

From Words to Wonders: An Introduction to NLP and LLMs

16/10/2024



GDG Carthage

By Mohammed Arbi Nsibi



MOHAMED ARBI NSIBI

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



<https://huggingface.co/Goodnight7>



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



mohammedarbinsibi@gmail.com

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 ?

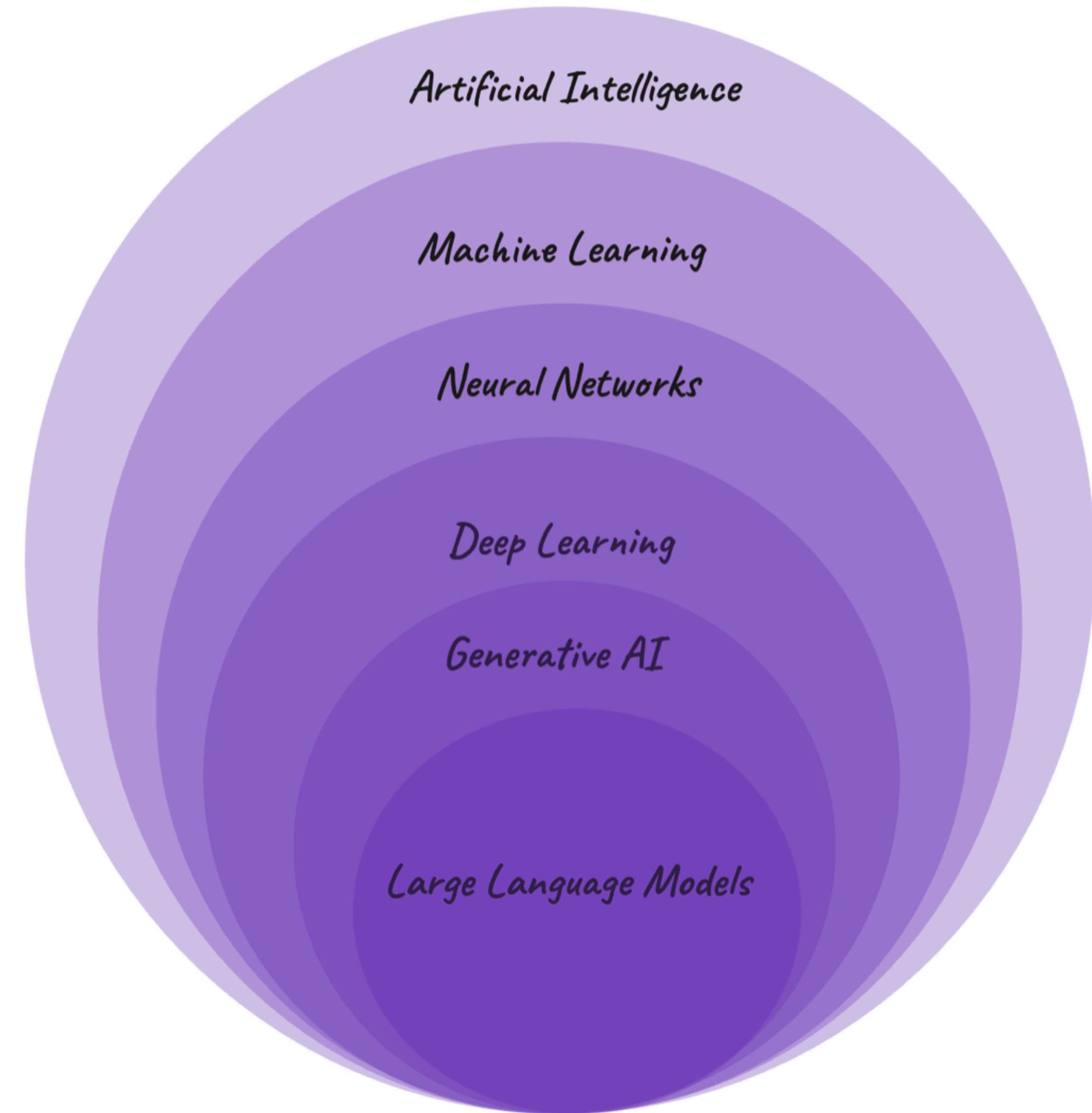
AI is a **branch of computer science** that deals with the creation of intelligent **agents** which are systems that can **reason** and learn (is the development of computer systems able to perform tasks normally requiring **human intelligence**)

What is GenAI ?

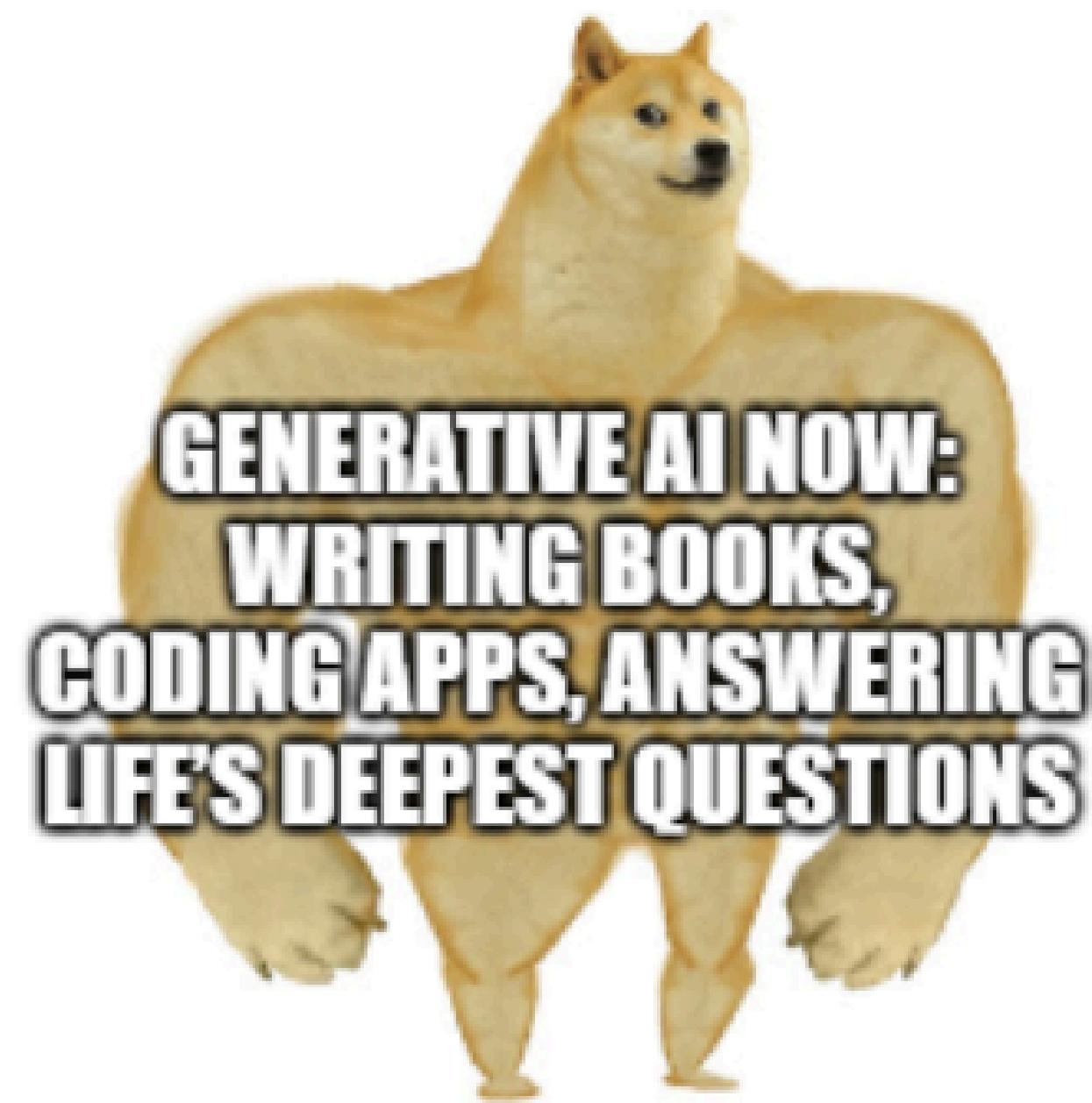
Gen AI is a type of AI technology that can **produce** various types of content , including text, imagery, audio and synthetic data .



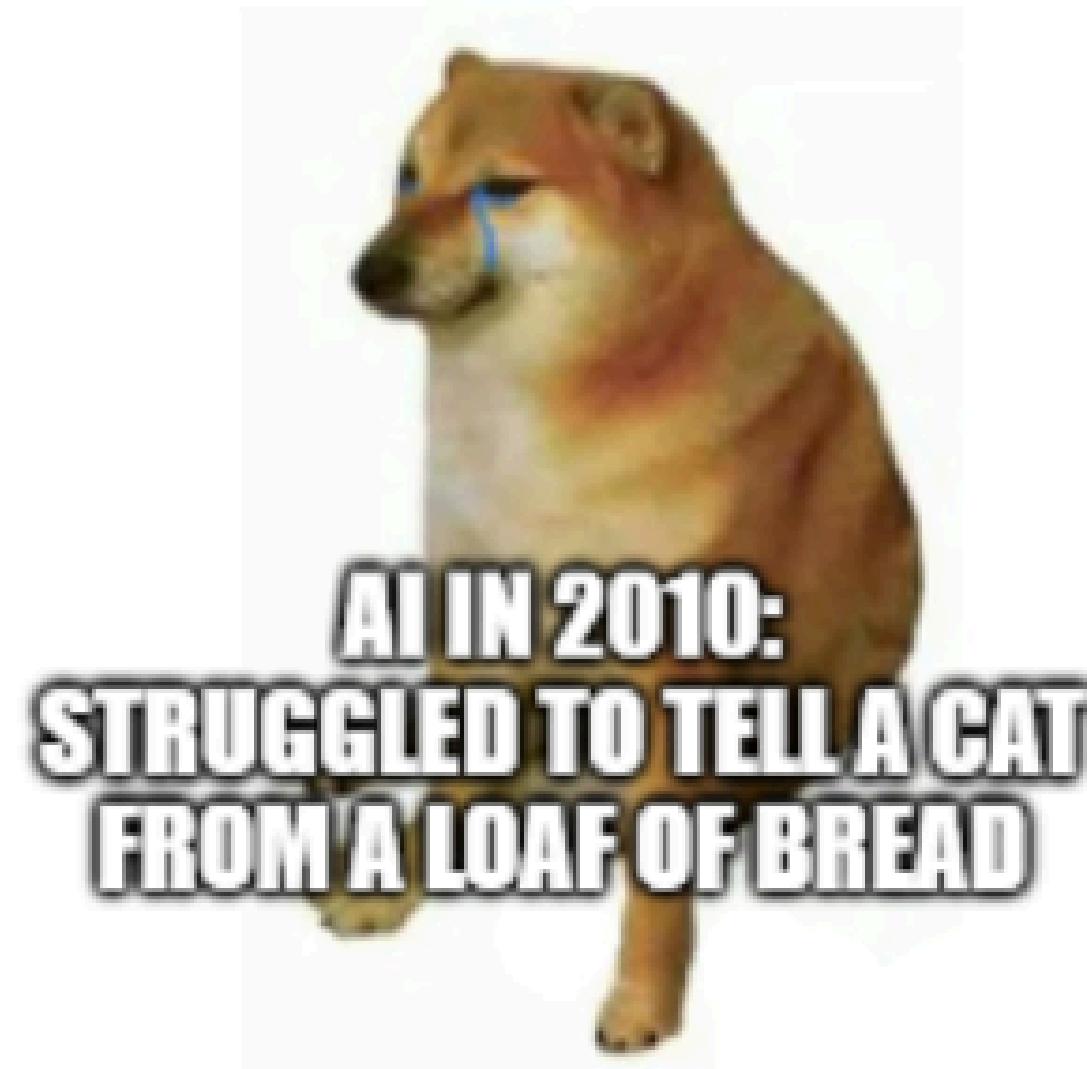
GDG Carthage



GDG Carthage



**GENERATIVE AI NOW:
WRITING BOOKS,
CODING APPS, ANSWERING
LIFE'S DEEPEST QUESTIONS**



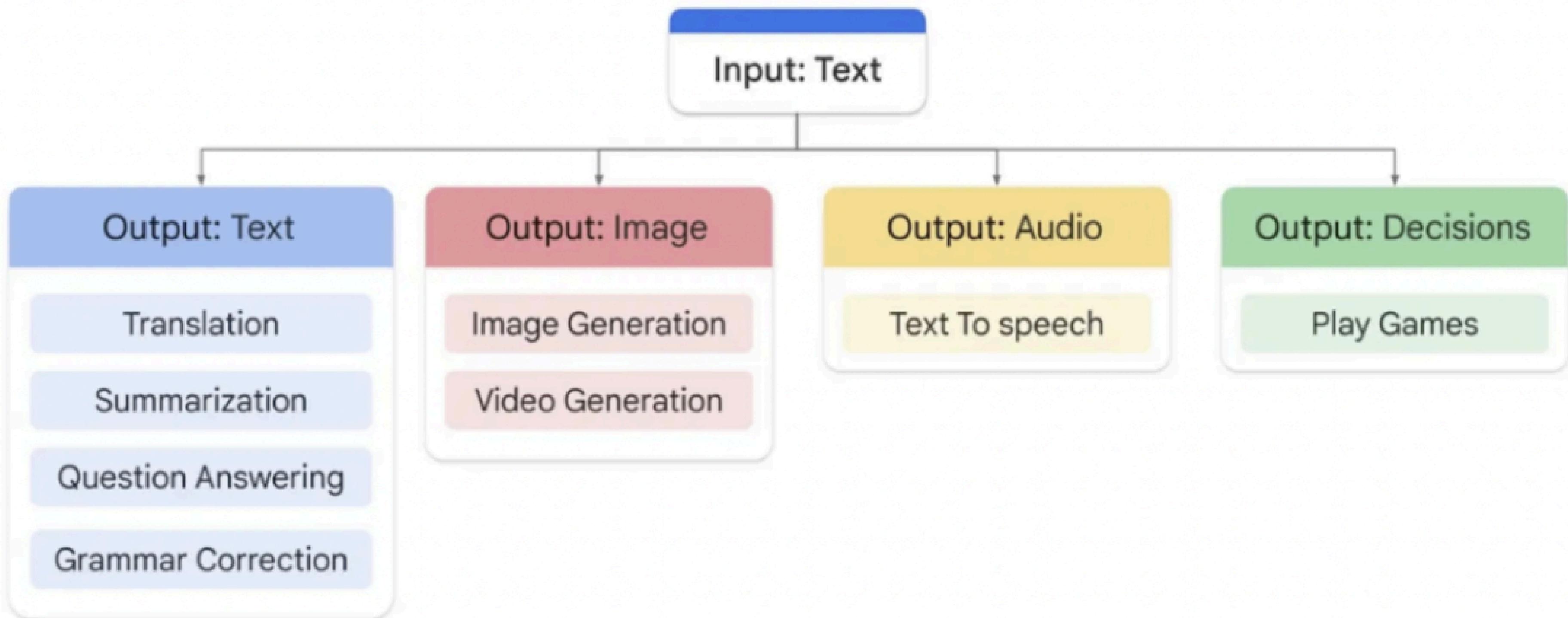
**AI IN 2010:
STRUGGLED TO TELL A CAT
FROM A LOAF OF BREAD**



GDG Carthage

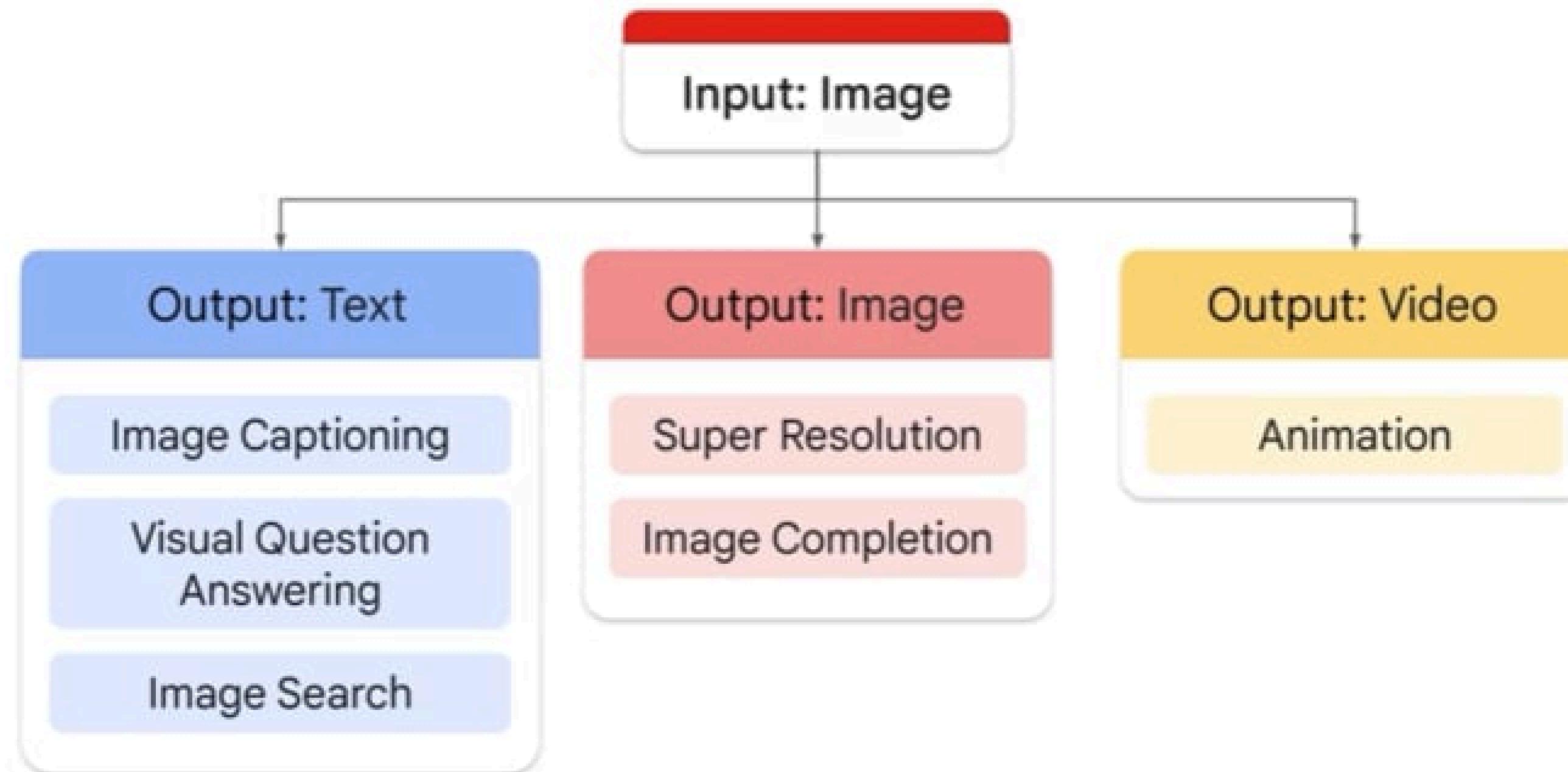
By Mohammed Arbi Nsibi

Types of Generative AI Based on Data

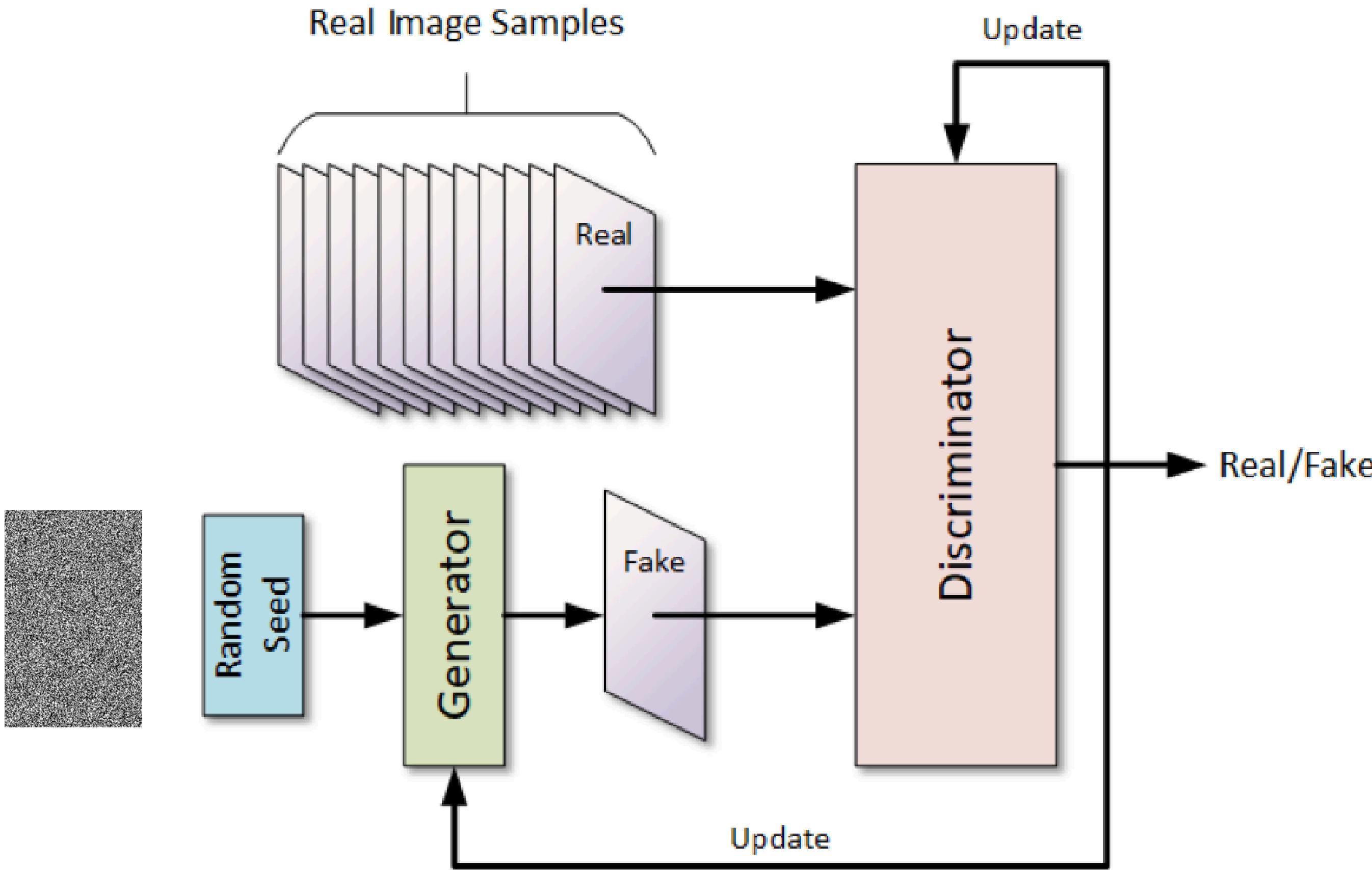


GDG Carthage

Types of Generative AI Based on Data



GANs: Generative adversarial network





Diffusion models :

DALL-E

A wolf running though the woods



Here are two depictions of the wolf running through the woods. Let me know if you'd like any adjustments!

Can you make it more wintery?



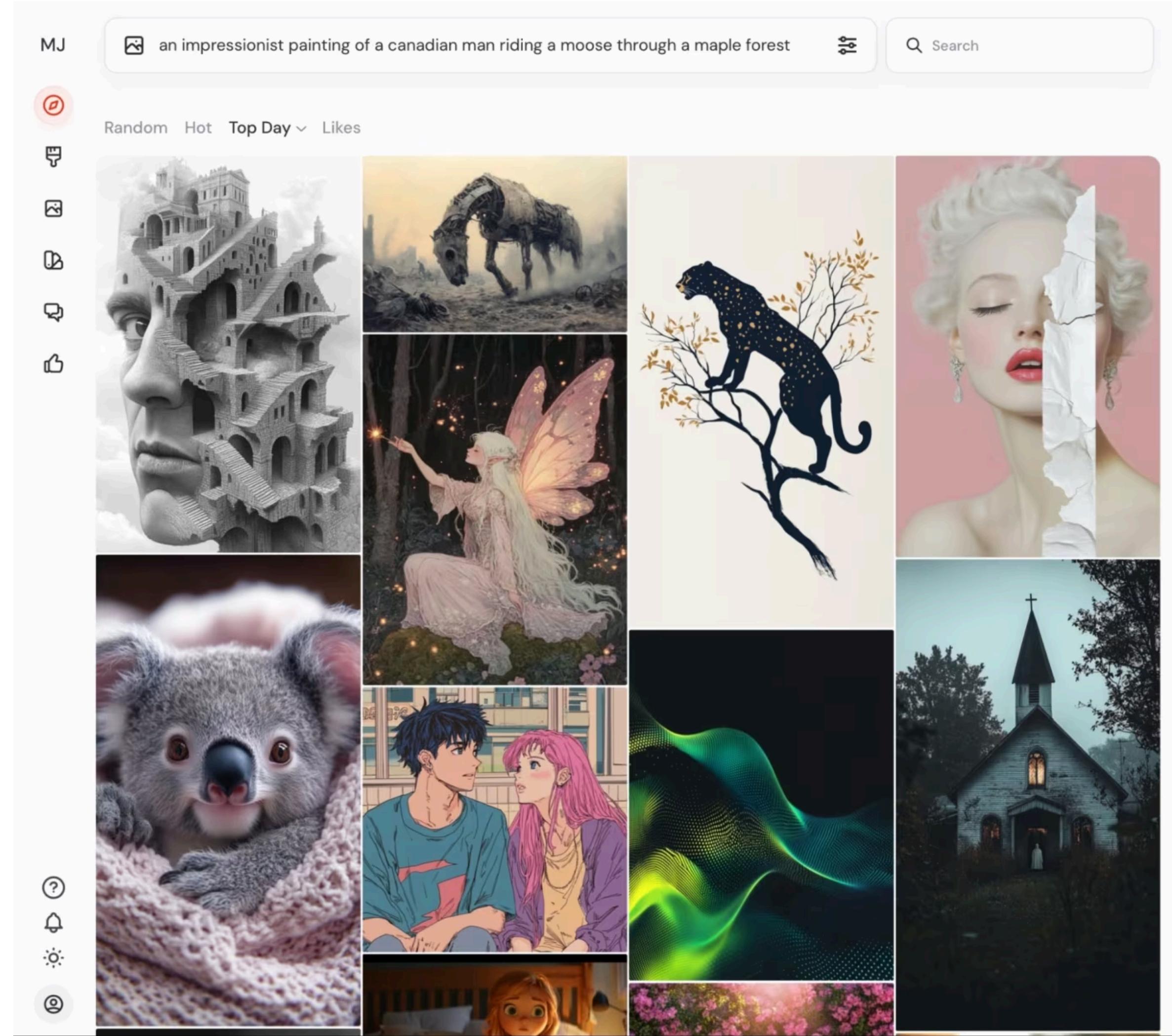
Here are the updated images with a more wintery atmosphere, featuring the wolf running through a snowy forest. Let me know what you think or if you'd like any further adjustments!

<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 Doc 1.5 Warning: Free Research Preview. Our goal is to make AI systems more natural and able to interact with. Your feedback will help us improve.



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/>



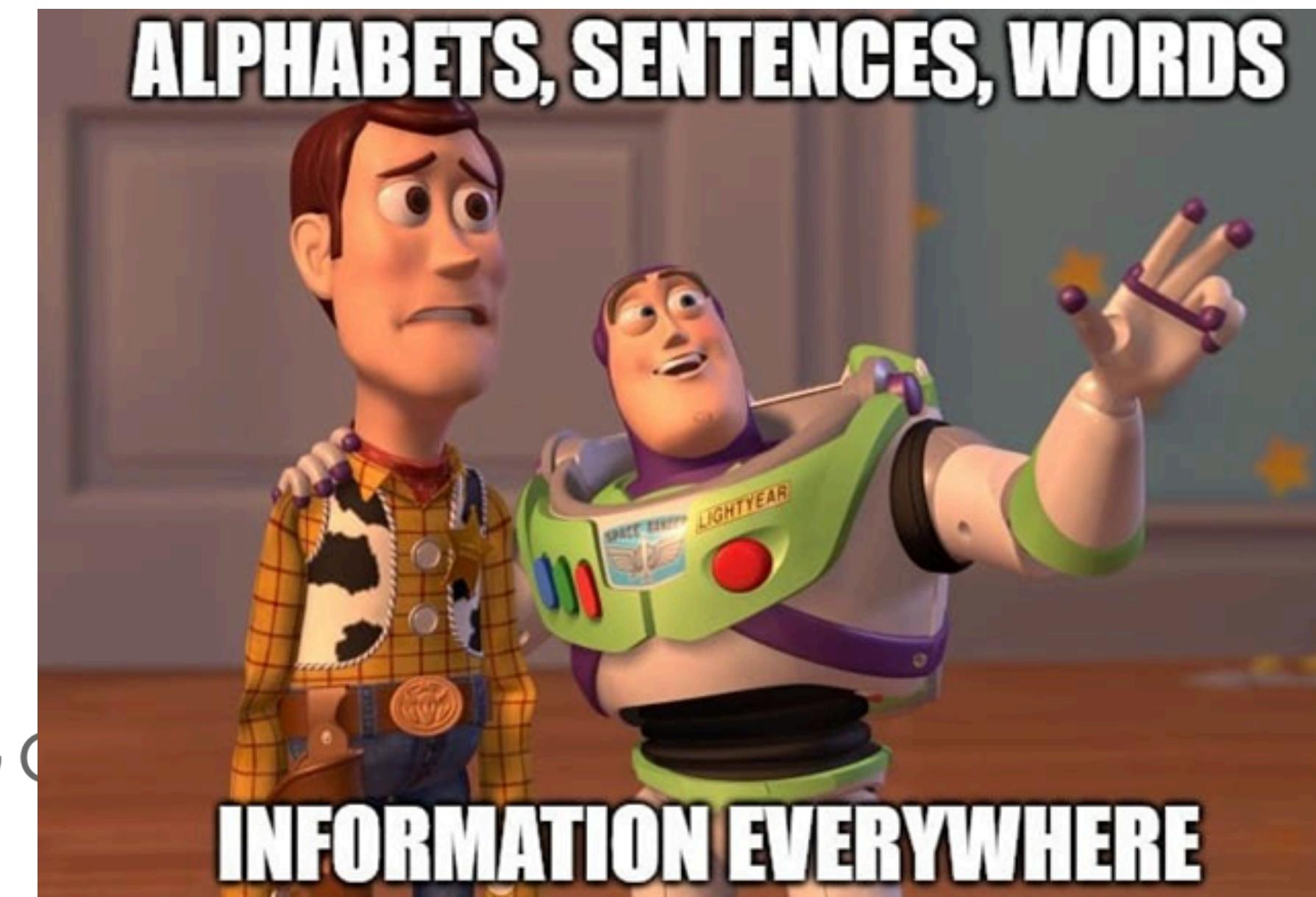
WHAT IS NLP ?

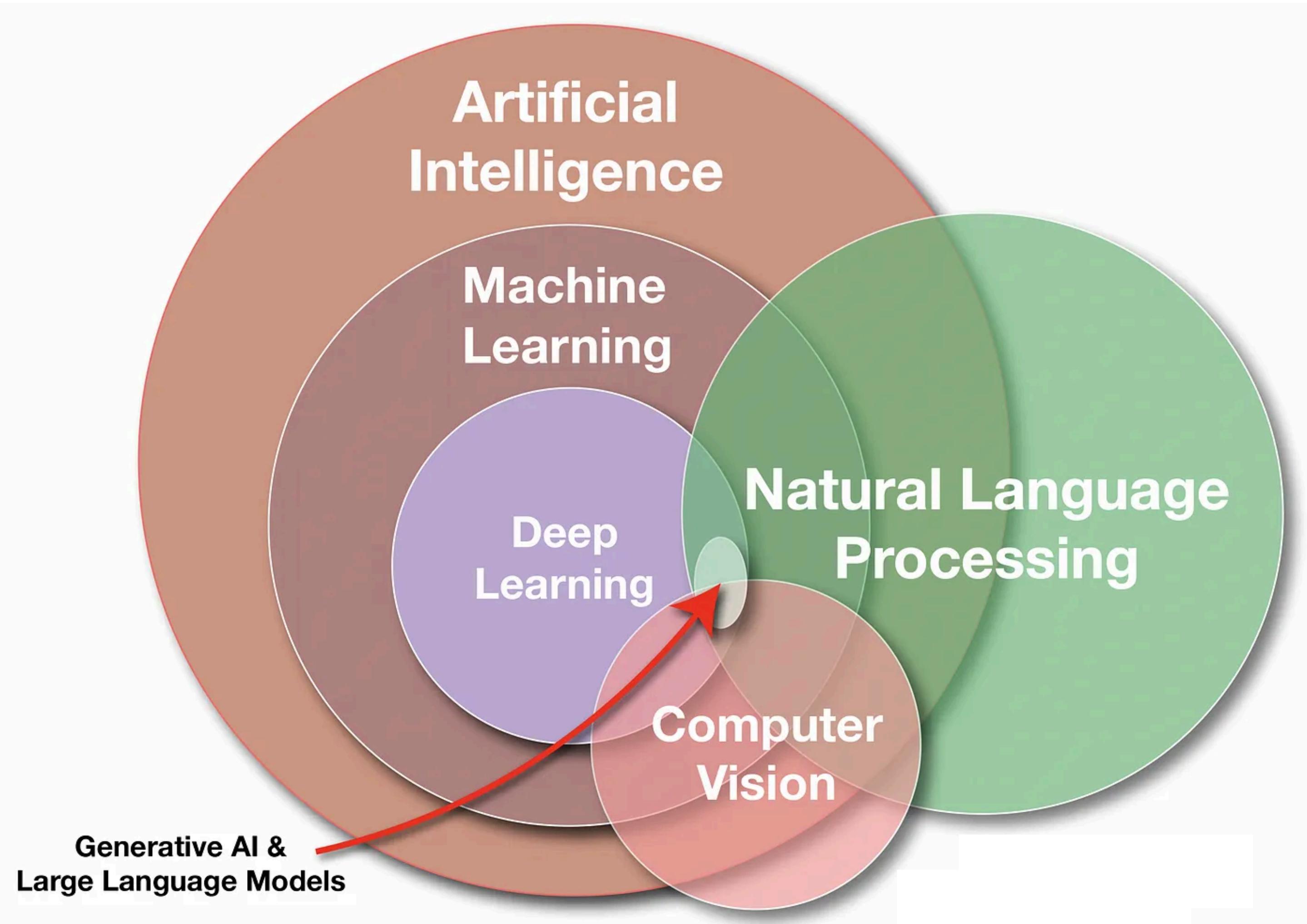


GDG Carthage

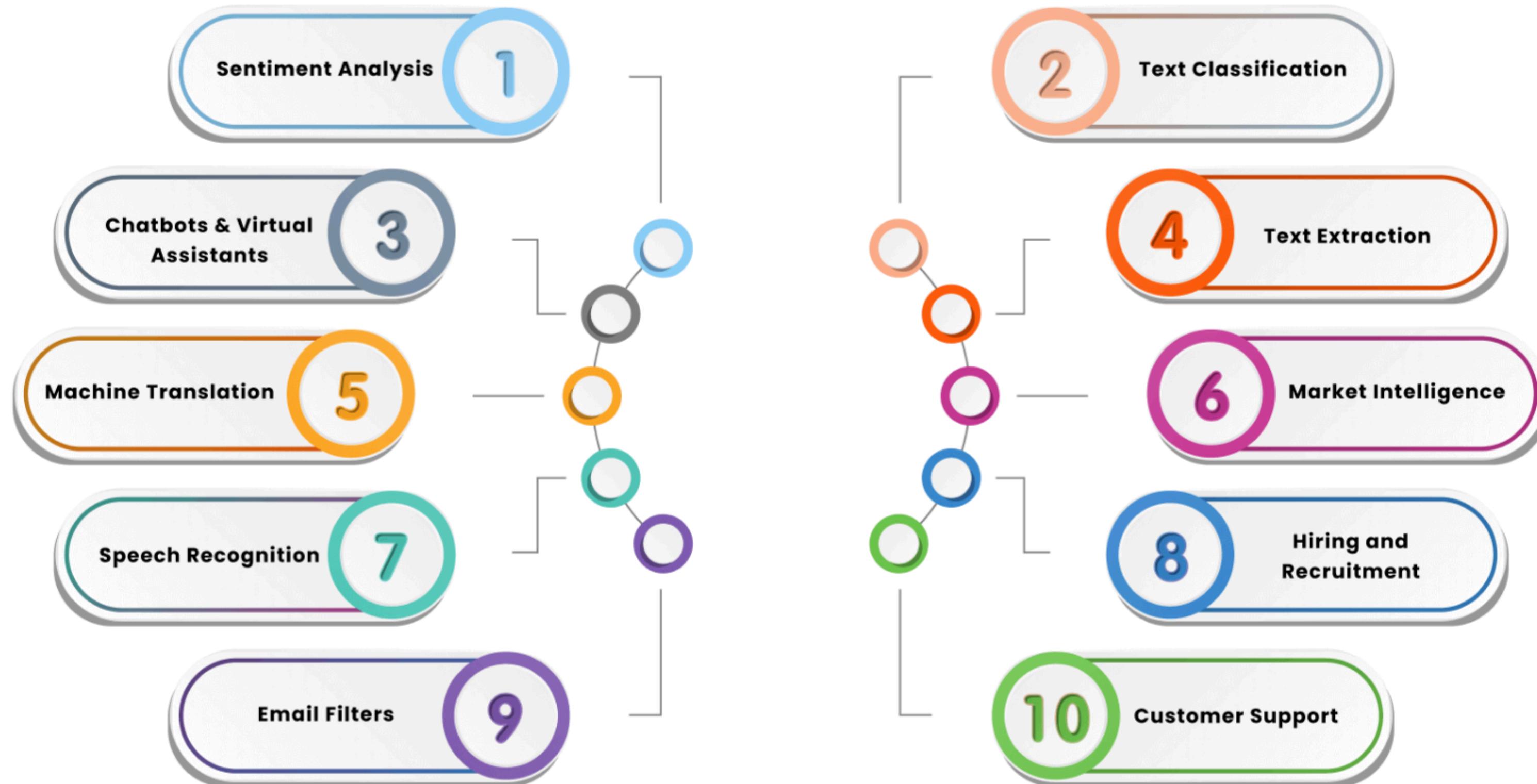
WHAT IS NLP ?

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.



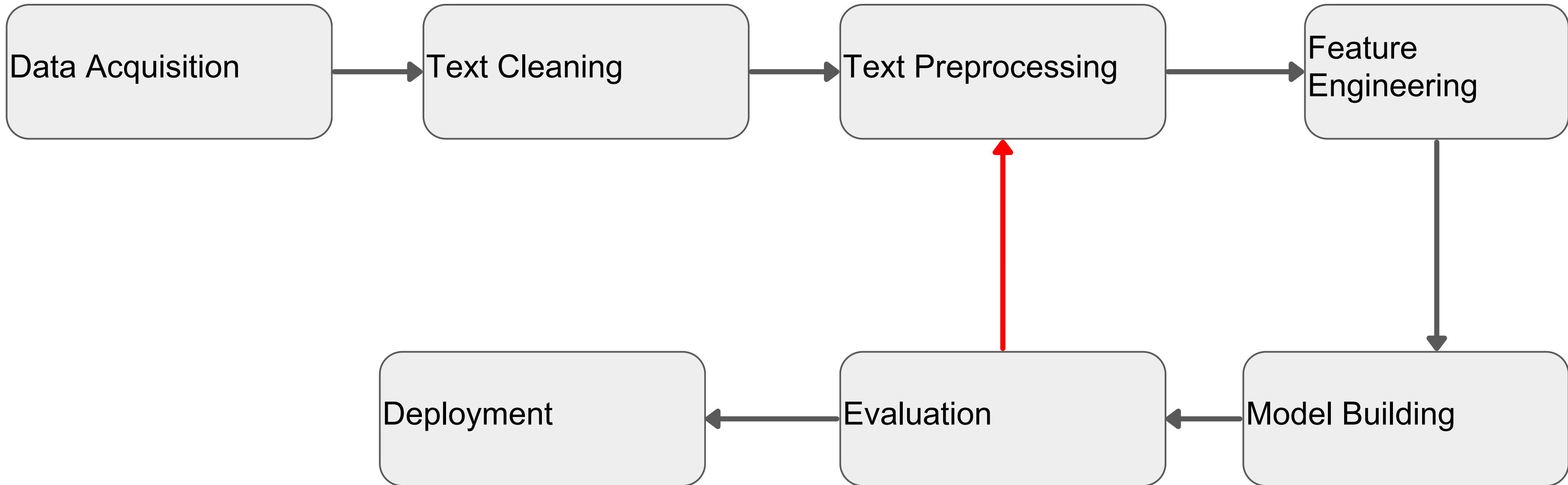


NLP applications



GDG Carthage

NLP Pipeline



GDG Carthage



By Mohammed Arbi Nsibi

Text Cleaning

Regex or Regular Expression

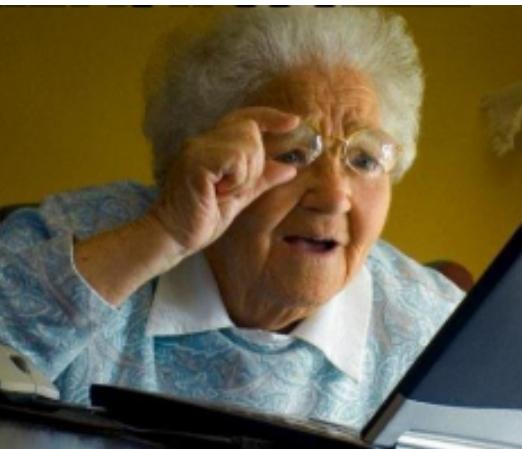
Spelling corrections

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)

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

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

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

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


GDG Carthage

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

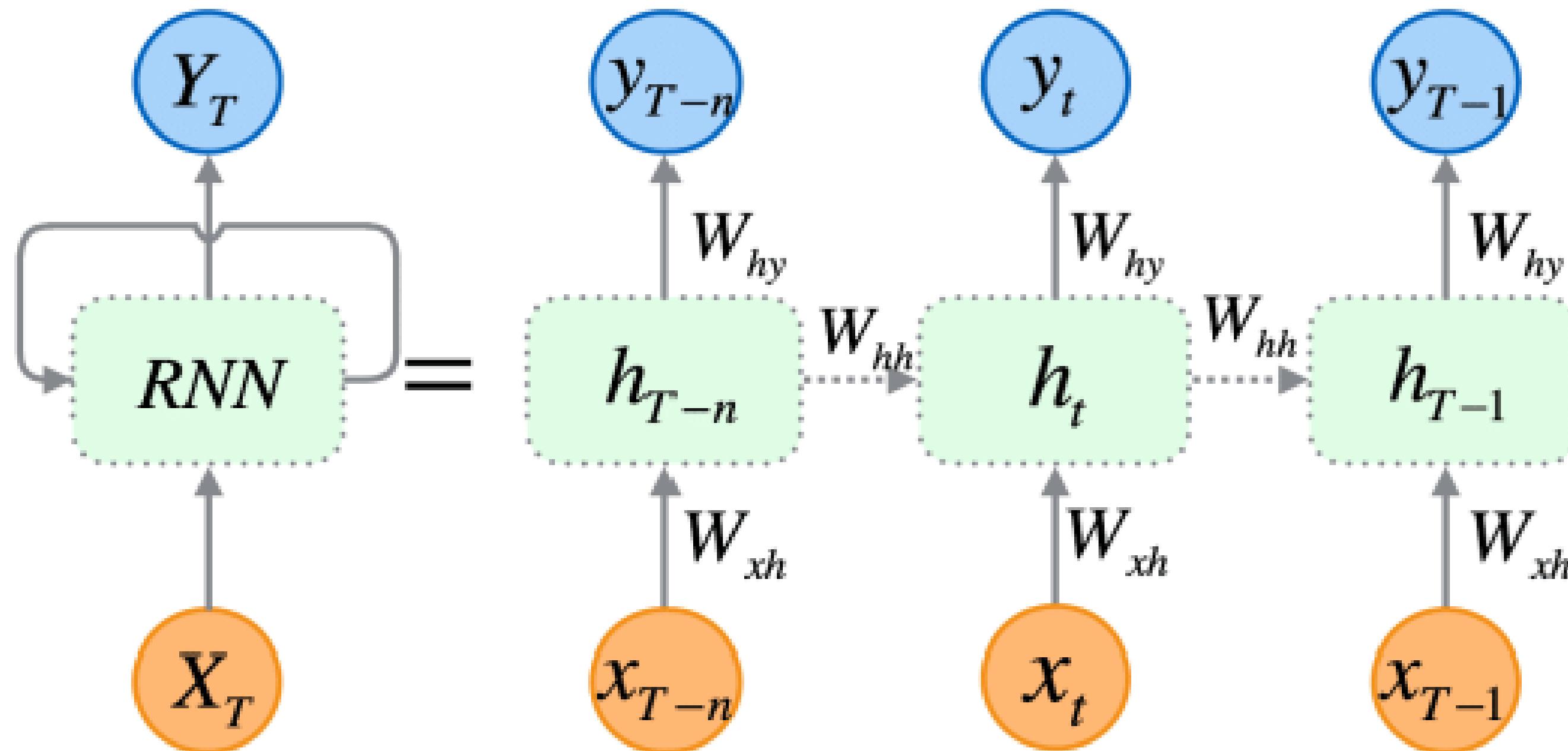
By Mohammed Arbi Nsibi

LET'S CODE



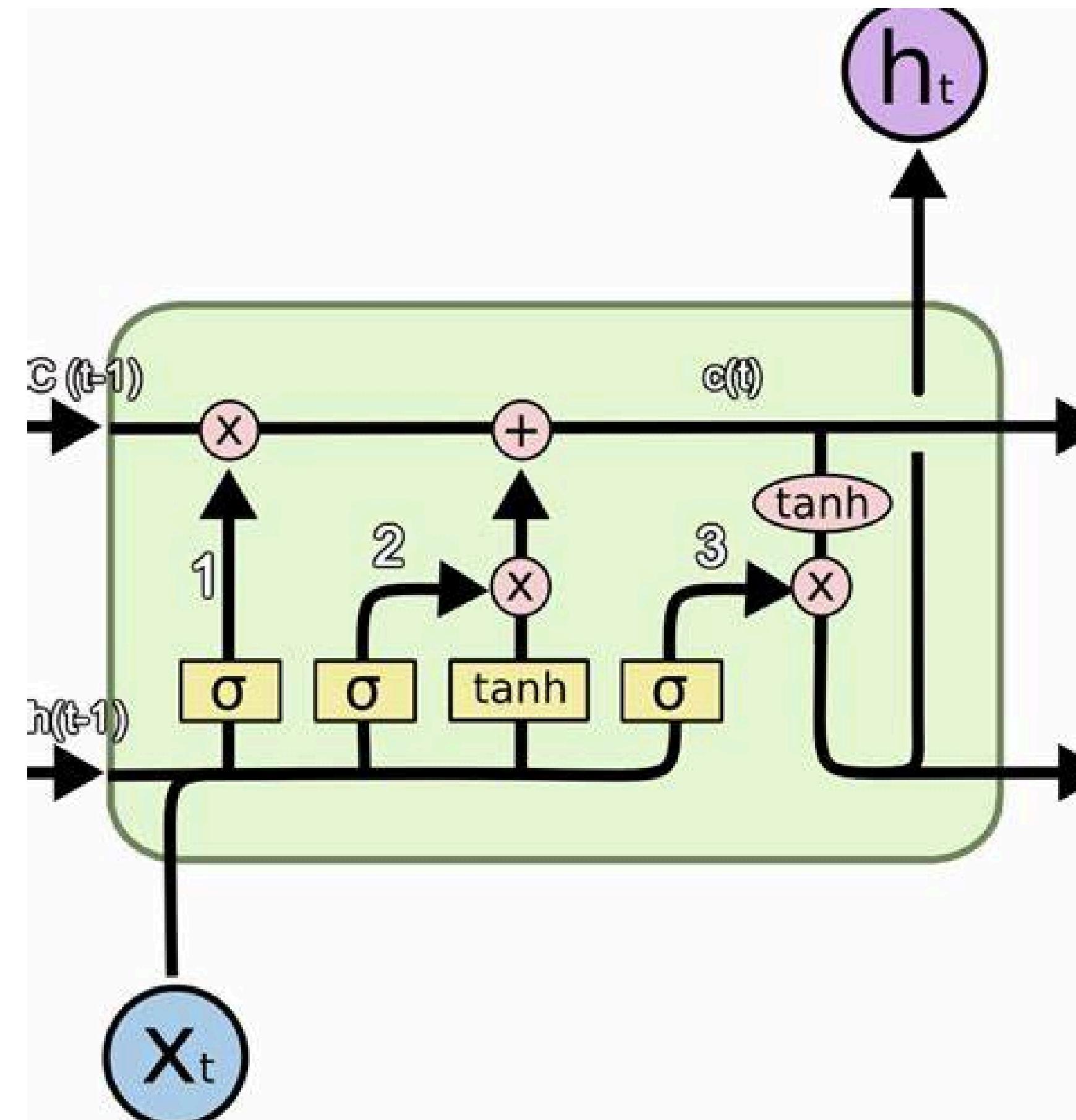
GDG Carthage

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**



Problem: 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.

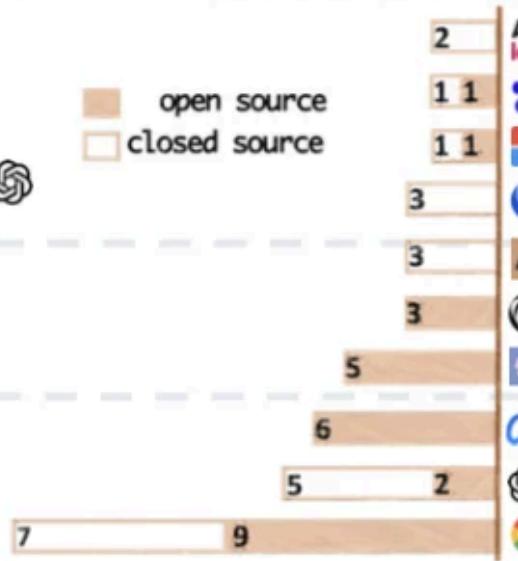
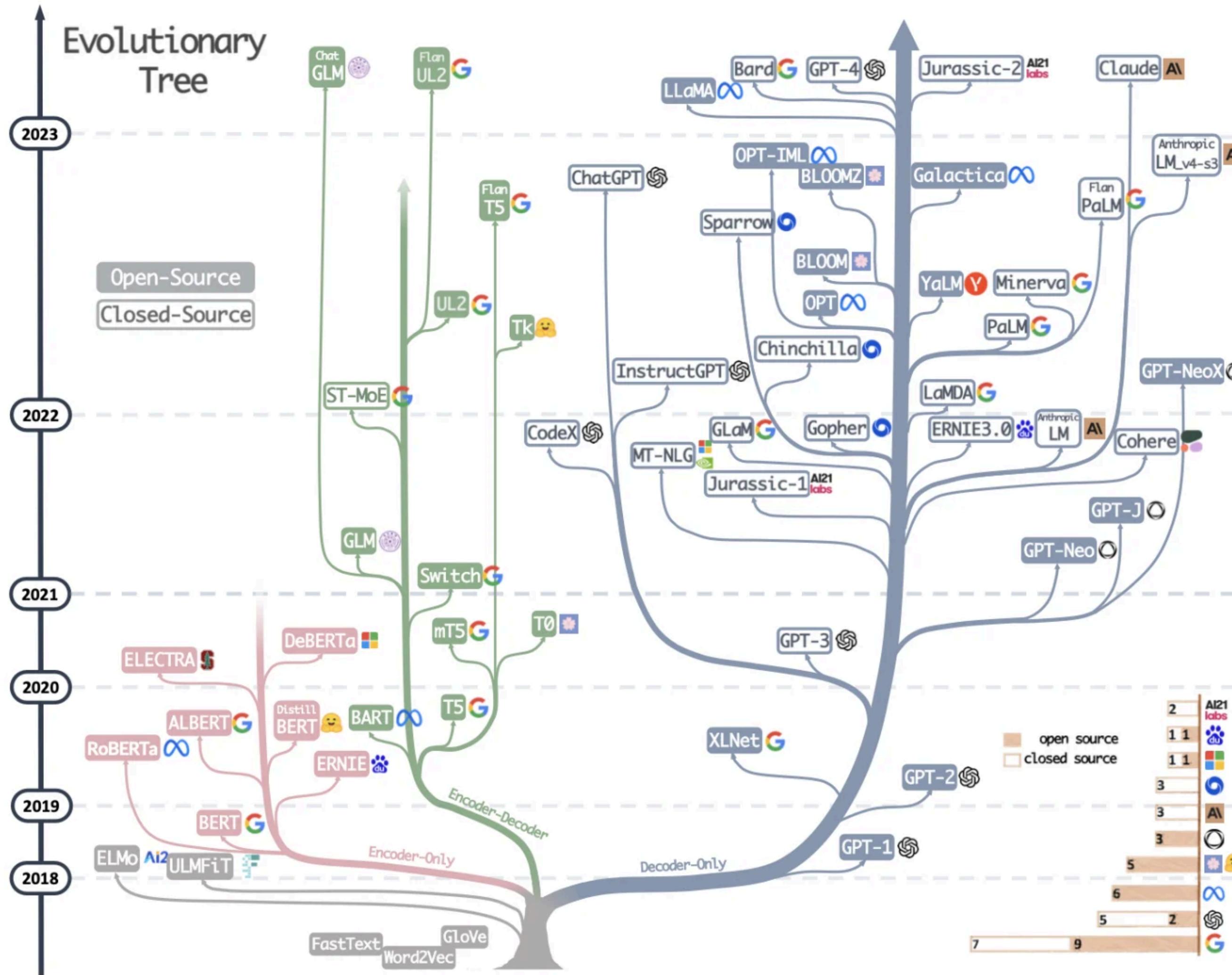


NLP vs LLM



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

GPT-2

124M to 1.5B

2023

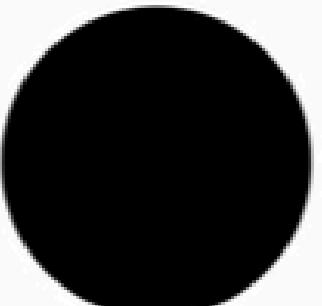
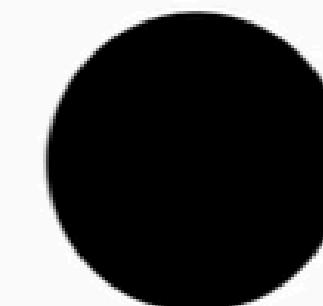
Llama-1

7B to 65B

2023

Llama-2

7B to 70B



GDG Carthage

Model size

2019

GPT-2

124M to 1.5B

2023

Llama-1

7B to 65B

2023

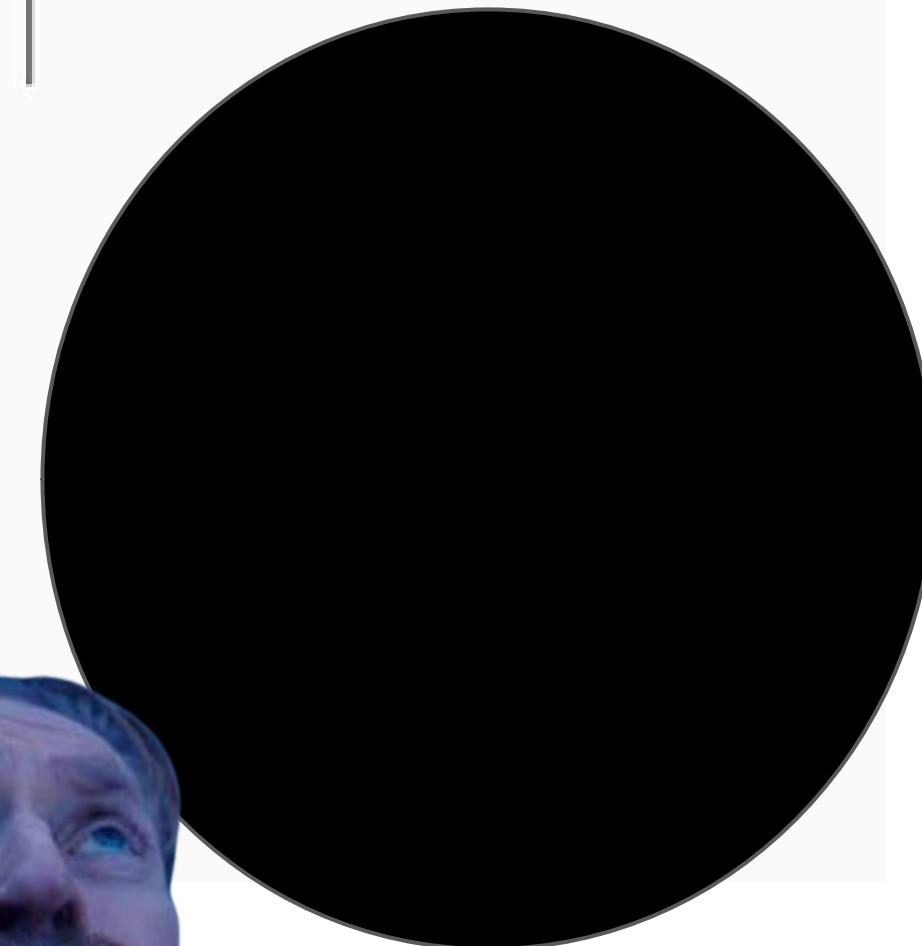
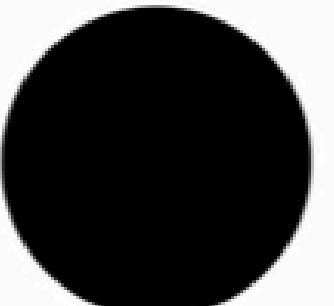
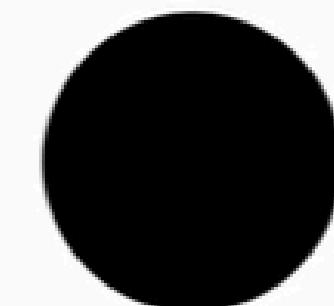
Llama-2

7B to 70B

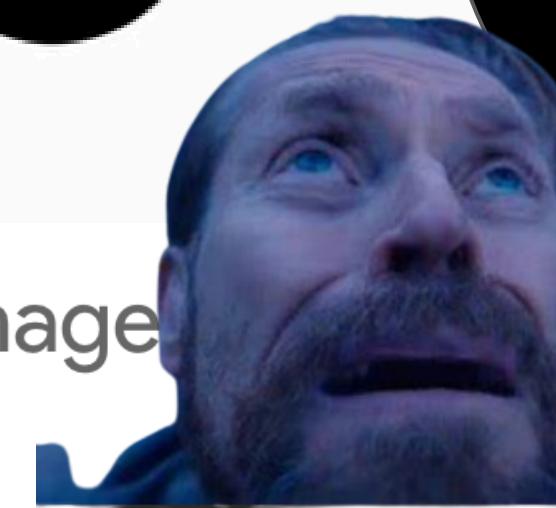
2024

Llama-3

8B to 405B



GDG Carthage



Dataset

2019

GPT-2

40B tokens

2023

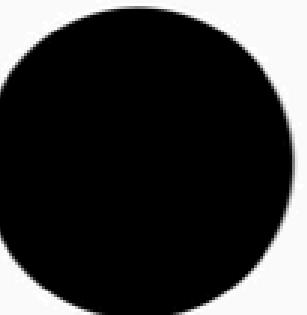
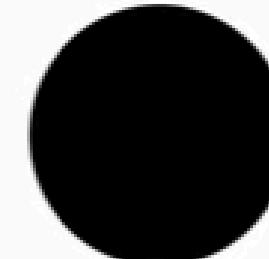
Llama-1

1.4T tokens

2023

Llama-2

2T tokens



GDG Carthage

Dataset

2019

GPT-2

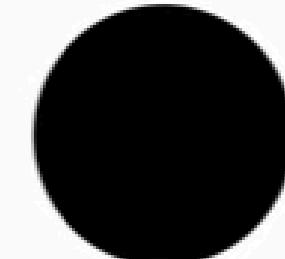
40B tokens



2023

Llama-1

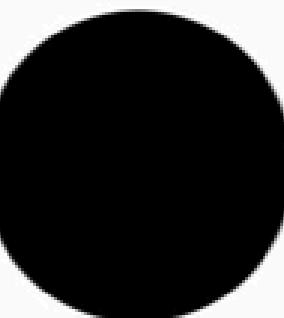
1.4T tokens



2023

Llama-2

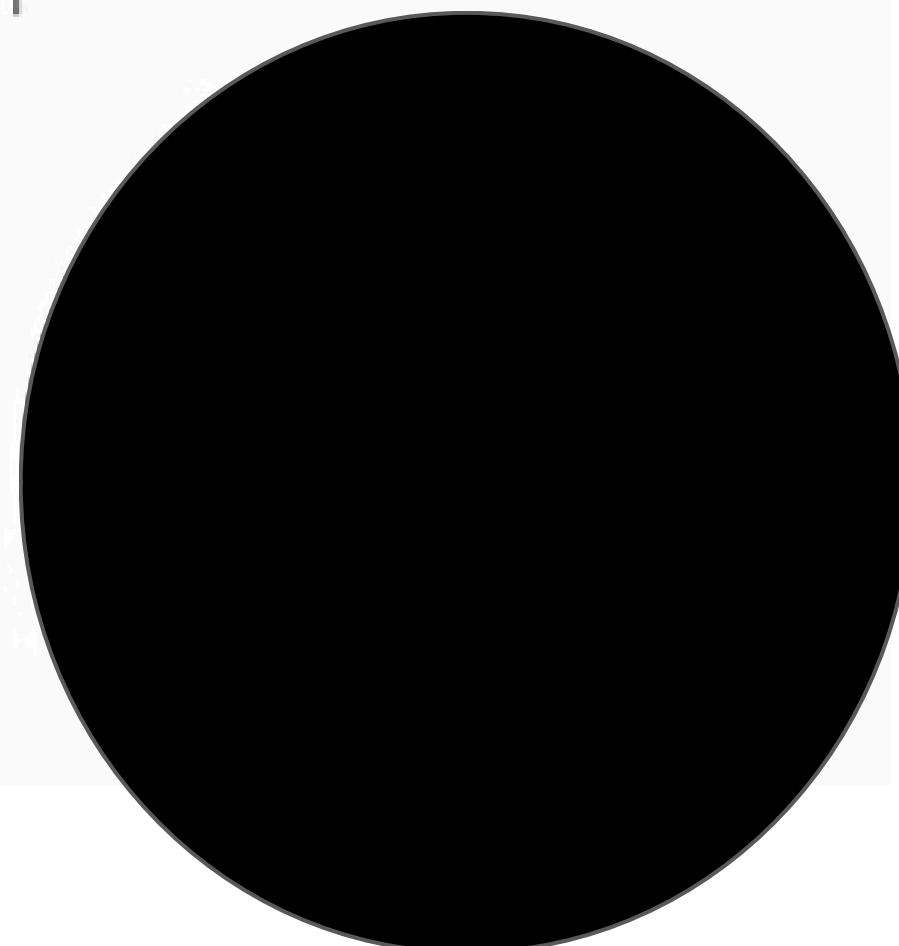
2T tokens



2024

Llama-3

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

Lukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

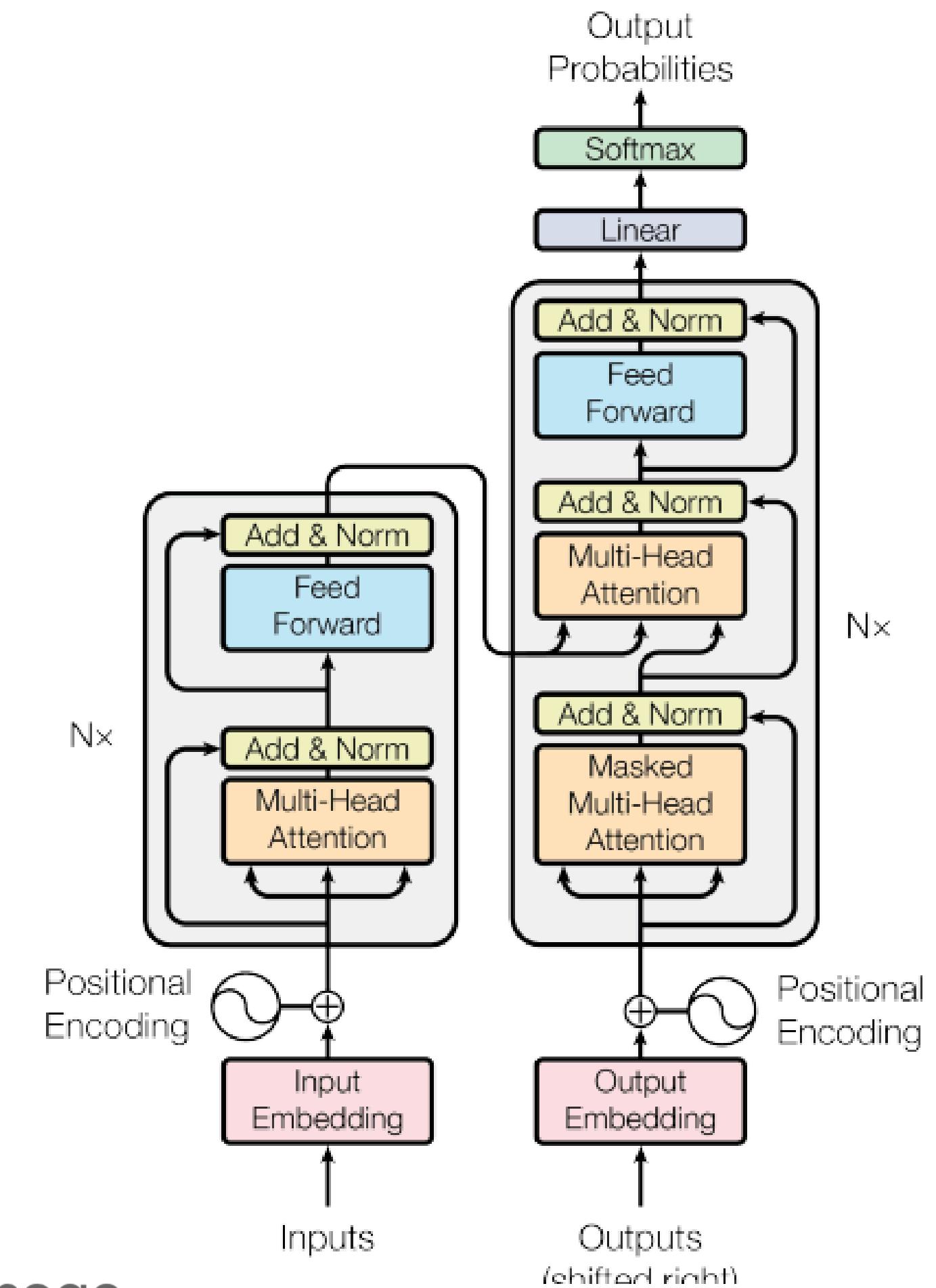
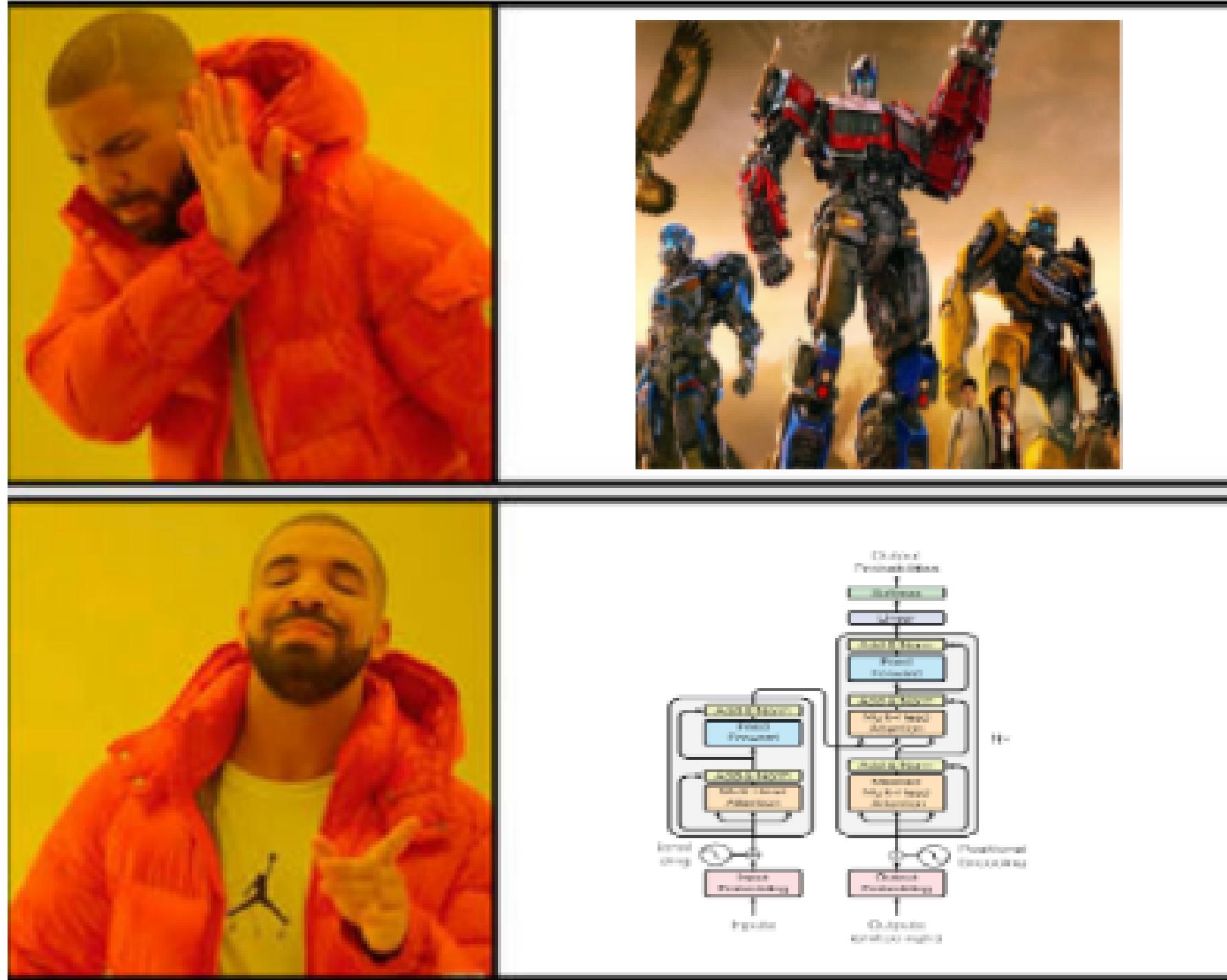
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, 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	$\sin(1/10000^0) \approx 0.8415$	$\cos(1/10000^0) \approx 0.5403$	$\sin(1/10000^0) \approx 0.8415$	$\cos(1/10000^0) \approx 0.5403$
2	$\sin(2/10000^0) \approx 0.9093$	$\cos(2/10000^0) \approx -0.4161$	$\sin(2/10000^0) \approx 0.9093$	$\cos(2/10000^0) \approx -0.4161$

$$\textcolor{brown}{E} = \textcolor{blue}{X} + \textcolor{brown}{PE}$$



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

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



GDG Carthage

Where to find pretrained LLMs ?



Hugging Face



GDG Carthage

By Mohammed Arbi Nsibi

Where to find pretrained LLMs ?

Models 1,028,261 Full-text search

 [openai/whisper-large-v3-turbo](#)
Automatic Speech Recognition • Updated 1 day ago • ↓ 10k • ⚡ • ❤ 324

 [black-forest-labs/FLUX.1-dev](#)
Text-to-Image • Updated Aug 16 • ↓ 1.14M • ⚡ • ❤ 5.03k

 [jasperai/Flux.1-dev-Controlnet-Upscaler](#)
Image-to-Image • Updated 3 days ago • ↓ 9.86k • ❤ 244

 [allenai/Molmo-7B-D-0924](#)
Image-Text-to-Text • Updated 1 day ago • ↓ 14.5k • ❤ 273

 [meta-llama/Llama-3.2-11B-Vision-Instruct](#)
Image-Text-to-Text • Updated 4 days ago • ↓ 139k • ⚡ • ❤ 479

 [nvidia/NVLM-D-72B](#)
Image-Text-to-Text • Updated about 18 hours ago • ↓ 860 • ❤ 242

 [meta-llama/Llama-3.2-1B](#)
Text Generation • Updated 3 days ago • ↓ 61.2k • ⚡ • ❤ 299

 [openbmb/MiniCPM-Embedding](#)
Feature Extraction • Updated 2 days ago • ↓ 130k • ❤ 204

Datasets 222,500 Full-text search

 [google/frames-benchmark](#)
Viewer • Updated about 17 hours ago • ↓ 824 • ↓ 562 • ❤ 122

 [FBK-MT/mosel](#)
Viewer • Updated 5 days ago • ↓ 51.1M • ↓ 21 • ❤ 42

 [openai/MMMLU](#)
Viewer • Updated 4 days ago • ↓ 393k • ↓ 5.33k • ❤ 374

 [argilla/FinePersonas-v0.1](#)
Viewer • Updated 19 days ago • ↓ 21.1M • ↓ 371 • ❤ 304

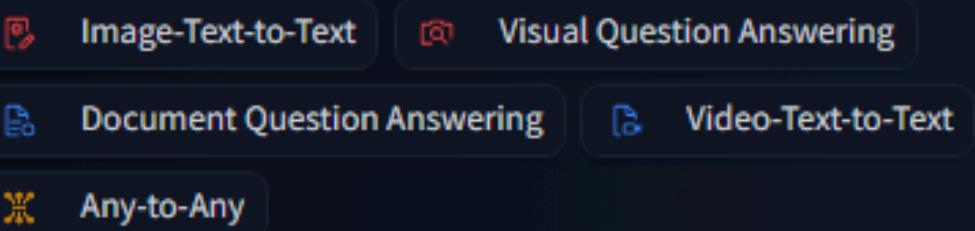
 [fka/awesome-chatgpt-prompts](#)
Viewer • Updated Sep 3 • ↓ 170 • ↓ 8.36k • ❤ 5.82k

 [migtissera/Synthia-v1.5-I](#)
Viewer • Updated 8 days ago • ↓ 20.7k • ↓ 99 • ❤ 39

 [Hacker Noon/where-startups-trend](#)
Preview • Updated 7 days ago • ↓ 19 • ❤ 36

 [k-mktr/improved-flux-prompts-photoreal-portrait](#)
Viewer • Updated 4 days ago • ↓ 20k • ↓ 54 • ❤ 62

Model types



Text-to-text

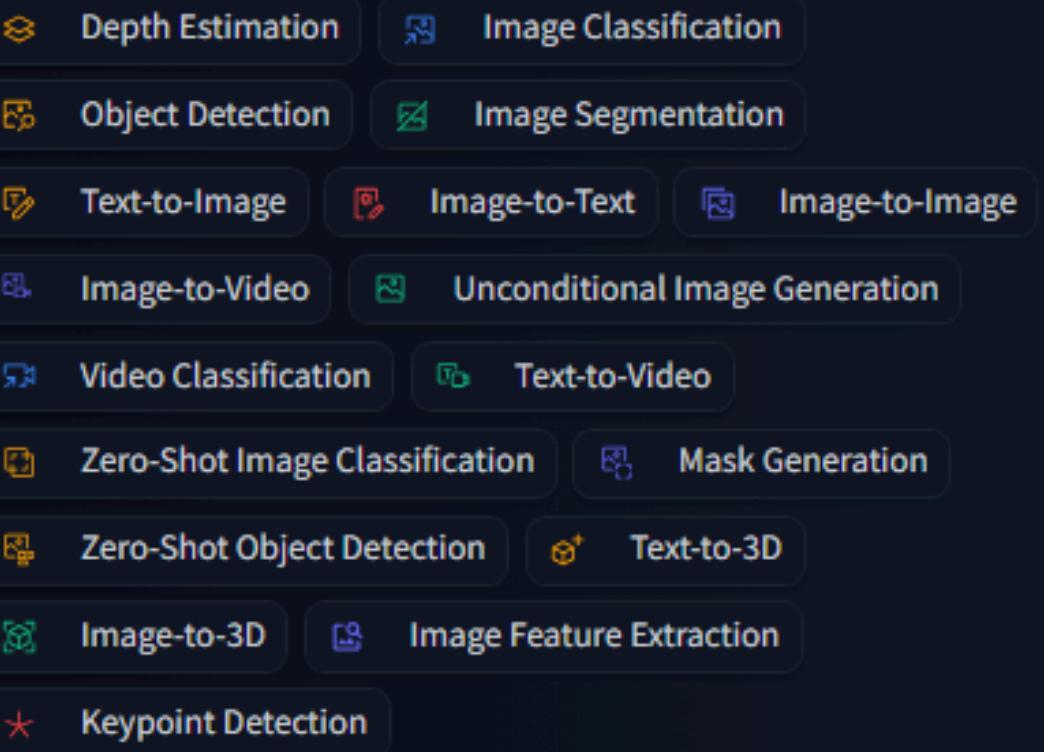
Text-to-image

Text-to-video

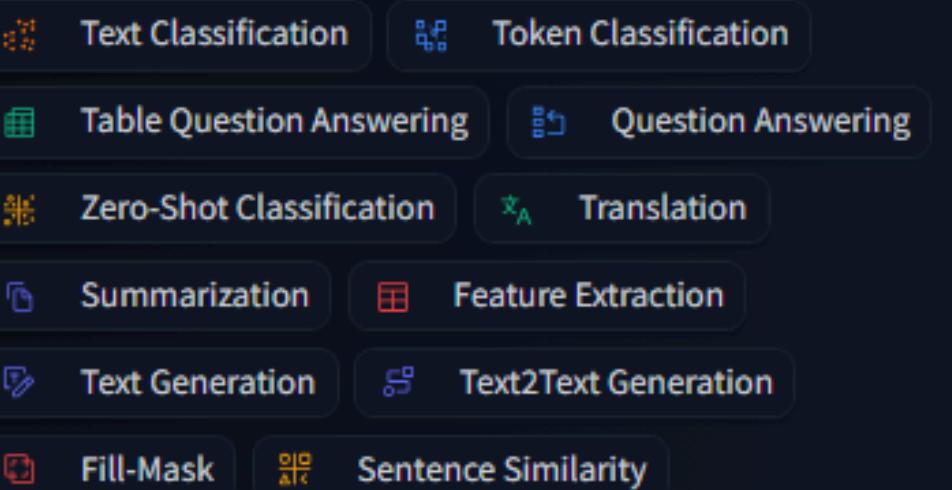
Sentence-similarity



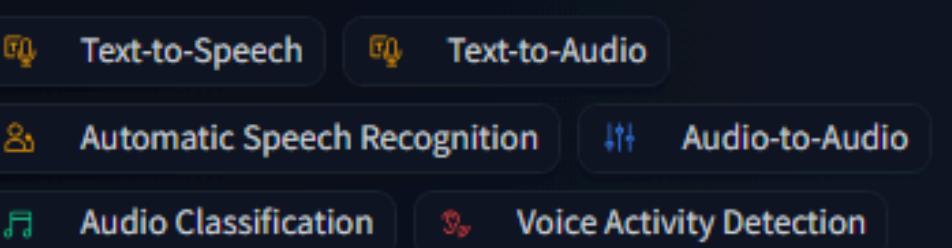
Computer Vision



Natural Language Processing



Audio



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.

LLM AFTER TRAINING
ON 90% OF THE INTERNET...



TIME TO HALLUCINATE SOME FACTS!



GDG Carthage

By Mohammed Arbi Nsibi

Hallucinations

Is 9677 a prime number?

No, 9677 is not a prime number.
It can be factored into 13 and 745, as $9677 = 13 \times 745$.

Is 9677 divisible by 13?

No

X

in a separate session,
GPT-4 recognizes its
claim as incorrect!

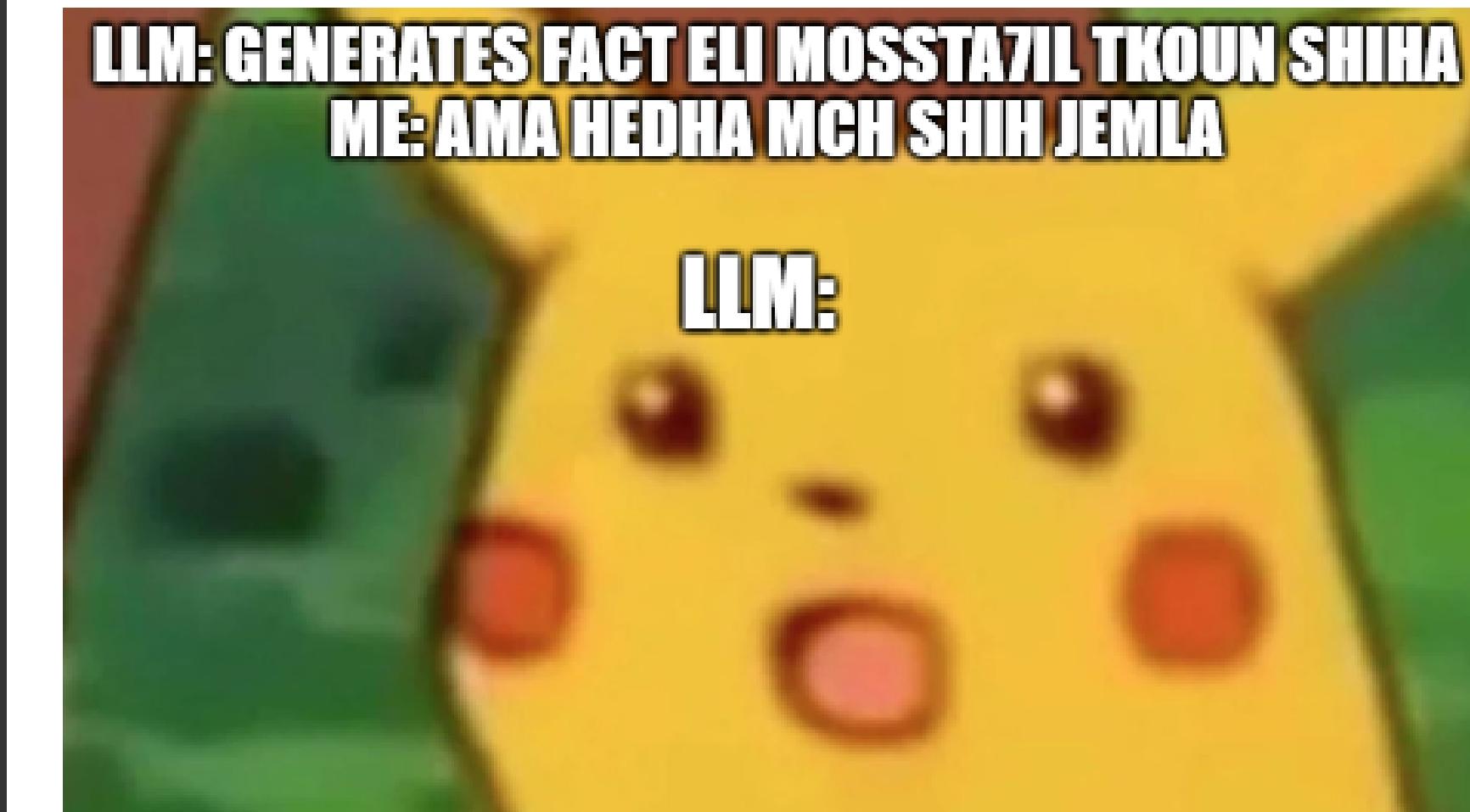
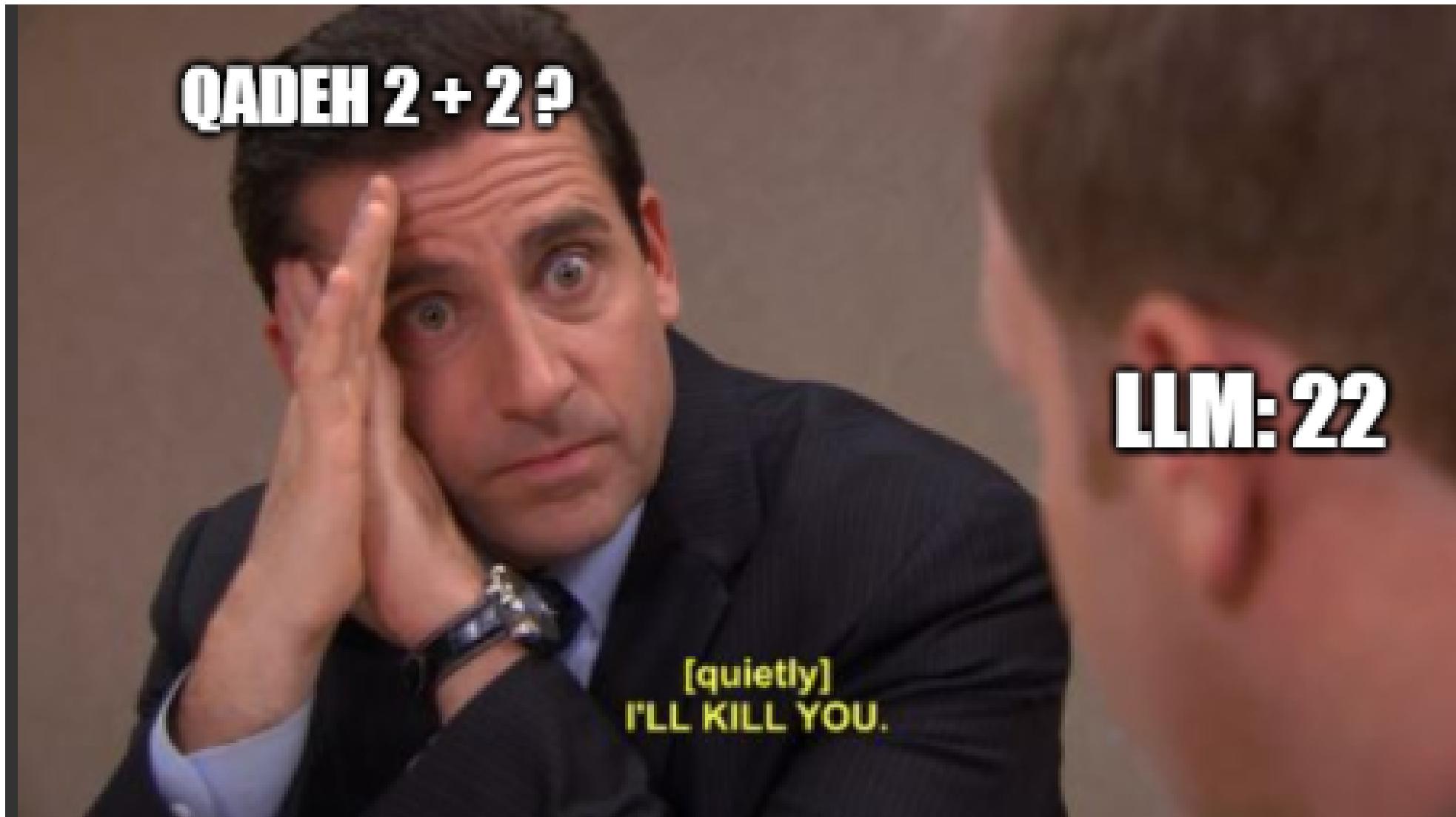
✓

} incorrect assertion
} snowballed
} hallucination

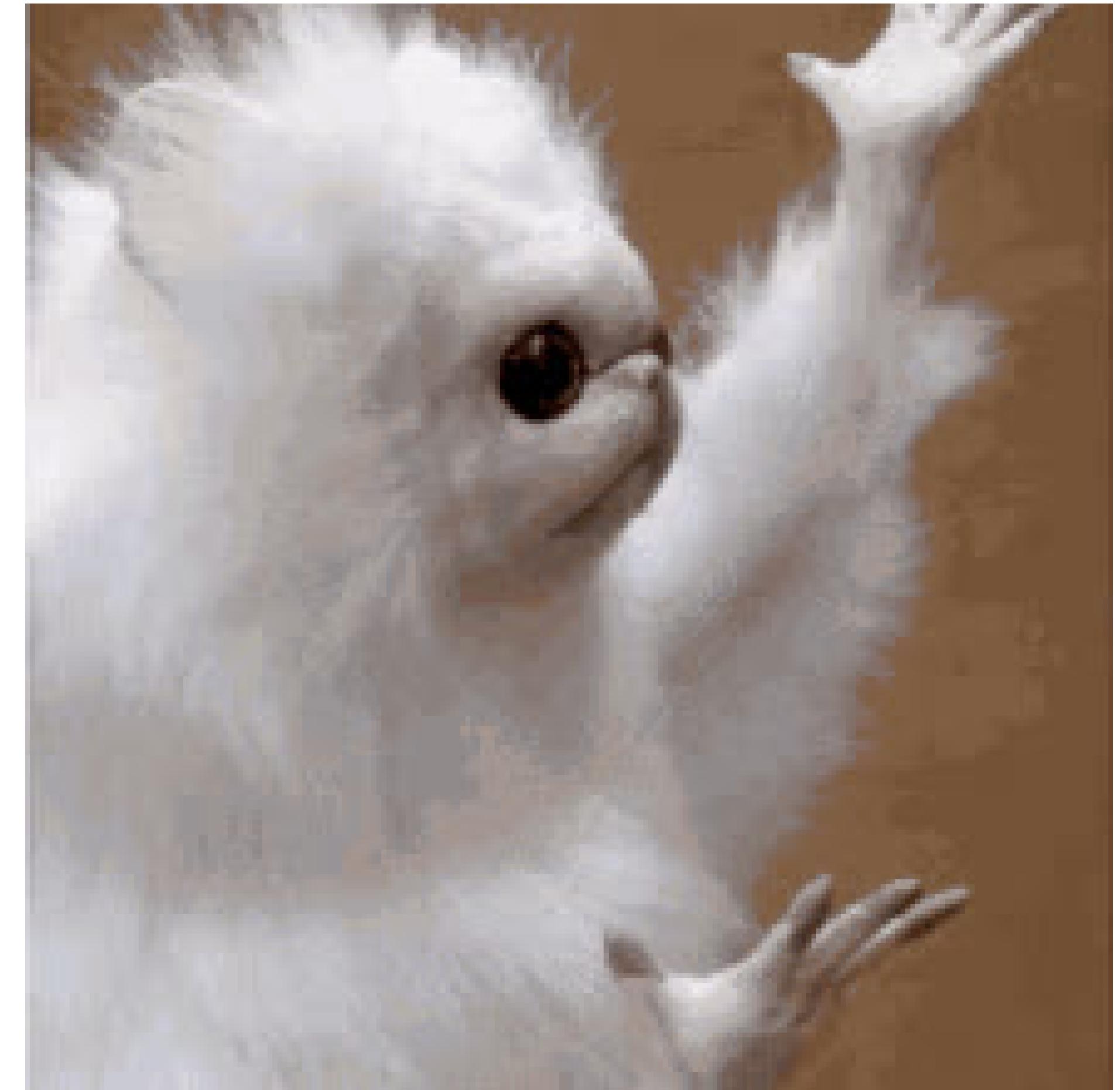


GDG Carthage

Hallucinations



Solutions ?

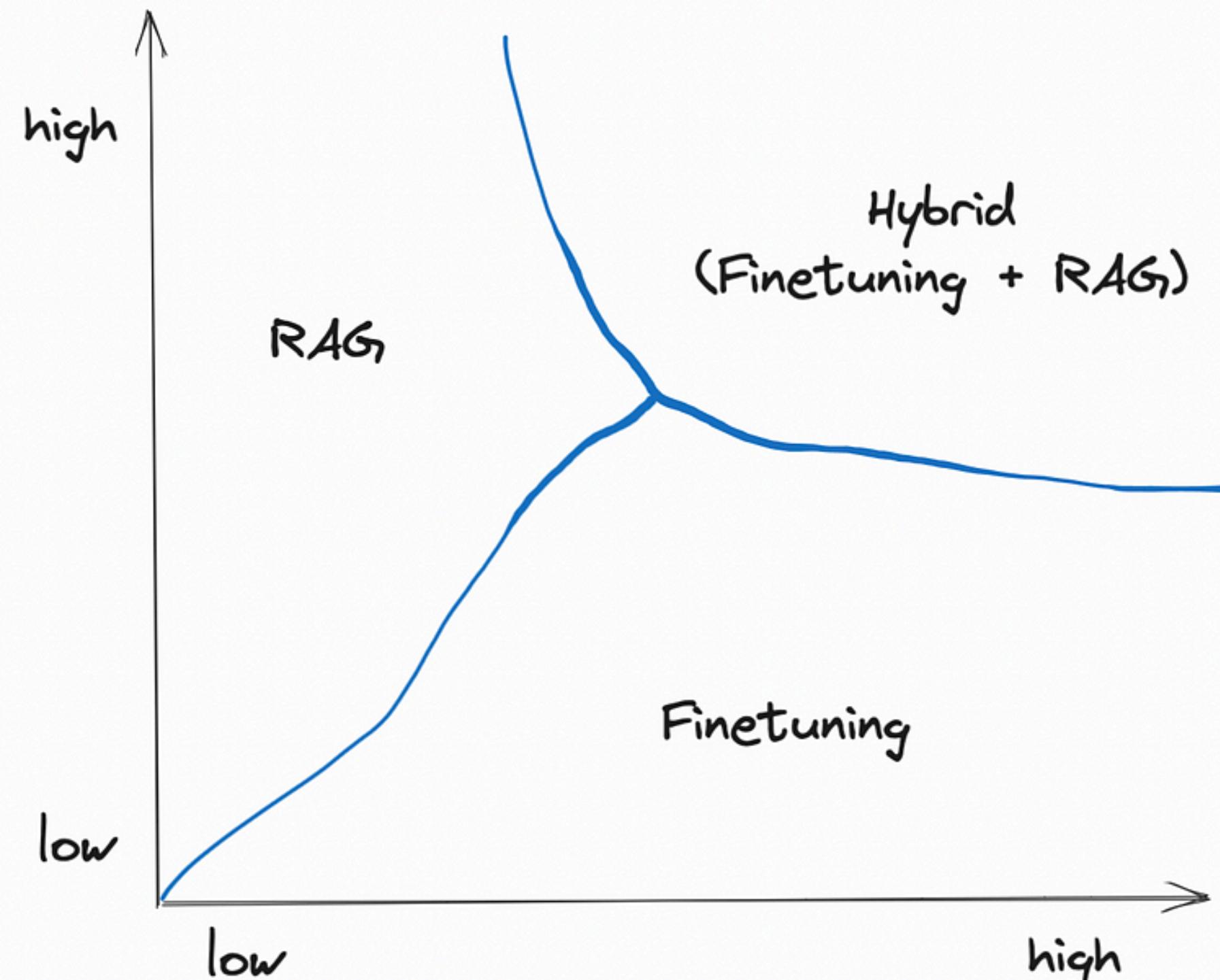


GDG Carthage

By Mohammed Arbi Nsibi

RAG / Fine-tuning

external knowledge
required



model adaptation required
(e.g. behaviour/
writing style/
vocabulary)

RAG (Retrieval-augmented generation)



Q&A



GDG Carthage

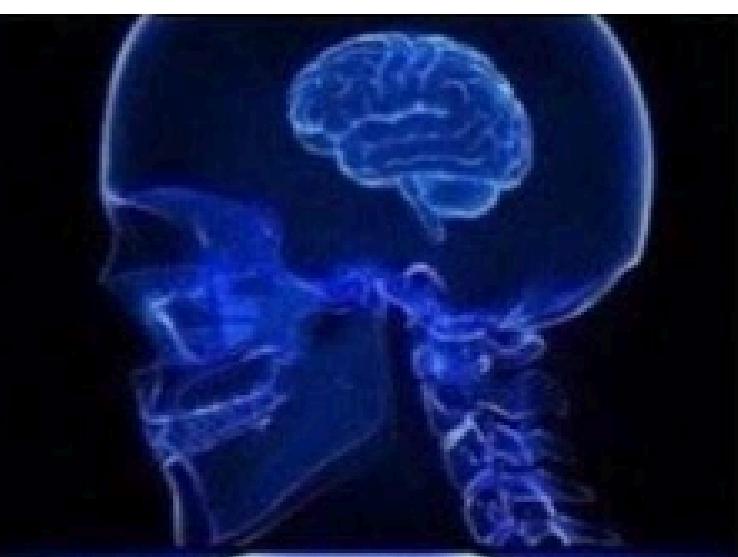
By Mohammed Arbi Nsibi



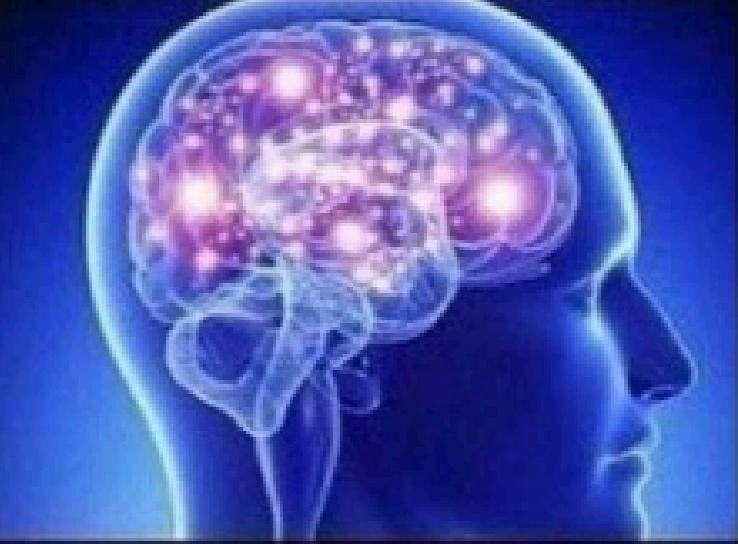
GDG Carthage

THANK YOU for you
attention!!

**USING AI
FOR CALCULATIONS**



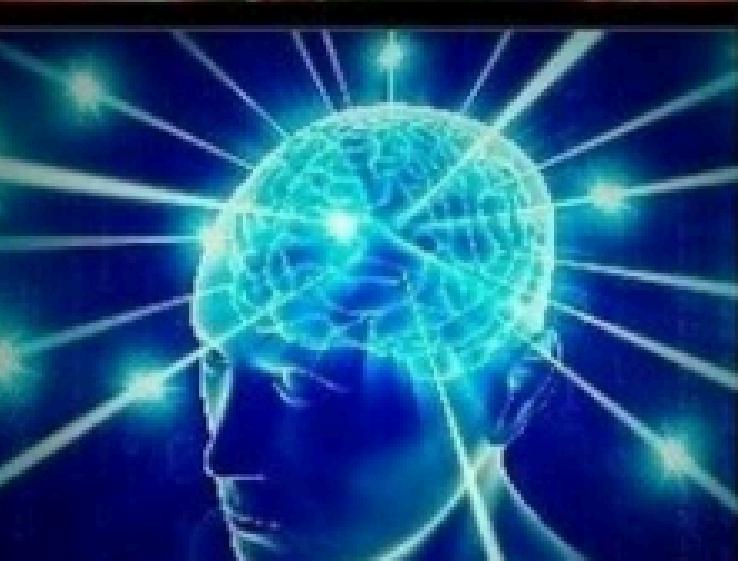
**USING AI TO
WRITE ESSAYS**



**USING AI TO
GENERATE CODE**

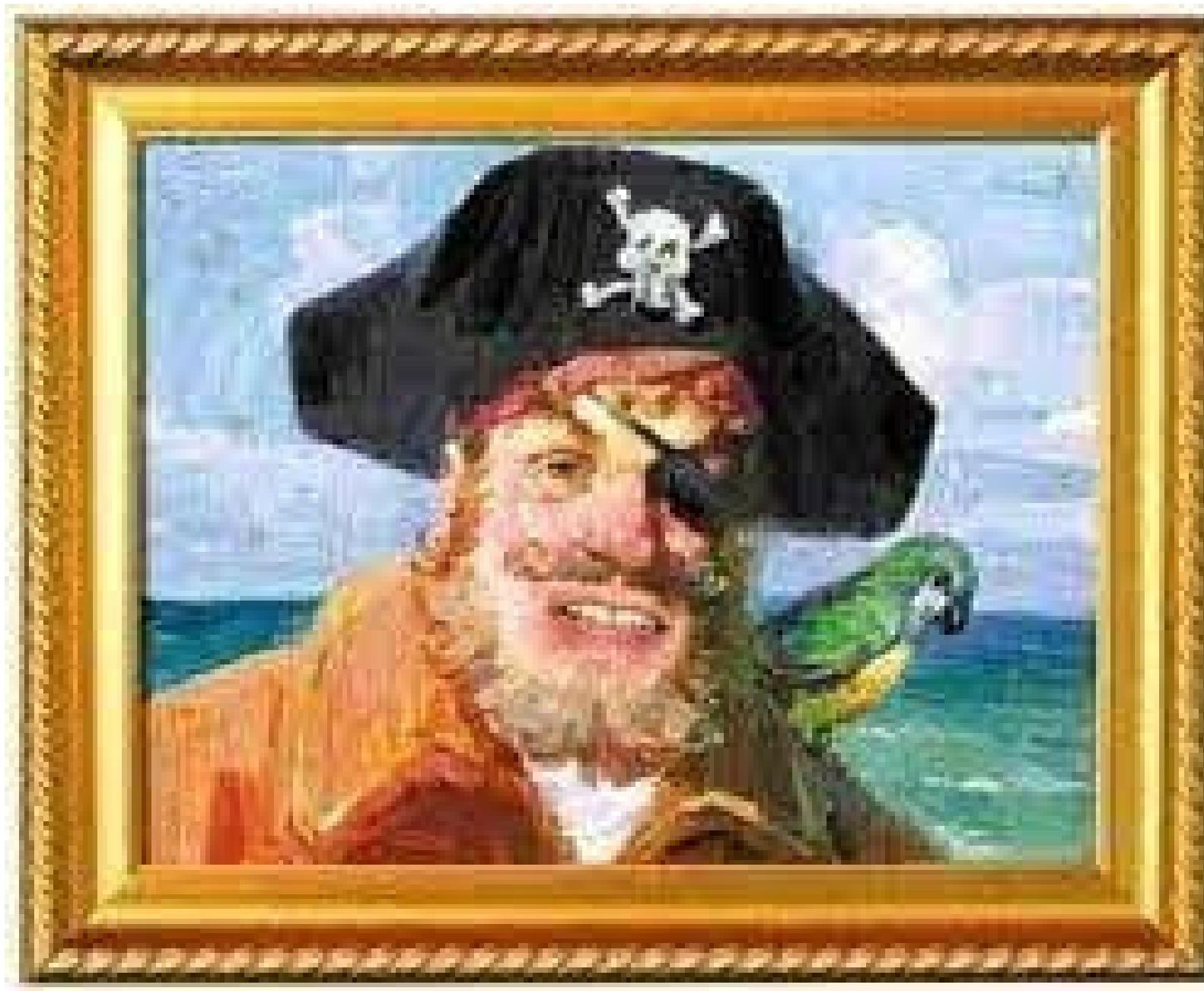


**USING AI
TO GENERATE
MEMES ABOUT AI**



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

QUIZ TIME



GDG Carthage

By Mohammed Arbi Nsibi