SPATIAL ECOLOGY

Using GPT as your coding partner

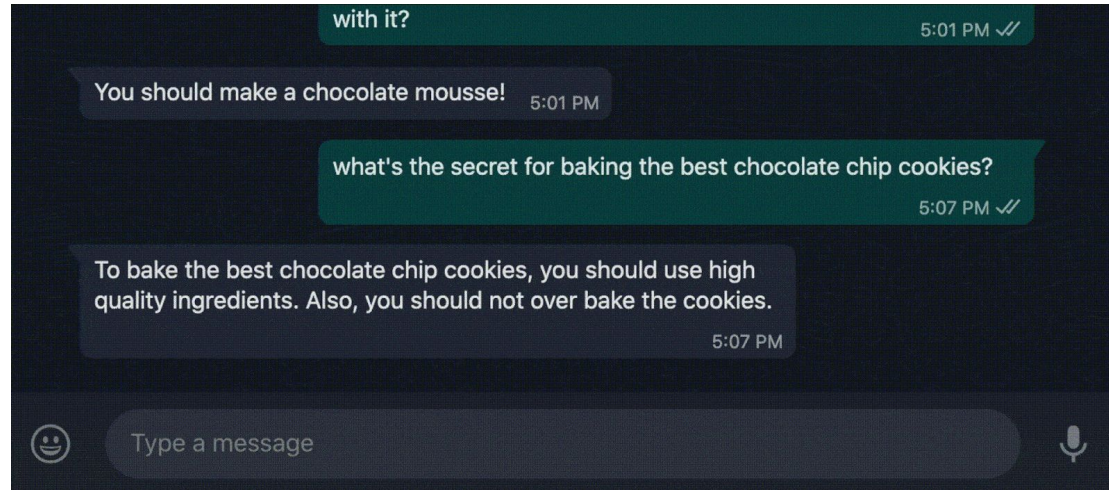Antonio Fonseca - Matera 2023

Agenda

- What is GPT?
- How does it work?
- Overview:
  - Pretraining
  - Prompting
  - Inference
- Tutorial

Generative AI and large datasets

Dall-E: generate image from prompt

ChatGPT: conversational agent

Introduction to GPT

What is GPT?

- GPT stands for Generative Pre-trained Transformer.
- It is a language model developed by OpenAI that uses deep learning to generate human-like text.
- Tasks:
    - Language translation
    - Text summarization
    - PowerPoint presentations (used to create these slides)
    - Coding

How does GPT work?

The basics

-   GPT uses a neural network to analyze large amounts of text data and learn patterns in language.
-   Once trained, it can generate text based on a given prompt.
-   The more data it is trained on, the better it becomes at generating human-like text.

GPT Overview

- Estimated cost of pretraining GPT-3: 355 GPU years (NVIDIA Tesla V100 GPU) and cost $4.6m.
- 175B parameters
- Massive training set
- Up to 2048 tokens (words) for context
- Main steps:
  - Pretraining,
  - Prompting
  - Inference
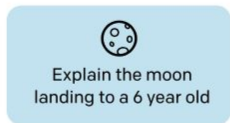
GPT-3 training data[1]:9

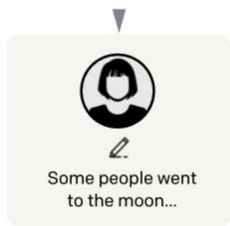| Dataset	# tokens	Proportion within training
Common Crawl	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

Step 1

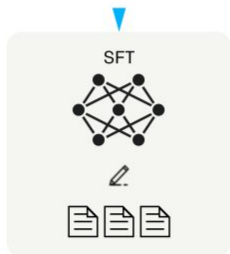**Collect demonstration data, and train a supervised policy.**

A prompt is sampled from our prompt dataset.
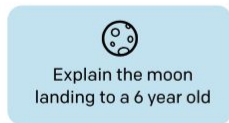
A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

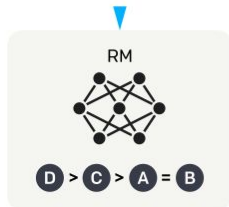**Collect comparison data, and train a reward model.**

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

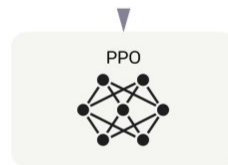This data is used to train our reward model.

Step 3

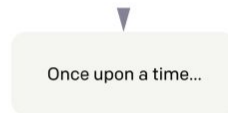**Optimize a policy against the reward model using reinforcement learning.**
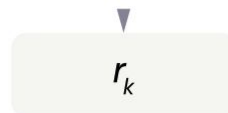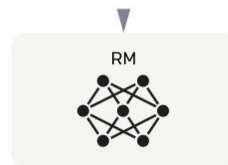
A new prompt is sampled from the dataset.

The policy generates an output.

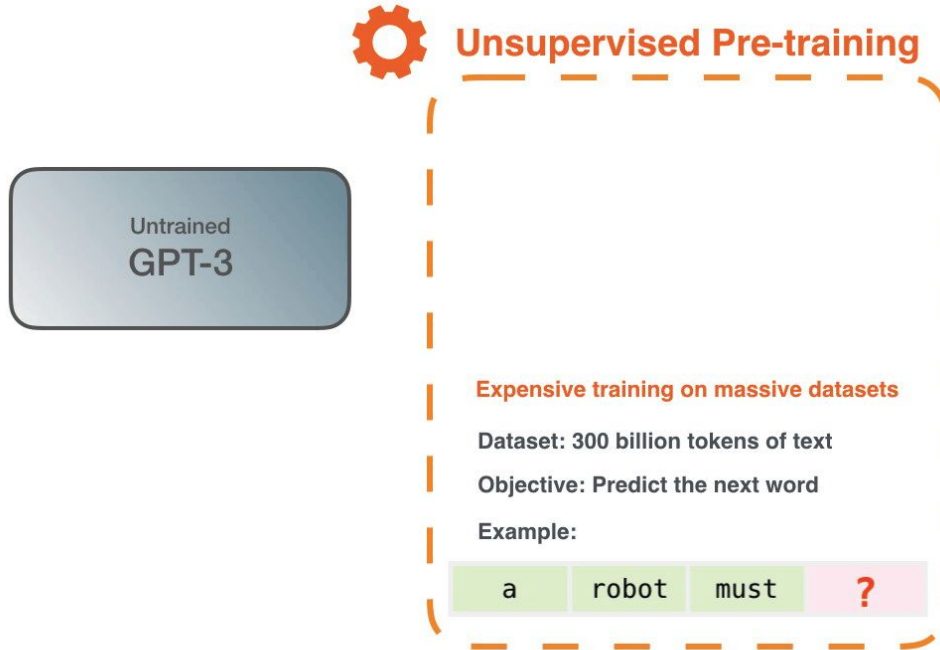The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

GPT Pretraining

**Unsupervised Pre-training**

Untrained
GPT-3

Expensive training on massive datasets

Dataset: 300 billion tokens of text

Objective: Predict the next word

Example:

| a | robot | must | ? |

GPT Pretraining

Text: Second Law of Robotics: A robot must obey the orders given it by human beings

↓

Generated training examples

Example #	Input (features)						Correct output (labels)
1	Second	law	of	robotics	:		a
2	Second	law	of	robotics	:	a	robot
3	Second	law	of	robotics	:	a	robot → must

...

GPT Pretraining

Unsupervised Pre-training

Input **(features)** | a | robot | must

Correct output (label): obey

GPT-3 (under training)
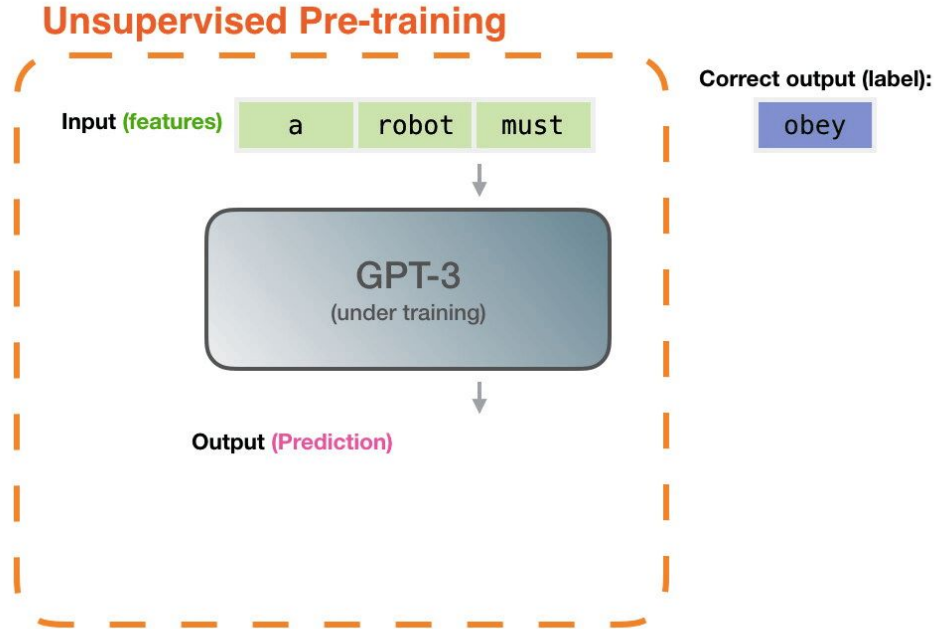
Output **(Prediction)**

Prompting

Input Prompt: Recite the first law of robotics

GPT-3

Output:

Inference: next token prediction

1 2 3 4 5 6   **... 2048**

Input: Recite the first law of robotics

GPT-3
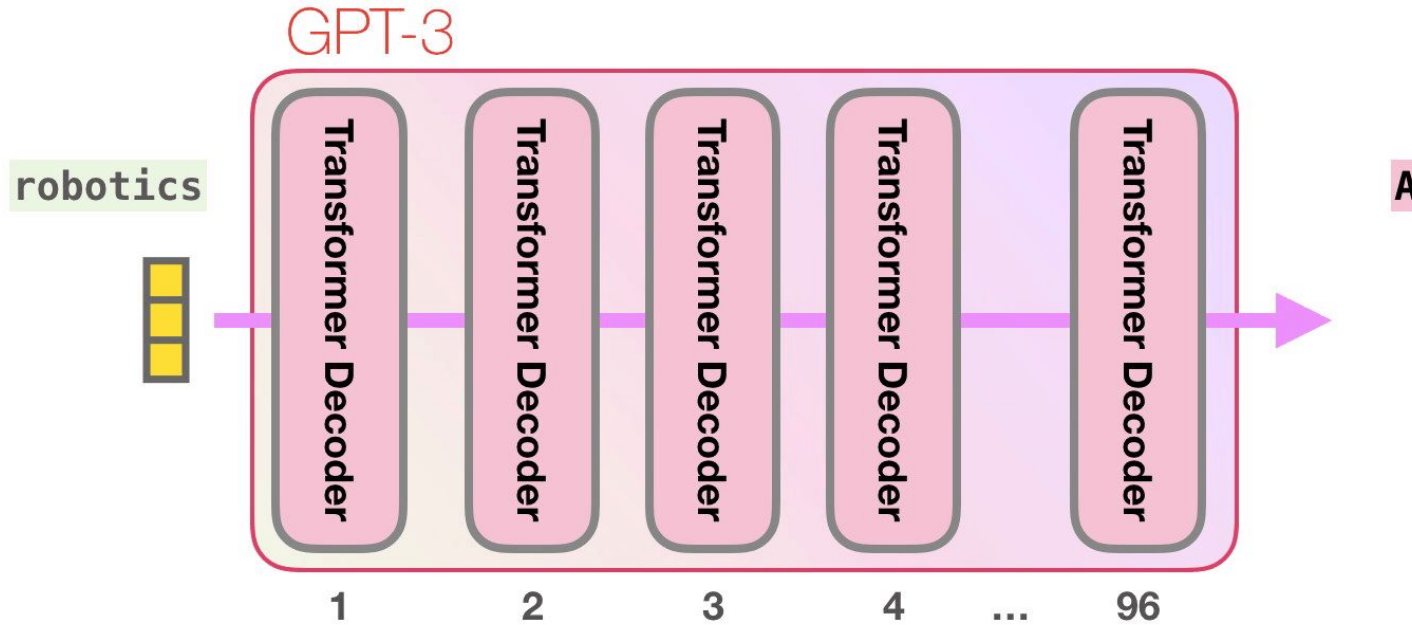
Output:
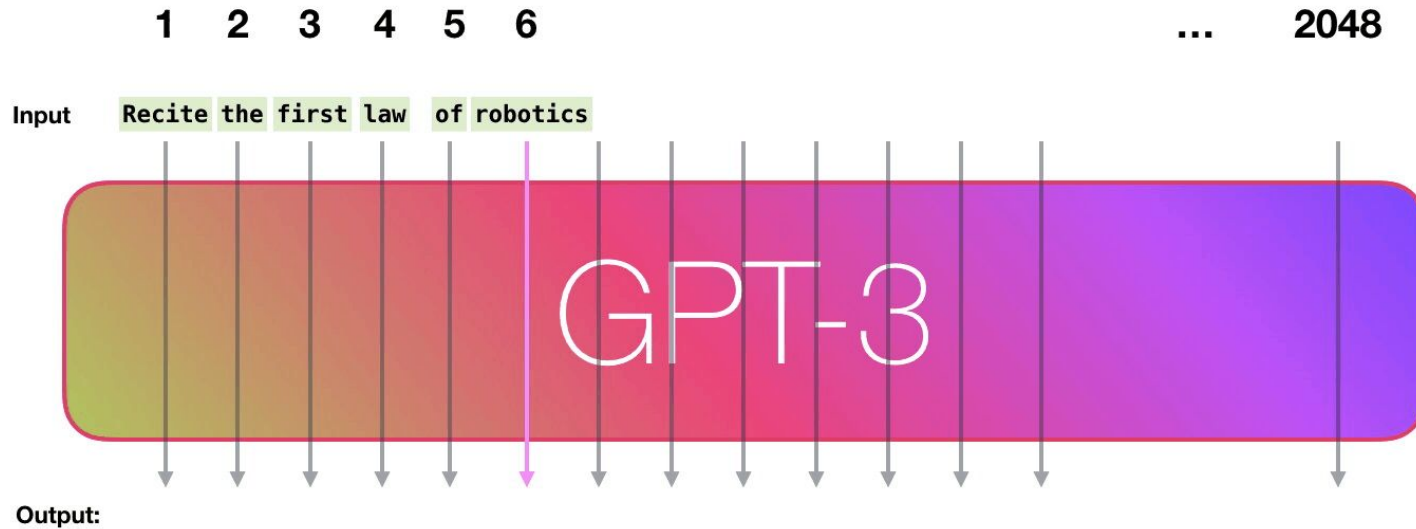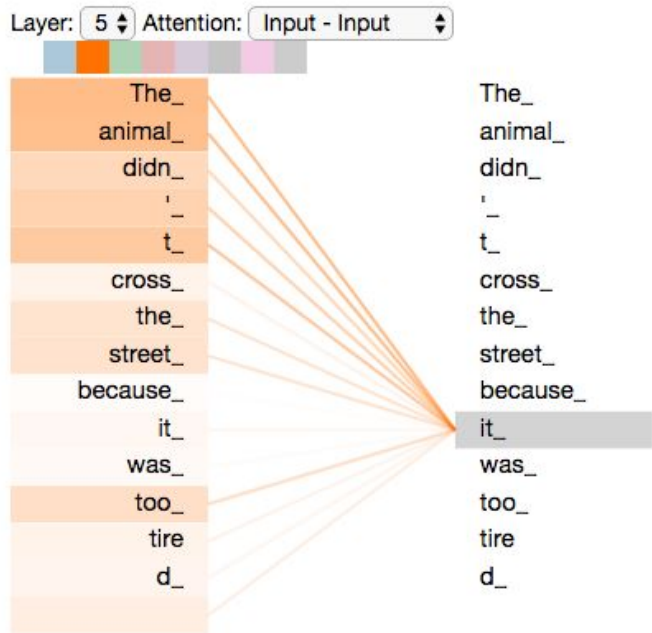
The "magic" (Transformers)

Transformer's secret sauce: self-attention

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

What does "it" in this sentence refer to?

Self-attention allows it to look at other positions in the input sequence for clues that can help lead to a better encoding for this word.

GPT-4

The latest milestone

- OpenAI has recently released GPT-4, the latest version of the language model.
- GPT-4 is a large (100T parameters) multimodal model that can accept both text and images as input.
- It is capable of generating text outputs in a variety of domains, including natural language and code
- Generates code in a variety of programming languages, including Python, JavaScript, and HTML.