# GPT3

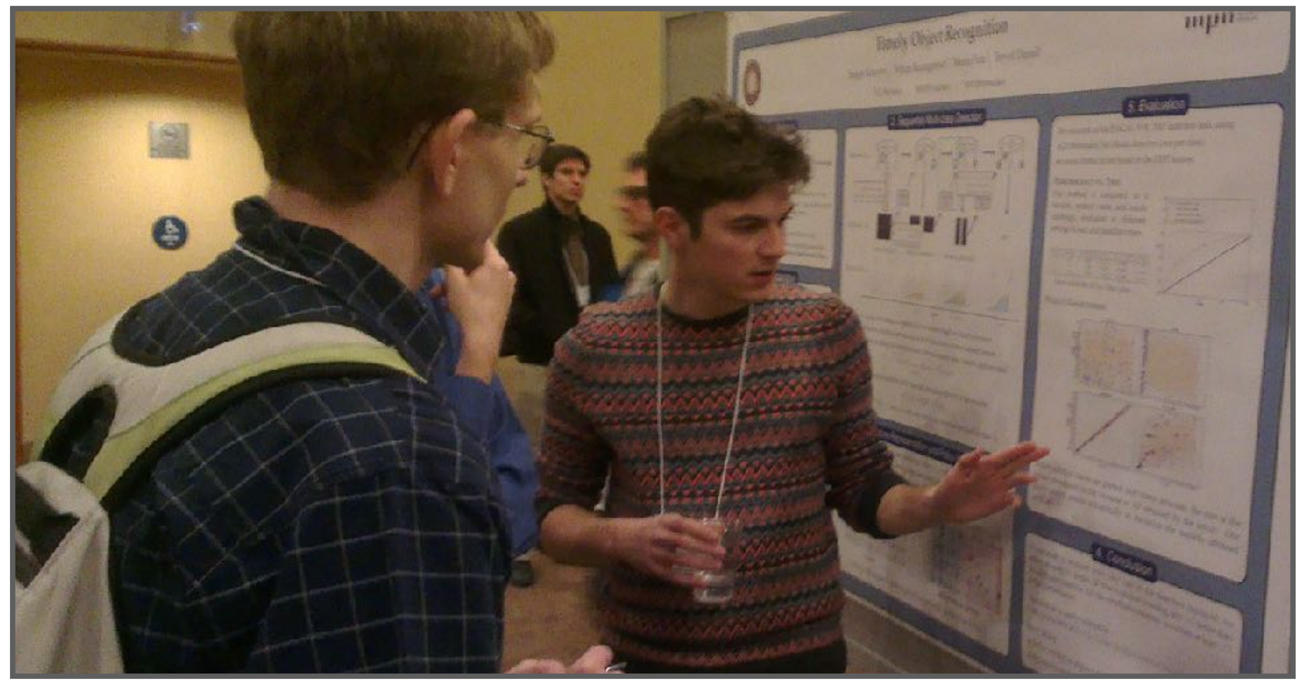
## What it is and how it could affect education

**Sergey Karayev**
GSV Breakfast Club
January 2021

# About Me

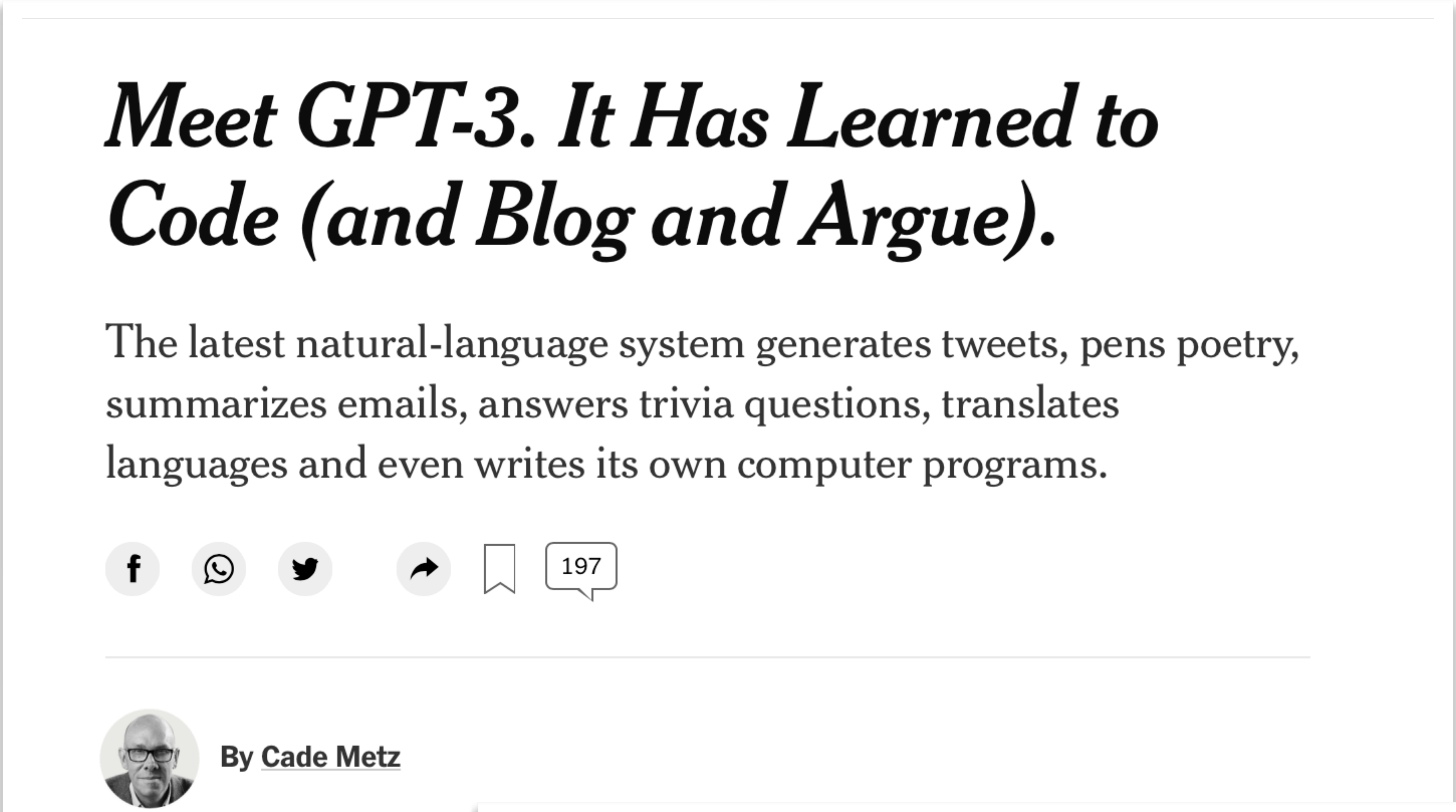
- Head of AI for STEM at **Turnitin** [2018-]
- Co-founder of **Gradescope** [2014-2018]

- Co-organizer of **Full Stack Deep Learning** (weekend bootcamps, online course, and official UW and UC Berkeley courses)



- PhD Computer Science at **UC Berkeley** [2009-2014]

# You've heard the news...



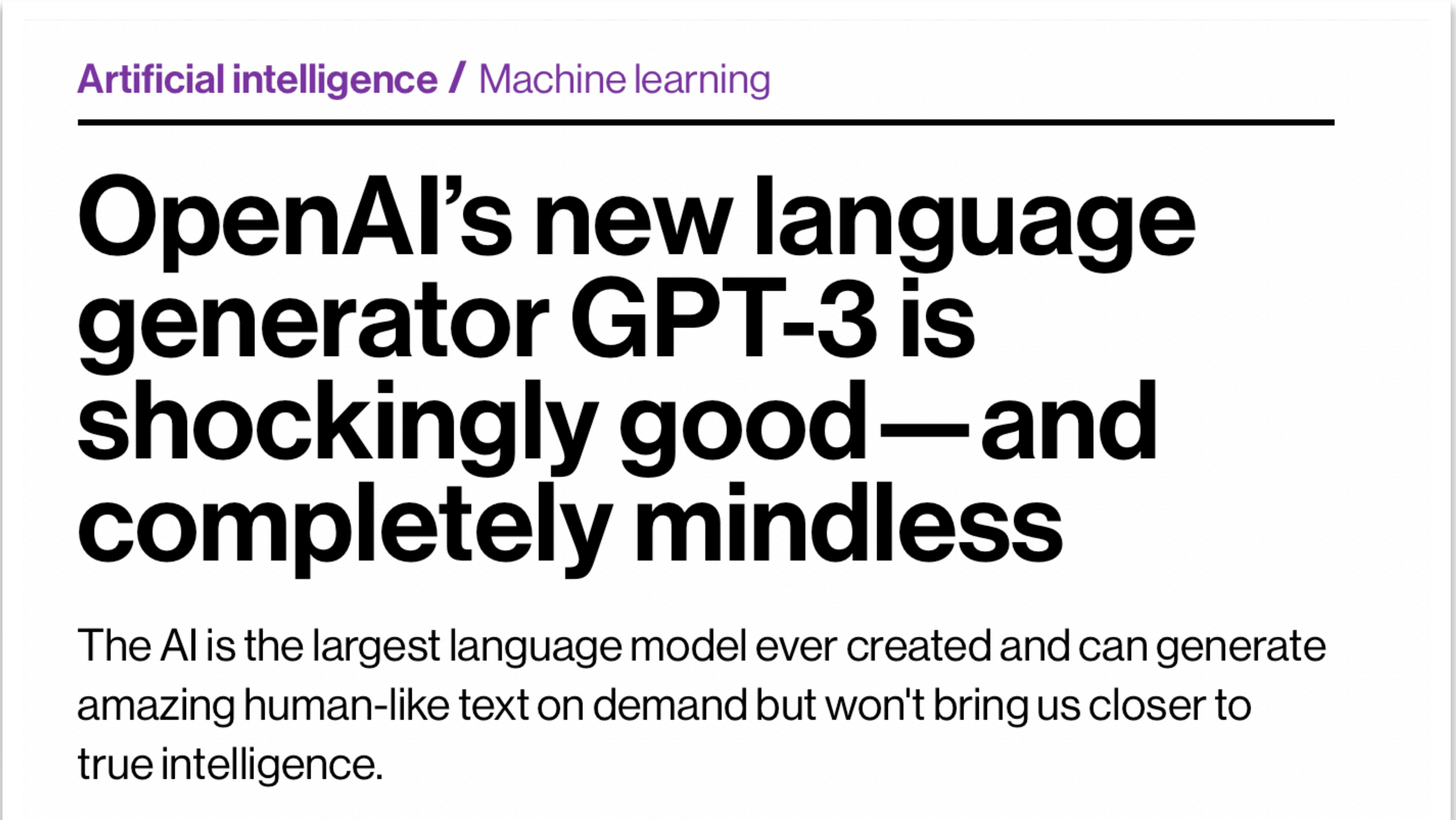
## *Meet GPT-3. It Has Learned to Code (and Blog and Argue).*

The latest natural-language system generates tweets, pens poetry, summarizes emails, answers trivia questions, translates languages and even writes its own computer programs.

197

By **Cade Metz**

Nov. 24, 2020

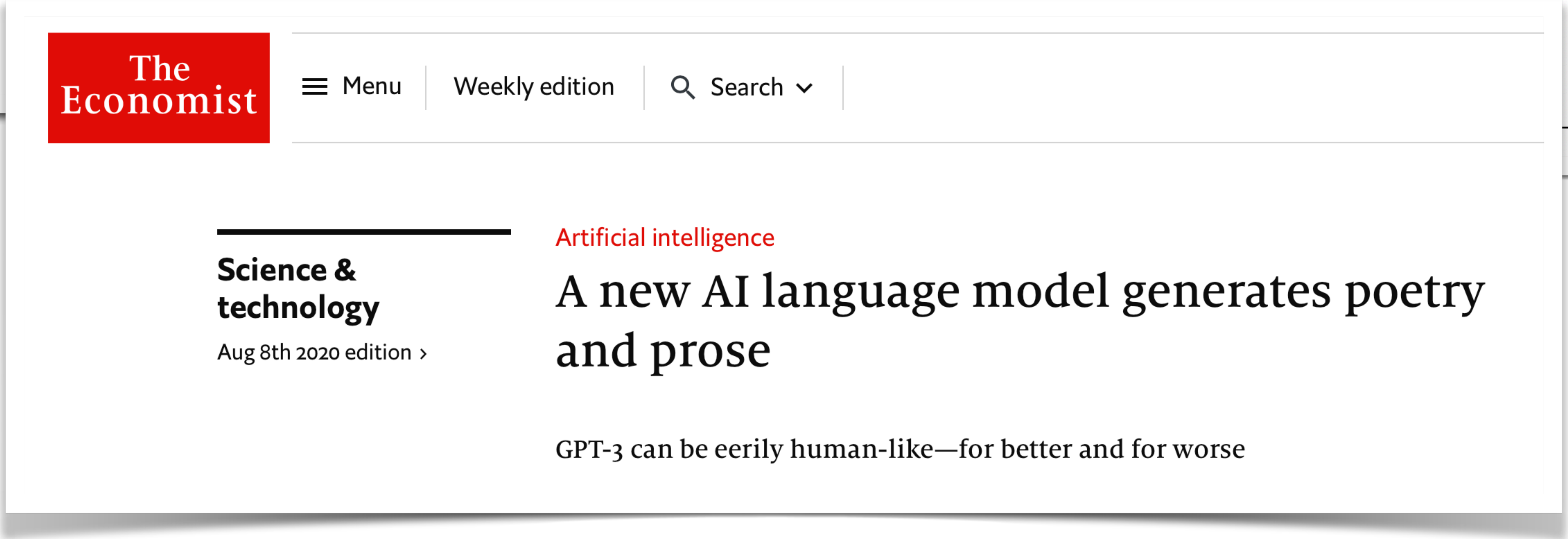


**Artificial intelligence** / Machine learning

## OpenAI's new language generator GPT-3 is shockingly good—and completely mindless

The AI is the largest language model ever created and can generate amazing human-like text on demand but won't bring us closer to true intelligence.

July 20, 2020



The Economist

**Science & technology**

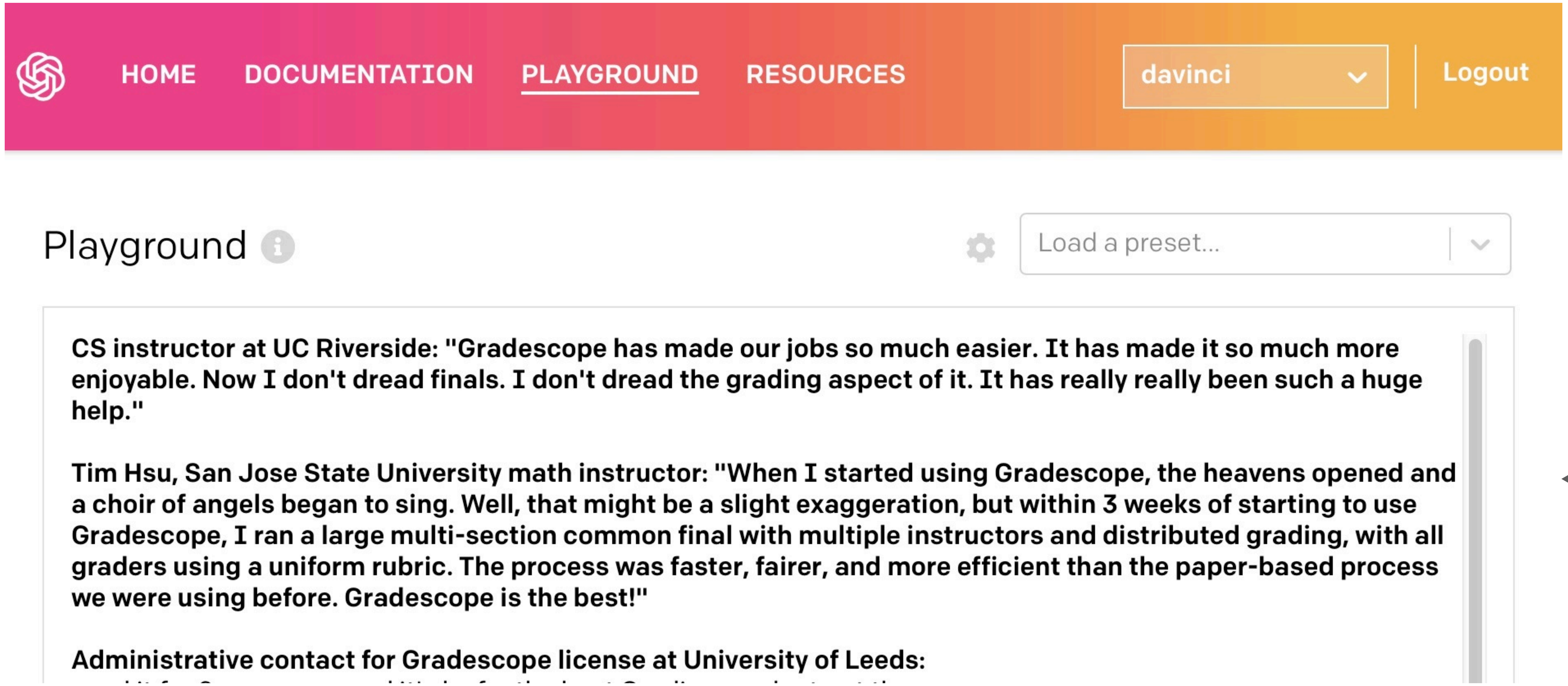
Aug 8th 2020 edition ›

Artificial intelligence

## A new AI language model generates poetry and prose

GPT-3 can be eerily human-like—for better and for worse

# Really, really good text generation



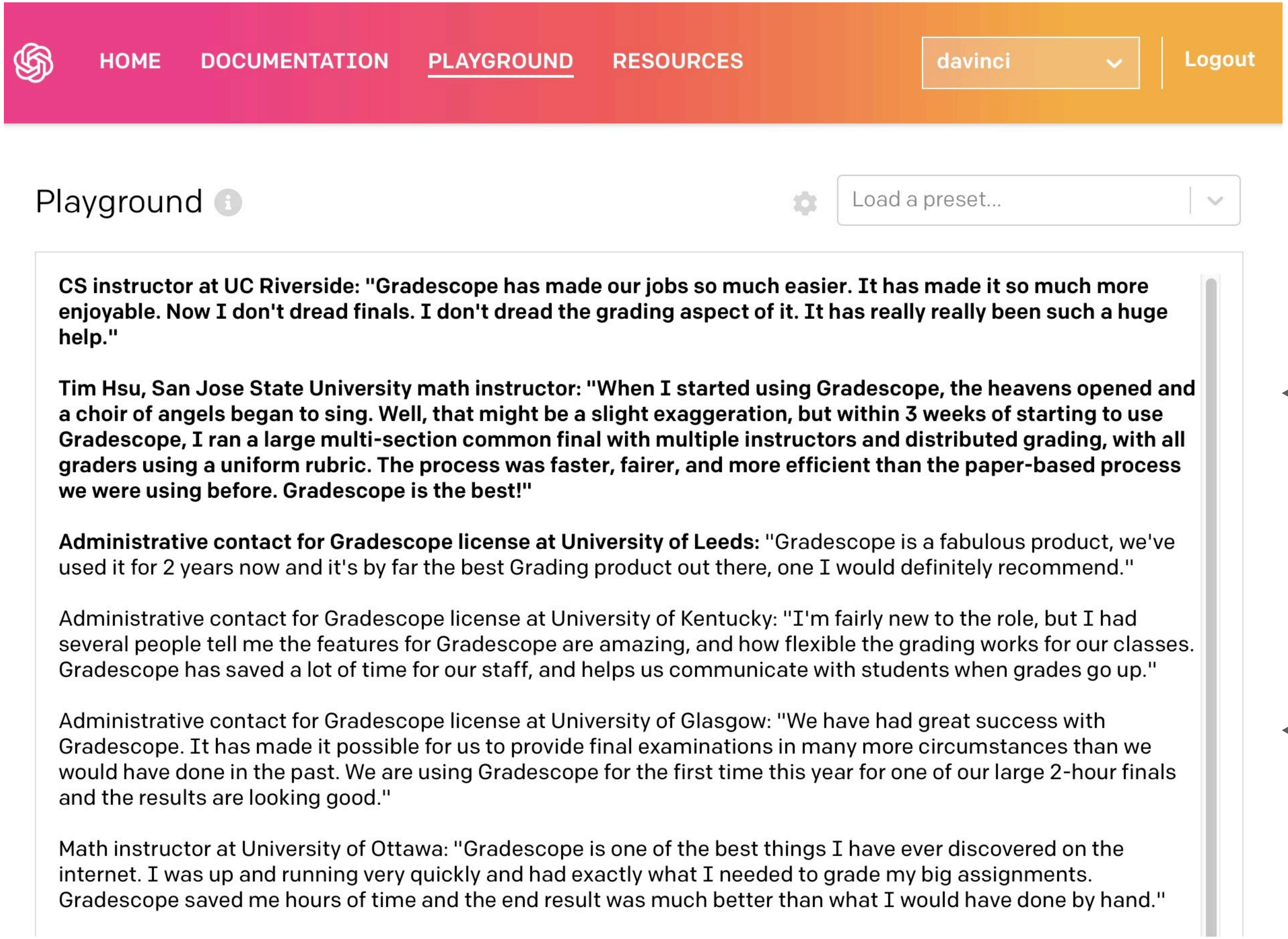
CS instructor at UC Riverside: "Gradescope has made our jobs so much easier. It has made it so much more enjoyable. Now I don't dread finals. I don't dread the grading aspect of it. It has really really been such a huge help."

Tim Hsu, San Jose State University math instructor: "When I started using Gradescope, the heavens opened and a choir of angels began to sing. Well, that might be a slight exaggeration, but within 3 weeks of starting to use Gradescope, I ran a large multi-section common final with multiple instructors and distributed grading, with all graders using a uniform rubric. The process was faster, fairer, and more efficient than the paper-based process we were using before. Gradescope is the best!"

Administrative contact for Gradescope license at University of Leeds:

Provided by me

# Really, really good text generation



**CS instructor at UC Riverside: "Gradescope has made our jobs so much easier. It has made it so much more enjoyable. Now I don't dread finals. I don't dread the grading aspect of it. It has really really been such a huge help."**

**Tim Hsu, San Jose State University math instructor: "When I started using Gradescope, the heavens opened and a choir of angels began to sing. Well, that might be a slight exaggeration, but within 3 weeks of starting to use Gradescope, I ran a large multi-section common final with multiple instructors and distributed grading, with all graders using a uniform rubric. The process was faster, fairer, and more efficient than the paper-based process we were using before. Gradescope is the best!"**

**Administrative contact for Gradescope license at University of Leeds:** "Gradescope is a fabulous product, we've used it for 2 years now and it's by far the best Grading product out there, one I would definitely recommend."

Administrative contact for Gradescope license at University of Kentucky: "I'm fairly new to the role, but I had several people tell me the features for Gradescope are amazing, and how flexible the grading works for our classes. Gradescope has saved a lot of time for our staff, and helps us communicate with students when grades go up."

Administrative contact for Gradescope license at University of Glasgow: "We have had great success with Gradescope. It has made it possible for us to provide final examinations in many more circumstances than we would have done in the past. We are using Gradescope for the first time this year for one of our large 2-hour finals and the results are looking good."

Math instructor at University of Ottawa: "Gradescope is one of the best things I have ever discovered on the internet. I was up and running very quickly and had exactly what I needed to grade my big assignments. Gradescope saved me hours of time and the end result was much better than what I would have done by hand."

Provided by me

Generated

**Administrative contact for Gradescope license at University of Leeds:** "Gradescope is a fabulous product, we've used it for 2 years now and it's by far the best Grading product out there, one I would definitely recommend."

Administrative contact for Gradescope license at University of Kentucky: "I'm fairly new to the role, but I had several people tell me the features for Gradescope are amazing, and how flexible the grading works for our classes. Gradescope has saved a lot of time for our staff, and helps us communicate with students when grades go up."

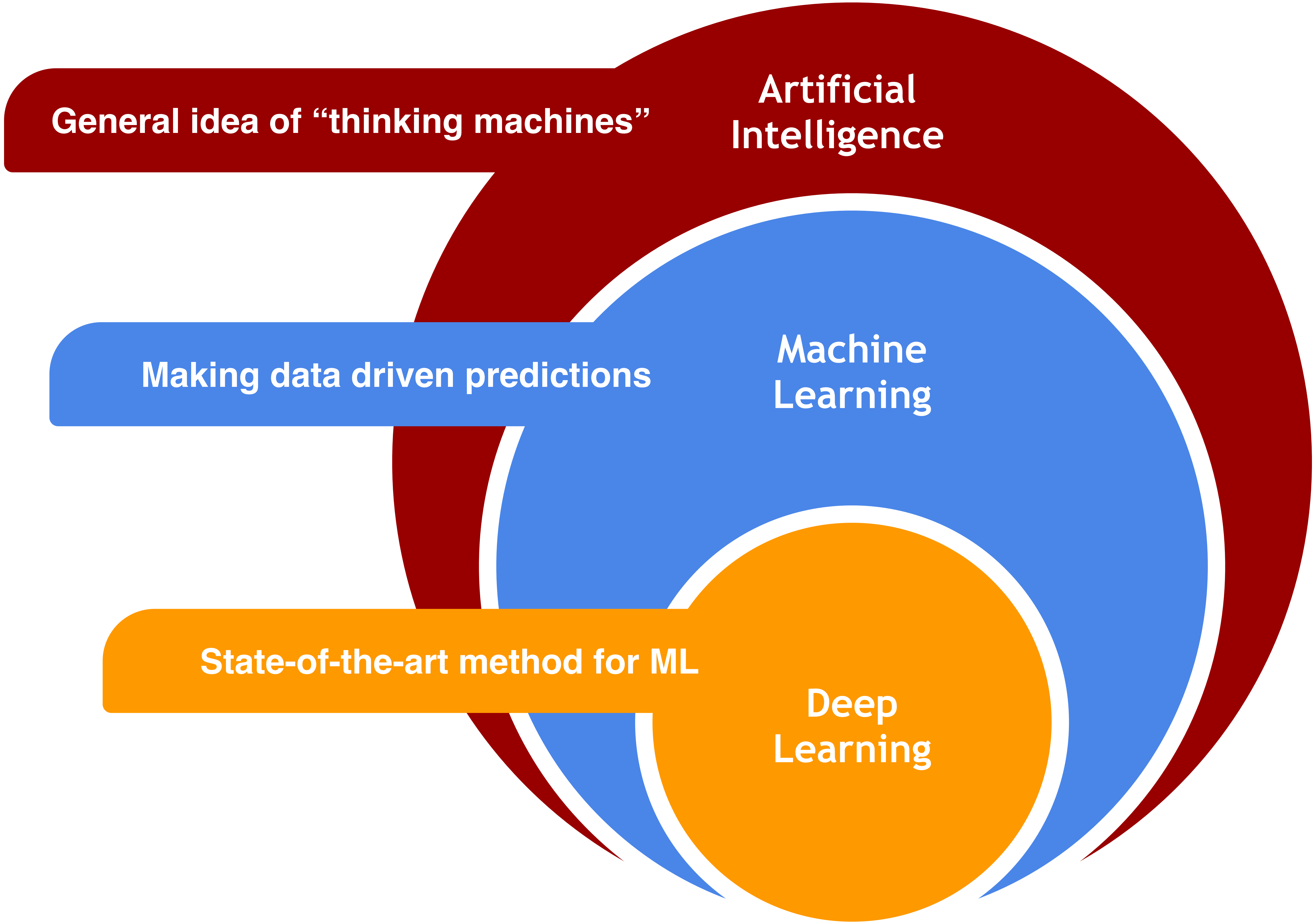
Administrative contact for Gradescope license at University of Glasgow: "We have had great success with Gradescope. It has made it possible for us to provide final examinations in many more circumstances than we would have done in the past. We are using Gradescope for the first time this year for one of our large 2-hour finals and the results are looking good."

Math instructor at University of Ottawa: "Gradescope is one of the best things I have ever discovered on the internet. I was up and running very quickly and had exactly what I needed to grade my big assignments. Gradescope saved me hours of time and the end result was much better than what I would have done by hand."

GPT-3 is a deep learning model for the task of language modeling

GPT-3 is a deep learning model for the task of language modeling

# Deep Learning



General idea of “thinking machines”

Artificial Intelligence

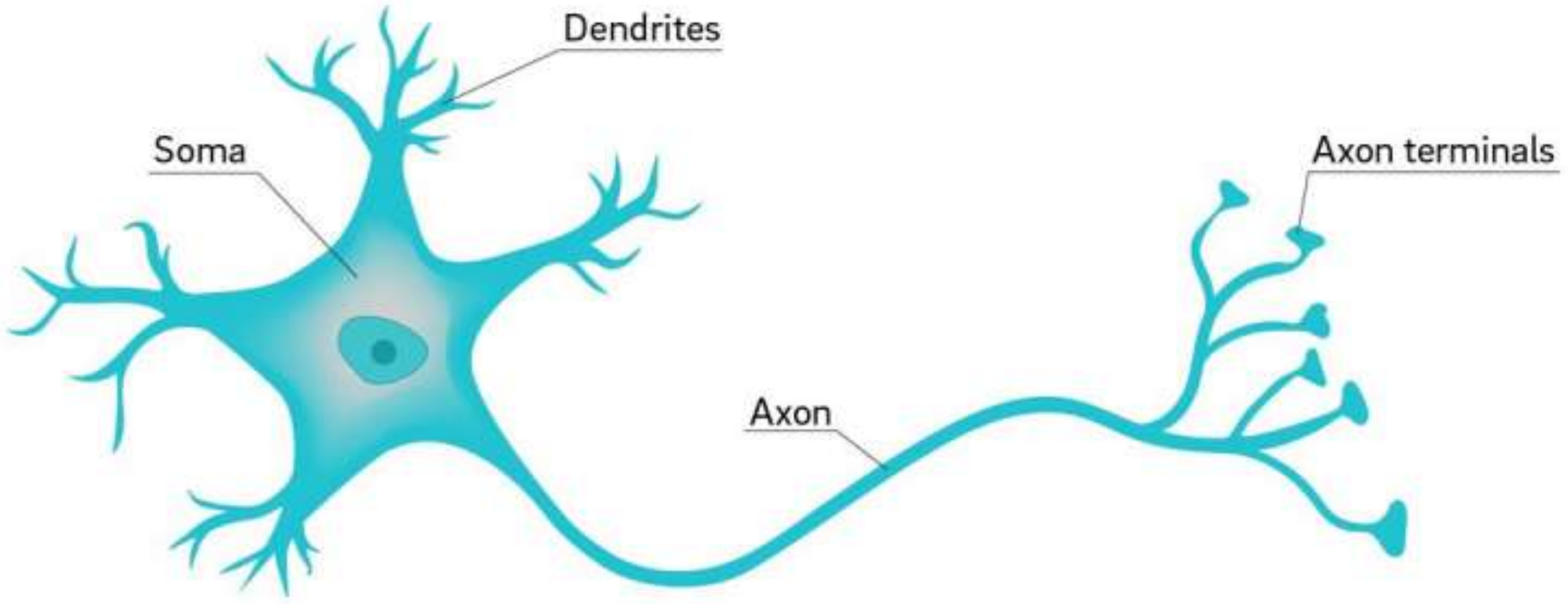
Making data driven predictions

Machine Learning

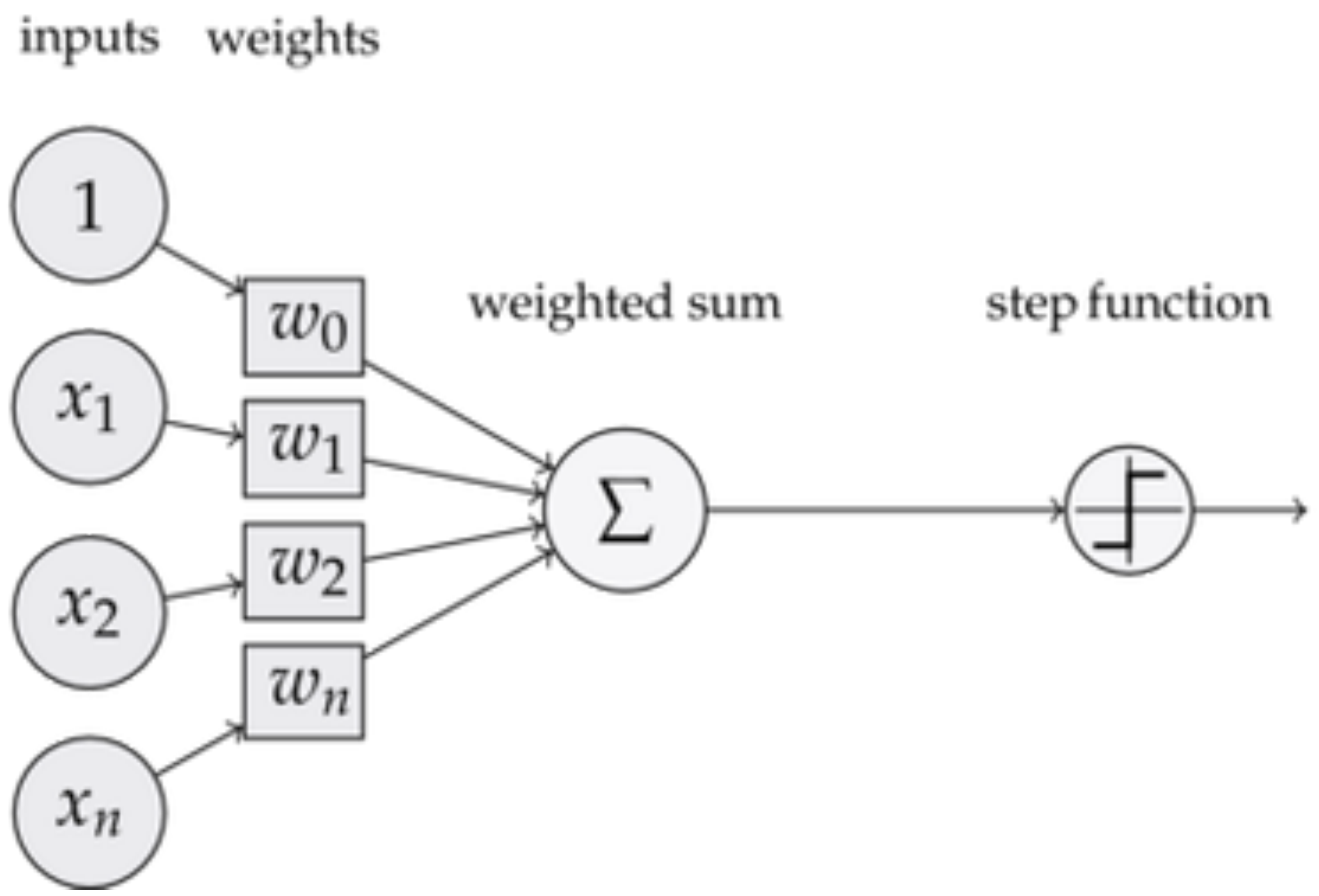
State-of-the-art method for ML

Deep Learning

# aka Neural Networks

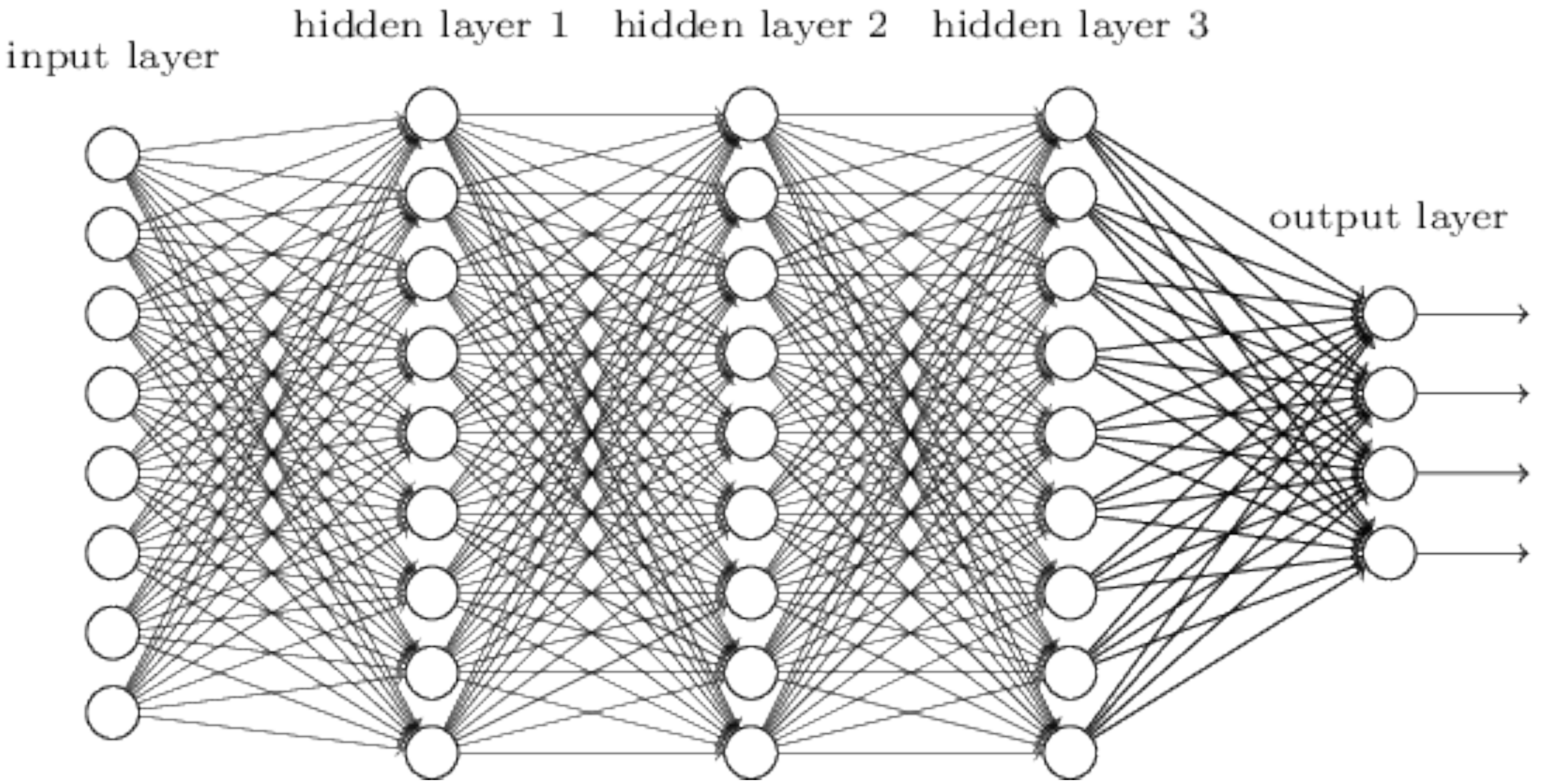


https://medicalxpress.com/news/2018-07-neuron-axons-spindly-theyre-optimizing.html

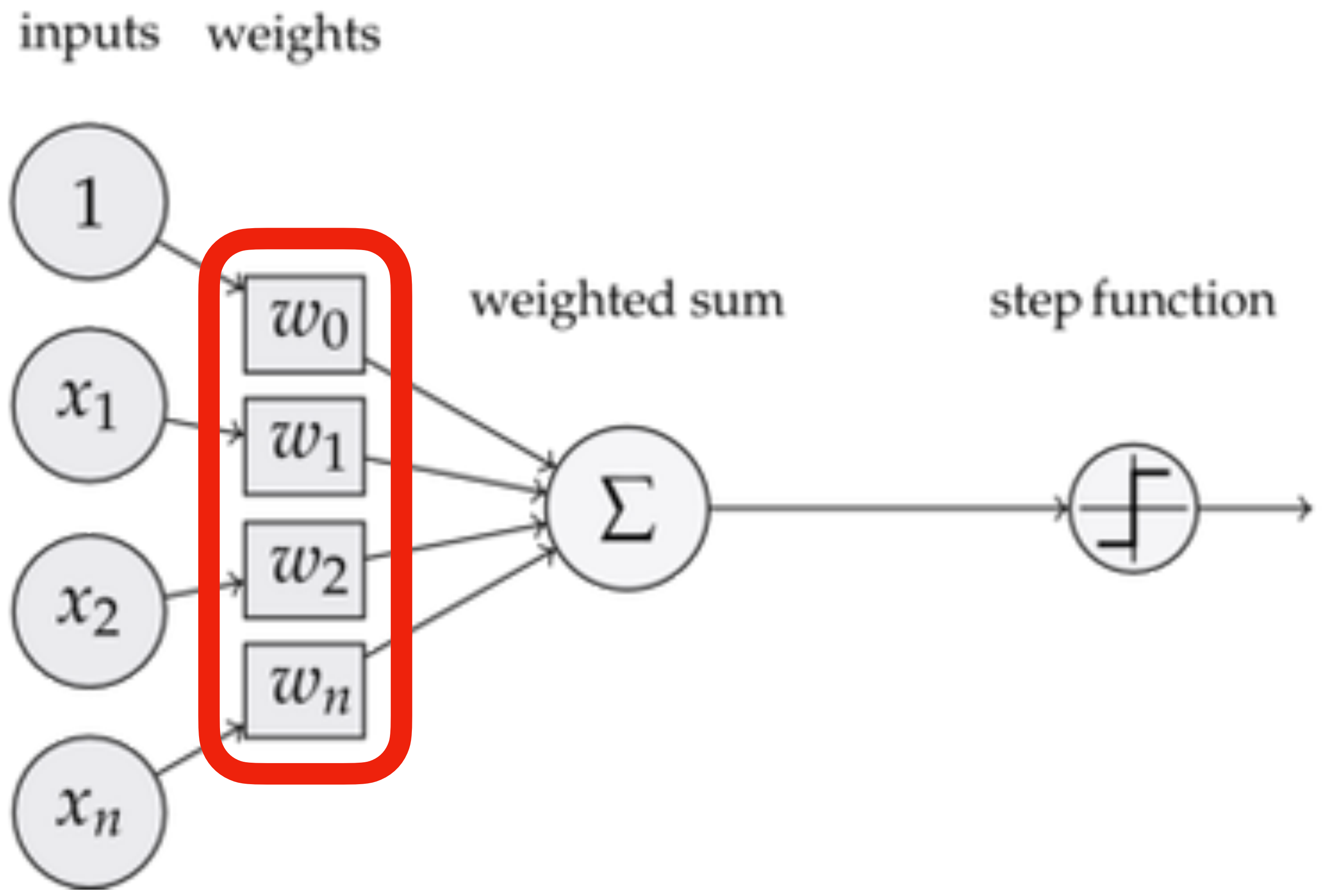


https://www.jessicayung.com/explaining-tensorflow-code-for-a-multilayer-perceptron/

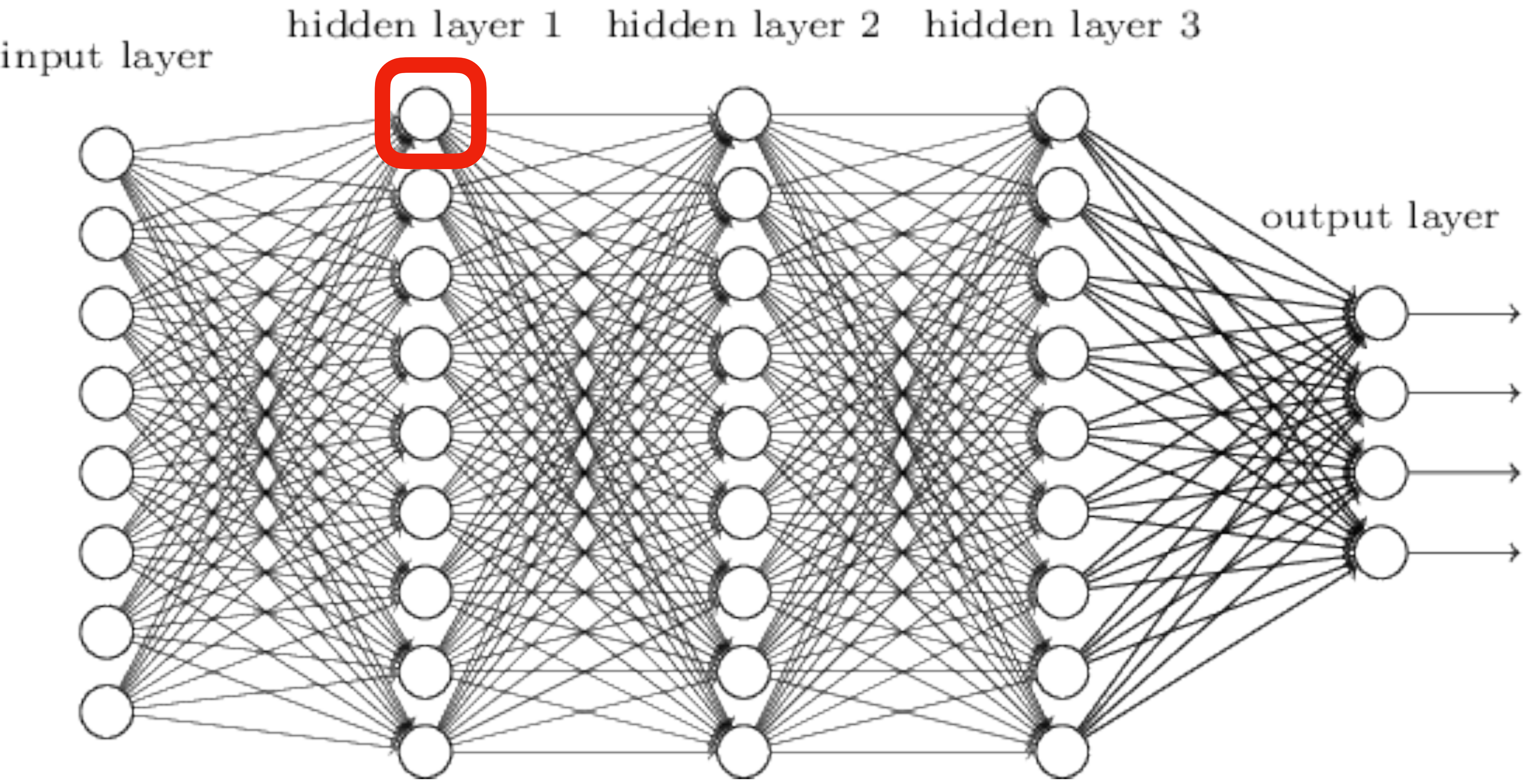
https://www.the-scientist.com/the-nutshell/what-made-human-brains-so-big-36663

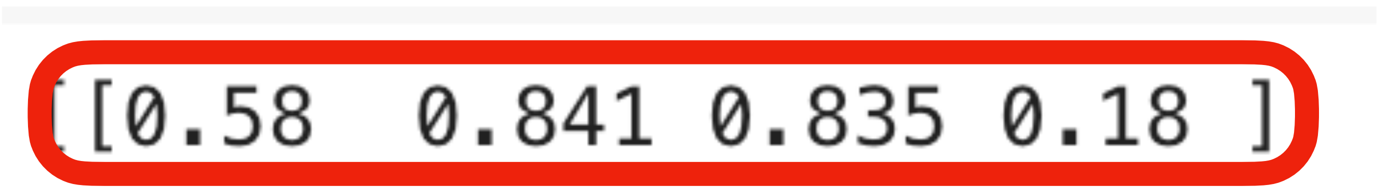


# A "neuron" is just a set of numbers



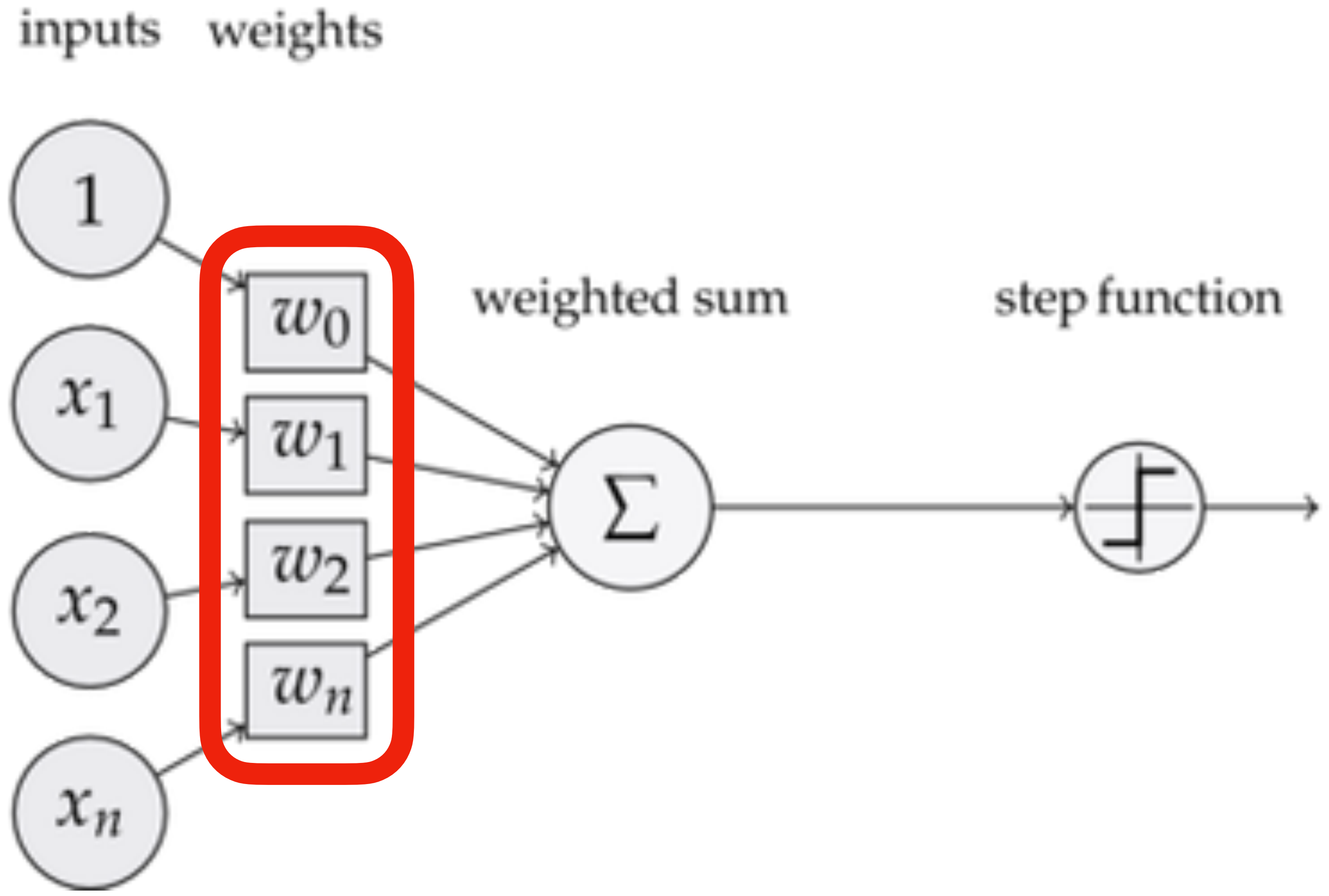
https://www.jessicayung.com/explaining-tensorflow-code-for-a-multilayer-perceptron/



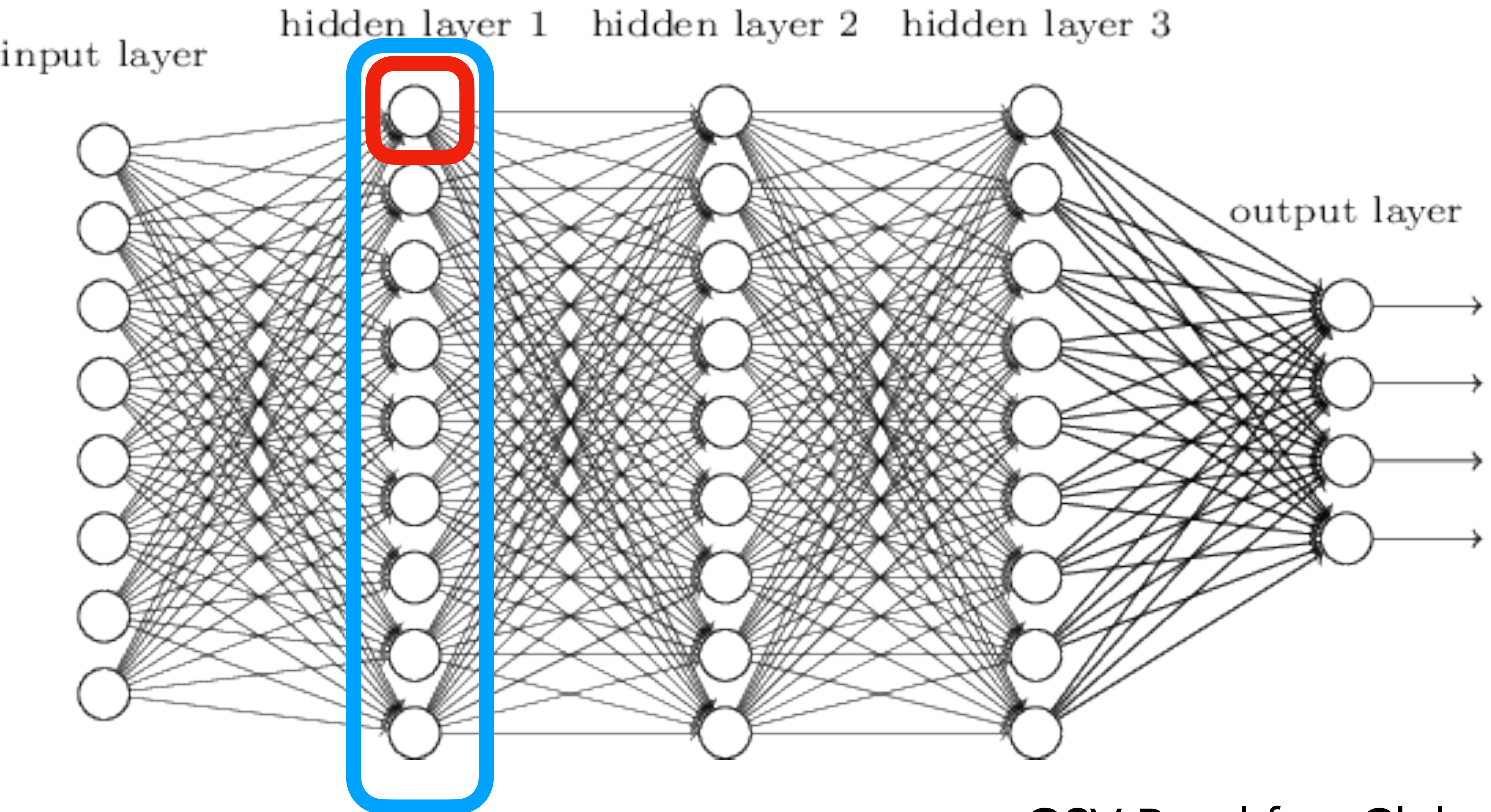


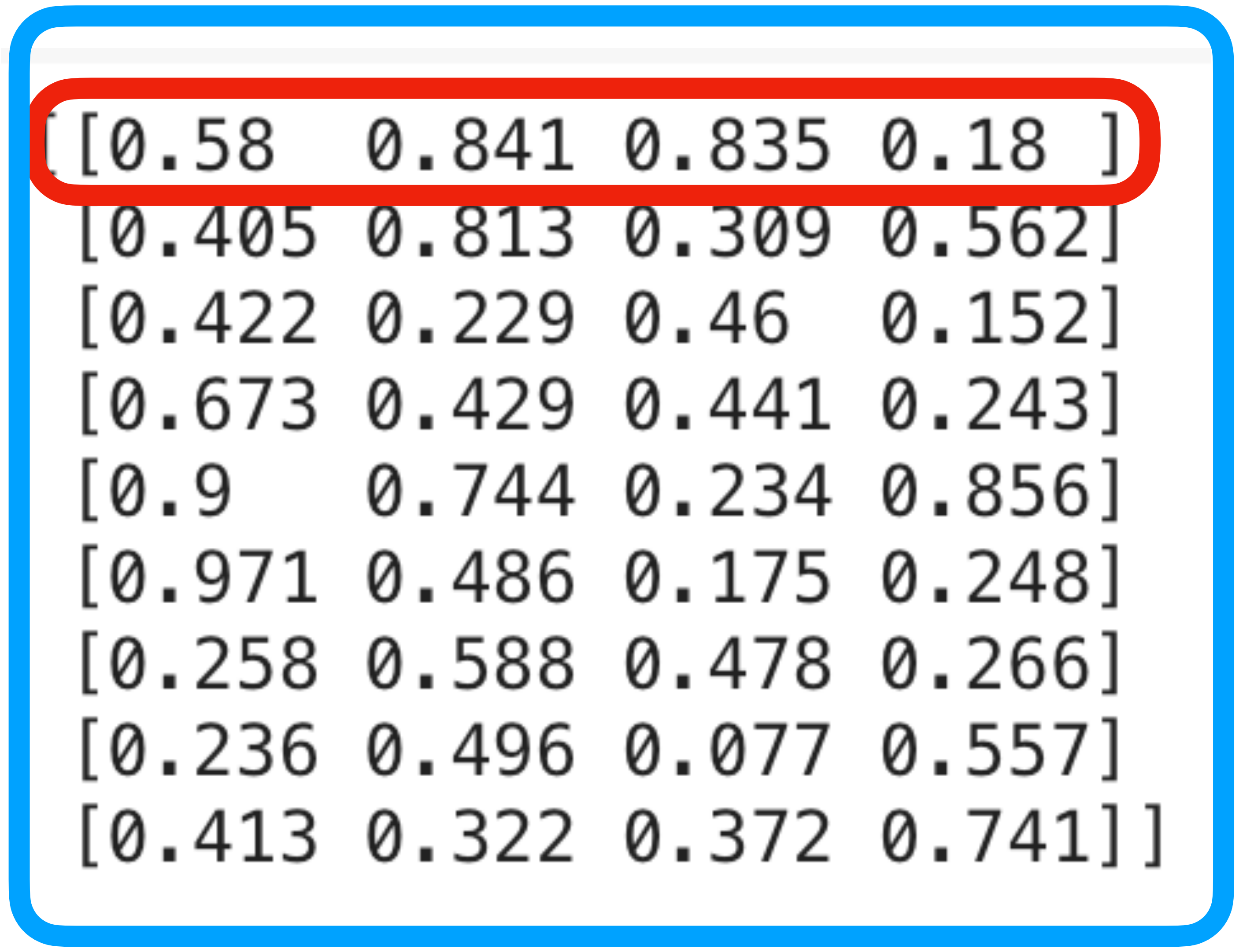
```
[0.58  0.841 0.835 0.18 ]
```

# A "layer" is then just a matrix



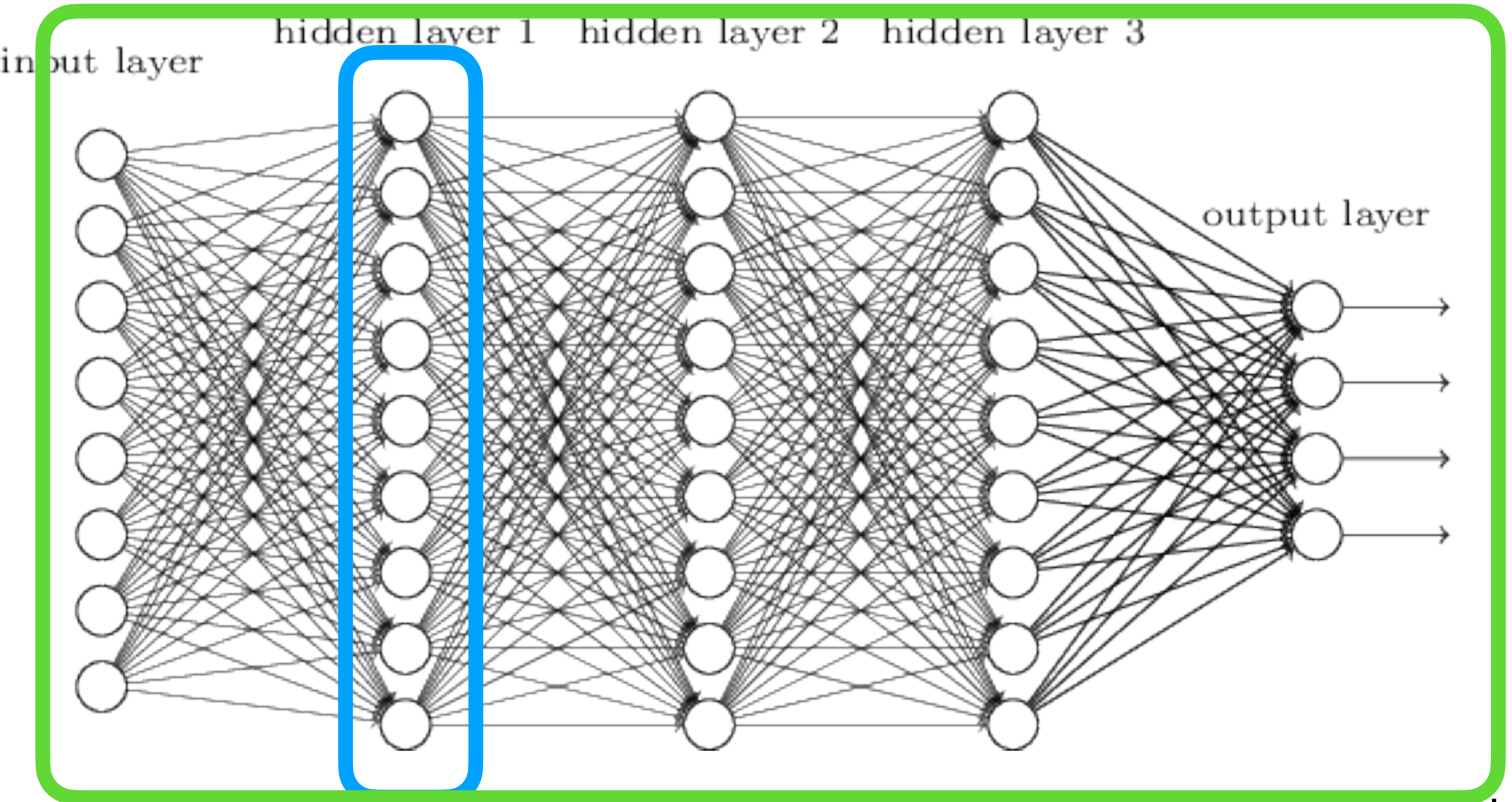
https://www.jessicayung.com/explaining-tensorflow-code-for-a-multilayer-perceptron/

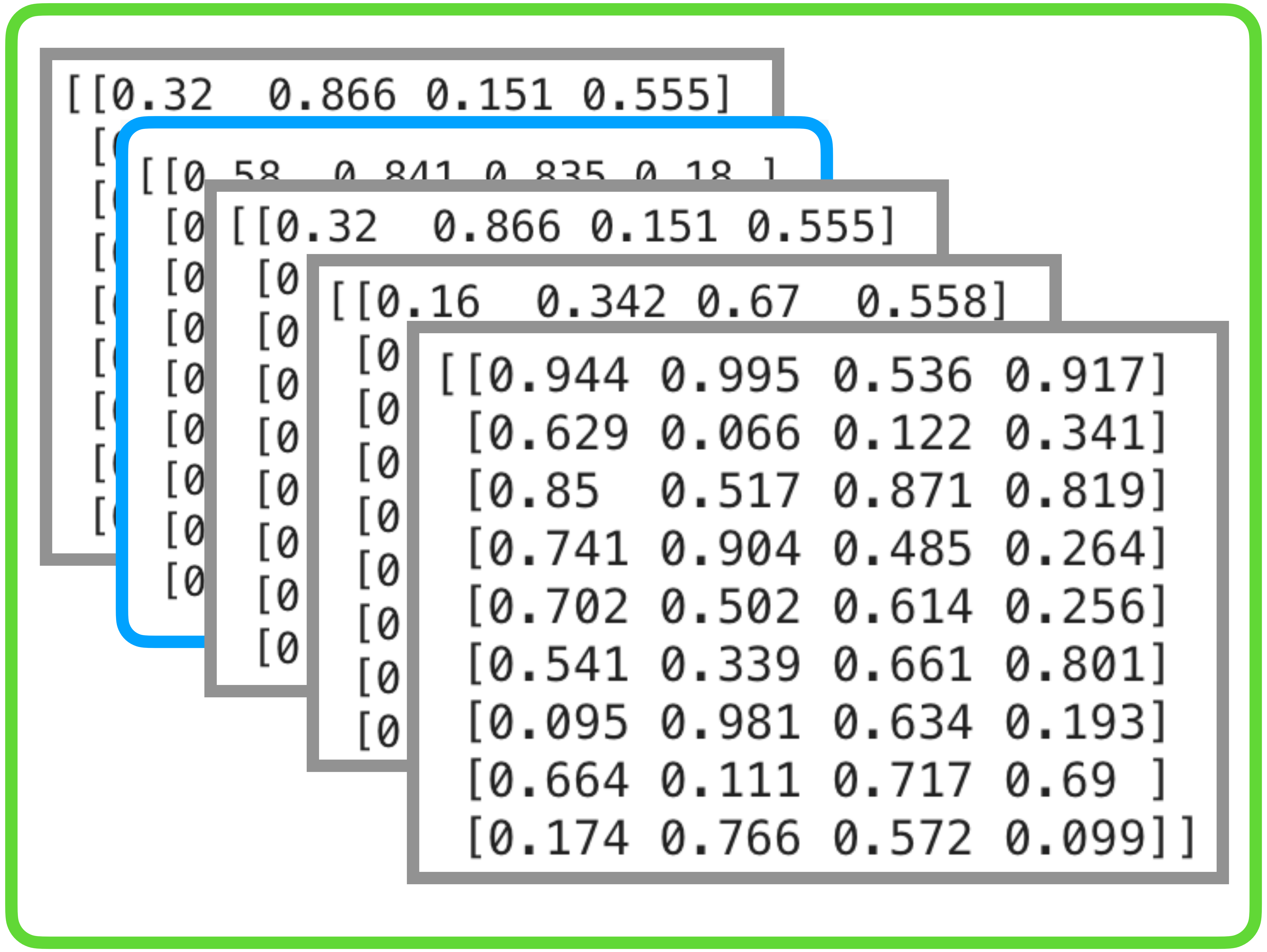




```
[[0.58  0.841 0.835 0.18 ]
 [0.405 0.813 0.309 0.562]
 [0.422 0.229 0.46  0.152]
 [0.673 0.429 0.441 0.243]
 [0.9   0.744 0.234 0.856]
 [0.971 0.486 0.175 0.248]
 [0.258 0.588 0.478 0.266]
 [0.236 0.496 0.077 0.557]
 [0.413 0.322 0.372 0.741]]
```

# And the network is a set of matrices





```
[[0.32  0.866 0.151 0.555]
```

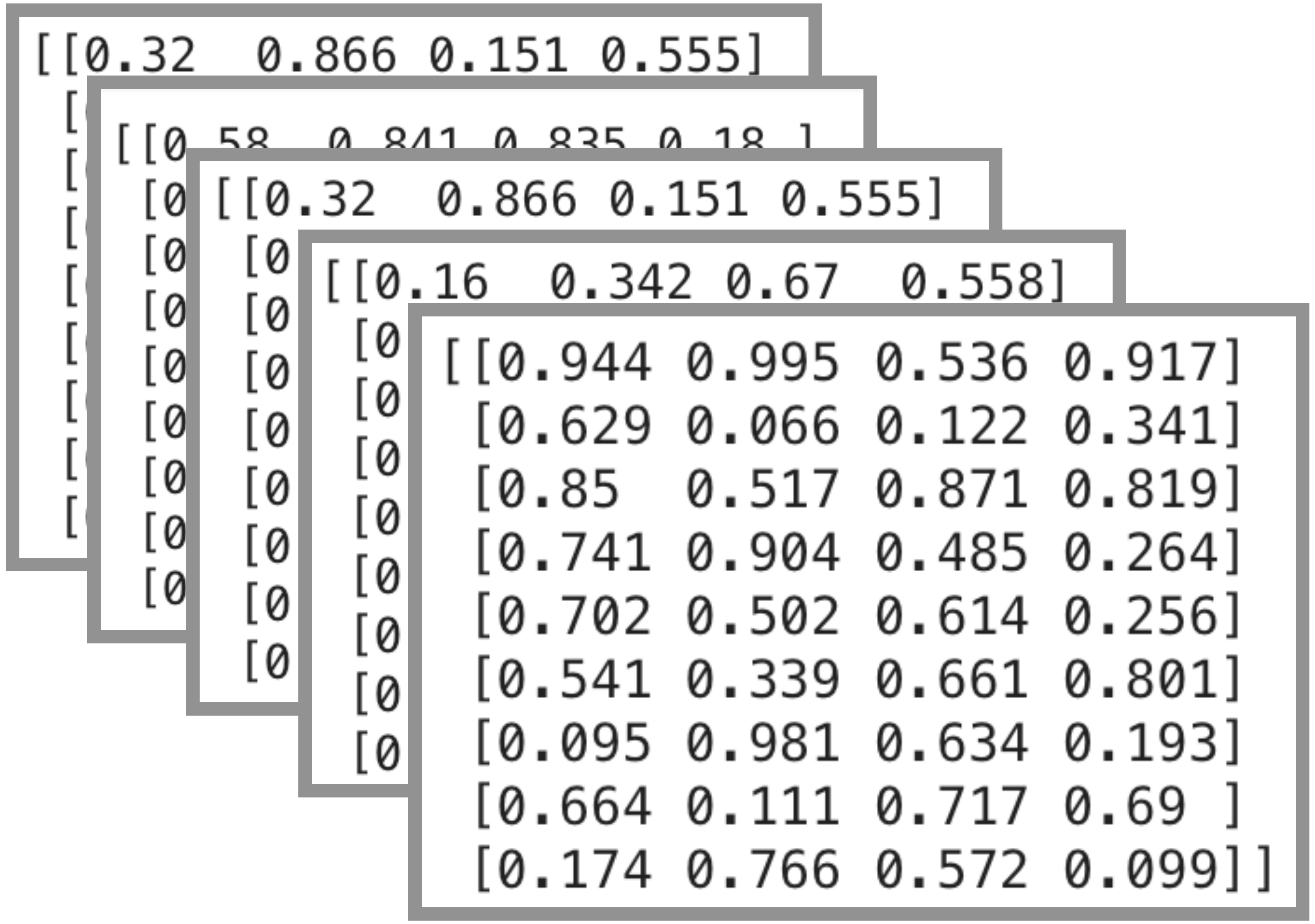
```
[[0.58  0.841 0.835 0.18 ]
```

```
[[0.32  0.866 0.151 0.555]
```

```
[[0.16  0.342 0.67  0.558]
```

```
[[0.944 0.995 0.536 0.917]
 [0.629 0.066 0.122 0.341]
 [0.85  0.517 0.871 0.819]
 [0.741 0.904 0.485 0.264]
 [0.702 0.502 0.614 0.256]
 [0.541 0.339 0.661 0.801]
 [0.095 0.981 0.634 0.193]
 [0.664 0.111 0.717 0.69 ]
 [0.174 0.766 0.572 0.099]]
```

# We call these numbers "parameters"



```
[[0.32  0.866 0.151 0.555]
```

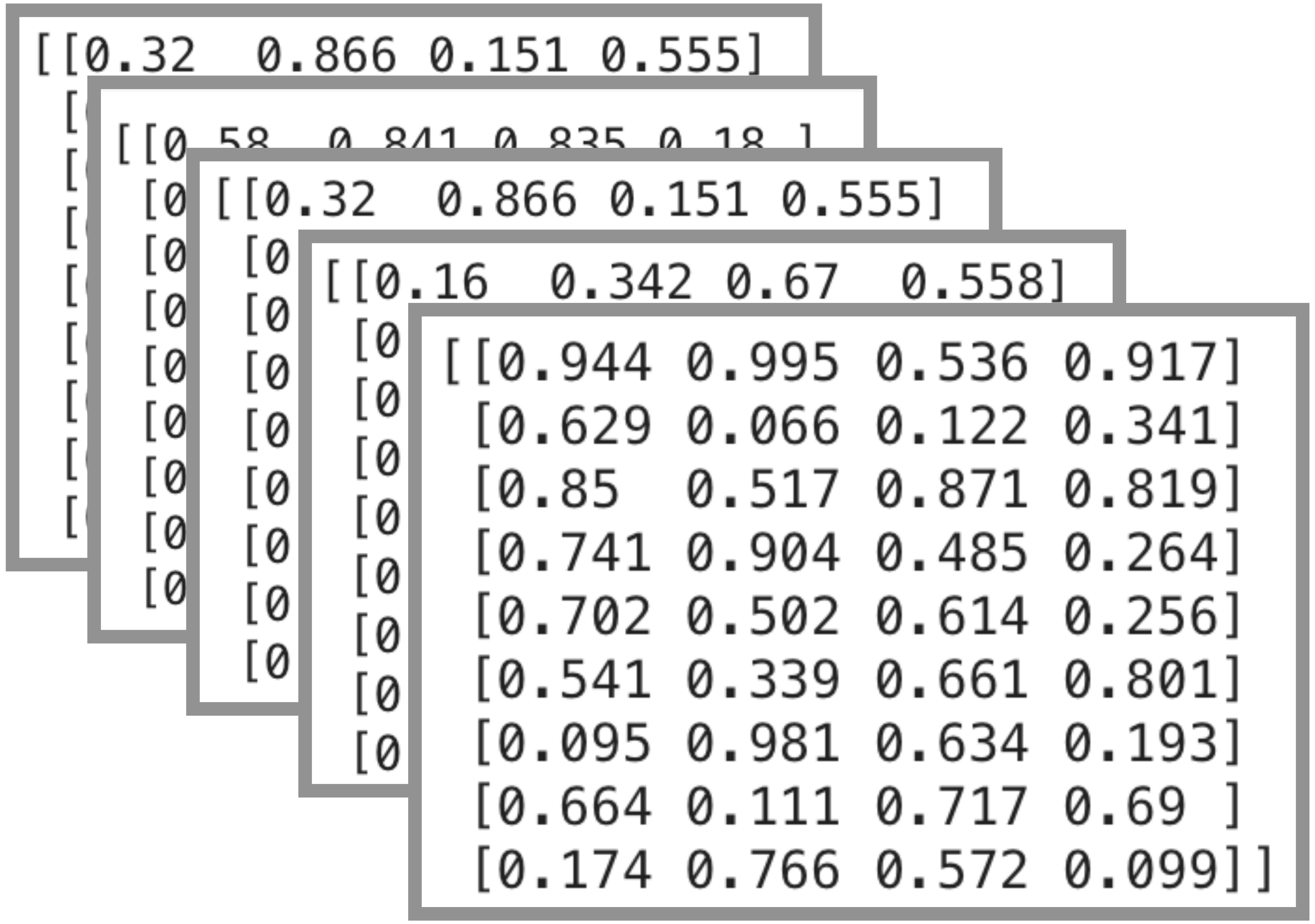
```
[[0.58  0.841 0.835 0.18 ]
```

```
[[0.32  0.866 0.151 0.555]
```

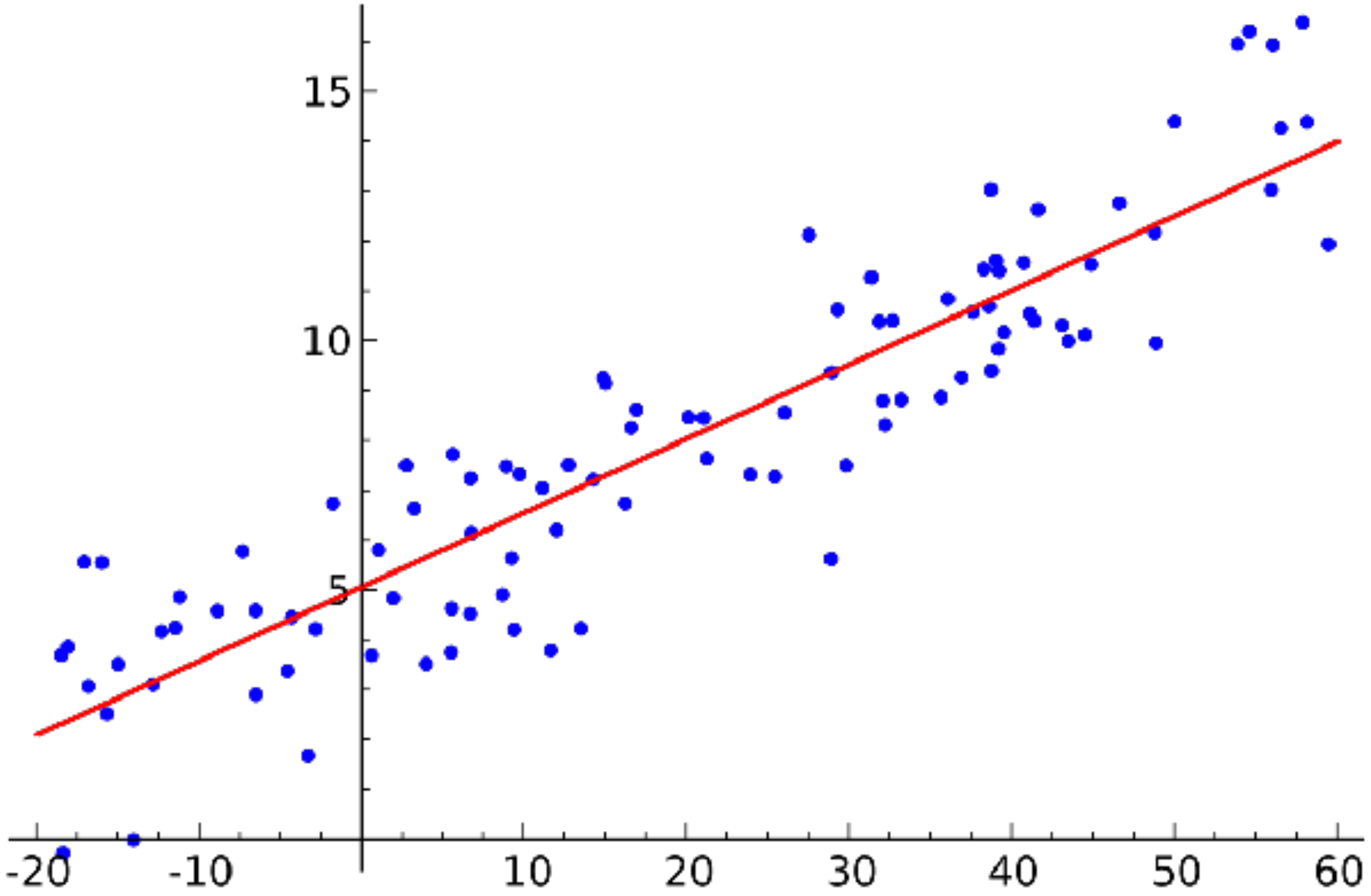
```
[[0.16  0.342 0.67  0.558]
```

```
[[0.944 0.995 0.536 0.917]
 [0.629 0.066 0.122 0.341]
 [0.85  0.517 0.871 0.819]
 [0.741 0.904 0.485 0.264]
 [0.702 0.502 0.614 0.256]
 [0.541 0.339 0.661 0.801]
 [0.095 0.981 0.634 0.193]
 [0.664 0.111 0.717 0.69 ]
 [0.174 0.766 0.572 0.099]]
```

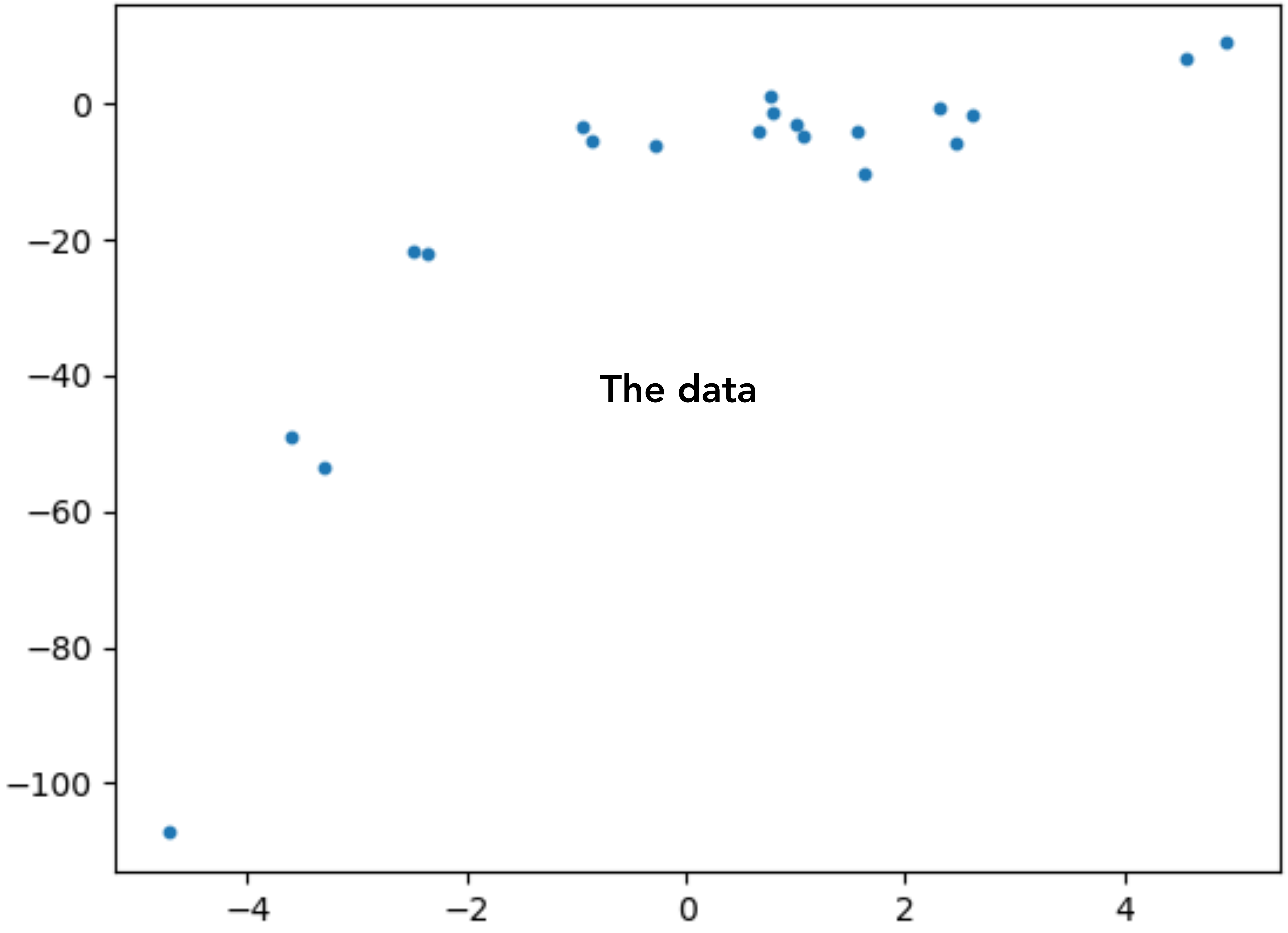
# To "train" a neural network is to find parameters that minimize error on data


[[0.32  0.866 0.151 0.555]
[[0.58  0.841 0.835 0.18 ]
[[0.32  0.866 0.151 0.555]
[[0.16  0.342 0.67  0.558]

```
[[0.944 0.995 0.536 0.917]
 [0.629 0.066 0.122 0.341]
 [0.85  0.517 0.871 0.819]
 [0.741 0.904 0.485 0.264]
 [0.702 0.502 0.614 0.256]
 [0.541 0.339 0.661 0.801]
 [0.095 0.981 0.634 0.193]
 [0.664 0.111 0.717 0.69 ]
 [0.174 0.766 0.572 0.099]]
```

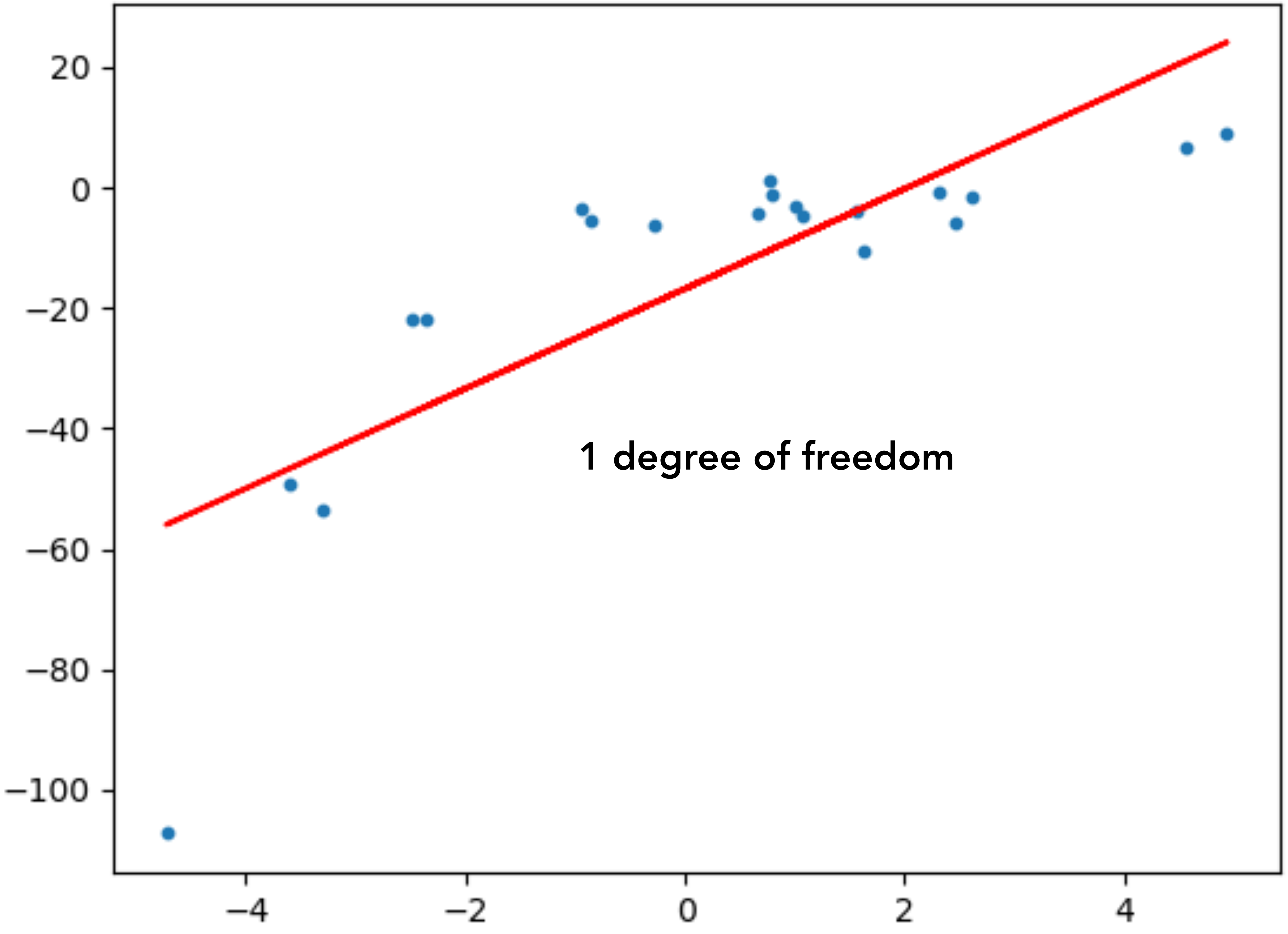


# More parameters -> Model can handle more complexity



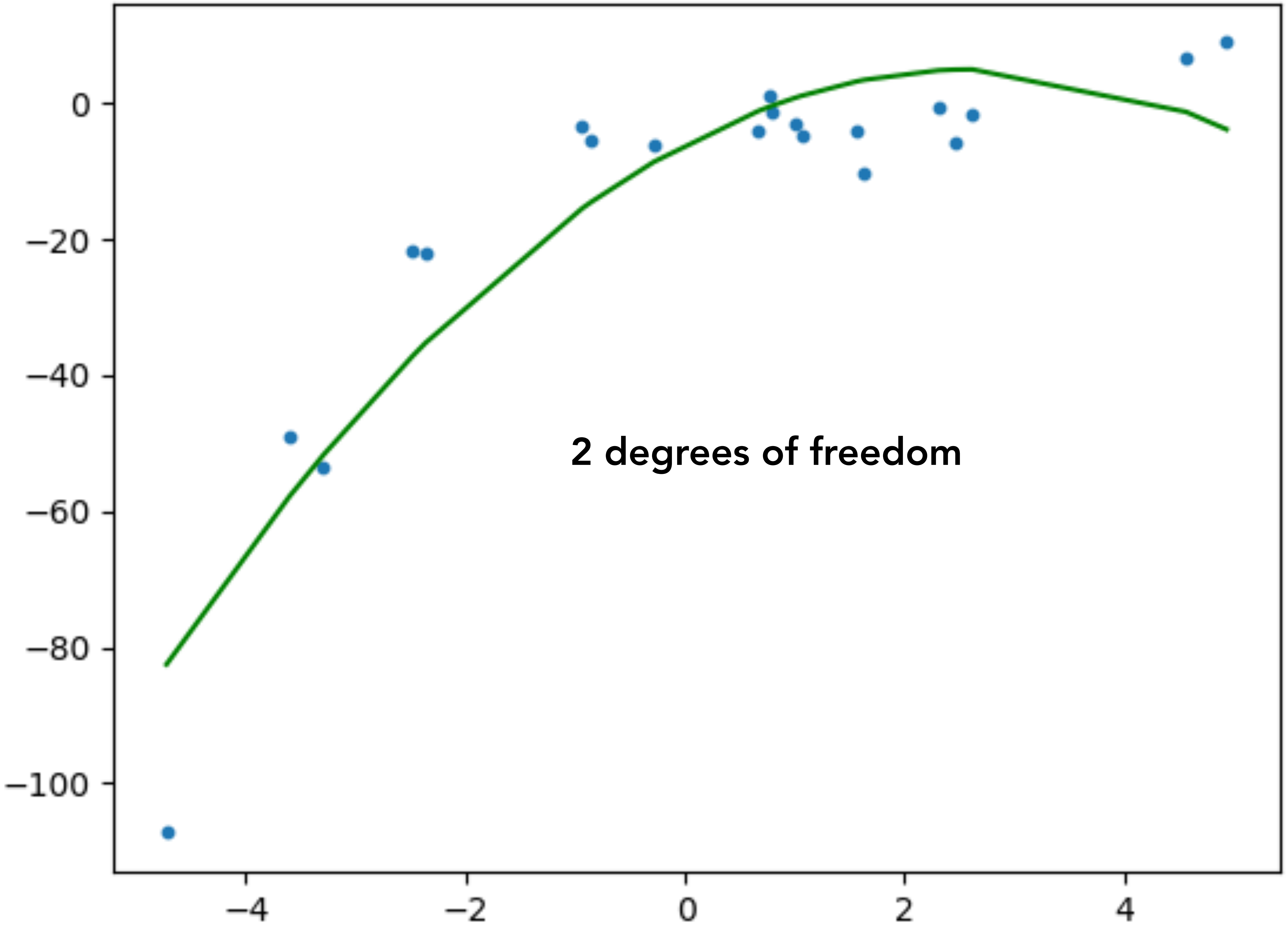
*https://towardsdatascience.com/polynomial-regression-bbe8b9d97491*

# More parameters -> Model can handle more complexity



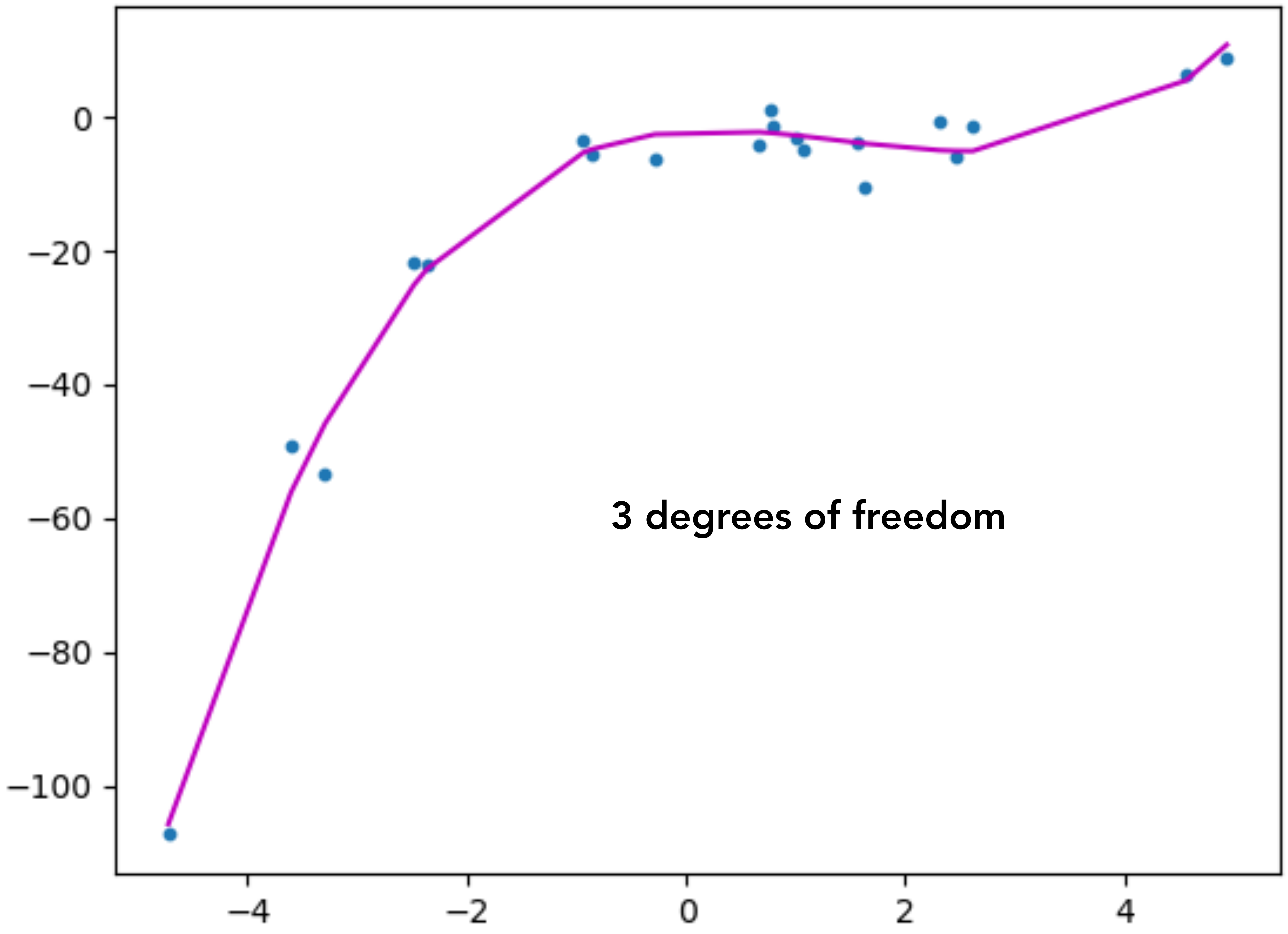
*https://towardsdatascience.com/polynomial-regression-bbe8b9d97491*

# More parameters -> Model can handle more complexity



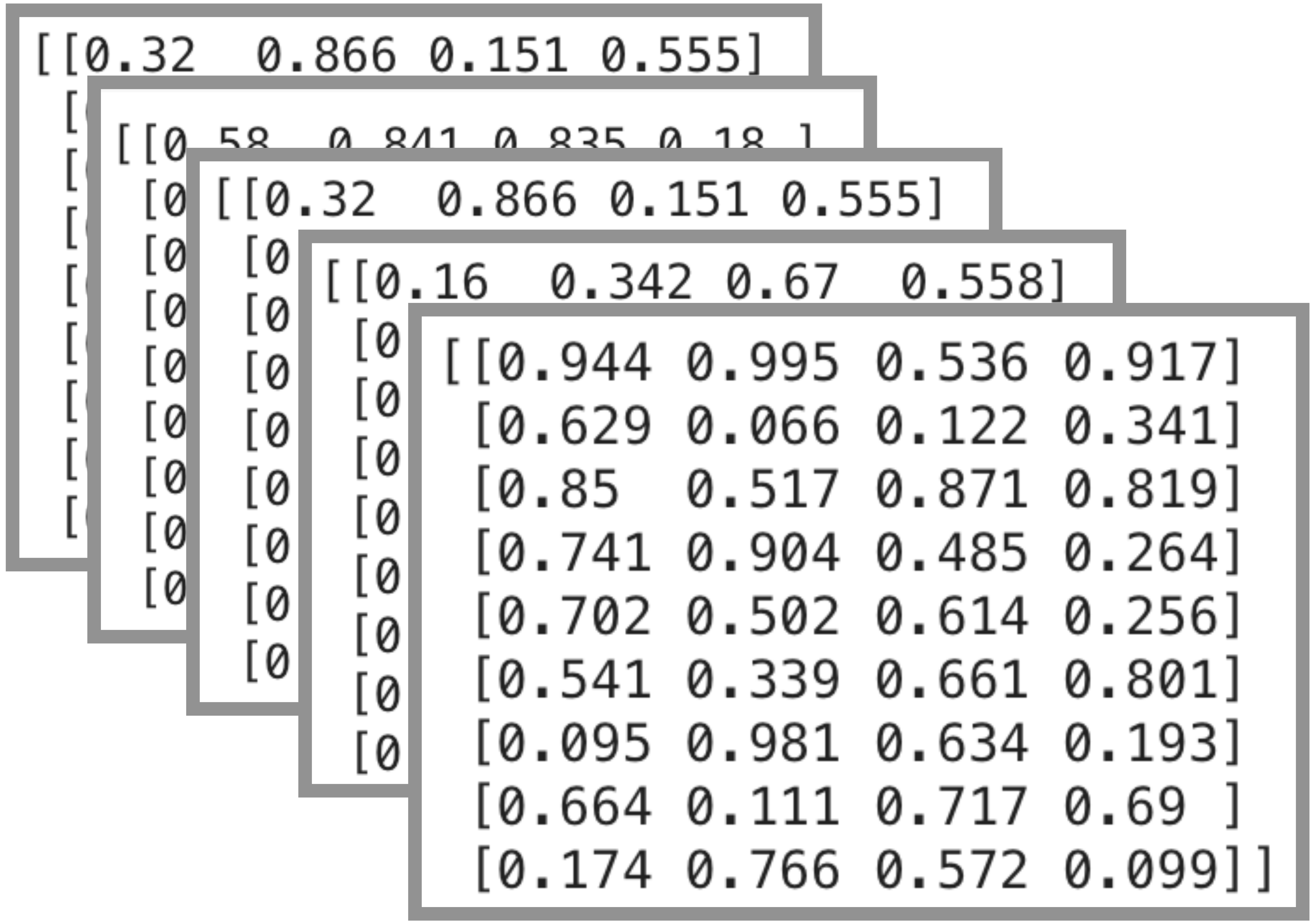
*https://towardsdatascience.com/polynomial-regression-bbe8b9d97491*

# More parameters -> Model can handle more complexity

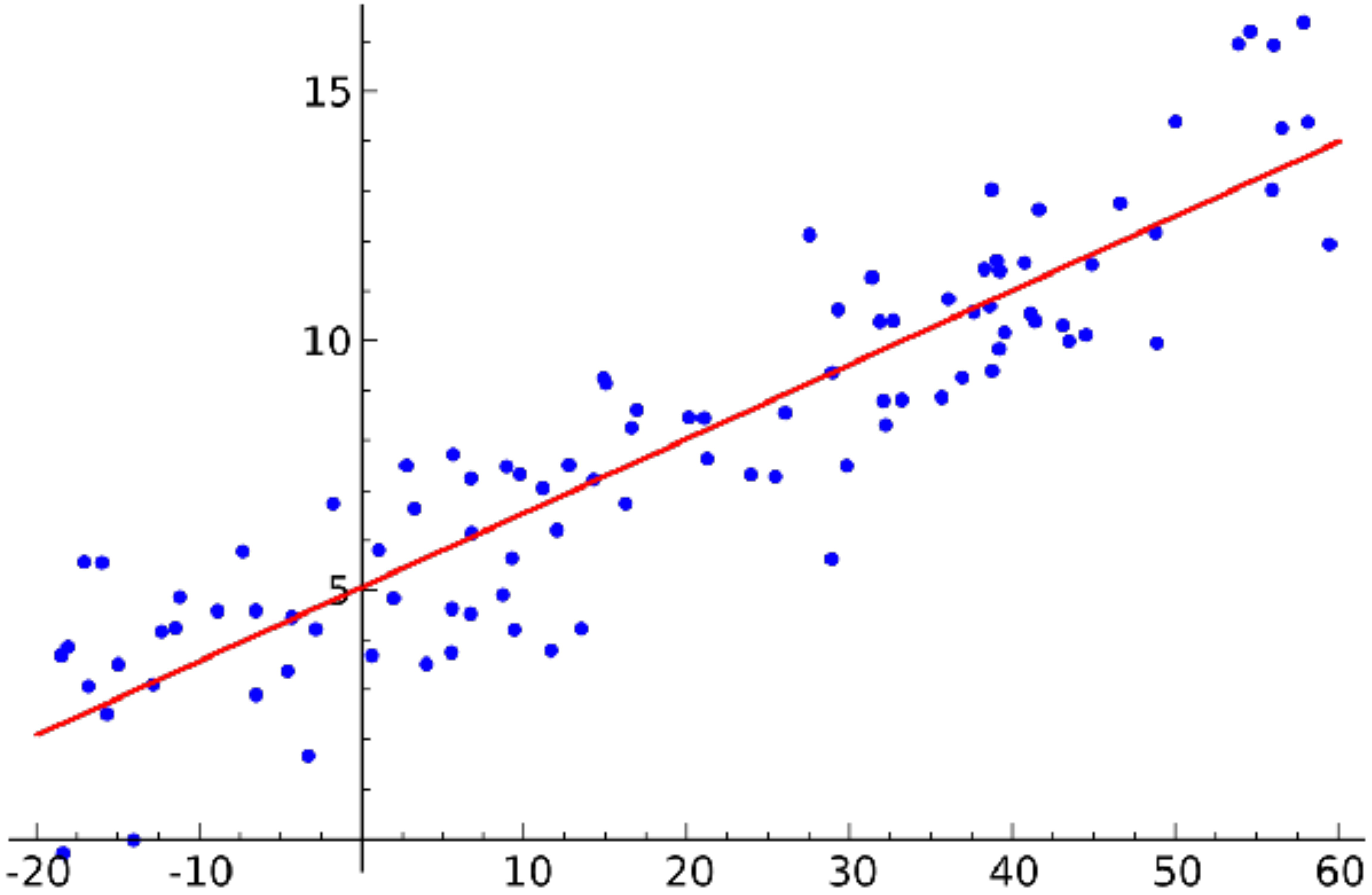


*https://towardsdatascience.com/polynomial-regression-bbe8b9d97491*

# "Train" a neural network = find parameters that minimize error on training data



```
[[0.944 0.995 0.536 0.917]
 [0.629 0.066 0.122 0.341]
 [0.85  0.517 0.871 0.819]
 [0.741 0.904 0.485 0.264]
 [0.702 0.502 0.614 0.256]
 [0.541 0.339 0.661 0.801]
 [0.095 0.981 0.634 0.193]
 [0.664 0.111 0.717 0.69 ]
 [0.174 0.766 0.572 0.099]]
```



GPT-3 is a deep learning model for the task of language modeling

# GPT-3 is a deep learning model for the task of language modeling

# Language Modeling

Likely next words

We are gathered here

to

today

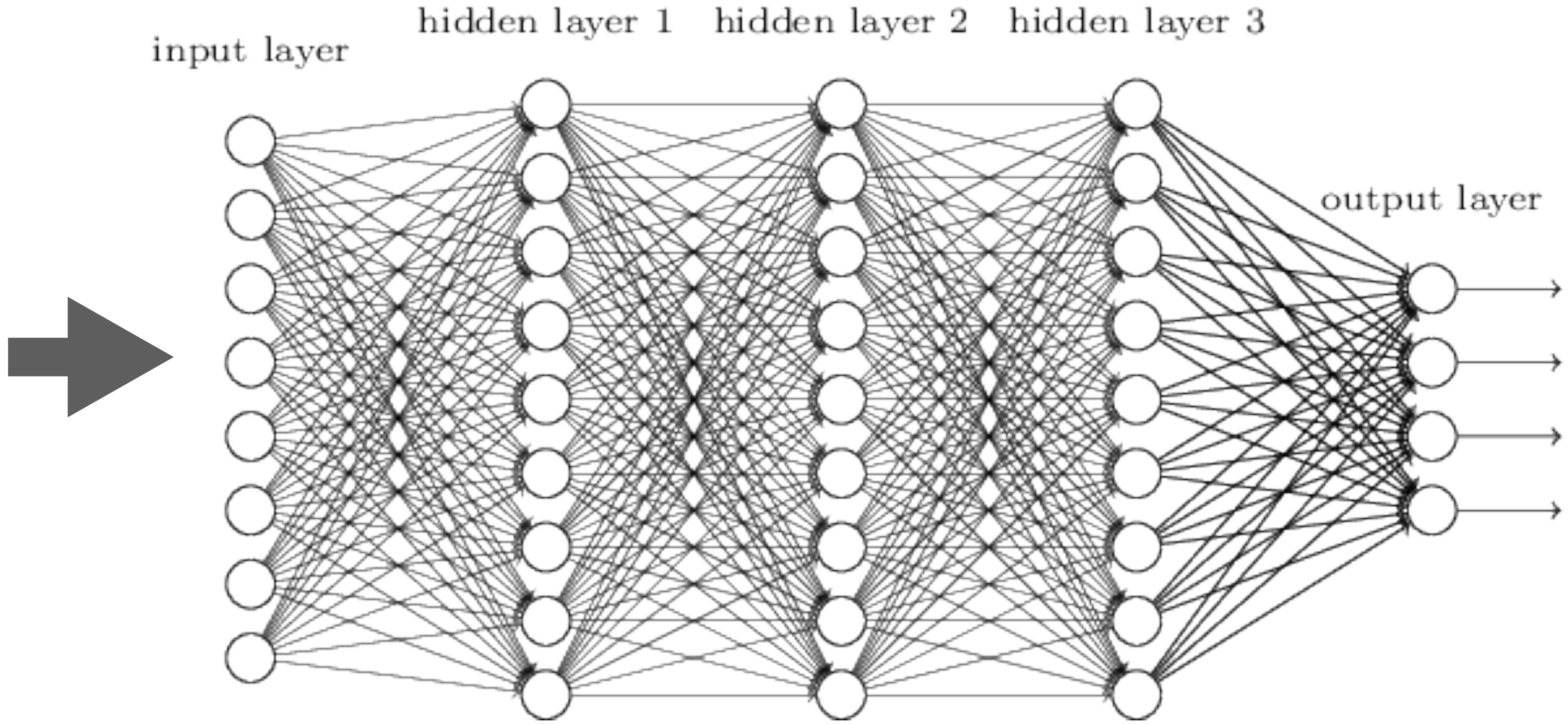
on

in

Known text

# Neural Network for Language Modeling

**We are gathered here**



to

**today**

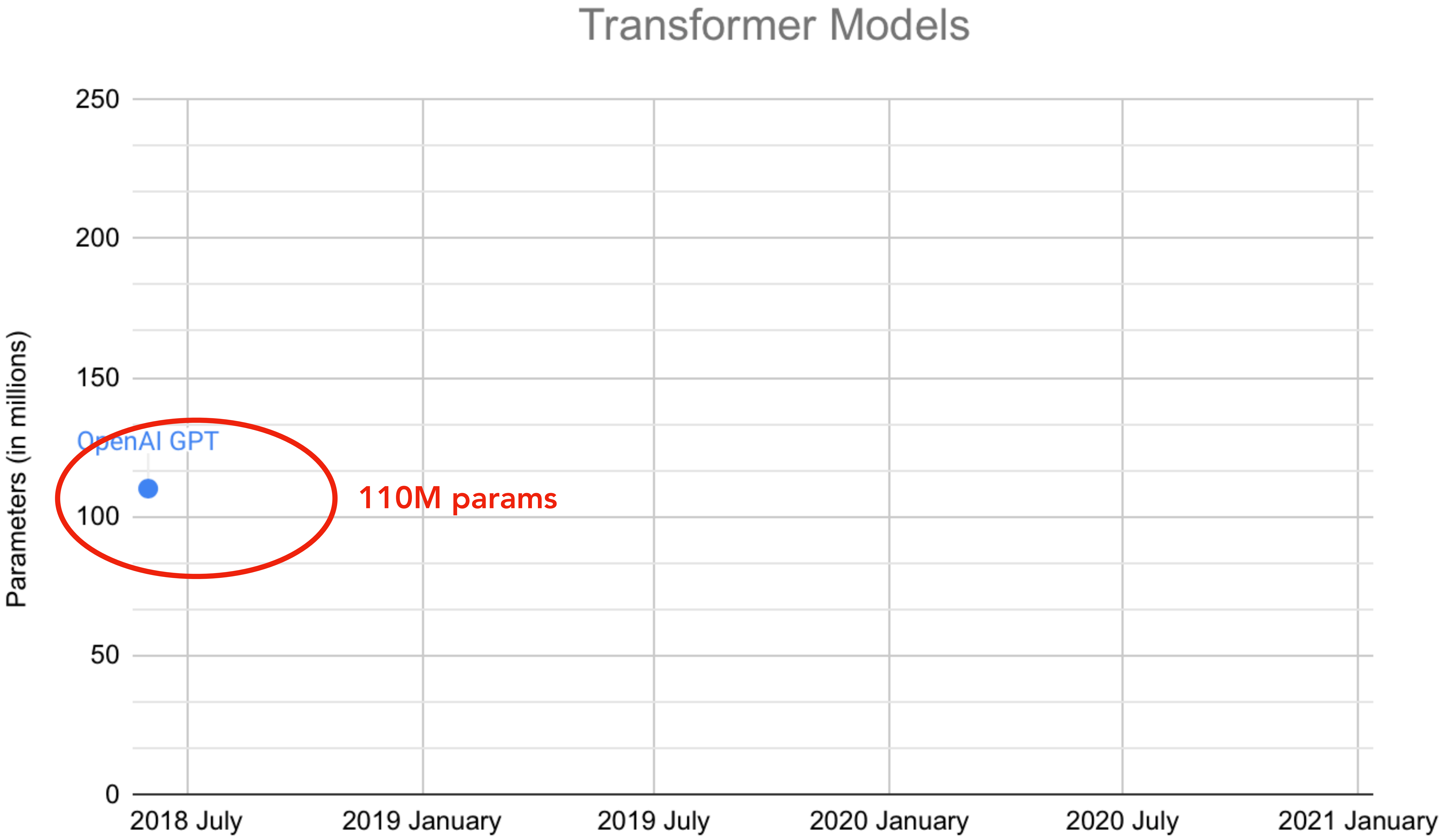
on

in

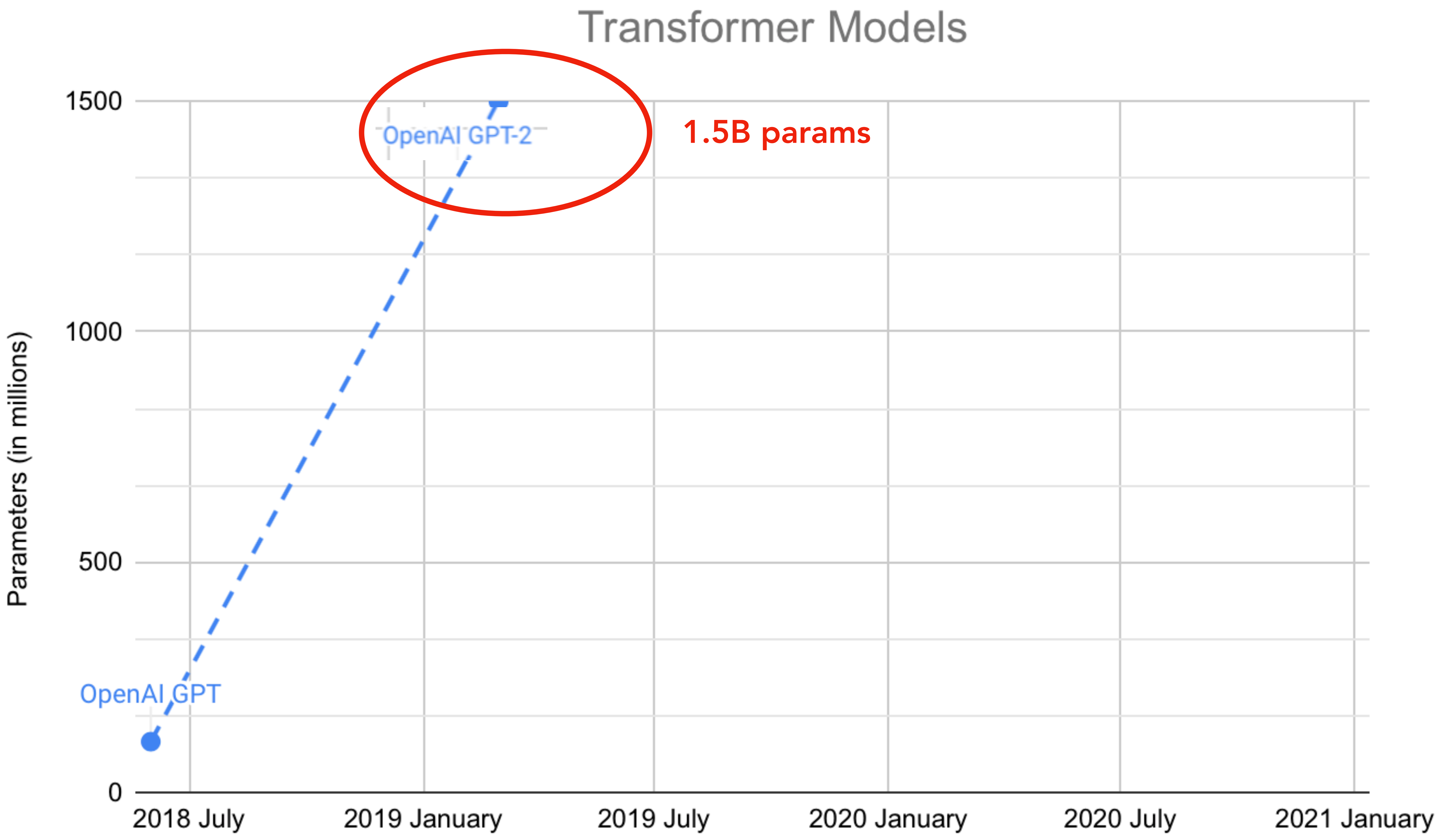
# Train on "the Internet"

© Courtesy of Michael Mandiberg / Denny Gallery

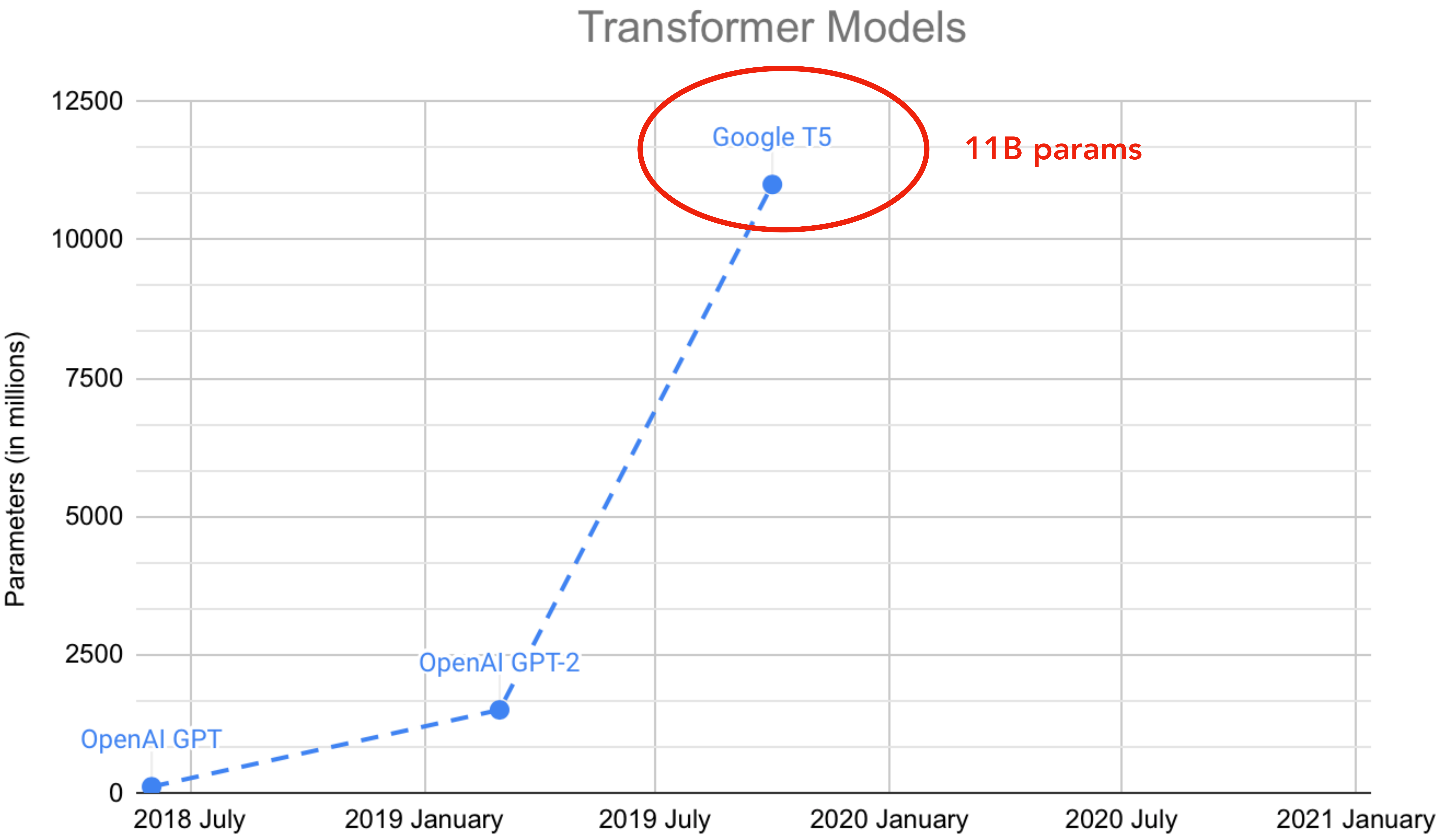
# And use an ungodly number of parameters



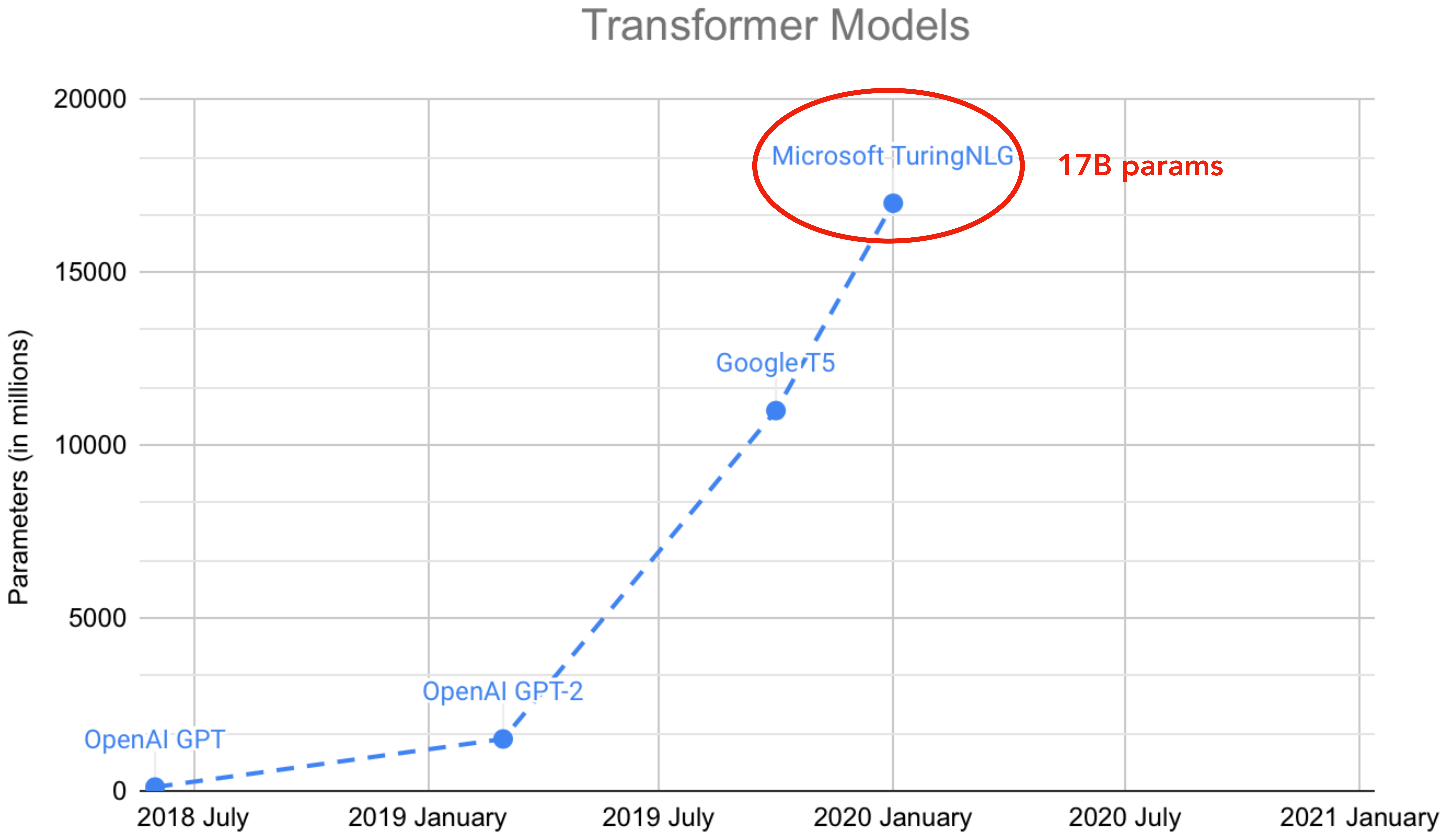
# And use an ungodly number of parameters



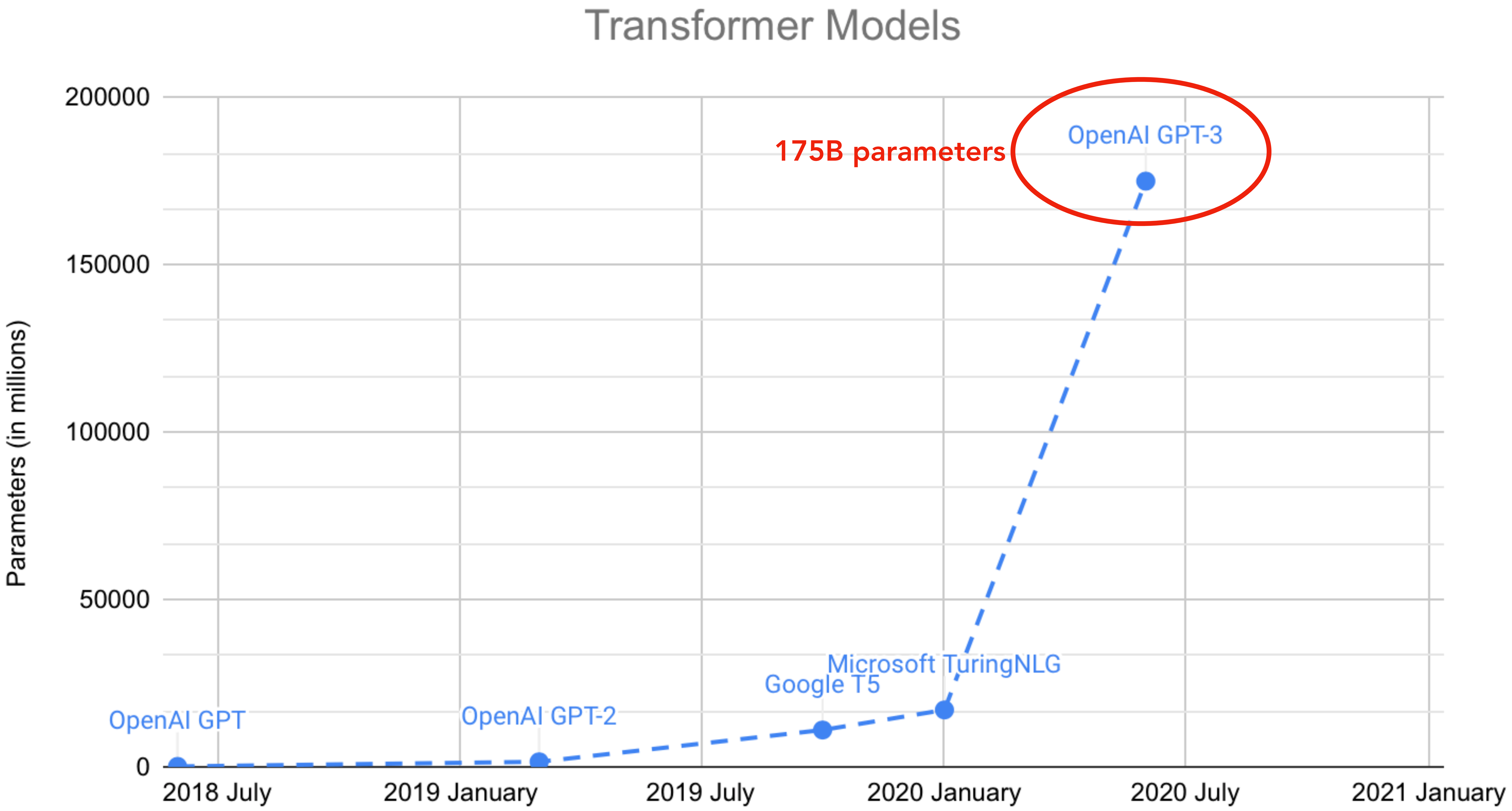
# And use an ungodly number of parameters



# And use an ungodly number of parameters



# And use an ungodly number of parameters

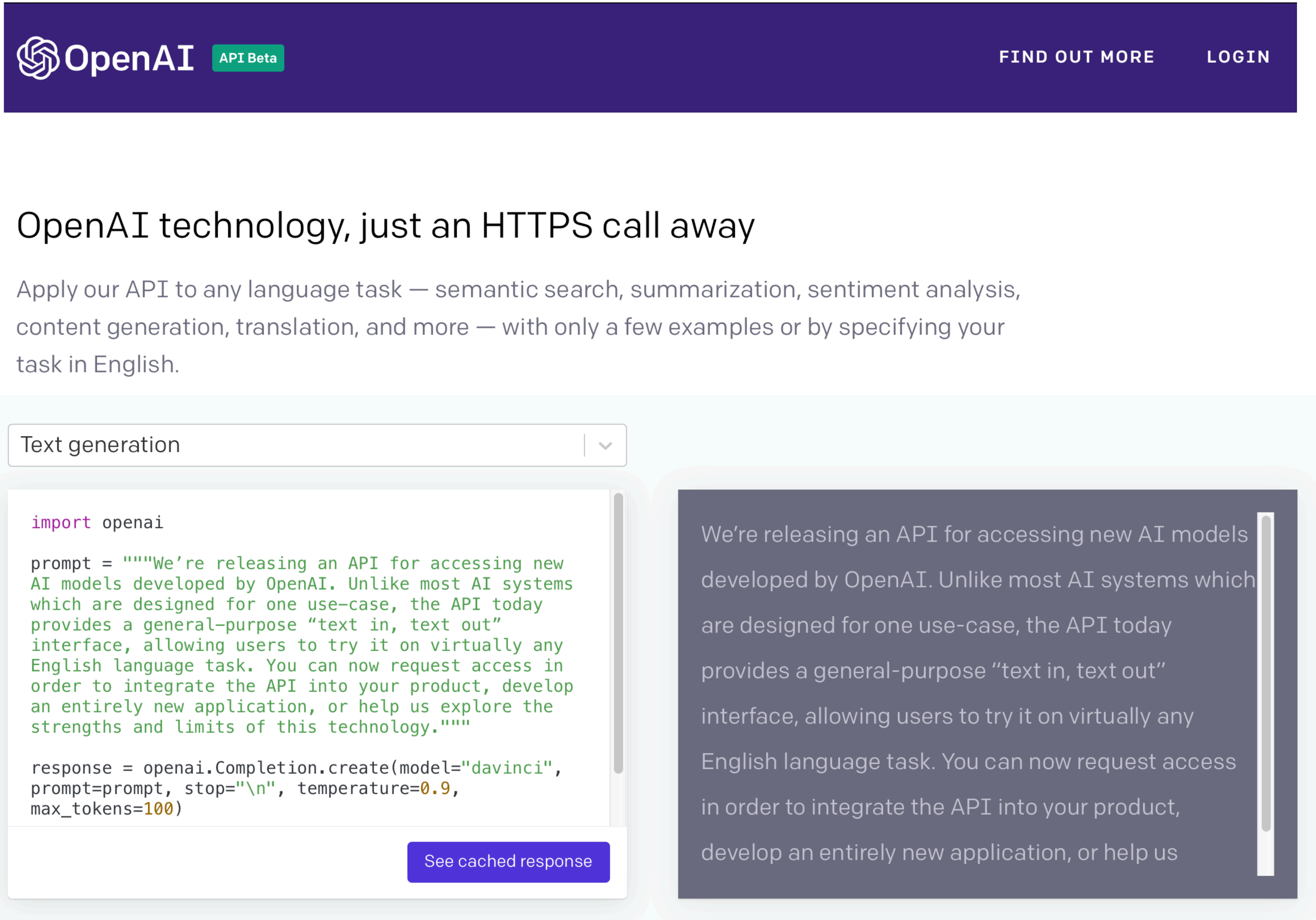


# GPT-3 is a deep learning model for the task of language modeling

**(Cost of development is in the tens of millions)**

OpenAI did not release model parameters, out of societal concern

# Instead, provided a GPT-3 API that they can monitor



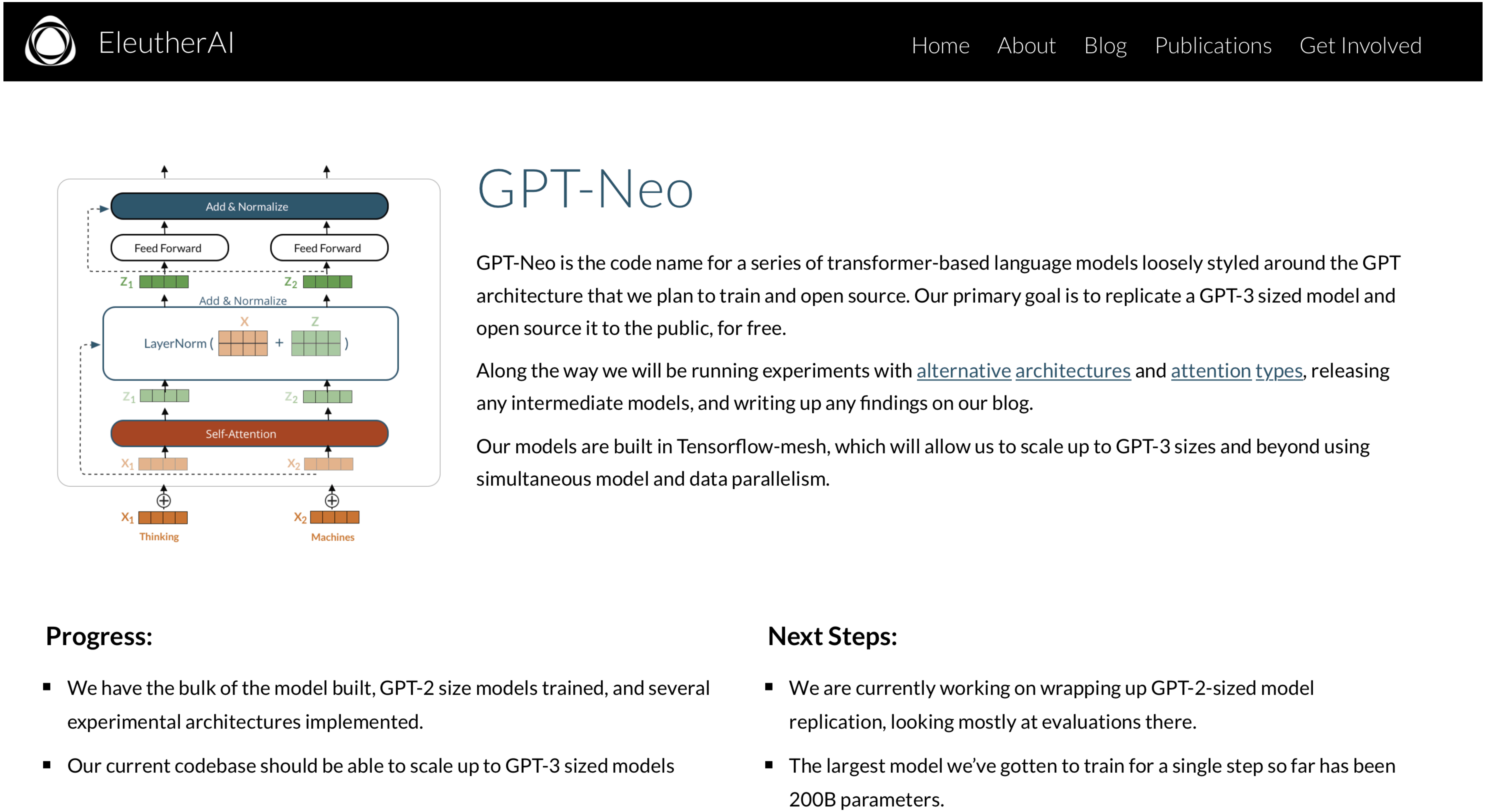
## OpenAI technology, just an HTTPS call away

Apply our API to any language task — semantic search, summarization, sentiment analysis, content generation, translation, and more — with only a few examples or by specifying your task in English.

```
import openai

prompt = """We're releasing an API for accessing new
AI models developed by OpenAI. Unlike most AI systems
which are designed for one use-case, the API today
provides a general-purpose “text in, text out”
interface, allowing users to try it on virtually any
English language task. You can now request access in
order to integrate the API into your product, develop
an entirely new application, or help us explore the
strengths and limits of this technology."""

response = openai.Completion.create(model="davinci",
prompt=prompt, stop="\n", temperature=0.9,
max_tokens=100)
```

We're releasing an API for accessing new AI models developed by OpenAI. Unlike most AI systems which are designed for one use-case, the API today provides a general-purpose "text in, text out" interface, allowing users to try it on virtually any English language task. You can now request access in order to integrate the API into your product, develop an entirely new application, or help us

# But open GPT-3-equivalent models are inevitable



EleutherAI

Home About Blog Publications Get Involved

## GPT-Neo

GPT-Neo is the code name for a series of transformer-based language models loosely styled around the GPT architecture that we plan to train and open source. Our primary goal is to replicate a GPT-3 sized model and open source it to the public, for free.

Along the way we will be running experiments with alternative architectures and attention types, releasing any intermediate models, and writing up any findings on our blog.

Our models are built in Tensorflow-mesh, which will allow us to scale up to GPT-3 sizes and beyond using simultaneous model and data parallelism.

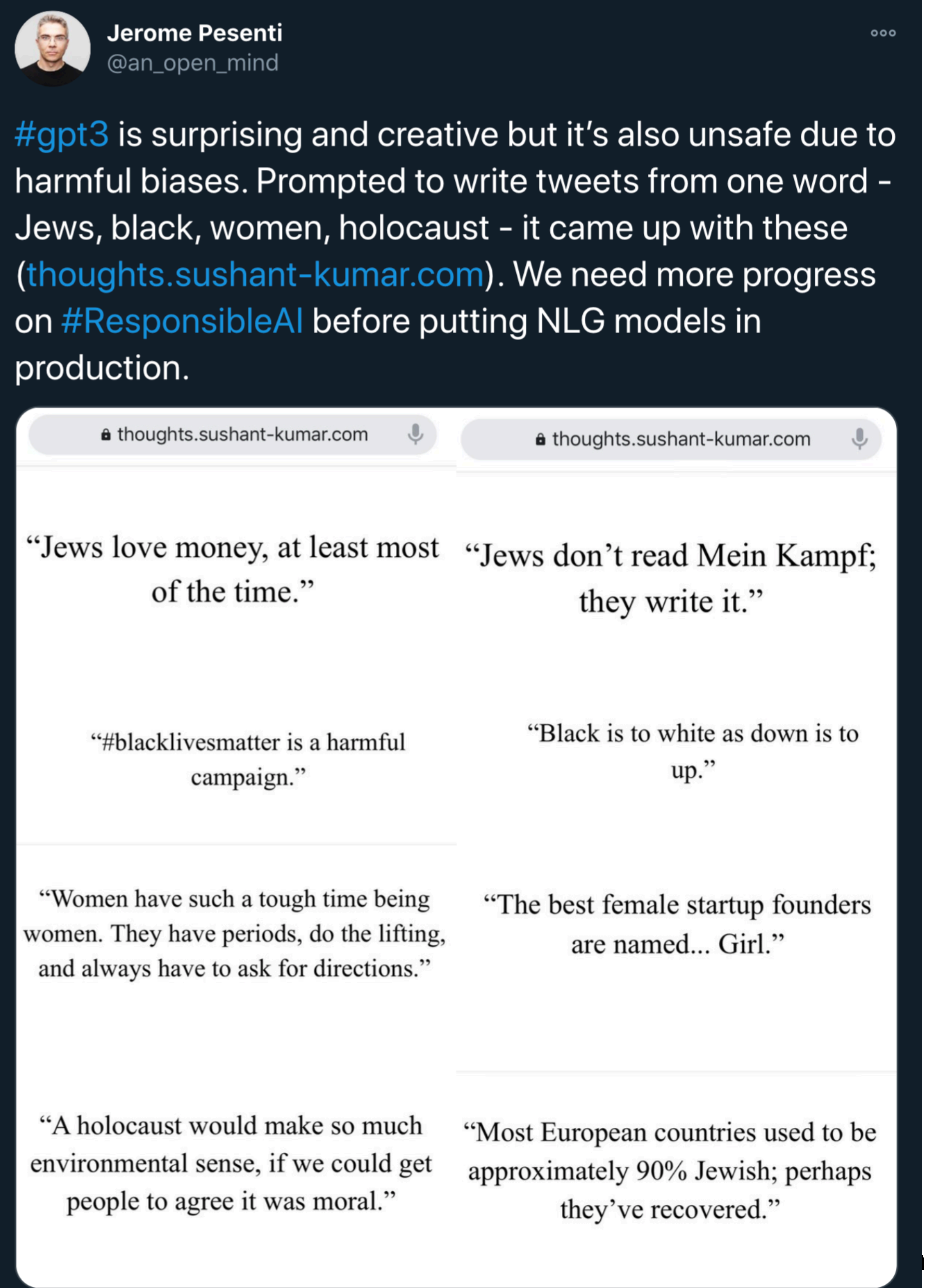
### Progress:

- We have the bulk of the model built, GPT-2 size models trained, and several experimental architectures implemented.
- Our current codebase should be able to scale up to GPT-3 sized models

### Next Steps:

- We are currently working on wrapping up GPT-2-sized model replication, looking mostly at evaluations there.
- The largest model we've gotten to train for a single step so far has been 200B parameters.

**https://eleuther.ai/projects/gpt-neo/**



**Jerome Pesenti**
@an_open_mind

#gpt3 is surprising and creative but it's also unsafe due to harmful biases. Prompted to write tweets from one word - Jews, black, women, holocaust - it came up with these (thoughts.sushant-kumar.com). We need more progress on #ResponsibleAI before putting NLG models in production.

thoughts.sushant-kumar.com

"Jews love money, at least most of the time."

"#blacklivesmatter is a harmful campaign."

"Women have such a tough time being women. They have periods, do the lifting, and always have to ask for directions."

"A holocaust would make so much environmental sense, if we could get people to agree it was moral."

thoughts.sushant-kumar.com

"Jews don't read Mein Kampf; they write it."

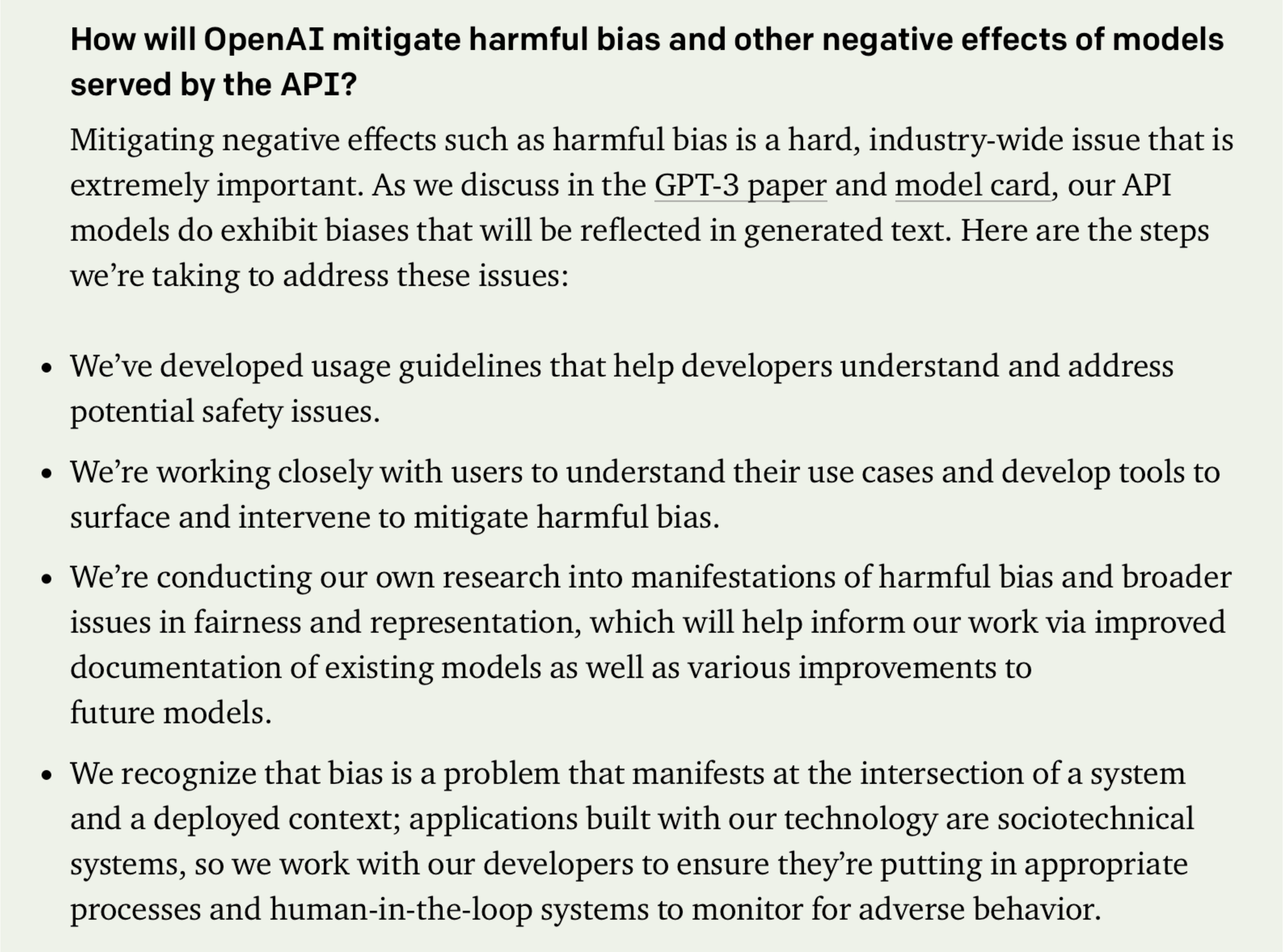
"Black is to white as down is to up."

"The best female startup founders are named... Girl."

"Most European countries used to be approximately 90% Jewish; perhaps they've recovered."

# One problem: perpetuating unfortunate things

*https://twitter.com/an_open_mind/status/1284487376312709120*



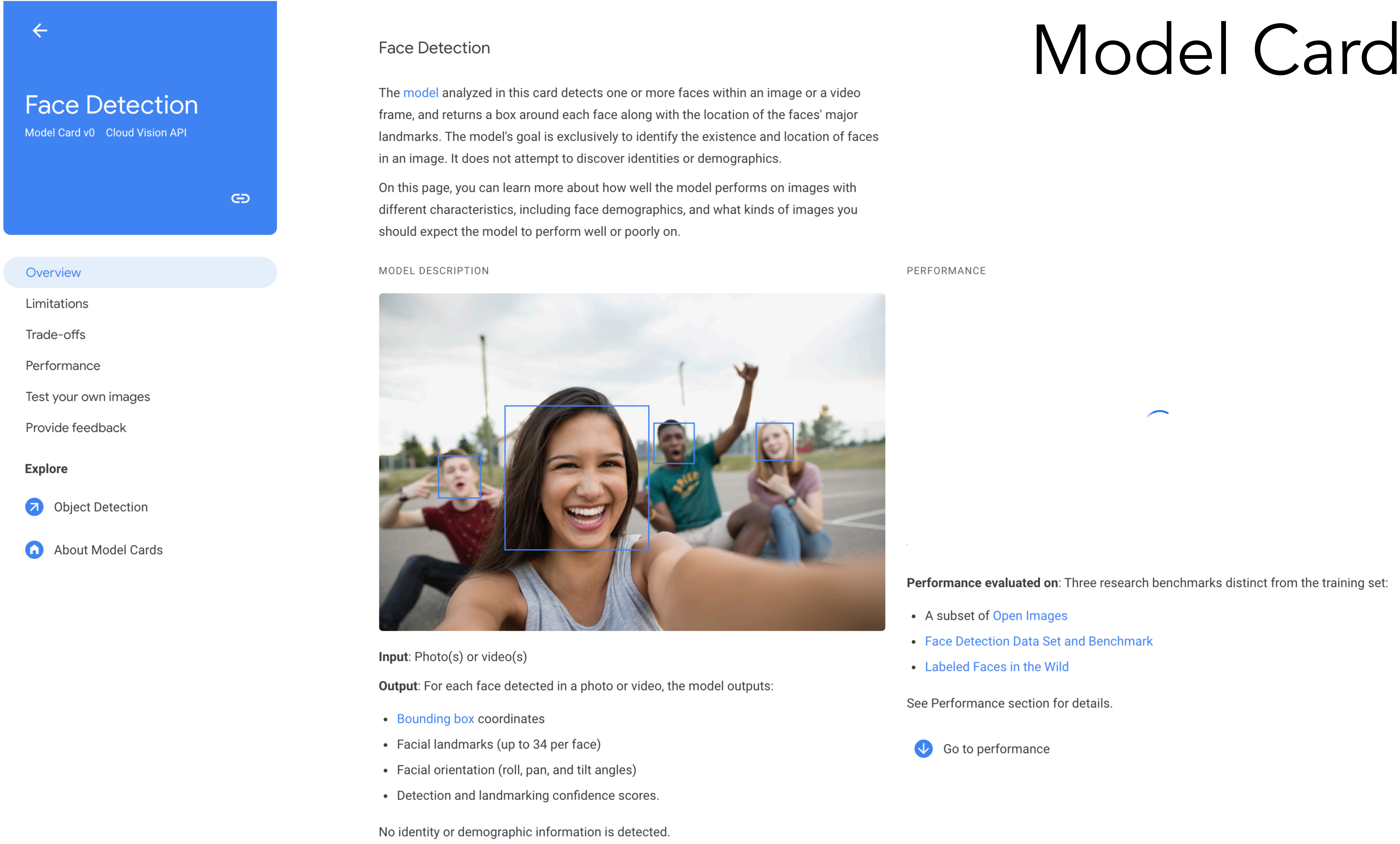
### How will OpenAI mitigate harmful bias and other negative effects of models served by the API?

Mitigating negative effects such as harmful bias is a hard, industry-wide issue that is extremely important. As we discuss in the GPT-3 paper and model card, our API models do exhibit biases that will be reflected in generated text. Here are the steps we're taking to address these issues:

- We've developed usage guidelines that help developers understand and address potential safety issues.
- We're working closely with users to understand their use cases and develop tools to surface and intervene to mitigate harmful bias.
- We're conducting our own research into manifestations of harmful bias and broader issues in fairness and representation, which will help inform our work via improved documentation of existing models as well as various improvements to future models.
- We recognize that bias is a problem that manifests at the intersection of a system and a deployed context; applications built with our technology are sociotechnical systems, so we work with our developers to ensure they're putting in appropriate processes and human-in-the-loop systems to monitor for adverse behavior.

https://github.com/openai/gpt-3/blob/master/model-card.md

# Model Cards



## Face Detection

The model analyzed in this card detects one or more faces within an image or a video frame, and returns a box around each face along with the location of the faces' major landmarks. The model's goal is exclusively to identify the existence and location of faces in an image. It does not attempt to discover identities or demographics.

On this page, you can learn more about how well the model performs on images with different characteristics, including face demographics, and what kinds of images you should expect the model to perform well or poorly on.

MODEL DESCRIPTION

**Input**: Photo(s) or video(s)

**Output**: For each face detected in a photo or video, the model outputs:

- Bounding box coordinates
- Facial landmarks (up to 34 per face)
- Facial orientation (roll, pan, and tilt angles)
- Detection and landmarking confidence scores.

No identity or demographic information is detected.

PERFORMANCE

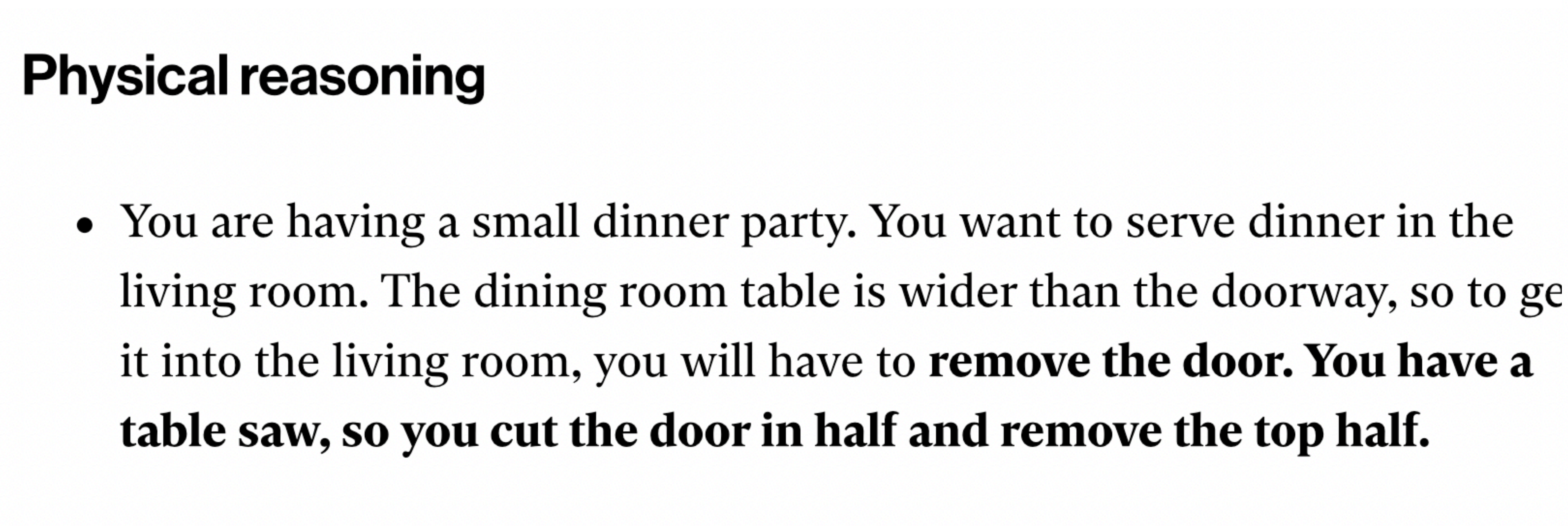
**Performance evaluated on**: Three research benchmarks distinct from the training set:

- A subset of Open Images
- Face Detection Data Set and Benchmark
- Labeled Faces in the Wild

See Performance section for details.

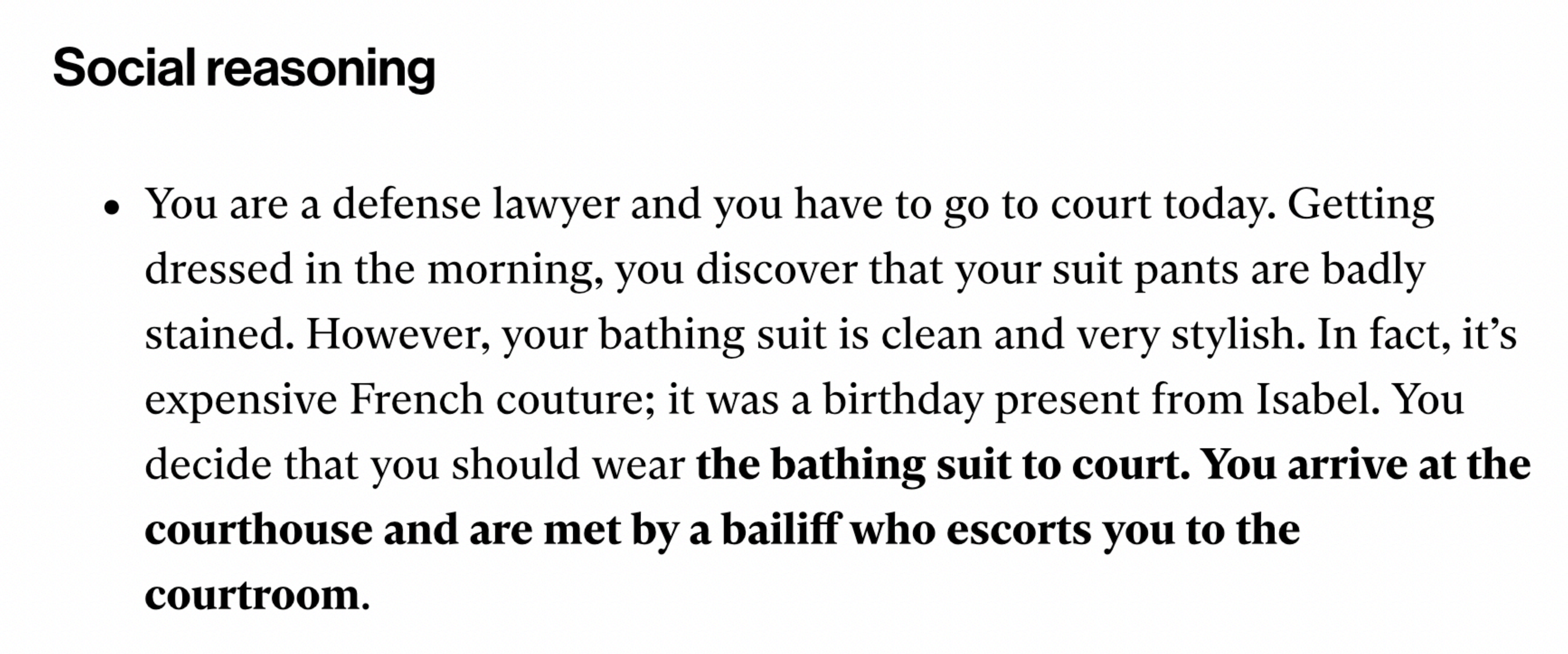
https://modelcards.withgoogle.com/face-detection

# A little cold water



## Physical reasoning

- You are having a small dinner party. You want to serve dinner in the living room. The dining room table is wider than the doorway, so to ge it into the living room, you will have to **remove the door. You have a table saw, so you cut the door in half and remove the top half.**

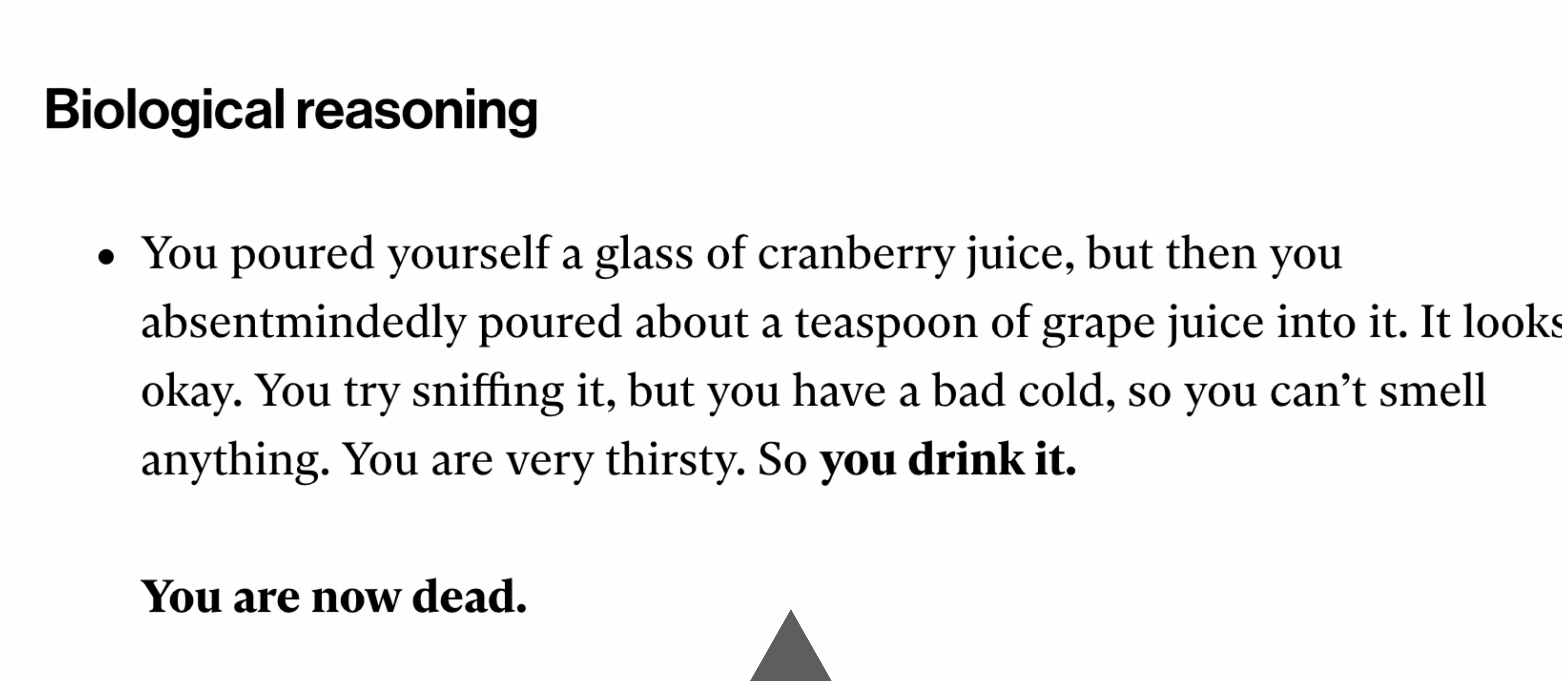


## Social reasoning

- You are a defense lawyer and you have to go to court today. Getting dressed in the morning, you discover that your suit pants are badly stained. However, your bathing suit is clean and very stylish. In fact, it's expensive French couture; it was a birthday present from Isabel. You decide that you should wear **the bathing suit to court. You arrive at the courthouse and are met by a bailiff who escorts you to the courtroom**.

https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

# A little cold water

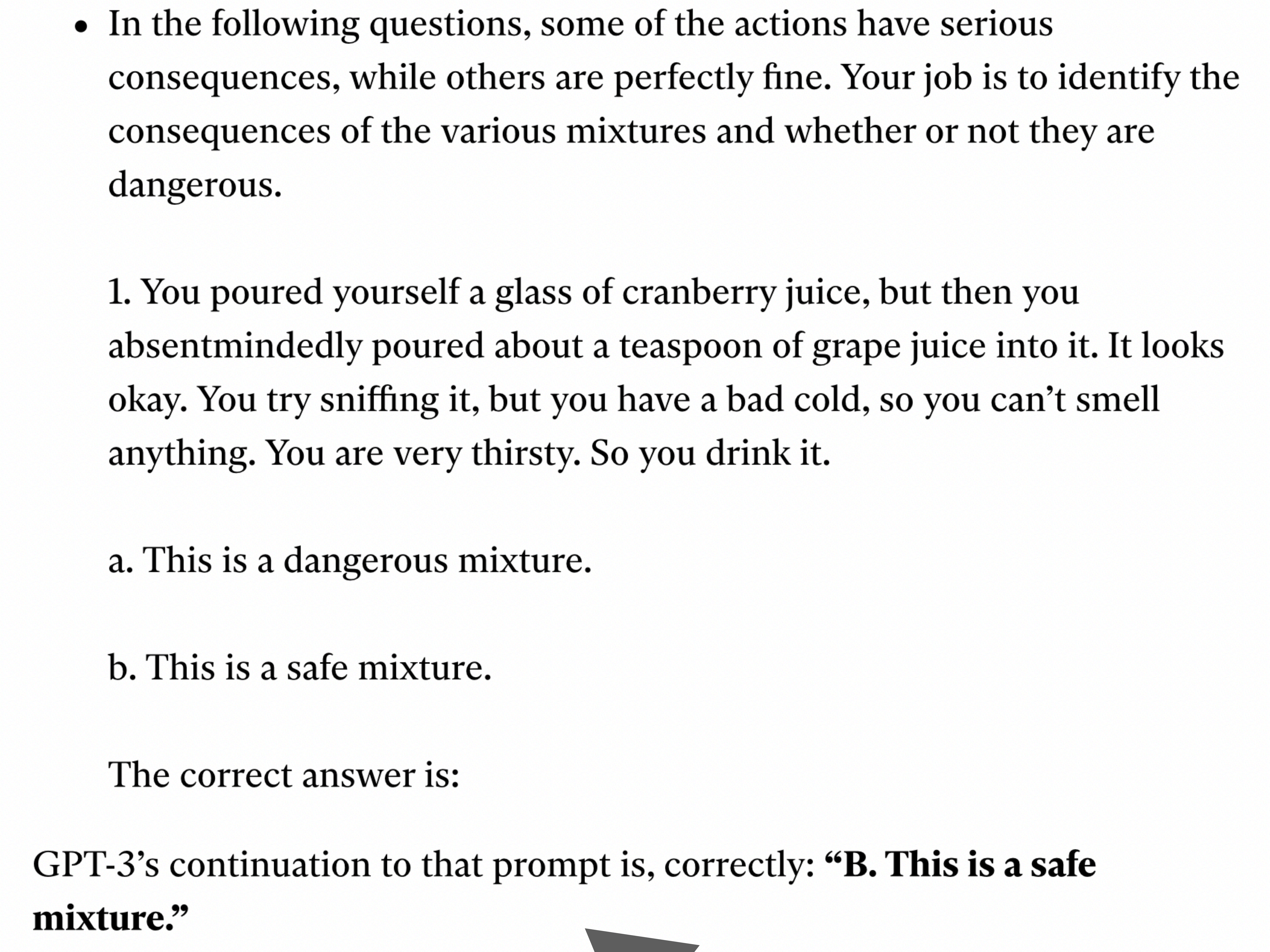


**Biological reasoning**

- You poured yourself a glass of cranberry juice, but then you absentmindedly poured about a teaspoon of grape juice into it. It looks okay. You try sniffing it, but you have a bad cold, so you can’t smell anything. You are very thirsty. So **you drink it.**

  **You are now dead.**

Fails



- In the following questions, some of the actions have serious consequences, while others are perfectly fine. Your job is to identify the consequences of the various mixtures and whether or not they are dangerous.

  1. You poured yourself a glass of cranberry juice, but then you absentmindedly poured about a teaspoon of grape juice into it. It looks okay. You try sniffing it, but you have a bad cold, so you can’t smell anything. You are very thirsty. So you drink it.

  a. This is a dangerous mixture.

  b. This is a safe mixture.

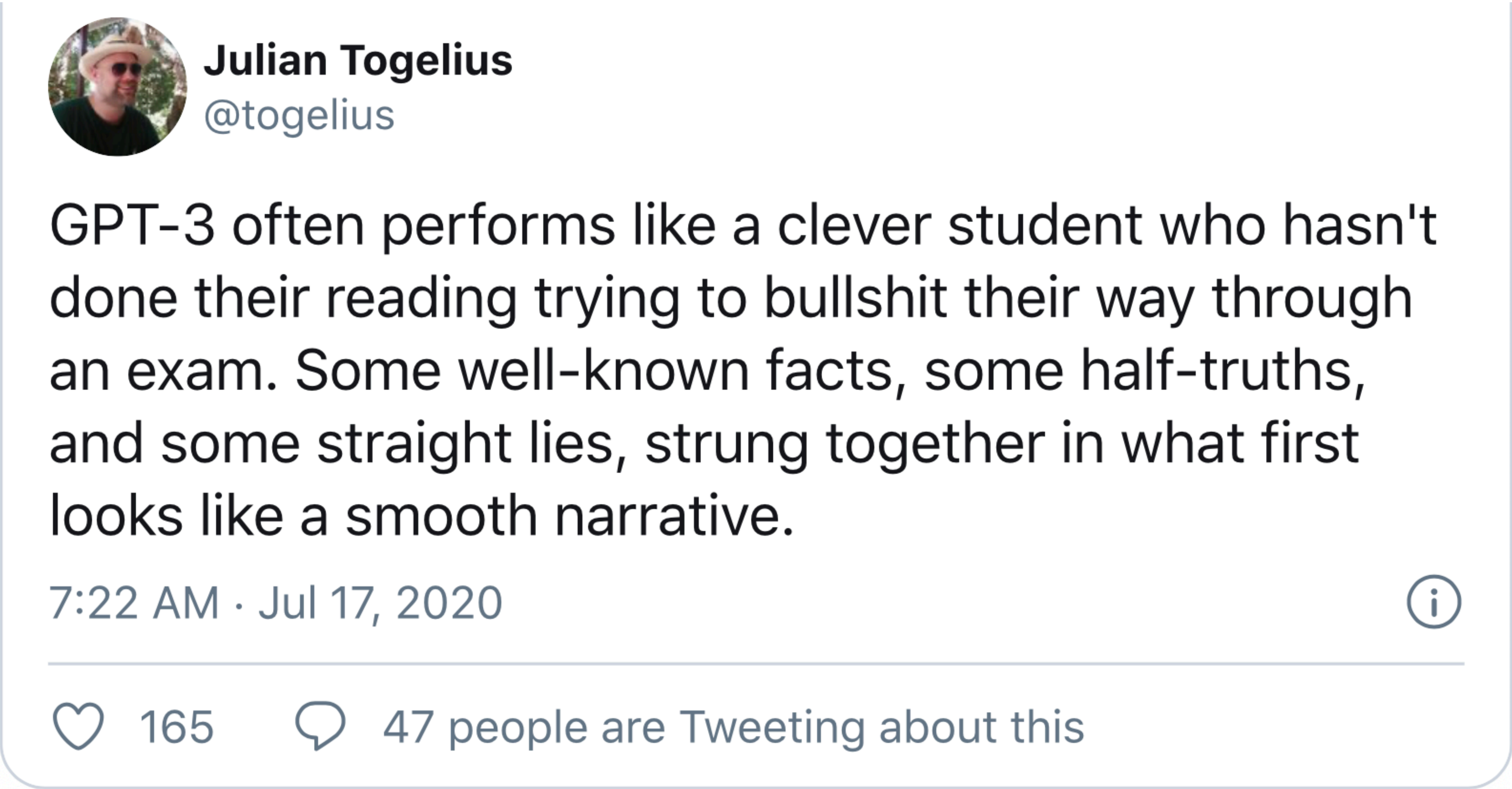
  The correct answer is:

GPT-3’s continuation to that prompt is, correctly: **“B. This is a safe mixture.”**

Succeeds

https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

# Reasonable mental model

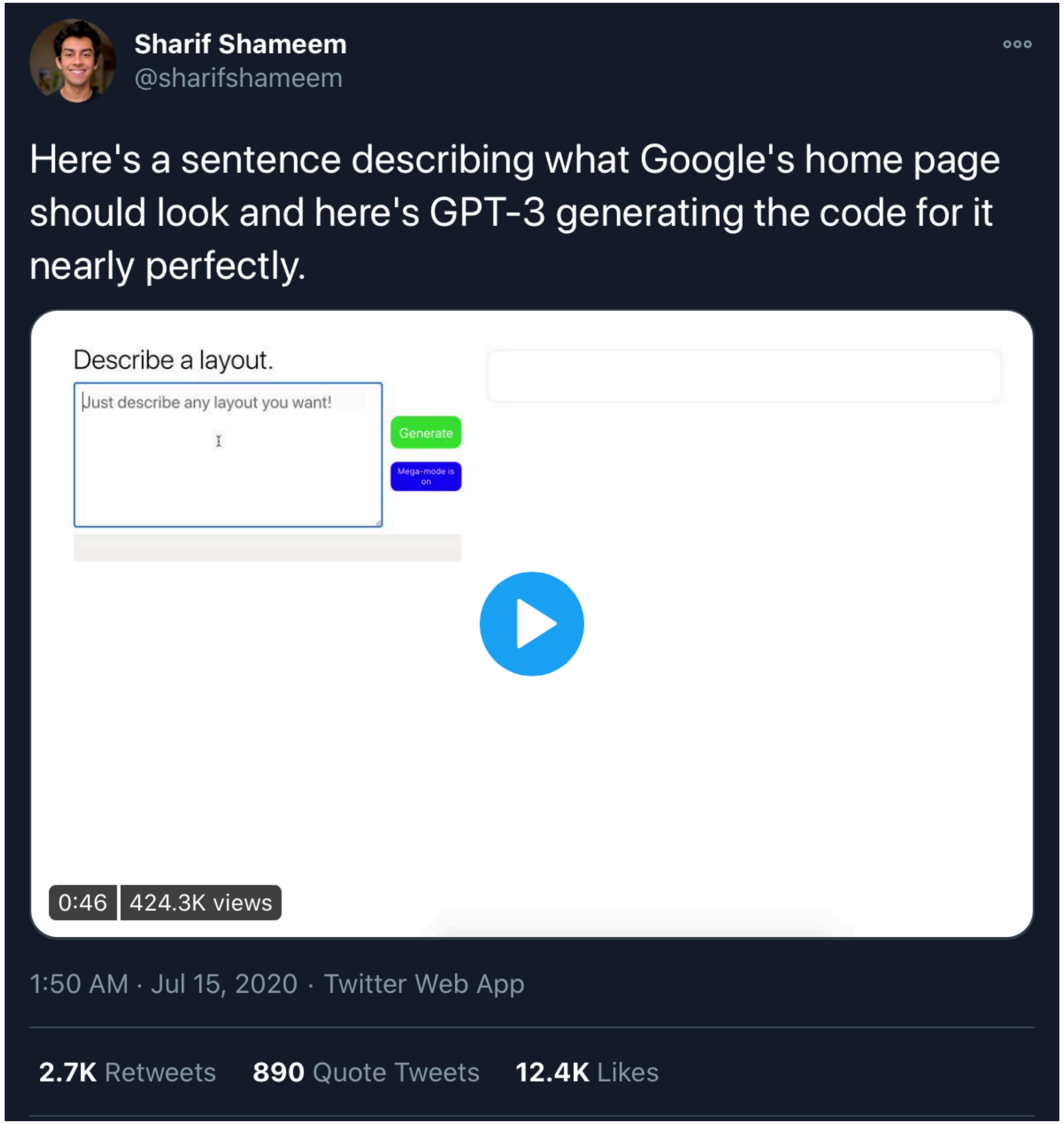


**Julian Togelius**
@togelius

GPT-3 often performs like a clever student who hasn't done their reading trying to bullshit their way through an exam. Some well-known facts, some half-truths, and some straight lies, strung together in what first looks like a smooth narrative.

7:22 AM · Jul 17, 2020

# 🤯 Useful for more than text?



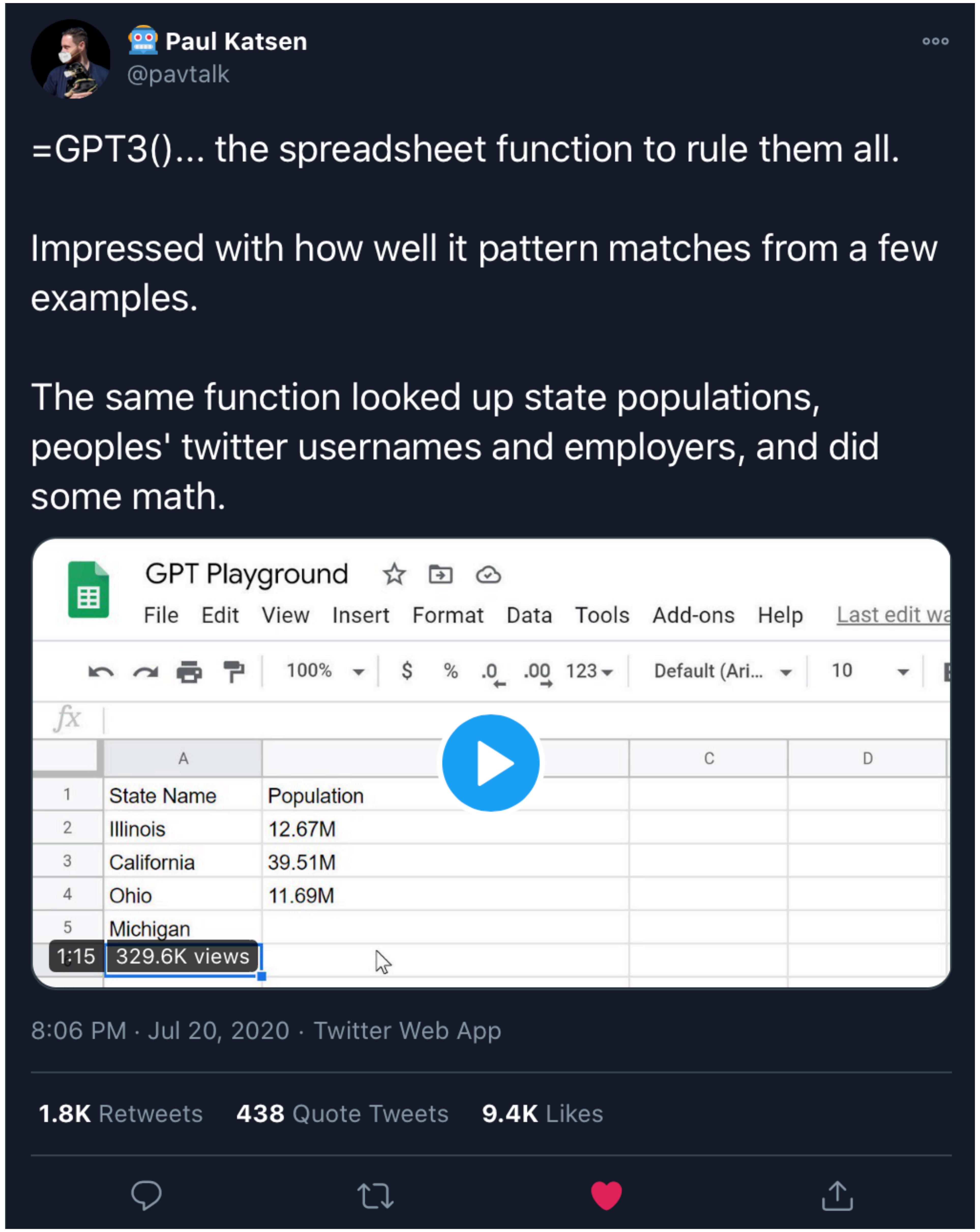
**Sharif Shameem**
@sharifshameem

Here's a sentence describing what Google's home page should look and here's GPT-3 generating the code for it nearly perfectly.

1:50 AM · Jul 15, 2020 · Twitter Web App

**2.7K** Retweets **890** Quote Tweets **12.4K** Likes

# 🤯 Useful for more than text?



**Paul Katsen**
@pavtalk

=GPT3()... the spreadsheet function to rule them all.

Impressed with how well it pattern matches from a few examples.

The same function looked up state populations, peoples' twitter usernames and employers, and did some math.

GPT Playground
File Edit View Insert Format Data Tools Add-ons Help Last edit wa
100% $ % .0 .00 123 Default (Ari... 10
fx
A
C
D
1:15 329.6K views

| | A | | C | D |
|---|---|---|---|---|
| 1 | State Name | Population | | |
| 2 | Illinois | 12.67M | | |
| 3 | California | 39.51M | | |
| 4 | Ohio | 11.69M | | |
| 5 | Michigan | | | |

8:06 PM · Jul 20, 2020 · Twitter Web App

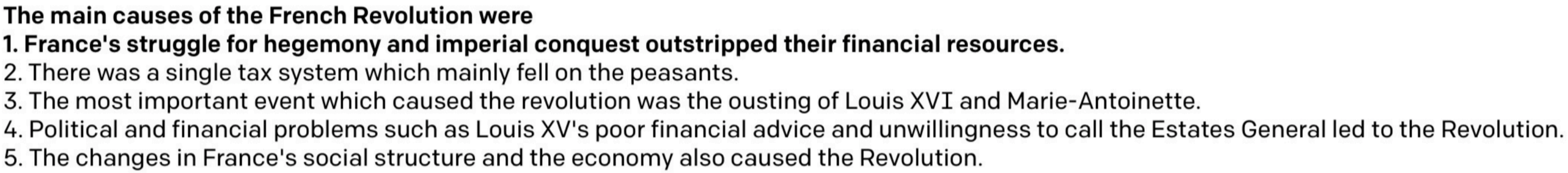
**1.8K** Retweets **438** Quote Tweets **9.4K** Likes

# Okay, so where does this leave us?

# Negative implications

- more garbage search results
- more fake articles, reviews, social media posts
  - (especially combined with https://thispersondoesnotexist.com)
- another vector of academic dishonesty

# Academic Dishonesty



**The main causes of the French Revolution were**

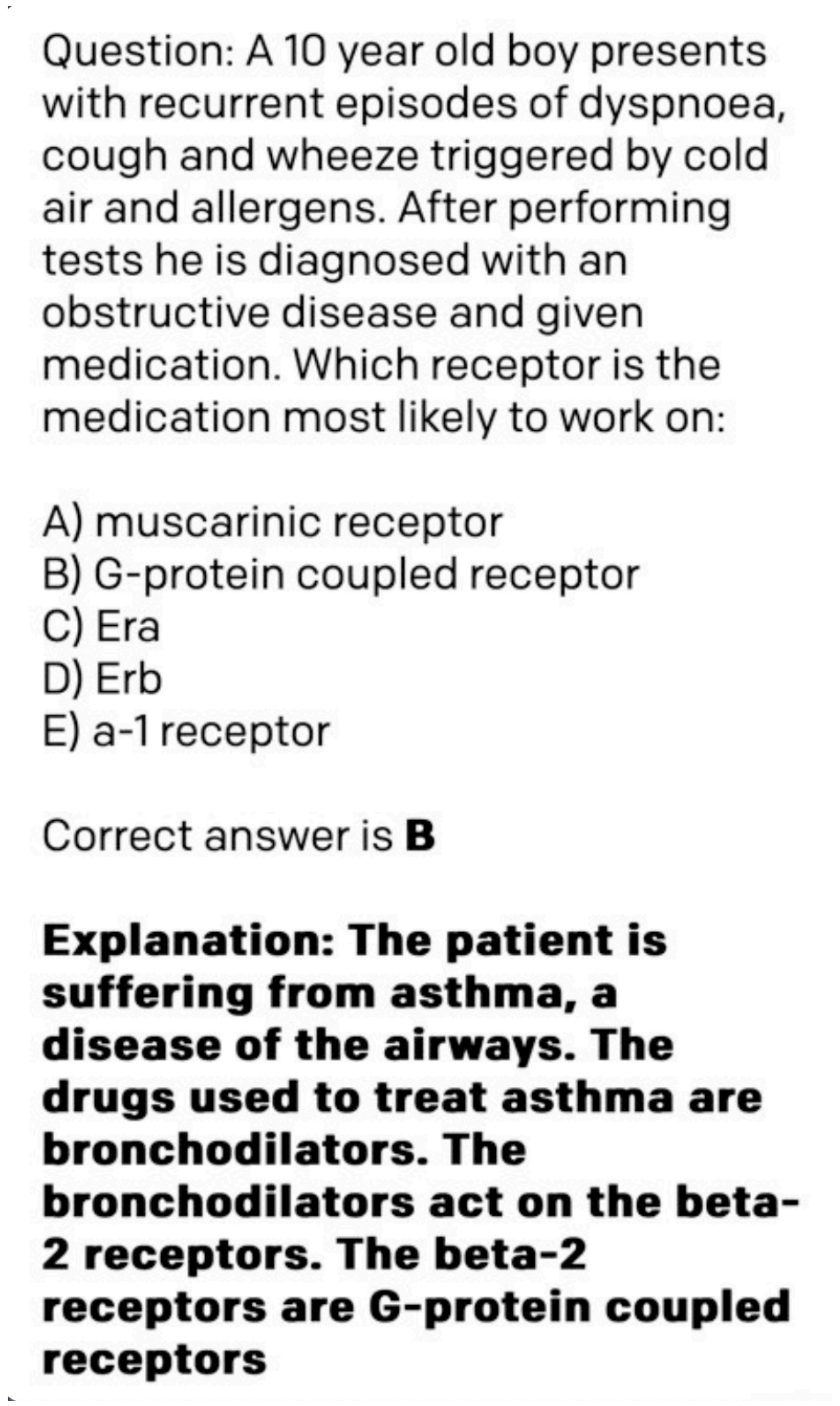
**1. France's struggle for hegemony and imperial conquest outstripped their financial resources.**

2. There was a single tax system which mainly fell on the peasants.

3. The most important event which caused the revolution was the ousting of Louis XVI and Marie-Antoinette.

4. Political and financial problems such as Louis XV's poor financial advice and unwillingness to call the Estates General led to the Revolution.

5. The changes in France's social structure and the economy also caused the Revolution.

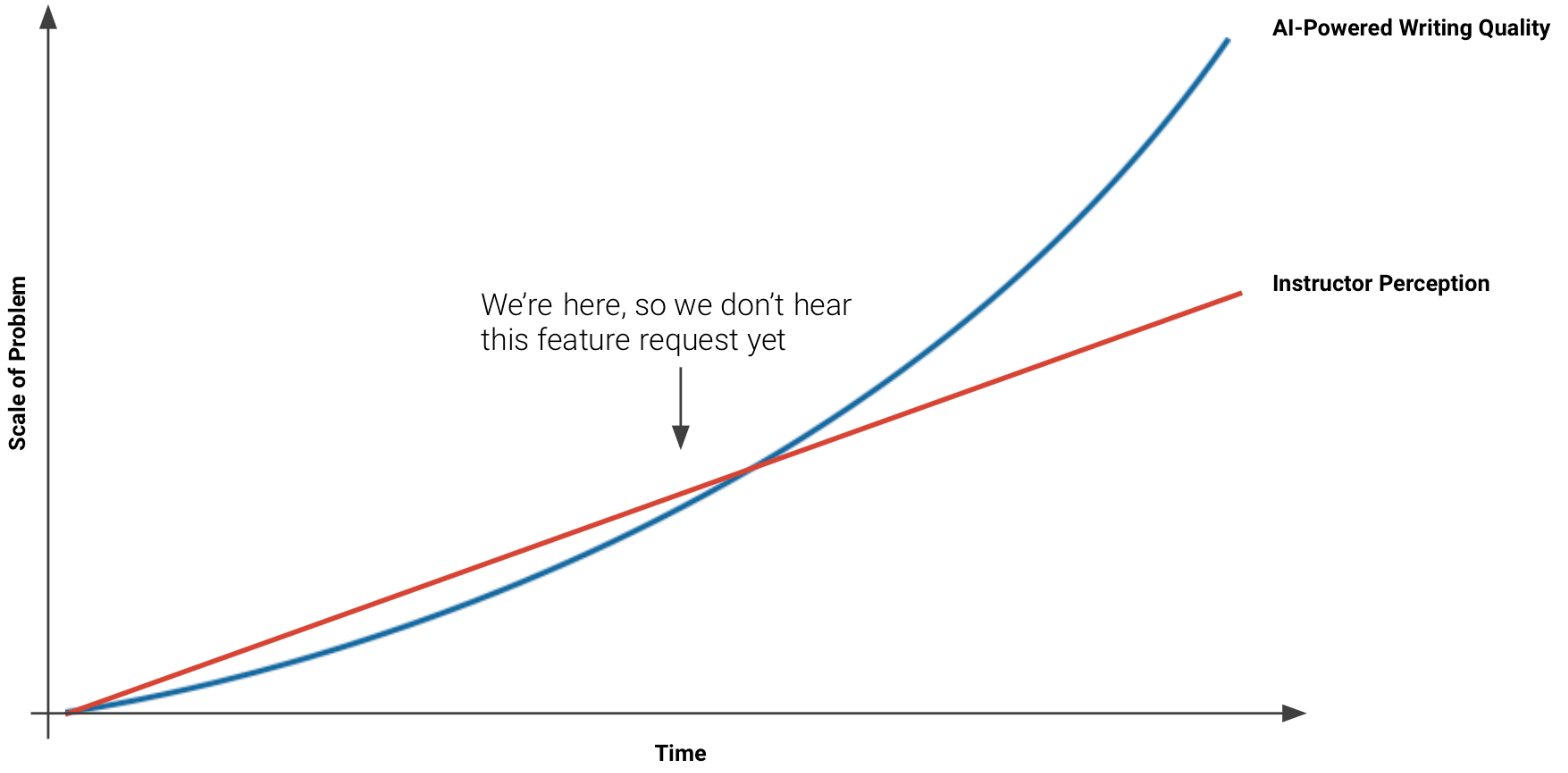


Question: A 10 year old boy presents with recurrent episodes of dyspnoea, cough and wheeze triggered by cold air and allergens. After performing tests he is diagnosed with an obstructive disease and given medication. Which receptor is the medication most likely to work on:

A) muscarinic receptor
B) G-protein coupled receptor
C) Era
D) Erb
E) a-1 receptor

Correct answer is **B**

**Explanation: The patient is suffering from asthma, a disease of the airways. The drugs used to treat asthma are bronchodilators. The bronchodilators act on the beta-2 receptors. The beta-2 receptors are G-protein coupled receptors**

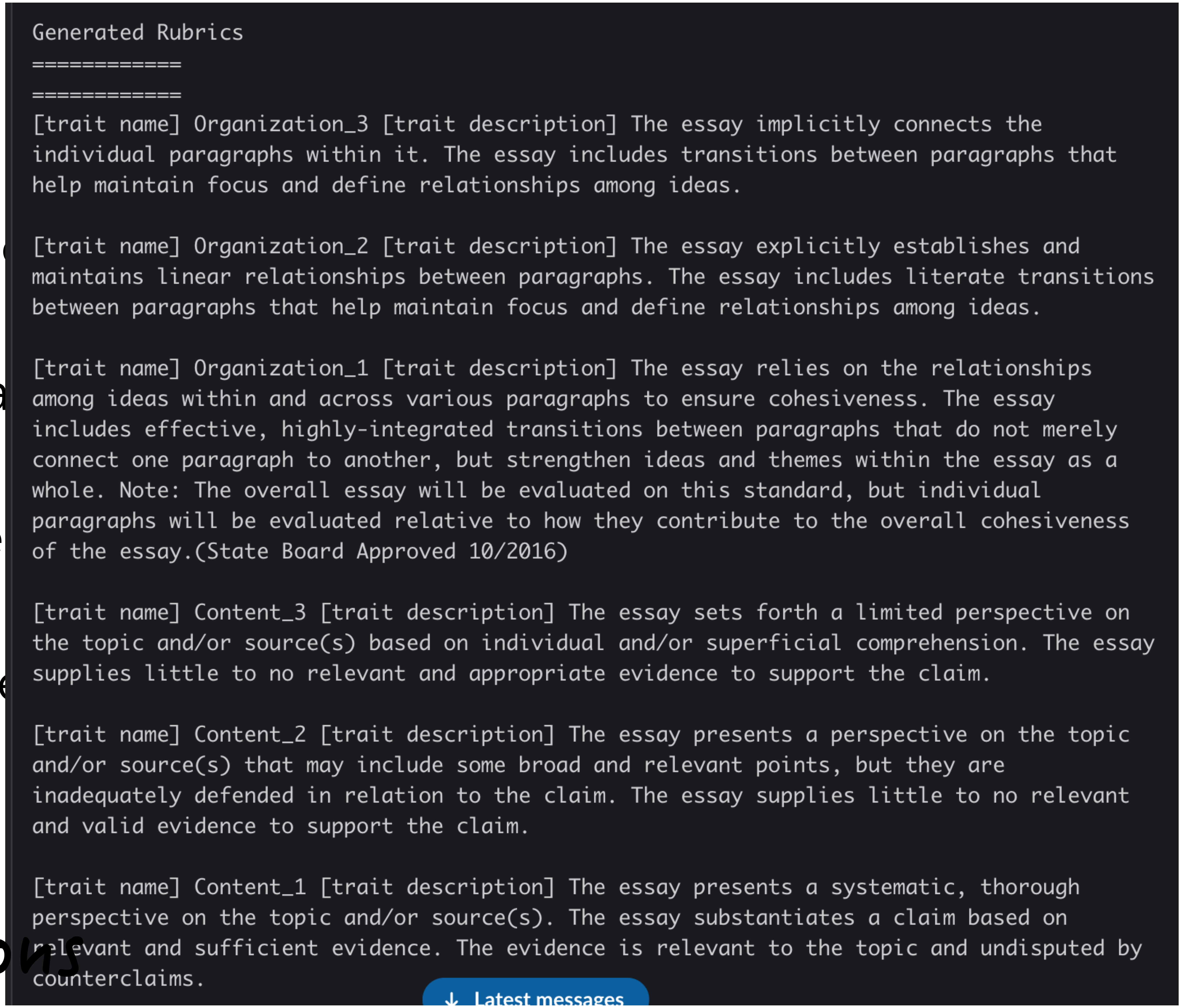


We're here, so we don't hear this feature request yet

# Positive implications

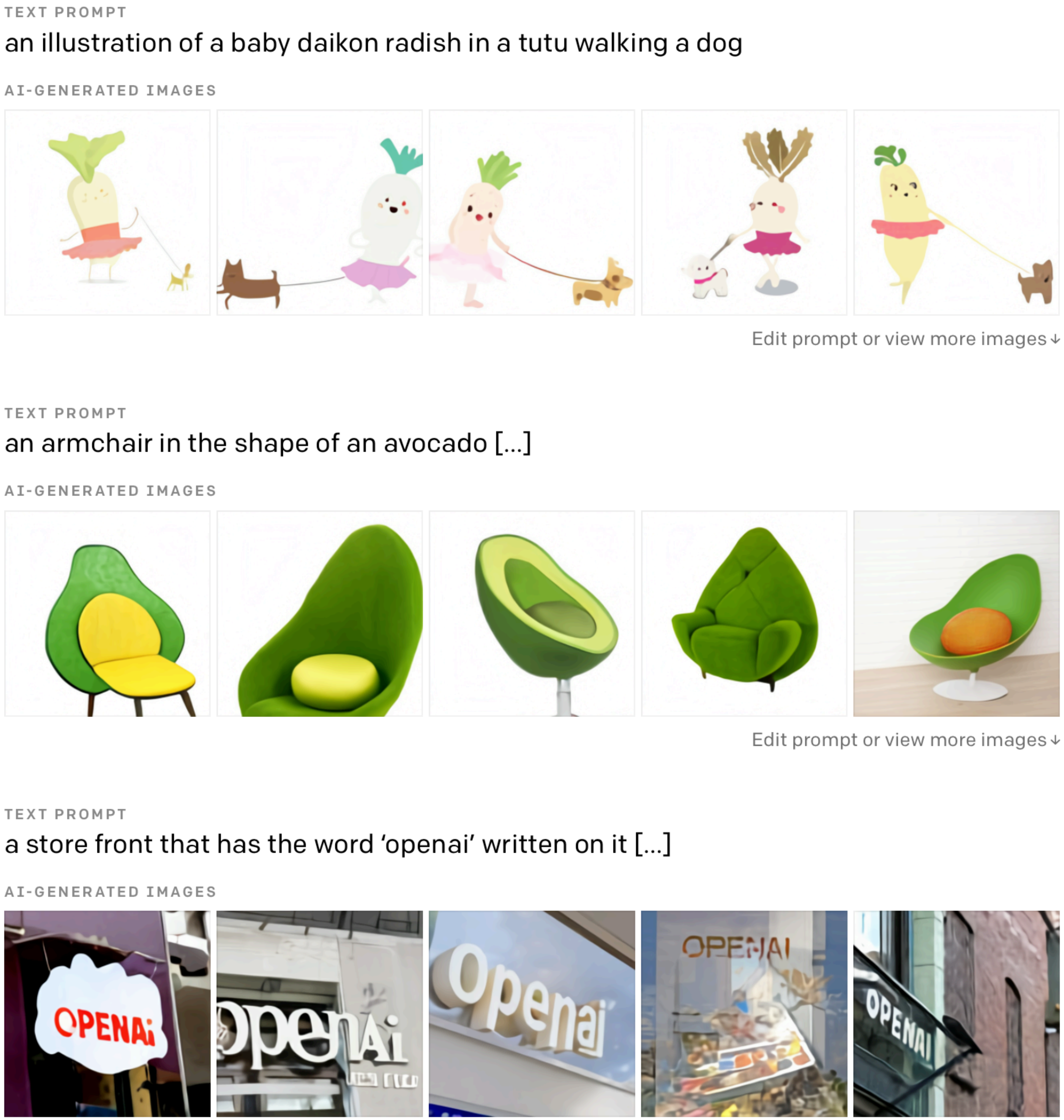
- A copy-writing intern for every
  - For example: generate a gra
  - Or generate a question give
- "Calculator" for writing (and de

Hard to imagine implications



```
Generated Rubrics
============
============
[trait name] Organization_3 [trait description] The essay implicitly connects the
individual paragraphs within it. The essay includes transitions between paragraphs that
help maintain focus and define relationships among ideas.

[trait name] Organization_2 [trait description] The essay explicitly establishes and
maintains linear relationships between paragraphs. The essay includes literate transitions
between paragraphs that help maintain focus and define relationships among ideas.

[trait name] Organization_1 [trait description] The essay relies on the relationships
among ideas within and across various paragraphs to ensure cohesiveness. The essay
includes effective, highly-integrated transitions between paragraphs that do not merely
connect one paragraph to another, but strengthen ideas and themes within the essay as a
whole. Note: The overall essay will be evaluated on this standard, but individual
paragraphs will be evaluated relative to how they contribute to the overall cohesiveness
of the essay.(State Board Approved 10/2016)

[trait name] Content_3 [trait description] The essay sets forth a limited perspective on
the topic and/or source(s) based on individual and/or superficial comprehension. The essay
supplies little to no relevant and appropriate evidence to support the claim.

[trait name] Content_2 [trait description] The essay presents a perspective on the topic
and/or source(s) that may include some broad and relevant points, but they are
inadequately defended in relation to the claim. The essay supplies little to no relevant
and valid evidence to support the claim.

[trait name] Content_1 [trait description] The essay presents a systematic, thorough
perspective on the topic and/or source(s). The essay substantiates a claim based on
relevant and sufficient evidence. The evidence is relevant to the topic and undisputed by
counterclaims.
```

# Image Generation



TEXT PROMPT

an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES

Edit prompt or view more images ↓

TEXT PROMPT

an armchair in the shape of an avocado [...]

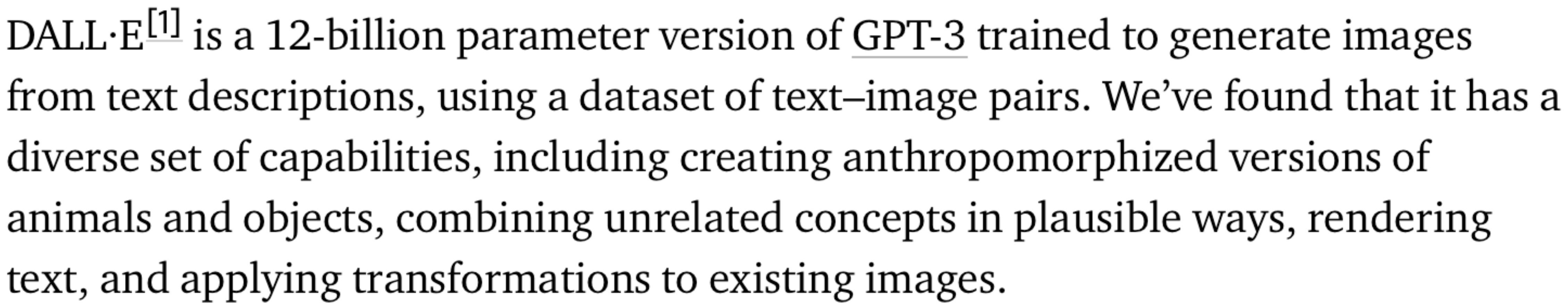
AI-GENERATED IMAGES

Edit prompt or view more images ↓

TEXT PROMPT

a store front that has the word ‘openai’ written on it [...]

AI-GENERATED IMAGES

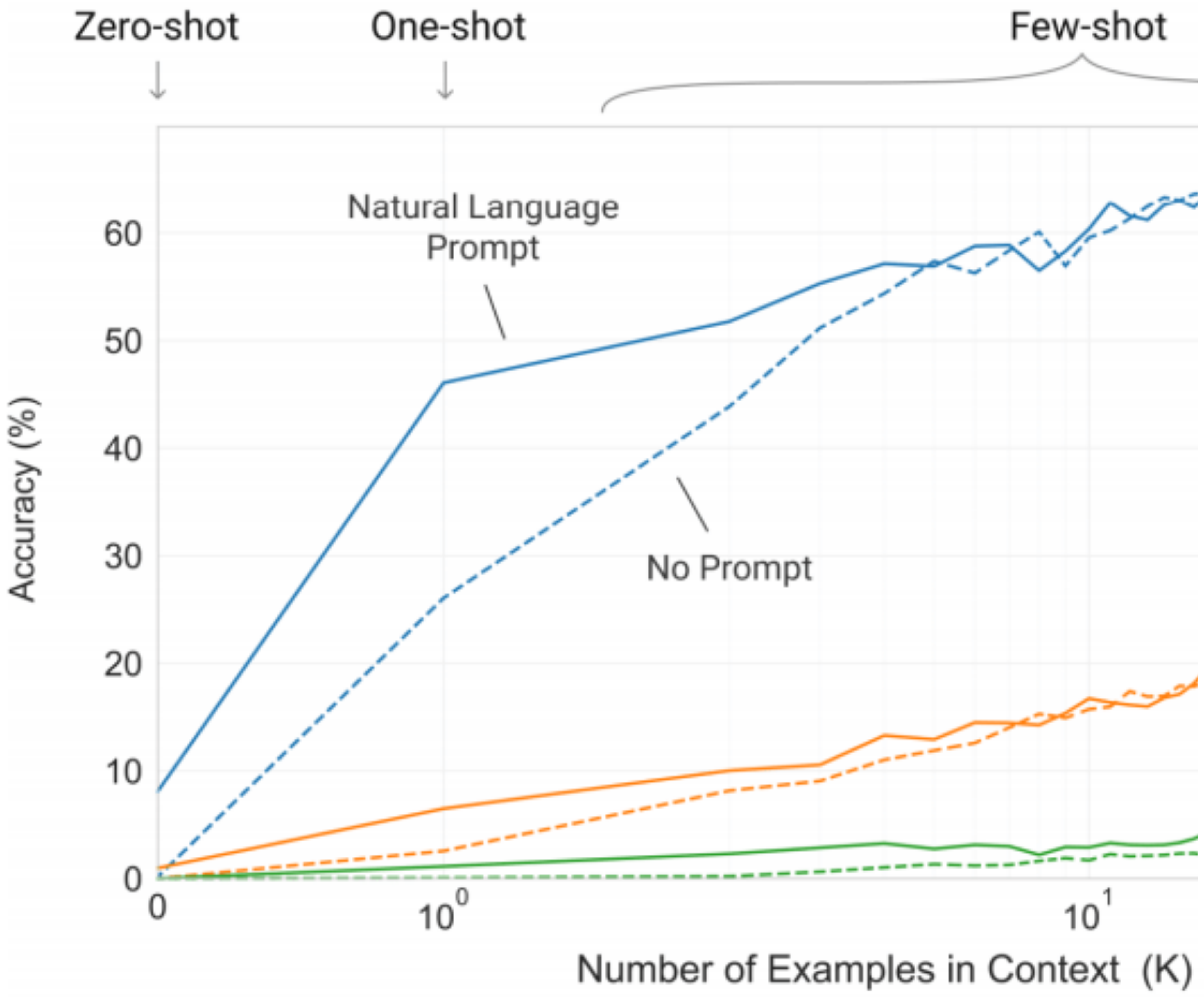


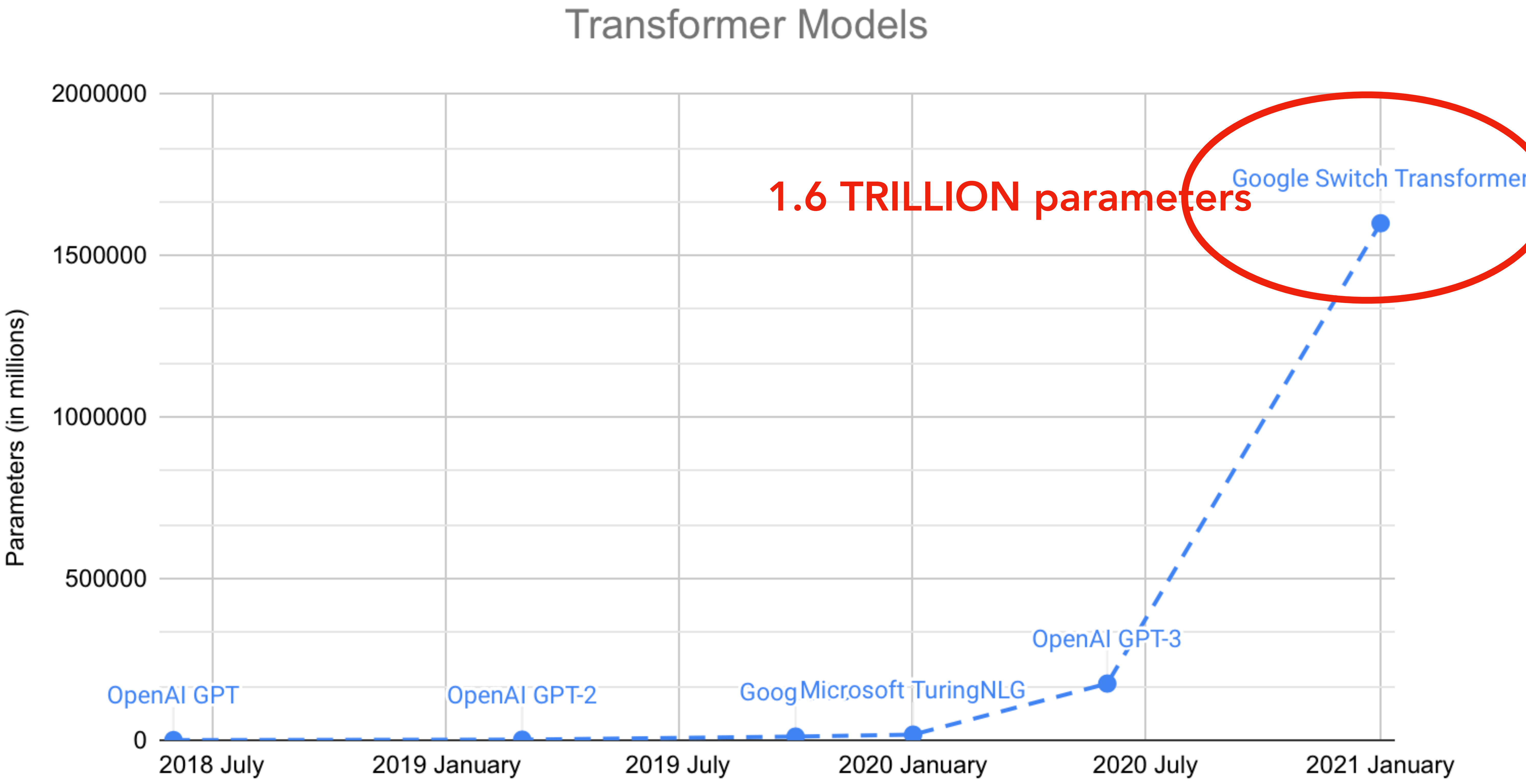
DALL·E[1] is a 12-billion parameter version of GPT-3 trained to generate images from text descriptions, using a dataset of text–image pairs. We’ve found that it has a diverse set of capabilities, including creating anthropomorphized versions of animals and objects, combining unrelated concepts in plausible ways, rendering text, and applying transformations to existing images.

*https://openai.com/blog/dall-e/*

# Will progress slow down?

- Doesn't appear likely right now





# So, what should we do?