



WELCOME!

The Course Structure

Concept

Project

Concept

Project

Concept

Project

The Founders

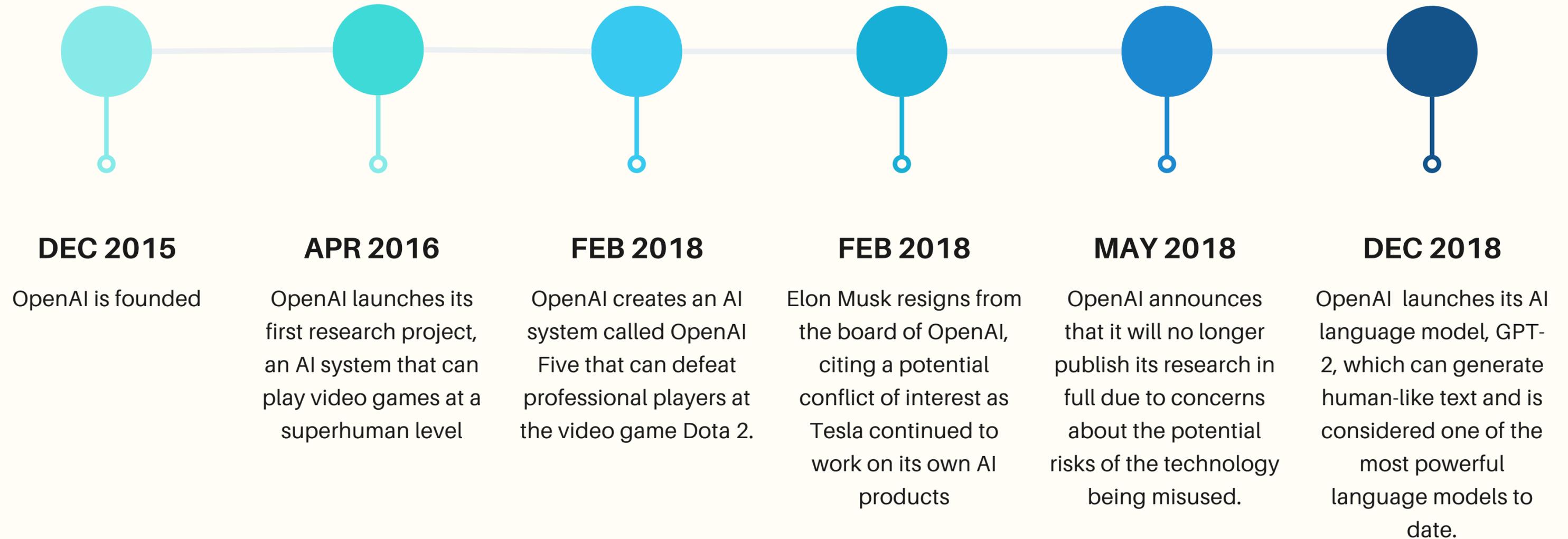
OpenAI was founded in 2015 by a group of high-profile individuals in tech, including:

- Elon Musk
- Sam Altman
- Greg Brockman
- Ilya Sutskever
- John Schulman
- Wojciech Zaremba



OPEN AI

A brief history of the early days



OPEN AI

A brief history of the recent stuff



JUNE 2020

OpenAI creates an AI language model, GPT-3. It is widely considered one of the most significant advances in AI language modeling to date.



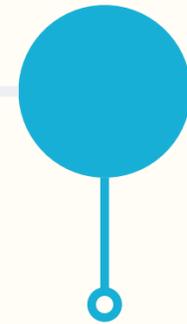
JUNE 2020

The company launches its first commercial project: an API that gives users access to GPT-3



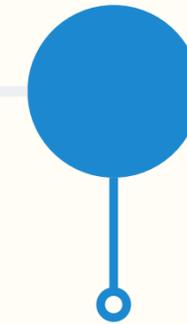
DEC 2020

OpenAI creates an AI system called DALL-E that can generate images from text.



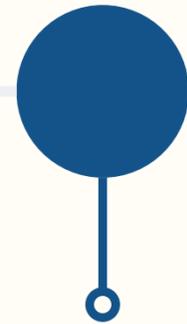
MAY 2021

OpenAI releases Codex, an AI system that can generate code based on language inputs, revolutionizing the field of software development



NOV 2022

OpenAI releases ChatGPT, a language model chatbot built on top of GPT-3. It quickly garnered attention around the world.



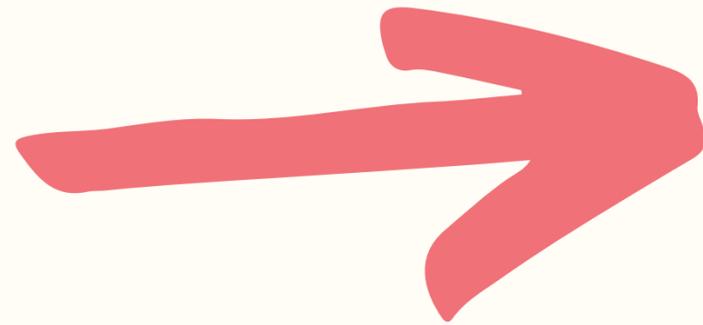
MAR 2023

OpenAI releases the ChatGPT API, giving developers access to ChatGPT in their own applications

MAR 2023

OpenAI releases GPT-4!

2015



2019

OpenAI was founded as a non-profit with the goal of creating and promoting "friendly" artificial intelligence

OpenAI became a for-profit company and announced its intentions to commercially license its technologies

GPT-3

Generative Pre-Trained Transformer

GPT-3.5

Generative Pre-Trained Transformer

Finely-Tuned version of GPT-3

GPT-4

Generative Pre-Trained Transformer

A whole new model!

What is it?

- A language model that uses deep learning to generate text
- We give it initial text (the prompt) and it produces text to continue the prompt



PROMPT

Why did the chicken cross the

OUTPUT

road?



PROMPT

My favorite food is:

OUTPUT

pizza!



PROMPT

Translate from English to French:

snails =>

OUTPUT

escargots



PROMPT

*Write a tagline for a doggy
day care*

OUTPUT

*Cozy, Safe and Fun for Your
Furry Friend!*



PROMPT

Write 2 sentences explaining acid rain to a toddler

OUTPUT

Acid rain is rain that has acids in it, which makes it a little bit yucky. Acid rain can hurt trees and other plants, and even give us coughs if the air is full of it.



PROMPT

Correct this to standard English:

She no have no money

OUTPUT

She doesn't have any money.



The 2 ways we can "talk" to GPT models

Completion

- Expects a single text prompt
- Does not support GPT-4

Chat

- Expects a list of messages in a chat-based format
- Supports GPT-4

How?

- GPT-4 is based on a type of neural network called a **transformer**
 - Transformers are a deep learning model that excel at processing sequential data (like natural language text!)



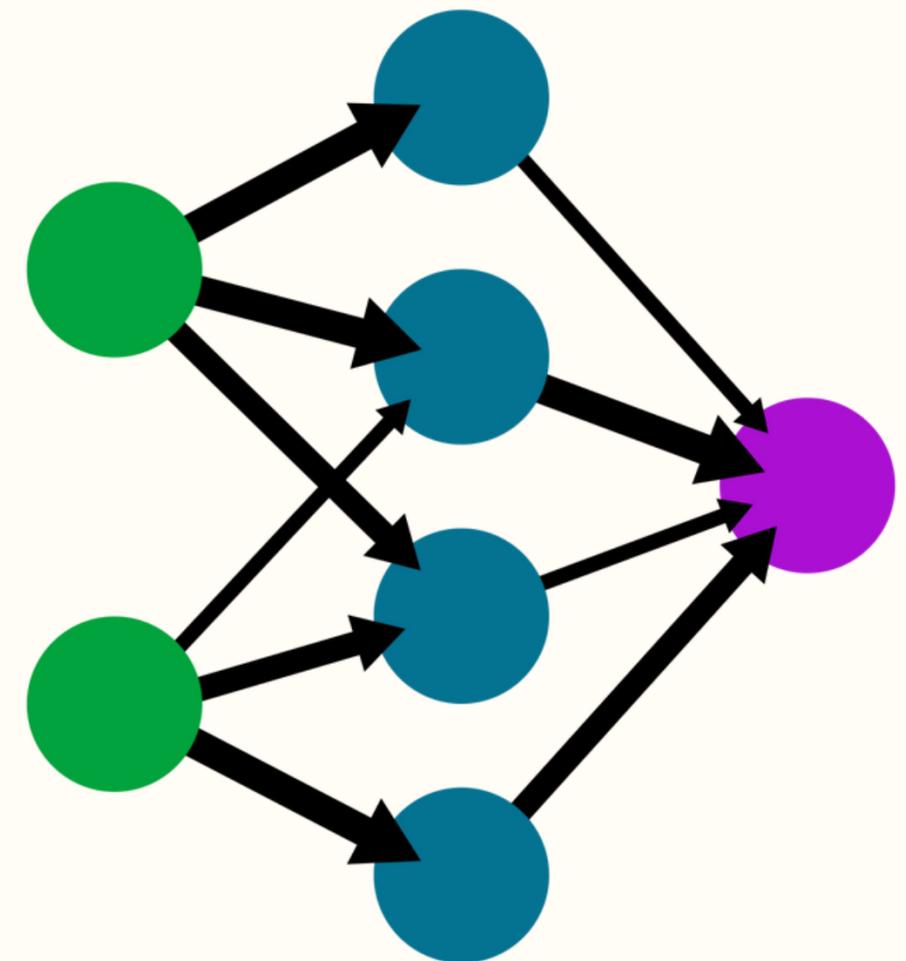
Neural Networks

Neural networks are a category of models that are very good at analyzing complicated data types

They consist of layers of connected nodes that can "fire" like neurons – passing data to other nodes

A simple neural network

input layer hidden layer output layer



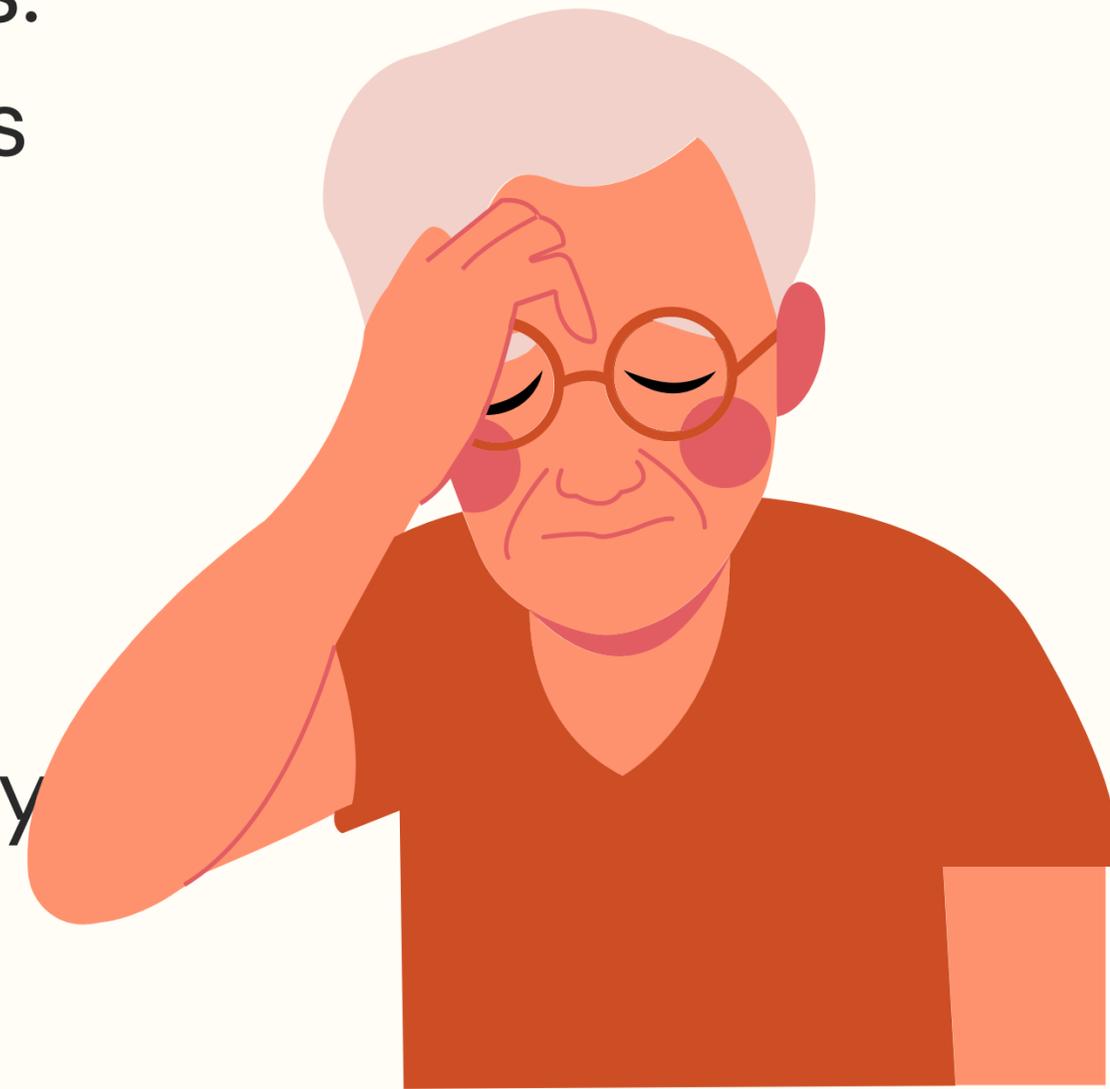
Neural Networks

- There are many different types of neural networks
 - **Convolutional Neural Networks** work great for analyzing images
 - **Recurrent Neural Networks** work well at text processing and translation
 - **Transformers** (what we care about)

Recurrent Neural Networks

RNNs work sequentially, processing text one word at a time, but they have some problems:

- They're not great at analyzing large pieces of text
- They're slow to train. This means they can't be trained on huge amounts of data easily
- Training can't be parallelized because they process sequentially



Enter...

Transformers

- Transformers are a relatively recent approach (2017)
- Transformers process the entire input at once, rather than sequentially, meaning there is less risk of "forgetting" previous context.
- This means they can be trained in parallel!



**How do
they work?**

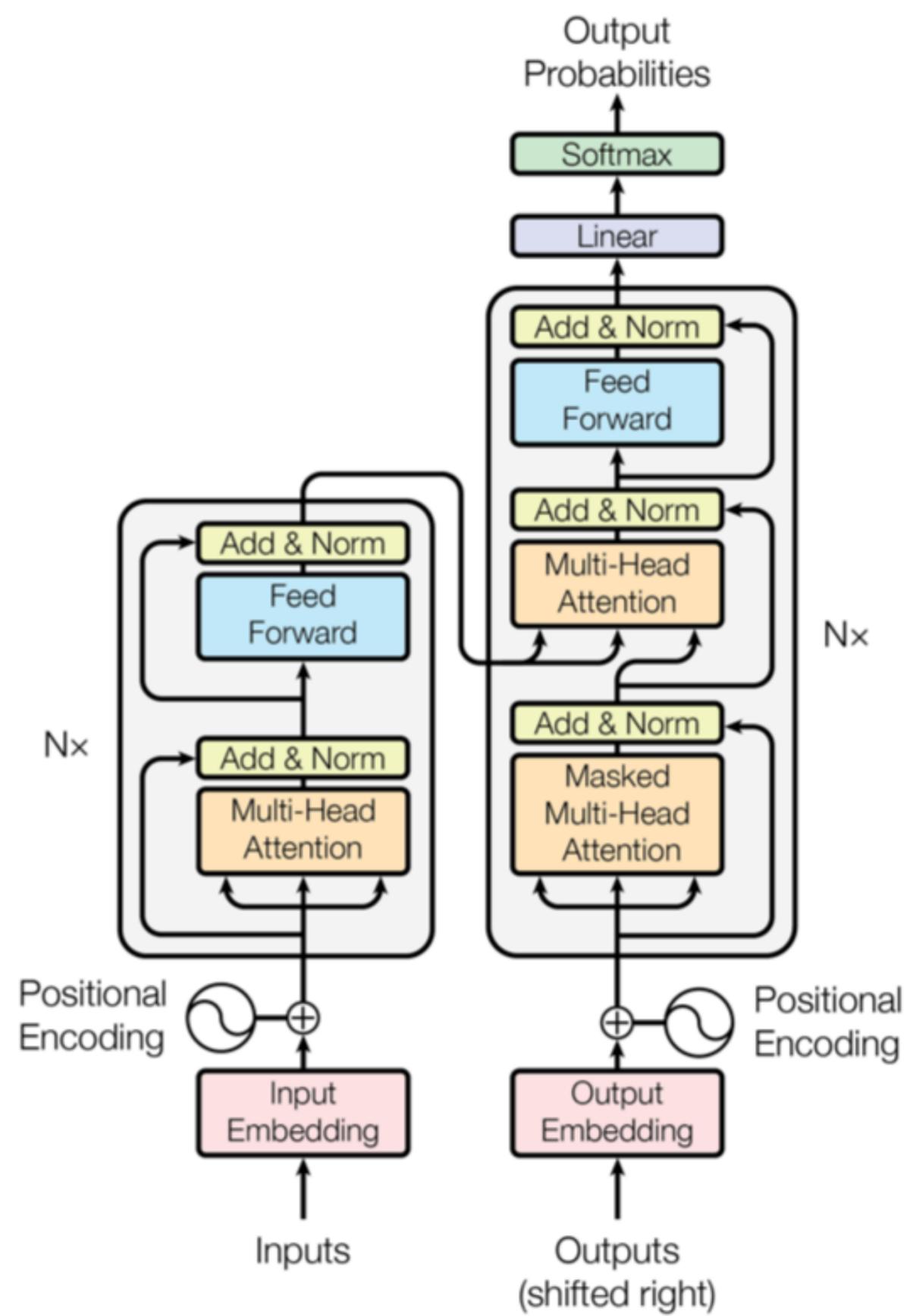


Figure 1: The Transformer - model architecture.

Transformers introduced
a couple key innovations

Positional **Self**
Encoding **&** **Attention**

Positional Encoding

- Instead of dealing with each word sequentially, one at a time, transformers **encode positional data**
- Before feeding each piece of the input into the neural network, we label it with positional information
- Word-order information is stored in the actual data itself rather than in the network structure
- The network learns the significance of word-order from the data itself

pickles are overrated

pickles

are

overrated

1

2

3

* This is a **vast** oversimplification. The model doesn't work with words.

Attention

- Attention in a neural network is a mechanism that allows the network to selectively focus on certain parts of input data, while ignoring others
- Think of how humans focus our attention on certain aspects of the world, while filtering out irrelevant information



Attention

- Attention allows the network to focus on parts of the input data and dynamically adjust its focus as it processes the data
- There are different attention mechanisms but most involve computing an attention score for each piece of input data
- These scores are then used to compute a weighted sum or average of the input elements



But How??

How does the model "know" what words it should "attend" to?

It's learned over time from lots and lots of data.

With enough data, the model learns the basics of grammar, word order, etc.

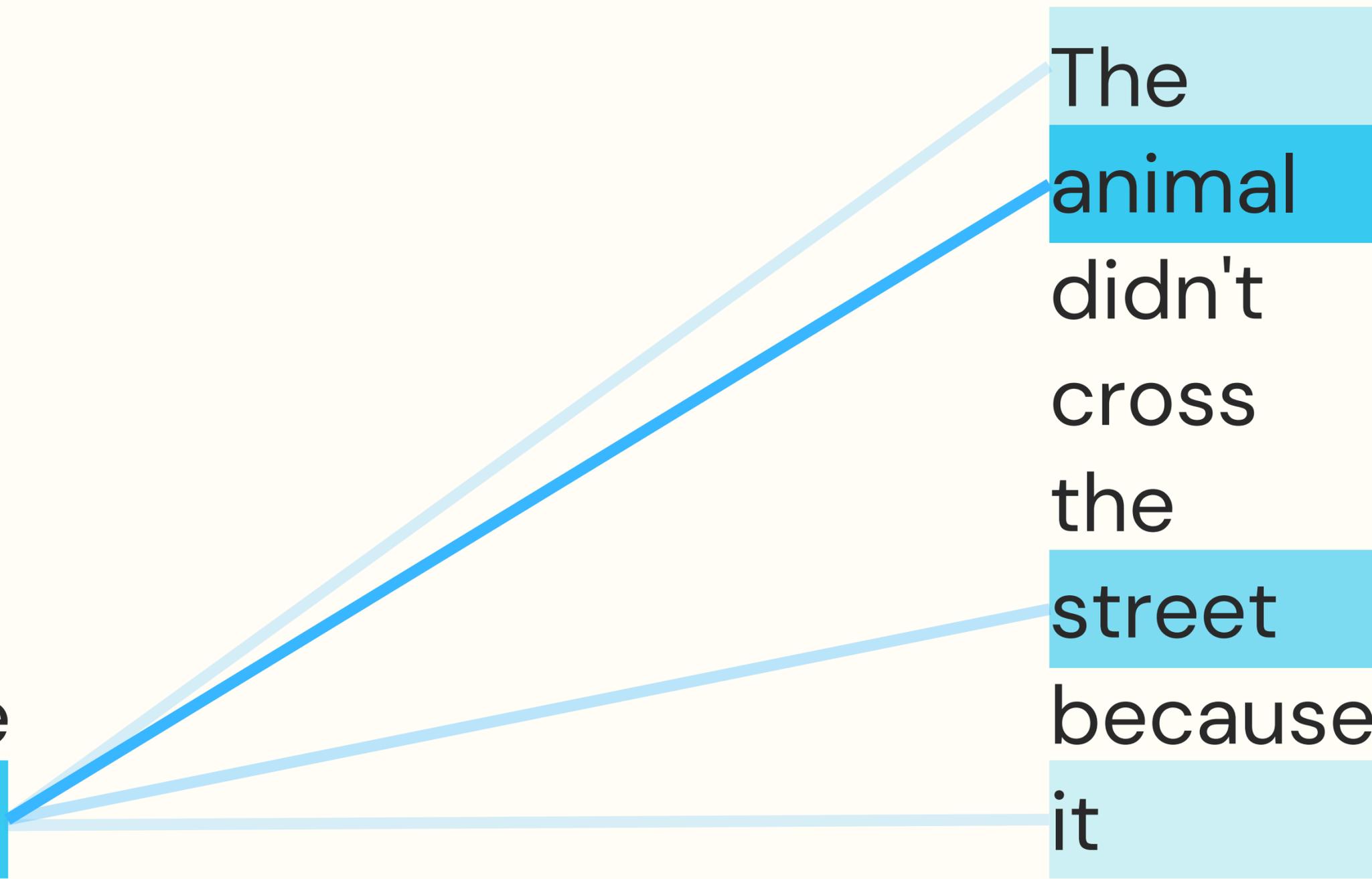
Self-Attention

- Self-attention is one of the key innovations that makes the transformer model so effective
- In self-attention, each element in the input sequence is compared to every other element in the sequence, and a set of attention weights is computed based on the similarity between each pair of elements
- The "self" in self-attention refers to the the same sequence which is currently being encoded.



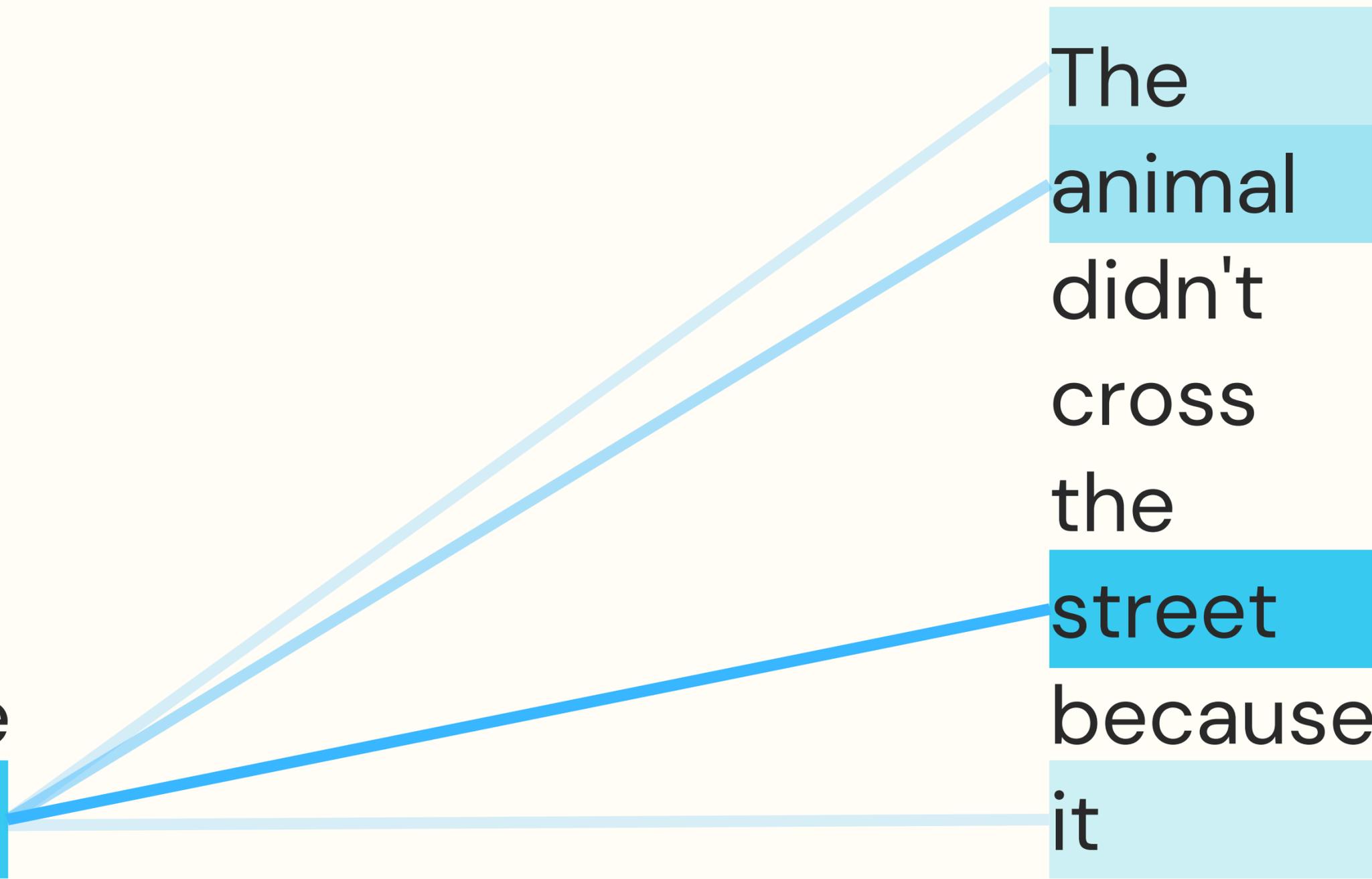
The
animal
didn't
cross
the
street
because
it
was
too
tired

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



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

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



Training

- GPT-3 was trained on over 45TB of text data:
 - A quality-filtered subset of the CommonCrawl Dataset
 - An expanded version of the *Webtext* dataset
 - All outbound links from Reddit with >3 Karma
 - Two databases of online books
 - English-language Wikipedia
- Nearly 500 billion tokens of training data
- Open AI has not released information on the training of GPT-4

Size

- GPT-3 is absolutely massive compared to GPT-2
- GPT-3 has 175 billion parameters and takes 800gb just to store the model itself
 - That's 800gb of basically numeric data that forms the model.
- It cost over \$4.6 million in GPU costs to initially train GPT-3
- OpenAI has not released the technical details of GPT-4

**Let's
Sign Up!**



**Our
First
Request!**



Completions



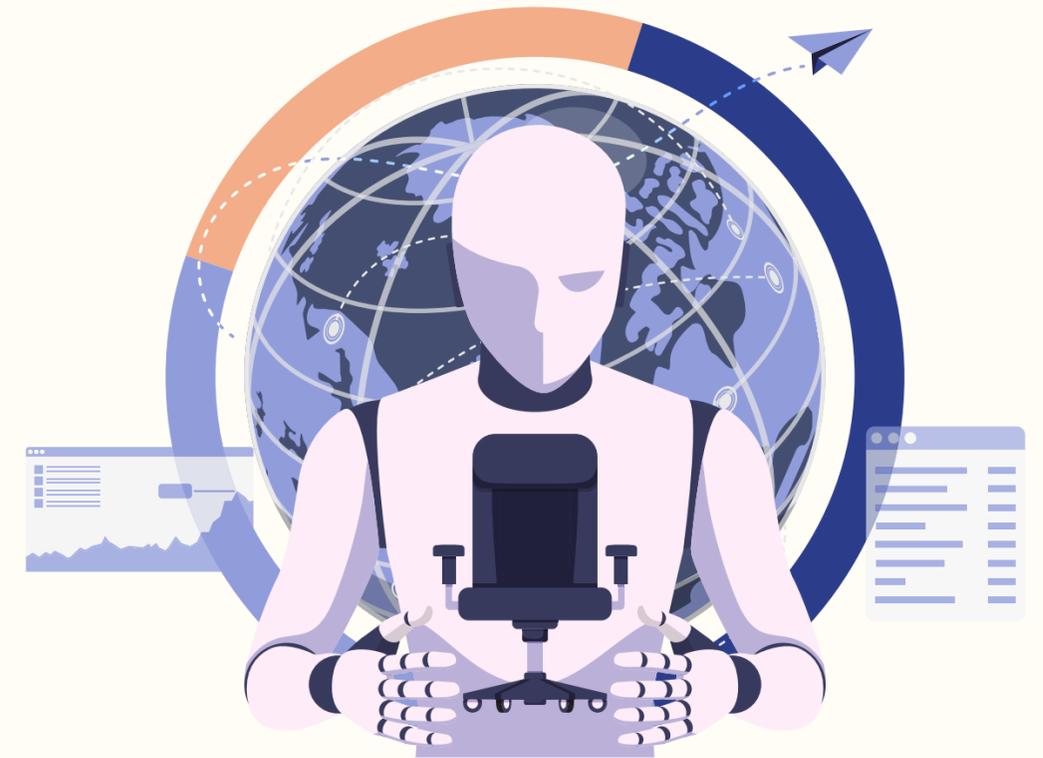
```
openai.Completion.create(  
    model="text-davinci-003",  
    prompt="tell me a joke"  
)
```


Models

- **DALL-E**: generates and edits images
- **Whisper**: converts audio to text
- **Moderation**: detects safe and unsensitive text
- **GPT-3**: understands and generates natural language
- **GPT-3.5**: set of models of that improve upon GPT-3
- **GPT-4**: The latest and most advanced version of OpenAi's large language model

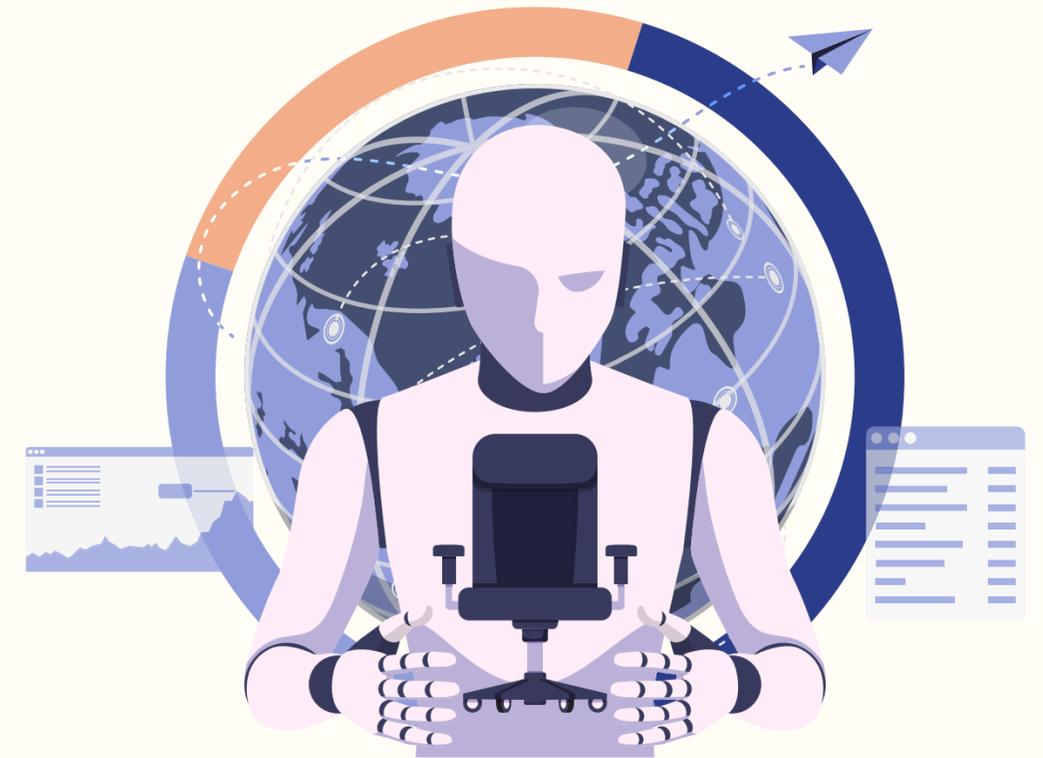
GPT-3.5

GPT-3.5 is **not an entirely new model**. It's a finely-tuned version of GPT-3 developed in 2022 and trained on data through 2021.



GPT-4

GPT-4 is a HUGE new model that is currently in beta. You must join a waitlist and be approved to gain access to GTP-4 via the APIs.



LATEST MODEL	DESCRIPTION	MAX REQUEST	TRAINING DATA
text-curie-001	Very capable, faster and lower cost than Davinci.	2,049 tokens	Up to Oct 2019
text-babbage-001	Capable of straightforward tasks, very fast, and lower cost.	2,049 tokens	Up to Oct 2019
text-ada-001	Capable of very simple tasks, usually the fastest model in the GPT-3 series, and lowest cost.	2,049 tokens	Up to Oct 2019
davinci	Most capable GPT-3 model. Can do any task the other models can do, often with higher quality.	2,049 tokens	Up to Oct 2019
curie	Very capable, but faster and lower cost than Davinci.	2,049 tokens	Up to Oct 2019
babbage	Capable of straightforward tasks, very fast, and lower cost.	2,049 tokens	Up to Oct 2019
ada	Capable of very simple tasks, usually the fastest model in the GPT-3 series, and lowest cost.	2,049 tokens	Up to Oct 2019

LATEST MODEL	DESCRIPTION	MAX REQUEST	TRAINING DATA
<code>gpt-3.5-turbo</code>	Most capable GPT-3.5 model and optimized for chat at 1/10th the cost of <code>text-davinci-003</code> . Will be updated with our latest model iteration.	4,096 tokens	Up to Sep 2021
<code>gpt-3.5-turbo-0301</code>	Snapshot of <code>gpt-3.5-turbo</code> from March 1st 2023. Unlike <code>gpt-3.5-turbo</code> , this model will not receive updates, and will only be supported for a three month period ending on June 1st 2023.	4,096 tokens	Up to Sep 2021
<code>text-davinci-003</code>	Can do any language task with better quality, longer output, and consistent instruction-following than the curie, babbage, or ada models. Also supports inserting completions within text.	4,097 tokens	Up to Jun 2021
<code>text-davinci-002</code>	Similar capabilities to <code>text-davinci-003</code> but trained with supervised fine-tuning instead of reinforcement learning	4,097 tokens	Up to Jun 2021
<code>code-davinci-002</code>	Optimized for code-completion tasks	8,001 tokens	Up to Jun 2021

LATEST MODEL	DESCRIPTION	MAX TOKENS	TRAINING DATA
gpt-4	More capable than any GPT-3.5 model, able to do more complex tasks, and optimized for chat. Will be updated with our latest model iteration.	8,192 tokens	Up to Sep 2021
gpt-4-0314	Snapshot of gpt-4 from March 14th 2023. Unlike gpt-4, this model will not receive updates, and will only be supported for a three month period ending on June 14th 2023.	8,192 tokens	Up to Sep 2021
gpt-4-32k	Same capabilities as the base gpt-4 mode but with 4x the context length. Will be updated with our latest model iteration.	32,768 tokens	Up to Sep 2021
gpt-4-32k-0314	Snapshot of gpt-4-32 from March 14th 2023. Unlike gpt-4-32k, this model will not receive updates, and will only be supported for a three month period ending on June 14th 2023.	32,768 tokens	Up to Sep 2021

Pricing

- **text-davinci-003**: \$0.02 / 1K tokens
- **text-curie-001**: \$0.002 / 1K tokens
- **text-babbage-001**: \$0.005 / 1K tokens
- **text-ada-001**: \$0.0004 / 1K Tokens
- **gpt-3.5-turbo**: \$0.002 / 1K tokens
- **GPT-4 Models**: \$0.06 – \$0.12 / 1K Tokens

Tokens

- GPT doesn't work with words, but instead uses a system of tokens.
- Tokens are essentially pieces of words (though some tokens are full words)
- A token on average is ~4 characters of English text



I like hamburgers

Pricing

- Open AI charges based on tokens
- It adds together the tokens in your prompt plus the tokens in the output it returns
- Different models are priced differently (we'll learn more about this later)



Prompt Design



Prompt Design

- **Main Instructions** – a task you want the model to perform
- **Data** – any input data (if necessary)
- **Output Instructions** – what type of output do you want? What format?

Provide clear instructions

Complete the sentence:

Humans are

Use a separator to designate instructions and input

Instruction

Translate the text below to French:

Text: "I am a huge idiot"

Reduce "fluffy" language. Be precise

In 3–4 sentences, explain the role of the Supreme Court in US politics.

The explanation should be geared towards middle-schoolers

Be specific about your desired output

Extract the place names in the following

Desired format:

Places: <comma_separated_list_of_places>

Input: " Airbnb, Inc.is an American San Francisco-based company. The company is credited with revolutionizing the tourism industry however it has also been the subject of intense criticism by residents of tourism hotspot cities like Barcelona, Venice, etc. for enabling an unaffordable increase in home rents, and for a lack of regulation."

Zero-Shot

Extract keywords from the below text.

Text: {text}

Keywords:

Few-Shot

Extract keywords from the corresponding texts below.

Text 1: Stripe provides APIs that web developers can use to integrate payment processing into their websites and mobile applications.

Keywords 1: Stripe, payment processing, APIs, web developers, websites, mobile applications
##

Text 2: OpenAI has trained cutting-edge language models that are very good at understanding and generating text. Our API provides access to these models and can be used to solve virtually any task that involves processing language.

Keywords 2: OpenAI, language models, text processing, API.
##

Text 3: {text}

Keywords 3:

Temperature

- A value from 0-2, though most often between 0 and 1
- Its default value is 1
- Controls the randomness of the output. Higher values are more random, lower values are more deterministic



Temperature

- Temperature works by scaling the logits (a measure of probability distribution over the possible next words)
- The logits are divided by the temperature value before applying the softmax function
- This results in a "softer" probability distribution with a higher temperature and a peaked distribution with low temp

Temperature

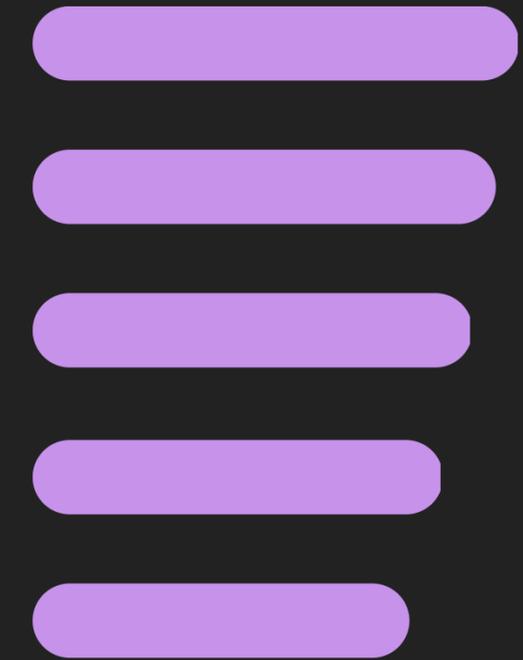
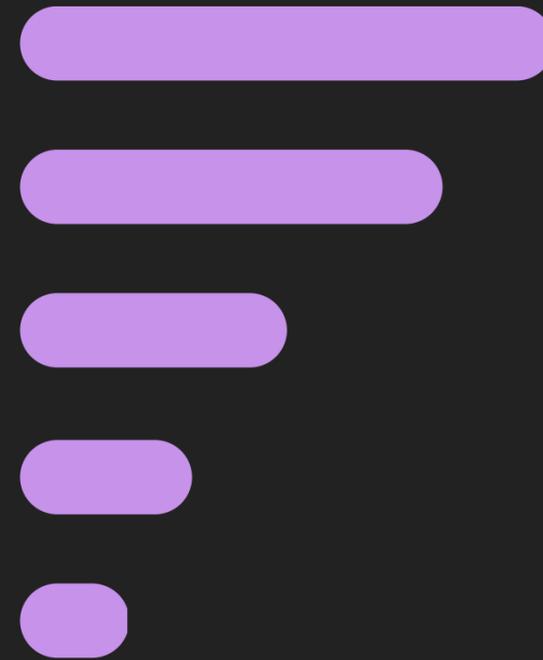
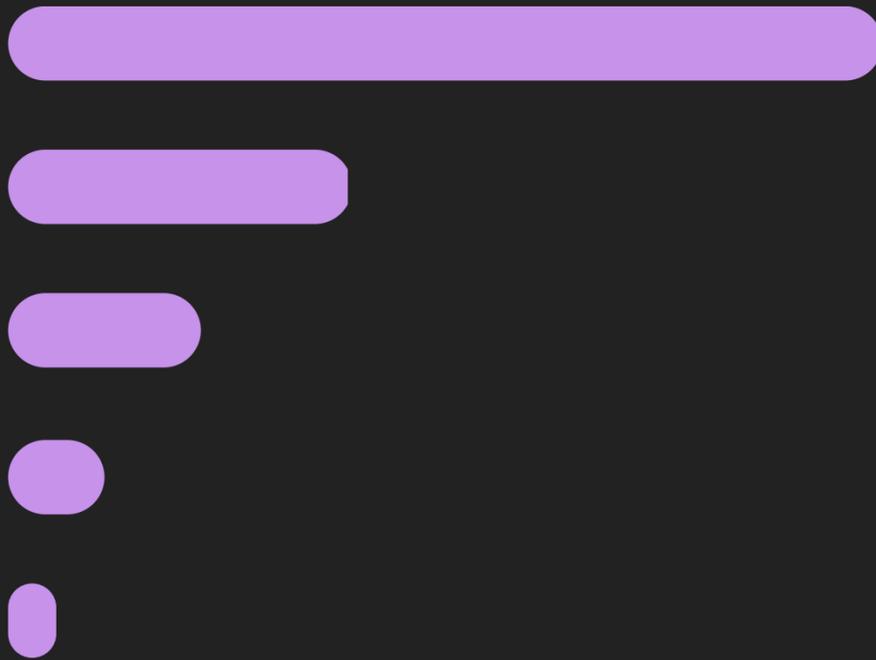
low temp

high temp

probabilities

probabilities

probabilities

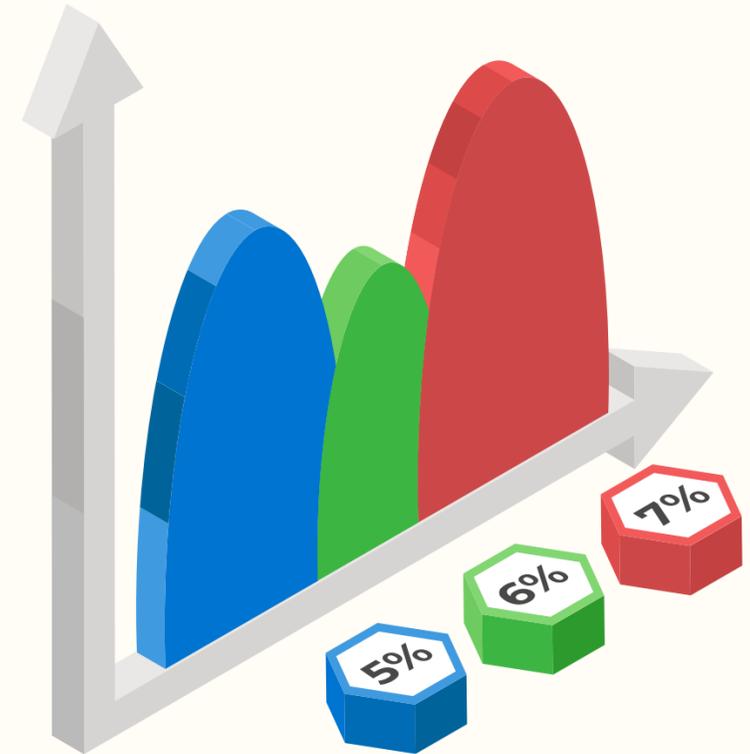


less random

more random

Top P

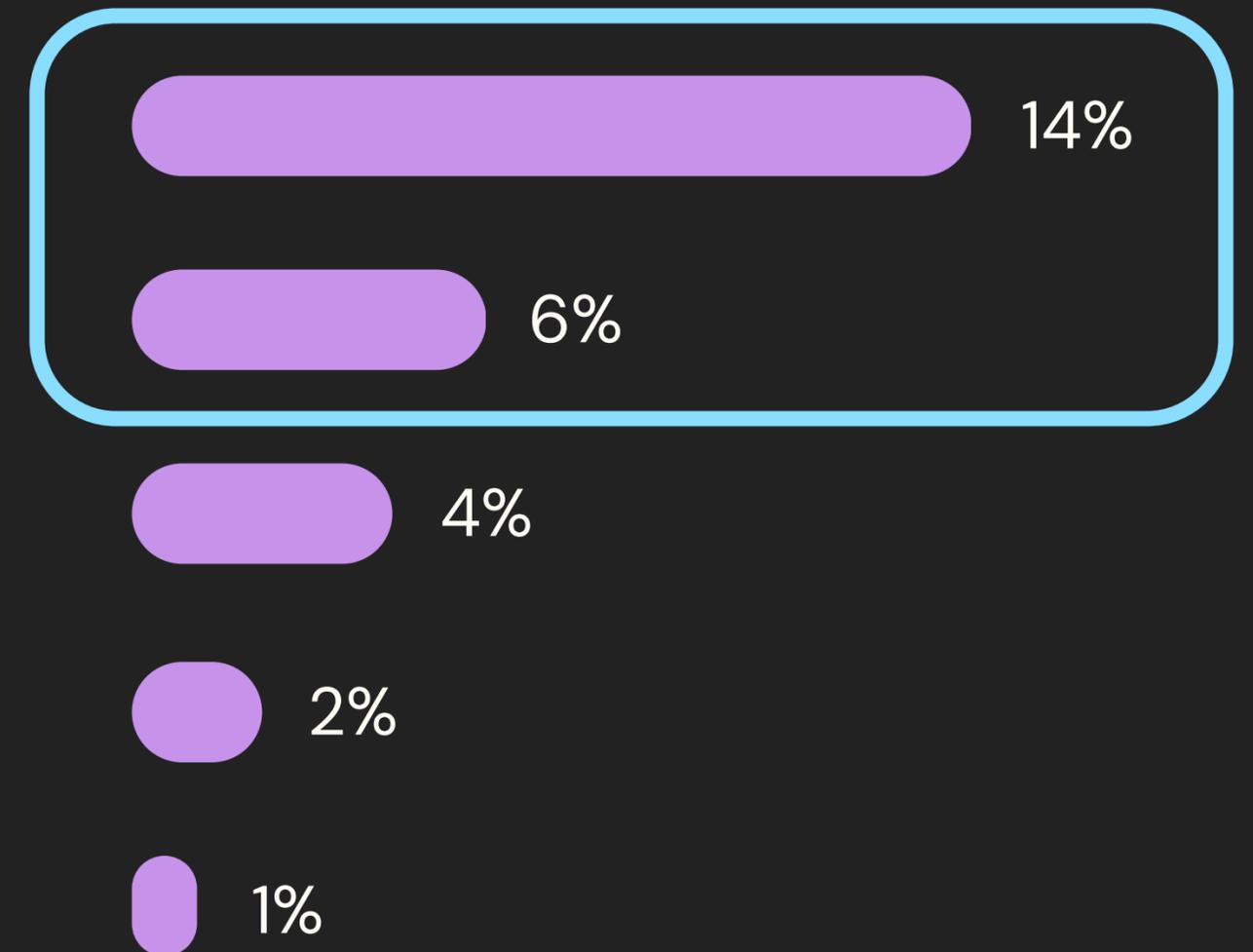
- An alternative to sampling with temperature, called nucleus sampling
- Its default value is 1
- Like temperature, top p alters the "creativity" and randomness of the output



Top P

- Top p controls the set of possible words/tokens that the model can choose from
- It restricts the candidate words to the smallest set whose cumulative probability is greater than or equal to a given threshold, "p"

top p = 0.2



Frequency Penalty

- A number from -2 to 2
- Defaults to 0
- Positive values penalize new tokens based on their existing frequency in the text so far, **decreasing the model's likelihood to repeat the same line verbatim.**



Frequency Penalty

- If you want to reduce repetitive samples, try a penalty from 0.1 – 1
- To strongly suppress repetition, you can increase it further BUT this can lead to bad quality outputs
- Negative values increase the likelihood of repetition



Presence Penalty

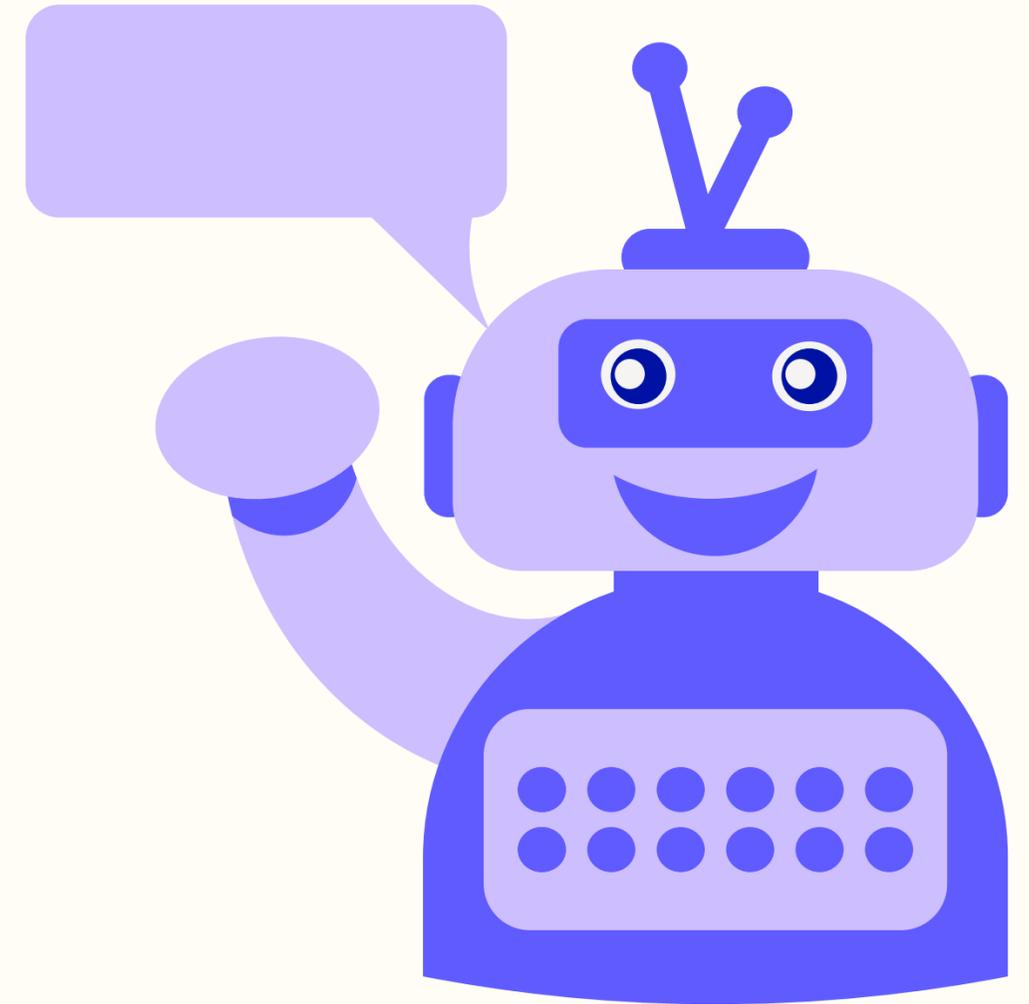
- A number from -2 to 2
- Defaults to 0
- Positive values penalize new tokens based on whether they appear in the text so far, **increasing the model's likelihood to talk about new topics.**



- Presence penalty is a **one-off additive contribution** that applies to all tokens that have been sampled at least once
- Frequency penalty is a contribution that is **proportional** to how often a particular token has already been sampled

ChatGPT API

- The ChatGPT API allows us to use `gpt-3.5-turbo` and `gpt-4`
- It uses a chat format designed to make multi-turn conversations easy
- It also can be used for any single-turn tasks that we've done with the completion API



Completions



```
openai.Completion.create(  
  model="text-davinci-003",  
  prompt="tell me a joke"  
)
```

ChatGPT API

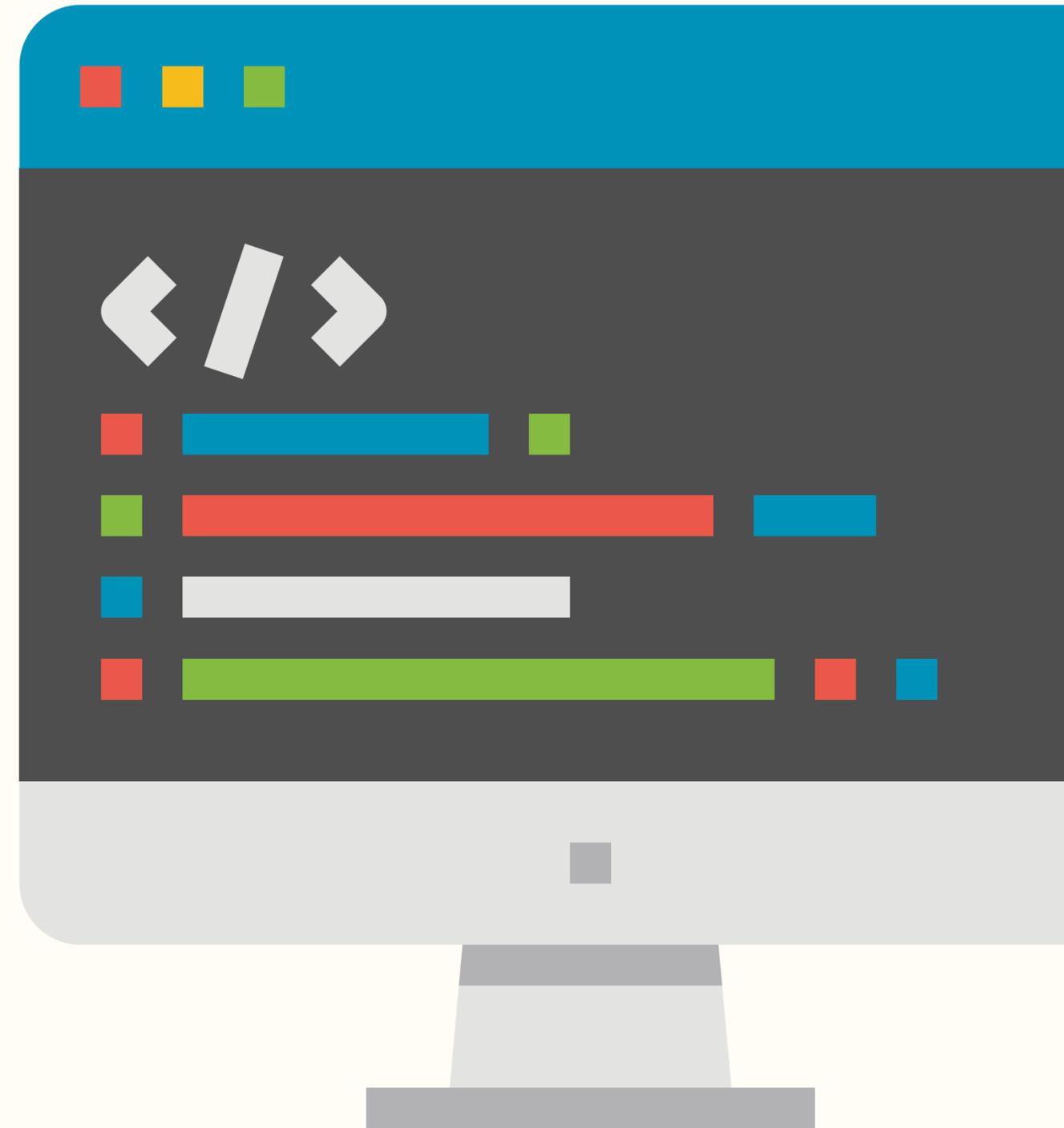


```
openai.ChatCompletion.create(  
  model="gpt-3.5-turbo",  
  messages=[  
    {"role": "user", "content": "tell me a joke"}  
  ]  
)
```

Messages

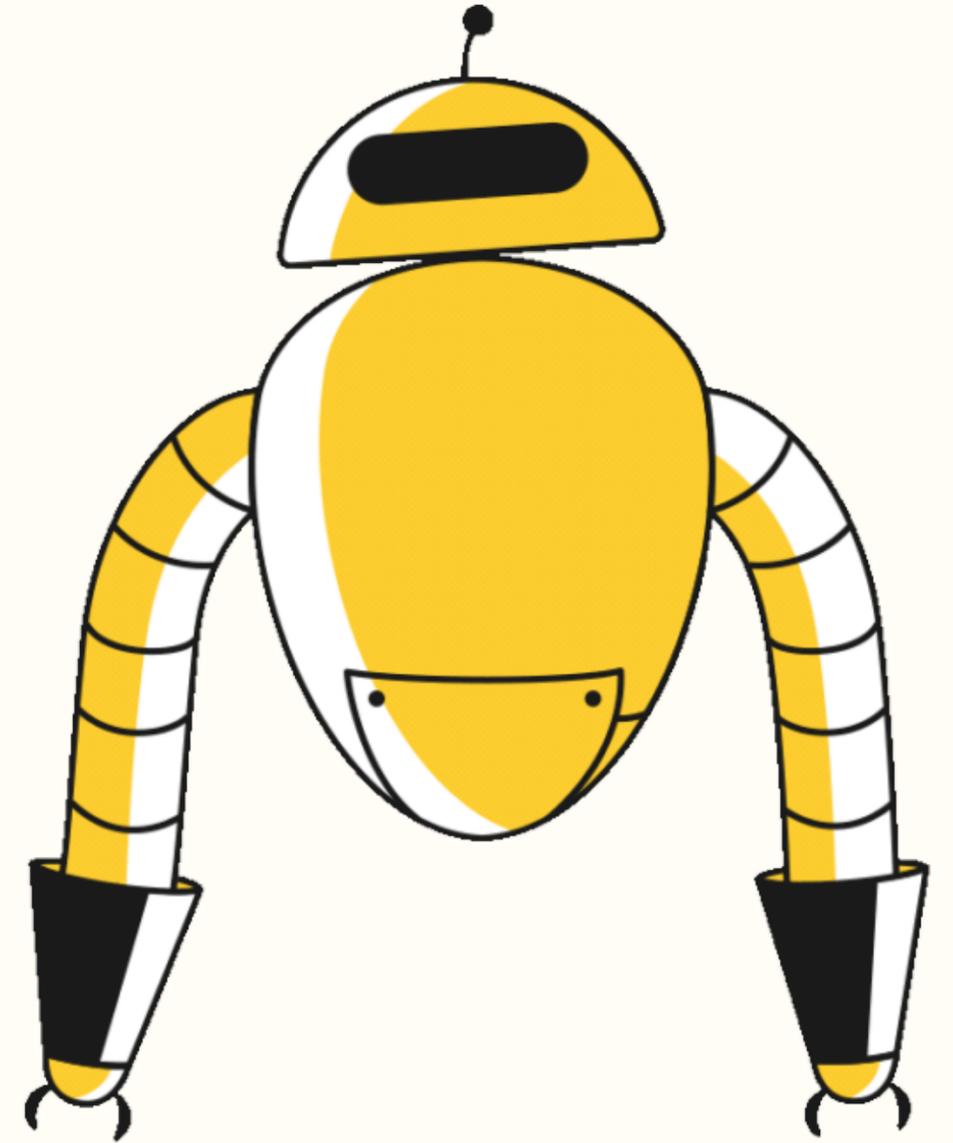
- The chat API expects a list of messages rather than a single text prompt
- Messages must be an array of message objects, where each object has:
 - a **"role"**, set to:
 - "system",
 - "user"
 - "assistant"
 - **"content"** (the content of the message)

Working With Code



DALL-E

- DALL-E is a neural network-based image generation system.
- It **generates images from text prompts**
- To train DALL-E, OpenAI used a dataset of over 250 million images and associated text descriptions

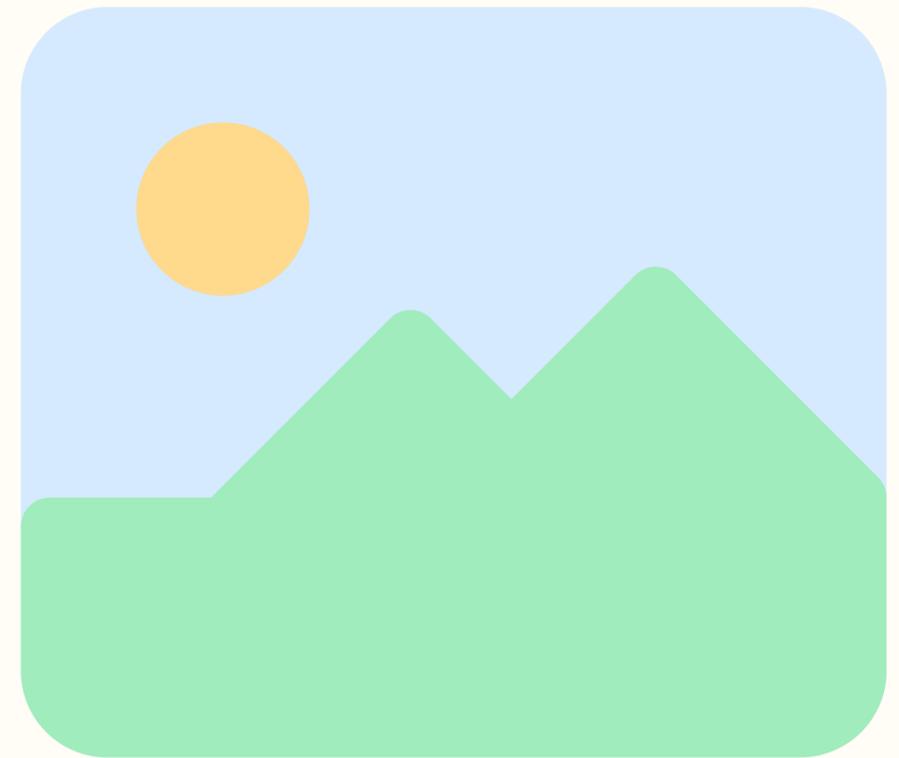




```
response = openai.Image.create(  
    prompt="a dancing pickle",  
    n=1,  
    size="1024x1024"  
)  
image_url = response['data'][0]['url']
```

Image Sizes

- The DALL-E API only supports square images of the following sizes:
 - 256x256 pixels
 - 512x512 pixels
 - 1024x1024 pixels
- The smaller the image, the faster it is to generate



DALL-E Pricing

Resolution	Price
1024×1024	\$0.020 / image
512×512	\$0.018 / image
256×256	\$0.016 / image

Whisper

- Whisper is OpenAI's speech recognition model.
- It can perform speech recognition, translation, and language identification
- It costs **\$0.006 / minute** (rounded to the nearest second)



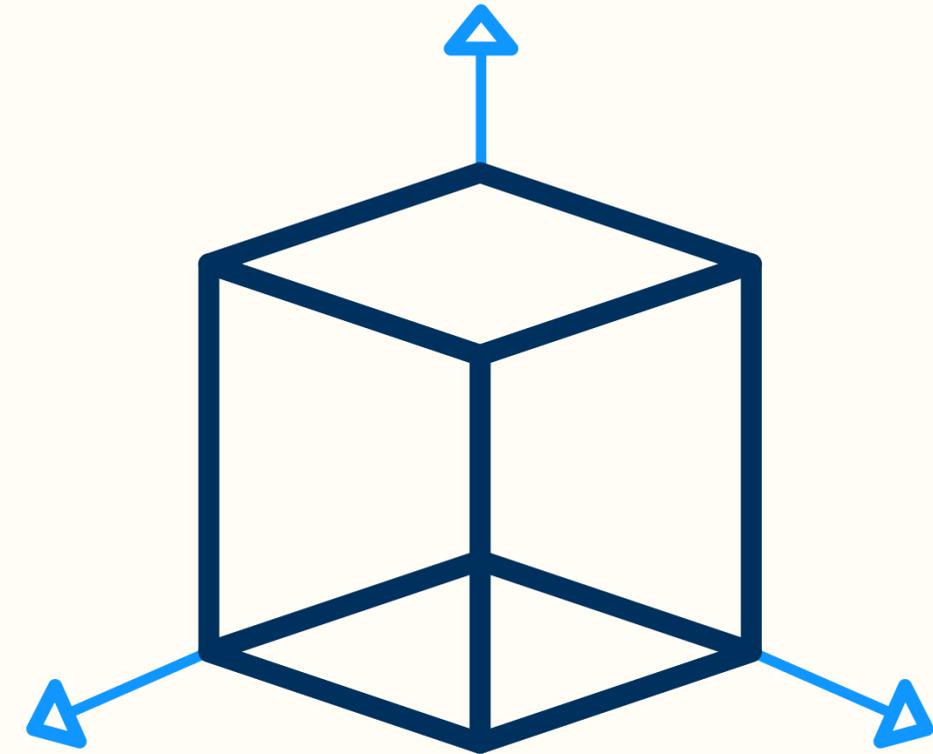
TikToken

Tokenizer Library



Embeddings

- Embeddings are numerical representations of text concepts converted to number sequences
- They make it easy for computers to understand the relationships between those concepts.



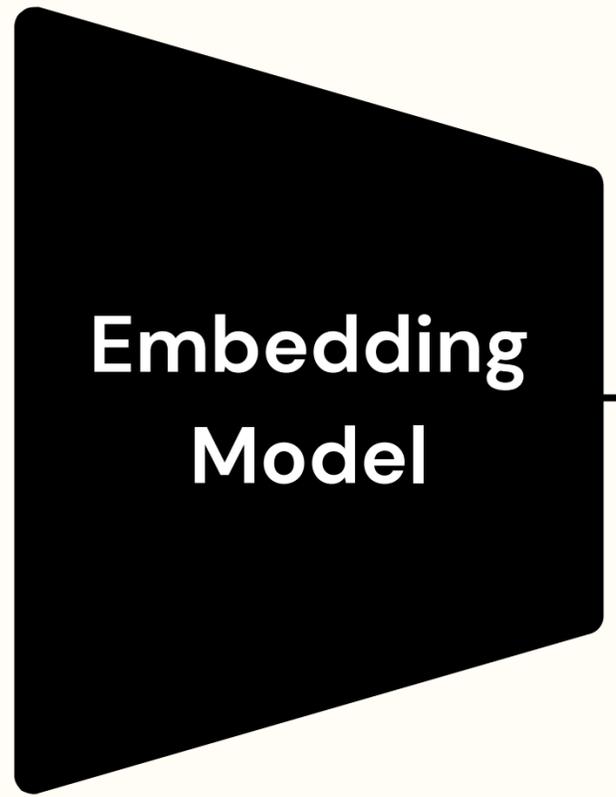
Embeddings

- OpenAI has an embedding model called **text-embedding-ada-002**
- Given some input text, it returns an embedding as a 1536 dimension vector
- We can store these embeddings and then use them to perform searches, recommendations, and more



pickled feet

Input Text



**Embedding
Model**



-0.004 -0.001 ... 0.032

Text As Vector

orange juice

Input Text

**Embedding
Model**

-0.002

0.012

...

0.022

Text As Vector

Embedding Q&A

1. Generate a bunch of embeddings on your own specific data
2. When a user asks a question, take their question and turn it into an embedding
3. Find the K nearest neighbors to that embedding
4. Include the matching text(s) in the prompt when you query the model



summarize()

summarize()

+

summarize()

summarize()

+

summarize()

summarize()

+

summarize()

By three bells that morning they were all stirring their stumps; for there was a big sea running; and Tootles, the bo'sun, was among them, with a rope's end in his hand and chewing tobacco. They all donned pirate clothes cut off at the knee, shaved smartly, and tumbled up, with the true nautical roll and hitching their trousers. It need not be said who was the captain. Nibs and John were first and second mate. There was a woman aboard. The rest were tars before the mast, and lived in the fo'c'sle. Peter had already lashed himself to the wheel; but he piped all hands and delivered a short address to them; said he hoped they would do their duty like gallant hearties, but that he knew they were the scum of Rio and the Gold Coast, and if they snapped at him he would tear them. The bluff strident words struck the note sailors understood, and they cheered him lustily. Then a few sharp

orders were given, and they turned the snip round, and nosed her for the mainland.

Captain Pan calculated, after consulting the ship's chart, that if this weather lasted they should strike the Azores about the 21st of June, after which it would save time to fly.

Some of them wanted it to be an honest ship and others were in favour of keeping it a pirate; but the captain treated them as dogs, and they dared not express their wishes to him even in a round robin. Instant obedience was the only safe thing. Slightly got a dozen for looking perplexed when told to take soundings. The general feeling was that Peter was honest just now to lull Wendy's suspicions, but that there might be a change when the new suit was ready, which, against her will, she was making for him out of some of Hook's wickedest garments. It was afterwards whispered among them that on the first night he wore this suit he sat long in the cabin with Hook's cigar-holder in his mouth and one hand clenched, all but for the forefinger, which he bent and held threateningly aloft like a hook.

Instead of watching the ship, however, we must now return to that desolate home from which three of our characters had taken heartless

flight so long ago. It seems a shame to have neglected No. 14 all this time; and yet we may be sure that Mrs. Darling does not blame us. If we had returned sooner to look with sorrowful sympathy at her, she would probably have cried, "Don't be silly; what do I matter? Do go back and keep an eye on the children." So long as mothers are like this their children will take advantage of them; and they may lay to that. Even now we venture into that familiar nursery only because its lawful occupants are on their way home; we are merely hurrying on in advance of them to see that their beds are properly aired and that Mr. and Mrs. Darling do not go out for the evening. We are no more than servants. Why on earth should their beds be properly aired, seeing that they left them in such a thankless hurry? Would it not serve them jolly well right if they came back and found that their parents were spending the week-end in the country? It would be the moral lesson they have been in need of ever since we met them; but if we contrived things in this way Mrs. Darling would never forgive us.

One thing I should like to do immensely, and that is to tell her, in the way authors have, that the children are coming back, that indeed they will be here on Thursday week. This would spoil so completely the surprise to which Wendy and John and Michael are looking forward. They have been planning it out on the ship: mother's rapture, father's shout of joy, Nana's leap through the air to embrace them first, when what they ought to be prepared for is a good hiding. How delicious to spoil it all by breaking the news in advance; so that when they enter grandly Mrs. Darling may not even offer Wendy her mouth, and Mr. Darling may exclaim pettishly, "Dash it all, here are those boys again." However, we should get no thanks even for this. We are beginning to know Mrs. Darling by this time, and may be sure that she would upbraid us for depriving the children of their little pleasure.