A Random Walk Down GPT Street

Presented by
David Greenwood

a light hearted exploration
of the ever accelerating
LLM landscape.

Overview

Hello!

I'm Dr David Greenwood, a Data Scientist & Systems Engineer based out of the green pastures of Wiltshire. I specialize in applying Data Science to urban regeneration, investment management & risk management use cases.

Back to Agenda Page

GPT? Why the hype?

Back to Agenda Page

"Why the hype?"

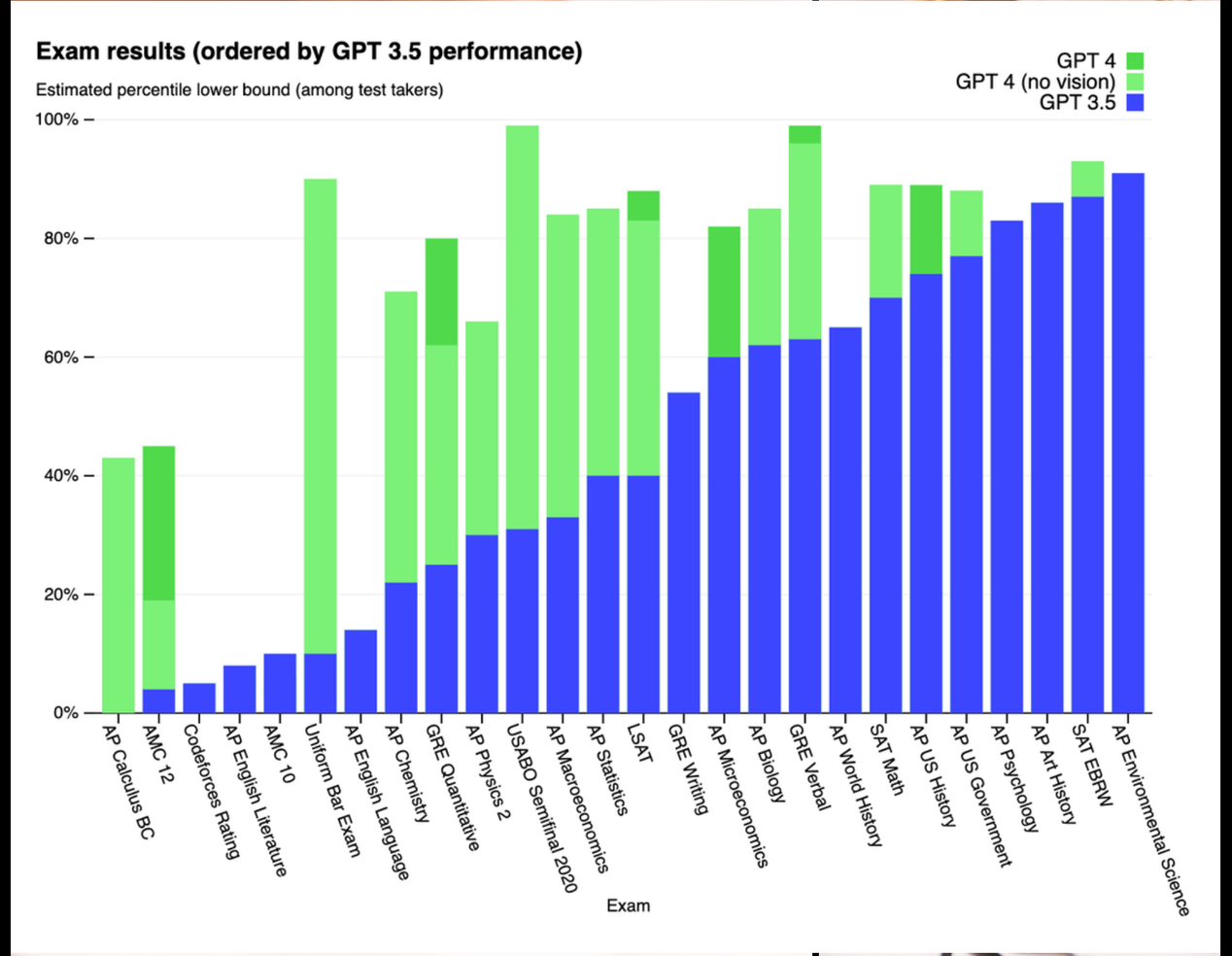
A polymath that augments your team and never gets tired.

How many people work for you?

How many have +80th percentile scores in Math, Verbal, Stats, Macroeconomics, Biology and the Bar exam?

How many are willing to work for almost free & never get tired?

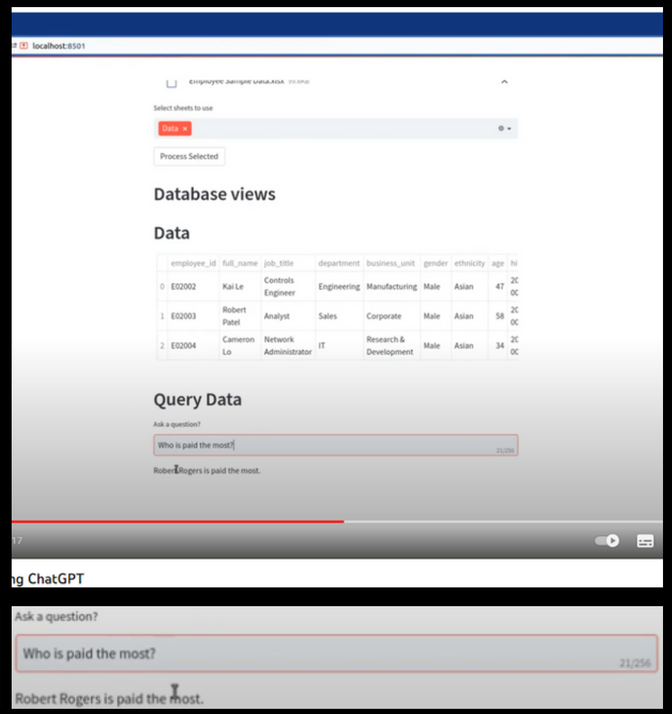
GPT-4 is a polymath, waiting for you to ask for help.



"Why the hype?"

The democratisation of Data Science

Classification

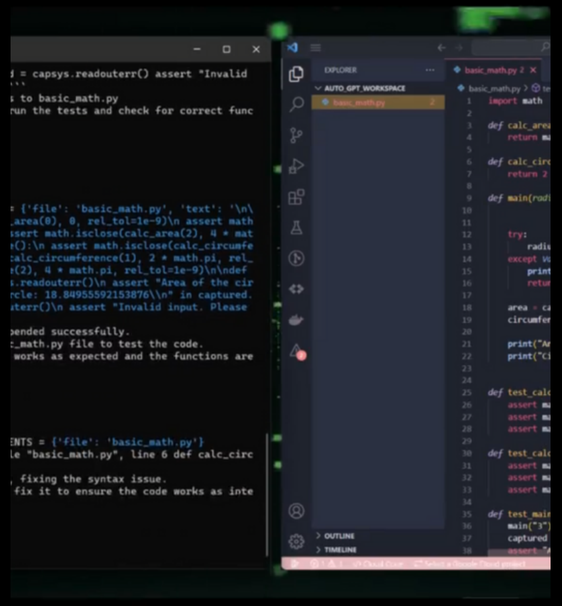


Data Analysis

Risk Analysis

"Why the hype?"

Democratisation of knowledge work



AutoGPT
https://twitter.com/i/status/1642181498278408193

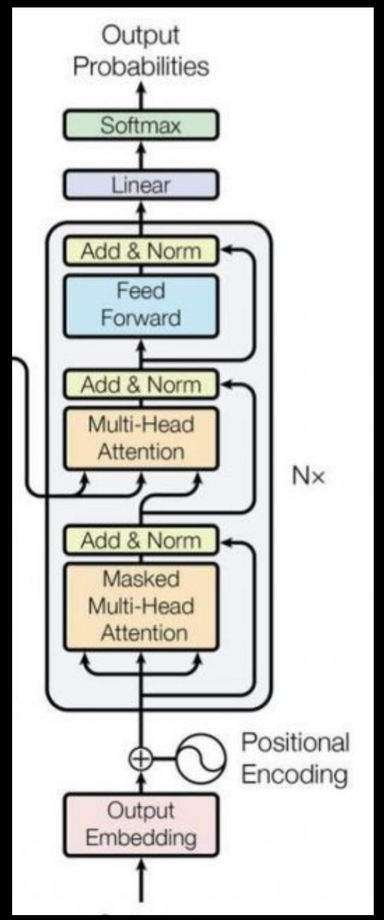
"Any sufficiently advanced technology is indistguishable from magic."

Arthur C. Clarke

What is a GPT, Why now?

Back to Agenda Page

What is GPT?



GPT

Generative

Produces content (from "joint probabilities")

Pre-trained

Comes out-of-the-box with "some ability"

Transformer

one type of input into another type of output

Back to Agenda Page

Why Now?

GPU's
+
novel Transformer architectures that are highly parallelizable
=
Unique moment in history

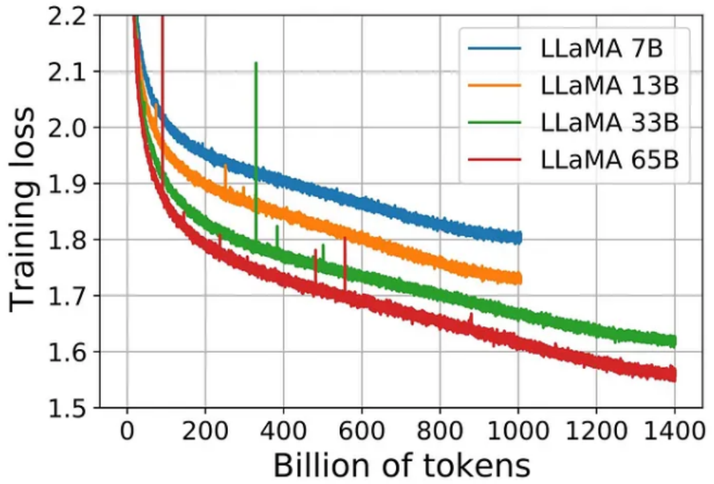
https://towardsdatascience.com/behind-the-millions-estimating-the-scale-of-large-language-models-97bd7287fb6b

Release	Model	Size	Paper
2019	GPT-2	1.5B	Language Models are Unsupervised Multitask Learners
2020	GPT-3	175B	Language Models are Few-Shot Learners
2021	Gopher	280B	Scaling Language Models: Methods, Analysis & Insights from Training Gopher
2022	PaLM	540B	PaLM: Scaling Language Modeling with Pathways
2022	Chinchilla	70B	Training Compute-Optimal Large Language Models
2022	OPT	175B	OPT: Open Pre-trained Transformer Language Models
2022	BLOOM	176B	BLOOM: A 176B-Parameter Open-Access Multilingual Language Model
2022	Galactica	120B	Galactica: A Large Language Model for Science
2023	LLaMA	65B	LLaMA: Open and Efficient Foundation Language Models

Some of the popular LLMs architectures. Image by Author

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

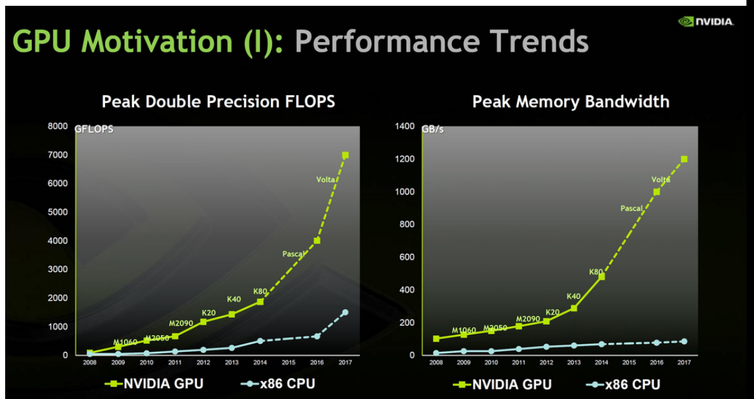
Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.



When training a 65B-parameter model, our code processes around 380 tokens/sec/GPU on 2048 A100 GPU with 80GB of RAM. This means that training over our dataset containing 1.4T tokens takes approximately 21 days.

2048 GPUs x $3.93 GPU per hour x 24 hours x 21 days =

4.05 million dollars



What is the Transformer architecture? How does it work?

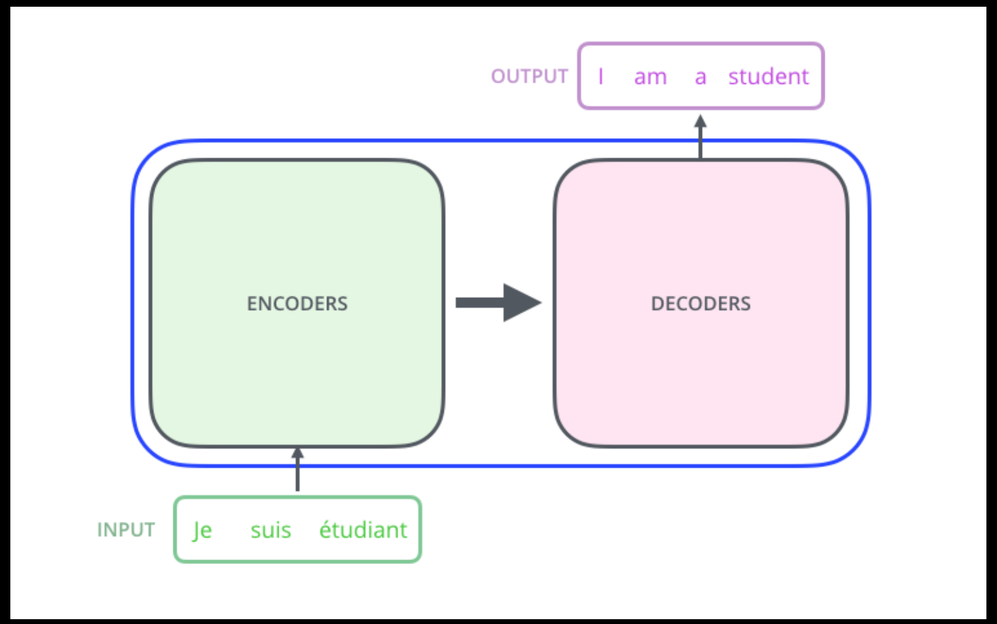
Back to Agenda Page

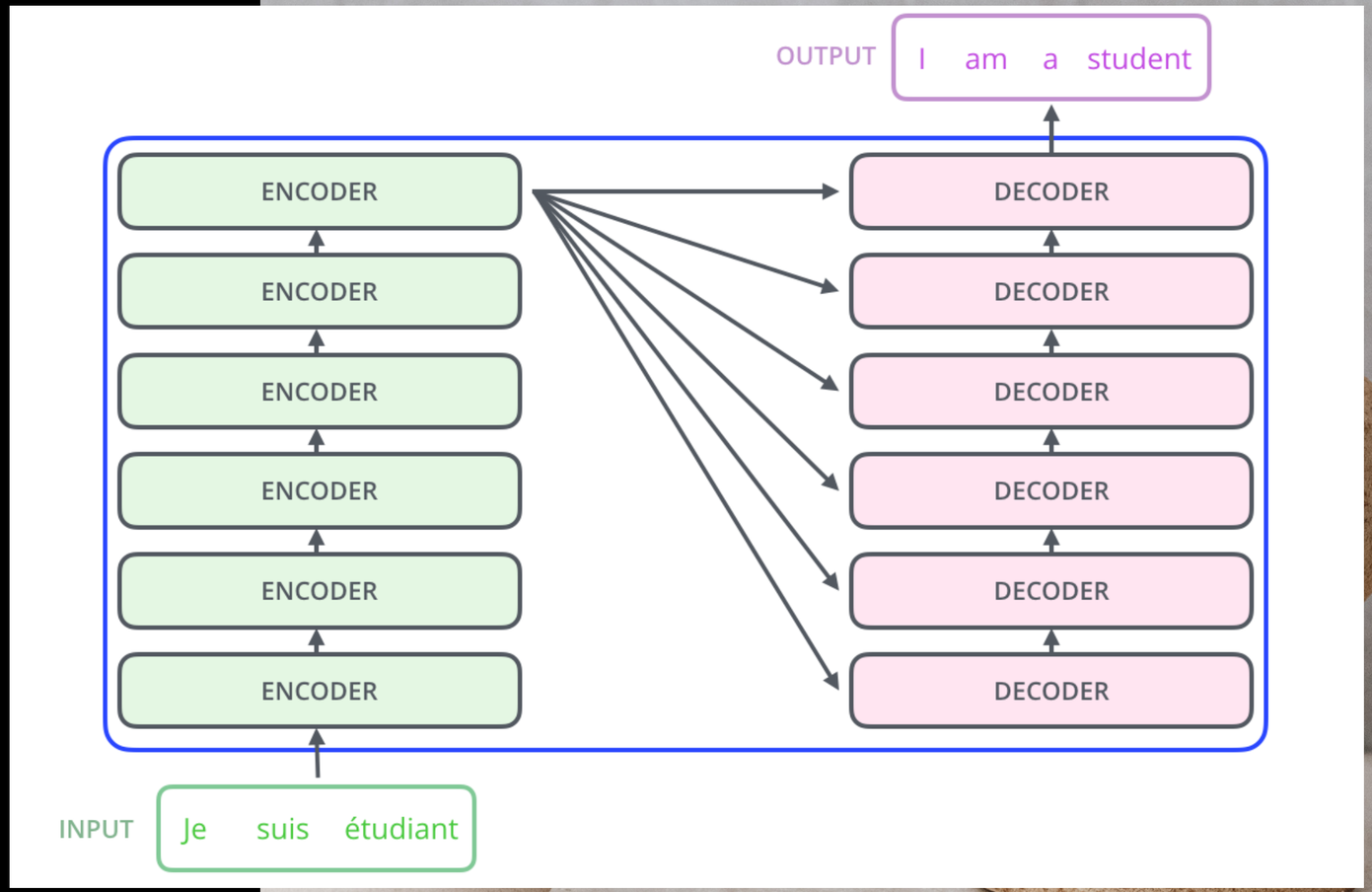
The Transformer

Input - Output

The Transformer

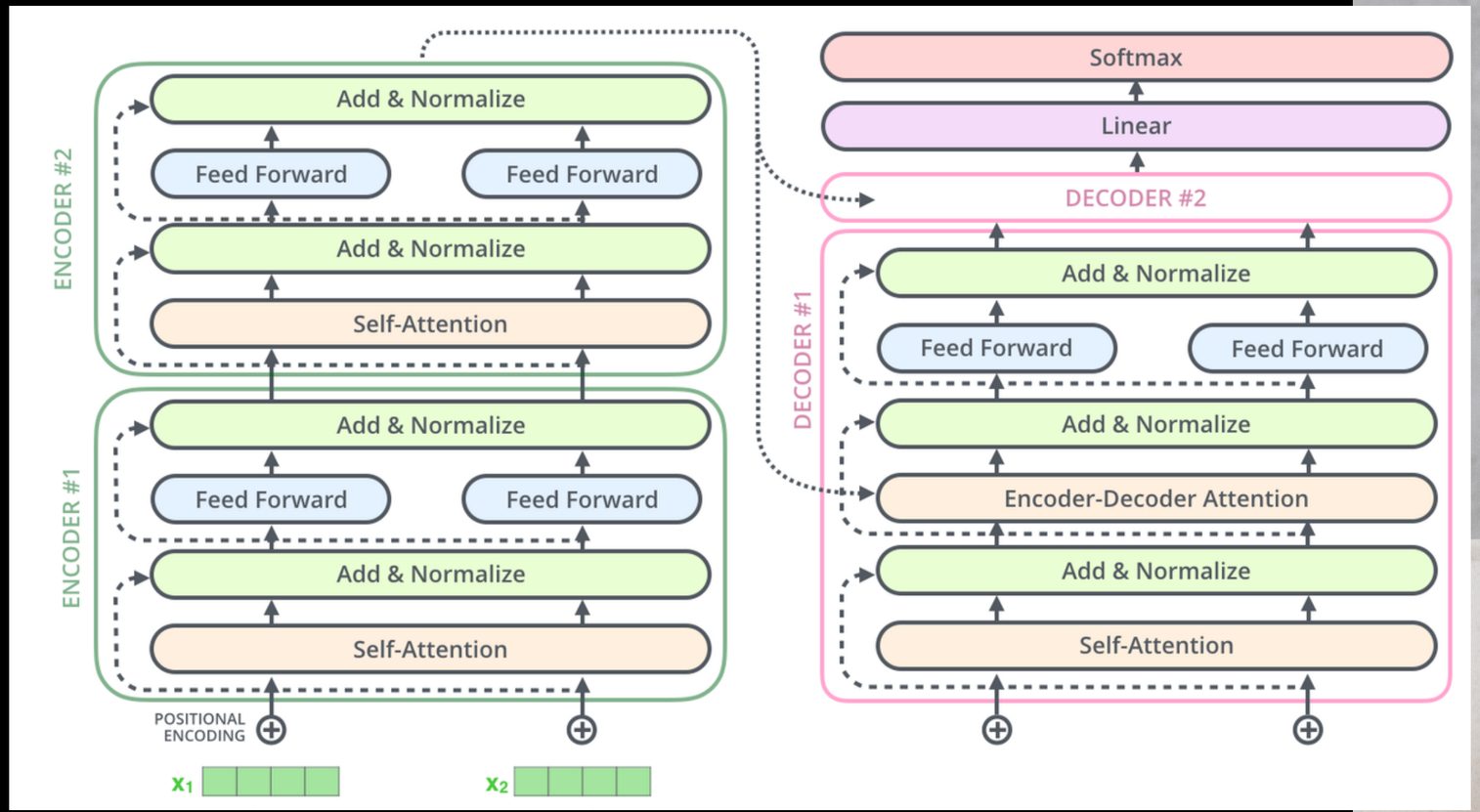
Basic Architecture

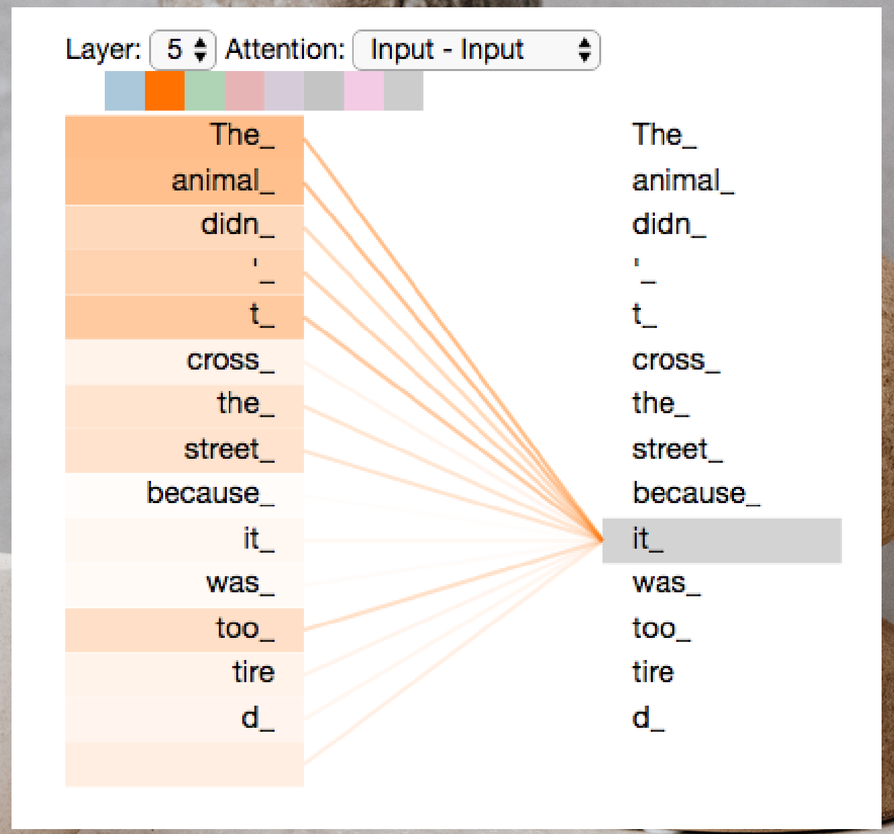




The Transformer

Architecture

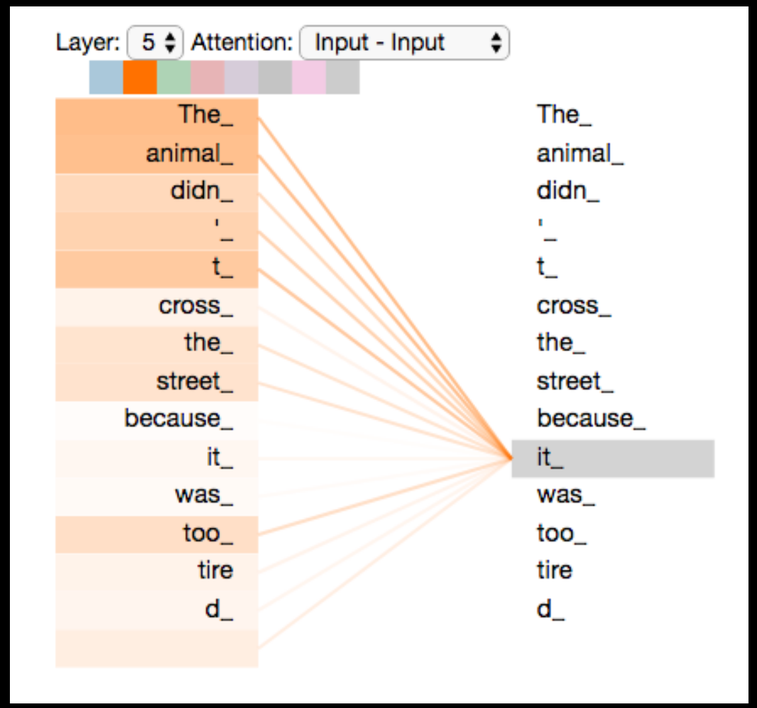


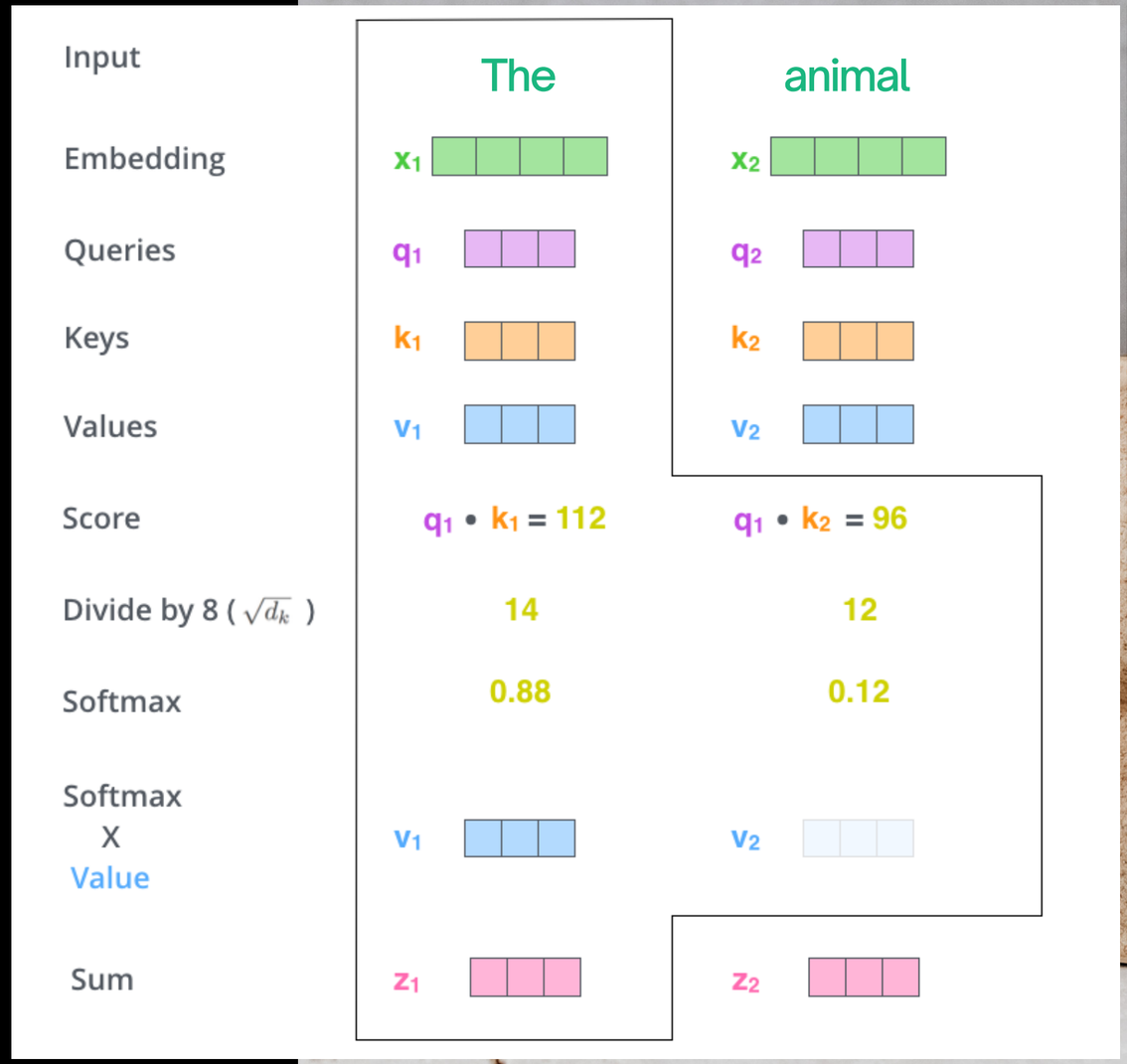


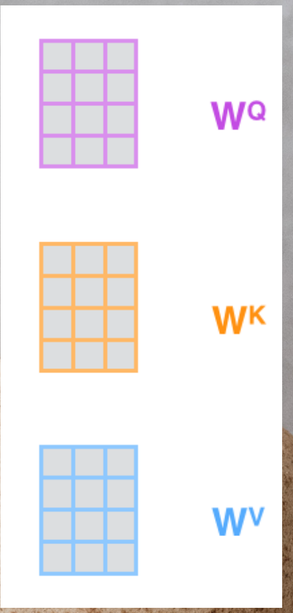
"The animal didn't cross the street because it was too tired"

The Transformer

Self Attention

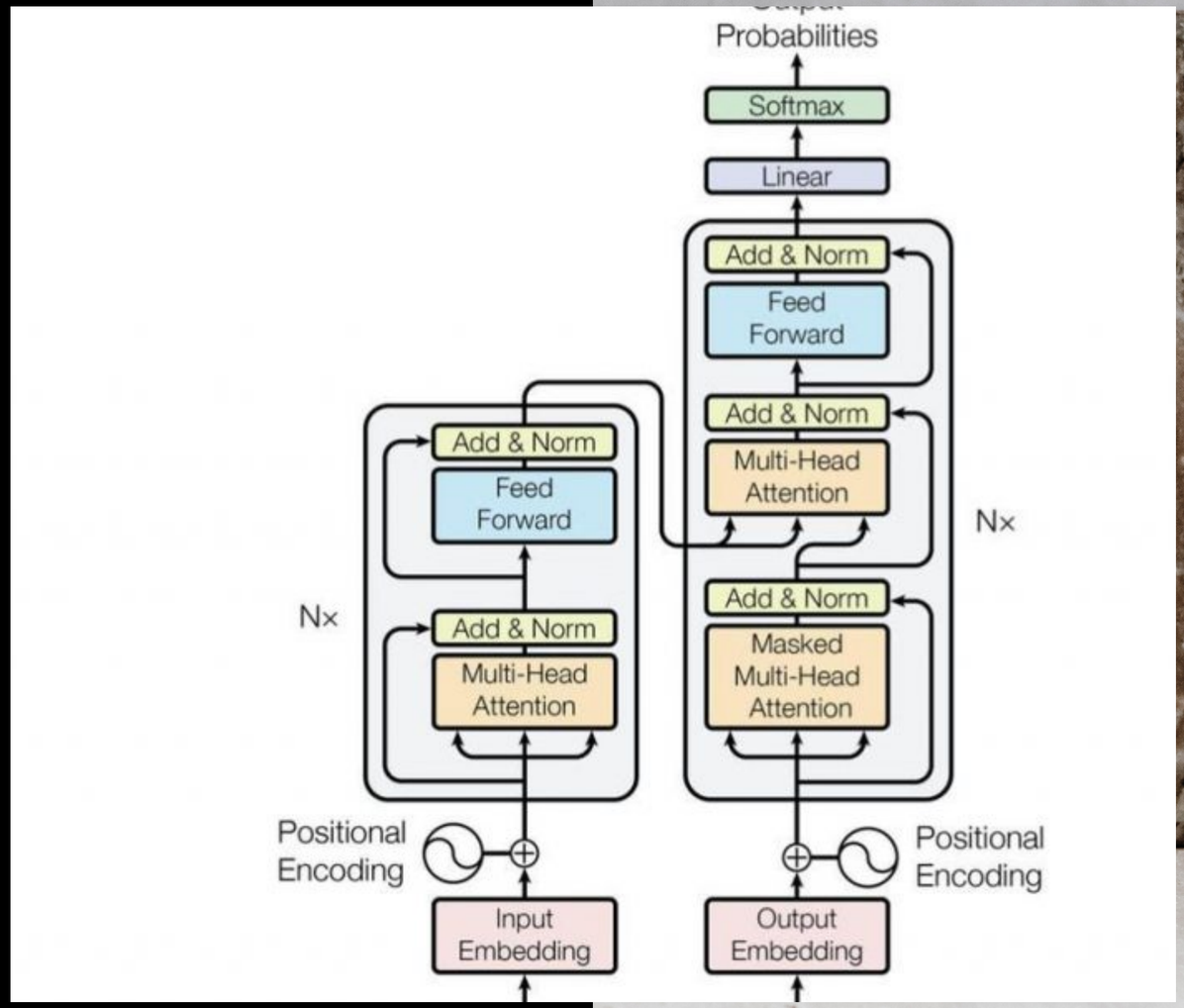






The Transformer

All Together



How do I give it a go?

Back to Agenda Page

Choices?

Different ways of "conditioning" your model depending on number of examples

Prompt engineering

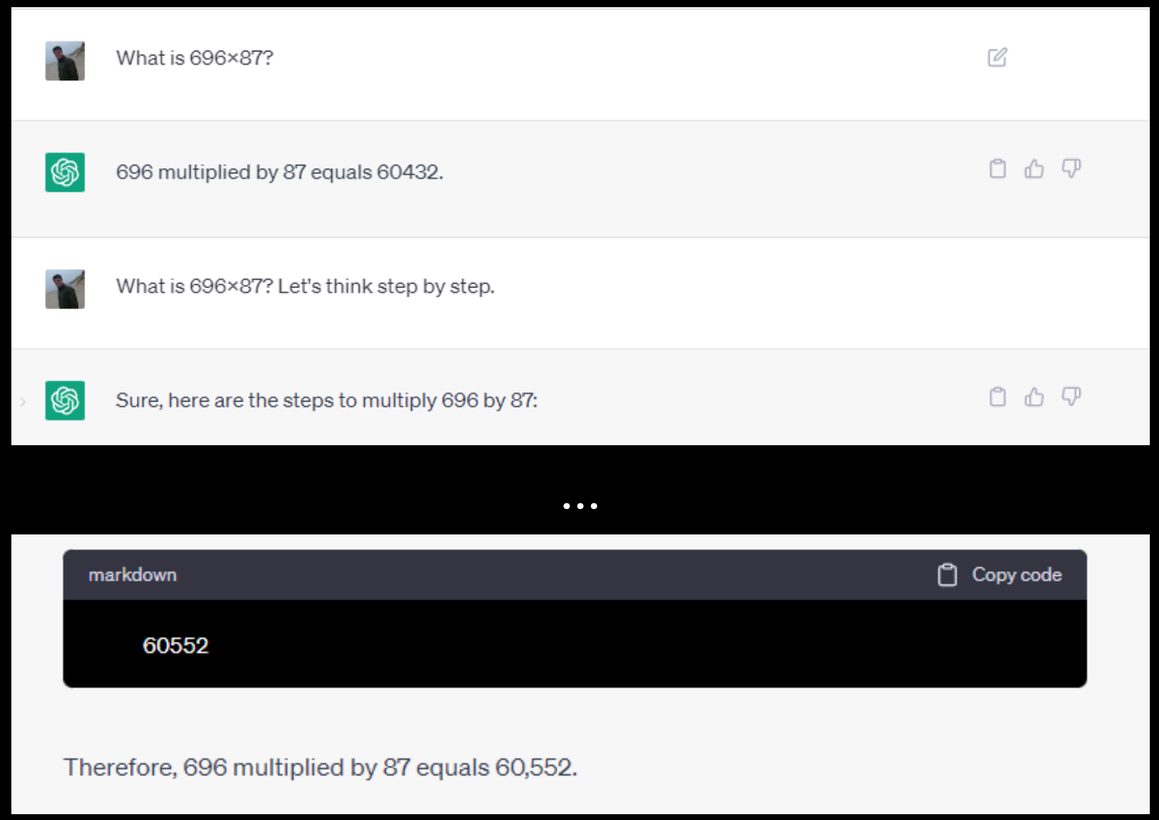
Fine tuning

```
{"prompt": "<prompt text>", "completion": "<ideal generated text>"}
{"prompt": "<prompt text>", "completion": "<ideal generated text>"}
```

Both are forms of model conditioning

Prompt Engineering Examples

Different ways of "conditioning" your model



5. Start with zero-shot, then few-shot (example), neither of them worked, then fine-tune

✅ Zero-shot

```
Extract keywords from the below text.

Text: {text}

Keywords:
```

✅ Few-shot - provide a couple of examples

```
Extract keywords from the corresponding texts below.

Text 1: Stripe provides APIs that web developers can use to integrate payment processing
Keywords 1: Stripe, payment processing, APIs, web developers, websites, mobile applicat
##
Text 2: OpenAI has trained cutting-edge language models that are very good at understan
Keywords 2: OpenAI, language models, text processing, API.
##
Text 3: {text}
Keywords 3:
```

https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api

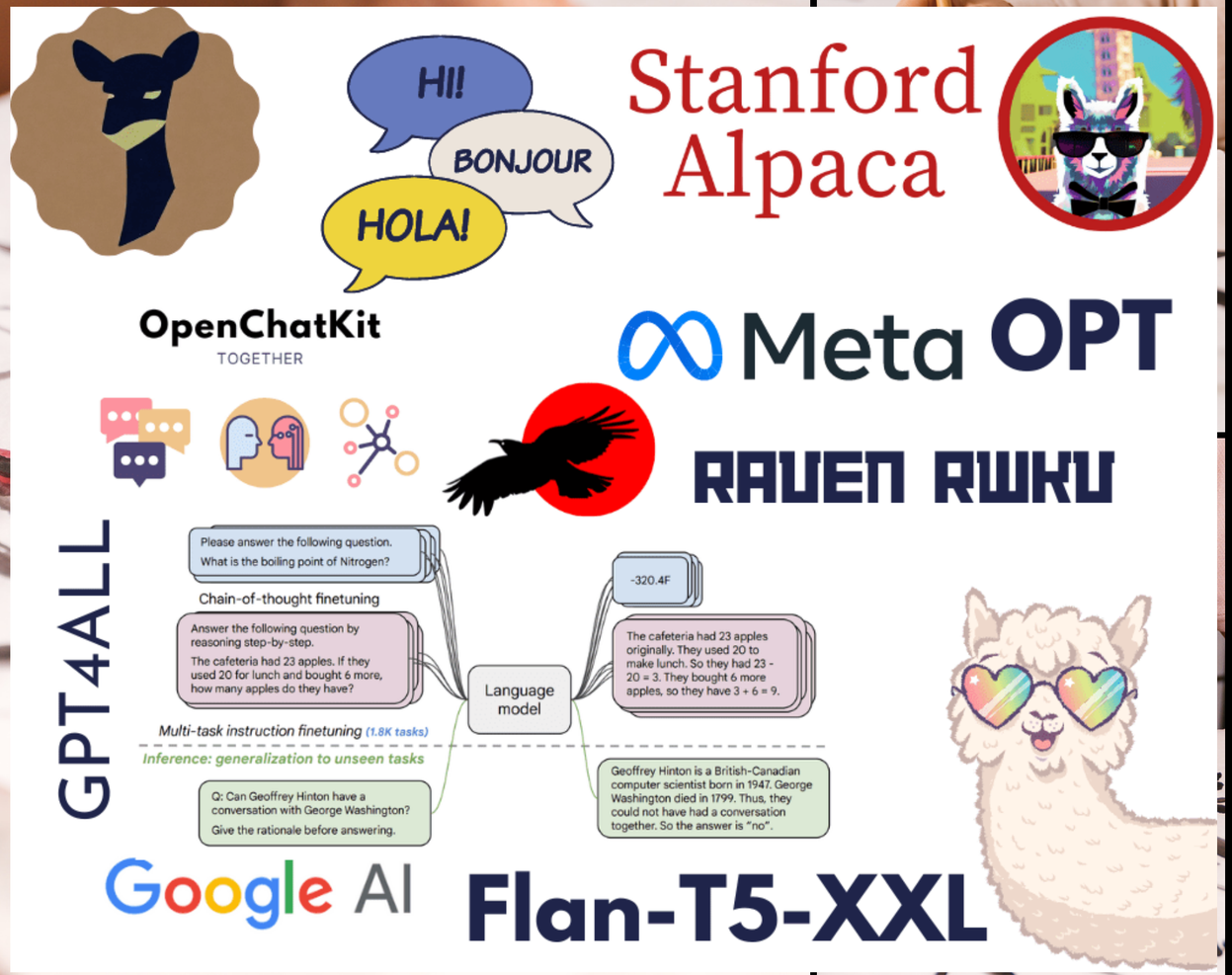
Open-source LLMs

There is a whole Zoo of stuff out there!

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	**88.0**	82.3	-	83.4	**81.1**	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	**80.0**	**57.8**	58.6
	65B	85.3	**82.8**	**52.3**	**84.2**	77.0	78.9	56.0	**60.2**

Table 3: **Zero-shot performance on Common Sense Reasoning tasks.**

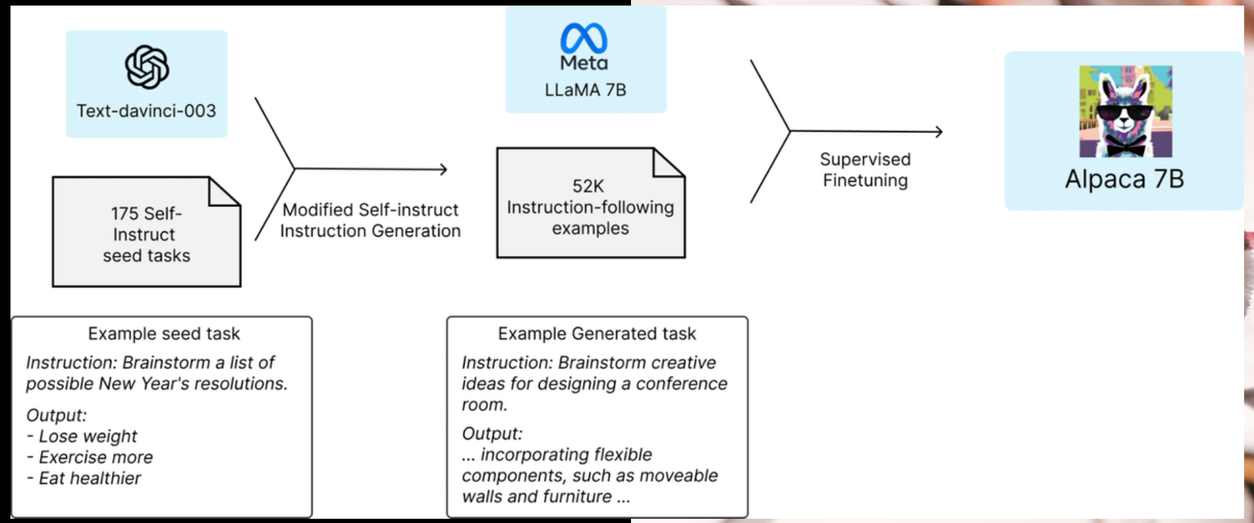
The Meta LLaMA release triggered a range of open source models from which to pick and choose.



Open-source LLMs

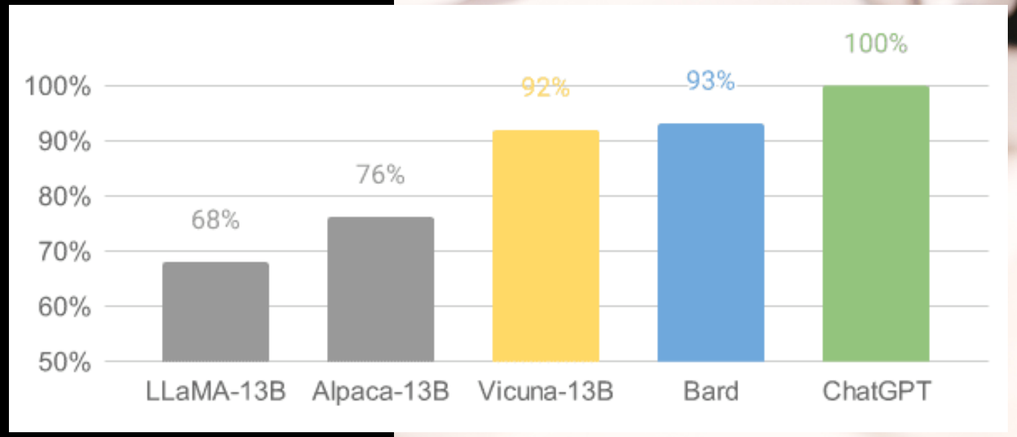
There is a whole Zoo of stuff out there!

Alpaca



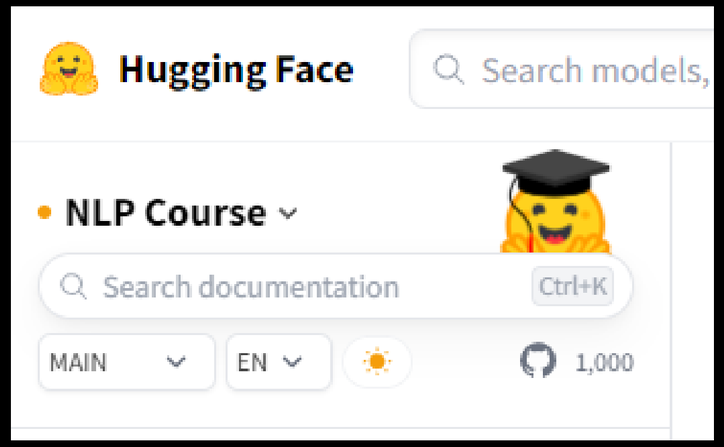
Vicuna

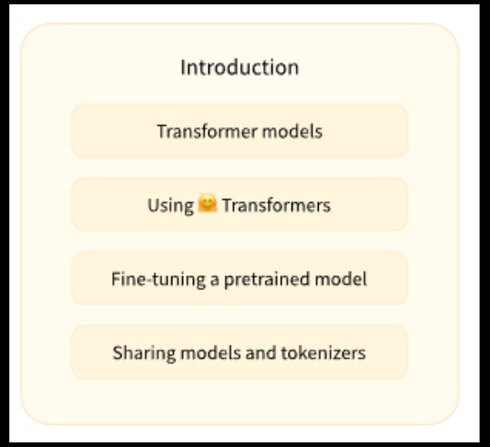
ShareGPT

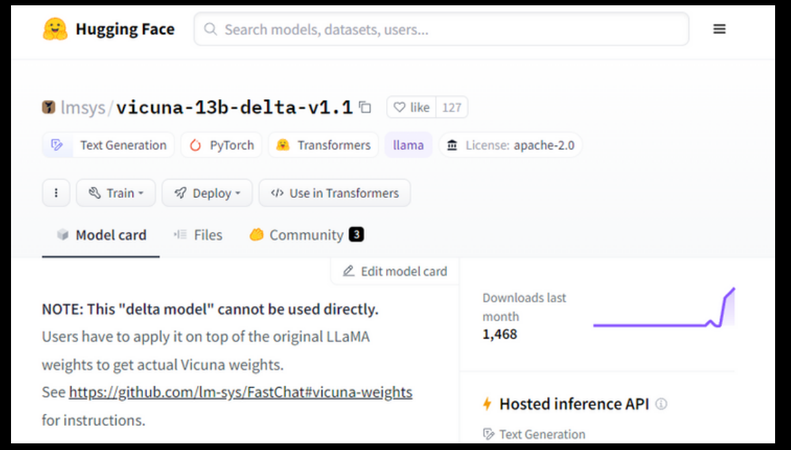


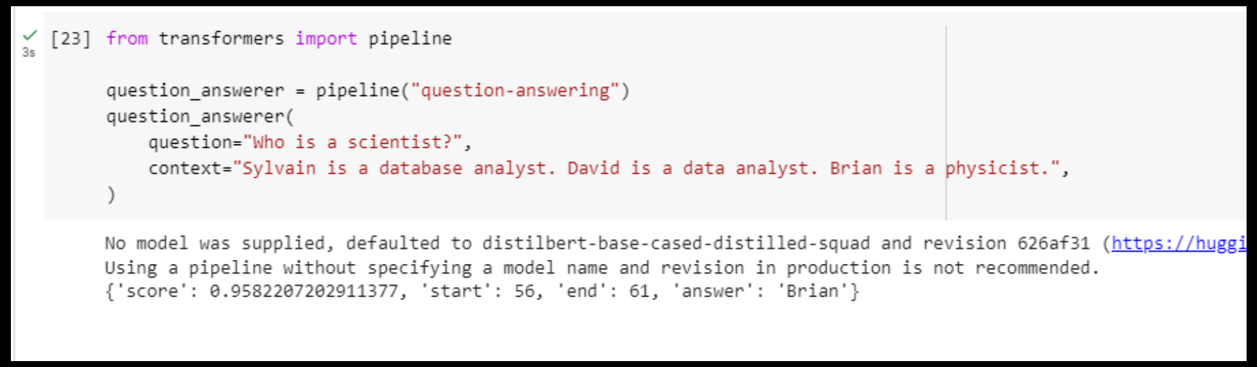
Open-source LLMs

A good place to get started?









HTTPS://HUGGINGFACE.CO/LEARN/NLP-COURSE/

GPT from the comfort of your laptop?

Back to Agenda Page

The mandatory live demo that goes tediously wrong, despite it previously working great

Resource Page

HuggingFace https://huggingface.co/learn/nlp-course/	**Understanding Large Language Models** https://magazine.sebastianraschka.com/p/understanding-large-language-models
Illustrated Transformer http://jalammar.github.io/illustrated-transformer/	**LLaMA: Open and Efficient Foundation Language Models** https://arxiv.org/abs/2302.13971
How to train your own Large Language Models https://blog.replit.com/llm-training	**Best practices for prompt engineering with OpenAI** https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api

to get in touch.

Linked-in

https://www.linkedin.com/in/dsgreenwood/

Email

dsg27@cantab.net