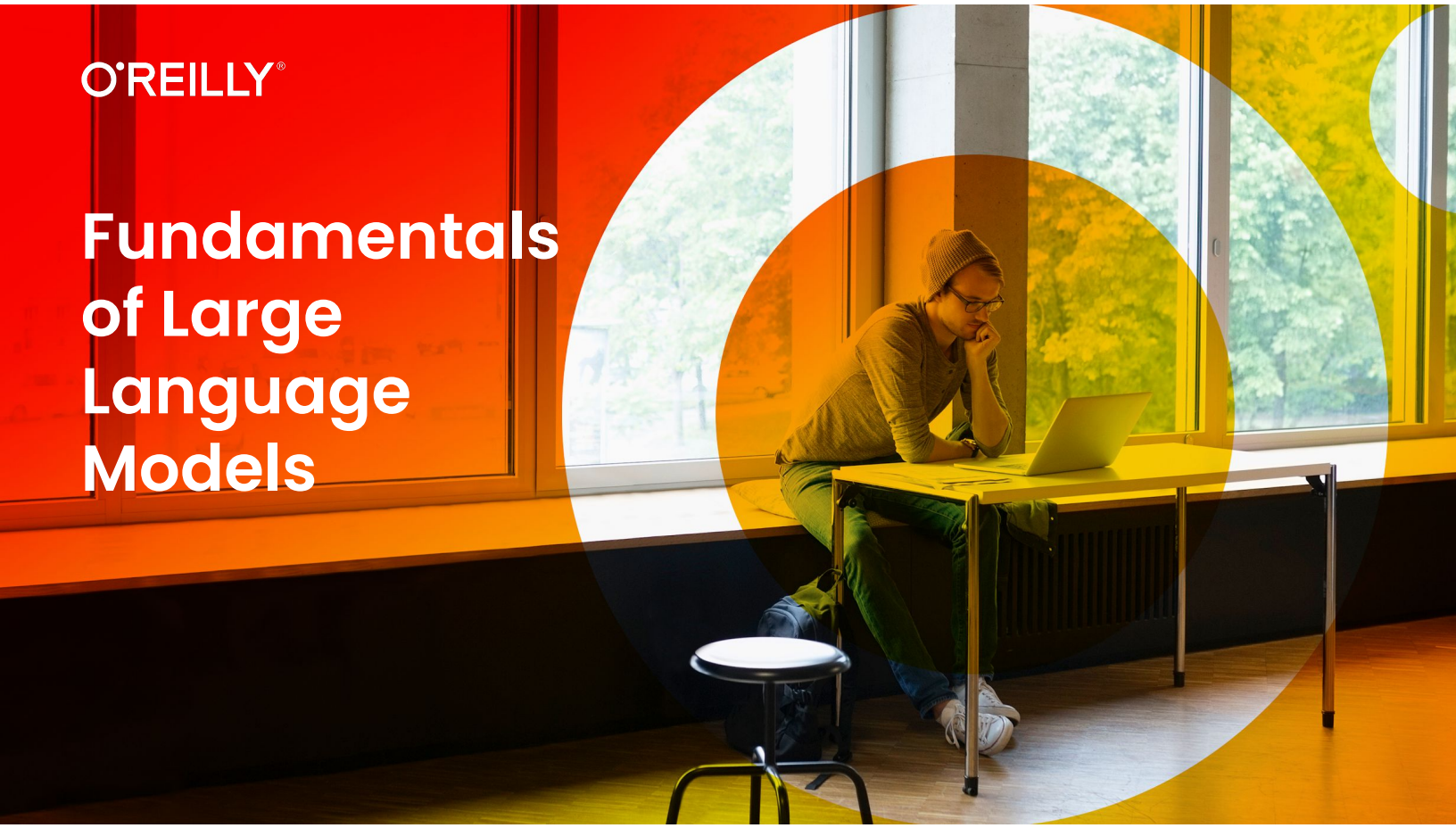
O'REILLY

Fundamentals of Large Language Models



Week 1 (today)

GPT-3 and why we needed GPT-4

- GPT-3, GPT-3.5, GPT-4, GPT-4-turbo
- Limitations of GPT-4
- RLHF
- Q&A
- Break

Transfer learning and fine-tuning

- BERT
- Transfer learning and fine-tuning
- Q&A
- Break

Transformer architecture overview

- Encoders and decoders
- Attention mechanism
- Q&A

Week 2 (next week)

LLM embeddings lab

- What are they?
- Exercises and demos
- Q&A
- Break

Benchmarks

- LLMs - HELM
- Q&A
- Break

LLMs 1-2 years after GPT-3

- Scaling laws Chinchilla
- BIG-Bench
- PaLM
- OPT and BLOOM and Llama2
- Mistral
- Q&A

This online training is always being updated.
Previous versions:

- **Falcon -> Llama 2 -> Mistral 7B**
- **GPT-4 -> GPT-4-turbo (today)**

Live Course

Hands-on GPT-4-Turbo

With Jonathan Fernandes

3h 0m April 25 • 5pm-8pm GMT+1

Live Course

Hands-on Retrieval Augmented Generation (RAG)

With Jonathan Fernandes

3h 0m April 29 • 5pm-8pm GMT+1

About me

Jonathan Fernandes Get verified

Generative AI | Large Language Models | NLP

United Kingdom · Contact info

University of Warwick - Warwick Business School

What does GPT stand for?

👍 Generative Pre-trained Transformer
👎 General Pre-trained Transformer

What are the parameters for a Large Language Model?
👍 The size of the model
👎 The variables that get adjusted during the training

What are tokens for a Large Language Model?

👍 These are subwords
👎 This is another word for parameter

What is the size of the GPT-4 model?

👍 175B
👎 This information wasn't released

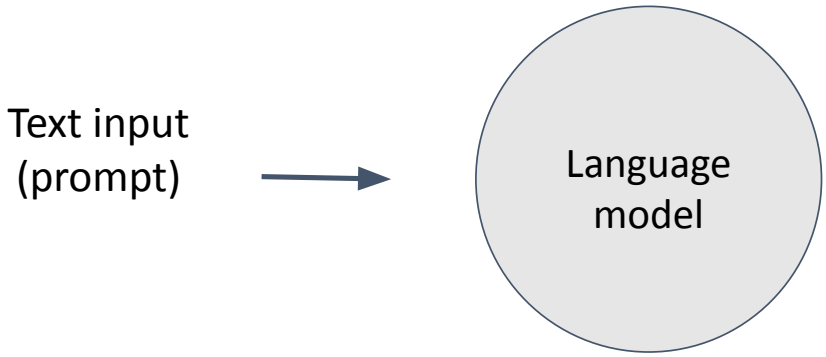
How is GPT-4 different from GPT-3 and GPT-3.5?

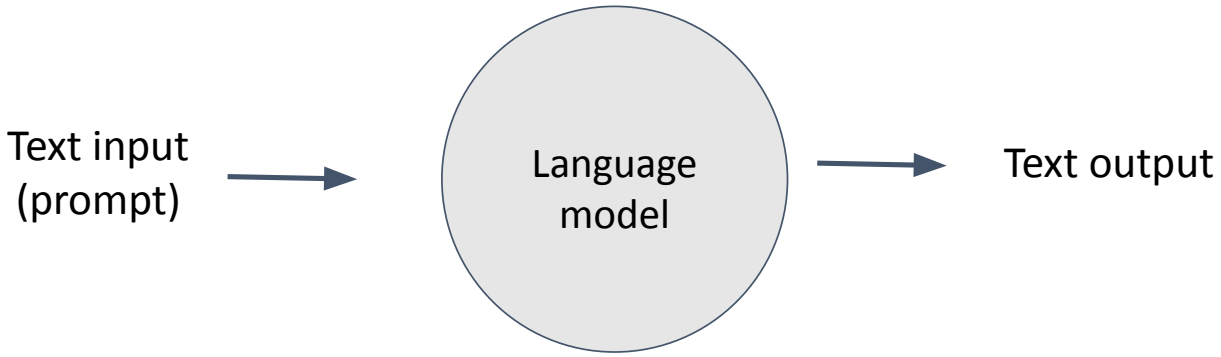
👍 GPT-4 is multi-modal
👎 GPT-4 is a decoder-based model

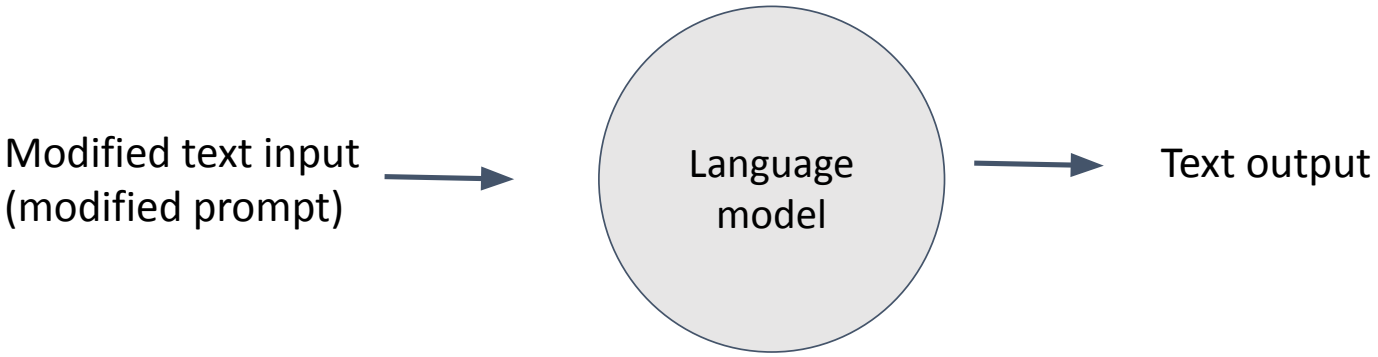
How can you ensure that GPT-4 won't hallucinate?

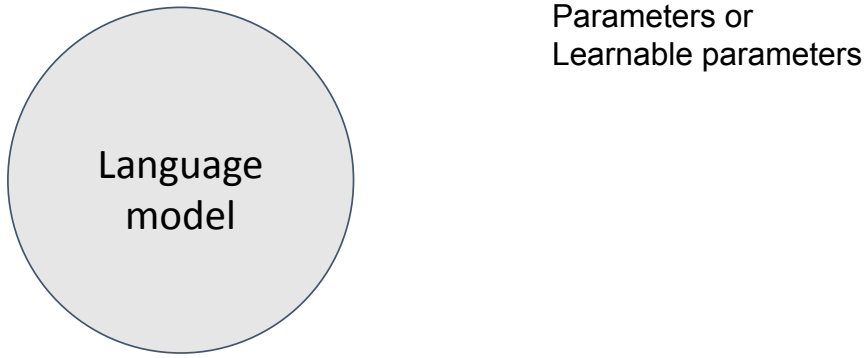
👍 Give it more training data
👎 No way to do this at the moment.

What Are Large Language Models and GPT?

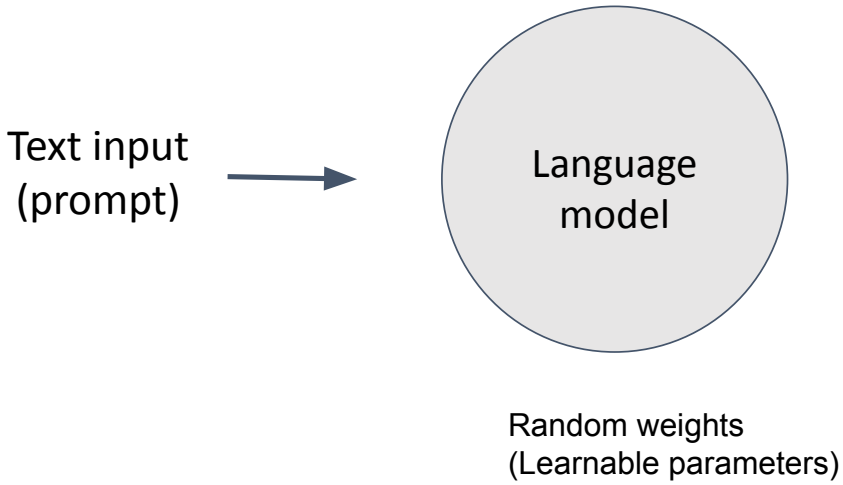



Text input
(prompt)
Language
model
Text output

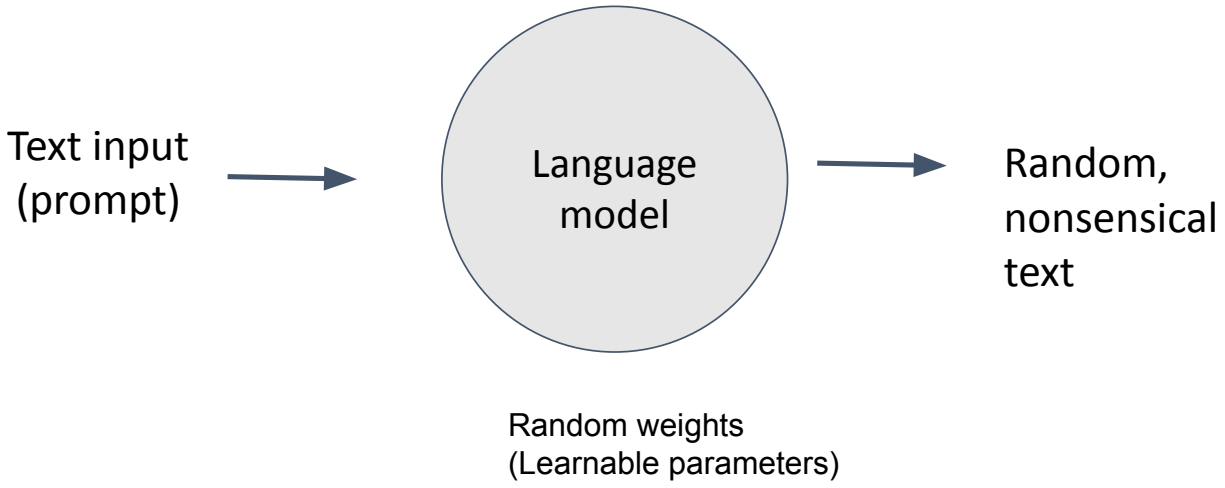

Modified text input
(modified prompt)
Language
model
Text output



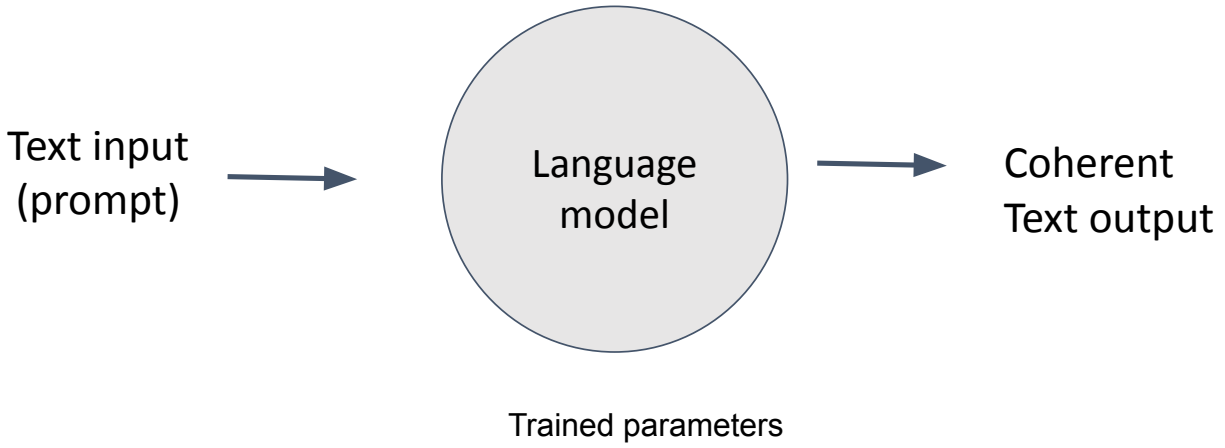
Initial model (Before training Language model)



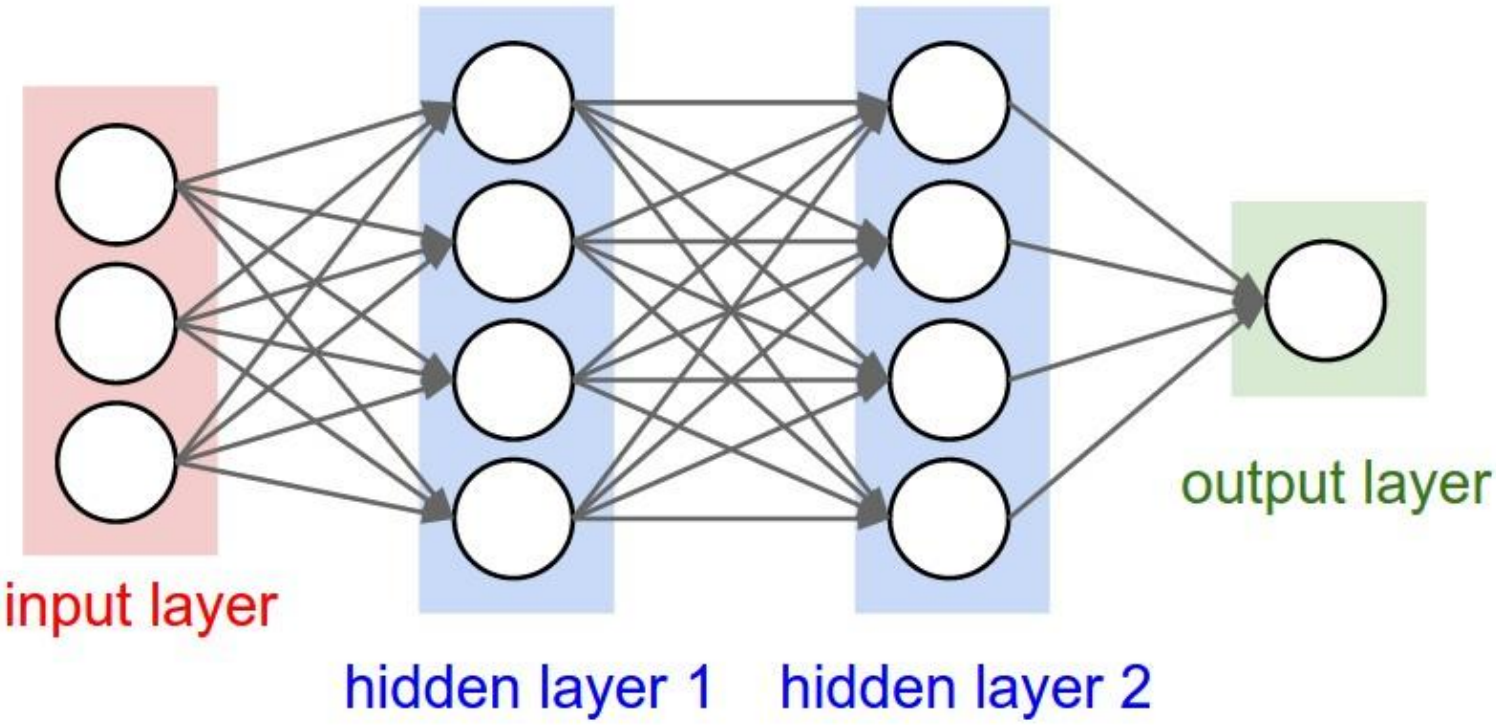
Initial model (Before training Language model)

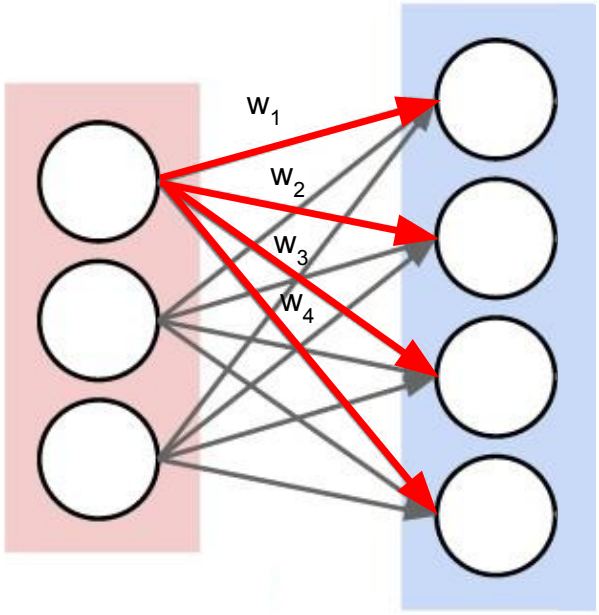


After training



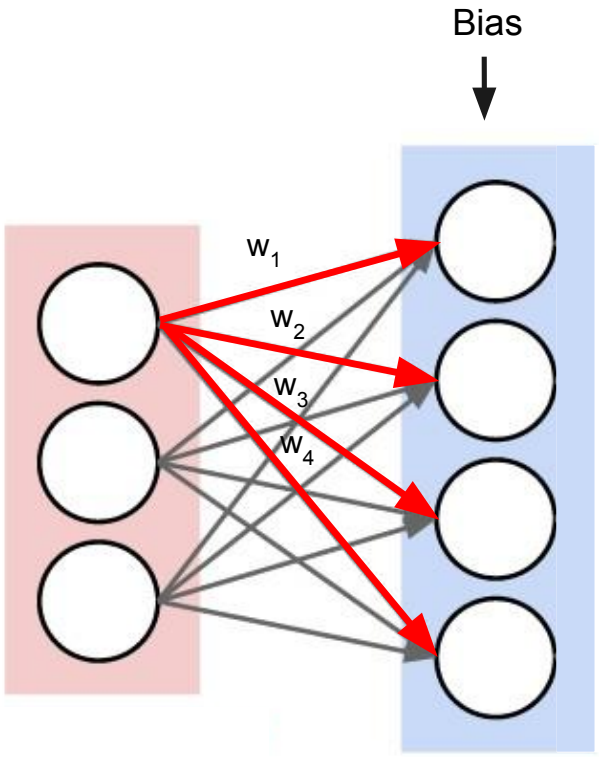
What are these (learnable) parameters?



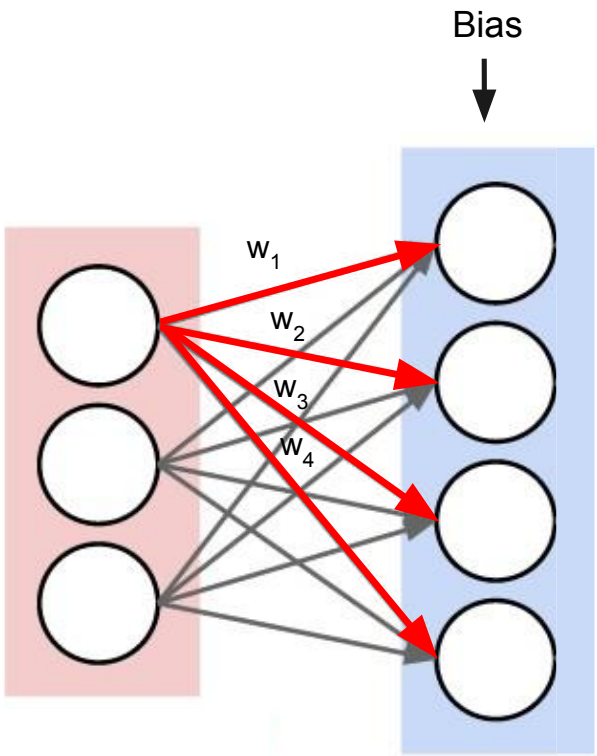


Weights:

Inputs x Outputs

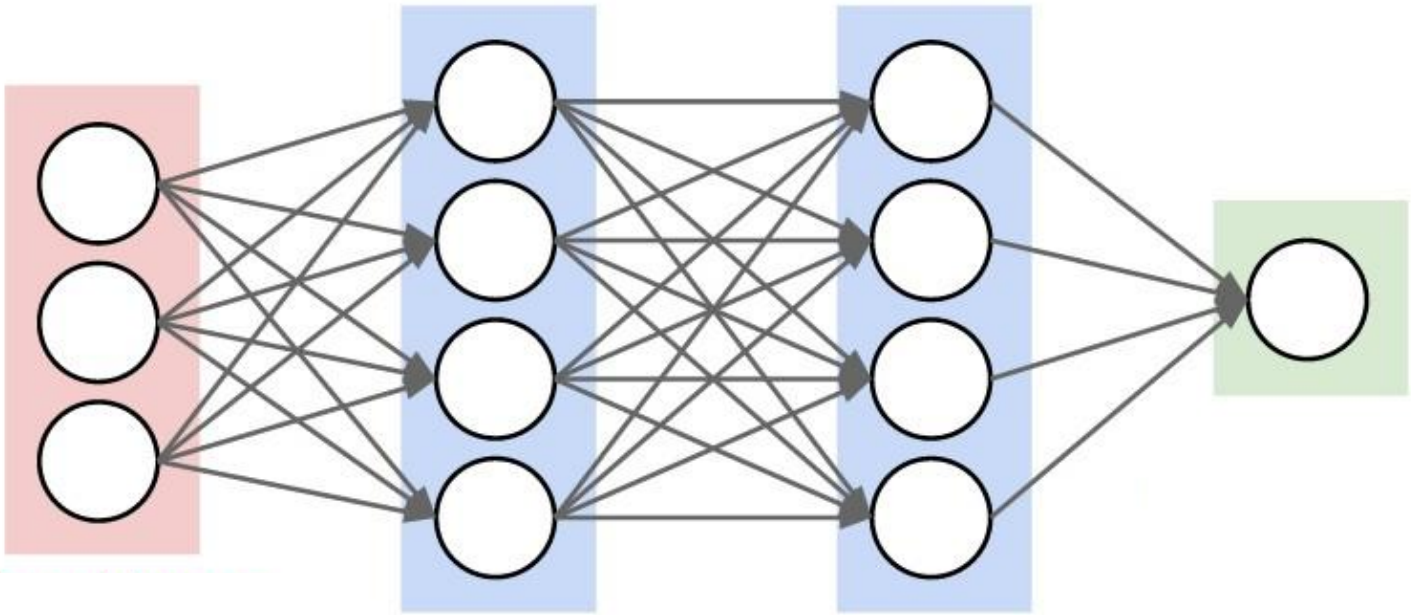


Bias



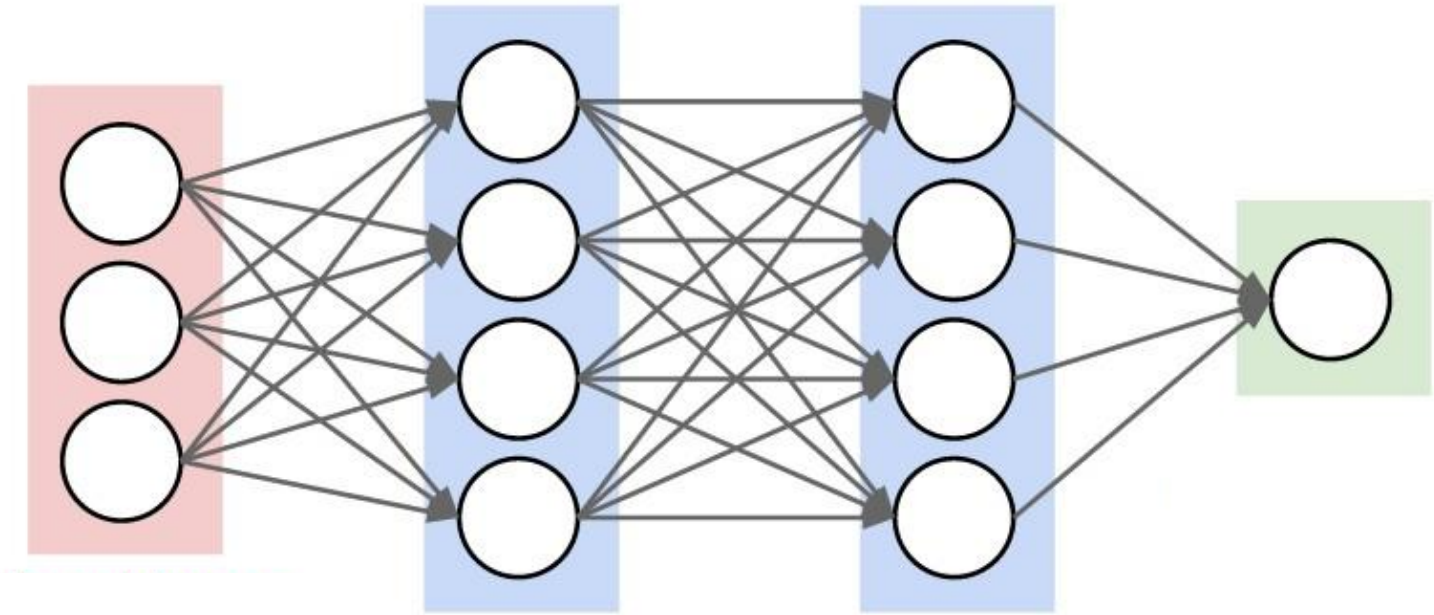
Number of parameters:
Number of weights + Biases
(3 x 4) + 4

3x4 + 4
16



3x4 + 4
16

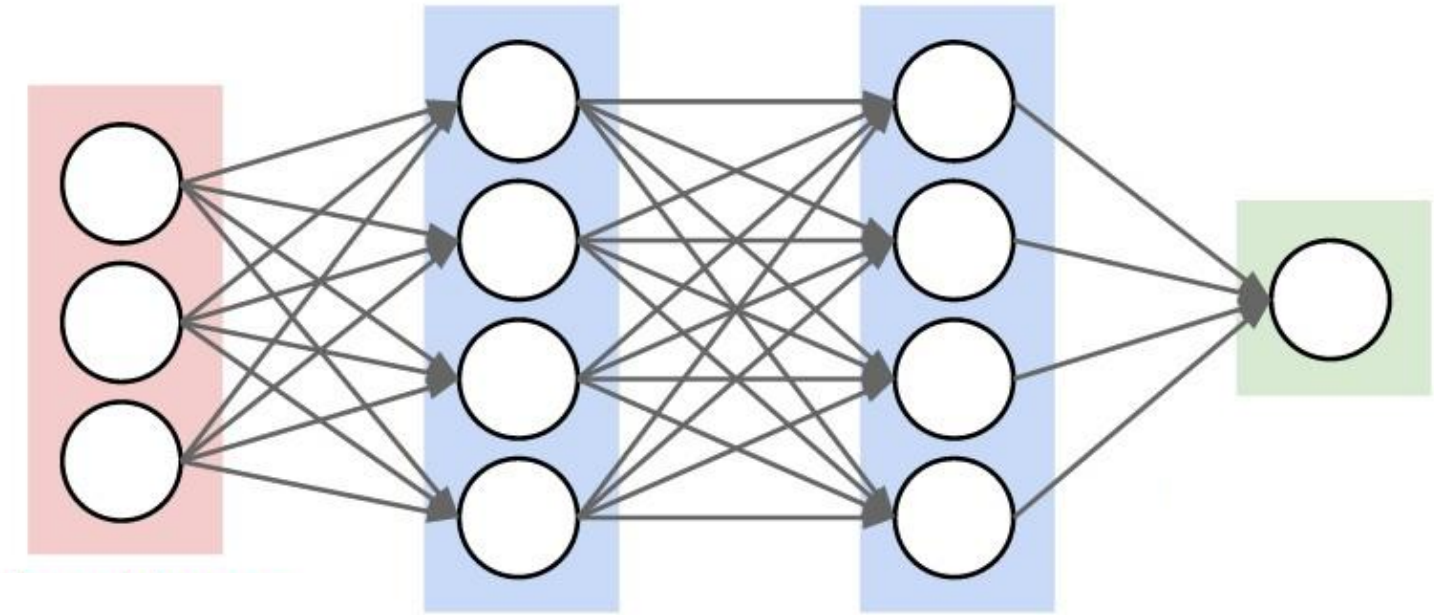
4x4 + 4
20



3x4 + 4
16

4x4 + 4
20

4x1 + 1
5

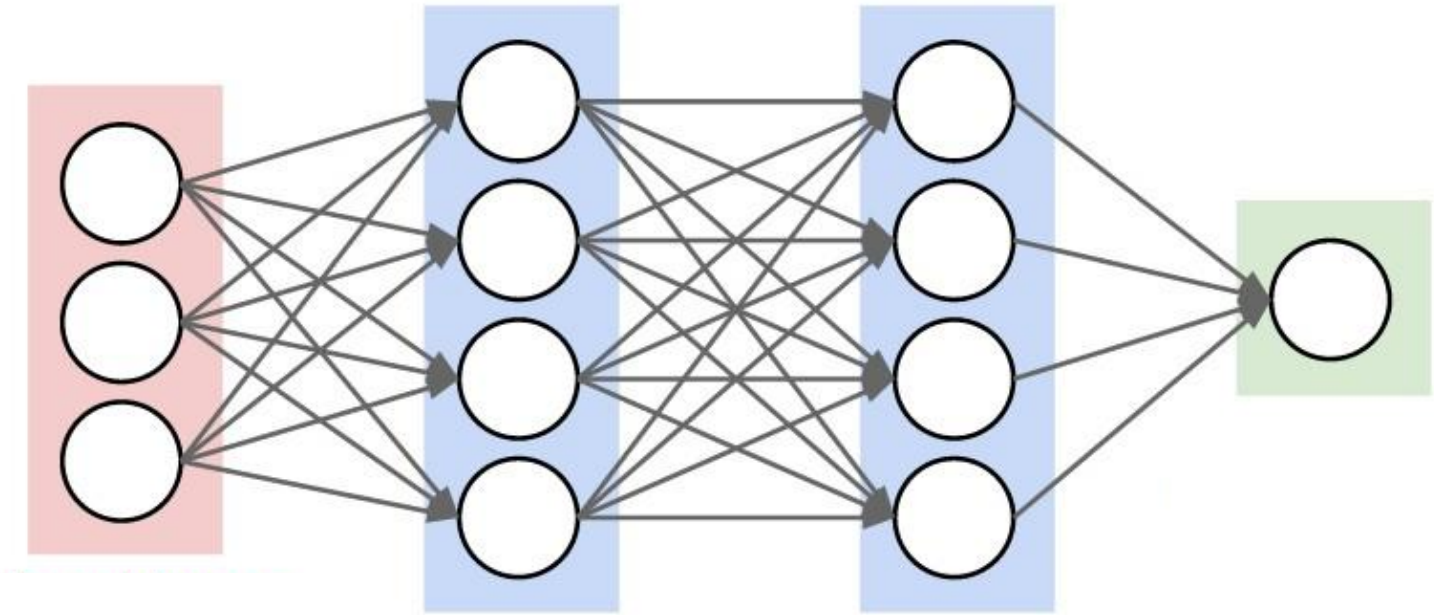


3x4 + 4
16

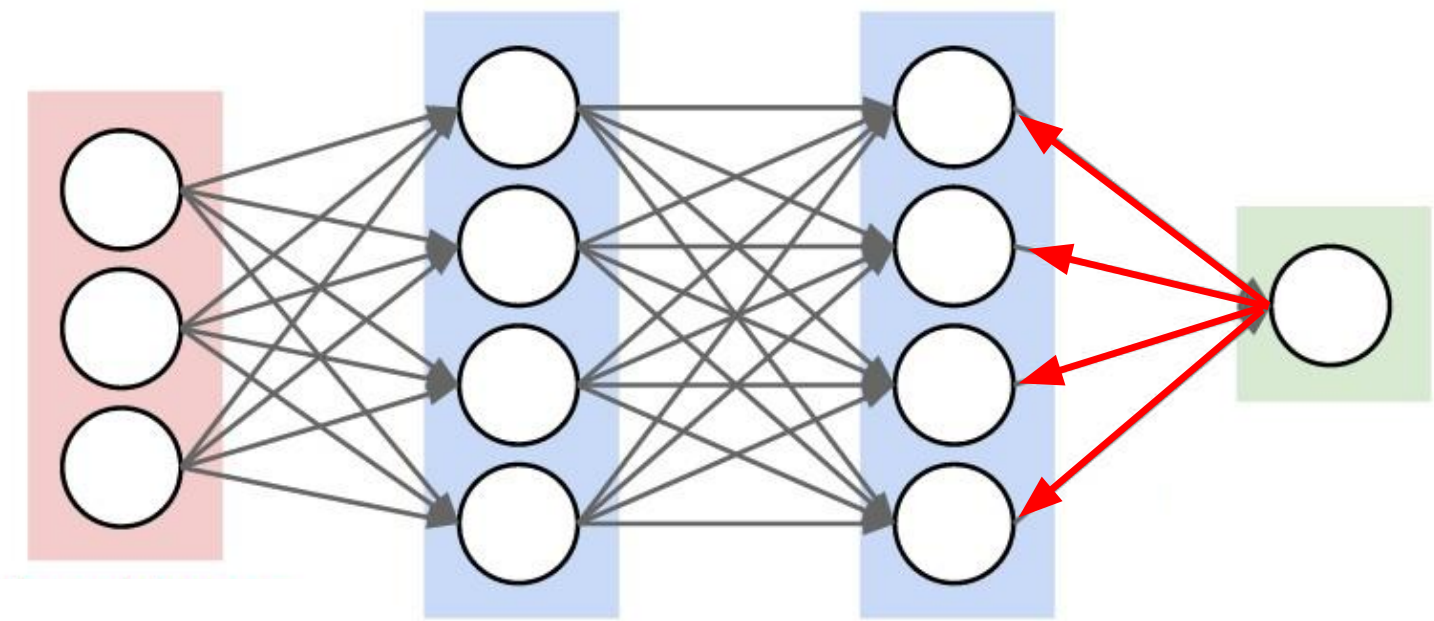
4x4 + 4
20

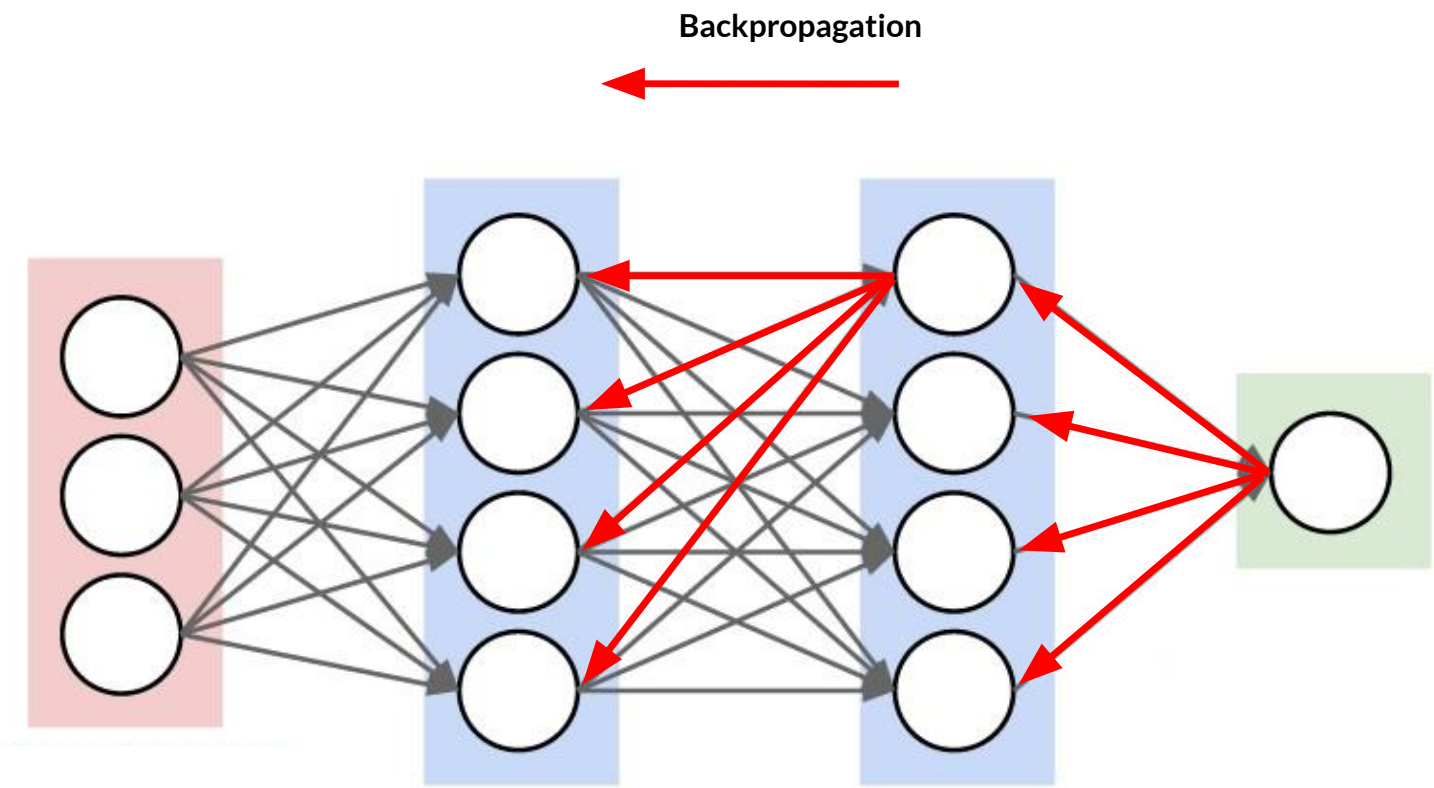
4x1 + 1
5

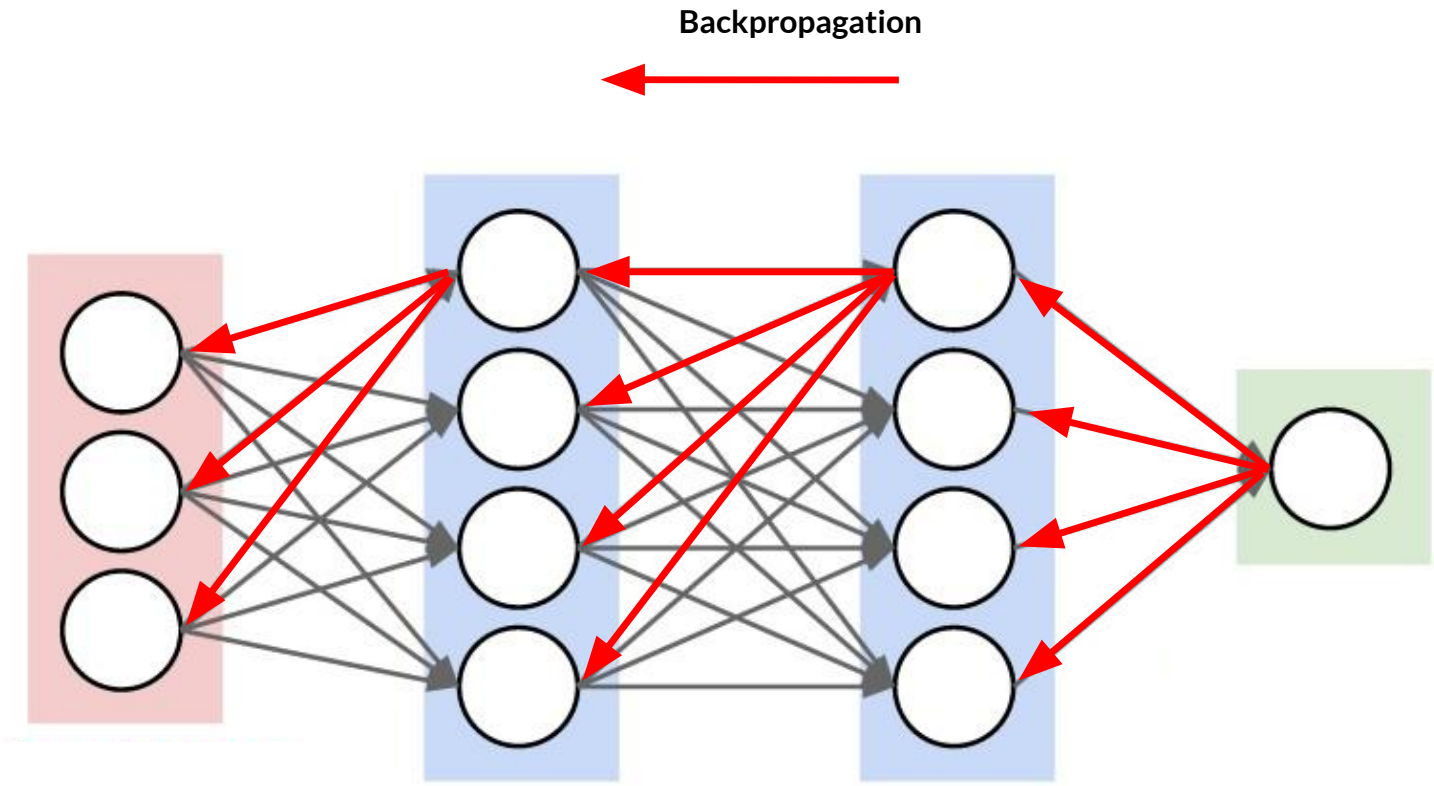
Total parameters:
41



Backpropagation




Backpropagation


Backpropagation

What are tokens?

What are tokens?

GPT-3 Codex

Tokenization is the process of splitting words into smaller chunks or tokens. These tokens are then converted into token ids. These are numbers that are inputted into a language model.

Clear Show example

Tokens	Characters
37	185

Tokenization is the process of splitting words into smaller chunks or tokens. These tokens are then converted into token ids. These are numbers that are inputted into a language model.

TEXT TOKEN IDS

Generative
Pre-trained
Transformer

Generative – predicting a future token, given past tokens
Pre-trained
Transformer

Generative – predicting a future token, given past tokens

Pre-trained – trained on a large corpus of data

Transformer

Generative – predicting a future token, given past tokens

Pre-trained – trained on a large corpus of data

Transformer – portion of transformer architecture

Objectives of GPT-3

- Predict the next token, given preceding tokens
- Causal language models
- Autoregressive language models

Roses _

Roses are _

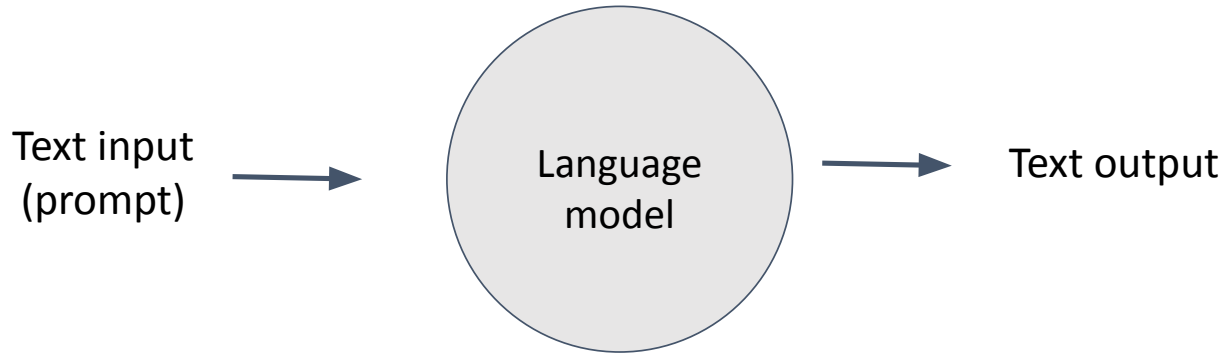
Roses are red _

Roses are red violets _

Roses are red violets are _

Roses are red violets are blue _

GPT-4


Text input
(prompt)
Language
model
Text output


Text input
(prompt)
OR
Text input
and image
GPT-4
Text output

College level exams

- GPT-4 matching human level performance at some exams
- Model performance progressing quickly

College level exams

GPT-4 technical report: https://arxiv.org/pdf/2303.08774.pdf

MMLU – Massive Multitask Language Understanding

- Multiple-choice questions in 57 subjects
- Includes STEM, humanities and the social sciences.

GPT-4 technical report: https://arxiv.org/pdf/2303.08774.pdf

MMLU: Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) 9.8 m/s²
(B) more than 9.8 m/s²
(C) less than 9.8 m/s²
(D) Cannot say unless the speed of throw is given.

MMLU: Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is

(A) 9.8 m/s² ✔
(B) more than 9.8 m/s² ✘
(C) less than 9.8 m/s² ✘
(D) Cannot say unless the speed of throw is given. ✘

MMLU: Microeconomics

One of the reasons that the government discourages and regulates monopolies is that

(A) producer surplus is lost and consumer surplus is gained.
(B) monopoly prices ensure productive efficiency but cost society allocative efficiency.
(C) monopoly firms do not engage in significant research and development.
(D) consumer surplus is lost with higher prices and lower levels of output.

MMLU: Microeconomics

One of the reasons that the government discourages and regulates monopolies is that
(A) producer surplus is lost and consumer surplus is gained. ✗
(B) monopoly prices ensure productive efficiency but cost society allocative efficiency. ✗
(C) monopoly firms do not engage in significant research and development. ✗
(D) consumer surplus is lost with higher prices and lower levels of output. ✓

MMLU: Medicine

A 33-year-old man undergoes a radical thyroidectomy for thyroid cancer. During the operation, moderate hemorrhaging requires ligation of several vessels in the left side of the neck. Postoperatively, serum studies show a calcium concentration of 7.5 mg/dL, albumin concentration of 4 g/dL, and parathyroid hormone concentration of 200 pg/mL. Damage to which of the following vessels caused the findings in this patient?
(A) Branch of the costocervical trunk
(B) Branch of the external carotid artery
(C) Branch of the thyrocervical trunk
(D) Tributary of the internal jugular vein

MMLU: Medicine

A 33-year-old man undergoes a radical thyroidectomy for thyroid cancer. During the operation, moderate hemorrhaging requires ligation of several vessels in the left side of the neck. Postoperatively, serum studies show a calcium concentration of 7.5 mg/dL, albumin concentration of 4 g/dL, and parathyroid hormone concentration of 200 pg/mL. Damage to which of the following vessels caused the findings in this patient?

(A) Branch of the costocervical trunk ✗
(B) Branch of the external carotid artery ✗
(C) Branch of the thyrocervical trunk ✓
(D) Tributary of the internal jugular vein ✗

Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar

Source: GPT-4 Technical Report

Why GPT-4?

Objectives of GPT-3

- Predict the next word

Challenges with GPT-3

- Doesn't follow user instructions
- Can generate toxic language
- Can make up facts

Objectives of GPT-3.5 / GPT-4

- Helpful and able to follow instructions
- Not toxic – for example, hateful speech, foul language
- Less likely to fabricate information or hallucinate

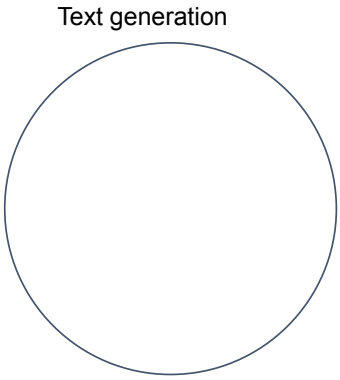
OpenAI playground

Create a shopping list

Comparing GPT-4 to GPT-3 and GPT-3.5

GPT-3?

- davinci

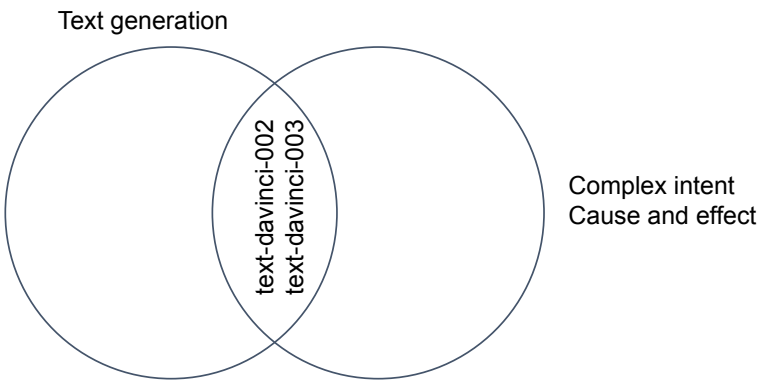


GPT-3.5

- text-davinci-002
- text-davinci-003
- code-davinci-002
- gpt-3.5-turbo

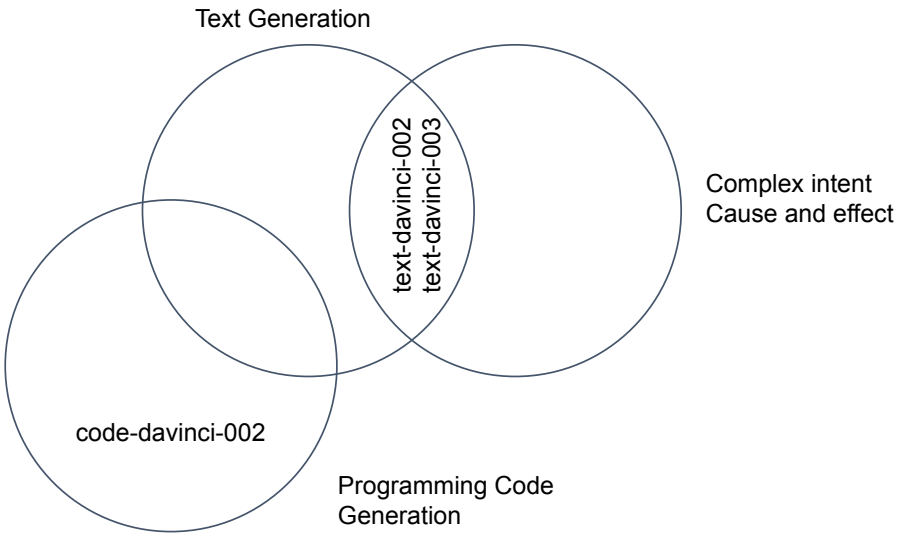
What Is GPT-3.5?

- text-davinci-002
- text-davinci-003
- code-davinci-002
- gpt-3.5-turbo



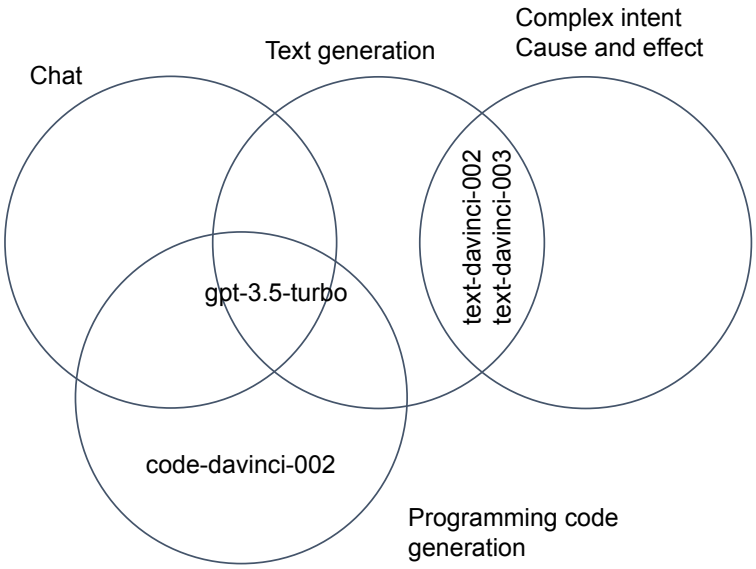
What Is GPT-3.5?

- text-davinci-002
- text-davinci-003
- code-davinci-002
- gpt-3.5-turbo



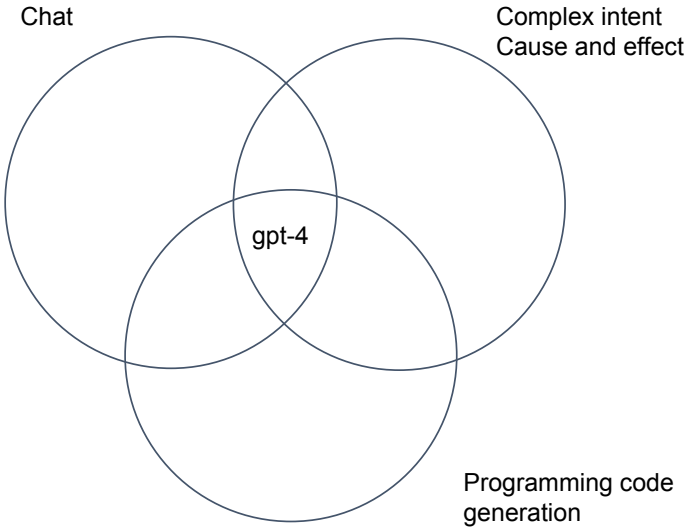
What Is GPT-3.5?

- text-davinci-002
- text-davinci-003
- code-davinci-002
- gpt-3.5-turbo

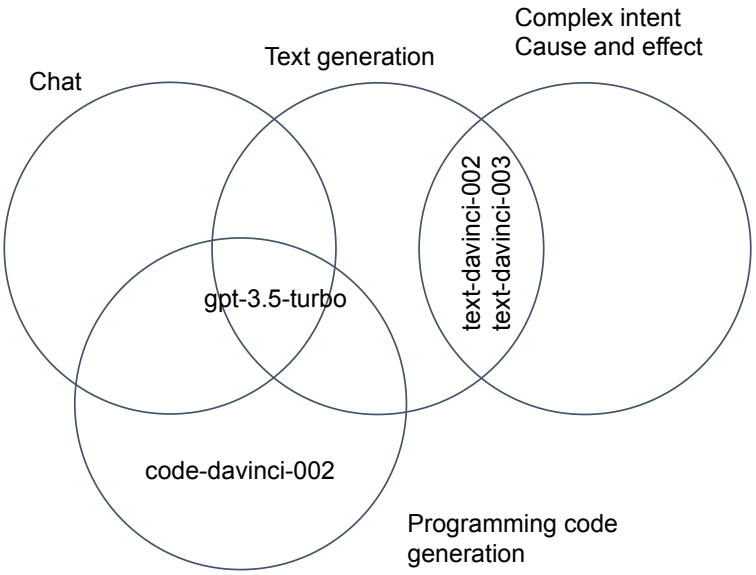


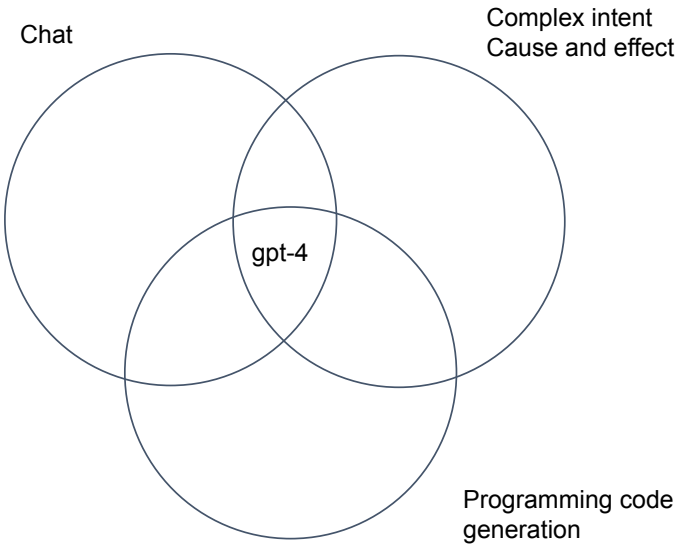
GPT-4

- gpt-4
- Gpt-4-32k
- gpt-4-turbo



What Is GPT-4?



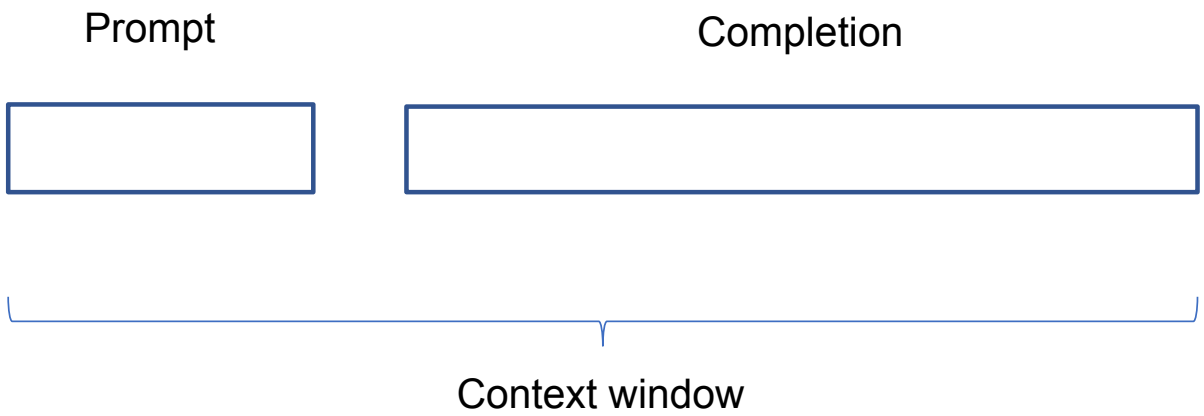


Differences between the GPT-4 models

Prompt

Differences between the GPT-4 models

Differences between the GPT-4 models



Differences between the GPT-4 models

gpt-4: 8,000 tokens
gpt-4-32k: 32,000 tokens

Differences between the GPT-4 models

gpt-4: 8,000 tokens
gpt-4-32k: 32,000 tokens

gpt-3: 2,000 tokens
gpt-3.5: 4,000 tokens

claude-3?

Thumbs up - Just right

Thumbs down - Too technical / Not technical enough

Pricing

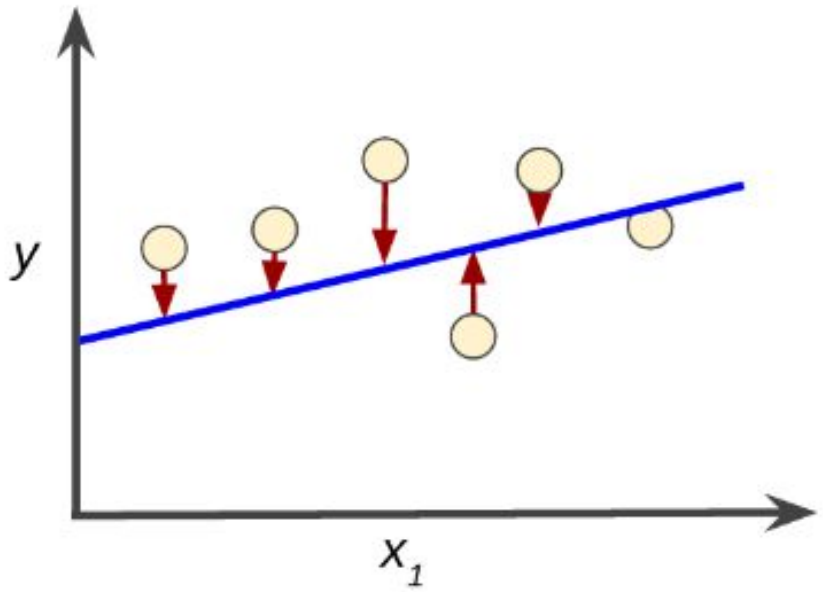
Model	PROMPT Price per 1,000 Tokens (About 750 words)	COMPLETION Price per 1,000 Tokens (About 750 words)
text-davinci-002 text-davinci-003 code-davinci-002	$0.02	$0.02
gpt-3.5-turbo	$0.002	$0.002
gpt-4 (8K context)	$0.03	$0.06
gpt-4 (32K context)	$0.06	$0.12

Model Architecture Differences

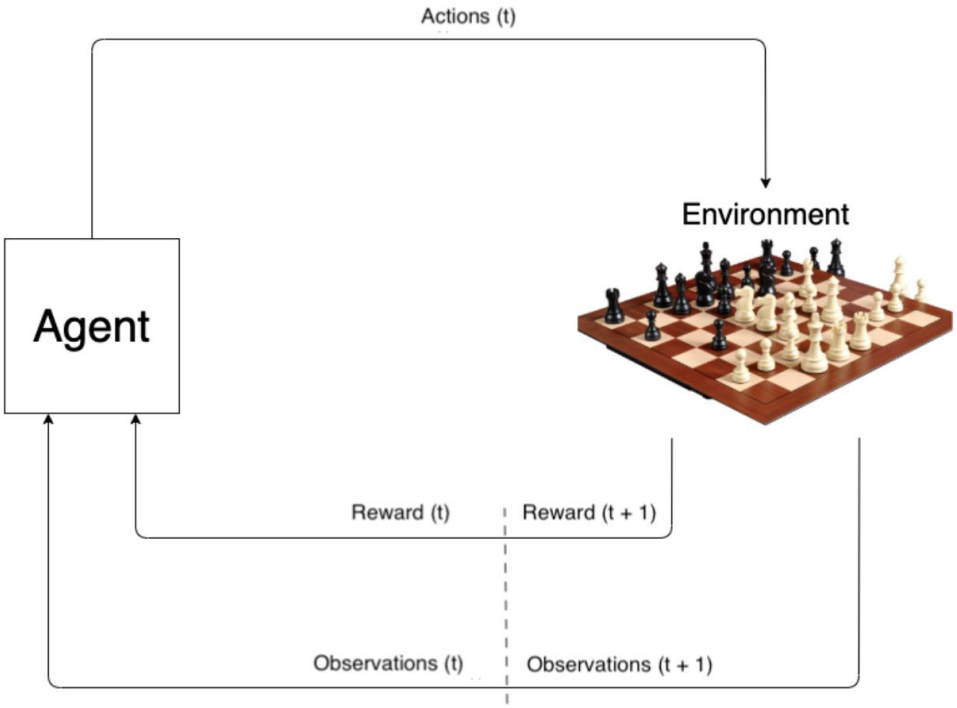
Model	COMPLETIONS: Price per 1,000 Tokens (About 750 words)	**Model Size (Number of Parameters)**
text-davinci-002 text-davinci-003 code-davinci-002	$0.02	
gpt-3.5-turbo	$0.002	
gpt-4	$0.06/$0.12	Unknown

How Was GPT-4 Trained?

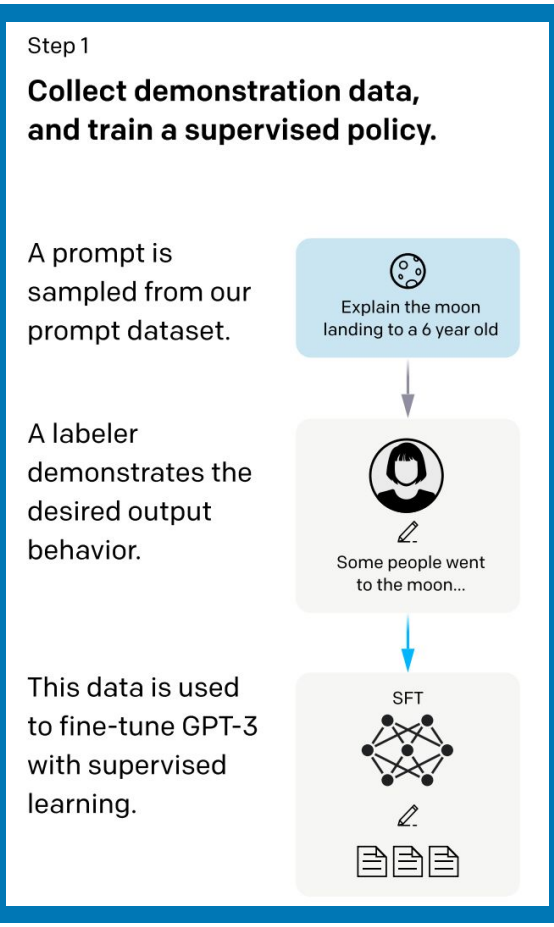
How ML models are trained



Reinforcement Learning



Reinforcement
Learning from
Human
Feedback



Training language models to follow instructions with human feedback - Long et al (Open AI)

Example: Summarize a News Article

Submit Skip « Page 3 / 11 » Total time: 05:39

Instruction

Include output

Summarize the following news article:

====
{article}
====

Output A

summary1

Rating (1 = worst, 7 = best)

1 2 3 4 5 6 7

Fails to follow the correct instruction / task **?**	Yes	No
Inappropriate for customer assistant **?**	Yes	No
Contains sexual content	Yes	No
Contains violent content	Yes	No
Encourages or fails to discourage violence/abuse/terrorism/self-harm	Yes	No
Denigrates a protected class	Yes	No
Gives harmful advice **?**	Yes	No
Expresses moral judgment	Yes	No

Notes

(Optional) notes

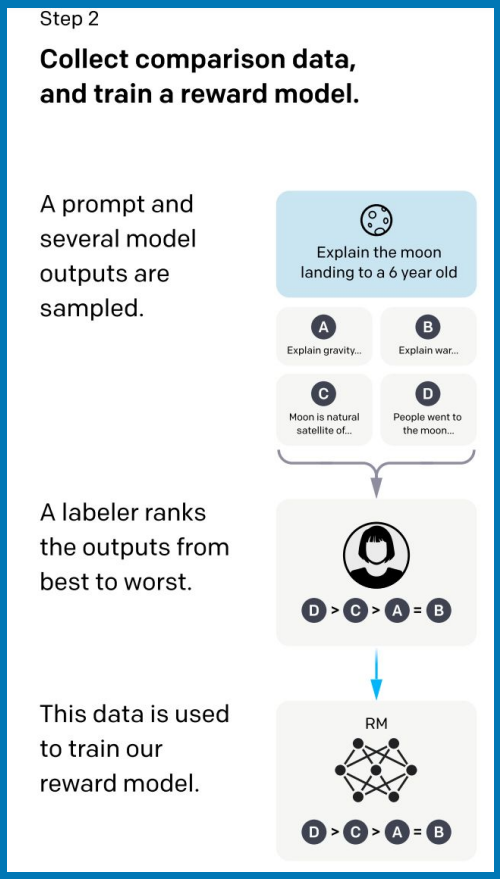
Ranking outputs

To be ranked

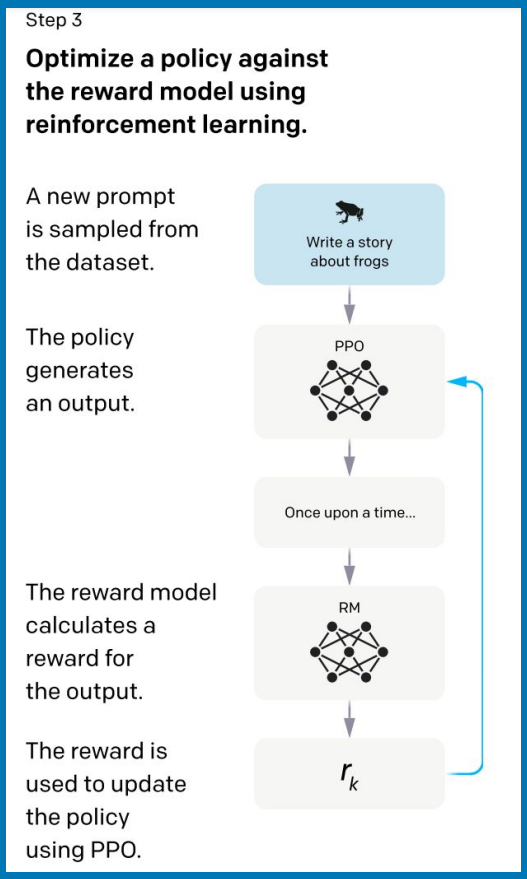
B A team of researchers from Yale University and University of California, Davis studied the vocalization patterns of several different types of parrots. They found that parrots like to mimic human speech, and can produce a wide range of sounds, such as whistles, squawks, and other types of vocalizations...

C Parrots have been found to have the ability to understand numbers. Researchers have found that parrots can understand numbers up to six. In a series of experiments, the parrots were able to identify the amount of food items under a number of cups...

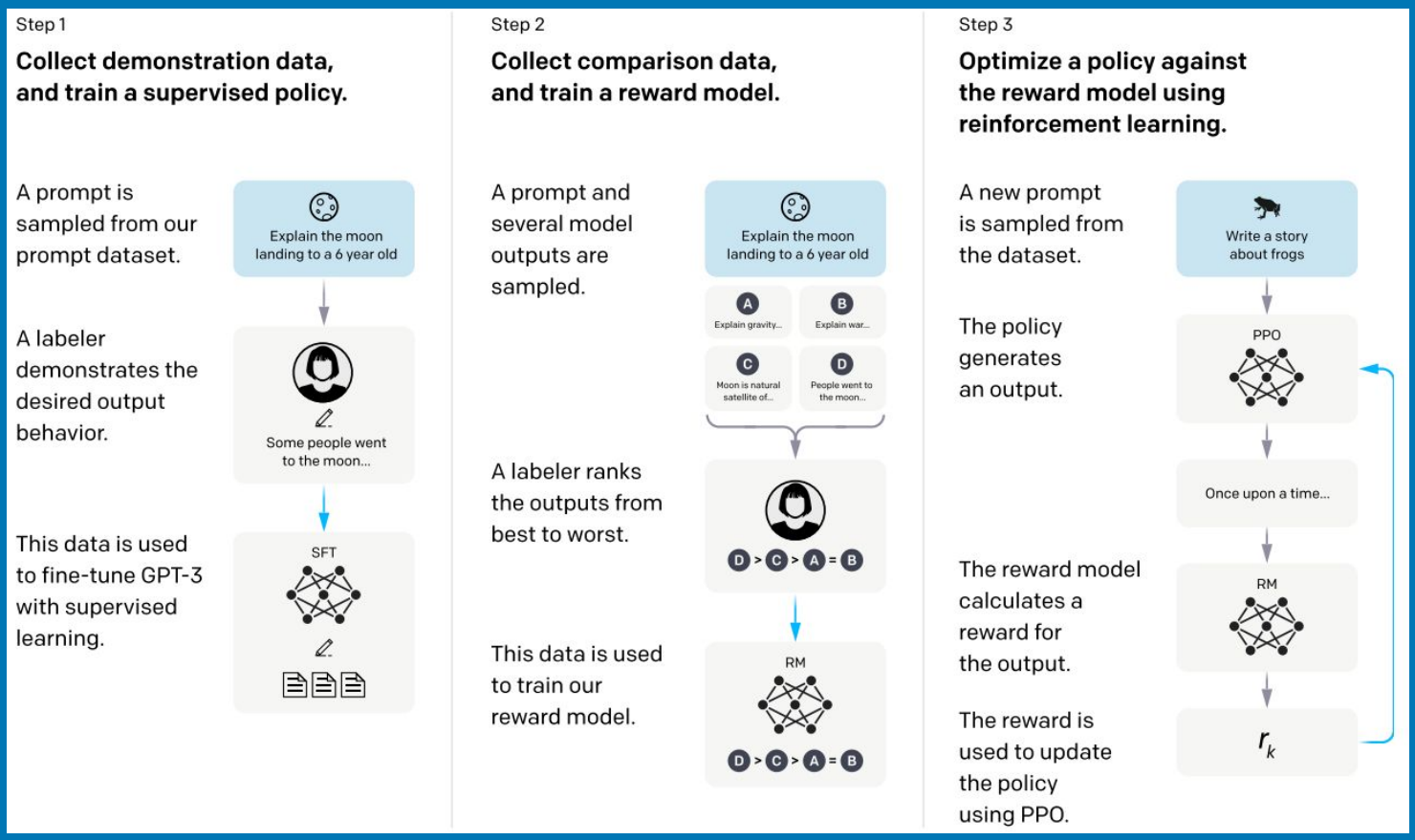
Rank 1 *(best)*	Rank 2	Rank 3	Rank 4	Rank 5 *(worst)*
A A research group in the United States has found that parrots can imitate human speech with ease, and some of them can even do so in the same way as humans. This group studied the sounds that parrots make in their natural habitats and found that they use their tongues and beaks in ways that are strikingly...		E Scientists have found that green-winged parrots can tell the difference between two noises that are the same except for the order in which they are heard. This is important because green-winged parrots are known to imitate sounds. This research shows that they are able to understand the difference between sounds.		
		D Current research suggests that parrots see and hear things in a different way than humans do. While humans see a rainbow of colors, parrots only see shades of red and green. Parrots can also see ultraviolet light, which is invisible to humans. Many birds have this ability to see ultraviolet light, an ability		



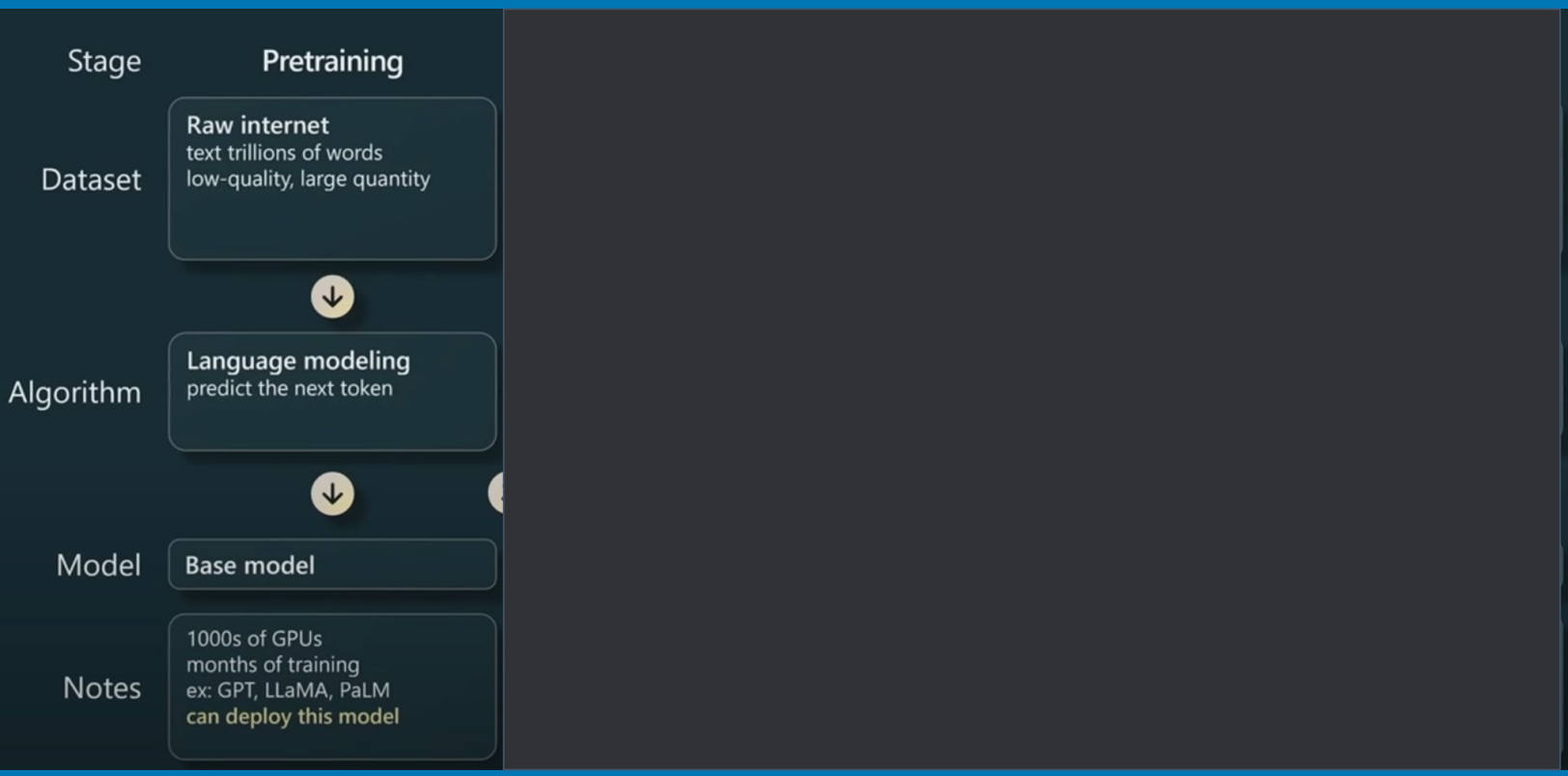
Training language models to follow instructions with human feedback - Long et al (Open AI)



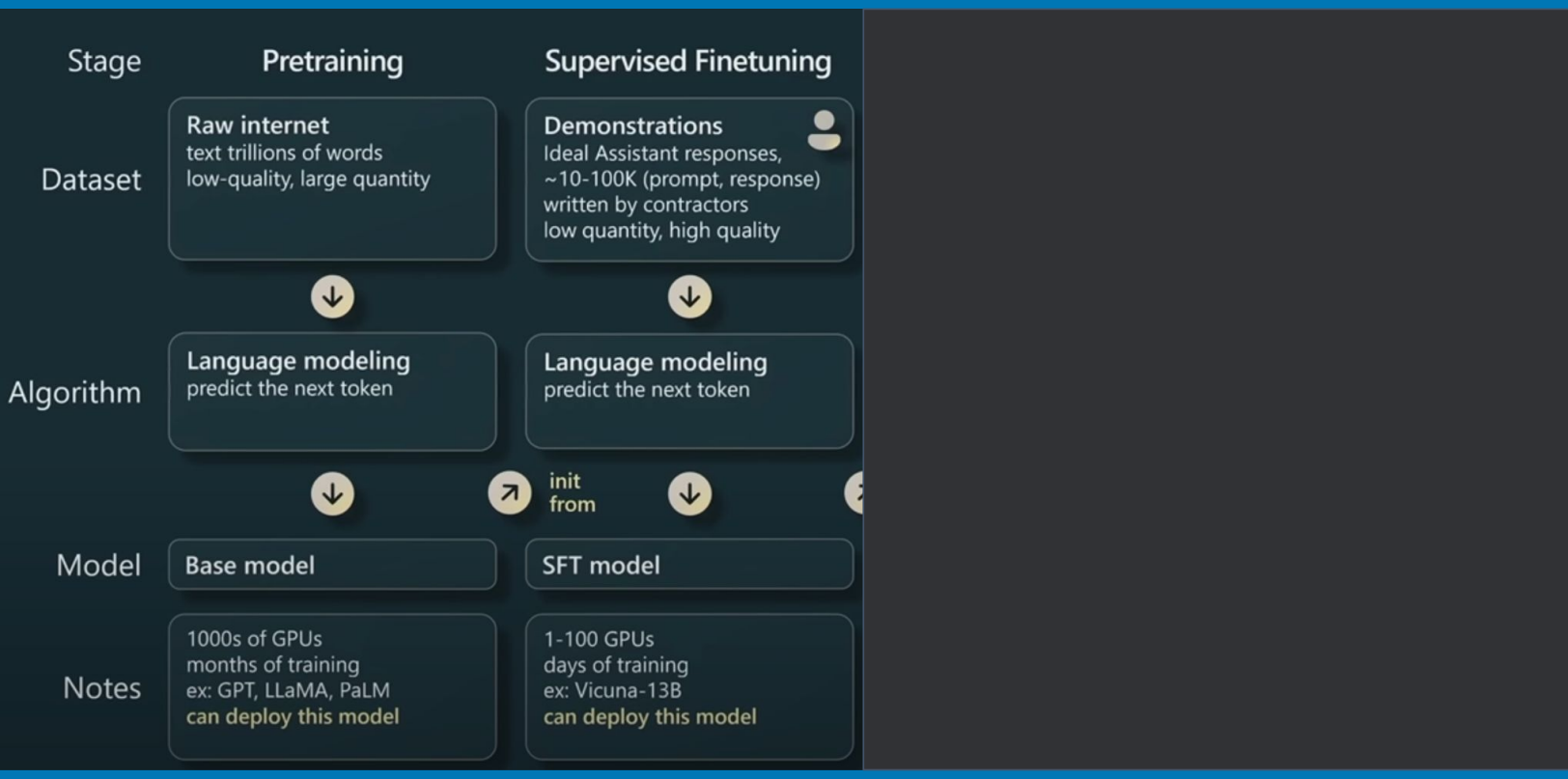
Training language models to follow instructions with human feedback - Long et al (Open AI)



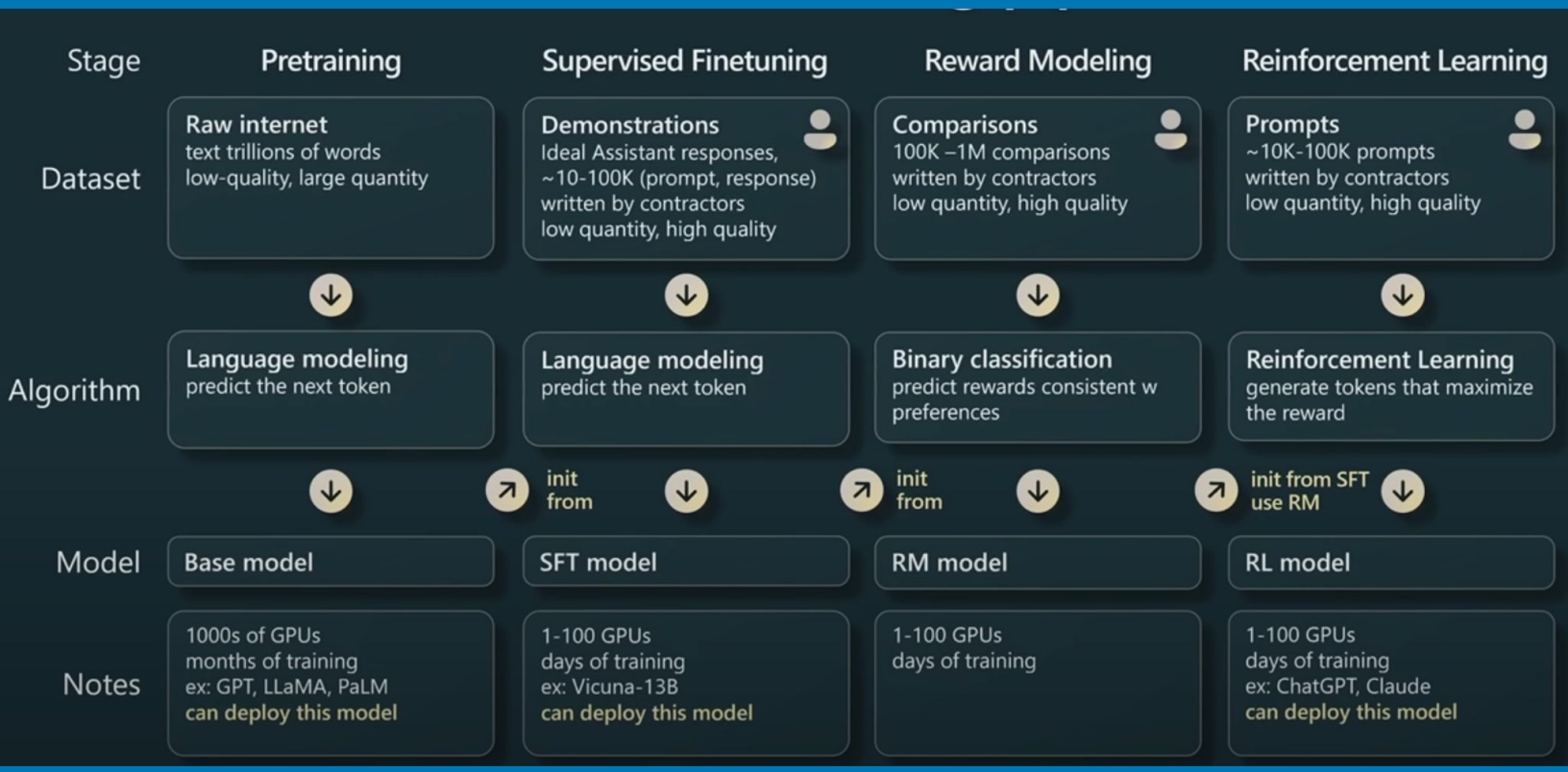
and 1.5 years later …



State of GPT - Karpathy



State of GPT - Karpathy



State of GPT - Karpathy

Limitations of GPT-4

Model Weaknesses

- Only recently been able to fine-tune (Not possible on release)
- Doesn’t update its knowledge in real-time.
- Makes up facts.

CNN BUSINESS Markets Tech Media Calculators Videos

Lawyer apologizes for fake court citations from ChatGPT

By Ramishah Maruf, CNN
Updated 3:28 PM EDT, Sun May 28, 2023

Why fine-tuning?

The Journal of Infectious Diseases
EDITORIAL COMMENTARY
IDSA Infectious Diseases Society of America
hivma hiv medicine association
OXFORD

Pneumococcal Carriage and Seroepidemiology Studies to Measure Current and Future Pneumococcal Conjugate Vaccine Effectiveness

Keith P. Klugman and Gail L. Rodgers
Pneumonia, Surveillance and Epidemic Control Programs, Bill & Melinda Gates Foundation, Seattle, Washington, USA

Two current questions in pneumococcal conjugate vaccine (PCV) development for children are whether immunization schedules are optimal to maintain direct and indirect protection of licensed vaccines, and what are the potential roles of next-generation higher-valent vaccines to further reduce pediatric pneumococcal disease. invasive pneumococcal disease (IPD) due to the vaccine serotypes, but this measure has been frustrated by the coronavirus disease 2019 (COVID-19) pandemic, which has reduced the incidence of IPD in the United Kingdom and many other countries, largely as a result of masking and restrictions on indoor children is that potentially key invasive serotypes, such as serotypes 8 and 12F, which have dominated adult disease after PCV13 introduction in children in the United Kingdom, were only rarely detected in carriage in this study. So while carriage among healthy children is useful to describe the potential distribution of

GPT-4 - hallucination

These are the topics that I am covering in an O'reilly online training on GPT-4. Write a catchy introduction.
What is GPT-4
Why GPT-4?
Comparing GPT-4 to GPT-3 and GPT-3.5
How was GPT-4 trained?
What are the limitations of GPT-4?
HELM

Colab notebook: GPT exercises

https://colab.research.google.com/drive/1-rSgn9pM5zyRXrxchzcxGZ_Wm_FTuo1K?usp=sharing

GPT-4 Turbo

GPT-4 Turbo with 128K context

- Knowledge of world events up to April 2023.
- 128k context window (300 pages of text)
- gpt-4-1106-preview
- Multiple function calls
- Respond with JSON mode
- Assistants API
  - Function calling (multiple function calling)
  - Code Interpreter
  - Retrieval
- GPT-4 Turbo with vision

Next week

Live Course

Hands-on GPT-4-Turbo

With Jonathan Fernandes

3h 0m April 25 • 5pm-8pm GMT+1

Pricing

Model	PROMPT Price per 1,000 Tokens (About 750 words)	COMPLETION Price per 1,000 Tokens (About 750 words)
GPT-3.5 Turbo 4K	$0.0015	$0.002
GPT-3.5 Turbo 16K	$0.001	$0.002
GPT-3.5 Turbo 4K fine-tuning	$0.012 (Training $0.008)	$0.003
GPT-3.5 Turbo 16K fine-tuning	$0.003 (Training $0.008)	$0.006

Pricing

Model	PROMPT Price per 1,000 Tokens (About 750 words)	COMPLETION Price per 1,000 Tokens (About 750 words)
GPT-4 Turbo 8K	$0.03	$0.06
GPT-4 Turbo 16K	$0.06	$0.12
GPT-4 Turbo 128K	$0.01	$0.03

Model Architecture Differences

Model	Model Size (Number of Parameters)
text-davinci-002 text-davinci-003 code-davinci-002	
gpt-3.5-turbo	
gpt-4	Unknown
gpt-4-turbo	Unknown

Transformers in production

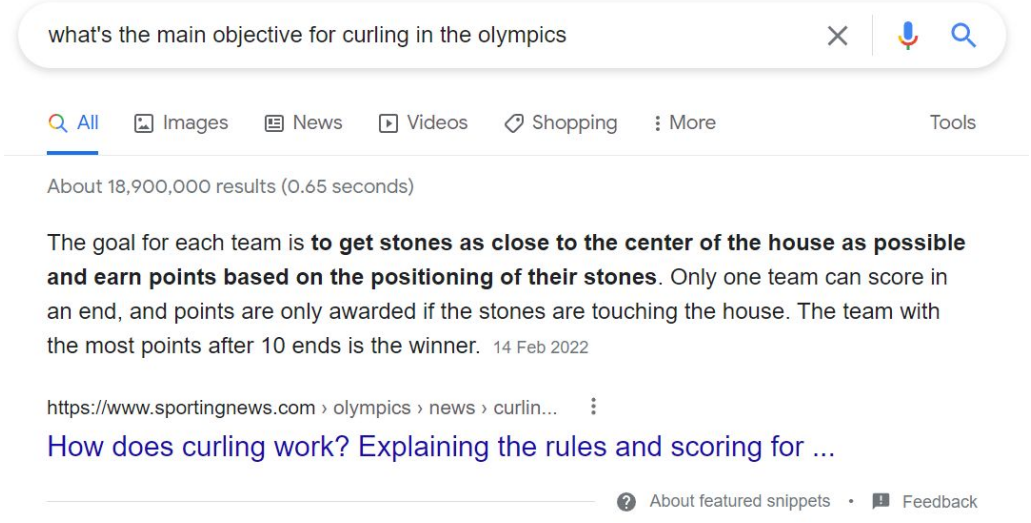
Google

curling objective

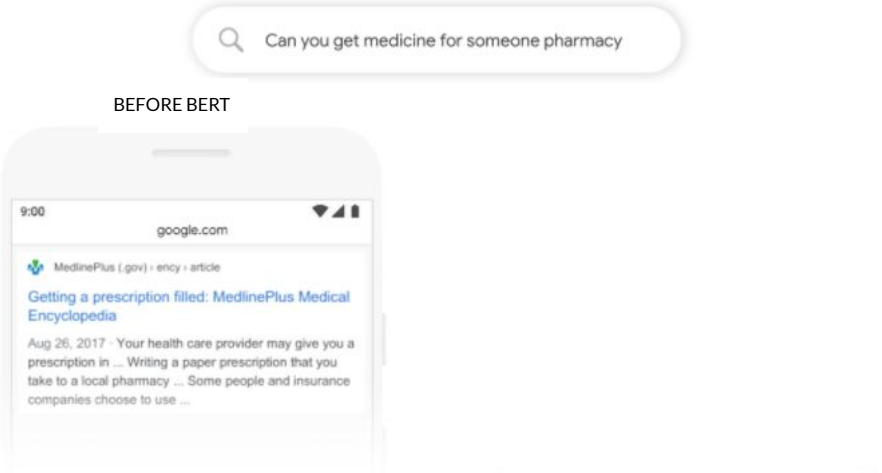
BERT

Bidirectional Encoder Representations from Transformers

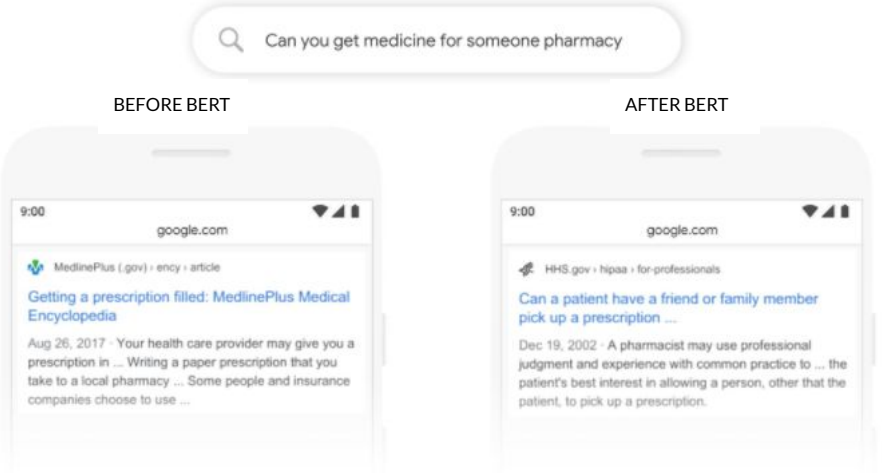
Where are Transformers used in production?



Transformers in production



Transformers in production



What was BERT trained on?

BERT - Wikipedia and BooksCorpus (11,000 unpublished books)

What tasks was BERT trained?

- Masked Language Model (MLM)
- Next Sentence Prediction (NSP)

The Tokyo Olympic games were <masked> from 2020 to 2021.

Masked Language Modelling (MLM)

The Tokyo Olympic games were <masked> from 2020 to 2021.

Masked Language Modelling (MLM)

The Tokyo Olympic games were postponed from 2020 to 2021.

Next sentence prediction (NSP)

The Tokyo Olympic games were postponed from 2020 to 2021. This is the first instance in the history of the Olympics as previous games had been cancelled but not rescheduled.

Why MLM and NSP?

BERT gets a good understanding of English language.

Transfer Learning

Transfer Learning

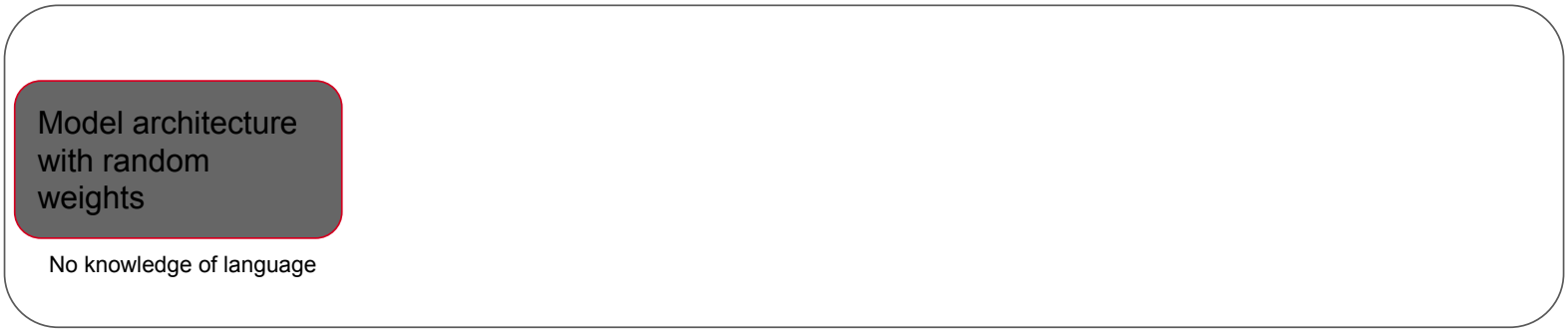
Transfer Learning is made up of 2 components.

- Pre-training
- Fine-tuning

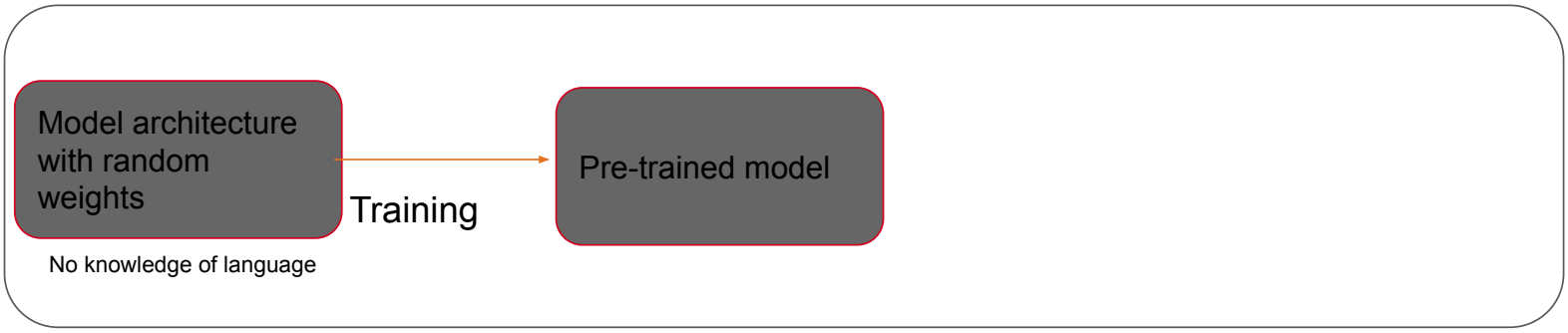
Transfer Learning

Model architecture with random weights

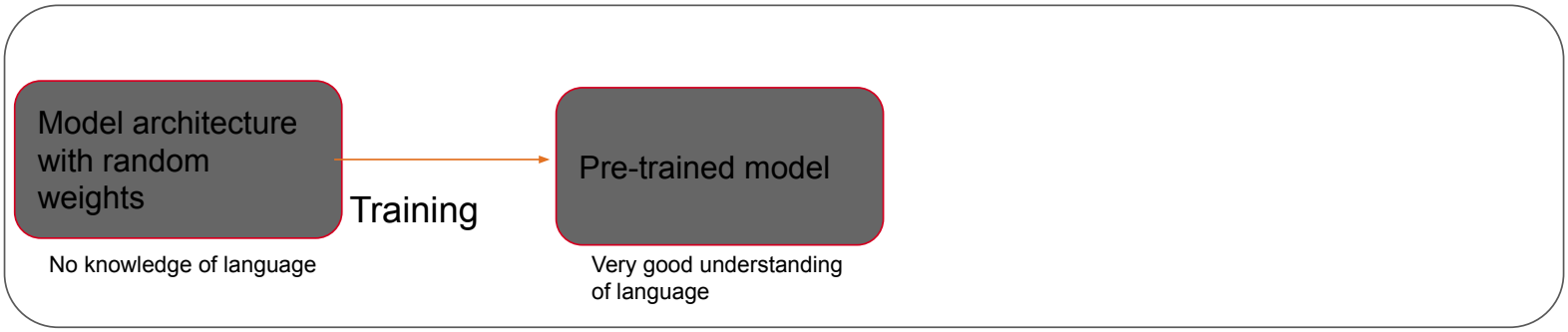
Transfer Learning



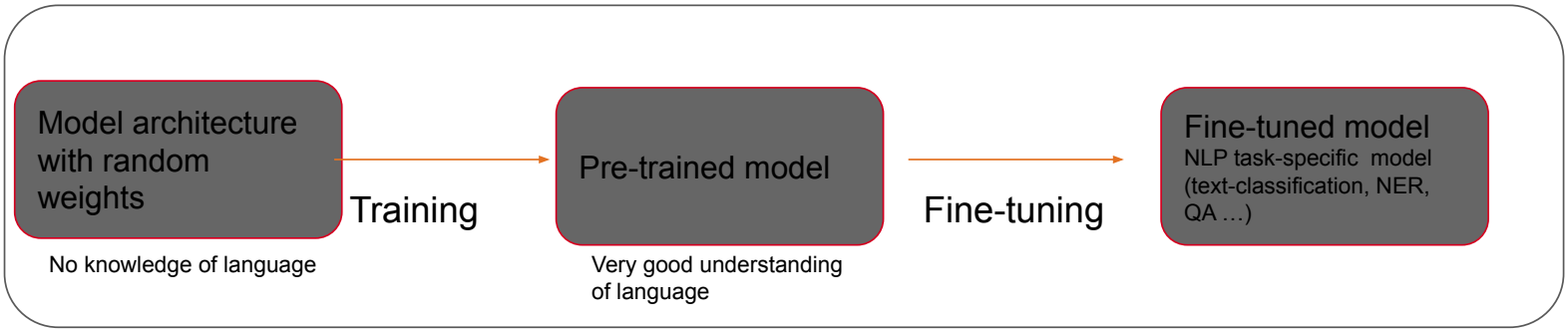
Transfer Learning



Transfer Learning



Transfer Learning



Fine-tuning

Text	Label
I love this place. It has just the right …	positive
This was a disappointing experience ..	negative
I did not enjoy my time here …	negative

Transfer Learning

Benefits of transfer learning

- Faster development
- Less data to fine-tune
- Excellent results

Pre-training: BERT

	BERT
Year	2018
Number of parameters	109M
Training time	12 days
Infrastructure	8 x V100 GPUs (*)
Size of dataset used for training	16GB
Training tokens (dataset)	250B
Dataset source	Wikipedia
	Book corpus

What are tokens?

1500 words is approximately equivalent to 2400 tokens

What are tokens?

1500 words is approximately equivalent to 2400 tokens

A word is approximately 1.4 tokens

What are tokens?

1500 words is approximately equivalent to 2400 tokens

A word is approximately 1.4 tokens

A novel is 100,000 words, or 140,000 tokens

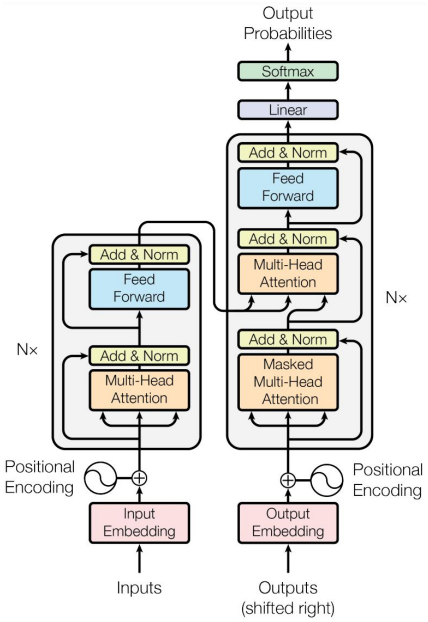
What are tokens?

BERT was trained on 250B tokens or:

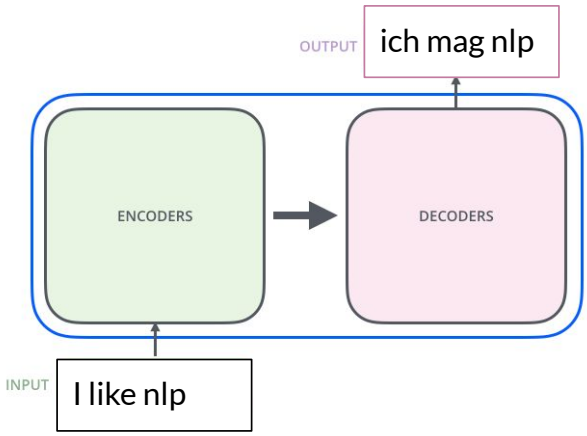
1.8 million novels

Transformer: Architecture Overview

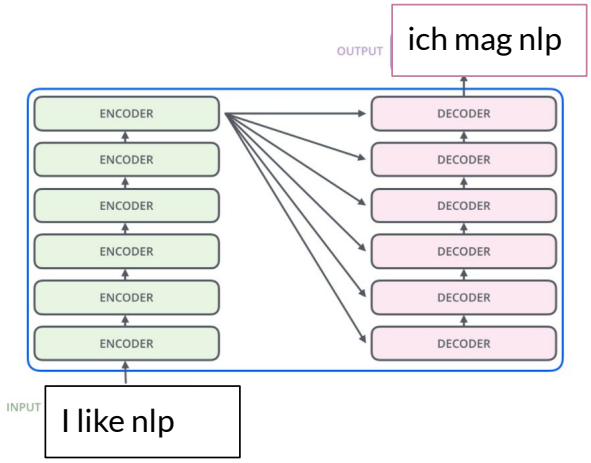
Transformer architecture



Transformer overview

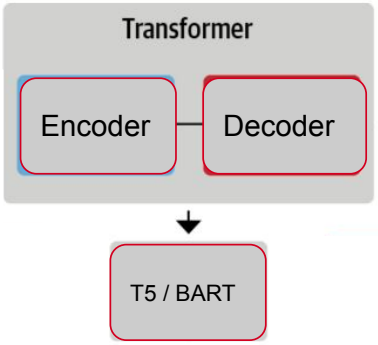


Transformer overview



Encoder-decoder model

- Generative tasks
- BART
- T5



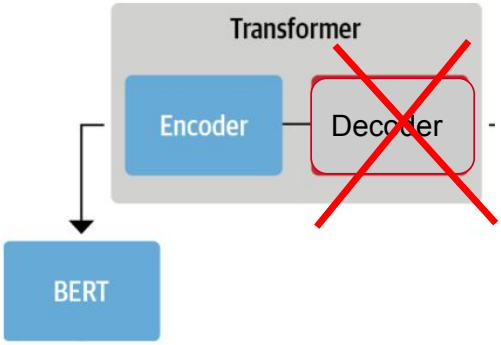
Encoder-only model

Understanding of input

- Sentence classification
- Named Entity Recognition

Family of BERT models:

- BERT, RoBERTa, DistilBERT ...

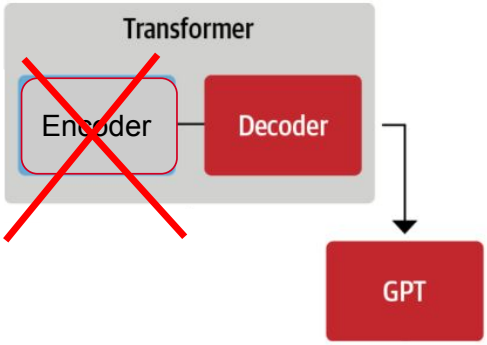


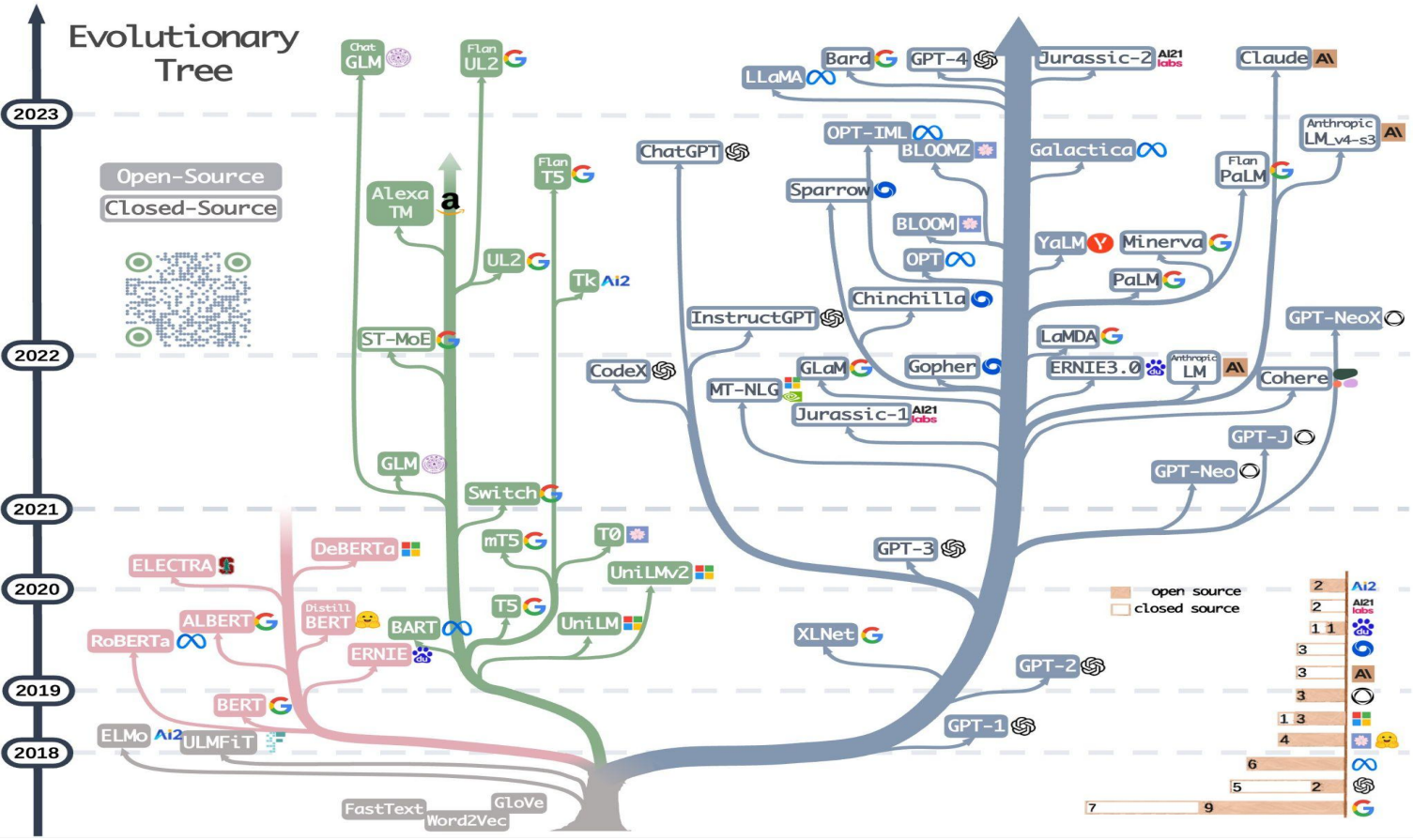
Decoder-only model

- Generative tasks

Examples:

- GPT
- GPT-2
- GPT-3

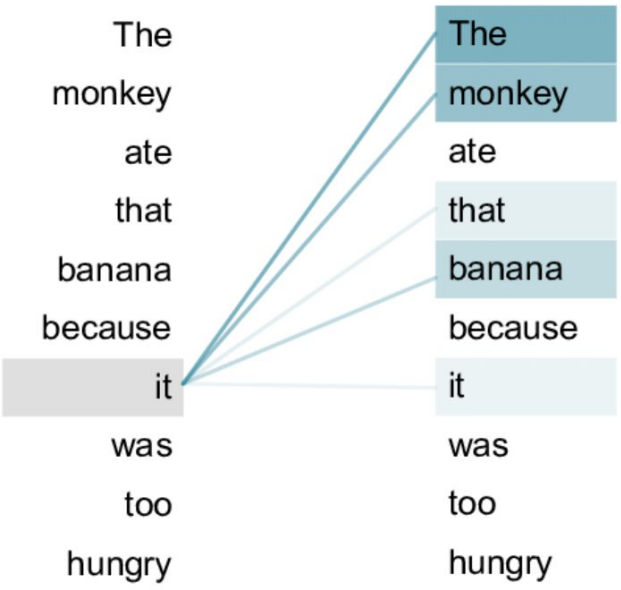


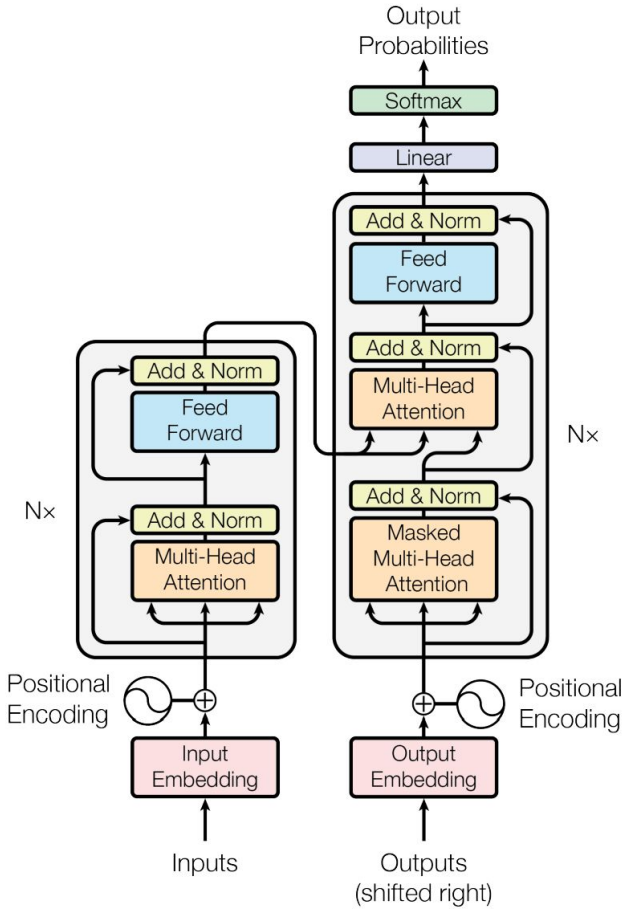


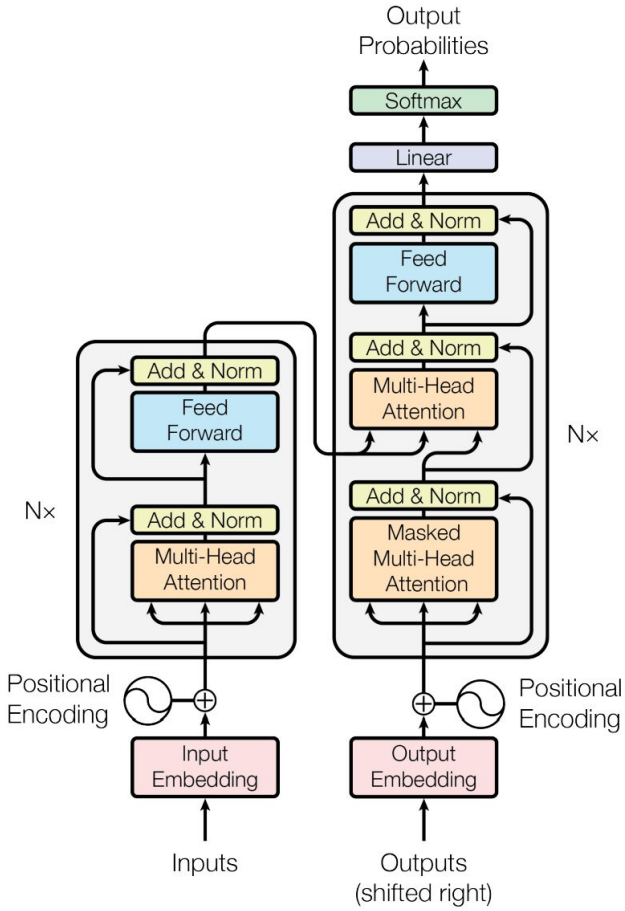
https://github.com/Mooler0410/LLMsPracticalGuide

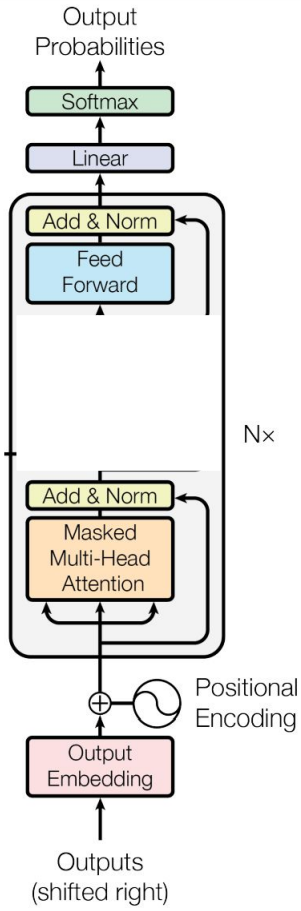
Self-attention

self-attention




Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)


Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)


Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Nx
Add & Norm
Masked
Multi-Head
Attention
Positional
Encoding
Output
Embedding
Outputs
(shifted right)

self-attention

Attention has 3 inputs:

- Q (query)
- K (key)
- V (value)

self-attention

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$$

self-attention

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{n}}$$

self-attention

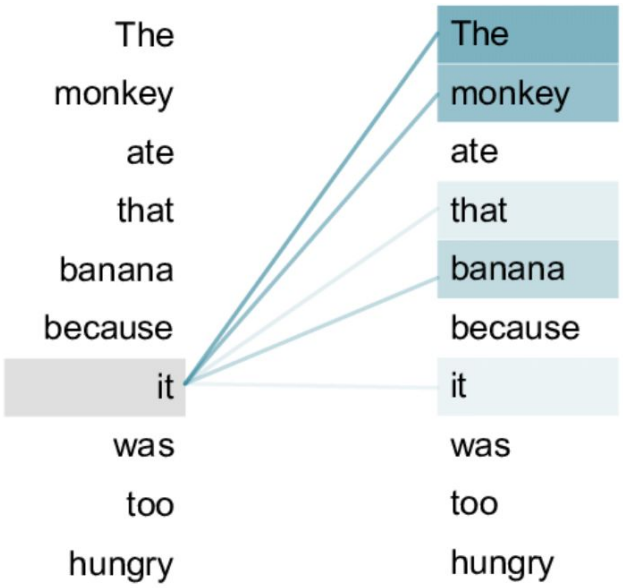
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{n}})$$

self-attention

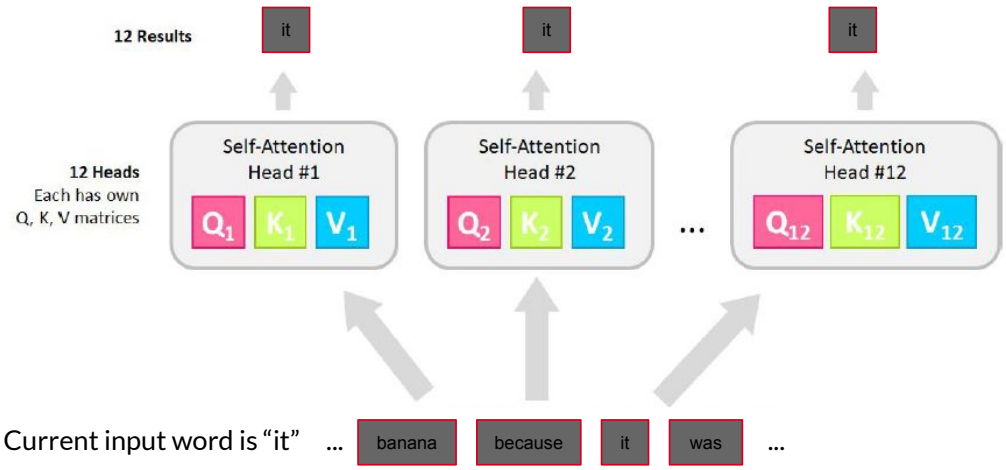
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{n}})\mathbf{V}$$

Multi-headed attention

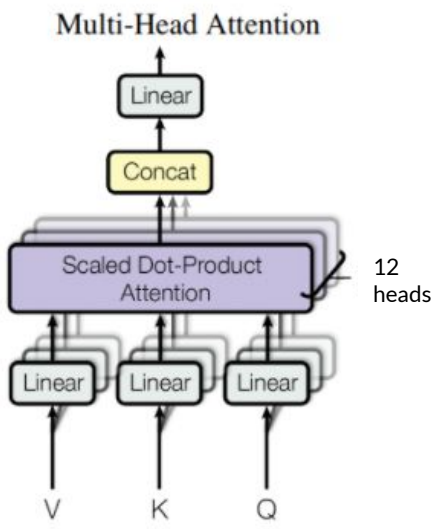
self-attention



Multi-headed attention



Multi-Head Attention



Week 1 (today)

GPT-3 and why we needed GPT-4

- GPT-3, GPT-3.5 and GPT-4
- Limitations of GPT-4
- RLHF
- Q&A
- Break

Transfer learning and fine-tuning

- BERT
- Transfer learning and fine-tuning
- Q&A
- Break

Transformer architecture overview

- Encoders and decoders
- Attention mechanism
- Q&A
- Break

O'REILLY®