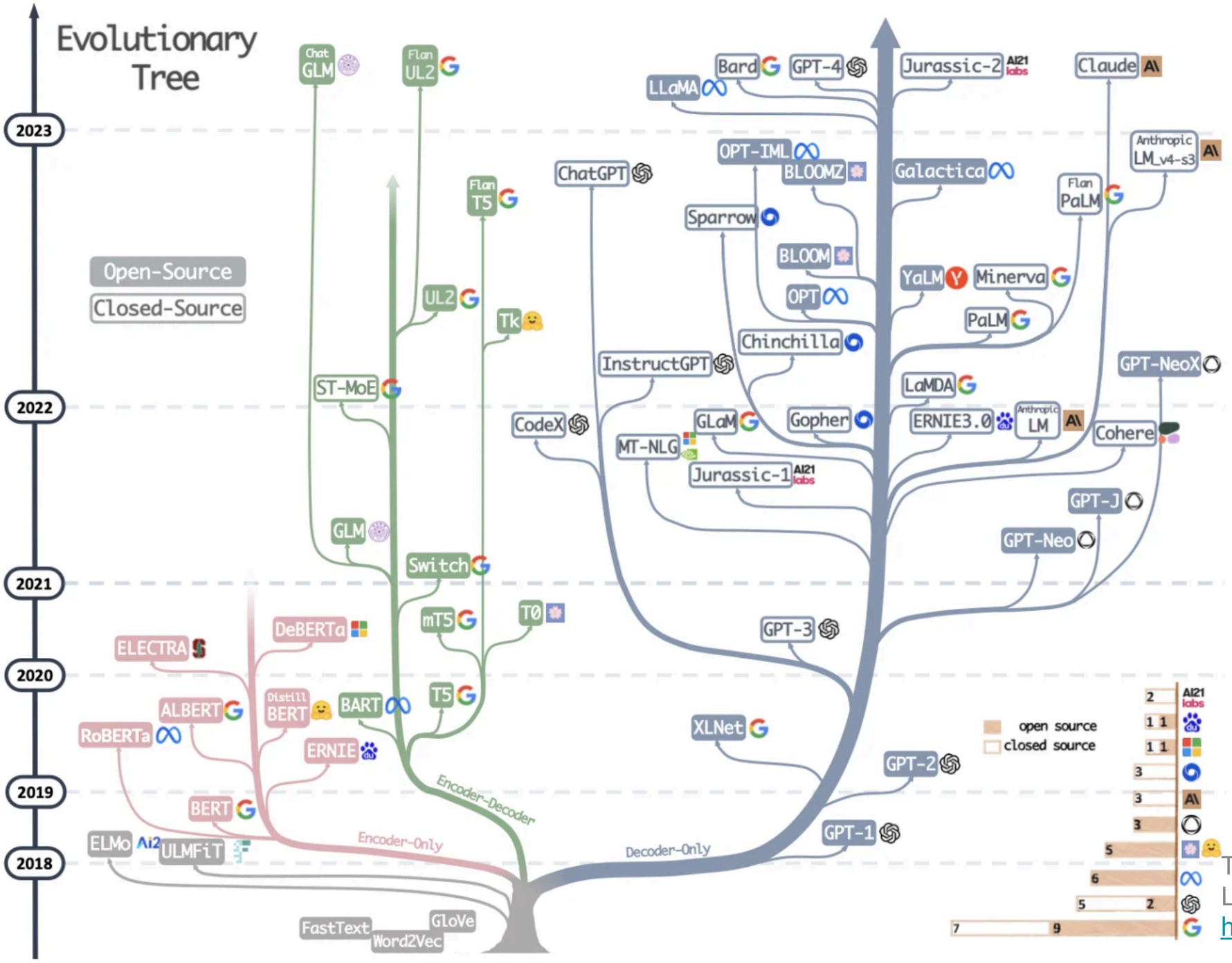
NLP vs LLM

GDG Carthage



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

GPT-1

(June 2018)

6 years: What has changed?

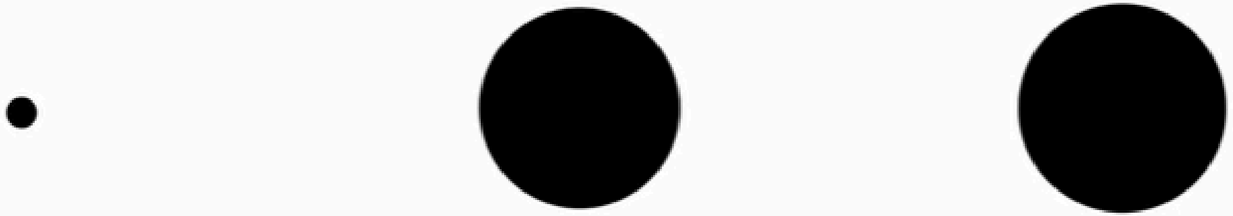
Llama **3.2**

(September 2024)

GDG Carthage

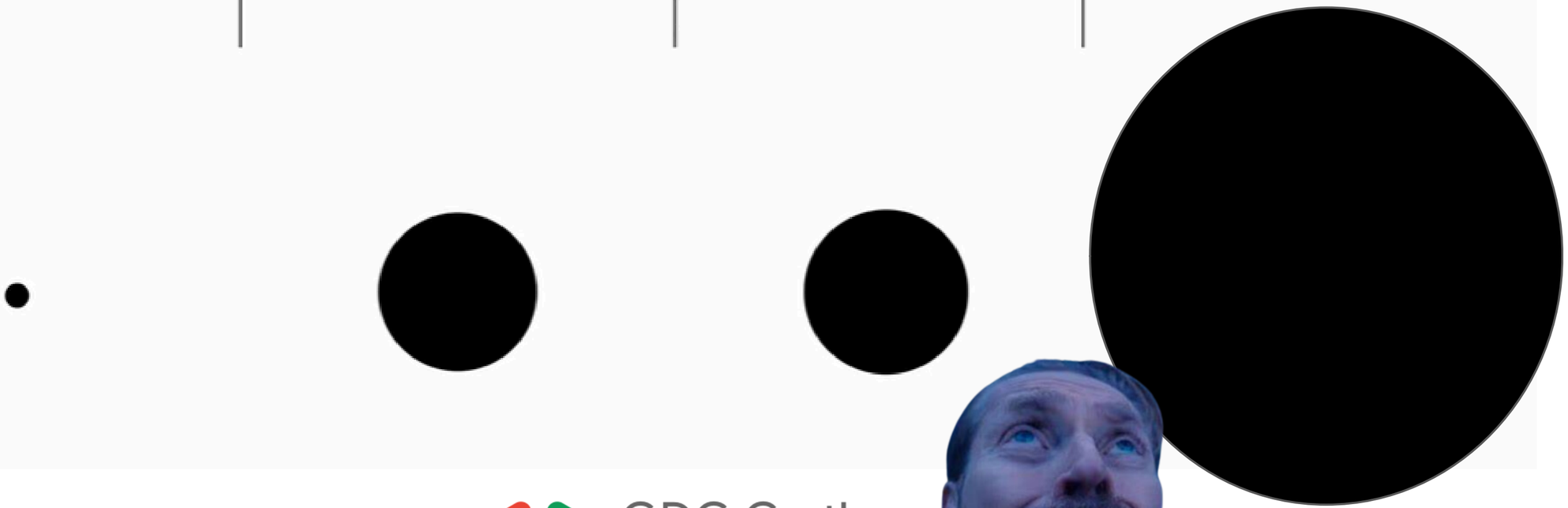
Model size

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



Model size

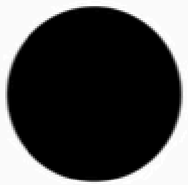
2019	2023	2023	2024
GPT-2	Llama-1	Llama-2	Llama-3
124M to 1.5B	7B to 65B	7B to 70B	8B to 405B

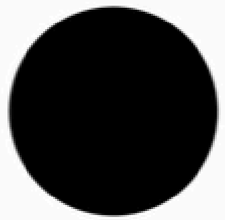


GDG Carthage

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

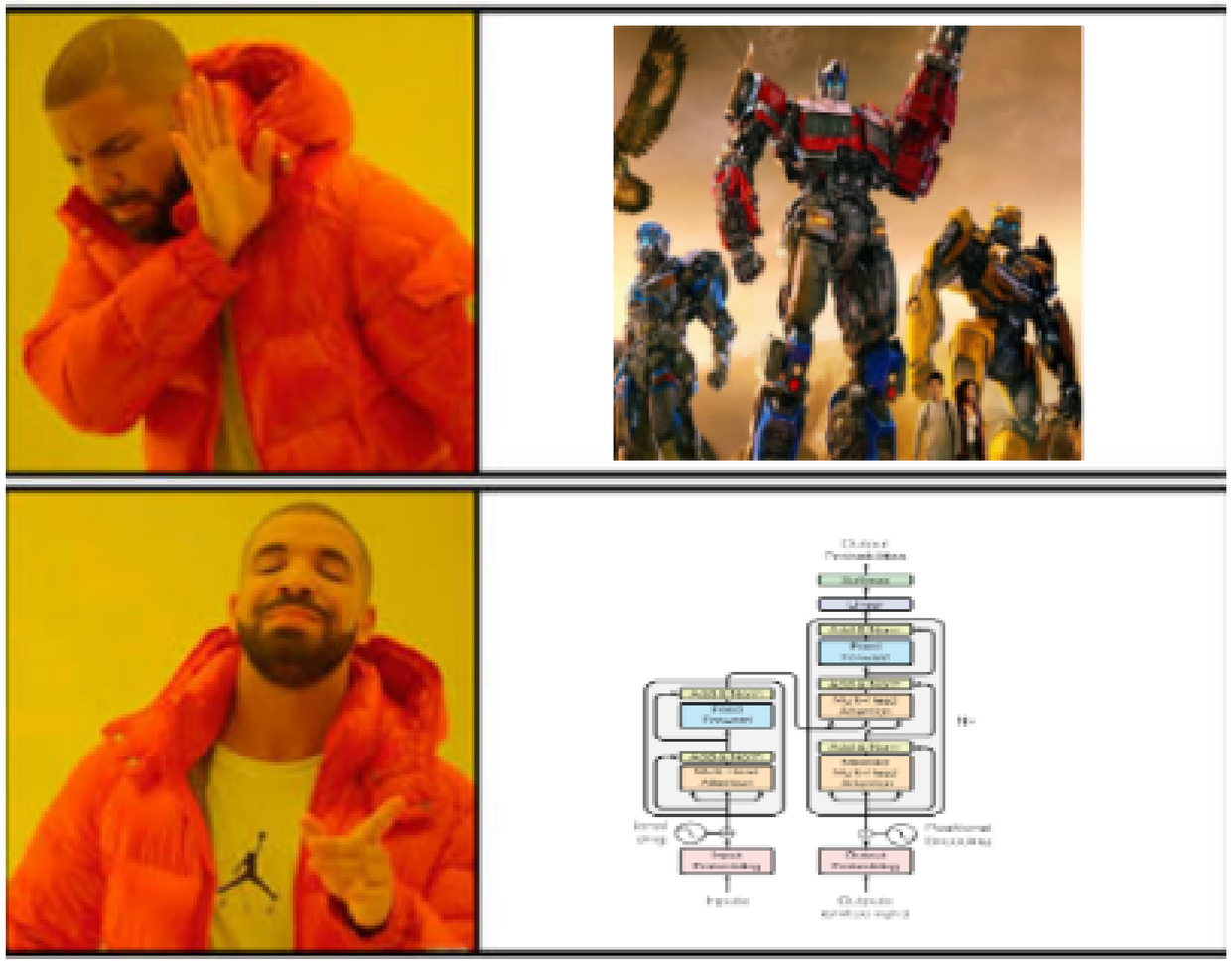
Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

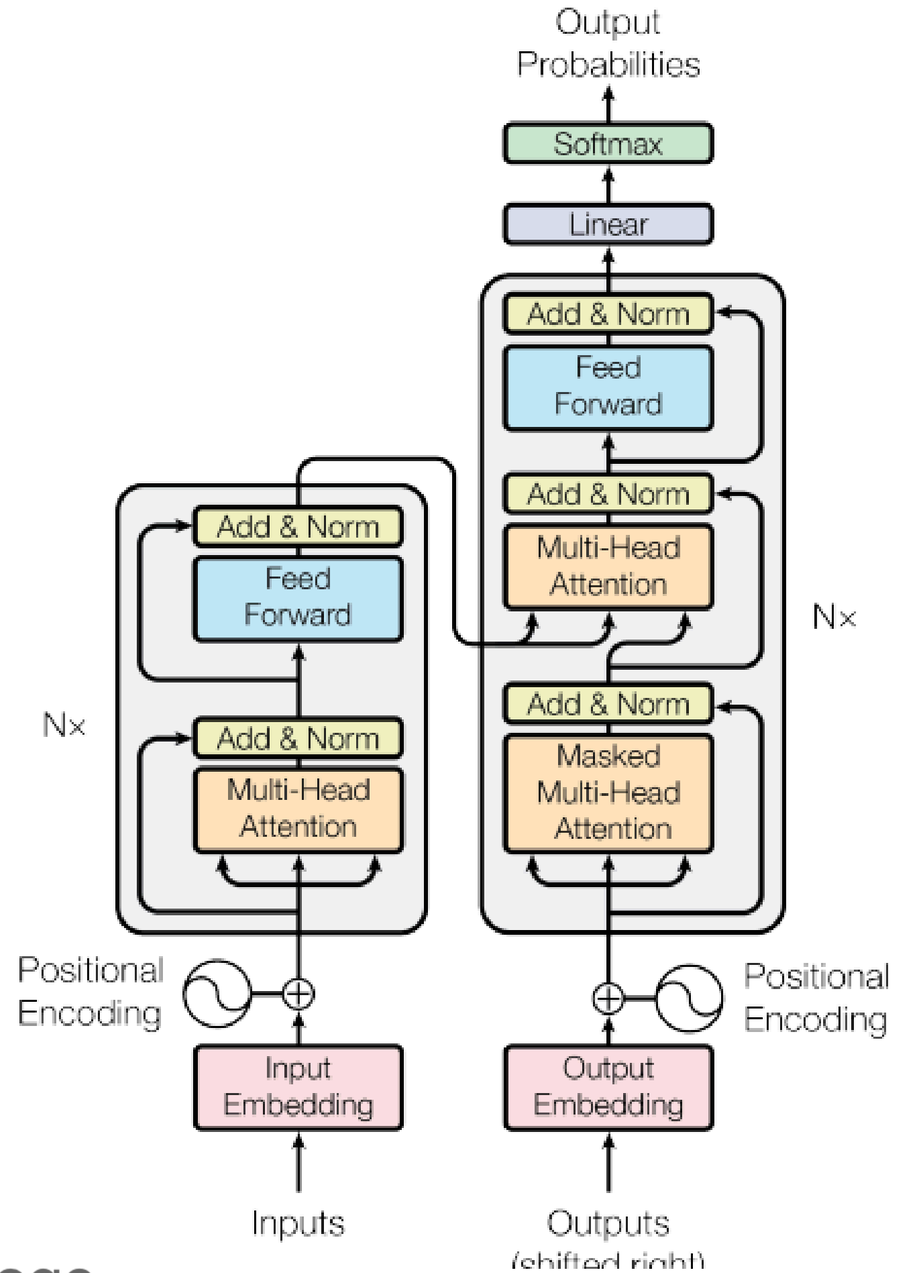
Abstract

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

Transformers





GDG Carthage

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


GDG Carthage

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:

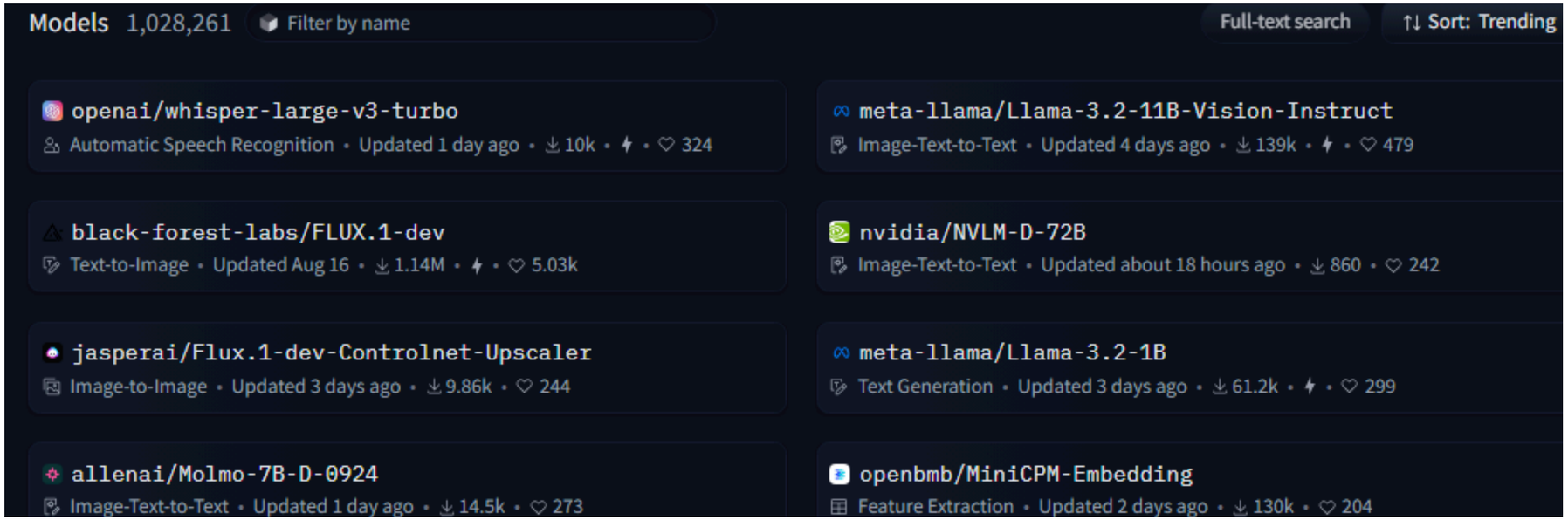
- What to focus on
- How much to focus on

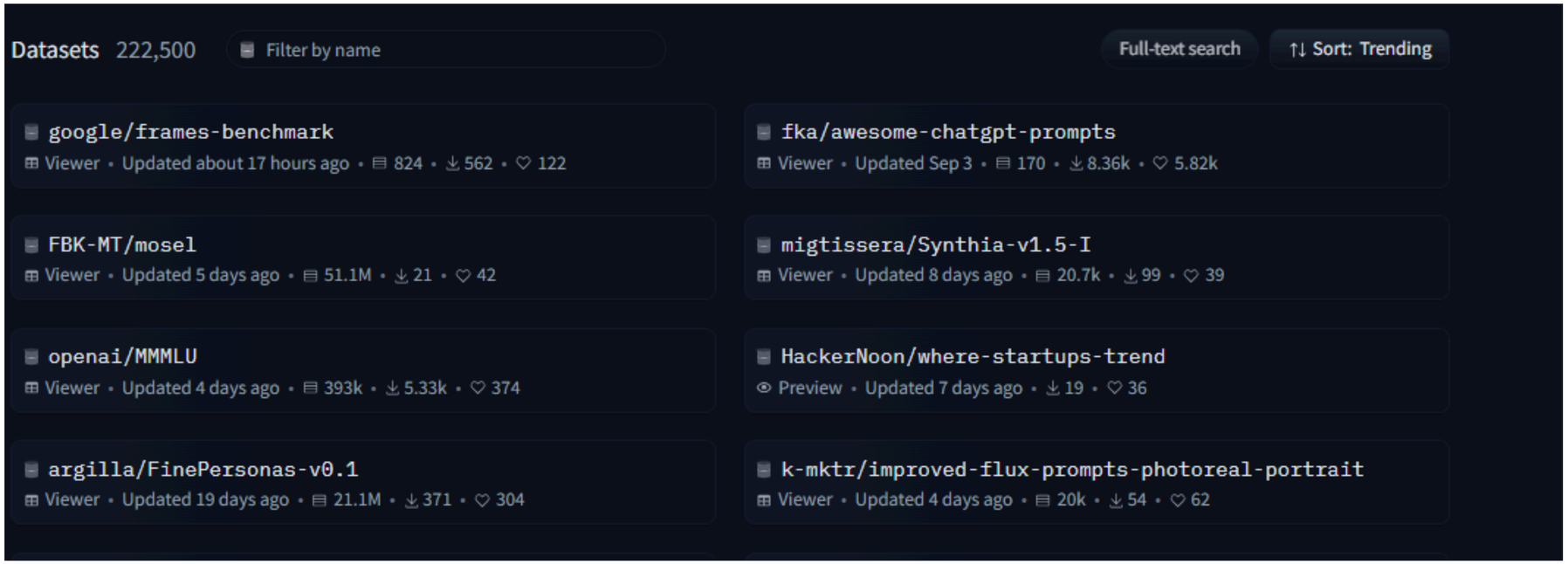
Where to find pretrained LLMs ?

Hugging Face

GDG Carthage

Where to find pretrained LLMs ?





Model types

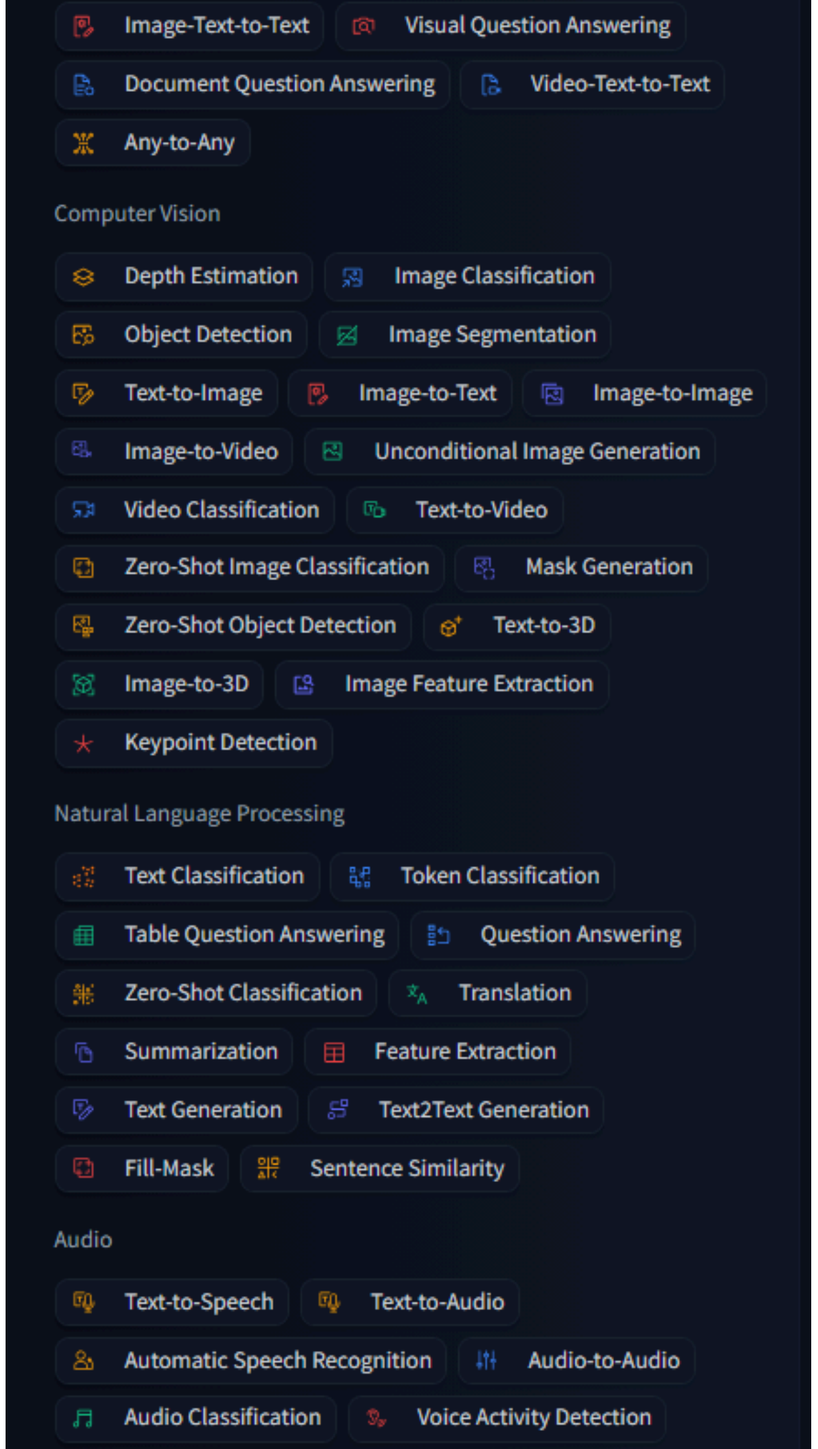
Text-to-text

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 

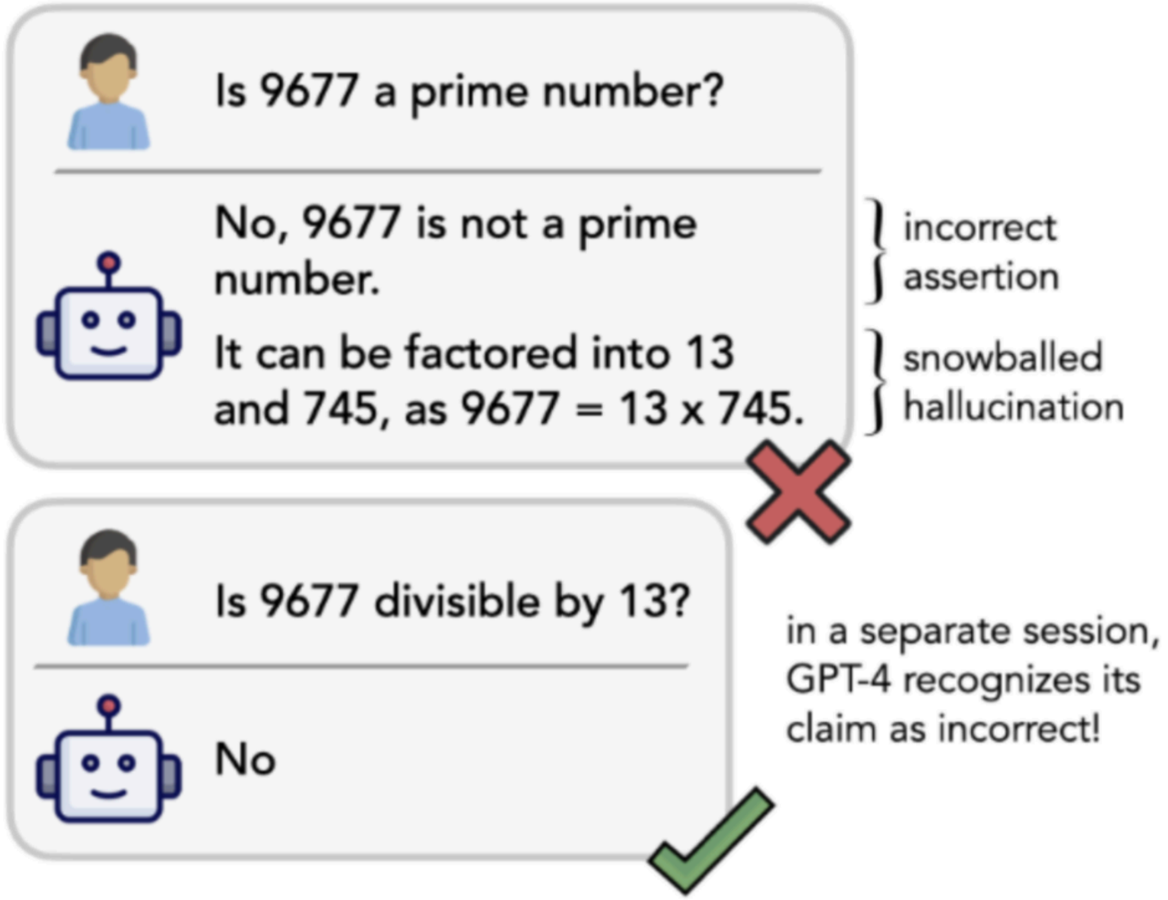
GDG Carthage

Hallucinations

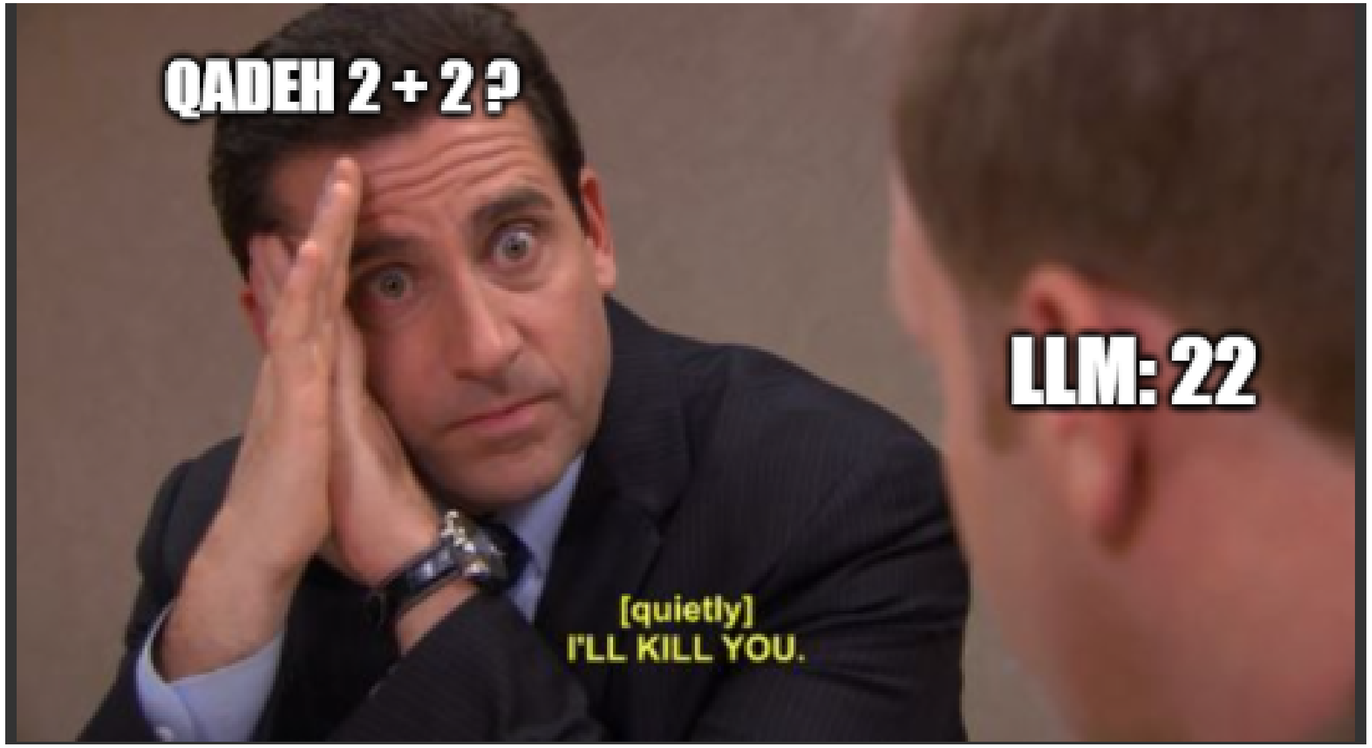
- The model is not trained on enough data.
- The model is trained on noisy or dirty data.
- The model is not given enough context .
- The model is not given enough constraints.

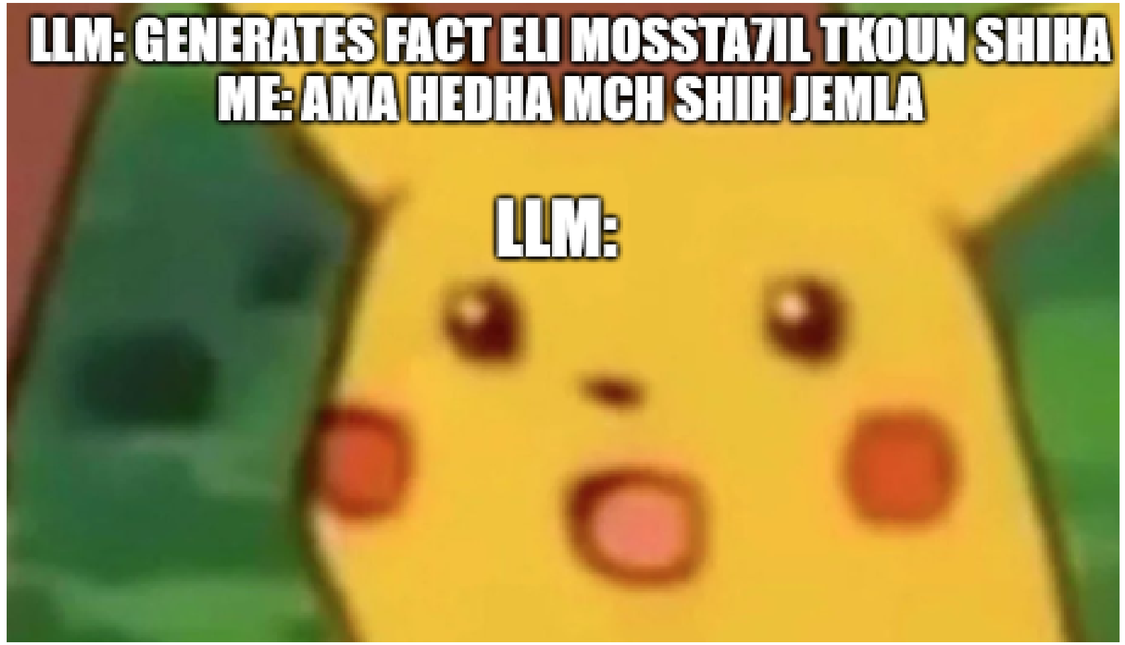
GDG Carthage

Hallucinations



Hallucinations

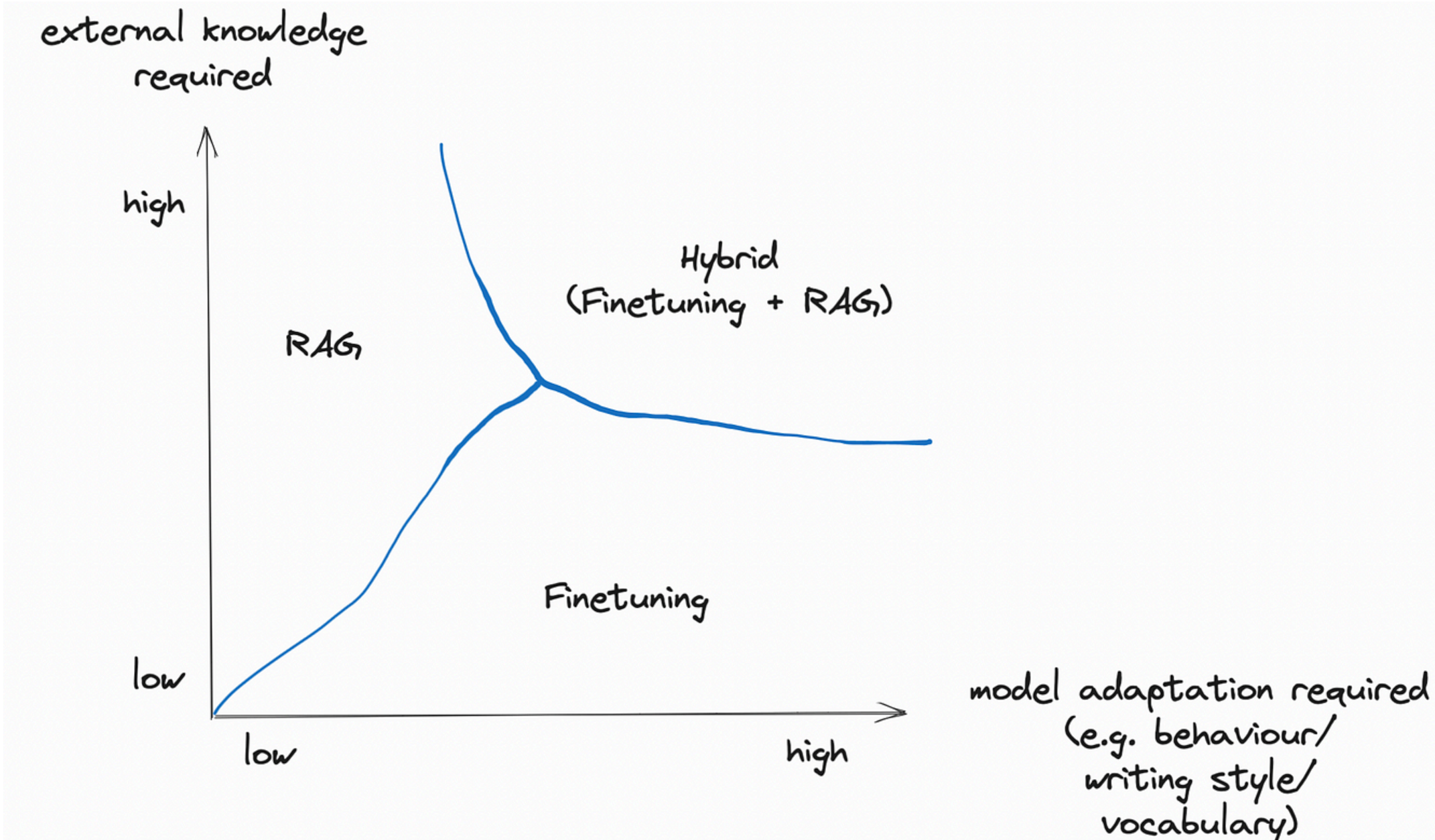




GDG Carthage

Solutions ?

RAG / Fine-tuning



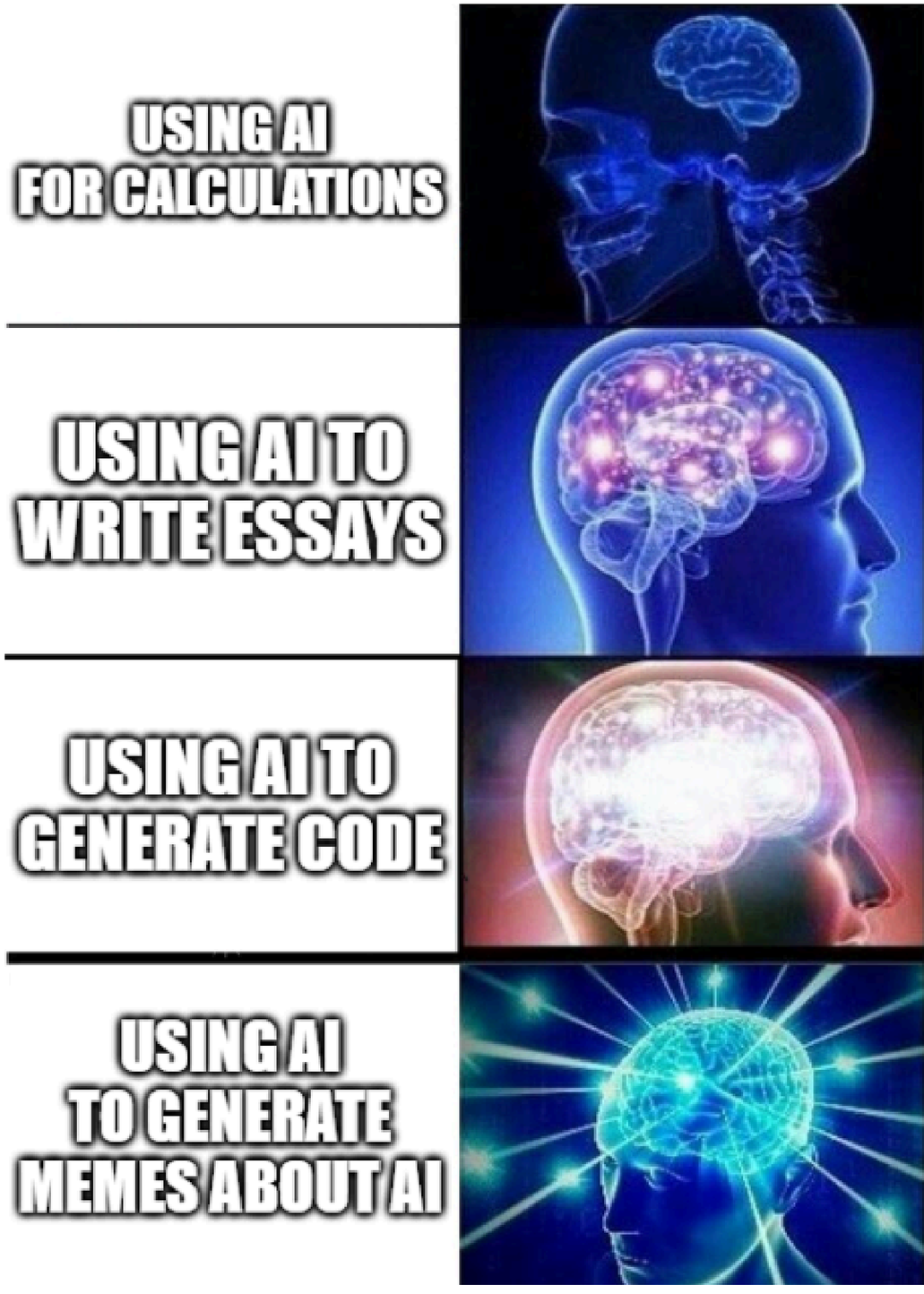
RAG (Retrieval-augmented generation)

Q&A

GDG Carthage

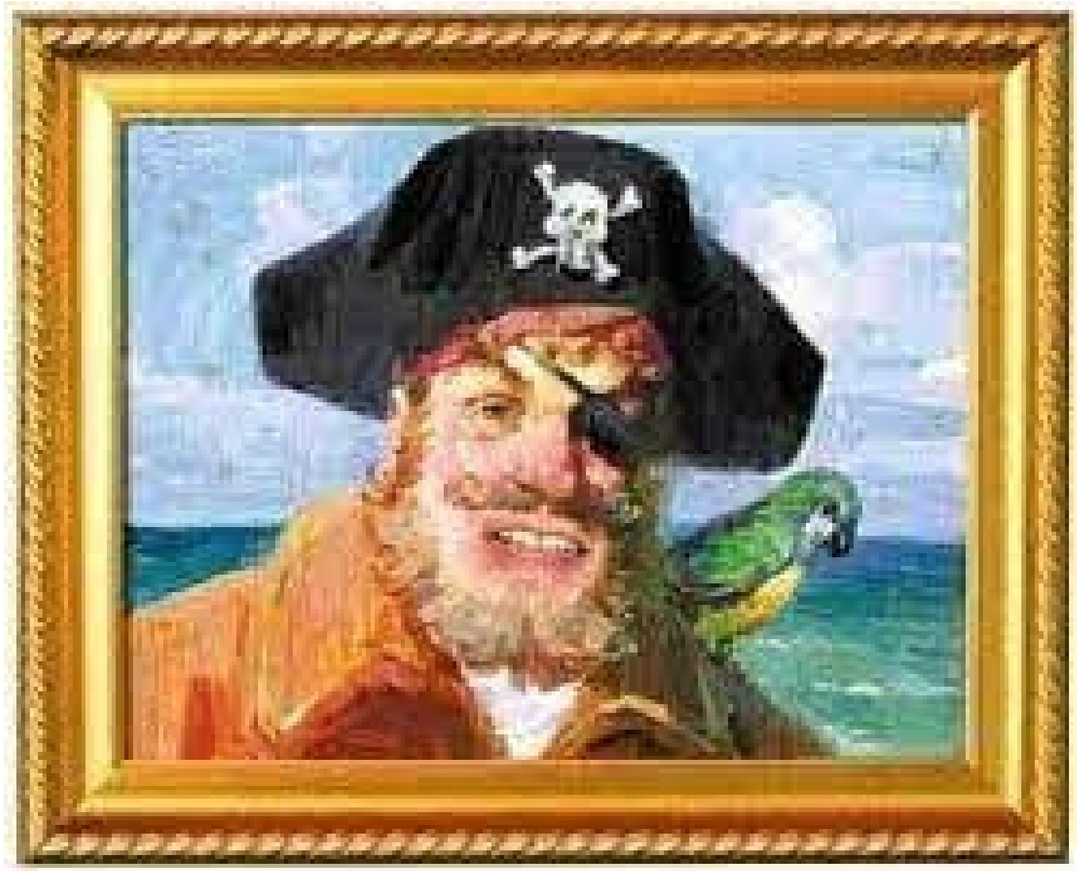
GDG Carthage

THANK YOU for you attention!!



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

QUIZ TIME 😄



GDG Carthage