE8jm1GzE&ab_channel=GoogleforDevelopers

By Mohammed Arbi Nsibi

LET'S CODE 

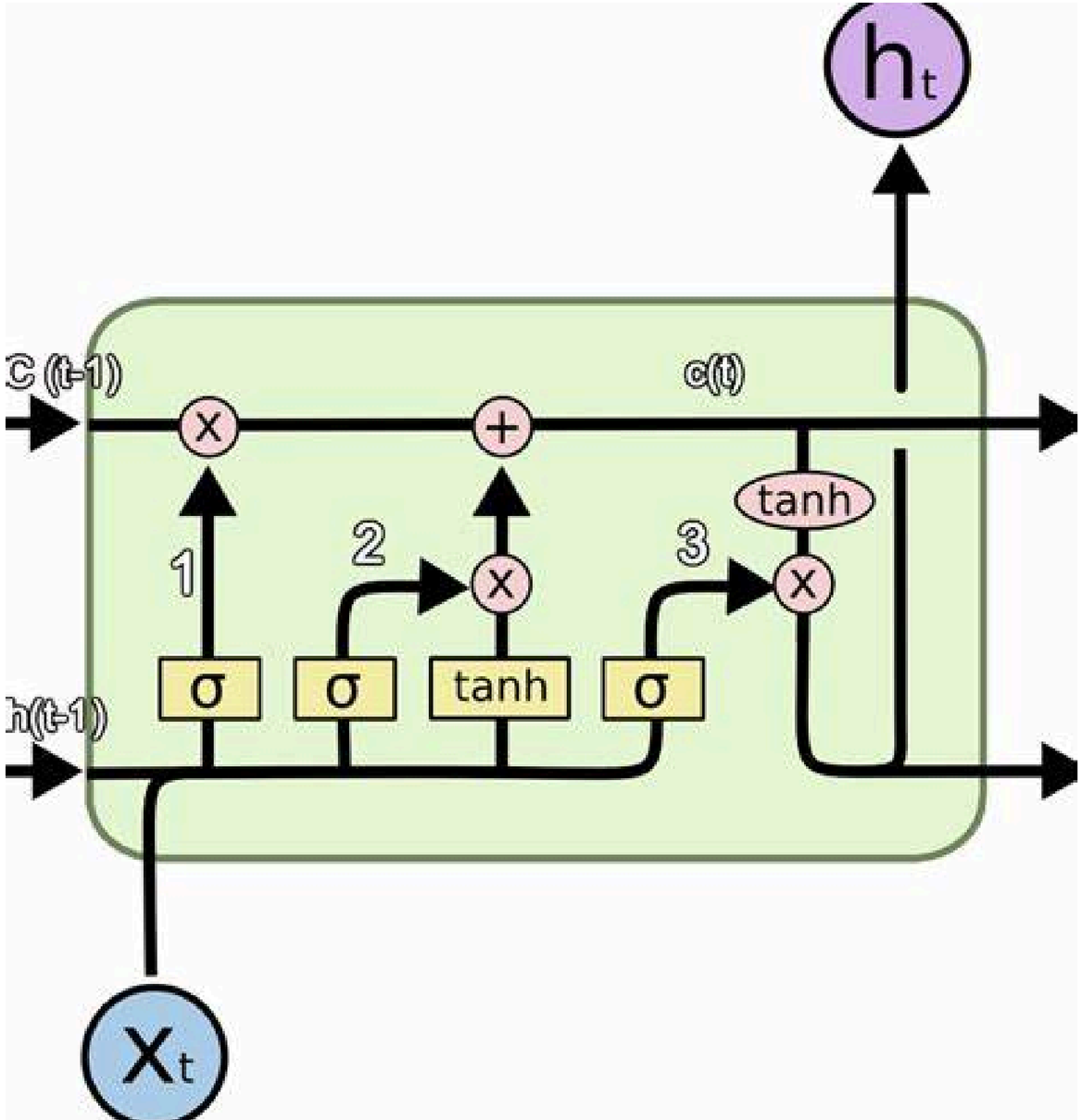
GDG Carthage

RNNs : RNN is a type of neural network designed to work with sequential data, like sentences or time series. The key feature is that it **remembers information from previous steps**



Prolem: vanishing gradient problem

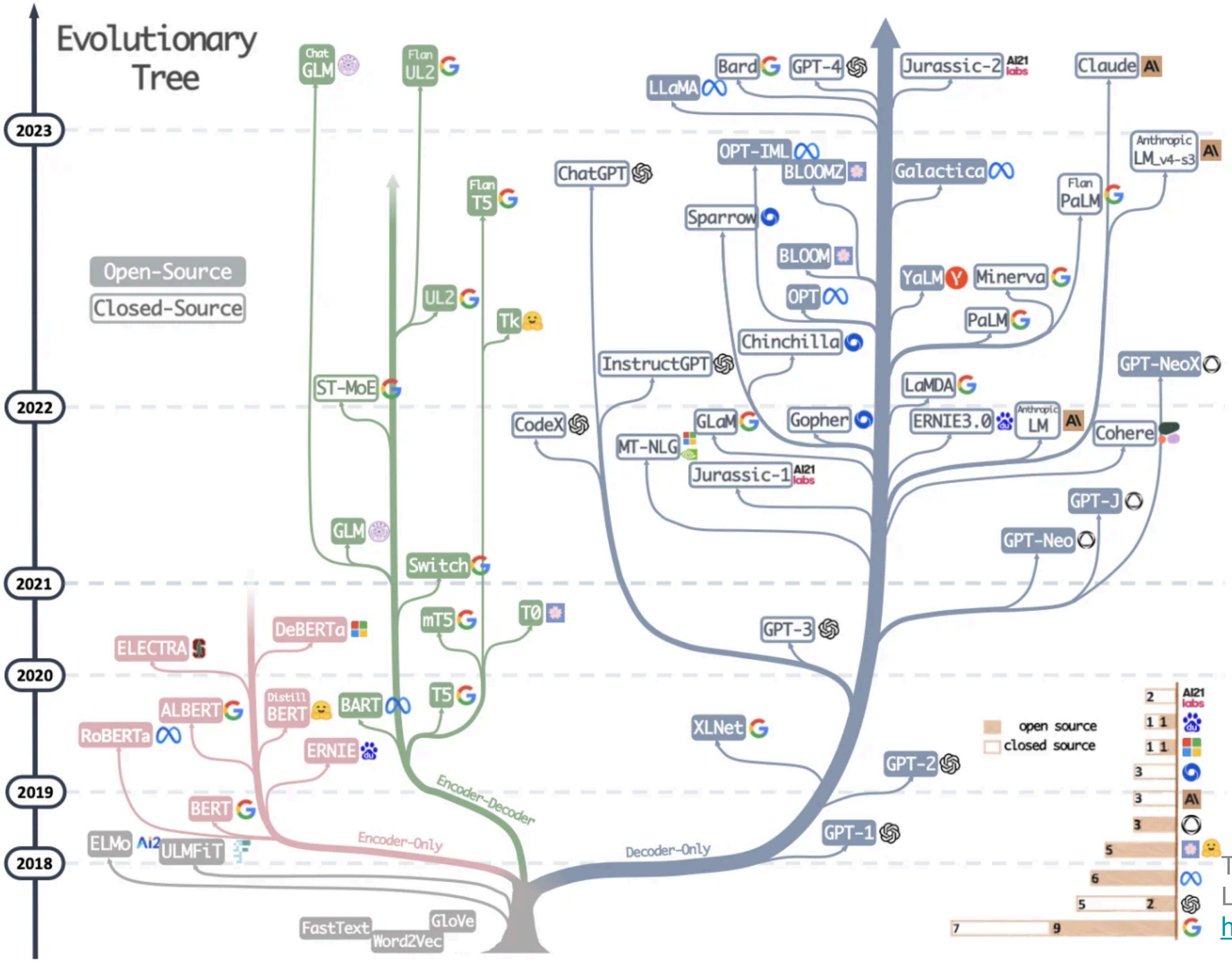
LSTM : is a special type of RNN designed to solve the problem of remembering things over long sequences. It has a more complex internal structure that lets it **decide what to remember and what to forget** more effectively.



NLP vs LLM

GDG Carthage

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?

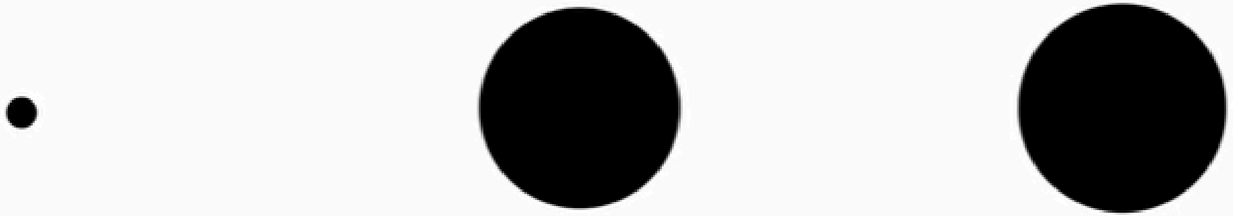
Llama **3.2**

(September 2024)

GDG Carthage

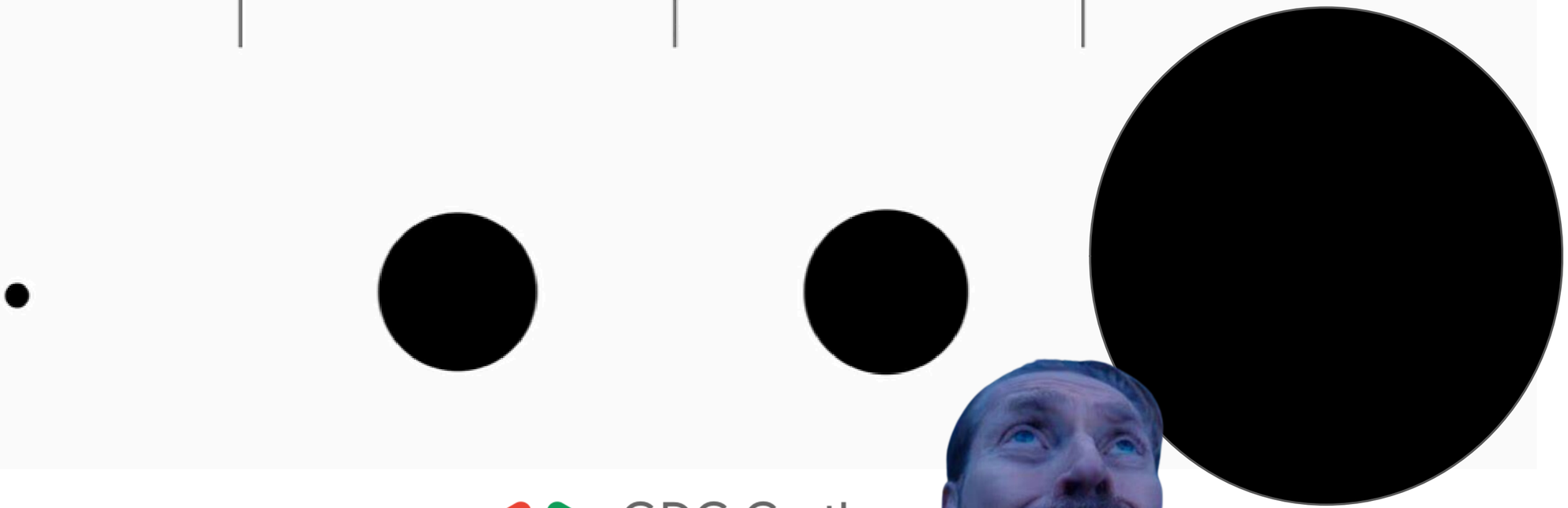
Model size

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



Model size

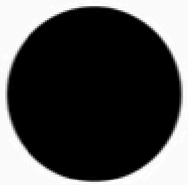
2019	2023	2023	2024
GPT-2	Llama-1	Llama-2	Llama-3
124M to 1.5B	7B to 65B	7B to 70B	8B to 405B

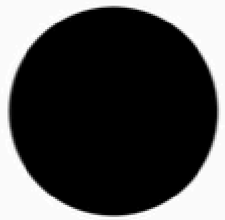


GDG Carthage

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

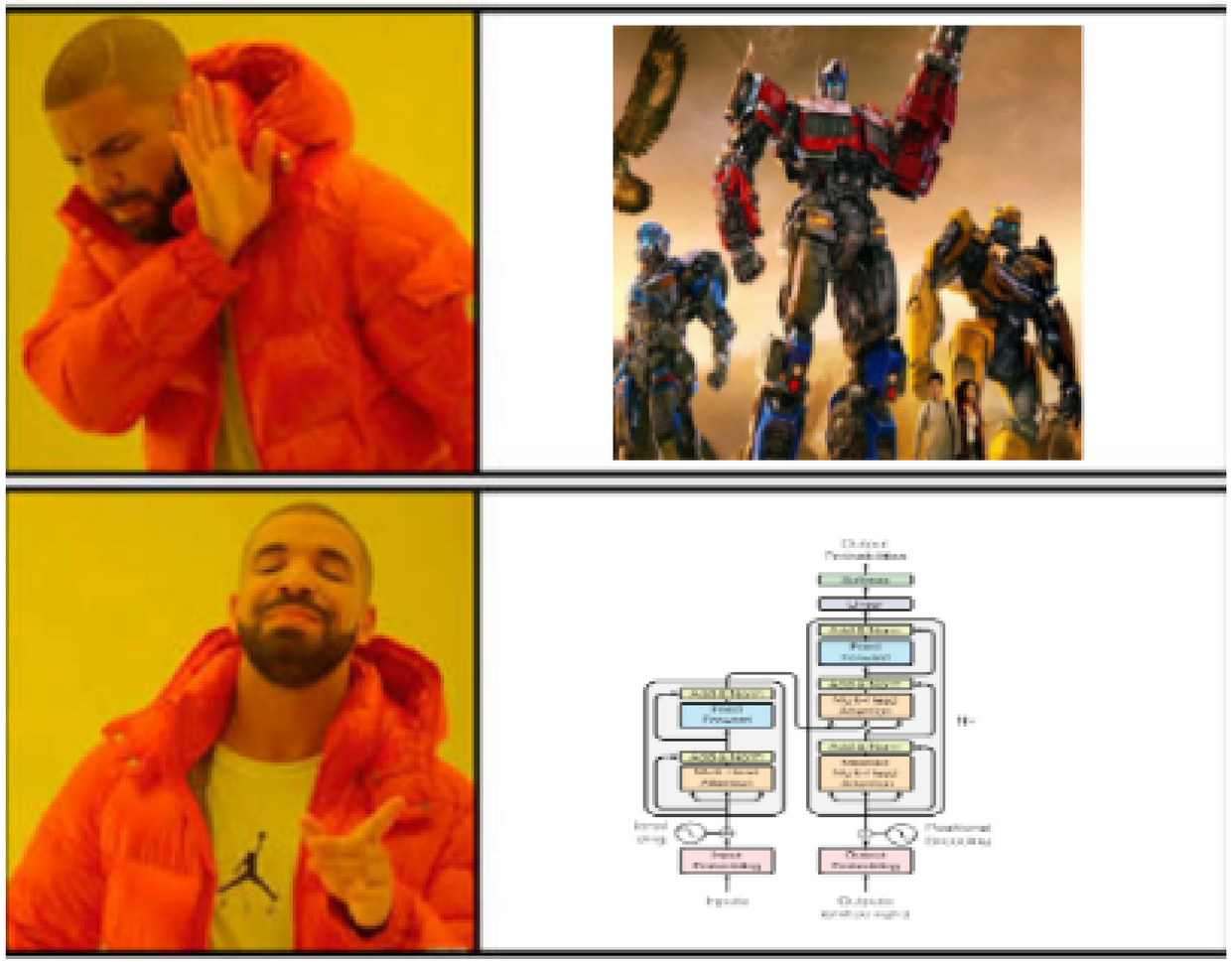
Abstract

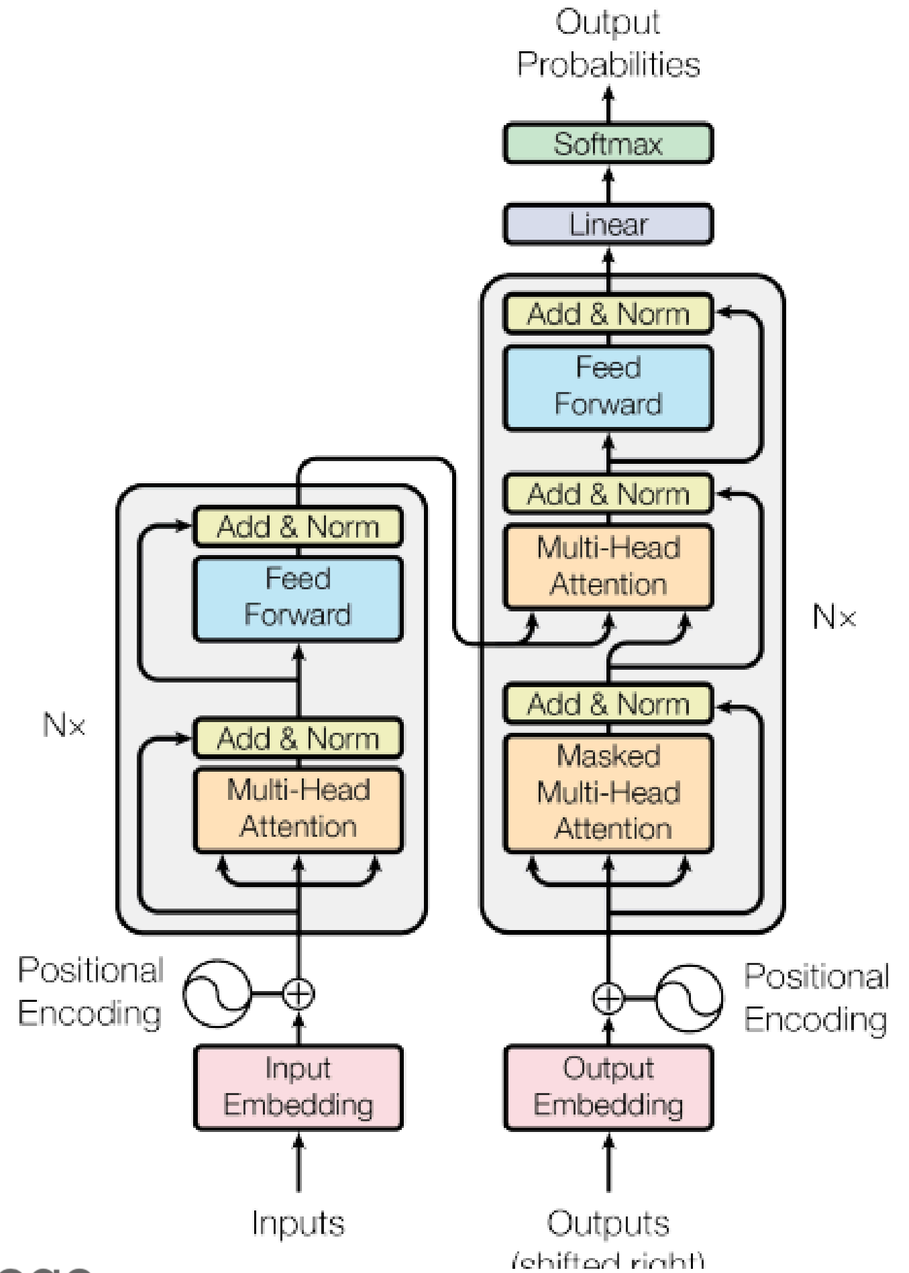
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, 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

$$E = X + PE$$

GDG Carthage

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

Query = What I want to know (question)

Key = label that tells others what they re good at (what we have available)

Value = important info they can share(content)

Attention helps the model decide:

- What to focus on
- How much to focus on

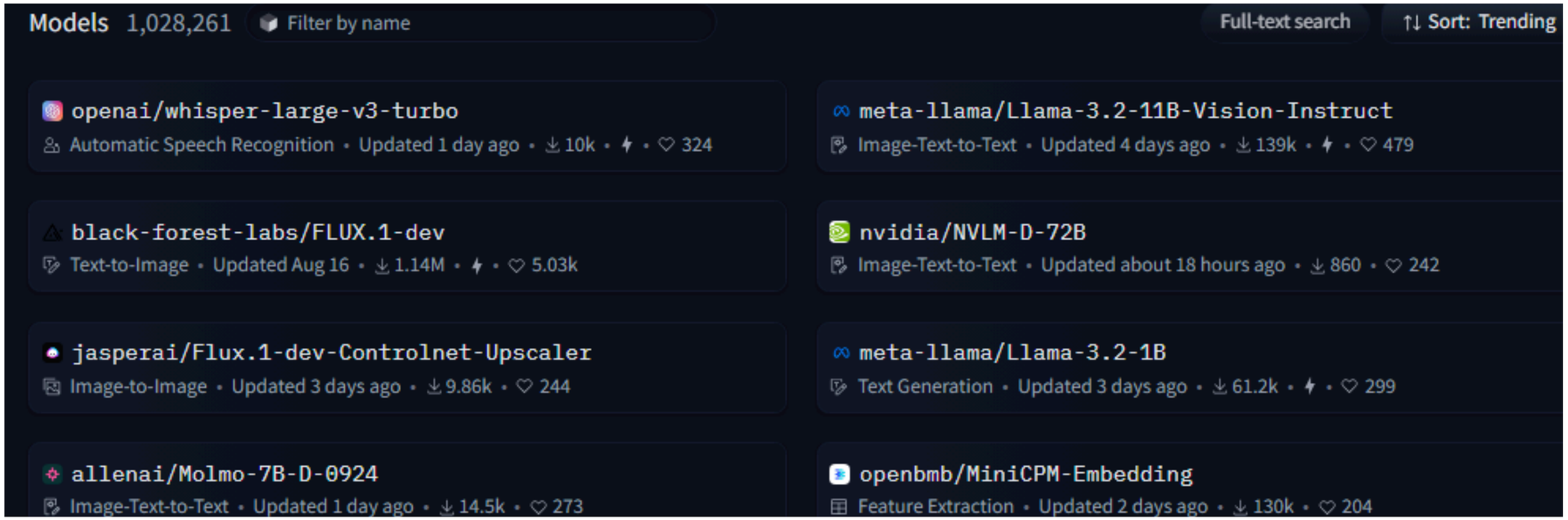
Where to find pretrained LLMs ?

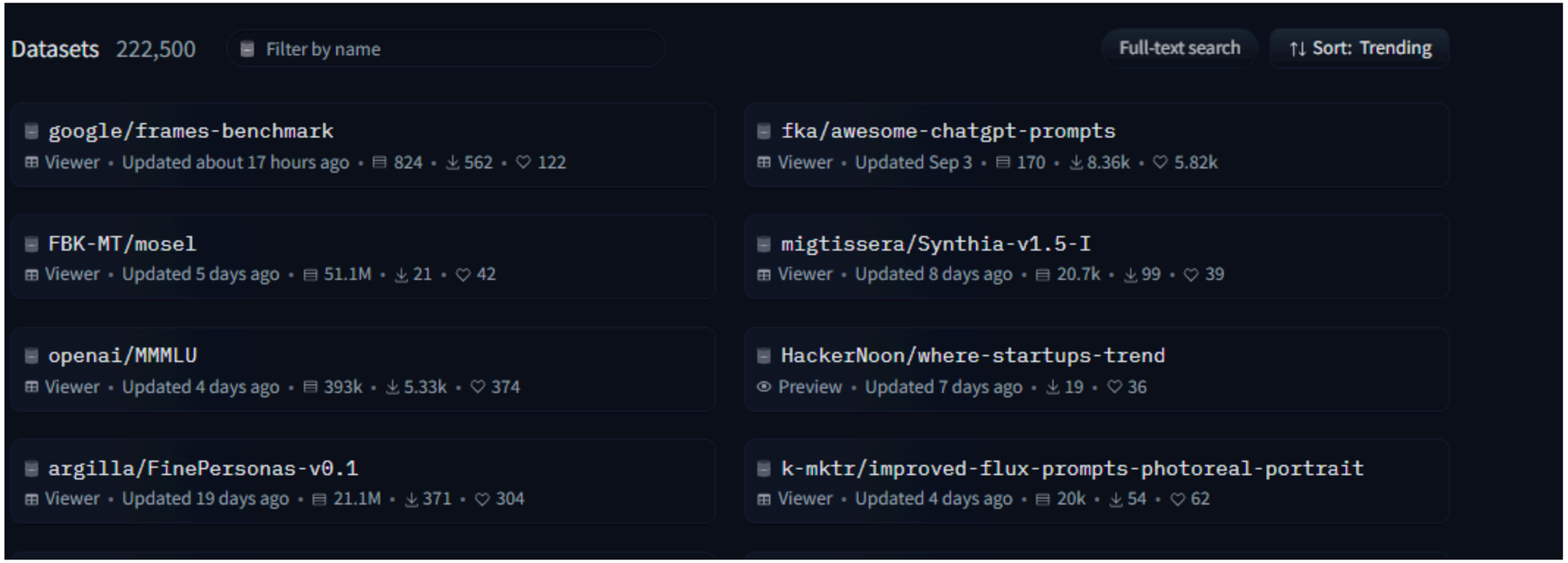
Hugging Face

GDG Carthage

By Mohammed Arbi Nsibi

Where to find pretrained LLMs ?





Model types

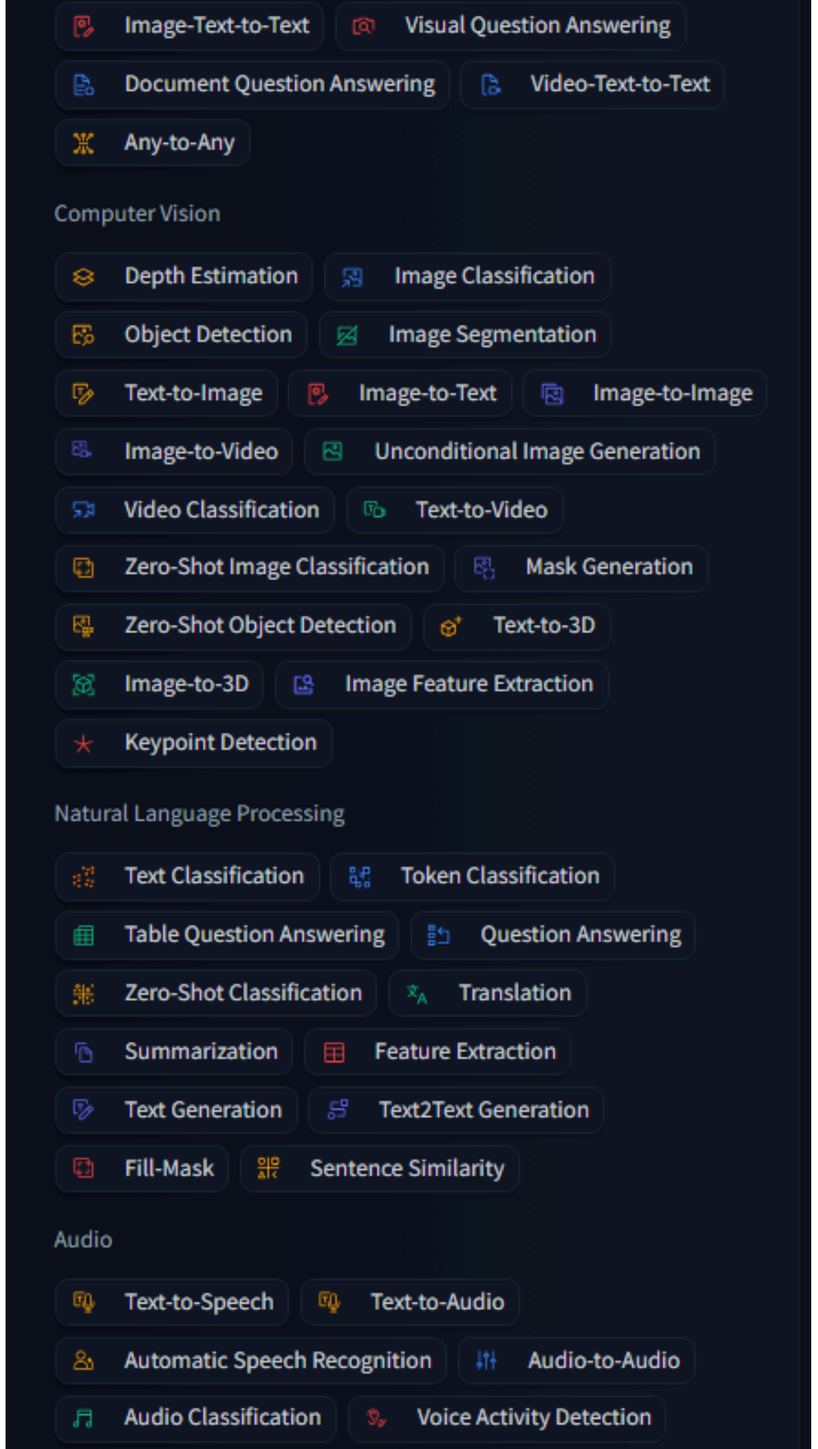
Text-to-text

Text-to-image

Text-to-video

Sentence-similarity

GDG Carthage



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 

GDG Carthage

By Mohammed Arbi Nsibi

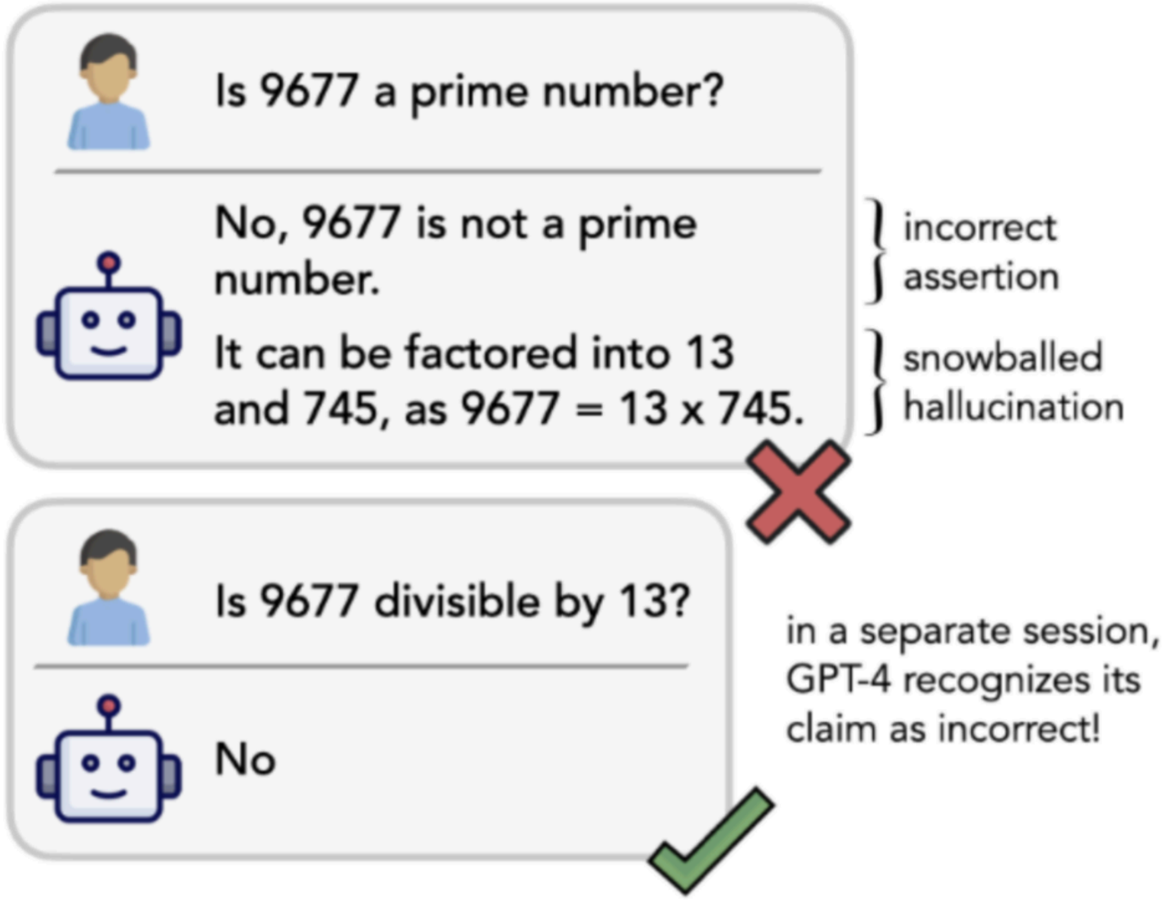
Hallucinations

- 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.

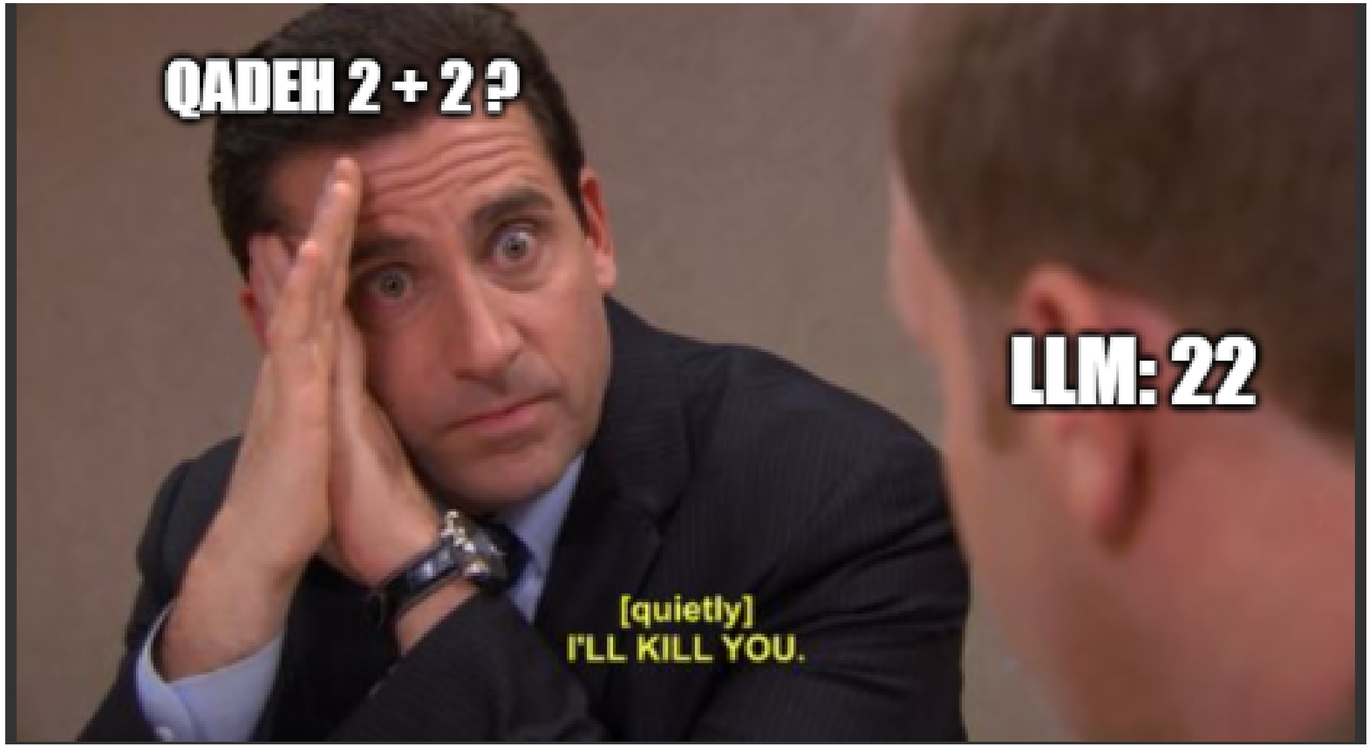
GDG Carthage

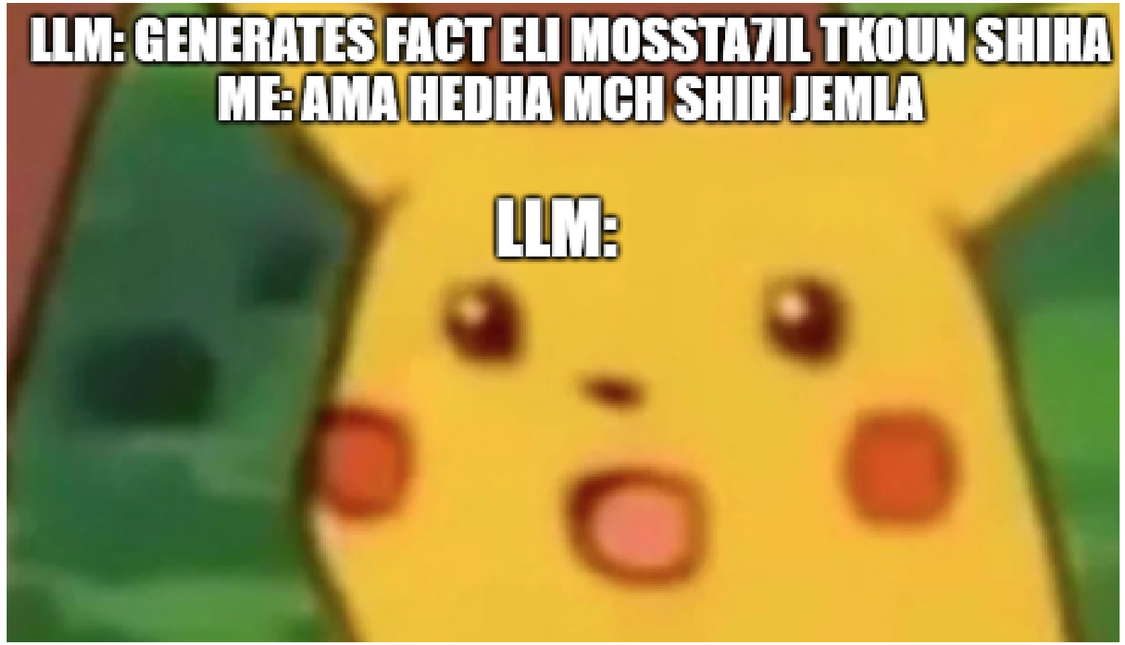
By Mohammed Arbi Nsibi

Hallucinations



Hallucinations





GDG Carthage

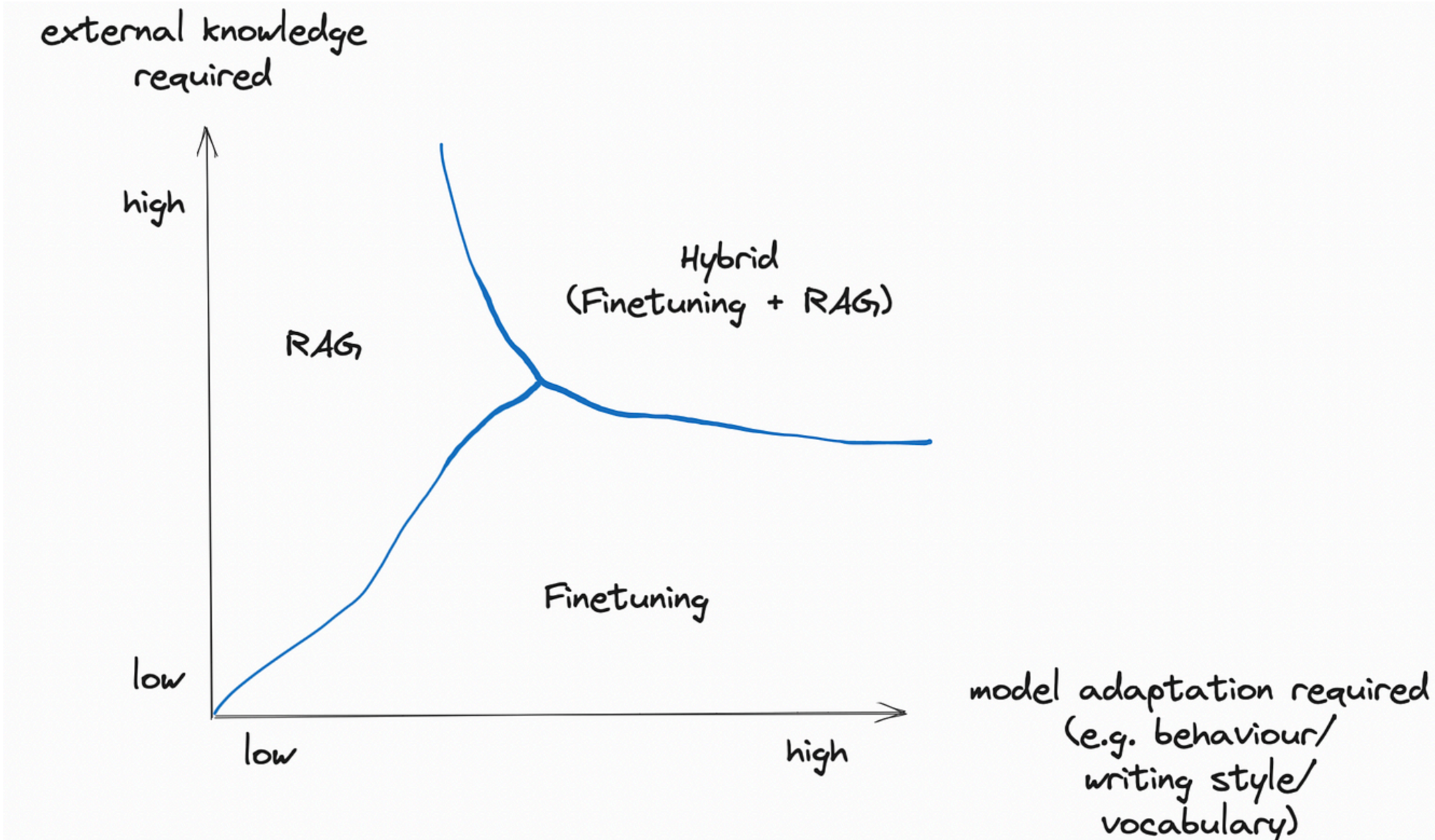
By Mohammed Arbi Nsibi

Solutions ?

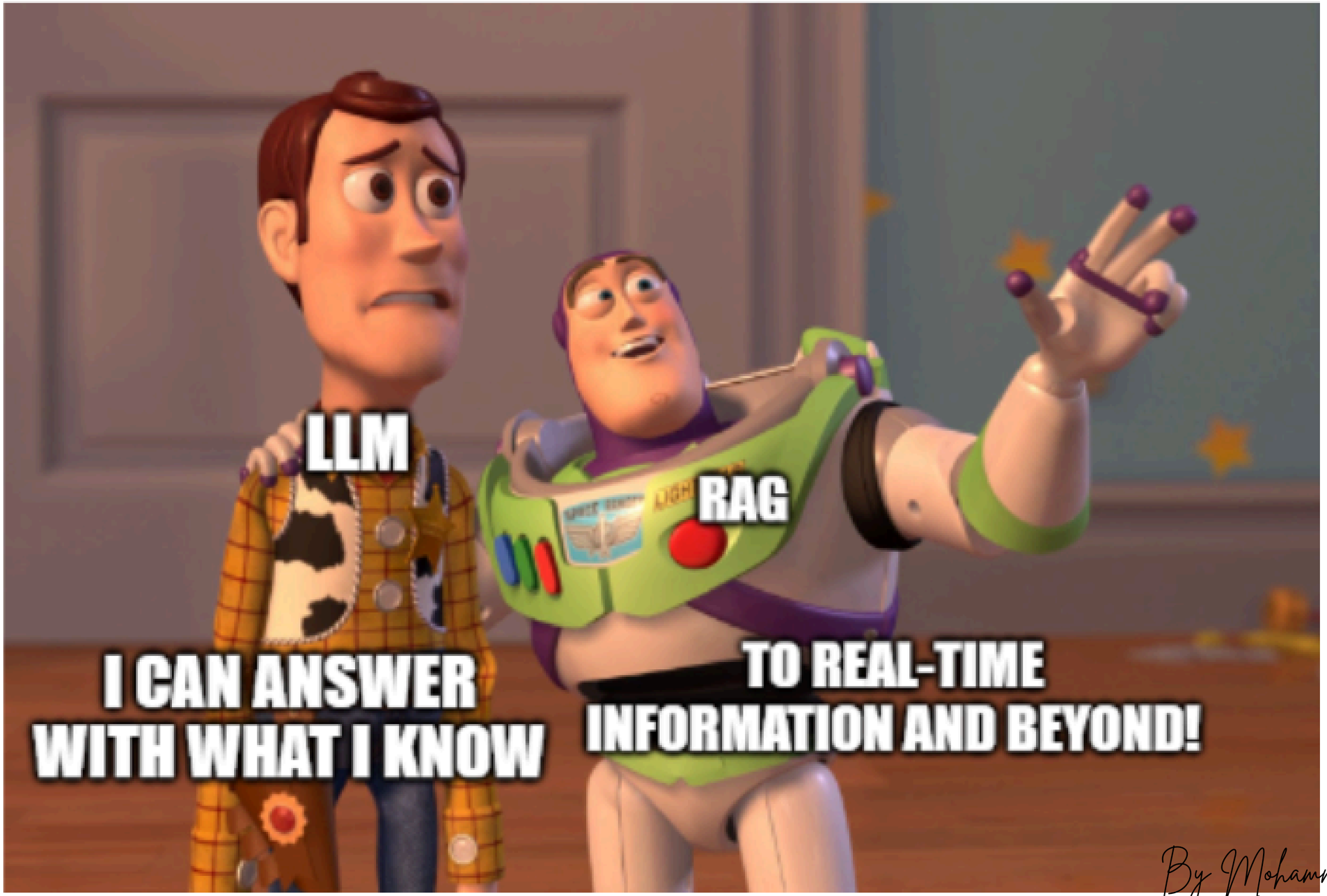
GDG Carthage

By Mohammed Arbi Nsibi

RAG / Fine-tuning



RAG (Retrieval-augmented generation)



By Mohammed Arbi Nsibi

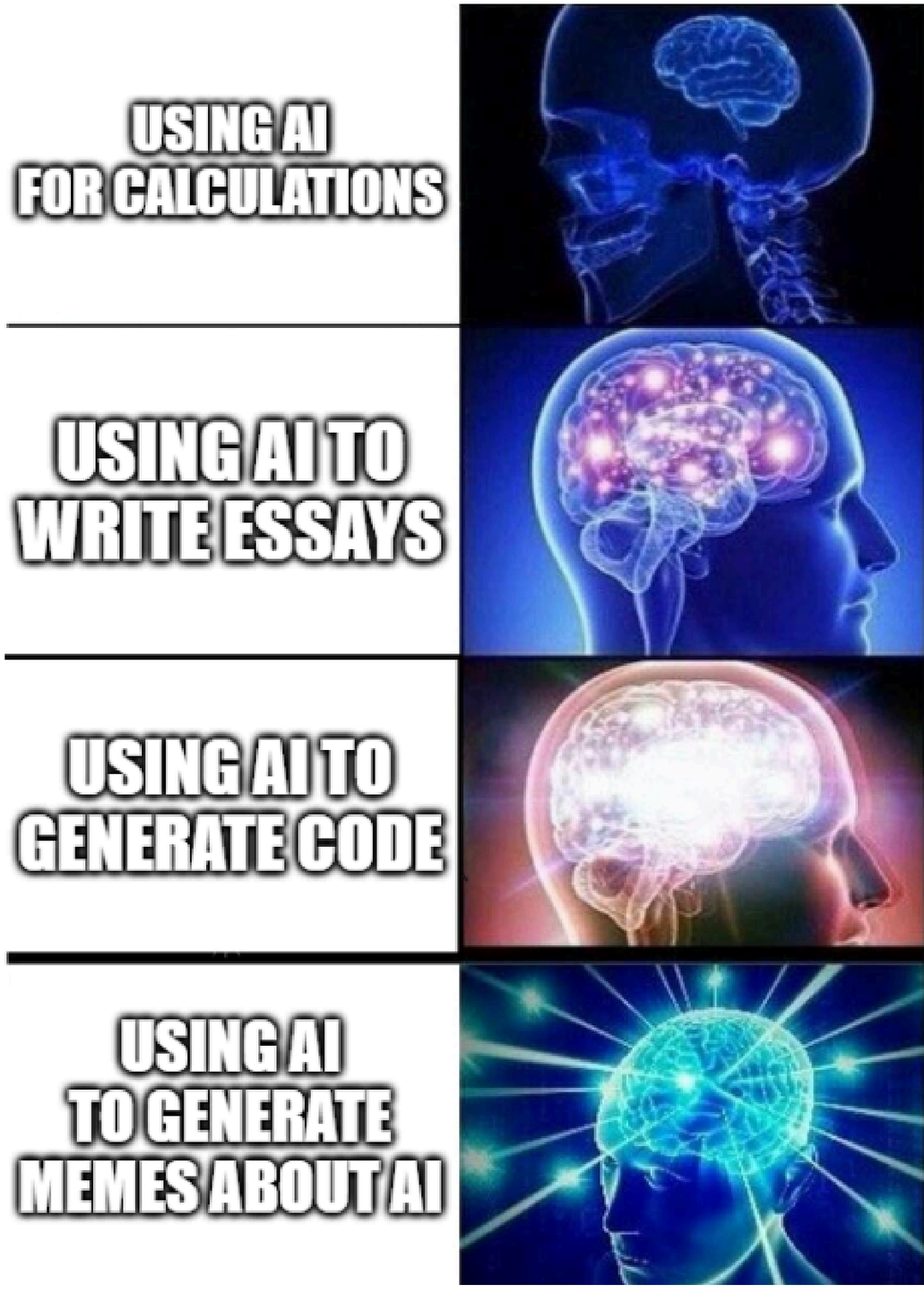
Q&A

GDG Carthage

By Mohammed Arbi Nsibi

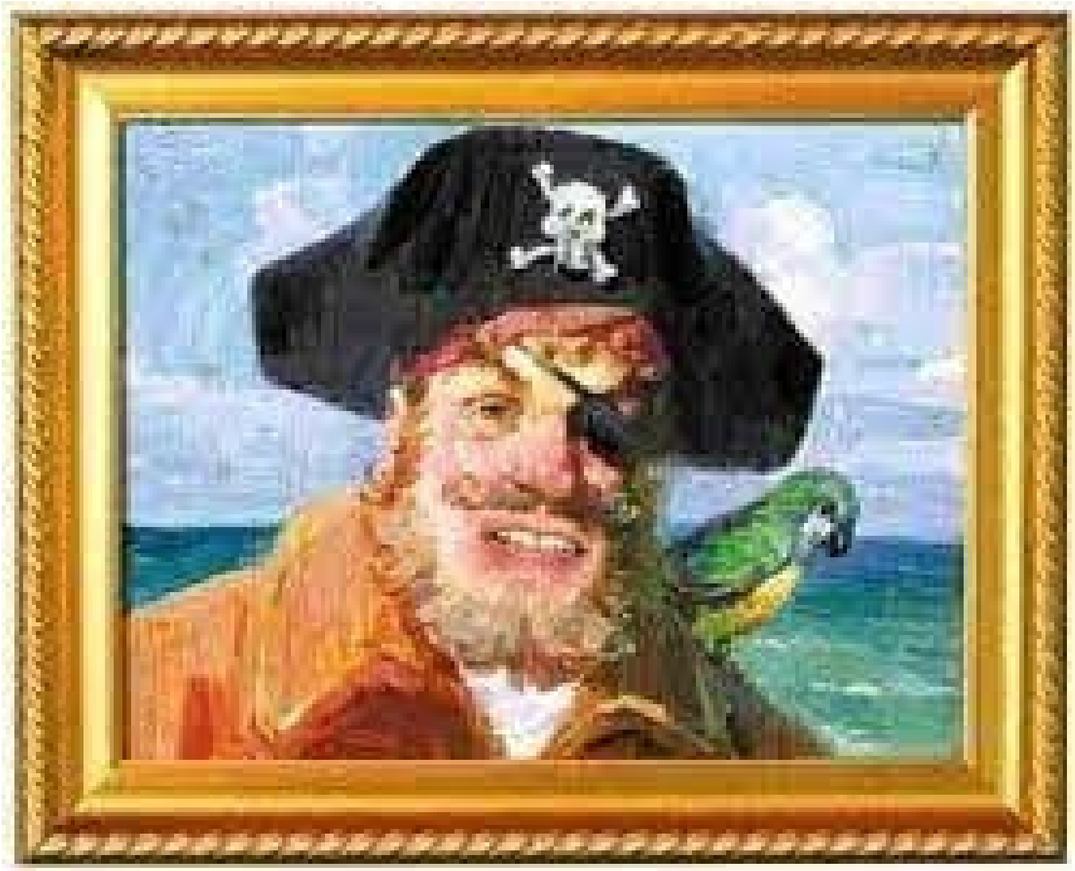
GDG Carthage

THANK YOU for you attention!!



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

QUIZ TIME 😄



GDG Carthage

By Mohammed Arbi Nsibi