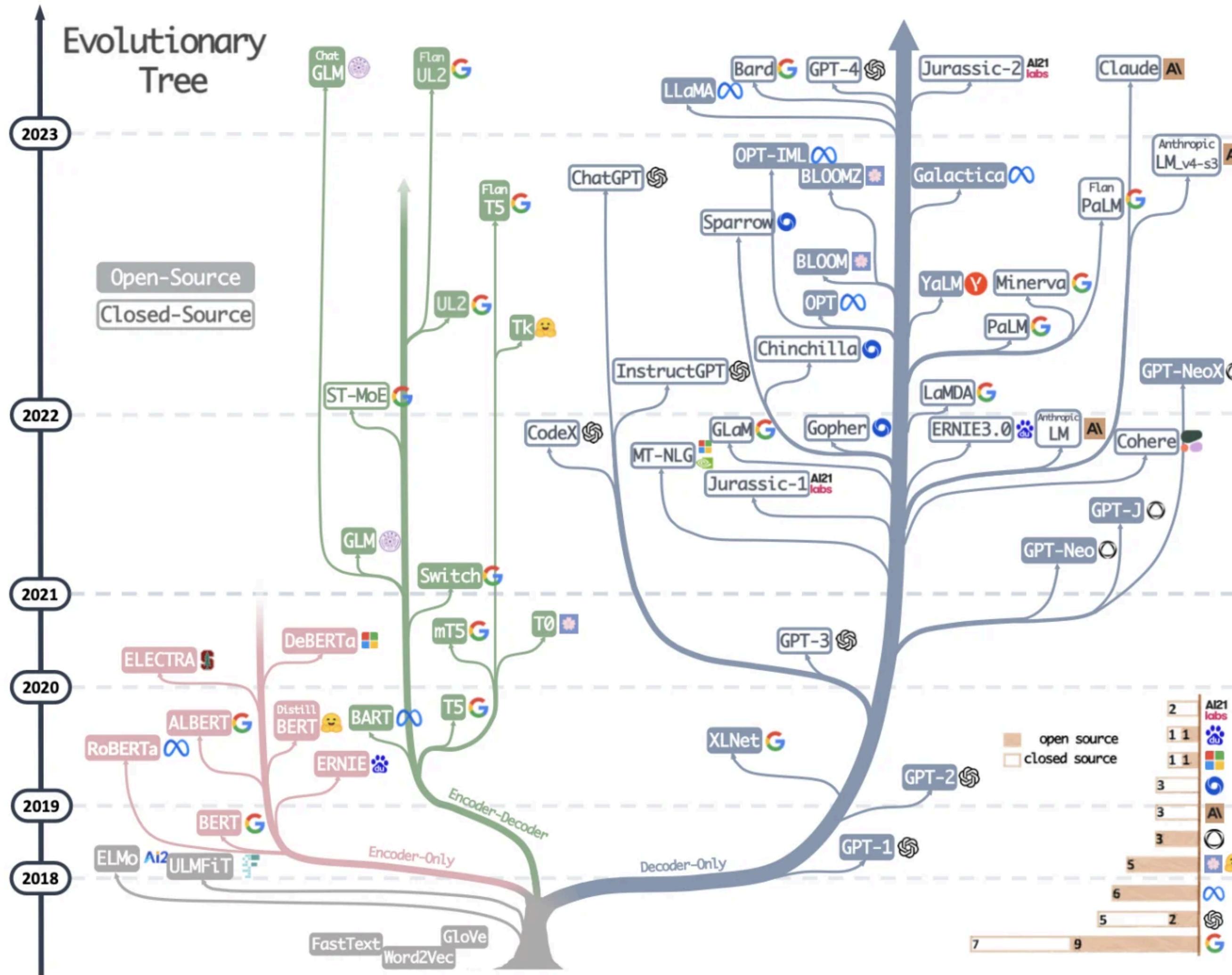


# NLP vs LLM





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

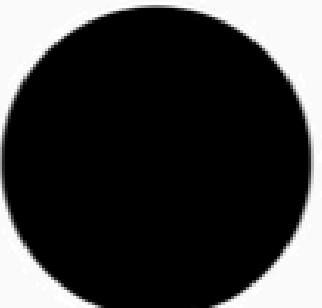
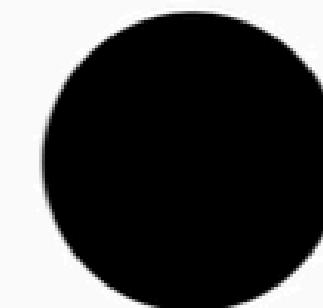
Llama-1

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

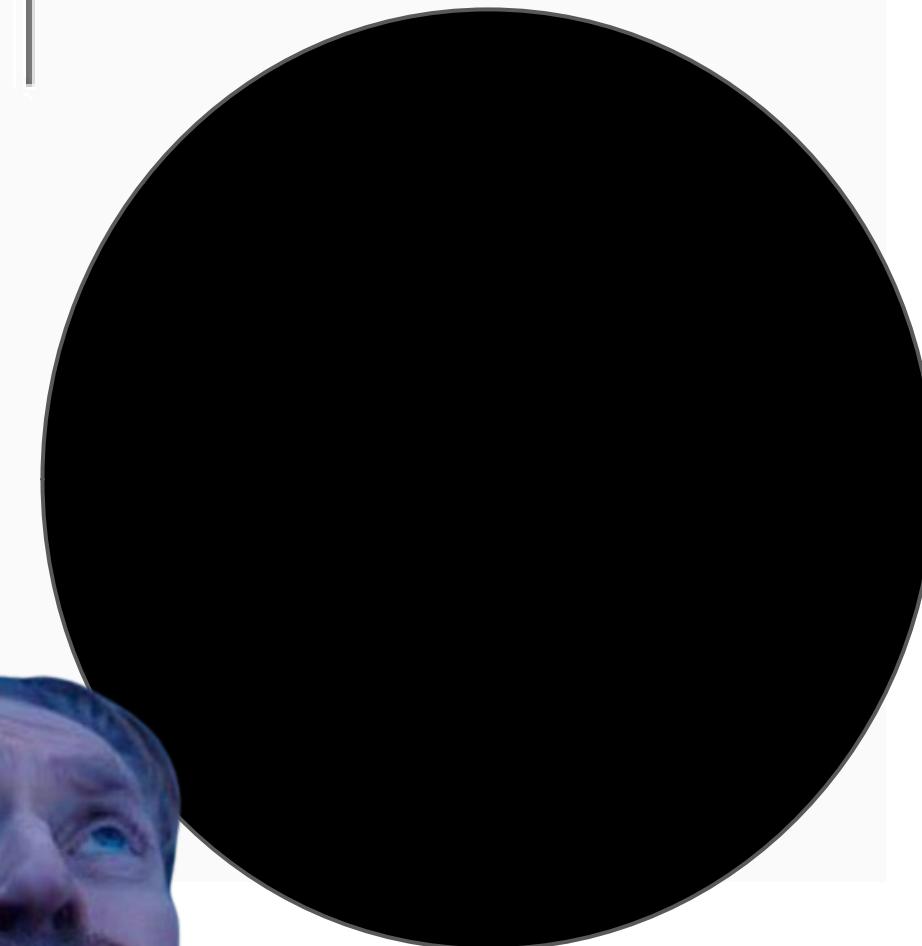
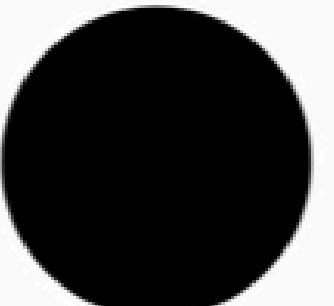
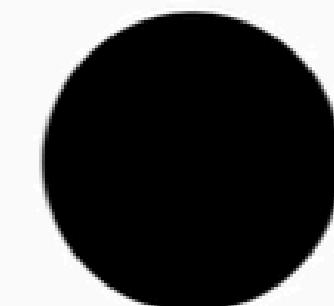
Llama-2

7B to 70B

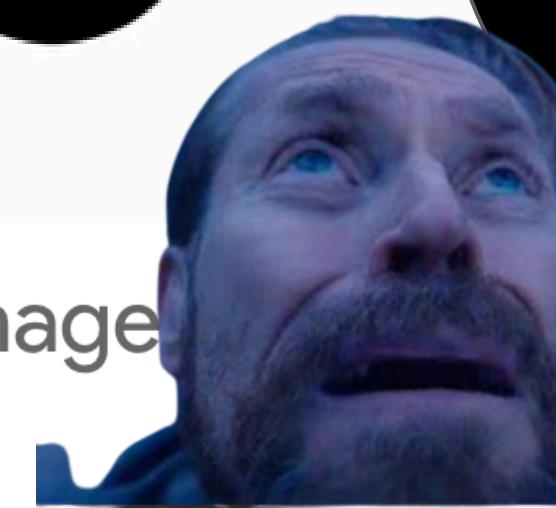
2024

Llama-3

8B to 405B



GDG Carthage



# Dataset

2019

GPT-2

40B tokens

2023

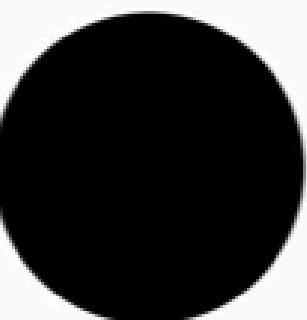
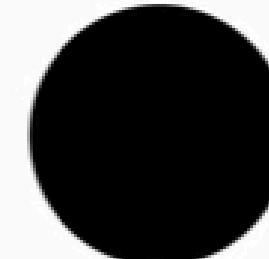
Llama-1

1.4T tokens

2023

Llama-2

2T tokens



GDG Carthage

# Dataset

2019

GPT-2

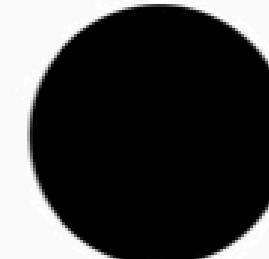
40B tokens



2023

Llama-1

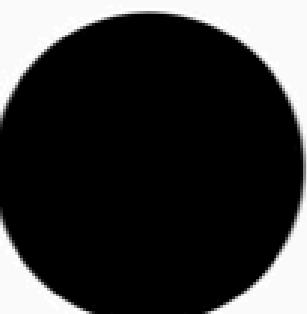
1.4T tokens



2023

Llama-2

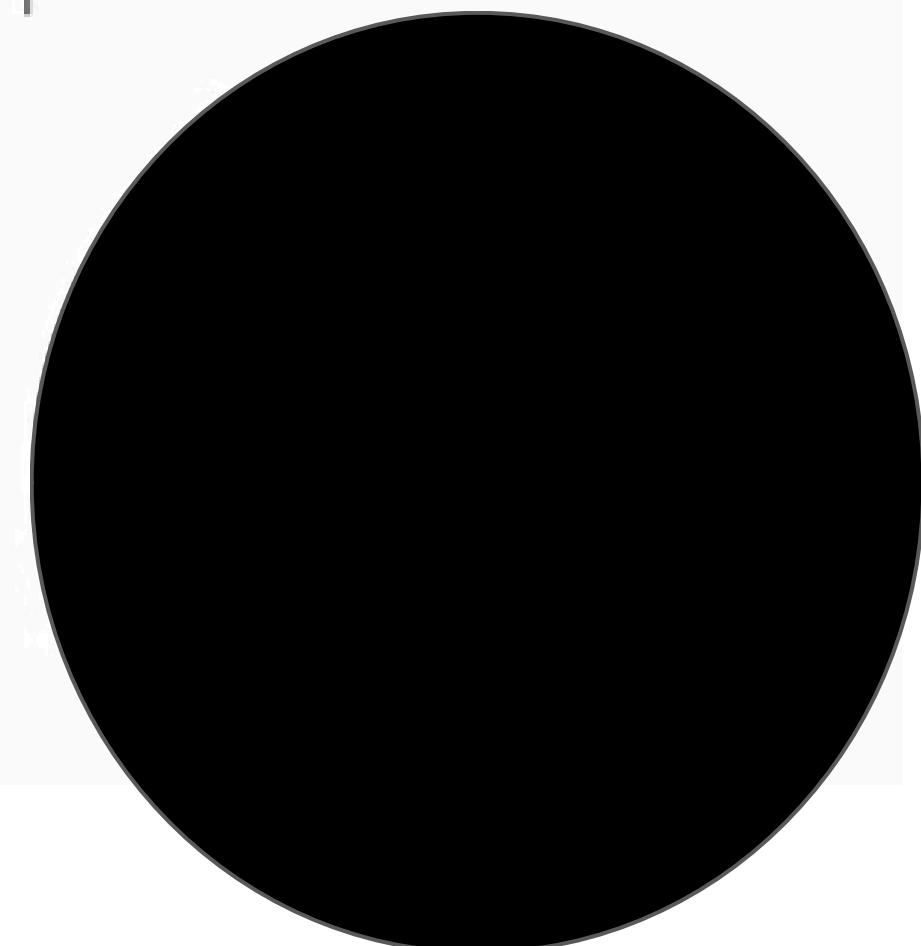
2T tokens



2024

Llama-3

15T tokens



GDG Carthage

# "Attention is All You Need"



## Attention Is All You Need

**Ashish Vaswani\***  
Google Brain  
[avaswani@google.com](mailto:avaswani@google.com)

**Noam Shazeer\***  
Google Brain  
[noam@google.com](mailto:noam@google.com)

**Niki Parmar\***  
Google Research  
[nikip@google.com](mailto:nikip@google.com)

**Jakob Uszkoreit\***  
Google Research  
[usz@google.com](mailto:usz@google.com)

**Llion Jones\***  
Google Research  
[llion@google.com](mailto:llion@google.com)

**Aidan N. Gomez\* †**  
University of Toronto  
[aidan@cs.toronto.edu](mailto:aidan@cs.toronto.edu)

**Lukasz Kaiser\***  
Google Brain  
[lukaszkaiser@google.com](mailto:lukaszkaiser@google.com)

**Illia Polosukhin\* ‡**  
[illia.polosukhin@gmail.com](mailto:illia.polosukhin@gmail.com)

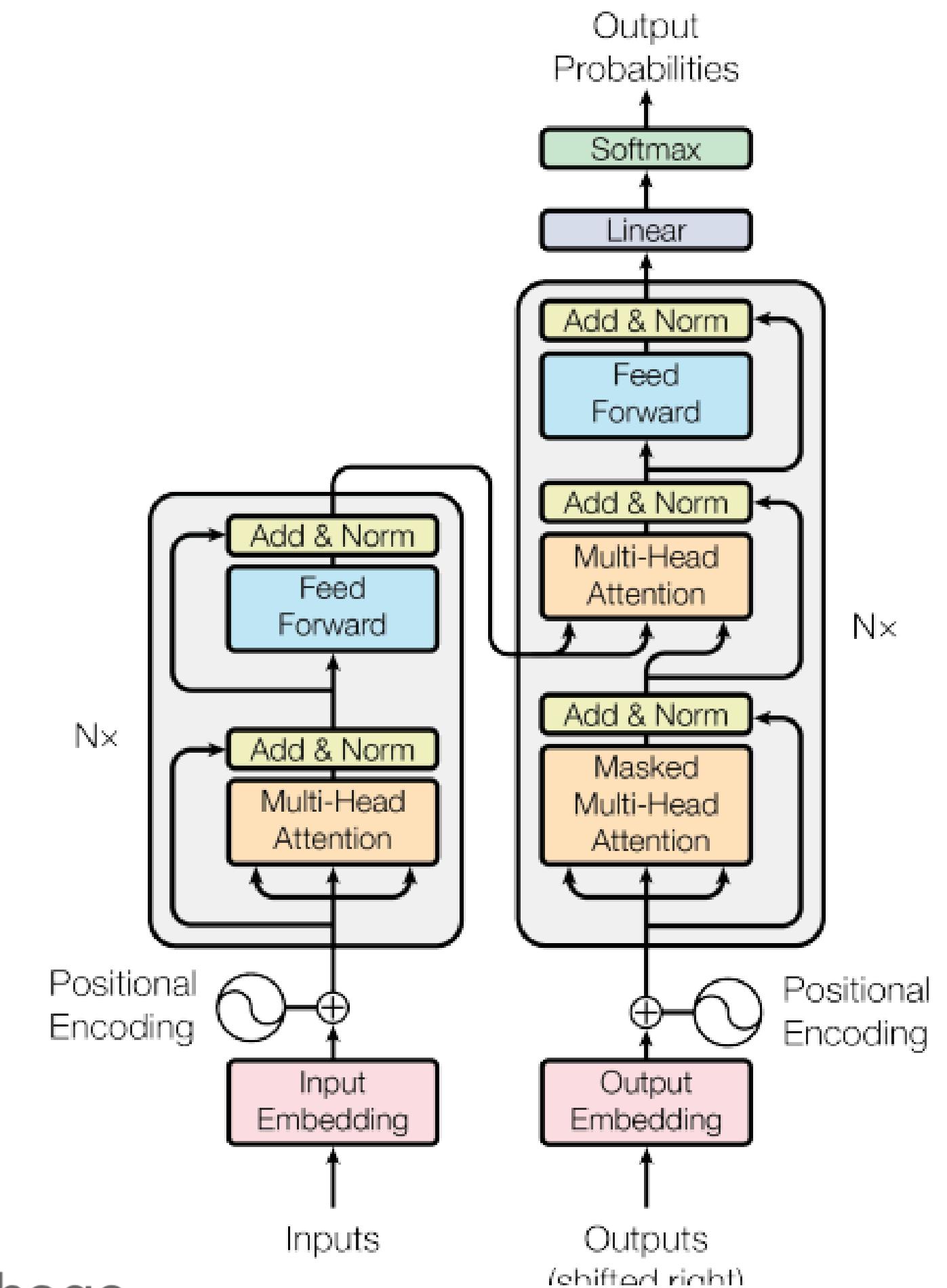
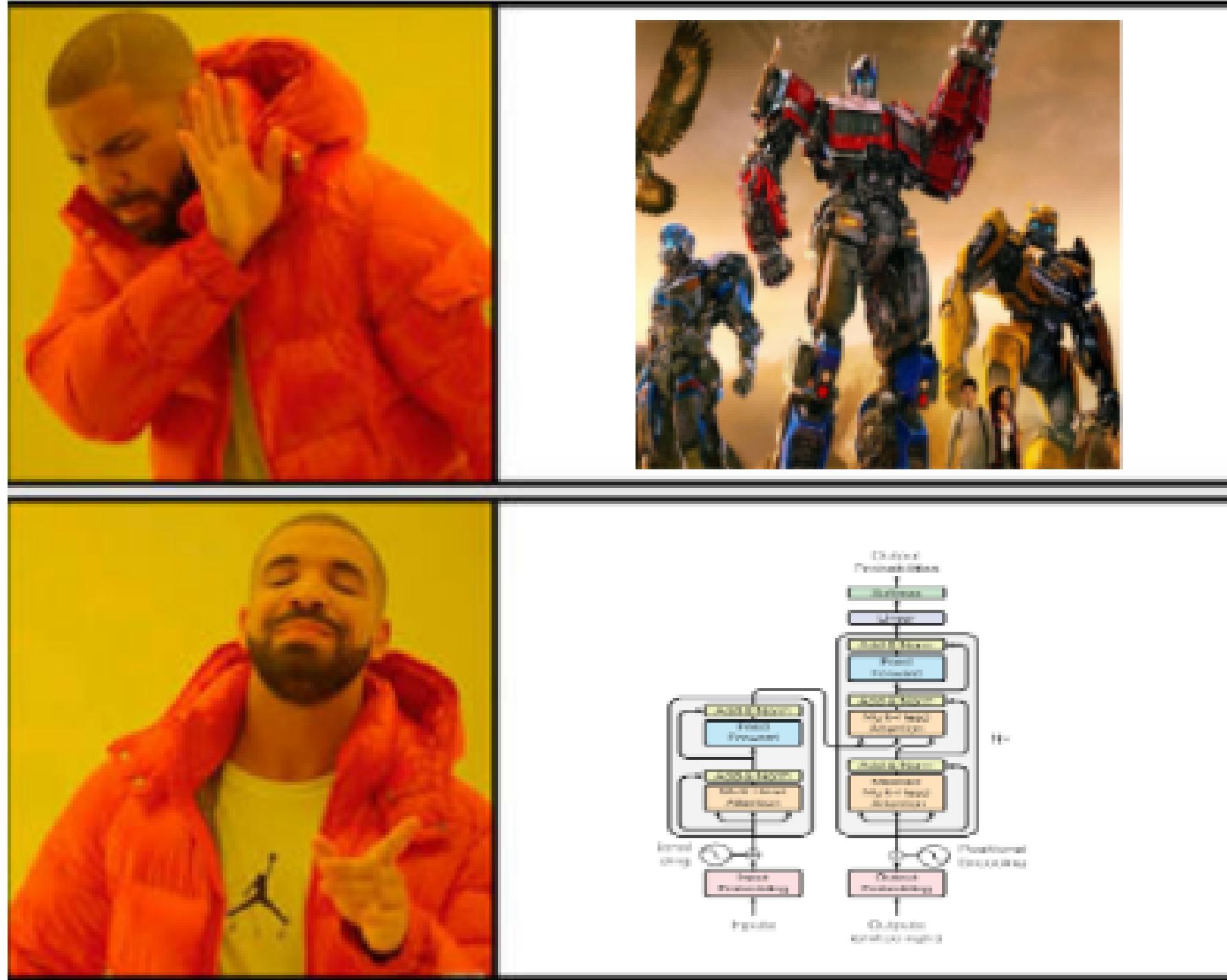
### Abstract

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	$\sin(1/10000^0) \approx 0.8415$	$\cos(1/10000^0) \approx 0.5403$	$\sin(1/10000^0) \approx 0.8415$	$\cos(1/10000^0) \approx 0.5403$
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

# 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

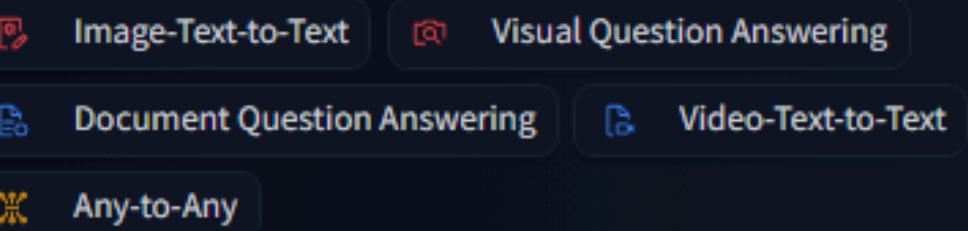
 [fka/awesome-chatgpt-prompts](#)  
Viewer • Updated Sep 3 • 170 • ↓ 8.36k • ❤ 5.82k

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

 [HackerNoon/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**

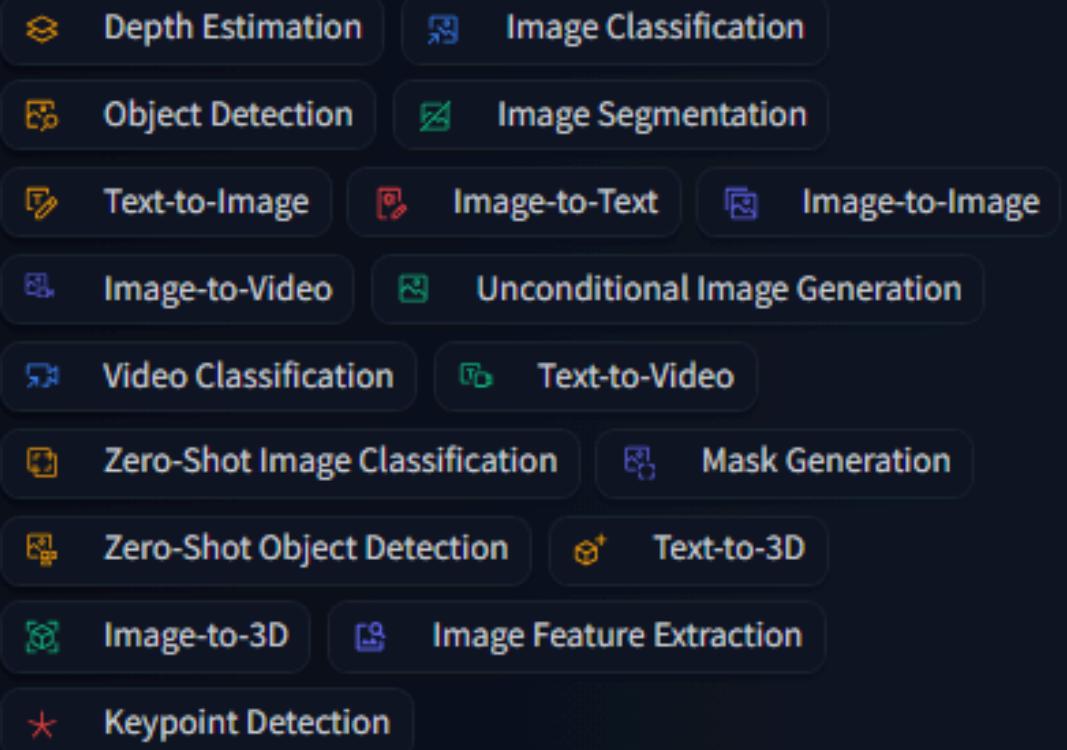
**Text-to-image**

**Text-to-video**

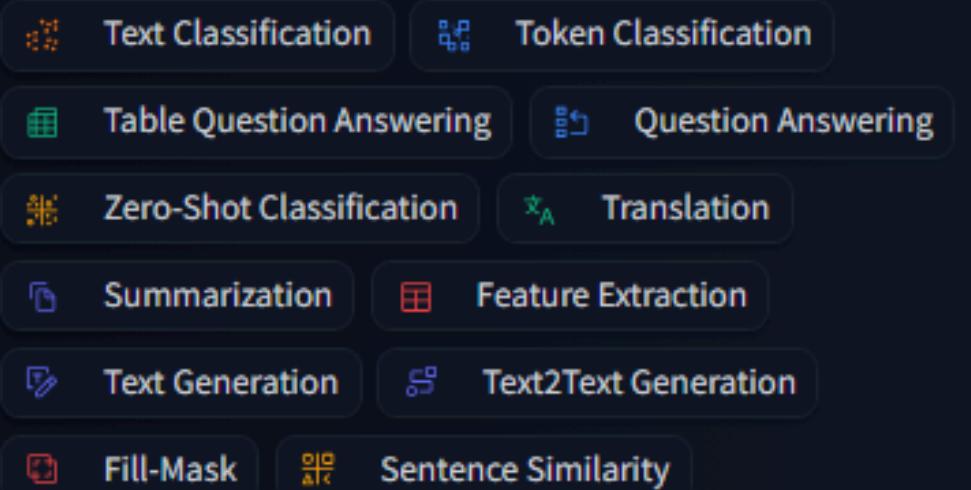
**Sentence-similarity**



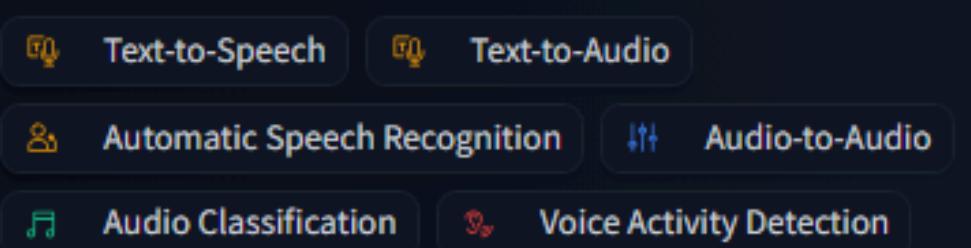
## Computer Vision



## 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

# 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

# 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**

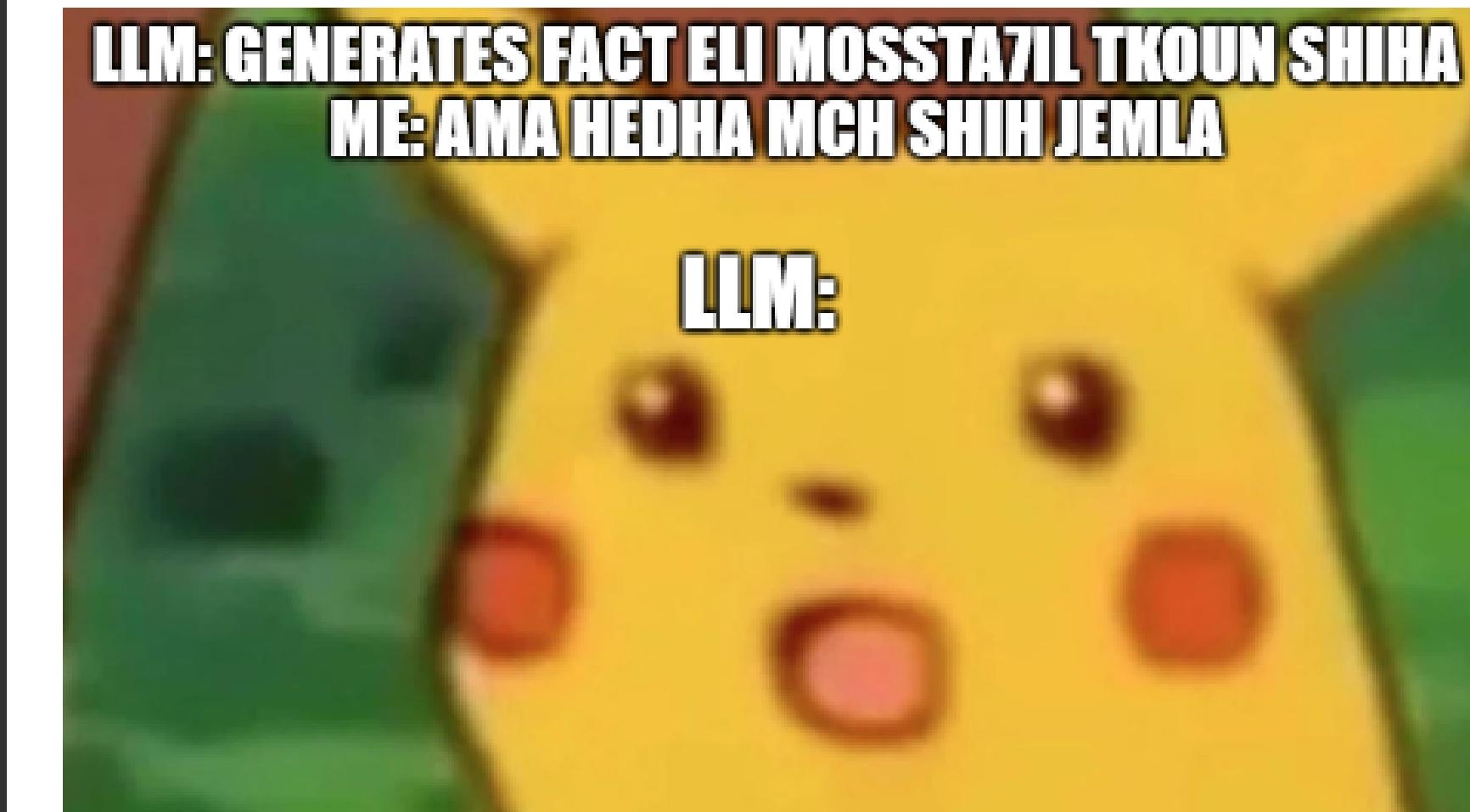
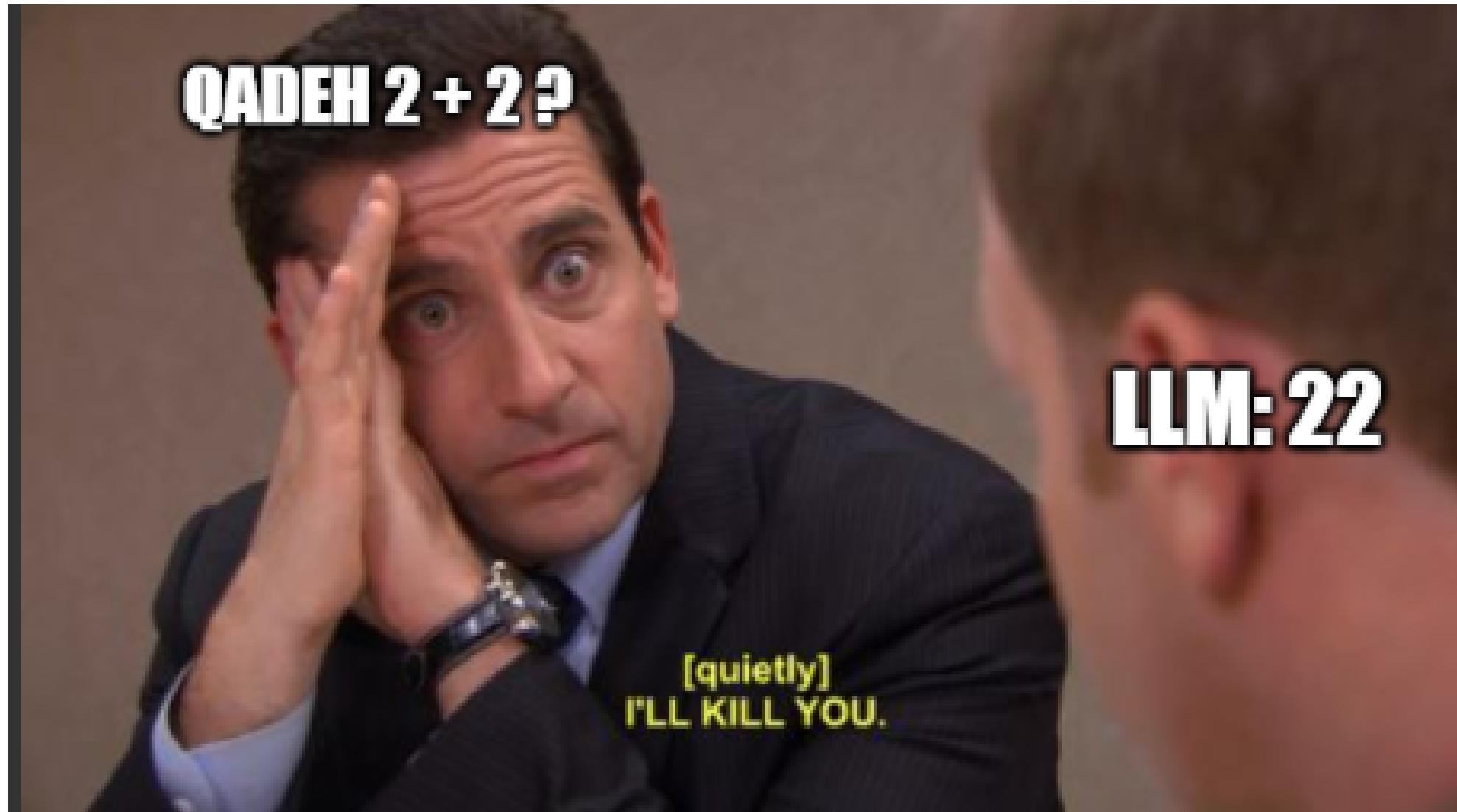
in a separate session, GPT-4 recognizes its claim as incorrect!

**✓**

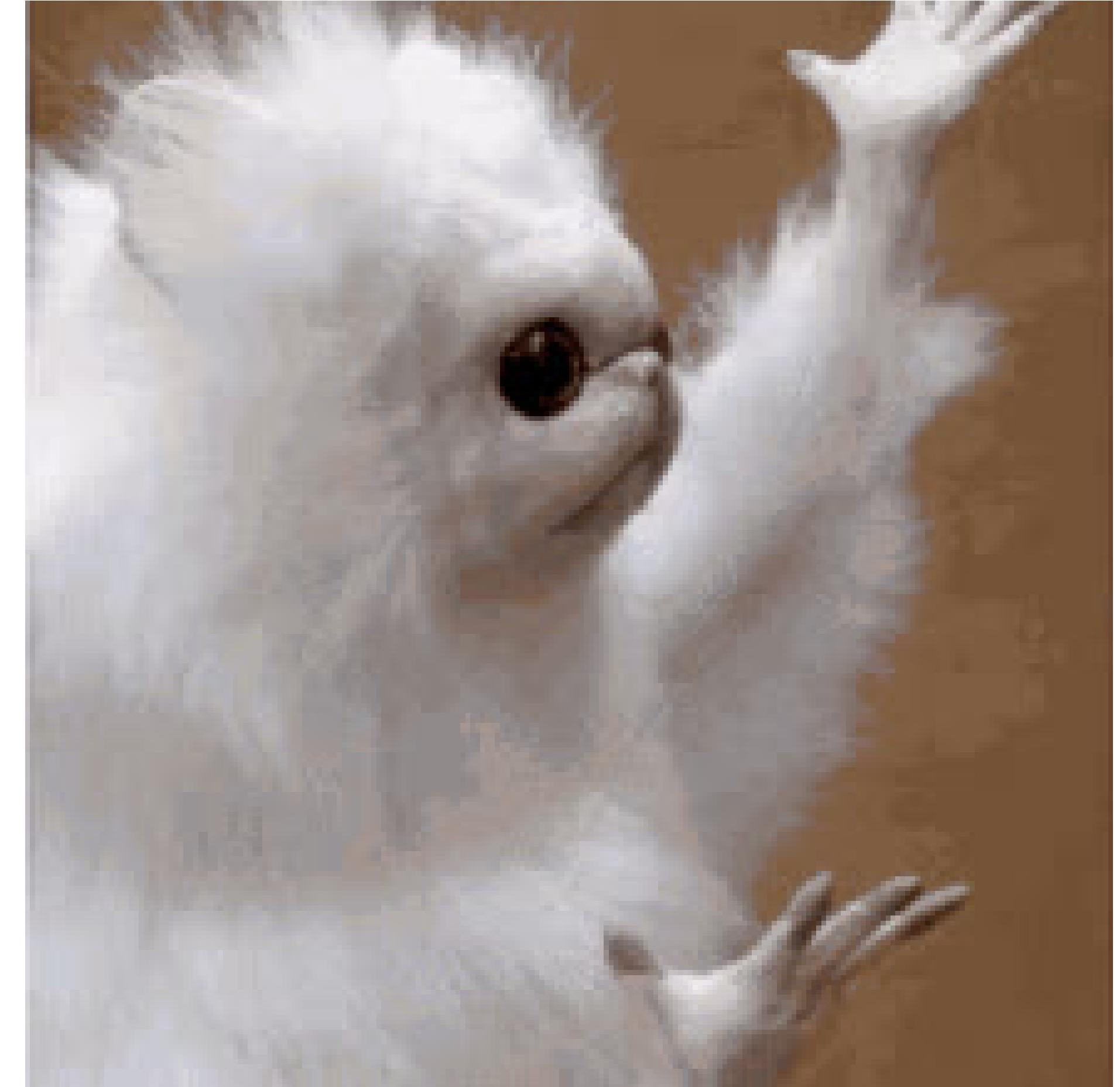
} incorrect assertion  
} snowballed hallucination



# Hallucinations



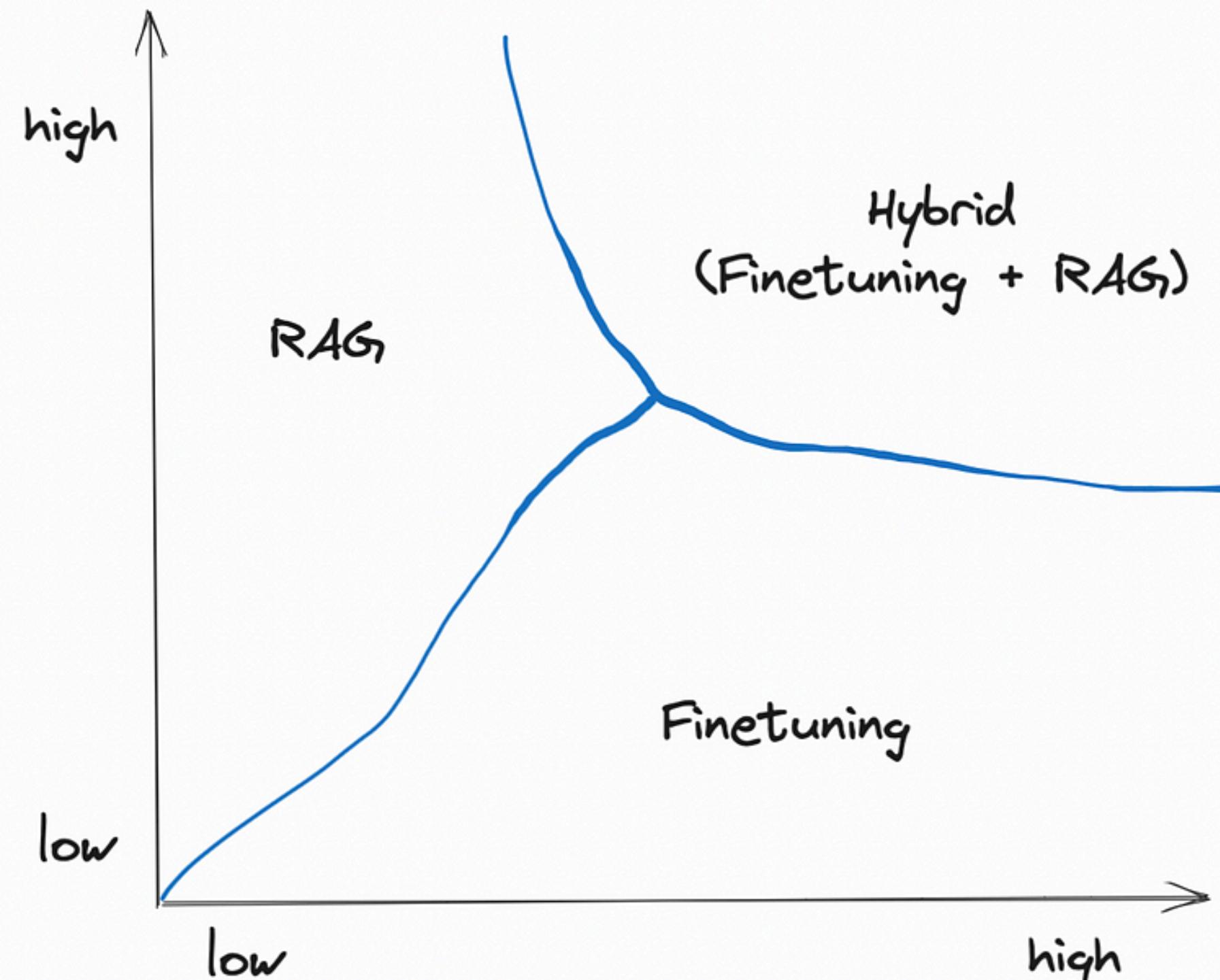
# Solutions ?



GDG Carthage

# RAG / Fine-tuning

external knowledge  
required



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

# RAG (Retrieval-augmented generation)



# Q&A



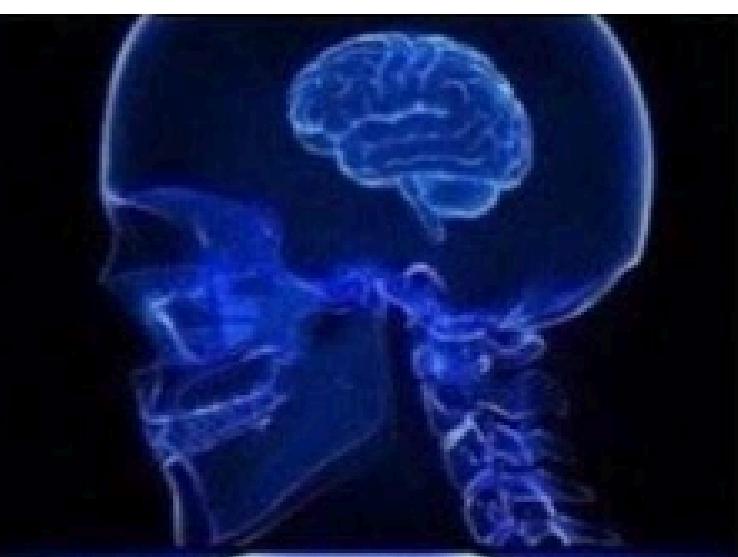
GDG Carthage



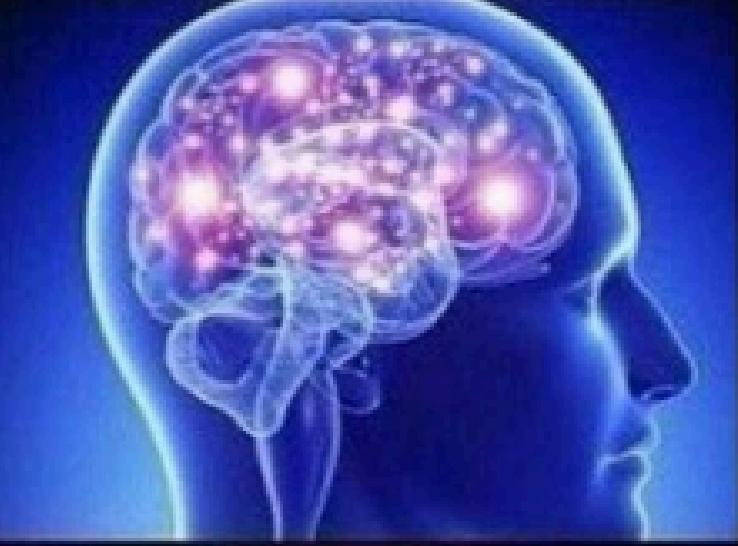
GDG Carthage

**THANK YOU** for you  
attention!!

**USING AI  
FOR CALCULATIONS**



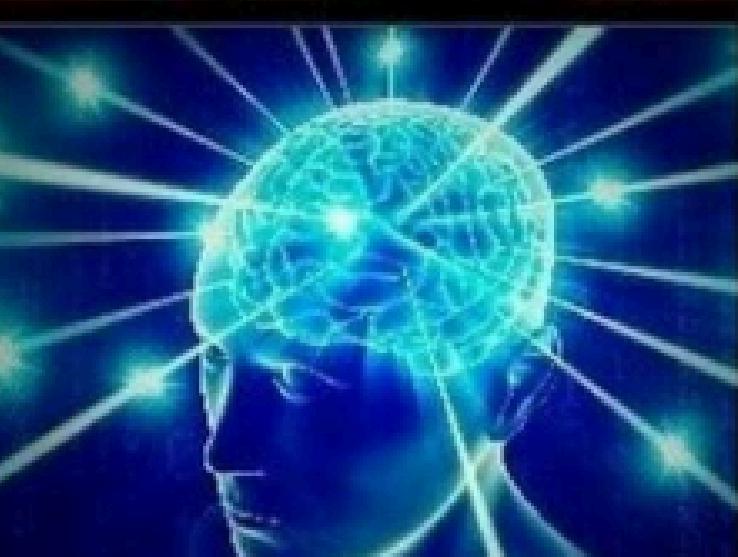
**USING AI TO  
WRITE ESSAYS**



**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