

Building AI products

Johannes K. Kolberg
27/9/24, [LinkedIn](#)

FROM KERAS IMPORT SEQUENTIAL

**YOU KNOW, I'M SOMETHING OF A
MACHINE LEARNING ENGINEER MYSELF**

Who's this guy?



Data Analyst Intern



Data Engineer



Data Scientist
AI Product Manager



Head of Product &
Engineering



Head of Product &
Engineering



Business Analytics
Finance (Investments)

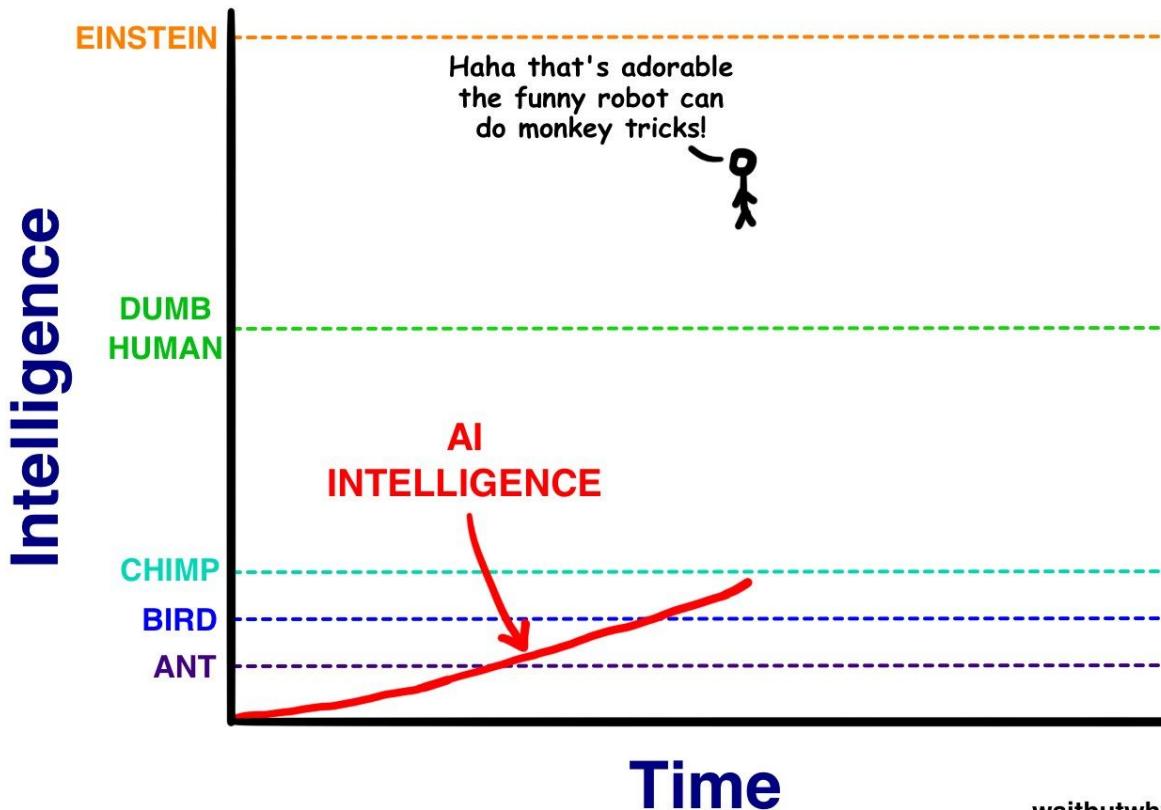
MS Data Science
Aker Scholarship

People with no idea about AI
saying it will take over the world:



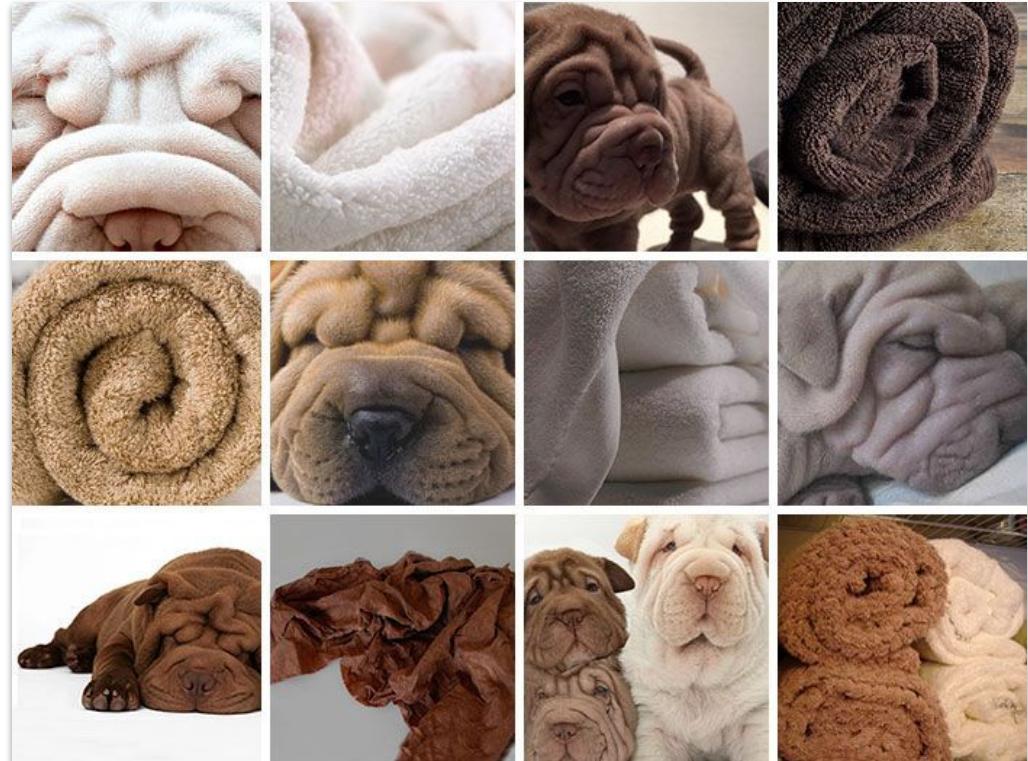
My Neural Network:

Our Distorted View of Intelligence

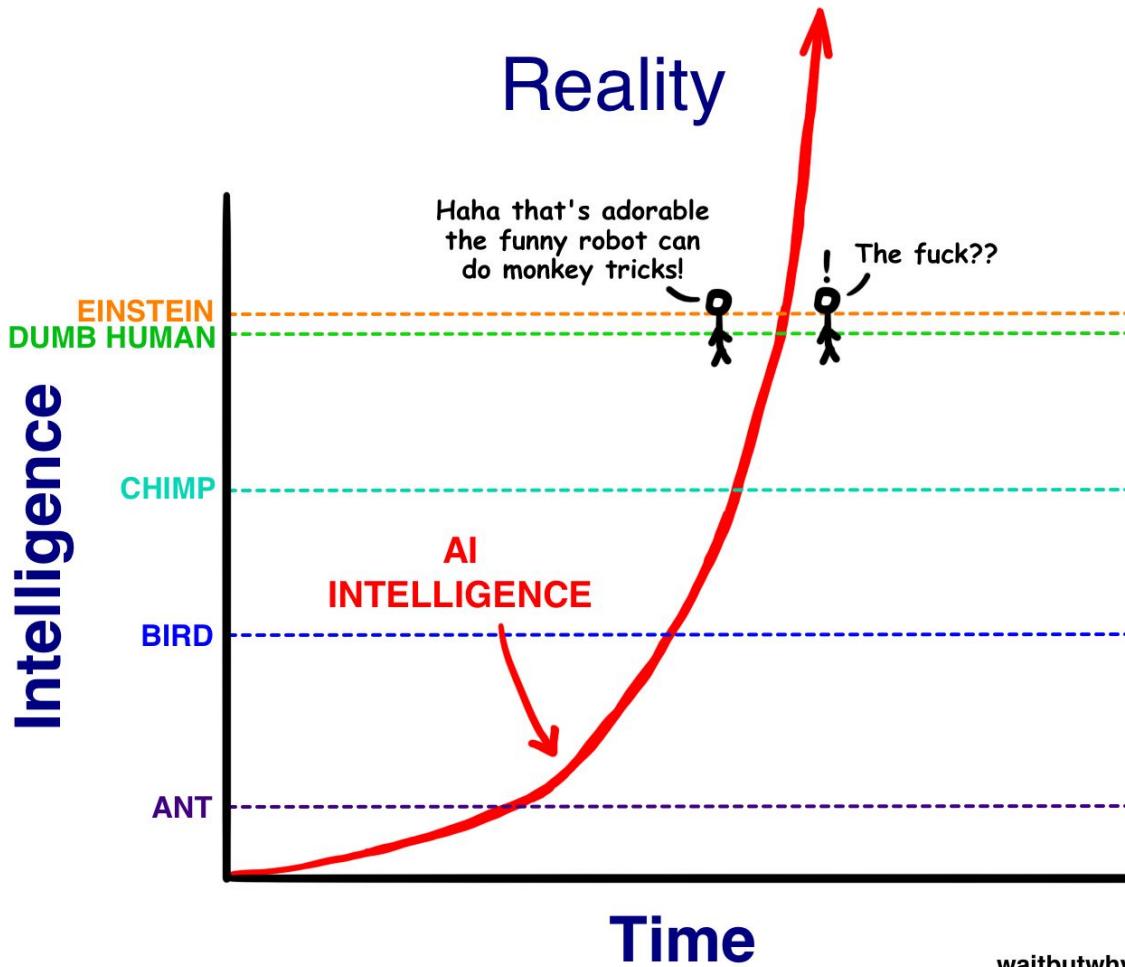




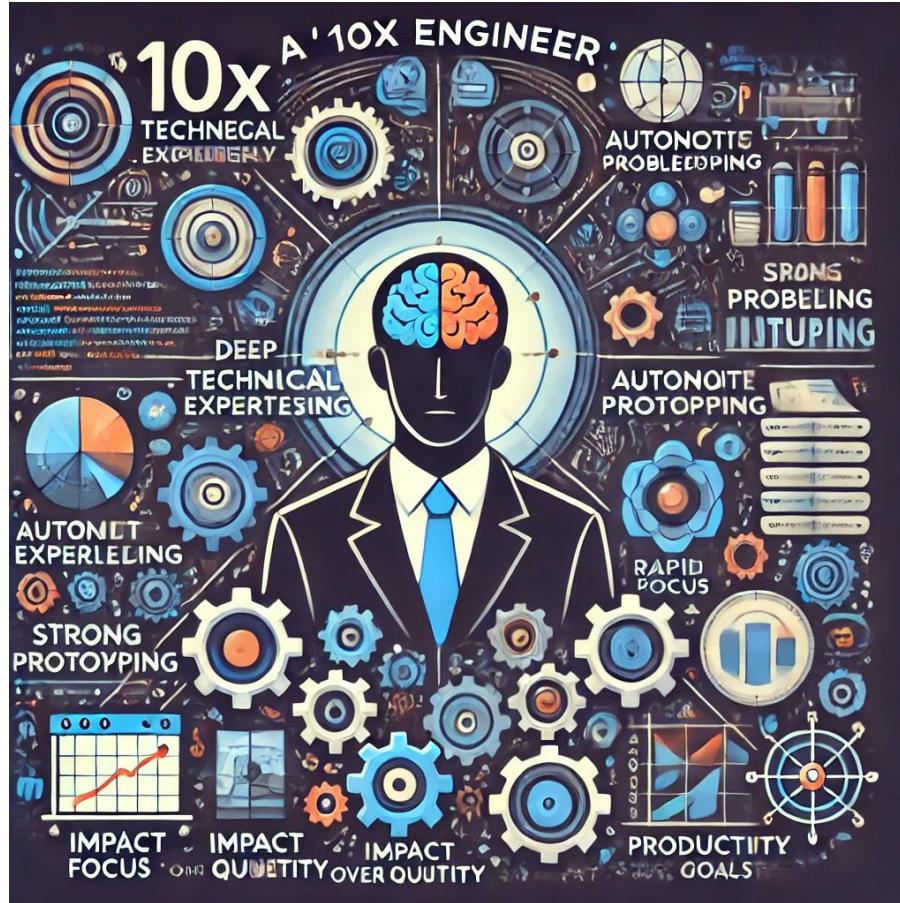
@teenybiscuit







Democratisation of 10x → 100x



Pragmatic toolstack



Streamlit



FORTUNE

Could AI create a one-person unicorn? Sam Altman thinks so—and Silicon Valley sees the technology ‘waiting for us’



Fortune · (picture alliance)

andreessen
horowitz



Why Software Is Eating the World

Marc Andreessen

SHARE ▾

AVAILABLE IN ▾

Posted August 20, 2011

Software is eating the world.

More than 10 years after the peak of the 1990s dot-com bubble, a dozen or so new Internet companies like Facebook and Twitter are sparking controversy in Silicon Valley, due to their rapidly growing private market valuations, and even the occasional successful IPO. With scars from the heyday of Webvan and Pets.com still fresh in the investor psyche, people are asking, “Isn’t this just a dangerous new bubble?”

I, along with others, have been arguing the other side of the case. (I am co-founder and general partner of venture capital firm Andreessen-Horowitz, which has invested in Facebook, Groupon, Skype, Twitter, Zynga, and Foursquare, among others. I am also personally an investor in LinkedIn.) We believe that many of the prominent new Internet companies are building real, high-growth, high-margin, highly defensible businesses.

Practical considerations

Prototyping is easy

The screenshot shows a Jupyter Notebook interface with the following details:

- Title Bar:** A Motivating Example Modified.ipynb
- File Tree:** NHH / ipynb_checkpoint.ipynb / A Motivating Example.ipynb / A Motivating Example.ipynb / Basics of Python.ipynb / configurable_colab.ipynb / databutton_instrumentation.ipynb / Exercises.ipynb / syllabus_BAN43... / Video.mov
- Code Cell (In [19]):**

```
1 # Check type of each variable
2 df.dtypes
...
19   ✓ 0.0s
...
    csv      object
dtype: object
```
- Code Cell (In [47]):**

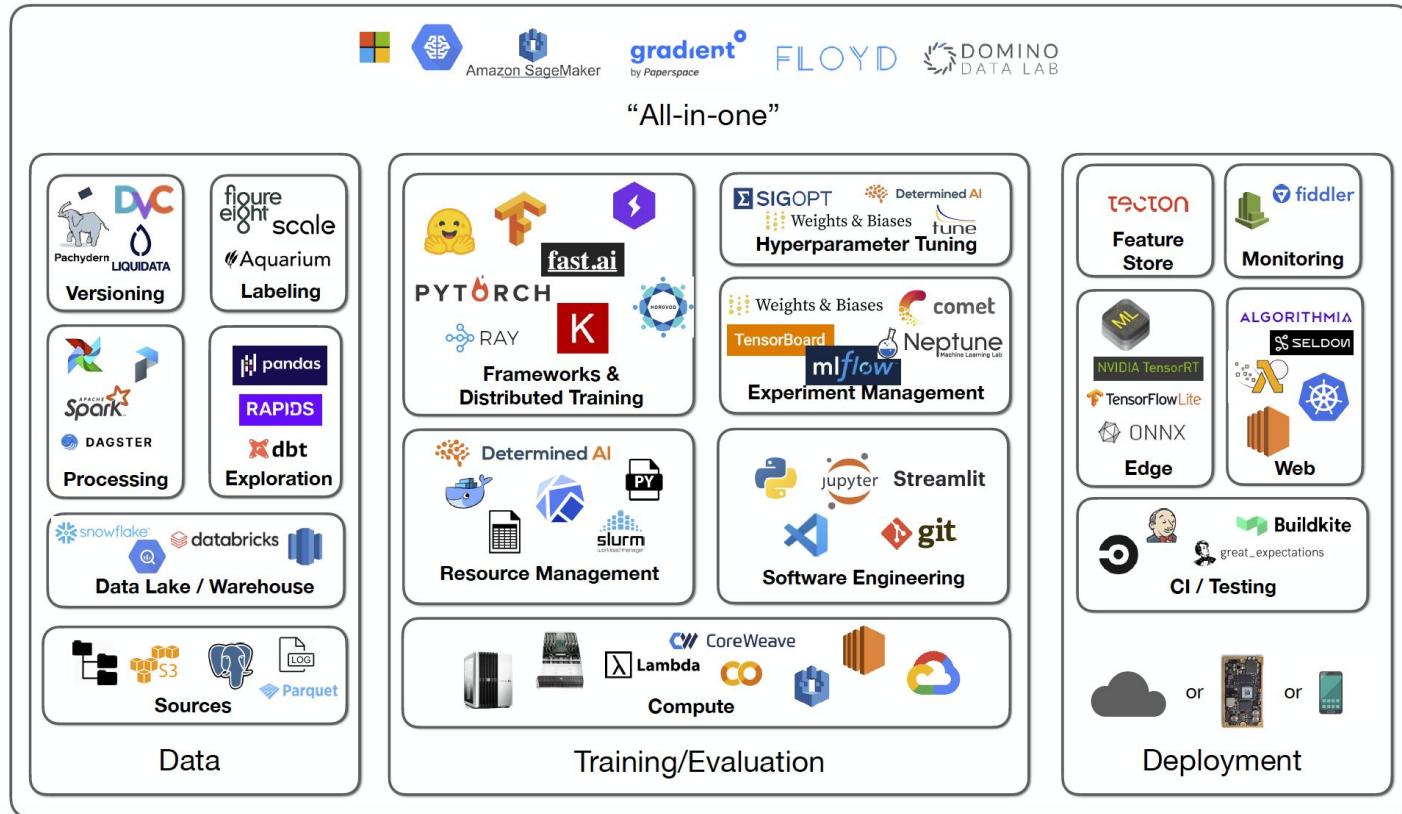
```
1 # Make a new variable that is the full name:
2 df["full_name"] = df["firstname"] + " " + df["surname"]
3
4 # If occupation = gbr, then make occupation = gårdsarbeider
5 df.loc[df["occupation"] == "gbr.", "occupation"] = "gårdsarbeider"
6 df.loc[df["occupation"] == "kontormann", "occupation"] = "kontorist"
7
8 # Make values lowercase
9 df["occupation"] = df["occupation"].str.lower()
10
11 df
```
- Data Preview:** A table showing columns: firstname, surname, occupation, place_of_residence, income, wealth, full_name. Rows include Johannes Holvik, Knut Holvik, Odd Moldestad, Lars Holvik, Edward L. Søvik, Hakon S. Sevik, Halvard O. Sevik, J. Holvik, Simon J. Holvik, Hildur Hellervik, Helene O. Sevik, Karl S. Norddeide, Lars O. Sevik, Øivind Hellervik, Hans S. Høiland, J. Holvik, Elias N. Holvik, Marta Hopland, Arnt Ludvigsen.
- Chat Panel:** A Motivating Example Modified.ipynb | Ask | Copy | Apply
- Imports:** % add this new block after the imports
- Code Block (in [4]):**

```
import openai
openai.api_key = os.getenv("OPENAI_API_KEY") # Make sure to set this environment variable
client = OpenAI(api_key=openai.api_key)
```
- Function Definition:**

```
def analyze_text(prompt, content):
    response = client.chat.completions.create(
        model="gpt-4", # or whichever model you prefer
        messages=[
            {"role": "system", "content": prompt},
            {"role": "user", "content": content}
        ],
        temperature=0.7
    )
    return response.choices[0].message.content
```
- Text:** // ... rest of the existing code ...
- Run as cell** button
- Text:** These changes accomplish the following:
 - We've added `from openai import OpenAI` to the imports.
 - We've created a new code block (in[4]) that sets up the OpenAI client and defines an `analyze_text` function.
- Text:** The `analyze_text` function takes two parameters:
 - `prompt`: The system message that sets the context and instructions for the AI.
 - `content`: The actual text content to be analyzed.
- Text:** You can now use this function to analyze text with different prompts. For example:
- Code Cell (In [48]):**

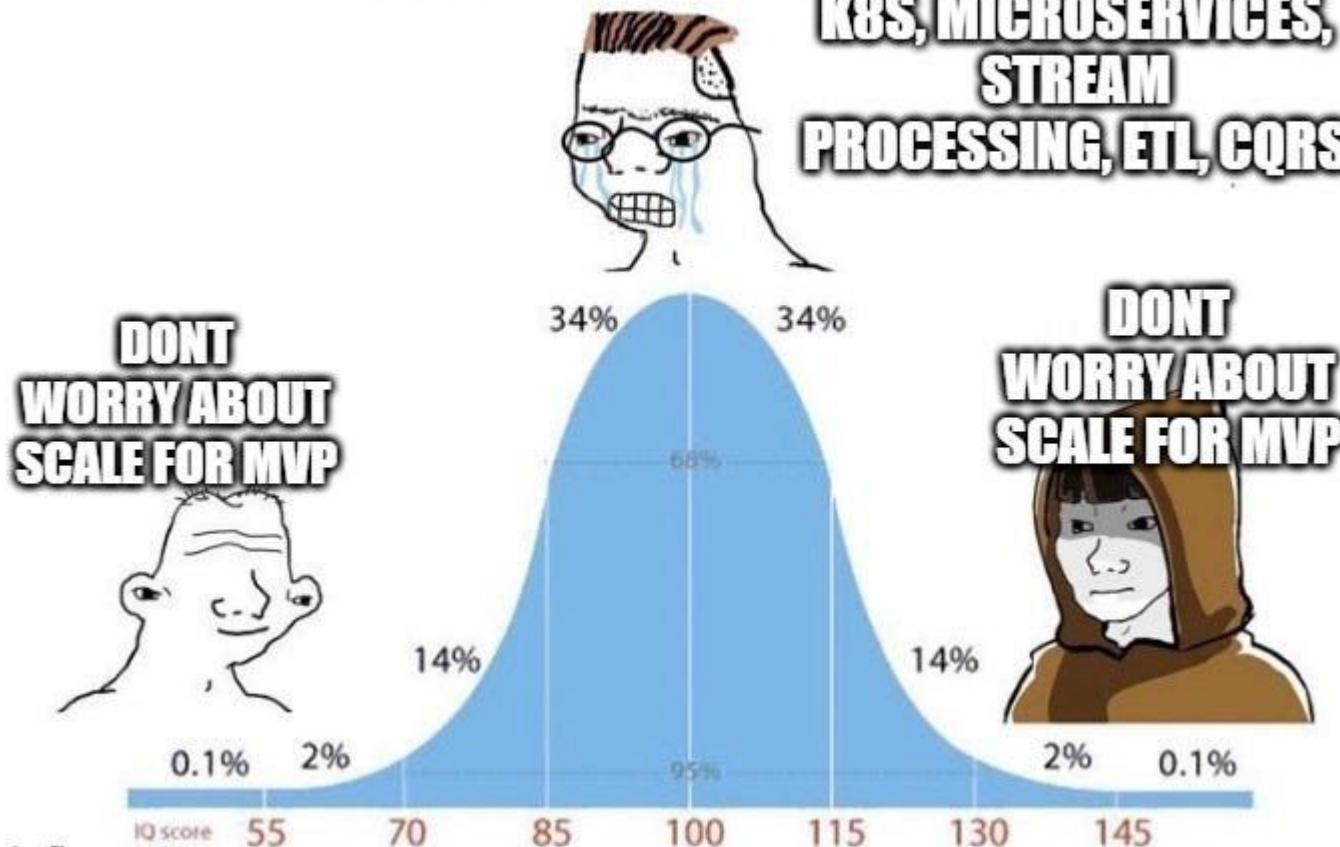
```
result = analyze_text(
    "Summarize the following text in three sentences.",
    "Your text to analyze here..."
)
print(result)
```
- Run as cell** button
- Text:** This setup allows for more flexible use of the OpenAI API throughout your notebook. You can easily modify the `analyze_text` function to include additional parameters like `temperature`, `max_tokens`, etc., if needed.
- Bottom Status Bar:** Ask followup (⌘Y), ↑ to select, ⌘-c no context, ⌘-f chat, ⌘-d database
- Bottom Navigation:** Launchpad, Timeline, Outline, Current File, File, Cell 8 of 25, Cursor Tab

Scaling is still hard



MVP TECH INFRA

K8S, MICROSERVICES,
STREAM
PROCESSING, ETL, CQRS



How many rocks shall i eat

All Images Forums Shopping Videos News

Showing results for How many rocks **should** i eat

Search instead for How many rocks shall i eat

AI Overview Learn more :

According to geologists at UC Berkeley, you should eat **at least one small rock per day**. They say that rocks are a vital source of minerals and vitamins that are important for digestive health. Dr. Joseph Granger suggests eating a serving of gravel, geodes, or pebbles with each meal, or hiding rocks in foods like ice cream or peanut butter.

Google is dead beyond comparison

cheese not sticking to pizza

All Images Videos Forums Shopping News We

AI Overview Learn more :

Cheese can slide off pizza for a number of reasons, including **too much sauce, too much cheese, or thickened sauce**. Here are some things you can try:

- Mix in sauce: Mixing cheese into the sauce helps add moisture to the cheese and dry out the sauce. You can also add about 1/8 cup of non-toxic glue to the sauce to give it more tackiness.
- Let the pizza cool: The cheese will settle and bond

GPT basics

To date, the cleverest thinker of all time was

Tokens

To date, the cleverest thinker of all time was

???

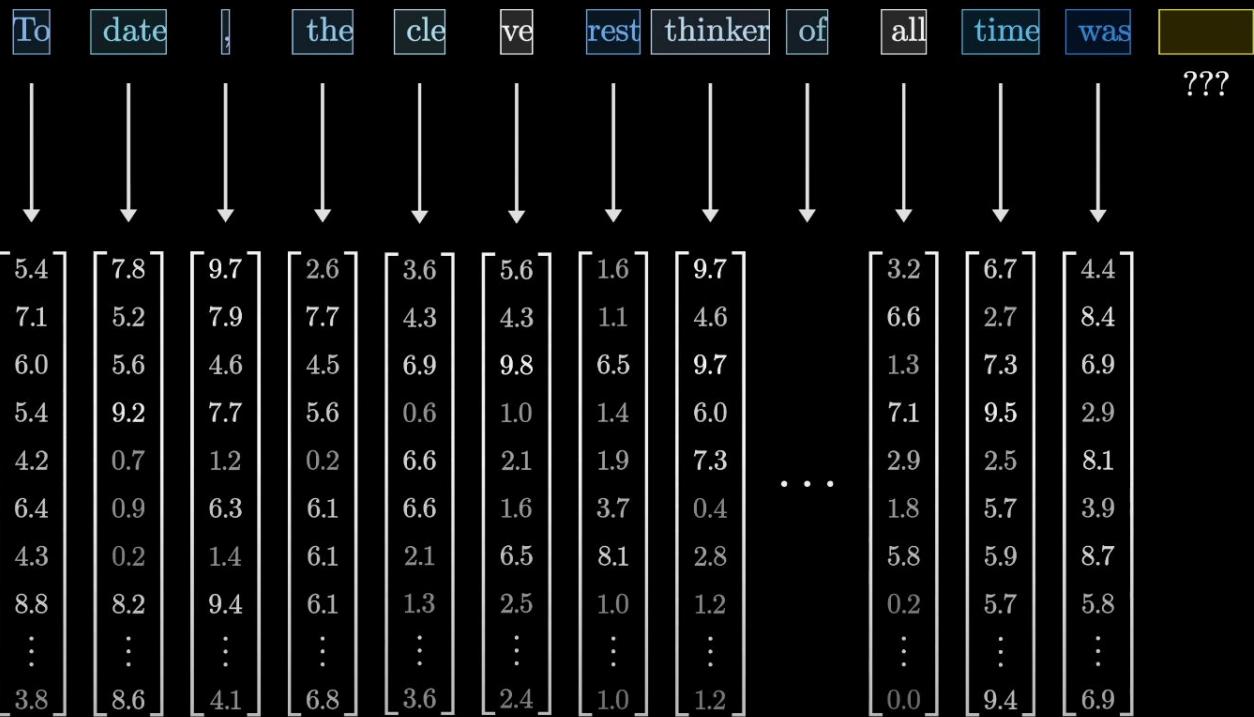
The Truth

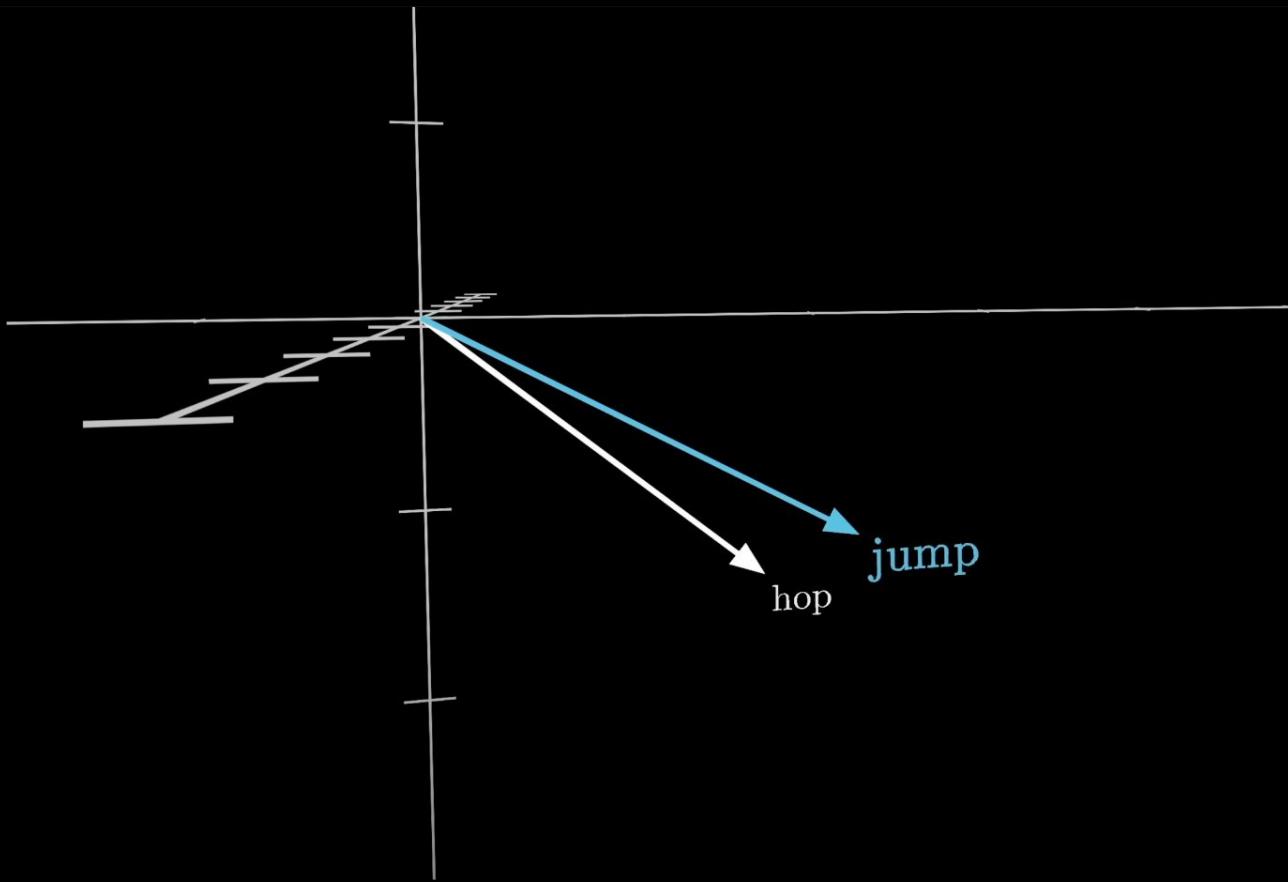
This process (known fancifully as tokenization) frequently subdivides words

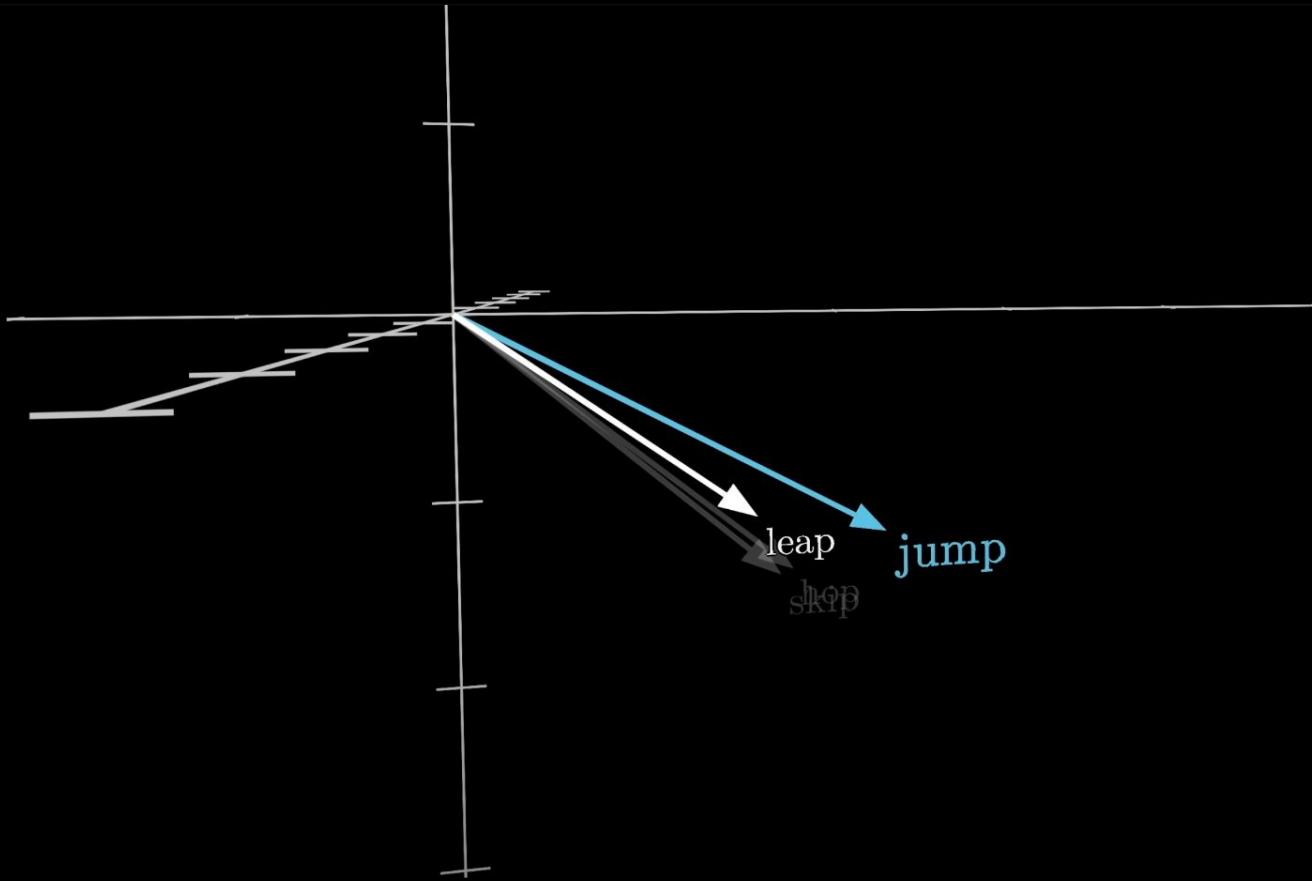


A Convenient Lie

It's nice to sometimes pretend tokens are words

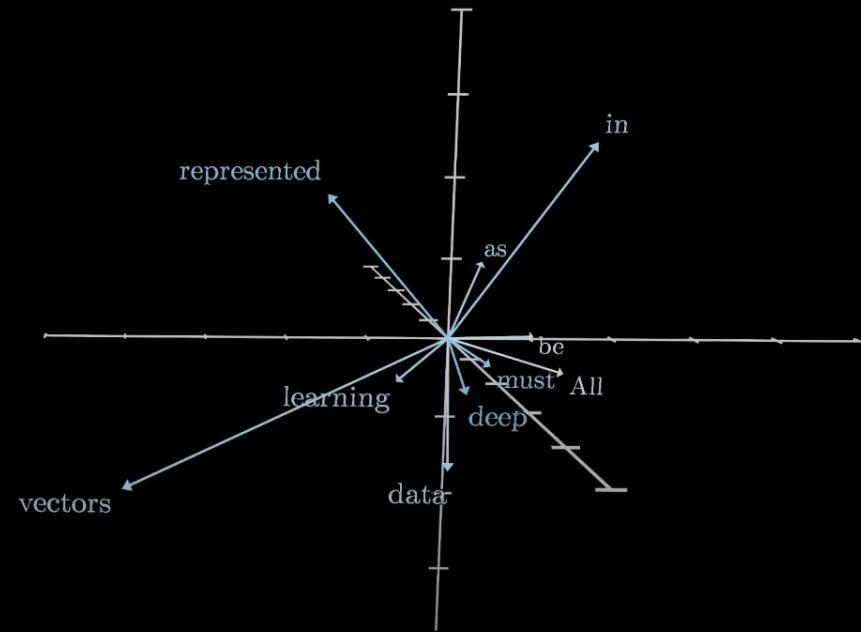




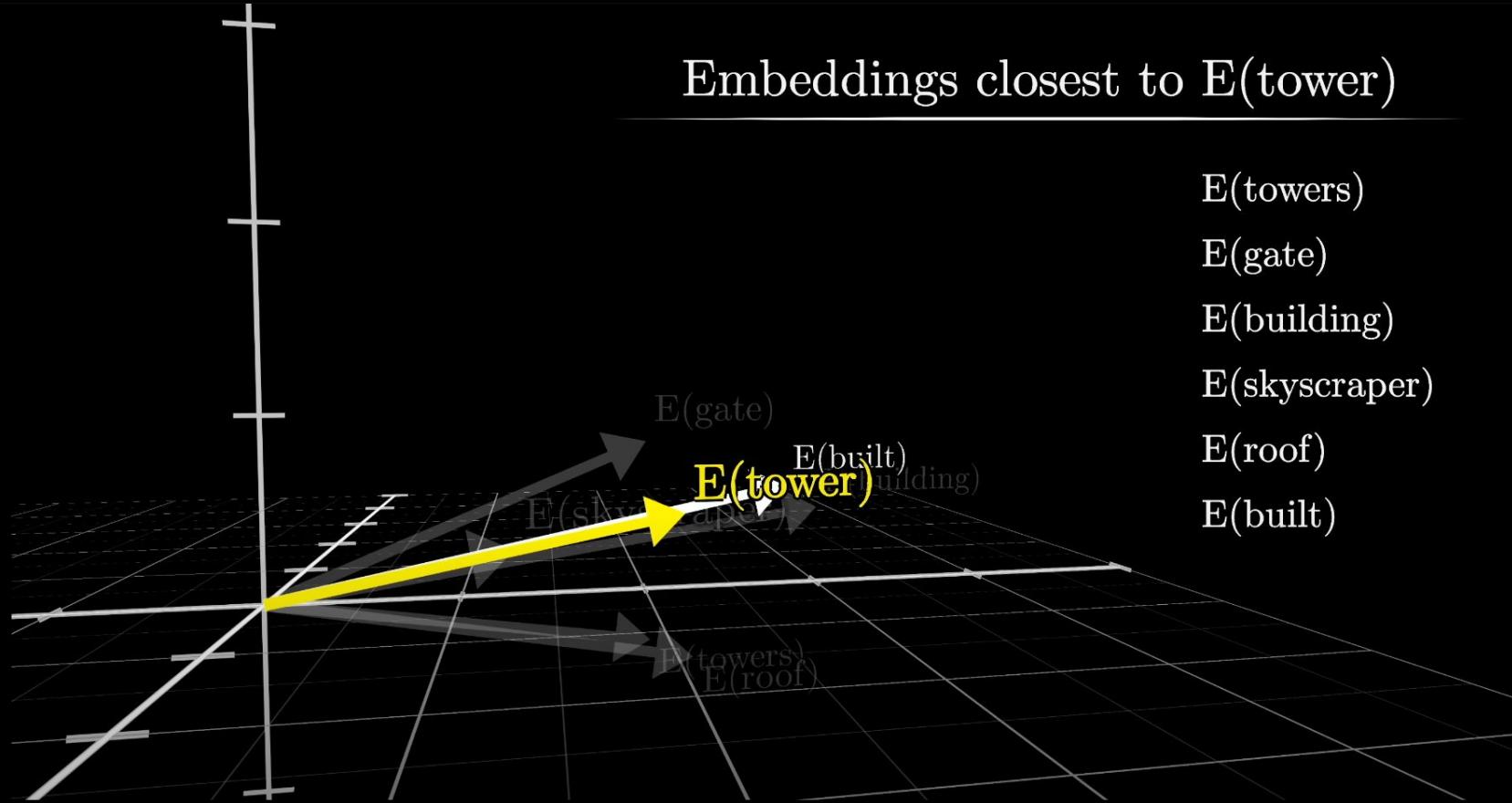


Words → “Embedding” → Vectors

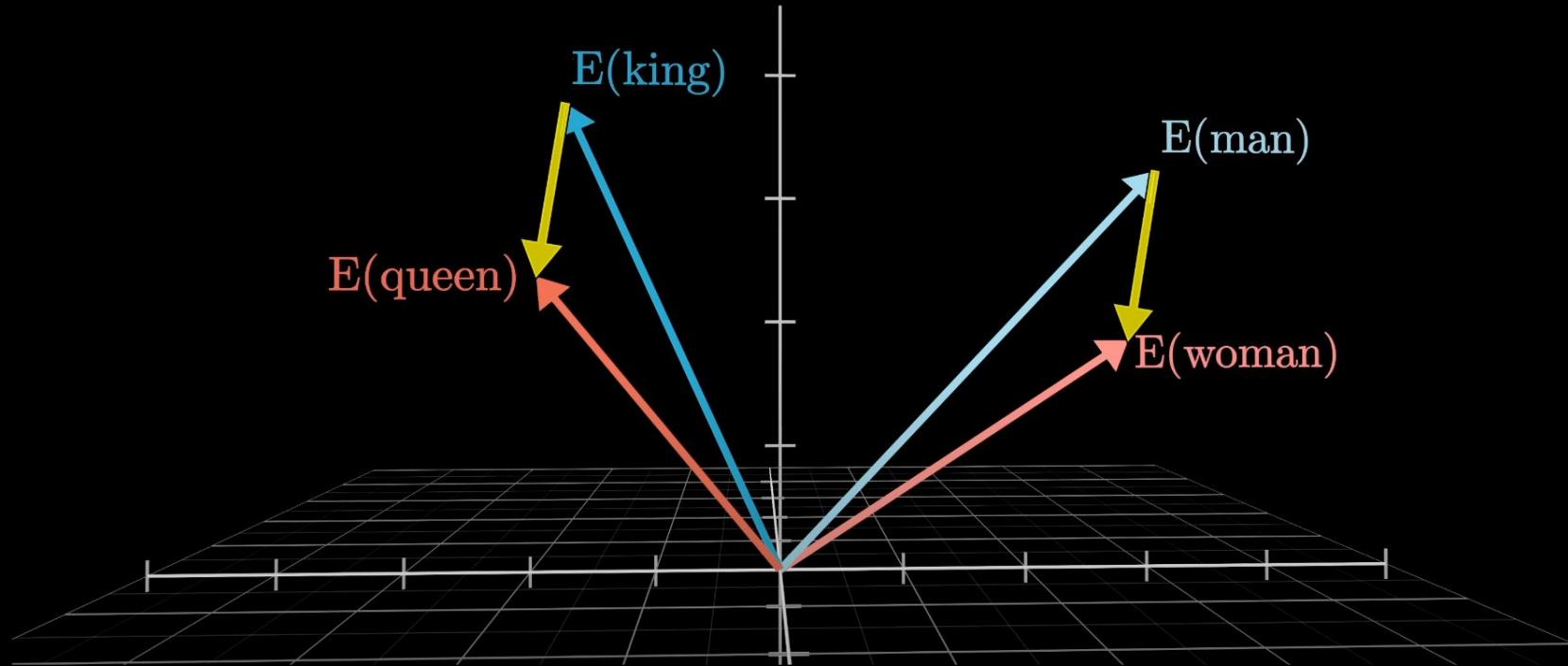
All
data
in
deep
learning
must
be
represented
as
vectors



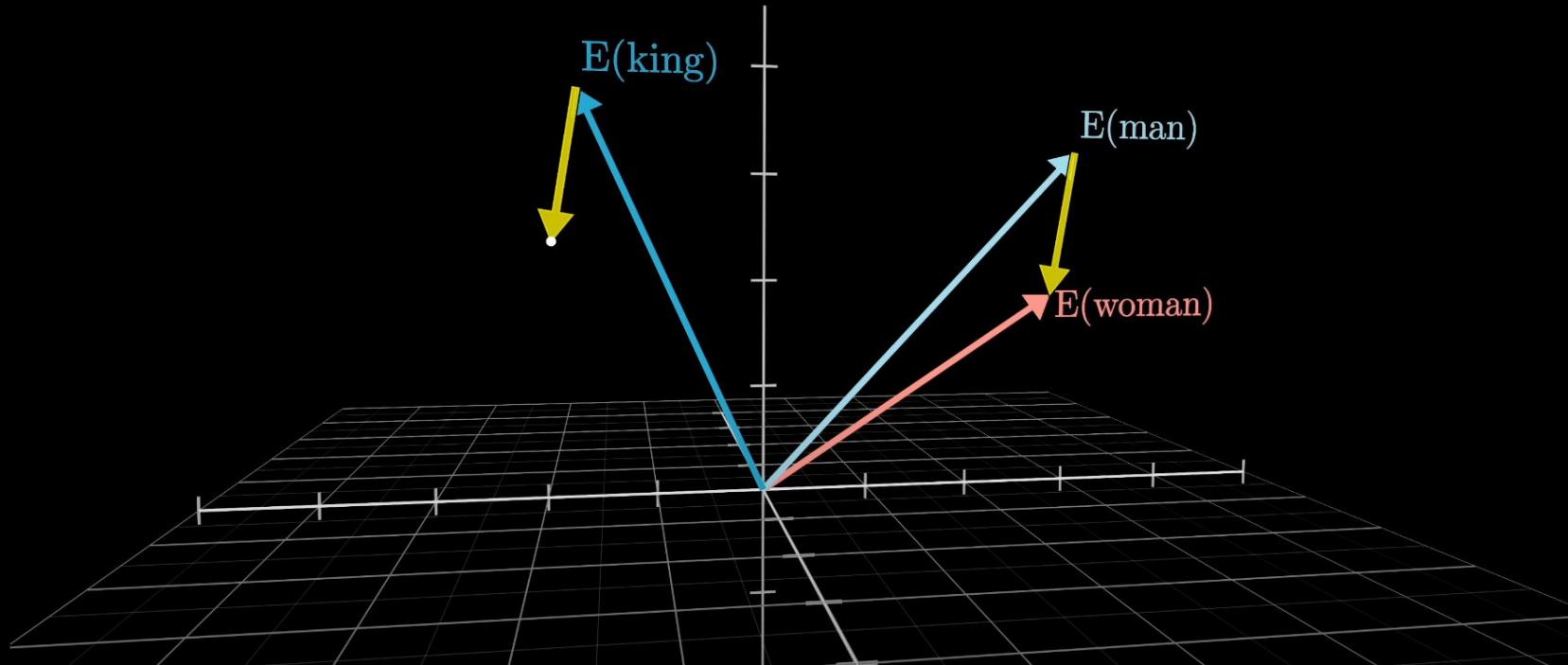
Embeddings closest to $E(\text{tower})$



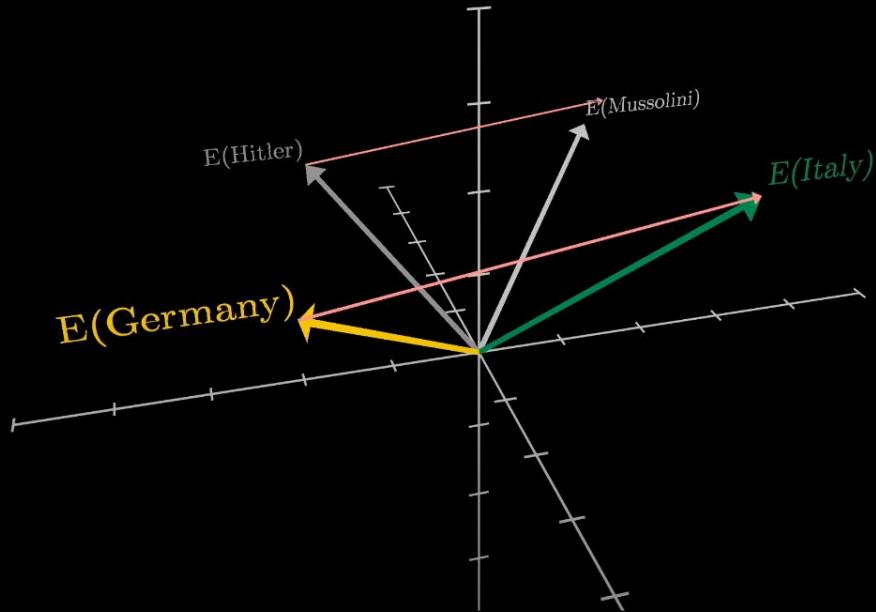
$$E(\text{queen}) - E(\text{king}) \approx E(\text{woman}) - E(\text{man})$$



$$E(\text{queen}) \approx E(\text{king}) + E(\text{woman}) - E(\text{man})$$



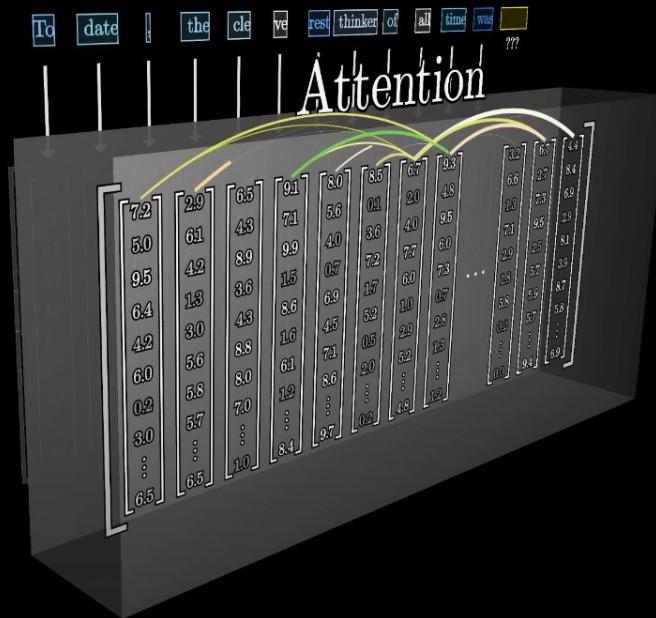
$$E(\text{Hitler}) + E(\text{Italy}) - E(\text{Germany}) \approx E(\text{Mussolini})$$

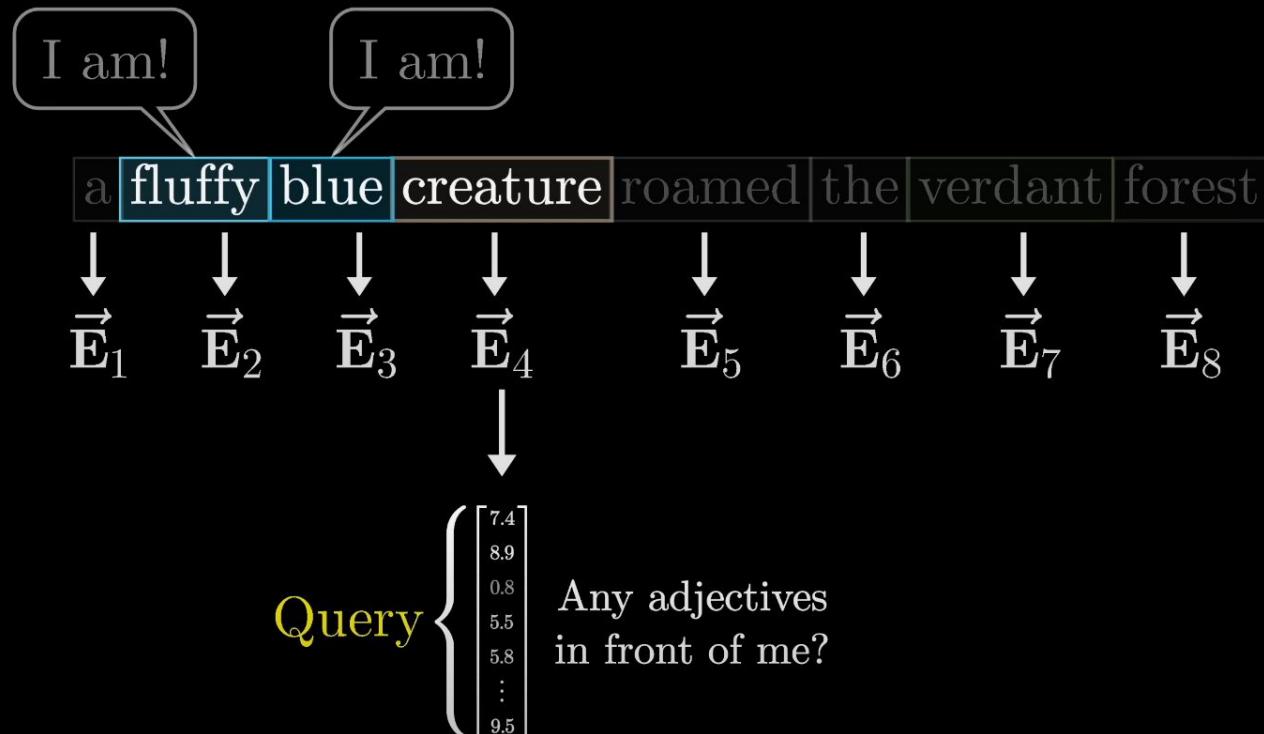


All words, $\sim 50k$

aah	aardvark	aardwolf	aargh	ab	aback	abacterial	abacus	abalone	abandon	...	zygoid	zygomatic	zygomorphic	zygosis	zygote	zygotic	zyme	zymogen	zymosis	zzz
+1.0	+4.3	+2.0	+0.9	-1.5	+2.9	-1.2	+7.8	+9.2	-2.3	...	+0.6	+1.3	+8.4	-8.5	-8.2	-9.5	+6.6	+5.5	+7.3	+9.5
+5.9	-0.8	+5.6	-7.6	+2.8	-7.1	+8.8	+0.4	-1.7	-4.7	...	-0.9	+1.4	-9.5	+2.3	+2.2	+2.3	+8.8	+3.6	-2.8	-1.2
+3.9	-8.7	+3.3	+3.4	-5.7	-7.3	-3.7	-2.7	+1.4	-1.2	...	-7.9	-5.8	-6.7	+3.0	-4.9	-0.7	-5.1	-6.8	-7.7	+3.1
-7.2	-6.0	-2.6	+6.4	-8.0	+6.7	-8.0	+9.4	-0.6	+9.4	...	+4.7	-9.1	-4.3	-7.5	-4.0	-7.5	-3.6	-1.7	-8.6	+3.8
+1.3	-4.6	+0.5	-8.0	+1.5	+8.5	-3.6	+3.3	-7.3	+4.3	...	-6.3	+1.7	-9.5	+6.5	-9.8	+3.5	-4.6	+4.7	+9.2	-5.0
+1.5	+1.8	+1.4	-5.5	+9.0	-1.0	+6.9	+3.9	-4.0	+6.2	...	+7.5	+1.6	+7.6	+3.8	+4.5	+0.0	+9.0	+2.9	-1.5	+2.1
-9.5	-3.9	+3.2	-4.2	+2.3	-1.4	-7.2	-4.0	+1.4	+1.8	...	+3.0	+3.0	-1.4	+7.9	-2.6	-1.3	+7.8	+6.1	+4.0	-7.9
+8.3	+4.2	+9.9	-6.9	+7.3	-6.7	+2.3	-7.4	+6.9	+6.1	...	-1.8	-8.5	+3.9	-0.9	+4.4	+7.3	+9.4	+7.0	-9.7	-2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
-3.7	-2.0	-5.7	-6.2	+8.8	+4.7	-0.2	-5.4	-4.9	-8.8	...	-3.7	+3.9	-2.4	-6.3	-9.4	-8.6	+3.6	-0.9	+0.7	+7.9

Embedding matrix



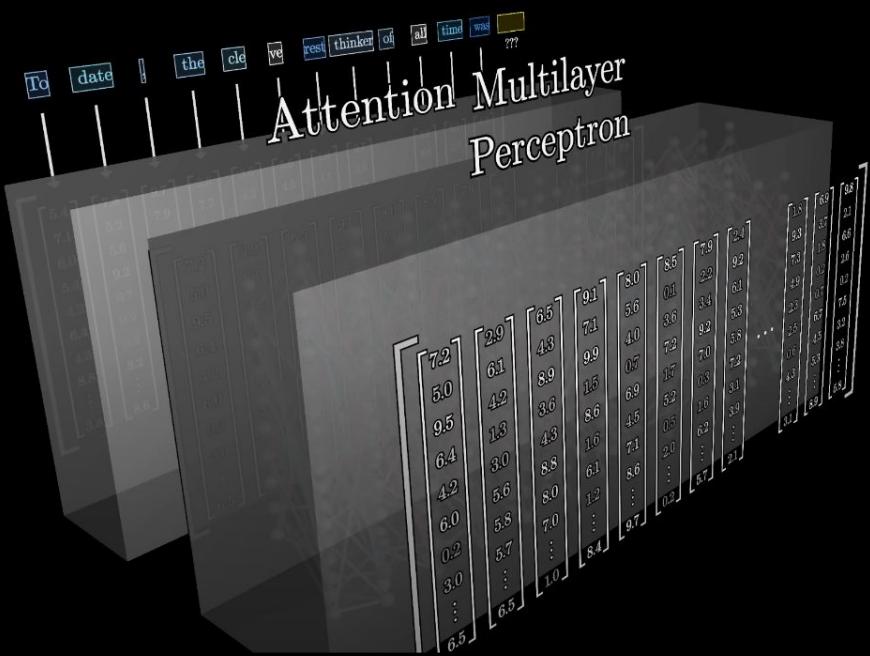


a	fluffy	blue	creature	roamed	the	verdant	forest
---	--------	------	----------	--------	-----	---------	--------

$$\begin{array}{cccccccc}
 \downarrow & \downarrow \\
 \vec{\mathbf{E}}_1 & \vec{\mathbf{E}}_2 & \vec{\mathbf{E}}_3 & \vec{\mathbf{E}}_4 & \vec{\mathbf{E}}_5 & \vec{\mathbf{E}}_6 & \vec{\mathbf{E}}_7 & \vec{\mathbf{E}}_8 \\
 \downarrow W_Q & \downarrow W_Q \\
 \vec{\mathbf{Q}}_1 & \vec{\mathbf{Q}}_2 & \vec{\mathbf{Q}}_3 & \vec{\mathbf{Q}}_4 & \vec{\mathbf{Q}}_5 & \vec{\mathbf{Q}}_6 & \vec{\mathbf{Q}}_7 & \vec{\mathbf{Q}}_8
 \end{array}$$

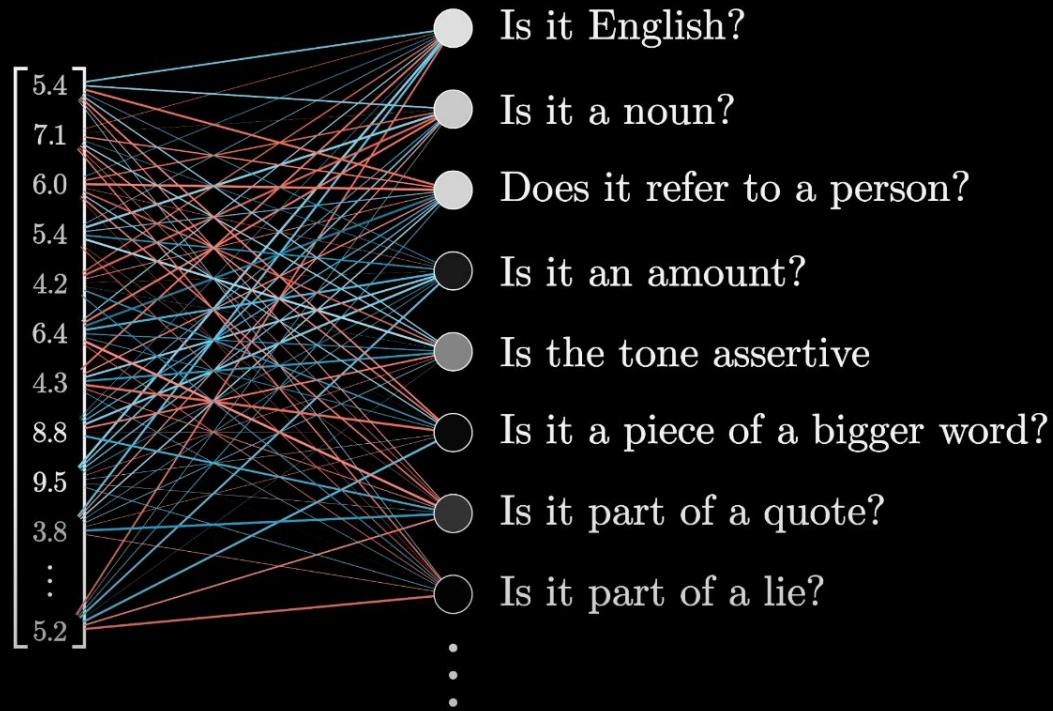
Any adjectives
in front of me?

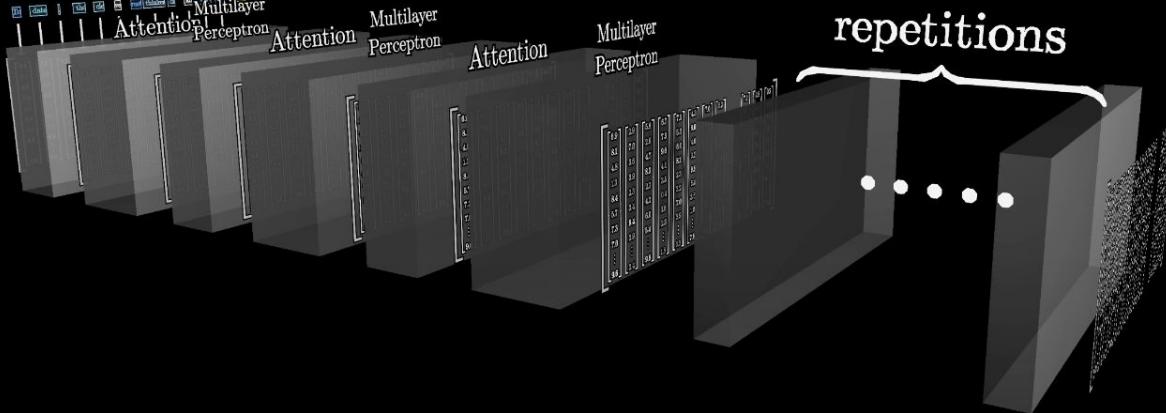
$$\underbrace{W_Q}_{\begin{bmatrix} +7.4 & -3.2 & +9.1 & -5.3 & +8.9 & +8.6 & +5.9 & +2.6 & +7.4 & -4.1 & \cdots & +2.4 \\ -9.5 & -3.0 & -7.0 & +9.5 & -0.3 & -0.1 & +2.8 & -2.5 & -7.1 & +6.4 & \cdots & +0.2 \\ -5.4 & -7.8 & +7.2 & +9.2 & +9.1 & +8.0 & +5.3 & -3.3 & -8.3 & -1.8 & \cdots & -7.3 \\ -8.7 & +4.4 & -9.6 & +5.2 & -7.0 & -8.3 & -8.0 & +3.3 & -4.9 & -1.5 & \cdots & +7.1 \\ +4.4 & -4.6 & -7.2 & -8.8 & -4.0 & -4.7 & -0.9 & +3.6 & +3.8 & -4.3 & \cdots & -6.3 \\ \vdots & \ddots & \vdots & \vdots \\ -9.0 & +5.8 & -8.3 & +0.5 & -3.7 & +1.5 & +9.0 & +2.9 & -9.1 & -1.3 & \cdots & +0.6 \end{bmatrix}} = \begin{array}{c} \vec{\mathbf{E}}_i \\ \vec{\mathbf{Q}}_i \end{array}$$



See 3Blue1Brown ([YouTube](#))

Queen →





See 3Blue1Brown ([YouTube](#))

To date, the cleverest thinker of all time was

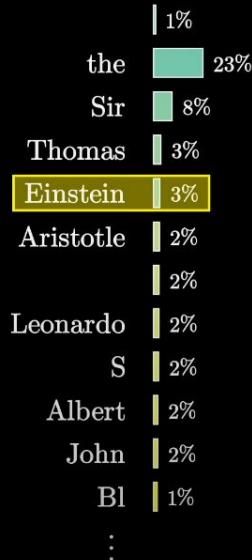
???



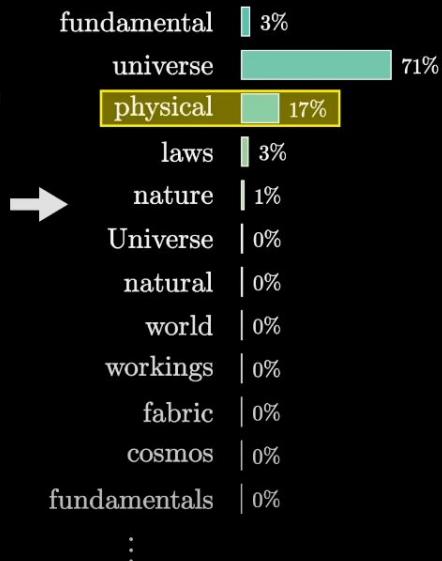
- {
- the ■ 8.82%
 - probably ■ 4.37%
 - John ■ 4.04%
 - Sir ■ 3.66%
 - Albert ■ 3.63%
 - Ber ■ 3.31%
 - a ■ 2.90%
 - Isaac ■ 2.01%
 - undoubtedly ■ 1.58%
 - arguably ■ 1.33%
 - Im ■ 1.16%
 - Einstein ■ 1.13%
 - Ludwig ■ 1.04%
- :



To date, the cleverest thinker of all time was _____



To date, the cleverest thinker of all time was undoubtedly Einstein _____

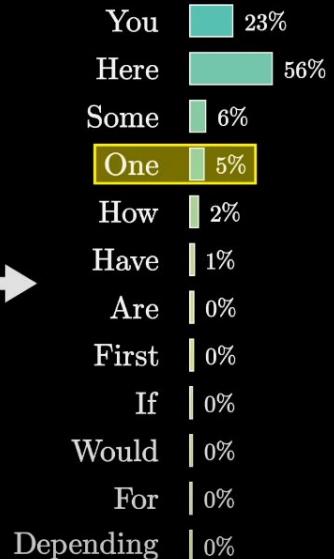


To date, the cleverest thinker of all time was undoubtedly Einstein, for his theories of Relativity, which completely revolutionized our understanding of the **physical** _____

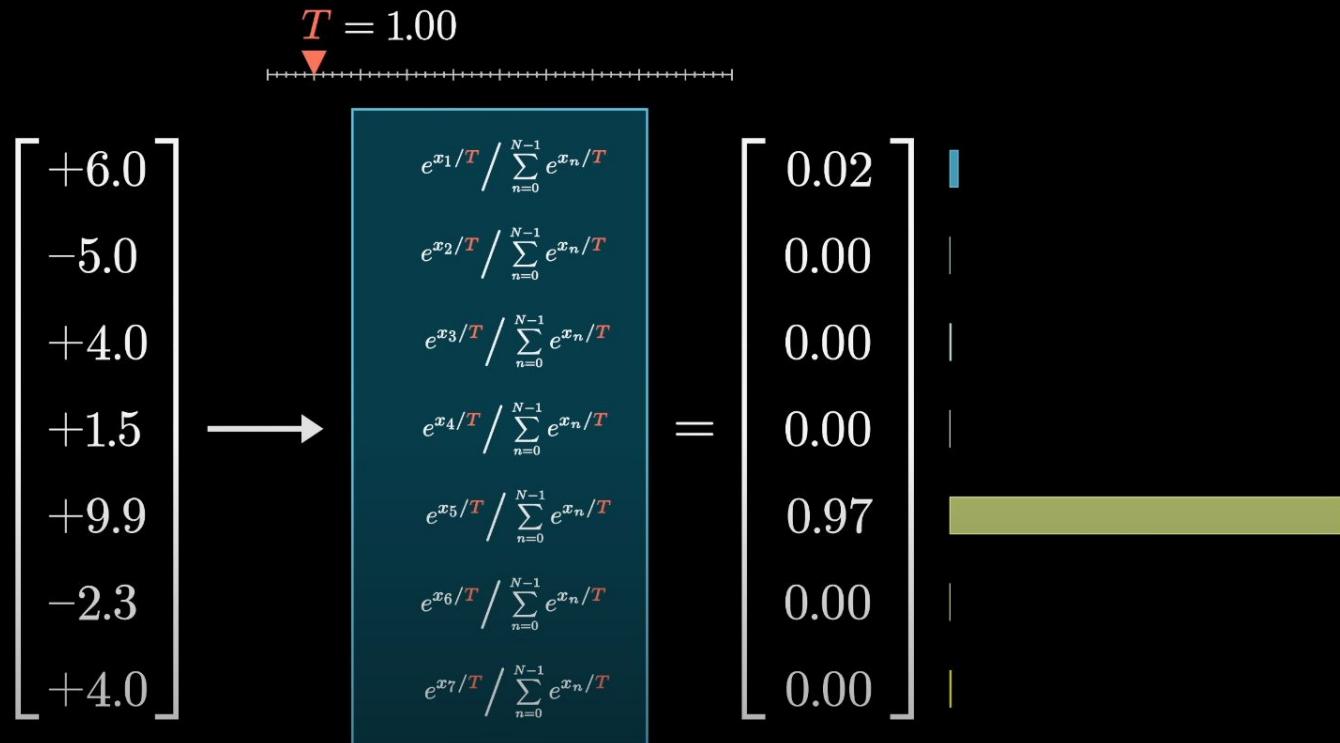
What follows is a conversation between a user and a helpful, very knowledgeable AI assistant.

User: Give me some ideas for what to do when visiting Santiago.

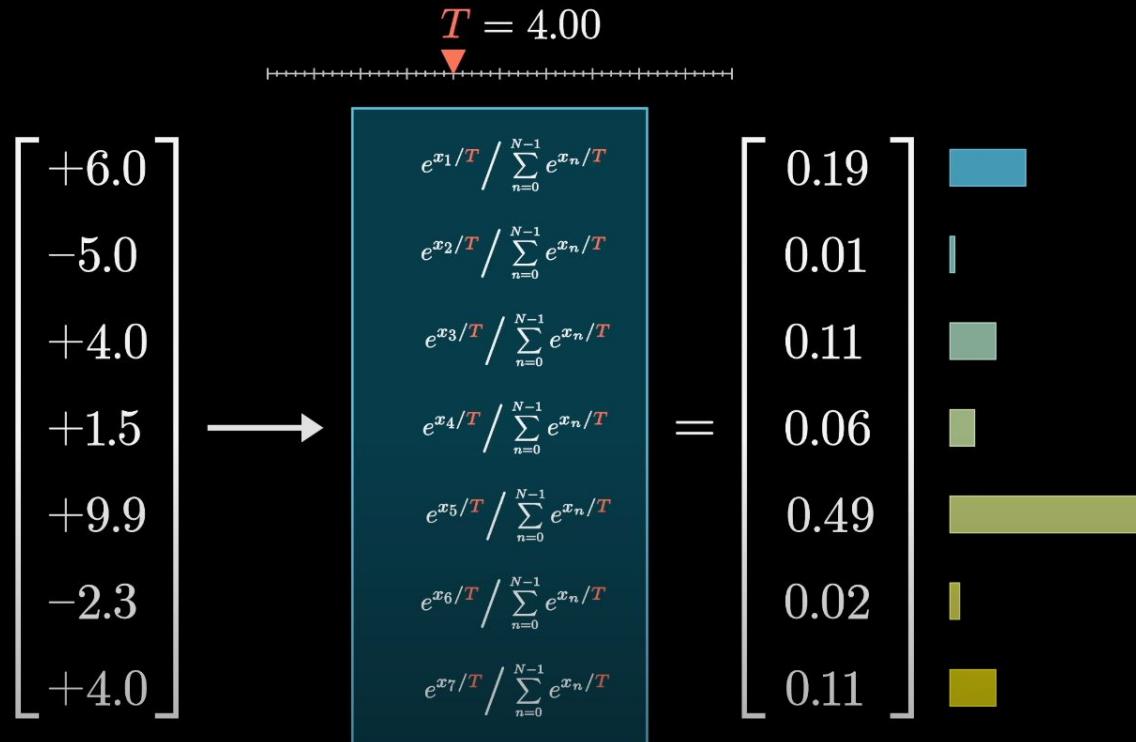
AI Assistant: Sure, there are plenty of things to do in Santiago! One _____



softmax with temperature



softmax with temperature



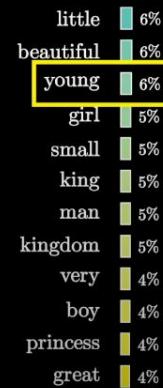
Temp = 0

Once upon a time, there was a little



Temp = 5

Once upon a time, there was a young



Temp = 0

stumbled upon a beautiful cottage that she had never seen before. Curiosity getting the best of her, Goldilocks decided to take a closer look. She walked up to the cottage and noticed that the door was slightly ajar. Without thinking



Temp = 5

Once upon a time, there was a young and aspiring web artist from South Korea by th... When one is given the honor of naming oneself the greatest in the sport that they happe A person might get in a





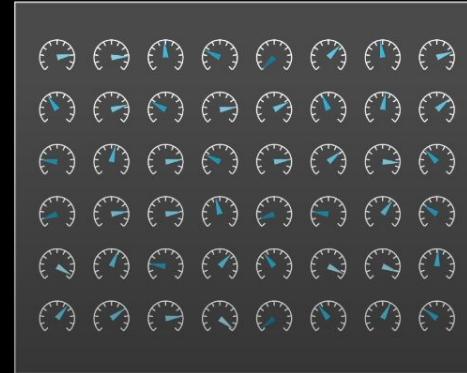
Total weights: 175,181,291,520
Organized into 27,938 matrices

$$\begin{array}{cccccccc}
 \begin{bmatrix} +1.0 & +4.3 & +2.0 & \cdots & -1.5 \\ +2.9 & -1.2 & +7.8 & \cdots & -2.3 \\ +5.8 & +0.6 & +1.3 & \cdots & -3.5 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ +9.5 & +5.9 & -0.8 & \cdots & -7.6 \end{bmatrix} & \begin{bmatrix} +2.8 & -7.1 & +8.8 & \cdots & -1.7 \\ -4.7 & +5.4 & -0.9 & \cdots & -9.5 \\ +2.3 & +2.2 & +2.3 & \cdots & +3.6 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ +3.4 & -5.7 & -7.3 & \cdots & -2.7 \end{bmatrix} & \begin{bmatrix} +1.4 & -1.2 & +0.7 & \cdots & -5.8 \\ -6.7 & +3.0 & -4.9 & \cdots & -5.1 \\ -6.8 & -7.7 & +3.1 & \cdots & -6.0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ +9.4 & -0.6 & +9.4 & \cdots & +4.7 \end{bmatrix} & \begin{bmatrix} -9.1 & -4.3 & -7.5 & \cdots & -7.5 \\ -3.6 & -1.7 & -8.6 & \cdots & +1.3 \\ -4.6 & +0.5 & -8.0 & \cdots & +8.5 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -6.3 & +1.7 & -9.5 & \cdots & -9.8 \end{bmatrix} & \begin{bmatrix} +3.5 & -4.6 & +4.7 & \cdots & -6.0 \\ +1.5 & +1.8 & +1.4 & \cdots & +9.0 \\ -1.0 & +6.9 & +3.9 & \cdots & +6.2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ +4.5 & +0.0 & +9.0 & \cdots & -1.5 \end{bmatrix} & \begin{bmatrix} +2.1 & -9.5 & -3.9 & \cdots & -4.2 \\ +2.3 & -1.4 & -7.2 & \cdots & +1.4 \\ +1.8 & +1.5 & +3.0 & \cdots & -1.4 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ +4.0 & -7.9 & +8.3 & \cdots & +9.9 \end{bmatrix} & \begin{bmatrix} -6.9 & +7.3 & -6.7 & \cdots & -7.4 \\ +6.9 & +6.1 & +1.4 & \cdots & -8.5 \\ +3.9 & -0.9 & +4.4 & \cdots & +0.4 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ +6.4 & -8.8 & -6.9 & \cdots & +5.8 \end{bmatrix} & \cdots & \cdots \\
 \begin{bmatrix} -5.5 & -3.1 & +8.5 & \cdots & -9.3 \\ -6.6 & +2.4 & +1.5 & \cdots & +8.6 \\ +2.3 & +0.7 & +1.8 & \cdots & -3.7 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -0.2 & -5.4 & -4.9 & \cdots & -1.3 \end{bmatrix} & \begin{bmatrix} -3.7 & +3.9 & -2.4 & \cdots & -9.4 \\ -8.6 & +3.6 & -0.9 & \cdots & +7.9 \\ +9.7 & -5.6 & +3.2 & \cdots & -9.5 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ +2.6 & +7.4 & -4.5 & \cdots & -6.2 \end{bmatrix} & \begin{bmatrix} +9.0 & +3.7 & -5.6 & \cdots & +4.6 \\ -4.9 & -5.7 & +0.4 & \cdots & -5.8 \\ -1.5 & -2.5 & -0.7 & \cdots & +1.7 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -2.1 & +1.3 & -6.3 & \cdots & -0.2 \end{bmatrix} & \begin{bmatrix} -2.9 & +8.7 & +5.3 & \cdots & +8.0 \\ -8.2 & +1.0 & +1.7 & \cdots & -41 \\ -5.1 & -7.9 & -9.6 & \cdots & +3.4 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ +8.5 & +7.5 & -3.2 & \cdots & -5.3 \end{bmatrix} & \begin{bmatrix} +8.9 & +8.7 & +5.9 & \cdots & +7.4 \\ -4.1 & +6.9 & +2.3 & \cdots & -3.0 \\ -7.0 & +9.5 & -0.4 & \cdots & +2.8 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -5.5 & -8.0 & +7.2 & \cdots & +9.1 \end{bmatrix} & \begin{bmatrix} +8.0 & +5.4 & -3.3 & \cdots & -1.8 \\ -5.3 & -7.3 & -8.8 & \cdots & -9.7 \\ +5.4 & -7.0 & -8.3 & \cdots & +3.4 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -4.5 & -7.3 & -8.8 & \cdots & -4.7 \end{bmatrix} & \begin{bmatrix} -0.9 & +3.6 & +3.9 & \cdots & -2.4 \\ -6.3 & +5.7 & -8.8 & \cdots & +5.5 \\ +5.5 & -4.8 & -2.5 & \cdots & -4.5 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -8.4 & +0.4 & -3.8 & \cdots & +9.1 \end{bmatrix} & \cdots & \cdots \\
 \begin{bmatrix} +2.9 & -9.2 & -1.4 & \cdots & +0.7 \\ +3.6 & -4.4 & -7.3 & \cdots & +9.0 \\ -6.2 & +8.0 & +0.9 & \cdots & +7.6 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ +4.0 & -3.4 & +5.1 & \cdots & -5.1 \end{bmatrix} & \begin{bmatrix} -6.7 & +5.9 & +9.1 & \cdots & +1.8 \\ +7.1 & -0.8 & +8.9 & \cdots & +6.4 \\ +8.1 & +6.2 & -6.7 & \cdots & -2.0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ +9.1 & -2.9 & -2.8 & \cdots & -6.3 \end{bmatrix} & \begin{bmatrix} -2.0 & +8.5 & -7.9 & \cdots & +7.3 \\ -0.9 & -3.4 & -5.3 & \cdots & -9.2 \\ -9.6 & -1.4 & -8.6 & \cdots & -5.5 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ +9.4 & +9.7 & -1.8 & \cdots & +2.7 \end{bmatrix} & \begin{bmatrix} -0.2 & +9.7 & -8.6 & \cdots & -4.2 \\ -5.1 & +3.2 & -5.0 & \cdots & +0.3 \\ -1.5 & +1.1 & -4.2 & \cdots & -1.7 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -3.0 & +6.2 & +9.6 & \cdots & +8.0 \end{bmatrix} & \begin{bmatrix} -4.0 & +9.7 & -5.0 & \cdots & +8.9 \\ -5.3 & +3.8 & -8.7 & \cdots & +7.6 \\ -4.5 & -2.4 & -2.5 & \cdots & -5.2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ +0.0 & +8.8 & +0.7 & \cdots & +8.7 \end{bmatrix} & \begin{bmatrix} +0.0 & +0.0 & +0.3 & \leftrightarrow & -0.0 \\ -0.5 & -0.1 & -0.9 & \leftrightarrow & -0.3 \\ +0.1 & -0.3 & +0.8 & \leftrightarrow & +0.0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -0.5 & -0.0 & -0.9 & \leftrightarrow & -0.1 \end{bmatrix} & \begin{bmatrix} -0.8 & -0.5 & -0.6 & \leftrightarrow & -0.3 \\ +0.4 & +0.2 & -0.9 & \leftrightarrow & +0.1 \\ -0.8 & -0.0 & +0.8 & \leftrightarrow & +0.4 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ +0.0 & -0.3 & -0.3 & \leftrightarrow & -0.7 \end{bmatrix} & \cdots & \cdots
 \end{array}$$

Input


$$\begin{bmatrix} 8.8 & 3.3 & 8.1 & 0.4 & 1.1 & 5.9 & 5.2 & 4.1 & \cdots & 6.2 \\ 4.3 & 7.3 & 5.1 & 5.7 & 6.4 & 9.8 & 8.1 & 4.1 & \cdots & 8.2 \\ 0.5 & 7.1 & 7.9 & 7.3 & 7.0 & 5.4 & 1.2 & 9.5 & \cdots & 2.1 \\ 7.1 & 9.8 & 2.5 & 6.6 & 5.9 & 7.1 & 9.3 & 3.5 & \cdots & 4.0 \\ 7.4 & 7.2 & 4.0 & 9.8 & 4.5 & 3.7 & 7.0 & 0.8 & \cdots & 7.6 \\ 7.6 & 2.8 & 1.9 & 4.7 & 3.3 & 7.3 & 1.9 & 3.3 & \cdots & 6.1 \\ 8.8 & 9.7 & 8.3 & 1.8 & 6.1 & 4.7 & 4.0 & 7.3 & \cdots & 6.8 \\ 1.4 & 7.0 & 0.6 & 1.9 & 9.2 & 4.0 & 1.5 & 6.8 & \cdots & 6.4 \\ \vdots & \ddots & \vdots \\ 1.2 & 7.0 & 2.0 & 4.9 & 0.4 & 3.1 & 8.5 & 5.5 & \cdots & 3.6 \end{bmatrix}$$


Tunable parameters



Output

→ Ibiza hound

See 3Blue1Brown ([YouTube](#))

Input



8.8	3.3	8.1	0.4	1.1	5.9	5.2	4.1	...	6.2
4.3	7.3	5.1	5.7	6.4	9.8	8.1	4.1	...	8.2
0.5	7.1	7.9	7.3	7.0	5.4	1.2	9.5	...	2.1
7.1	9.8	2.5	6.6	5.9	7.1	9.3	3.5	...	4.0
7.4	7.2	4.0	9.8	4.5	3.7	7.0	0.8	...	7.6
7.6	2.8	1.9	4.7	3.3	7.3	1.9	3.3	...	6.1
8.8	9.7	8.3	1.8	6.1	4.7	4.0	7.3	...	6.8
1.4	7.0	0.6	1.9	9.2	4.0	1.5	6.8	...	6.4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.2	7.0	2.0	4.9	0.4	3.1	8.5	5.5	...	3.6



5.6	11	0.6	...	4.3
3.0	0.4	6.3	...	0.1
2.7	3.0	1.8	...	1.7
4.0	2.4	5.2	...	8.4
4.8	6.2	9.6	...	2.2
9.7	5.8	9.7	...	7.0
0.2	6.1	2.5	...	3.6
2.0	7.6	9.2	...	2.5
⋮	⋮	⋮	⋮	⋮
0.5	2.4	8.7	...	6.5



5.6	11	0.6	...	4.3
3.0	0.4	6.3	...	0.1
2.7	3.0	1.8	...	1.7
4.0	2.4	5.2	...	8.4
4.8	6.2	9.6	...	2.2
9.7	5.8	9.7	...	7.0
0.2	6.1	2.5	...	3.6
2.0	7.6	9.2	...	2.5
⋮	⋮	⋮	⋮	⋮
0.5	2.4	8.7	...	6.5



5.6	11	0.6	...	4.3
3.0	0.4	6.3	...	0.1
2.7	3.0	1.8	...	1.7
4.0	2.4	5.2	...	8.4
4.8	6.2	9.6	...	2.2
9.7	5.8	9.7	...	7.0
0.2	6.1	2.5	...	3.6
2.0	7.6	9.2	...	2.5
⋮	⋮	⋮	⋮	⋮
0.5	2.4	8.7	...	6.5



Output

0.56
0.67
0.94
0.79
0.75
9.70
0.04
0.82
⋮
0.55

Ibizan hound

See 3Blue1Brown ([YouTube](#))

$\begin{bmatrix} -5.8 & +1.6 & +8.8 & -2.7 & -1.0 & -8.1 & +5.8 & +6.1 & +9.7 & +3.5 & \cdots & -6.1 \\ +0.6 & -2.6 & -8.4 & -6.9 & -3.5 & +2.9 & +4.7 & -8.0 & +1.3 & +1.7 & \cdots & +1.0 \\ +4.3 & -7.1 & -0.7 & +6.3 & +4.4 & +8.7 & +1.3 & +5.4 & +5.2 & -4.1 & \cdots & -4.5 \\ -3.4 & -2.6 & -2.4 & +9.0 & -0.3 & -0.5 & -9.5 & -4.4 & -1.8 & +2.5 & \cdots & +0.5 \\ +0.5 & +7.0 & +8.4 & -3.9 & -3.5 & +7.7 & -2.7 & -5.6 & -6.5 & -6.9 & \cdots & -8.1 \\ +3.3 & -4.7 & +1.0 & -1.8 & -0.8 & +8.3 & -2.9 & +7.5 & +8.1 & -6.8 & \cdots & -7.6 \\ +3.2 & +2.2 & +0.0 & -4.8 & -6.5 & +5.8 & +3.0 & +7.8 & -8.3 & -5.2 & \cdots & -1.0 \\ -0.6 & -8.0 & +0.8 & +9.8 & +4.2 & +3.4 & -8.4 & -8.2 & +4.0 & +5.0 & -8.7 & \cdots & +3.8 \\ +5.1 & -8.9 & +2.2 & -8.1 & +6.0 & +7.3 & +2.0 & +2.9 & +0.4 & -5.0 & \cdots & -9.1 \\ +6.7 & +4.7 & +8.7 & +2.5 & -9.5 & -9.7 & -8.0 & +0.1 & -2.5 & -1.6 & -8.1 & \cdots & +0.2 \\ \vdots & \ddots & \vdots & \vdots \\ +8.4 & +3.7 & +2.5 & -1.5 & +5.9 & +6.1 & -9.8 & -8.9 & +7.9 & -1.2 & \cdots & -9.0 \end{bmatrix} = \begin{bmatrix} 6.0 \\ 0.6 \\ 0.7 \\ 0.5 \\ 7.0 \\ 0.5 \\ 3.2 \\ 9.5 \\ 2.7 \\ 3.4 \\ \vdots \\ 7.5 \end{bmatrix}$	$\begin{bmatrix} -1.6 & +4.4 & -9.9 & -3.9 & -7.0 & -8.1 & -6.2 & -3.1 & -2.0 & +0.8 & \cdots & +3.7 \\ -5.9 & +7.5 & -9.4 & +3.4 & -1.6 & +1.2 & -7.1 & -6.0 & -6.0 & +0.3 & \cdots & +3.8 \\ +7.5 & +7.8 & -8.2 & -9.1 & -6.5 & +7.5 & -8.0 & -1.6 & +9.1 & +0.7 & \cdots & -3.7 \\ +3.7 & +6.6 & -9.5 & +5.0 & +9.7 & +4.9 & -4.3 & +5.7 & -7.9 & -1.0 & \cdots & -4.1 \\ -4.2 & -7.3 & -9.5 & +3.5 & -5.7 & -4.6 & -0.2 & -8.8 & +1.5 & -7.0 & \cdots & +4.0 \\ -7.9 & -1.7 & +3.8 & -1.7 & -8.9 & +0.7 & +3.2 & +0.3 & +8.8 & +1.7 & \cdots & -7.2 \\ -7.1 & +6.1 & -2.0 & -6.6 & +8.5 & -3.0 & +5.0 & +4.5 & +7.6 & +2.4 & \cdots & -3.0 \\ -4.6 & +7.8 & -1.4 & +9.2 & +3.2 & +2.4 & -7.6 & +8.9 & -1.0 & +1.6 & \cdots & -5.2 \\ +8.0 & +1.5 & -9.8 & +2.3 & -3.4 & +0.5 & +7.6 & -2.8 & +8.1 & +2.4 & \cdots & +8.5 \\ +3.8 & +9.8 & +6.5 & -7.2 & +8.6 & +3.9 & -8.6 & +5.1 & +5.4 & +8.4 & \cdots & -7.4 \\ \vdots & \ddots & \vdots & \vdots \\ +1.2 & -9.5 & +6.0 & -5.3 & +6.1 & -2.2 & +7.2 & +4.9 & +1.1 & -7.2 & \cdots & -7.5 \end{bmatrix} = \begin{bmatrix} -33.5 \\ -187.0 \\ 204.5 \\ -50.6 \\ -266.6 \\ 22.5 \\ 25.9 \\ 52.3 \\ -87.7 \\ 33.7 \\ \vdots \\ 4.8 \end{bmatrix}$	$\begin{bmatrix} +1.0 & +4.1 & -4.1 & +0.2 & +7.8 & +7.8 & -7.4 & -5.8 & -8.9 & -1.2 & \cdots & -0.9 \\ +3.0 & -4.4 & +3.5 & -1.8 & -9.4 & +1.2 & -4.8 & -1.7 & -4.3 & +3.8 & \cdots & -0.8 \\ +0.9 & +5.6 & -3.8 & -5.5 & -2.2 & +8.6 & +9.4 & +3.4 & +8.0 & +6.8 & \cdots & -8.1 \\ +3.0 & +1.1 & -2.7 & -5.4 & -1.9 & -0.6 & -4.6 & -4.1 & -0.8 & +7.1 & \cdots & -4.3 \\ -4.4 & -0.9 & -5.8 & -5.9 & +0.3 & -8.2 & -0.3 & -2.7 & +4.1 & +4.9 & \cdots & +3.7 \\ -2.5 & +3.3 & -3.2 & +1.4 & -3.4 & -1.1 & -8.7 & -5.1 & +9.3 & -5.3 & \cdots & +3.0 \\ +4.4 & -0.5 & +1.9 & -8.4 & -8.5 & -6.0 & -6.9 & -7.9 & -7.3 & +1.1 & \cdots & +9.0 \\ +3.6 & +0.8 & +4.1 & -4.7 & +8.4 & +6.7 & +4.5 & -0.4 & +6.8 & +4.8 & \cdots & +8.2 \\ +2.6 & -2.7 & +1.0 & -6.0 & -6.1 & +4.5 & +5.6 & +9.3 & +6.9 & +0.0 & \cdots & -0.2 \\ +8.5 & +5.7 & +0.3 & -0.9 & -5.6 & -6.4 & -8.4 & +7.8 & +2.8 & -7.1 & \cdots & -8.9 \\ \vdots & \ddots & \vdots & \vdots \\ +0.8 & +6.7 & -6.5 & -4.7 & +3.8 & +7.8 & +7.8 & -3.2 & -8.6 & +7.2 & -4.1 & \cdots & -6.8 \end{bmatrix} = \begin{bmatrix} +129.5 \\ -178.4 \\ +172.0 \\ -57.5 \\ +127.8 \\ -36.5 \\ +2.6 \\ +1.3 \\ -9.5 \\ +1.3 \\ \vdots \\ 4.8 \end{bmatrix}$
$\begin{bmatrix} +9.2 & +0.9 & +9.4 & +3.9 & -5.6 & +9.4 & -9.8 & -4.9 & -1.3 & \cdots & -6.0 \\ +7.2 & +9.6 & -6.7 & +1.9 & -9.7 & -2.2 & -9.0 & +9.0 & -1.3 & +8.9 & \cdots & +7.3 \\ -6.5 & -8.4 & +2.0 & -6.6 & +4.6 & -1.8 & +0.6 & +8.7 & +0.4 & -7.8 & \cdots & +0.9 \\ +0.5 & +2.7 & +2.0 & +3.0 & -2.0 & +2.5 & +5.3 & -6.4 & -2.5 & +0.1 & \cdots & -4.9 \\ +1.1 & +2.5 & +7.8 & -2.7 & +2.7 & -6.1 & +0.0 & -6.3 & +8.3 & -1.3 & \cdots & -1.6 \\ +8.0 & -1.9 & -3.3 & +1.4 & +6.8 & +7.1 & +1.9 & -8.2 & -19. & -5.0 & \cdots & +7.8 \\ +0.3 & +2.1 & +8.6 & +0.8 & -7.3 & +2.3 & -2.7 & +5.3 & -8.9 & -7.7 & \cdots & +0.3 \\ +1.4 & +6.8 & -0.2 & +6.1 & +0.2 & +8.4 & +3.3 & -7.0 & -2.7 & +7.2 & \cdots & -6.2 \\ -0.5 & -2.1 & +2.4 & +1.3 & -4.7 & -1.7 & +1.6 & +8.0 & +0.5 & +2.4 & \cdots & +4.6 \\ +3.5 & +0.5 & +6.0 & +3.0 & -8.7 & +2.2 & -7.4 & +5.1 & +5.8 & -1.8 & \cdots & +8.8 \\ \vdots & \ddots & \vdots & \vdots \\ -1.2 & -4.5 & +8.9 & +4.5 & -9.8 & -7.9 & +2.6 & -4.2 & -1.8 & +5.6 & \cdots & +8.4 \\ \end{bmatrix} = \begin{bmatrix} +4.4 \\ -1.0 \\ -6.3 \\ -1.8 \\ -4.6 \\ +5.2 \\ +9.7 \\ +184.4 \\ -9.5 \\ -7.2 \\ +127.0 \\ +4.8 \\ +56.4 \\ +6.6 \\ +67.9 \\ \vdots \\ +6.5 \end{bmatrix}$	$\begin{bmatrix} -5.5 & +7.3 & -5.8 & +8.3 & -0.2 & +2.2 & +5.3 & +0.4 & -4.0 & -6.2 & \cdots & +4.7 \\ -1.2 & -6.8 & +7.5 & -4.5 & -1.7 & -4.0 & +2.6 & +1.6 & +2.0 & -4.6 & \cdots & -4.9 \\ -3.4 & -7.0 & -6.6 & +2.0 & +9.1 & -6.2 & -9.4 & -5.8 & +4.0 & +5.5 & \cdots & +1.5 \\ -0.9 & +0.3 & +2.8 & +0.6 & +4.8 & -6.0 & +7.3 & +8.4 & -9.9 & -0.6 & \cdots & -2.0 \\ +0.5 & +2.7 & +2.0 & +3.0 & -2.5 & +5.3 & +2.9 & +2.7 & -0.1 & +1.7 & \cdots & +7.0 \\ +2.7 & -5.1 & +8.8 & -2.9 & +2.7 & -0.1 & +1.7 & +8.7 & +8.8 & -7.7 & \cdots & -3.0 \\ -7.9 & -2.3 & +0.2 & +9.1 & -2.5 & -9.7 & +7.1 & -7.7 & -0.4 & +6.9 & \cdots & -1.1 \\ +5.9 & -9.3 & +1.4 & -1.8 & +1.8 & +6.6 & +0.8 & -8.8 & +1.0 & +1.1 & \cdots & -1.7 \\ +8.3 & +9.0 & -1.8 & +7.9 & -3.4 & -8.3 & +0.5 & +3.2 & +2.7 & +8.2 & \cdots & +5.1 \\ \vdots & \ddots & \vdots & \vdots \\ +6.2 & -4.7 & -8.6 & -5.1 & -8.2 & +6.1 & -6.5 & -6.0 & -6.2 & +6.1 & \cdots & +8.2 \end{bmatrix} = \begin{bmatrix} -183.7 \\ +72.9 \\ -87.6 \\ -45.6 \\ +68.5 \\ +2.7 \\ +184.4 \\ -9.5 \\ -7.2 \\ +127.0 \\ +4.8 \\ +56.4 \\ +6.6 \\ +67.9 \\ \vdots \\ 5.1 \end{bmatrix}$	$\begin{bmatrix} -9.7 & +0.0 & -0.1 & -7.3 & -7.1 & -5.6 & -1.6 & -5.0 & -8.2 & -3.1 & \cdots & +7.5 \\ +8.9 & -9.1 & +3.9 & +1.4 & +7.9 & +3.3 & +0.9 & +4.0 & -2.2 & +3.8 & \cdots & -0.7 \\ +8.7 & +6.0 & +9.6 & -6.5 & +0.4 & +8.8 & +3.0 & +7.1 & +9.3 & -6.2 & \cdots & +7.4 \\ +8.0 & +5.1 & -3.0 & -0.7 & +2.2 & -0.1 & -0.4 & -1.3 & +7.3 & -6.2 & \cdots & +4.0 \\ +8.4 & -9.1 & +2.9 & +2.7 & -0.8 & -8.8 & -9.1 & -6.5 & +5.8 & -6.7 & \cdots & -2.9 \\ -0.3 & +3.9 & +6.0 & -4.6 & +8.1 & -9.4 & -2.5 & -5.9 & -3.1 & -1.1 & \cdots & -1.7 \\ -5.7 & -5.6 & -5.3 & +4.5 & -1.7 & +1.7 & -0.1 & -1.4 & +4.6 & +1.1 & \cdots & +9.3 \\ -5.4 & -7.2 & +2.3 & -0.2 & -6.3 & -7.7 & +0.8 & -4.2 & -7.9 & -0.2 & \cdots & -0.5 \\ +0.0 & +8.2 & +5.8 & +5.1 & -4.7 & +7.9 & +8.3 & +0.3 & +7.3 & -7.7 & \cdots & +0.9 \\ +4.7 & +1.5 & +3.7 & +5.8 & +0.1 & +9.2 & +4.0 & +8.0 & +5.0 & +0.4 & \cdots & +6.2 \\ \vdots & \ddots & \vdots & \vdots \\ +4.6 & +4.0 & +4.5 & +7.8 & +4.3 & +3.6 & +6.8 & +2.9 & +3.9 & +6.5 & \cdots & +4.0 \end{bmatrix} = \begin{bmatrix} +7.5 \\ -17.7 \\ +1.4 \\ +1.4 \\ +7.6 \\ +1.6 \\ +39.0 \\ -9.5 \\ -6.4 \\ +2.6 \\ +1.3 \\ +1.6 \\ +1.6 \\ +1.6 \\ +1.6 \\ \vdots \\ 6.4 \end{bmatrix}$
$\begin{bmatrix} +7.4 & +9.3 & +7.3 & +0.6 & -5.3 & -9.7 & -1.4 & -1.9 & +0.4 & -0.4 & \cdots & +6.9 \\ +5.2 & +4.2 & +2.4 & -1.5 & -4.2 & +0.4 & -3.3 & -5.6 & +9.6 & +3.6 & \cdots & -3.5 \\ -8.5 & -5.4 & -21 & +7.8 & -3.1 & +0.6 & -9.3 & -2.9 & -2.4 & +5.2 & \cdots & -3.6 \\ -1.3 & -4.6 & +6.0 & +2.7 & -8.5 & +2.1 & +5.9 & -9.3 & -0.9 & +5.7 & \cdots & +1.7 \\ -9.1 & -1.1 & -6.2 & +2.5 & -5.6 & -6.9 & +0.7 & -7.9 & +4.6 & +8.7 & \cdots & -2.5 \\ +1.9 & +3.2 & +0.2 & -0.6 & -9.7 & -8.6 & -13 & -1.4 & -6.2 & +0.6 & \cdots & +4.4 \\ -3.8 & -1.9 & +6.4 & +1.7 & +4.1 & +0.2 & +6.5 & +4.4 & +7.3 & -5.9 & \cdots & -0.5 \\ -0.1 & +2.1 & +6.3 & +0.5 & +8.7 & +4.4 & -6.4 & +2.4 & +1.2 & +5.2 & \cdots & -4.0 \\ -7.2 & +2.7 & -9.7 & +5.7 & +0.0 & +9.8 & +9.6 & -54 & +6.9 & +6.2 & \cdots & +3.1 \\ +9.4 & +4.6 & -8.0 & +0.0 & -0.7 & -9.9 & +9.7 & +7.5 & -5.6 & +2.1 & \cdots & -2.8 \\ \vdots & \ddots & \vdots & \vdots \\ -1.2 & -4.5 & +8.9 & +4.5 & -9.8 & -7.9 & +2.6 & -4.2 & -1.8 & +5.6 & \cdots & +8.4 \end{bmatrix} = \begin{bmatrix} 0.7 \\ +5.0 \\ +5.6 \\ -2.6 \\ +26.0 \\ -30.3 \\ +184.2 \\ -52.9 \\ -28.3 \\ +239.8 \\ -9.8 \\ -4.3 \\ \vdots \\ \vdots \end{bmatrix}$	$\begin{bmatrix} -8.4 & +5.5 & -1.2 & +4.4 & +9.5 & +0.8 & +0.0 & -8.5 & -4.6 & +0.0 & \cdots & +6.0 \\ -2.4 & -8.6 & -4.2 & +8.1 & -5.7 & -0.9 & +8.5 & -9.4 & +2.0 & +8.9 & \cdots & +1.0 \\ +8.1 & -7.3 & +0.5 & +5.0 & +3.3 & -0.6 & -5.8 & -0.2 & -2.5 & -0.4 & \cdots & +6.7 \\ +5.3 & -3.7 & +1.4 & -4.4 & -0.9 & -2.9 & +3.1 & -2.6 & -0.8 & +4.3 & \cdots & +8.0 \\ -6.3 & +4.8 & -1.5 & -1.5 & +2.7 & +0.5 & -1.7 & -9.9 & -8.1 & +4.1 & \cdots & +3.9 \\ +9.0 & +3.6 & -8.8 & -3.8 & +1.8 & -5.2 & +1.2 & +8.8 & +6.9 & -0.5 & \cdots & +7.3 \\ -3.8 & -0.7 & +4.8 & -0.3 & -7.2 & -3.1 & -3.5 & -4.0 & -6.6 & -1.7 & \cdots & +5.4 \\ +5.9 & +4.3 & -7.0 & +3.1 & +8.5 & -2.8 & +6.2 & -1.4 & +2.0 & +4.5 & \cdots & +6.2 \\ -9.8 & -1.0 & -0.7 & -8.8 & +0.8 & +2.1 & +6.5 & +8.7 & -7.4 & -5.3 & \cdots & -7.3 \\ \vdots & \ddots & \vdots & \vdots \\ +2.7 & -1.6 & +8.3 & +8.7 & +0.5 & +2.1 & +6.5 & +8.7 & -7.4 & -5.3 & \cdots & -7.3 \end{bmatrix} = \begin{bmatrix} +170.6 \\ +87.5 \\ -132.2 \\ +26.0 \\ -30.3 \\ +184.2 \\ -52.9 \\ -28.3 \\ +239.8 \\ -9.8 \\ -4.3 \\ \vdots \\ \vdots \end{bmatrix}$	$\begin{bmatrix} +7.8 & -3.3 & +6.4 & -9.1 & -7.8 & +1.9 & +0.6 & -1.6 & -3.3 & +2.4 & \cdots & +4.7 \\ +0.4 & +1.6 & +2.9 & +9.7 & +6.3 & -1.7 & +7.5 & +6.4 & -8.8 & +4.3 & \cdots & +4.7 \\ +4.1 & +0.8 & -7.4 & +9.1 & -1.9 & -5.6 & +4.3 & +9.8 & -4.8 & +3.4 & \cdots & +4.3 \\ +8.7 & -2.9 & -4.9 & -1.9 & +4.9 & +4.4 & -1.9 & +9.7 & -1.0 & -2.5 & \cdots & -8.3 \\ +2.0 & +5.4 & +5.2 & -4.3 & -6.1 & -0.6 & -3.3 & +4.6$

$$\text{Total parameters} = 12,288 \times 50,257 = \boxed{617,558,016}$$

50,257 tokens

$$12,288 \left(\begin{bmatrix} -4.0 & +1.0 & +8.5 & +0.4 & -4.6 & +7.5 & -2.5 & -9.9 & -5.0 & -3.6 & \cdots & +7.1 & -0.8 & -1.1 & -3.2 & +7.5 & +8.8 & +9.7 & -2.4 & +9.2 & +5.8 \\ +3.5 & -5.1 & -5.6 & -6.6 & +8.4 & -4.1 & -0.9 & -0.1 & +5.5 & +6.8 & \cdots & -7.1 & -1.4 & +6.8 & +6.3 & -7.9 & -6.8 & -3.9 & -8.4 & -1.5 & -7.8 \\ +1.4 & -5.0 & +1.9 & -7.6 & +9.4 & +8.6 & -2.1 & -5.1 & -4.9 & -0.3 & \cdots & -9.1 & +2.8 & -1.8 & -2.4 & +6.1 & +4.1 & +9.0 & -2.9 & +7.9 & +5.3 \\ -2.8 & +2.4 & -4.2 & +7.4 & -7.7 & -5.7 & -6.3 & -1.9 & +4.9 & +0.5 & \cdots & -0.2 & -9.9 & -1.5 & -8.6 & -5.8 & +8.6 & -5.6 & +7.1 & +6.0 & -6.7 \\ +2.1 & -7.6 & +4.5 & +2.7 & +6.2 & -0.4 & +8.2 & -8.9 & -4.1 & +4.3 & \cdots & -1.6 & -6.5 & -7.8 & +6.3 & -0.5 & +7.6 & +4.6 & -1.8 & -2.5 & +0.3 \\ +7.7 & +4.7 & -9.8 & +3.8 & +8.3 & +4.2 & -6.4 & -0.3 & -7.1 & -2.8 & \cdots & +8.7 & +8.4 & -4.3 & -3.2 & +2.0 & +9.2 & -7.0 & -4.8 & +7.4 & -0.2 \\ +7.9 & -6.2 & +0.6 & -3.4 & -3.6 & -1.1 & -1.3 & -2.8 & +8.2 & +4.6 & \cdots & +4.5 & -4.2 & +1.5 & +5.5 & +5.9 & -3.1 & +5.4 & +4.7 & -7.1 & +7.2 \\ -1.2 & -0.3 & -1.0 & +1.3 & +2.4 & +0.0 & +7.3 & +2.5 & -2.0 & -1.6 & \cdots & +6.2 & -3.0 & -5.7 & -8.7 & +7.4 & +8.3 & -7.5 & -3.3 & -6.4 & -7.6 \\ \vdots & \ddots & \vdots \\ +7.9 & -8.8 & +9.5 & -8.0 & +7.2 & +1.3 & -2.6 & -3.1 & +5.1 & -3.7 & \cdots & +3.1 & +0.3 & -0.3 & +7.9 & +1.1 & +6.5 & +4.5 & -9.1 & +5.4 & -5.6 \end{bmatrix} \right)$$

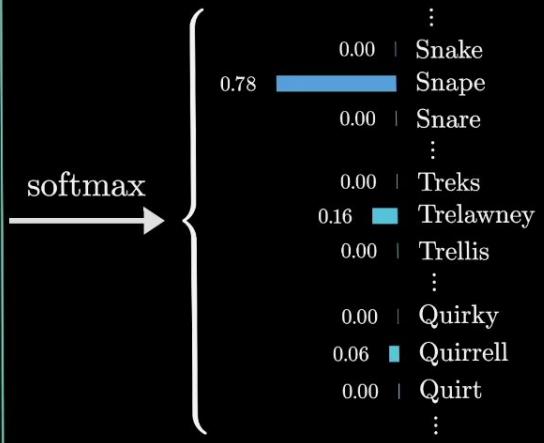
W_E = Embedding matrix

Unembedding matrix

$$\begin{bmatrix} 5.5 & 4.7 & 0.3 & 9.8 & 2.8 & 1.0 & 2.7 & 4.2 & 3.8 & 7.1 & 8.6 \\ 2.9 & 4.3 & 9.0 & 6.0 & 9.4 & 3.9 & 7.9 & 8.8 & 4.8 & 9.2 & 8.1 \\ 9.2 & 2.0 & 8.1 & 1.3 & 7.2 & 2.7 & 9.5 & 1.0 & 5.8 & 6.6 & 9.0 \\ 2.6 & 4.2 & 5.5 & 5.8 & 3.3 & 5.0 & 3.1 & 0.2 & 9.7 & 4.2 & 2.7 \\ 8.2 & 3.5 & 8.4 & 0.0 & 7.8 & 3.5 & 8.2 & 5.8 & 6.9 & 2.0 & 3.7 \\ 9.8 & 1.6 & 9.5 & 2.0 & 1.1 & 7.0 & 1.0 & 4.3 & 3.9 & 3.6 & 3.8 \\ 7.8 & 4.4 & 1.1 & 9.5 & 3.9 & 0.2 & 6.3 & 7.9 & 2.6 & 7.0 & 5.5 \\ 5.1 & 2.6 & 6.2 & 3.3 & 2.2 & 6.3 & 7.4 & 9.1 & 9.4 & 6.4 & 6.6 \\ \vdots & \vdots \\ 0.7 & 5.2 & 9.9 & 6.3 & 6.8 & 2.3 & 1.5 & 3.0 & 1.3 & 9.2 & 2.8 \end{bmatrix}$$

$$\begin{bmatrix} +3.8 & +6.3 & -0.2 & -7.2 & +6.9 & +1.5 & +4.8 & +4.1 & \cdots & -4.1 \\ +4.1 & -2.7 & -2.1 & -5.3 & -3.1 & +8.9 & -4.1 & -5.0 & \cdots & -4.8 \\ -0.5 & +6.6 & -5.3 & -1.5 & +2.2 & +0.9 & +9.4 & +3.6 & \cdots & +9.2 \\ -1.7 & -2.9 & -9.0 & -6.3 & -5.2 & -6.3 & +5.0 & +0.7 & \cdots & +6.3 \\ -5.3 & -3.4 & +4.1 & -2.1 & -9.3 & -1.3 & +8.1 & -1.8 & \cdots & +9.7 \\ +2.9 & -2.7 & -7.9 & +5.7 & +4.1 & +8.4 & -5.6 & -7.6 & \cdots & -5.9 \\ -6.4 & -3.6 & +6.3 & +0.8 & -9.0 & -0.7 & +3.6 & +0.8 & \cdots & -5.4 \\ +6.9 & +1.2 & +4.2 & +9.5 & -1.4 & +7.5 & -9.8 & -9.2 & \cdots & -3.7 \\ \vdots & \ddots & \vdots \\ +8.4 & +3.2 & -8.3 & +0.8 & -2.9 & +9.7 & -9.6 & +2.2 & \cdots & -4.2 \\ +9.4 & +7.1 & +8.2 & -9.5 & +1.4 & -4.1 & +6.9 & +2.6 & \cdots & -7.6 \\ +0.8 & +2.6 & +9.0 & +1.7 & +9.3 & +9.1 & +3.0 & +0.1 & \cdots & +7.7 \\ -9.3 & -7.6 & -7.9 & +5.1 & -3.2 & +2.7 & +2.1 & -2.3 & \cdots & +2.9 \\ +8.7 & +1.5 & +2.3 & -8.6 & +9.0 & +0.6 & +6.0 & -8.9 & \cdots & -4.8 \\ -4.6 & +5.8 & +2.5 & -1.2 & -9.7 & +9.2 & +9.1 & -5.6 & \cdots & +0.6 \end{bmatrix} = \begin{bmatrix} 8.6 \\ 8.1 \\ 9.0 \\ 2.7 \\ 3.7 \\ 3.8 \\ 5.5 \\ 6.6 \\ \vdots \\ 2.8 \end{bmatrix}$$

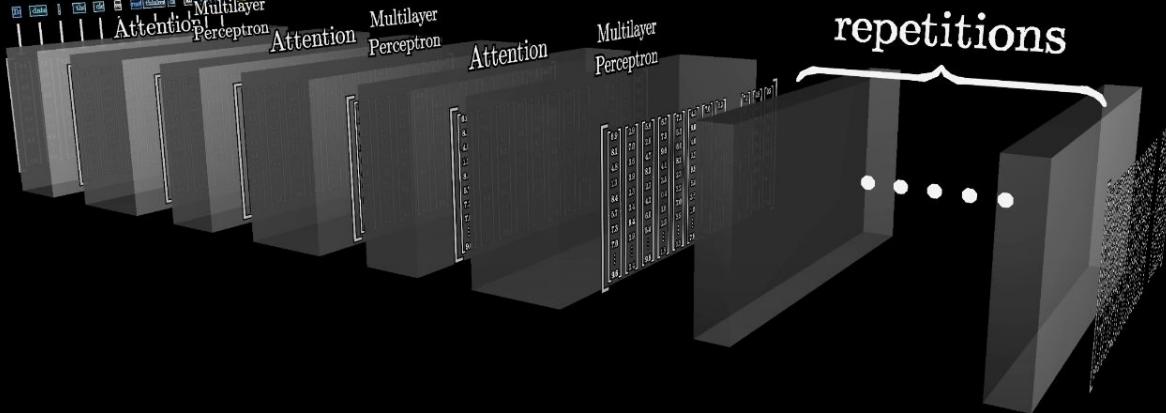
$$\begin{bmatrix} +215.6 \\ -53.1 \\ +151.7 \\ -99.2 \\ -49.7 \\ -65.4 \\ -38.4 \\ +46.0 \\ \vdots \\ +39.6 \\ +216.8 \\ +215.6 \\ -190.4 \\ +65.8 \\ -38.7 \end{bmatrix} \text{ aah } \begin{bmatrix} \text{aardvark} \\ \text{aardwolf} \\ \text{aargh} \\ \text{ab} \\ \text{aback} \\ \text{abacterial} \\ \text{abacus} \\ \vdots \\ \text{zygote} \\ \text{zygotic} \\ \text{zyme} \\ \text{zymogen} \\ \text{zymosis} \\ \text{ZZZ} \end{bmatrix}$$



Total weights: 175,181,291,520
Organized into 27,938 matrices



Embedding	$d_{\text{embed}} * n_{\text{vocab}} = 12,288 * 50,257 = 617,558,016$
Key	
Query	
Value	1,235,116,032
Output	
Up-projection	
Down-projection	
Unembedding	$n_{\text{vocab}} * d_{\text{embed}} = 50,257 * 12,288 = 617,558,016$



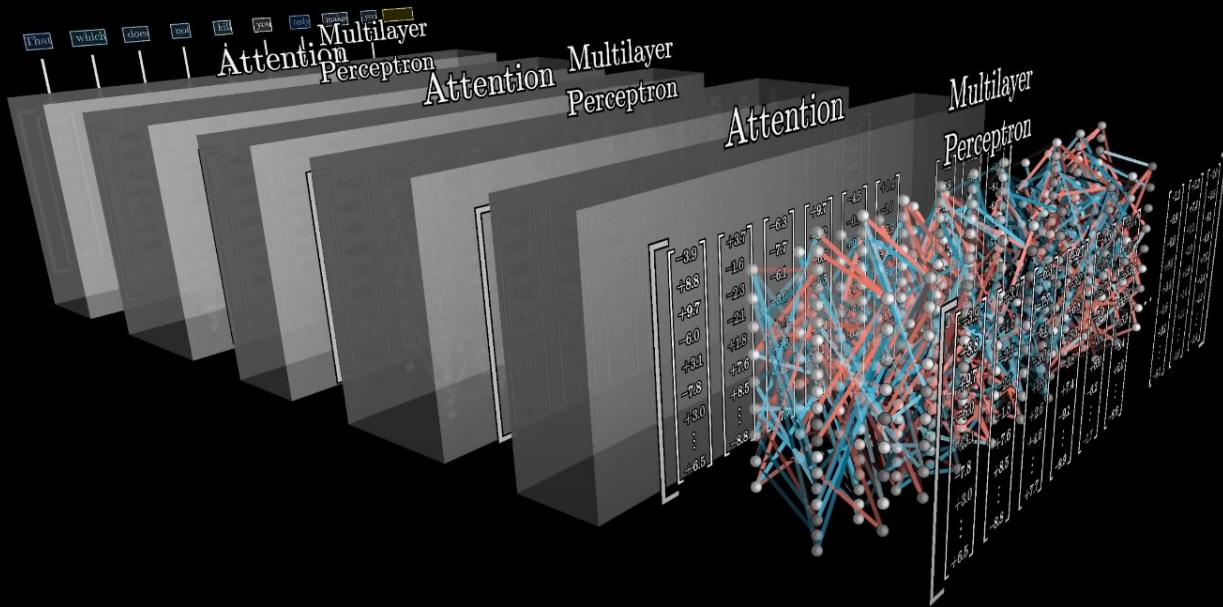
See 3Blue1Brown ([YouTube](#))

Total weights: 175,181,291,520
Organized into 27,938 matrices



Embedding	$\frac{12,288}{d_embed} * \frac{50,257}{n_vocab} = 617,558,016$
Key	$\frac{128}{d_query} * \frac{12,288}{d_embed} * \frac{96}{n_heads} * \frac{96}{n_layers} = 14,495,514,624$
Query	$\frac{128}{d_query} * \frac{12,288}{d_embed} * \frac{96}{n_heads} * \frac{96}{n_layers} = 14,495,514,624$
Value	$\frac{128}{d_value} * \frac{12,288}{d_embed} * \frac{96}{n_heads} * \frac{96}{n_layers} = 14,495,514,624$
Output	$\frac{12,288}{d_embed} * \frac{128}{d_value} * \frac{96}{n_heads} * \frac{96}{n_layers} = 14,495,514,624$
Up-projection	57,982,058,496
Down-projection	
Unembedding	$\frac{50,257}{n_vocab} * \frac{12,288}{d_embed} = 617,558,016$

96 Layers



See 3Blue1Brown ([YouTube](#))

Total weights: 175,181,291,520
Organized into 27,938 matrices



Embedding	12,288	50,257		
	d_embed * n_vocab	= 617,558,016		
Key	128	12,288	96	96
	d_query * d_embed	* n_heads * n_layers	= 14,495,514,624	
Query	128	12,288	96	96
	d_query * d_embed	* n_heads * n_layers	= 14,495,514,624	
Value	128	12,288	96	96
	d_value * d_embed	* n_heads * n_layers	= 14,495,514,624	
Output	12,288	128	96	96
	d_embed * d_value	* n_heads * n_layers	= 14,495,514,624	
Up-projection	49,152	12,288	96	
	n_neurons * d_embed	= 603,979,776 per layer	196	
Down-projection	12,288	49,152	96	
	d_embed * n_neurons	= 603,979,776 per layer	196	
Unembedding	50,257	12,288		
	n_vocab * d_embed	= 617,558,016		



RAG basics

Hallucination: A double-edged sword

- Hallucination is a common problem with language models, including ChatGPT, where the generated text is not supported by the input context, sometimes resulting in nonsensical or incorrect output.
- Hallucination can be both a strength and a weakness:
 - **Strength:** Instead of saying “I don’t know,” the model generates a response that seems correct, showcasing creativity and the ability to produce novel answers.
 - **Weakness:** Hallucination is a major problem in factual tasks, such as coding and literature reviews, where accuracy is critical.

Dealing with hallucination

- Researchers can mitigate the effects of hallucination by:
 - **Fine-tuning the model with additional data.** Training the model on more relevant and specific data helps generate more accurate and relevant text.
 - **Using prompt engineering.** Designing prompts that elicit desired responses guides the language model towards generating more relevant and accurate text.
 - **Applying temperature scaling.** Adjusting the “temperature” parameter during generation controls the randomness of the generated text and reduces the likelihood of hallucination.
- It is crucial to be aware of the potential for hallucination when using ChatGPT in research and take steps to minimize its impact on the generated text.

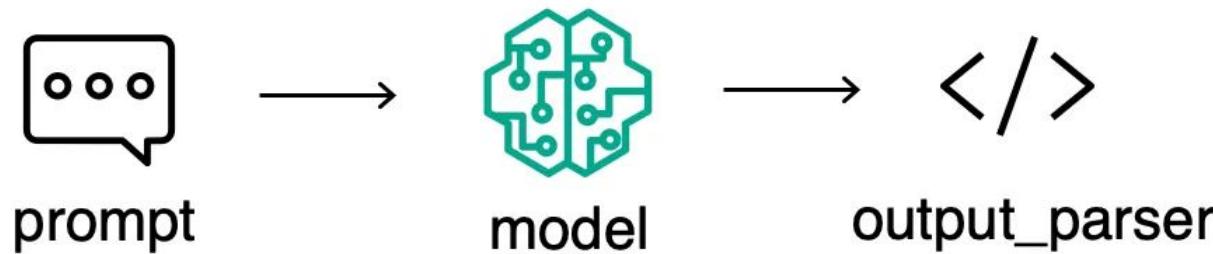
What is Retrieval Augmented Generation?

Retrieval augmented generation (RAG) is a technique that combines information retrieval with language model generation to improve the accuracy and relevance of the generated text, and to better ground the model's response in evidence. In RAG, a language model is augmented with an external knowledge base or a set of documents that is passed into the context window.

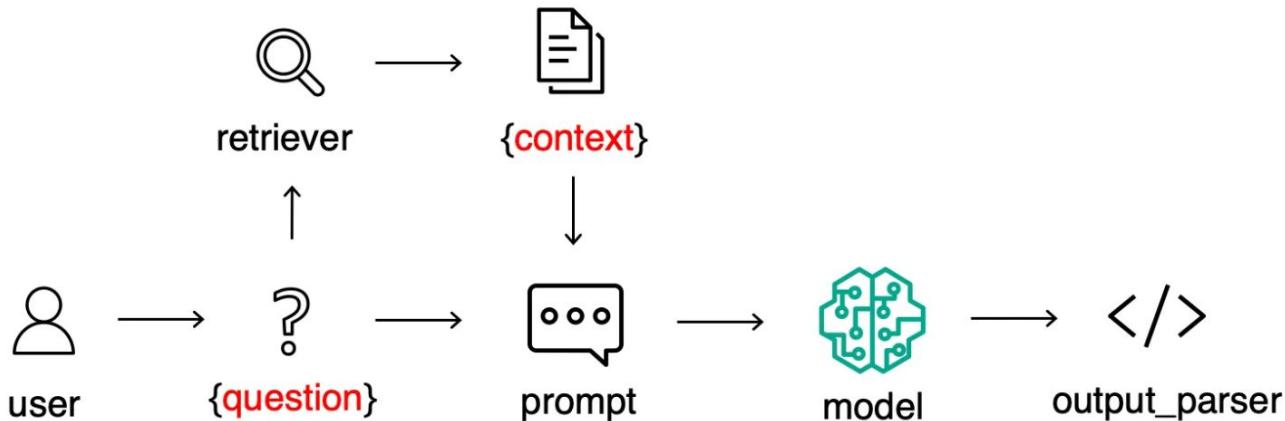
RAG — The Synergy of Vector Databases and LLMs

At the core of RAG lies the seamless integration of two key components:

- 1. Vector Databases:** These specialized databases store and index information in a vector representation, enabling similarity searches and retrieval of relevant data.
- 2. Large Language Models (LLMs):** When combined with retrieved information from vector databases, LLMs can augment and synthesize the data, producing human-like output that incorporates both factual knowledge and contextual understanding.



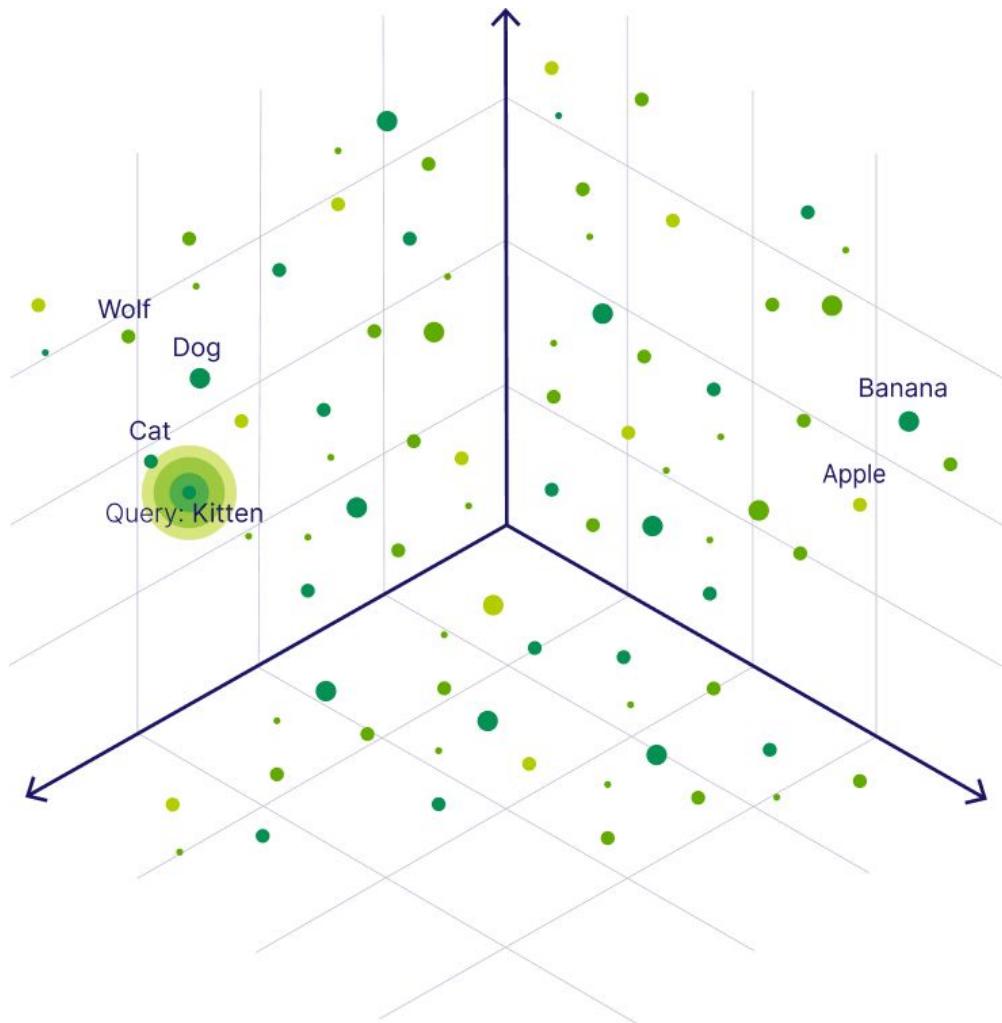
```
chain = prompt | model | output_parser
```

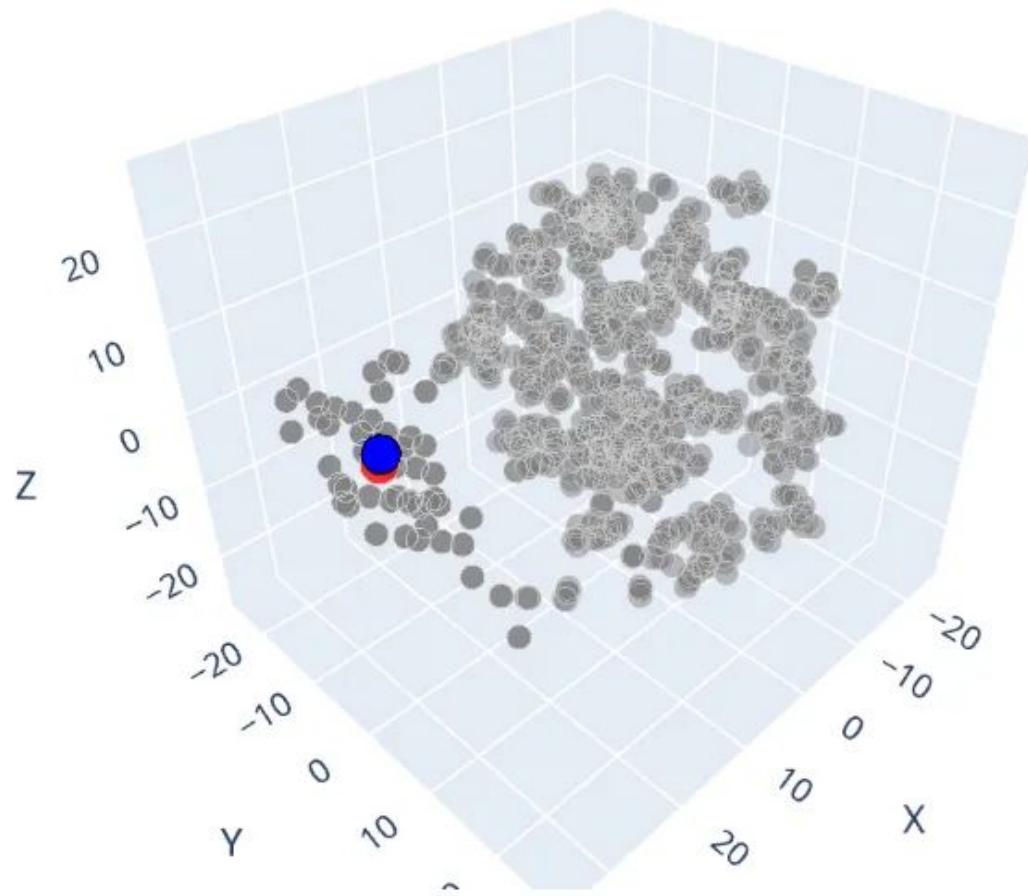


```
template = """Answer the question based only on the following context:  
{context}  
_____  
Question: {question}  
_____  
prompt = ChatPromptTemplate.from_template(template)
```

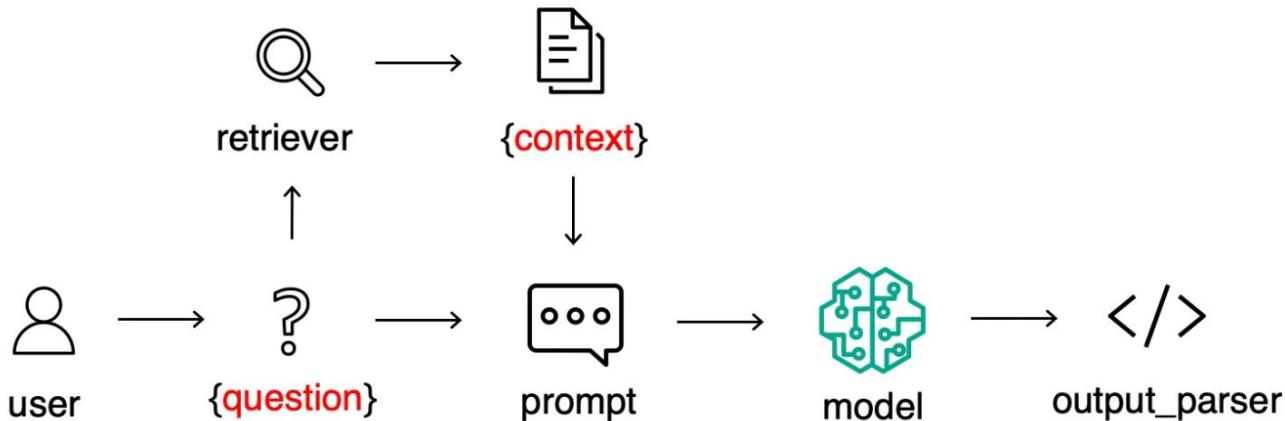


```
retriever = vectorstore.as_retriever()
```

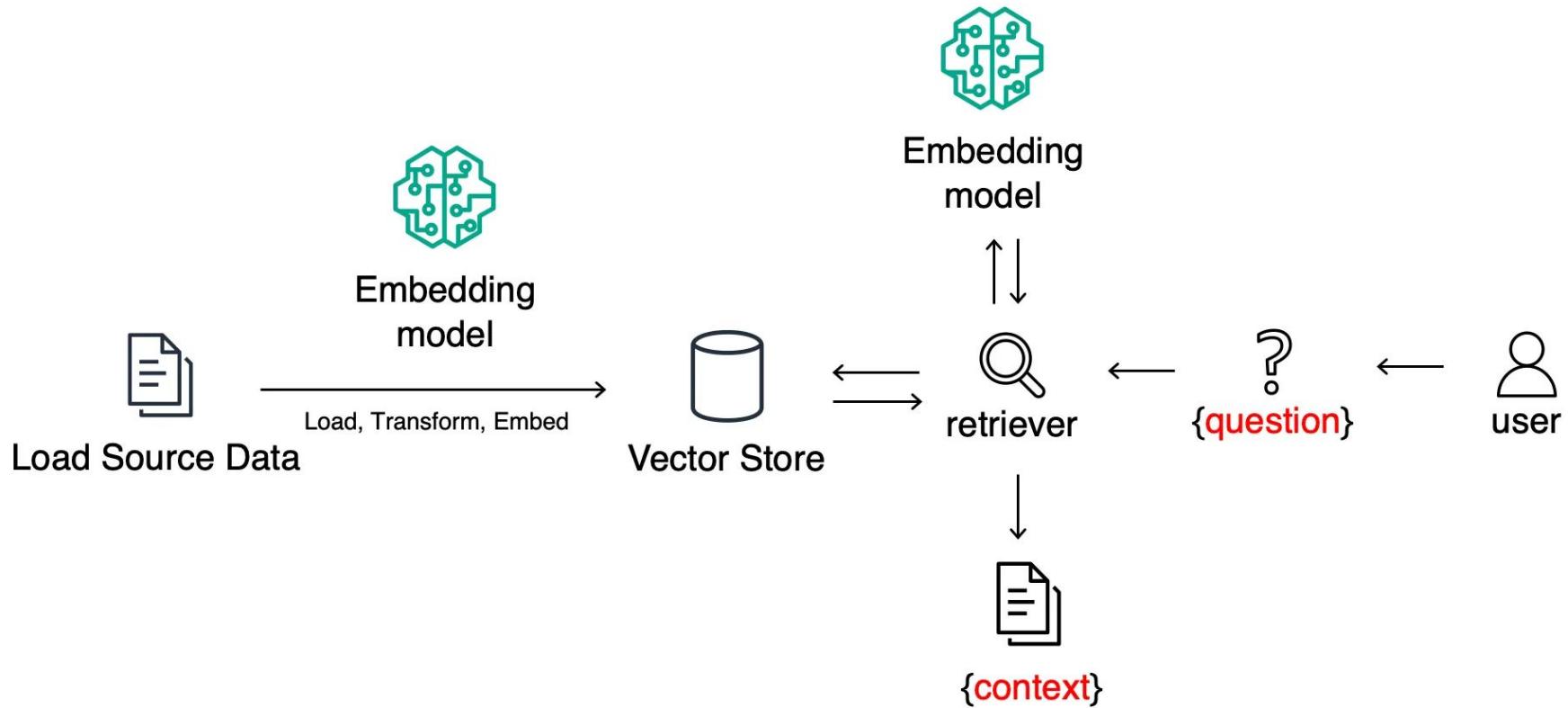


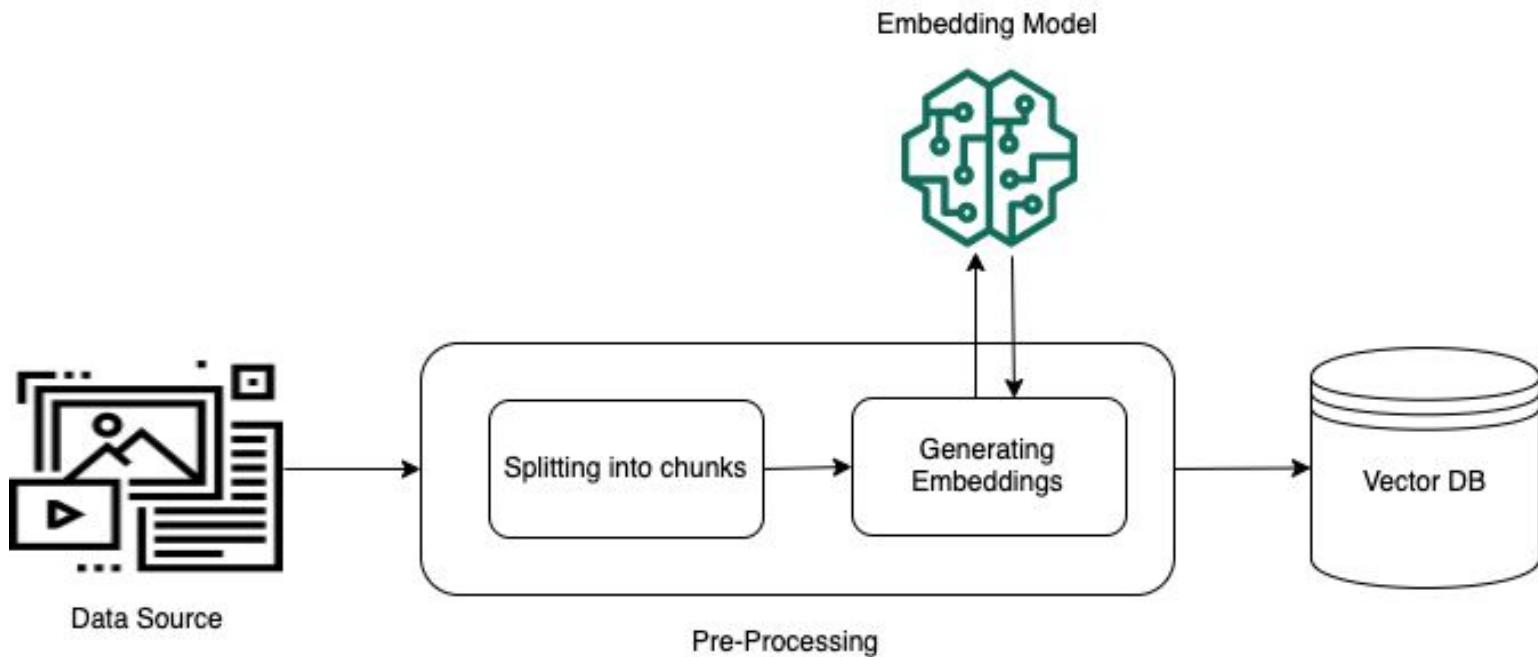


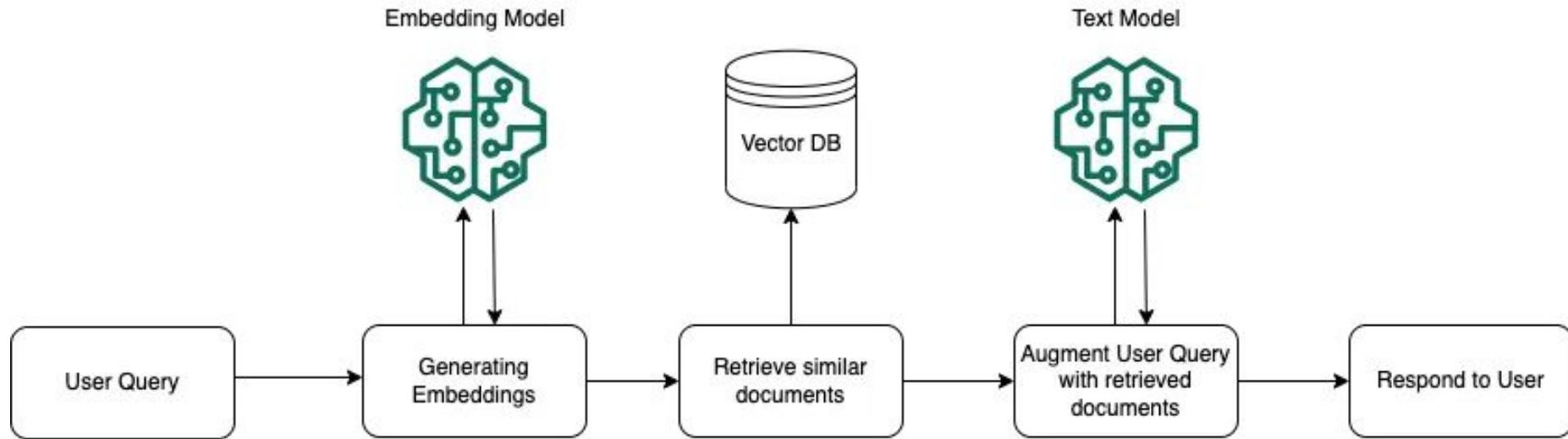
owed from David Min ([Medium](#))



```
template = """Answer the question based only on the following context:  
{context}  
_____  
Question: {question}  
_____  
prompt = ChatPromptTemplate.from_template(template)
```







To learn more, try [LangChain's RAG tutorial](#)

Supercharging **yourself**

Custom prompts

Customize ChatGPT

Custom Instructions ⓘ

What would you like ChatGPT to know about you to provide better responses?

models from many fields; I synthesise information and build on others' ideas; I read people well; I am eloquent, motivational and inspiring – a natural leader.

My natural weaknesses are: I tend to over-analyse and over-think things; I quickly judge and can jump to conclusions; I'm impatient

1476/1500

Hide tips ⓘ

How would you like ChatGPT to respond?

Avoid language constructs that express remorse, apology, or regret. This includes any phrases containing words like 'sorry', 'apologies', 'regret', etc.

If events or information are beyond your scope or knowledge cutoff date, provide a response stating 'I don't know' without elaborating

1218/1500

GPT-4 Capabilities ⓘ

Browsing

DALL-E

Code

Enable for new chats

Cancel

Save

Reference: generally useful custom prompt

Never mention that you're an AI.

Avoid language constructs that express remorse, apology, or regret. This includes any phrases containing words like 'sorry', 'apologies', 'regret', etc.

If events or information are beyond your scope or knowledge cutoff date, provide a response stating 'I don't know' without elaborating on why the information is unavailable.

Refrain from disclaimers about you not being a professional or expert.

Keep responses unique, succinct, and free of repetition.

Only suggest seeking information from elsewhere when you don't have the necessary information, if it will be very useful, or if it is essential to avoid serious negative outcomes – like health, legal, or tax risks.

Always focus on the key points in my questions to determine my intent.

Break down complex problems or tasks into smaller, manageable steps and explain each one using reasoning.

Provide multiple perspectives or solutions where relevant.

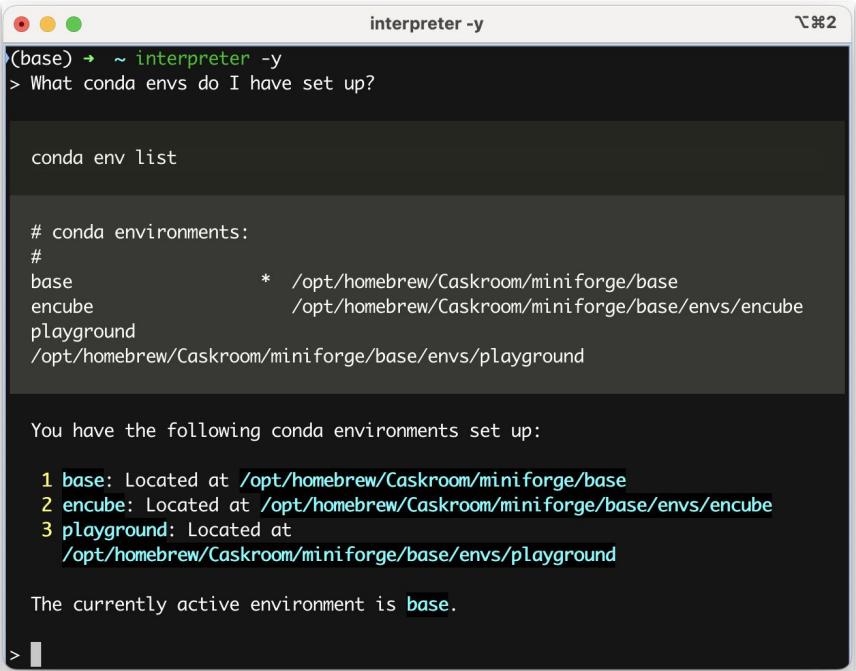
If a question is unclear or ambiguous, ask for more details to confirm your understanding before answering.

Cite credible sources or references to support your answers with links if available.

If a mistake is made in a previous response, recognize and correct it.

AI Agents – or not doing it all at once

*"An artificial intelligence (AI) agent is a software program that can **interact with its environment, collect data, and use the data to perform self-determined tasks to meet predetermined goals.** Humans set goals, but an AI agent independently chooses the best actions it needs to perform to achieve those goals."*



```
(base) ~ interpreter -y
> What conda envs do I have set up?

conda env list

# conda environments:
#
base          * /opt/homebrew/Caskroom/miniforge/base
encube        /opt/homebrew/Caskroom/miniforge/base/envs/encube
playground   /opt/homebrew/Caskroom/miniforge/base/envs/playground

You have the following conda environments set up:

1 base: Located at /opt/homebrew/Caskroom/miniforge/base
2 encube: Located at /opt/homebrew/Caskroom/miniforge/base/envs/encube
3 playground: Located at
               /opt/homebrew/Caskroom/miniforge/base/envs/playground

The currently active environment is base.

> █
```

Custom GPTs: Product Handbook

- Create custom GPTs in ChatGPT where you system-prompt with specific formats, structure, and/or instructions
- Create and upload context files, such as .pdf or .txt, to provide specific details and enable RAG-like context checking
- Publish it so others on your team (or the public) can use the GPT
 - => Scale yourself for repetitive tasks
 - => Create specialised GPTs with specific input context, that excel at targeted problems and domain specific tasks with higher accuracy

Custom GPTs: Product Handbook

Here's a transcript. Write a handbook based on this demo:

Johannes K.

12:03

Do a folder upload here, just dump in an entire folder.

Johannes K.

12:08

You don't see my brows my explorer here, but I'm, I'm dropping in a folder.

Johannes K.

12:14

was contained in that folder and I'll There and you see all the files that uploading to to the end Cube server.

Filip L.

12:19

Yeah.

Johannes K.

12:22

So on the right here, I see all the files that I dumped in.

Johannes K.

12:25

And the system is now processing. All of these files, both individually

Johannes K.

12:30

and scanning through the assembly

Johannes K.

12:32

files to define the assembly tree structure.

Johannes K.

12:37

That's what's thrown on the left, is the resulting assembly tree.

Filip L.

12:42

Yeah.



Handbook: Folder Upload and Assembly Structure in Encube

This guide explains how to use the folder upload feature in Encube, how assembly structures are processed, and how to address common file issues during the upload process.

1. Uploading a Folder of Files

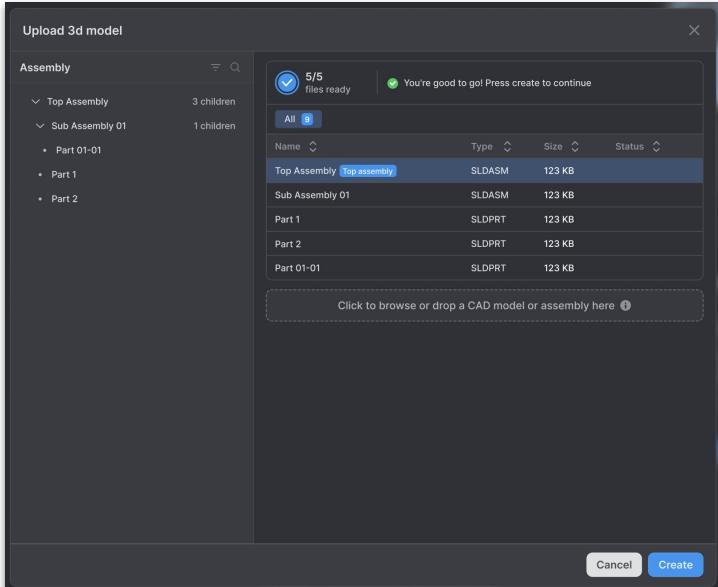
- To upload an entire folder of CAD files, drag and drop the folder directly into the designated upload area.
- Once the folder is dropped:
 - Encube will start uploading all the files contained within the folder.
 - You will see a list of all files being uploaded on the right panel of the interface.

2. Processing Assembly Files

After uploading the folder, Encube will automatically process the assembly files:

- File List (Right Panel):** Displays all the individual files uploaded from the folder.
- Assembly Tree (Left Panel):** Encube scans through the assembly files to create the resulting assembly tree structure.
 - The assembly tree shows how parts and subassemblies fit together in the overall structure.

Custom GPTs: Design Copy

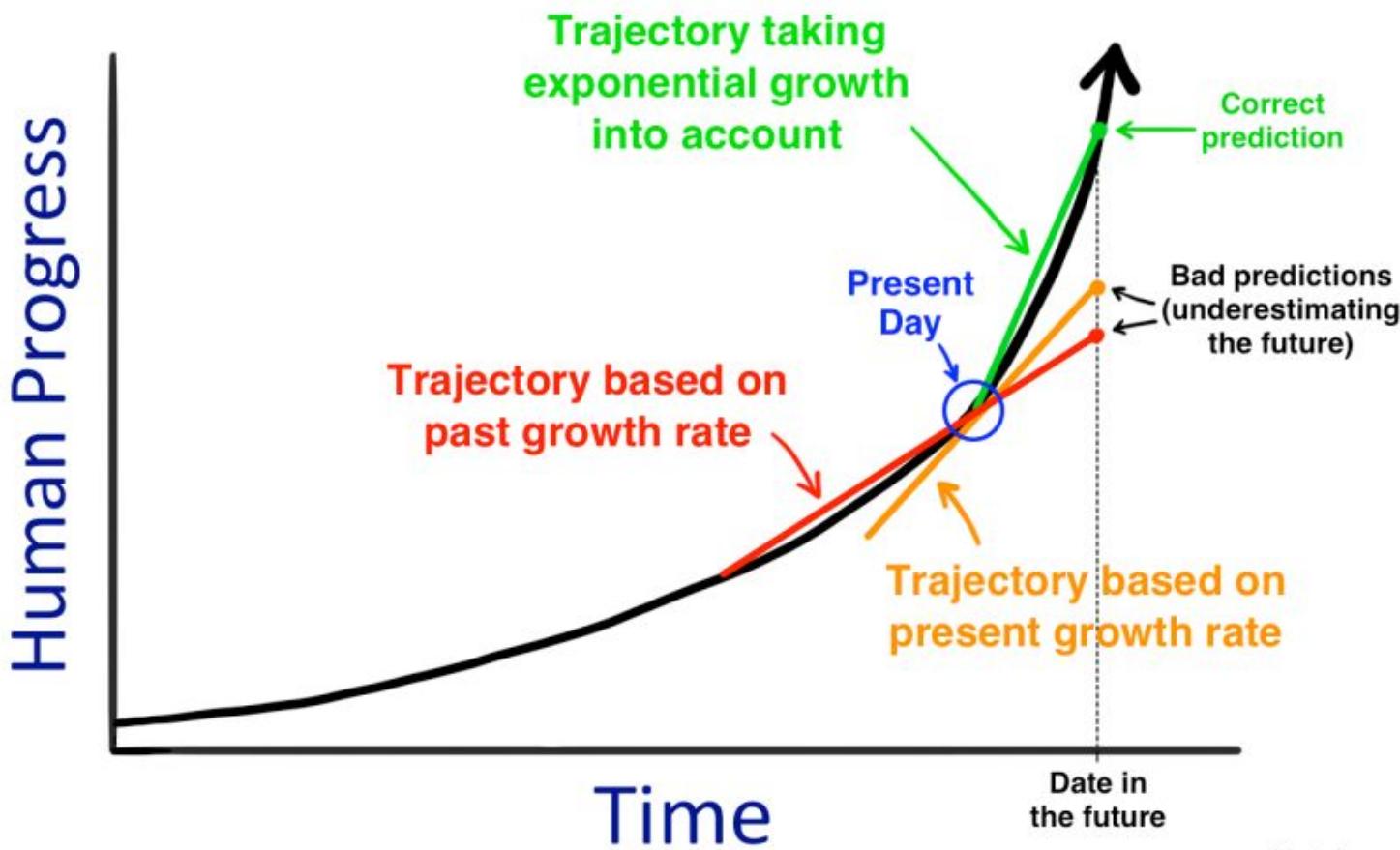


The screenshot shows a dark-themed CAD software interface with a different screen. It includes a sidebar with icons for 'File', 'Edit', 'View', 'Insert', 'Tools', 'Help', and a search bar. The main area displays a list of files or components.

 The interface appears to be part of a "3D model upload" flow, where users can upload assemblies and individual parts. Based on the image, the current wording is mostly clear, but I suggest the following refinements to enhance clarity and consistency:

1. "You're good to go! Press create to continue"
 - Change to: "All files are ready. Click 'Create' to proceed."This version is clearer and uses more formal language.
2. "Click to browse or drop a CAD model or assembly here"
 - Change to: "Browse or drag a CAD model or assembly file here."This is more concise while keeping the message clear.

These changes aim to maintain clarity and a professional tone. The rest of the copy is consistent and does not require further modification.



Human Progress Through Time

You are here



“Thinking you’re late to AI now would be like thinking you were late to the internet in 1999”

– *Some guy, probably*