# Demystifying AI

## Part 1 - AI State of the Art

Petronela Cretu 10.05.2024

# Outline

- AI -  what is in the current context
- How does AI work
- List of Generative AI models
- Energy consumption

# AI - what is in the current context

- deep learning performs so well - not known as of 2023.
- no new discovery or theoretical breakthrough
- 2 factors:
    - increase in computer power
    - the availability of vast amounts of training data
- In **2019**, generative pre-trained transformer (or "GPT") language models began to generate coherent text, and by 2023 these models were able to get human-level scores on the bar exam, SAT test, GRE test, and many other real-world applications.

https://en.wikipedia.org/wiki/Artificial_intelligence

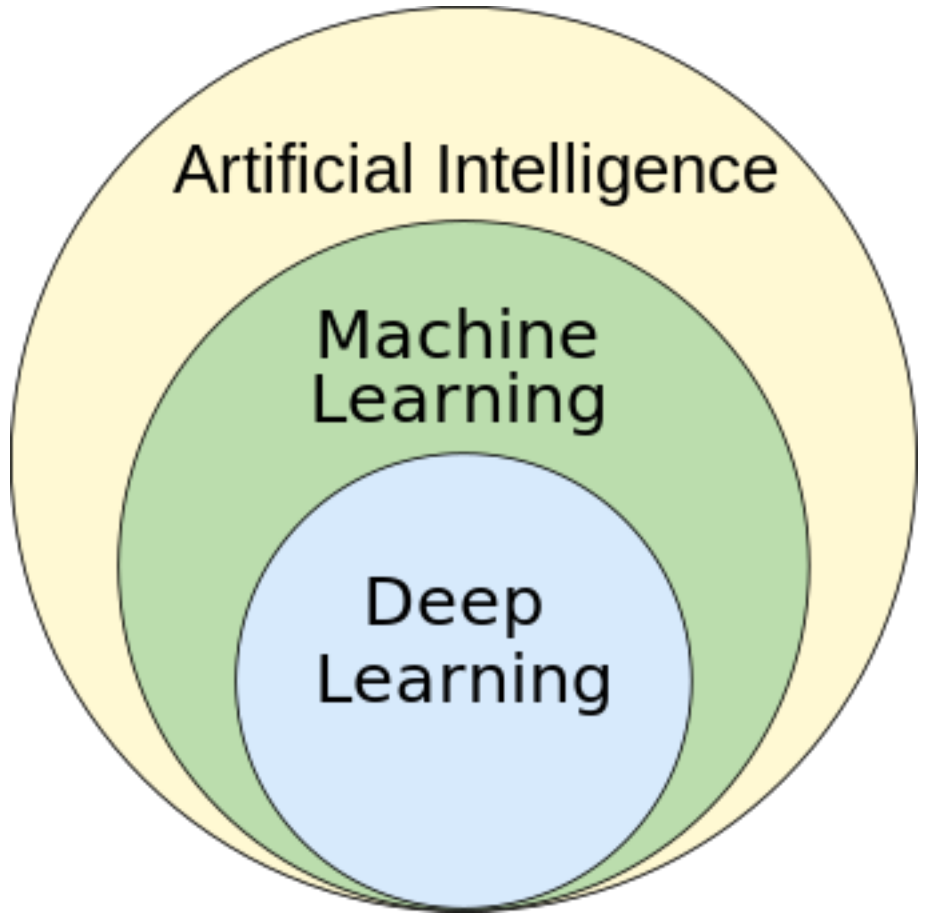
# How does AI work

- a massive amount of **data** is collected and applied to mathematical models - > recognise **patterns** and make **predictions** in a process called **training**[1]. The models continuously learn from and **adapt** to new data

- Tasks performed by AI: speech and image recognition, language processing, data analysis,  computer vision;

- The primary approach to building AI systems is through machine learning (ML)[2]:

    - Supervised learning: classification and regression
    - Unsupervised learning
    - Semi-supervised learning and Weakly supervised learning
    - Self-supervised learning
    - Reinforcement learning



[1] https://builtin.com/artificial-intelligence
[2] https://en.wikipedia.org/wiki/Machine_learning

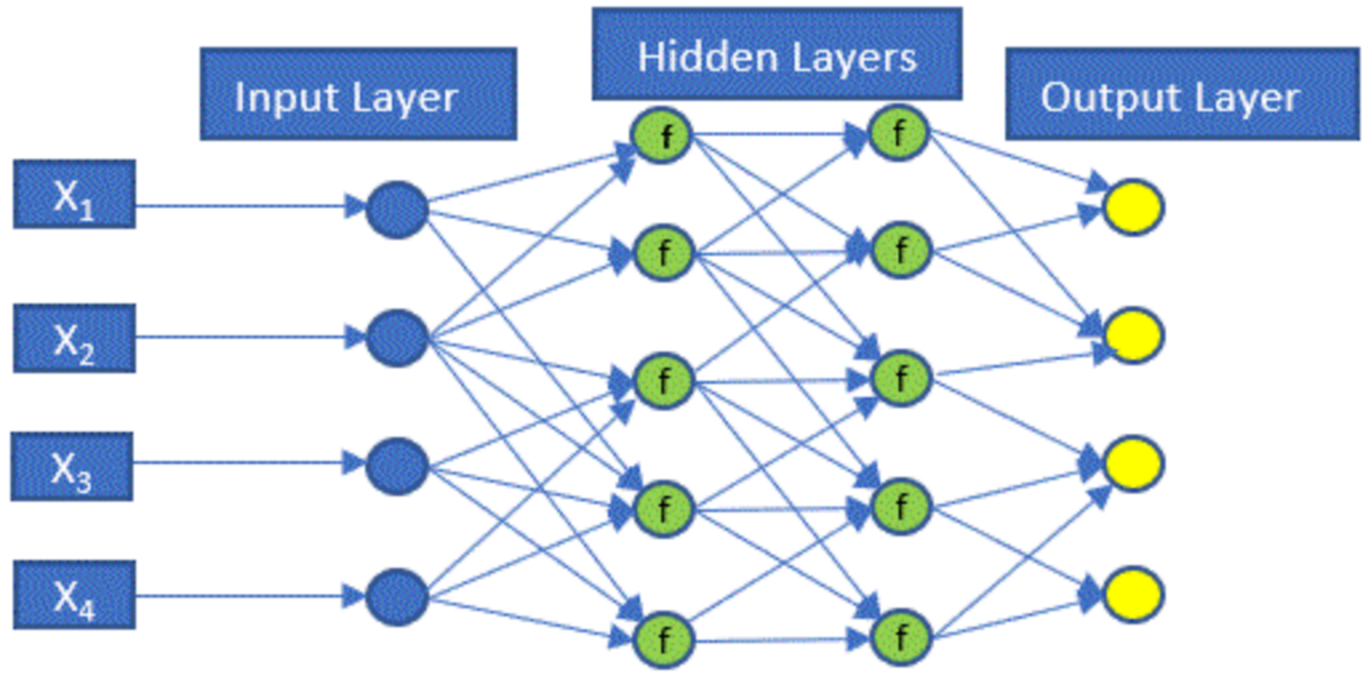
# Neural Networks

- Deep neural networks

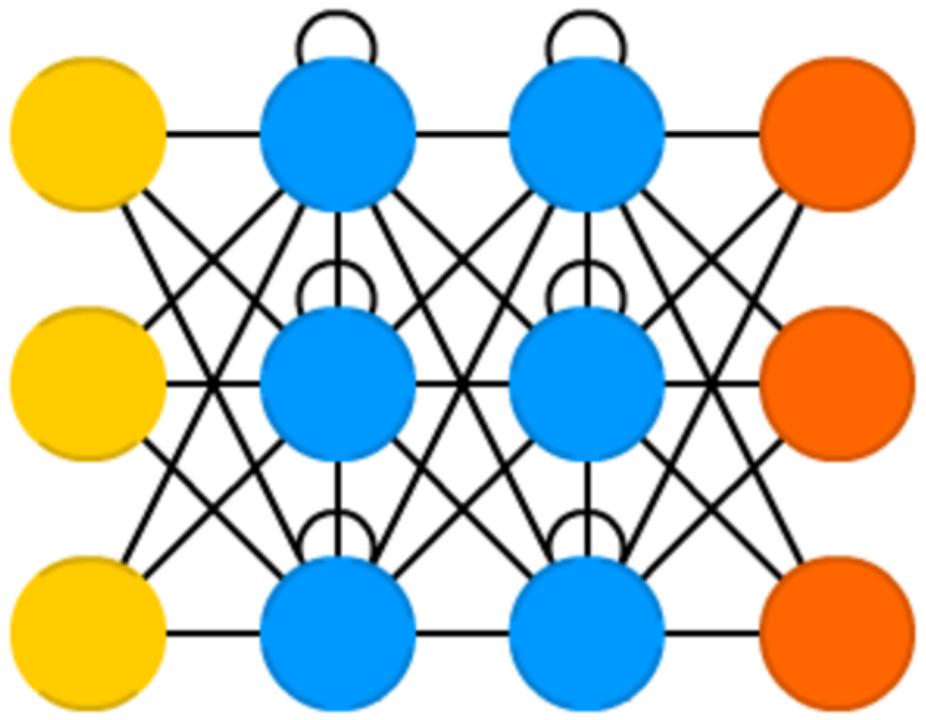
- Recurrent neural network (RNN)
  - characterised by direction of the flow of information between its layers
  - unsegmented, connected handwriting recognition
  - speech recognition

- feed forward neural network
  - a multi-layer neural network as all information is only passed forward

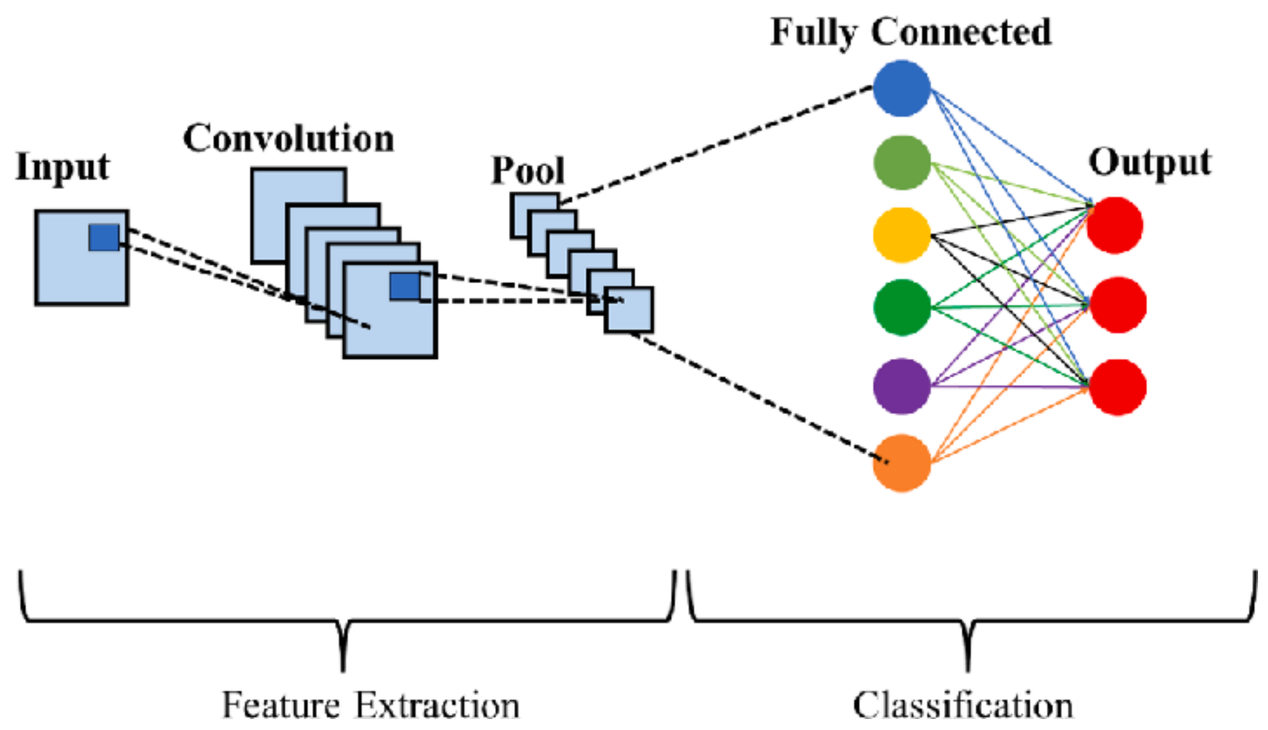
- Convolutional neural network (CNN)
  - image and video recognition,
  - recommender systems,
  - image classification,
  - image segmentation,
  - medical image analysis,
  - natural language processing,
  - brain–computer interfaces,
  - financial time series.



Deep neural network - layers nr > 1



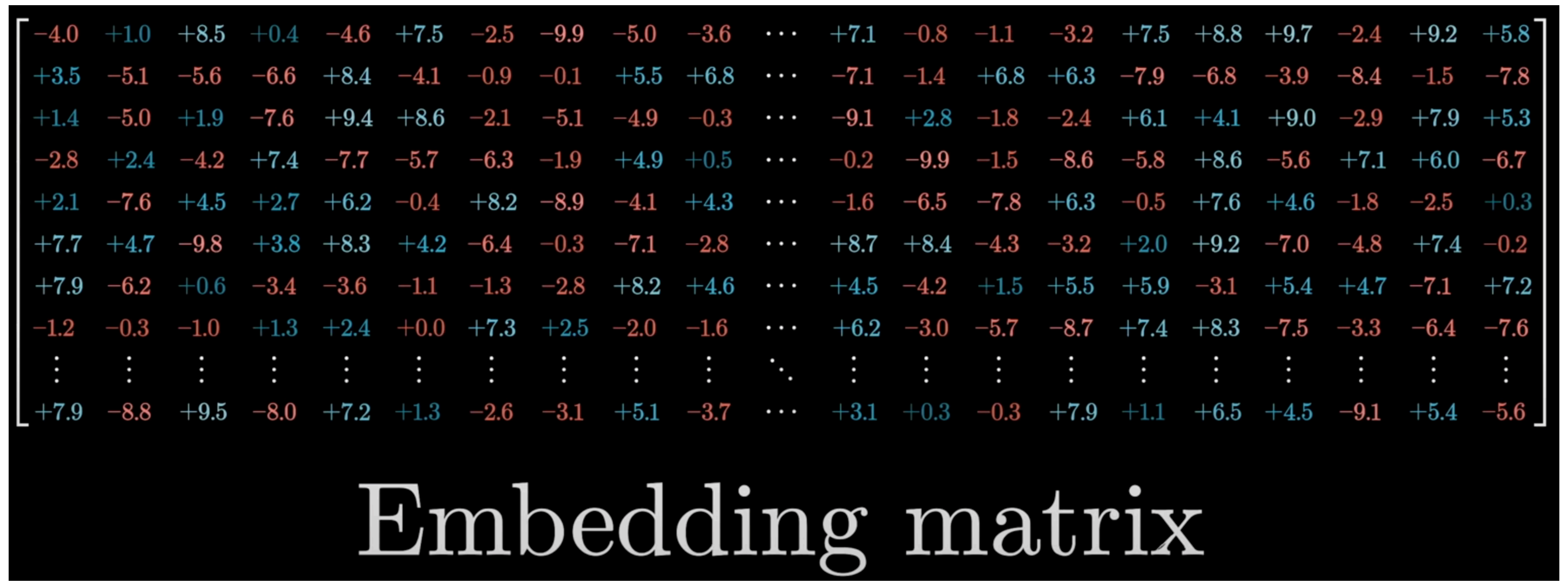
Recurrent Neural Network



Convolutional Neural Network

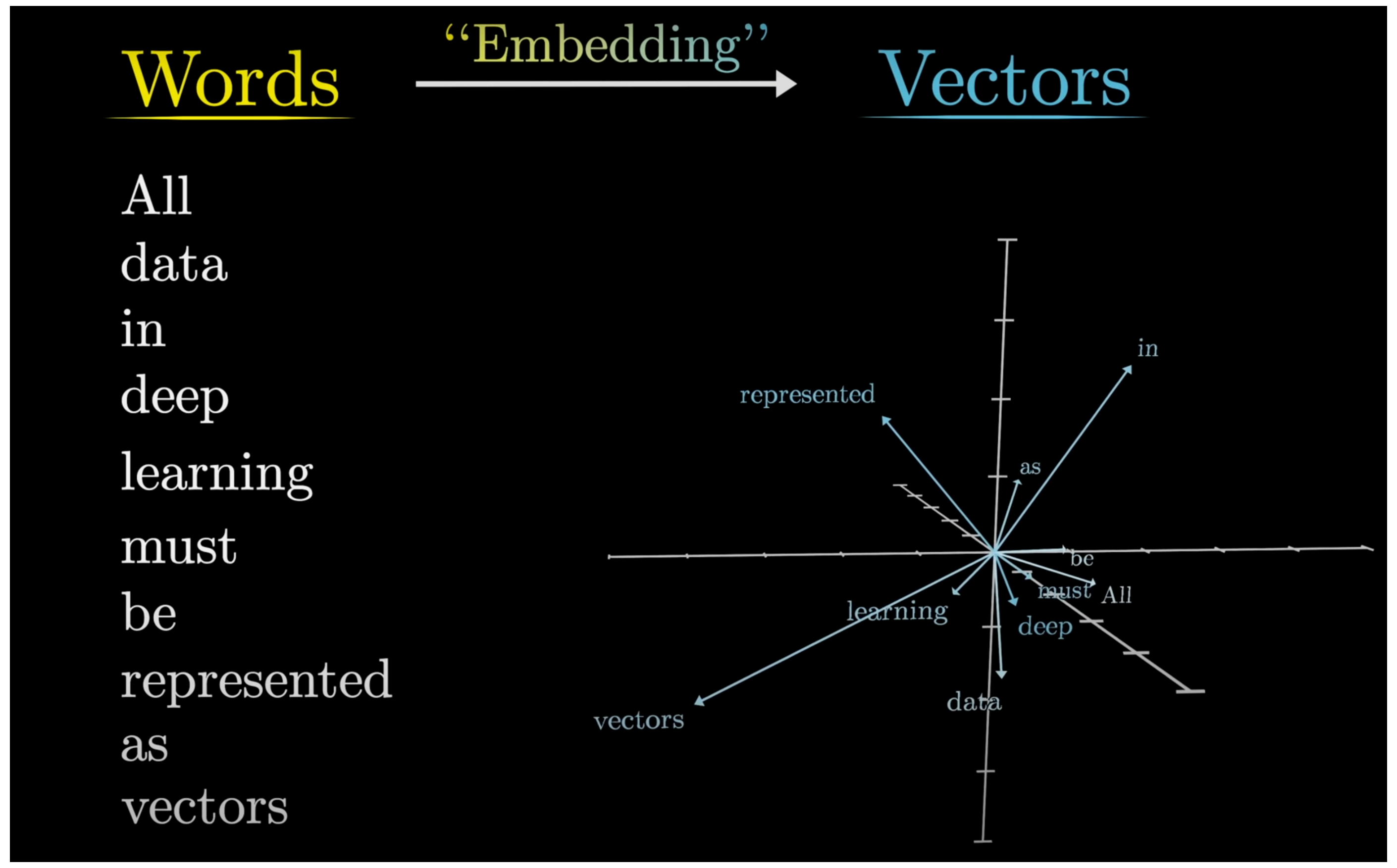
* The mostly complete chart of Neural Networks, explained:
https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464

# How does it work, actually?



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

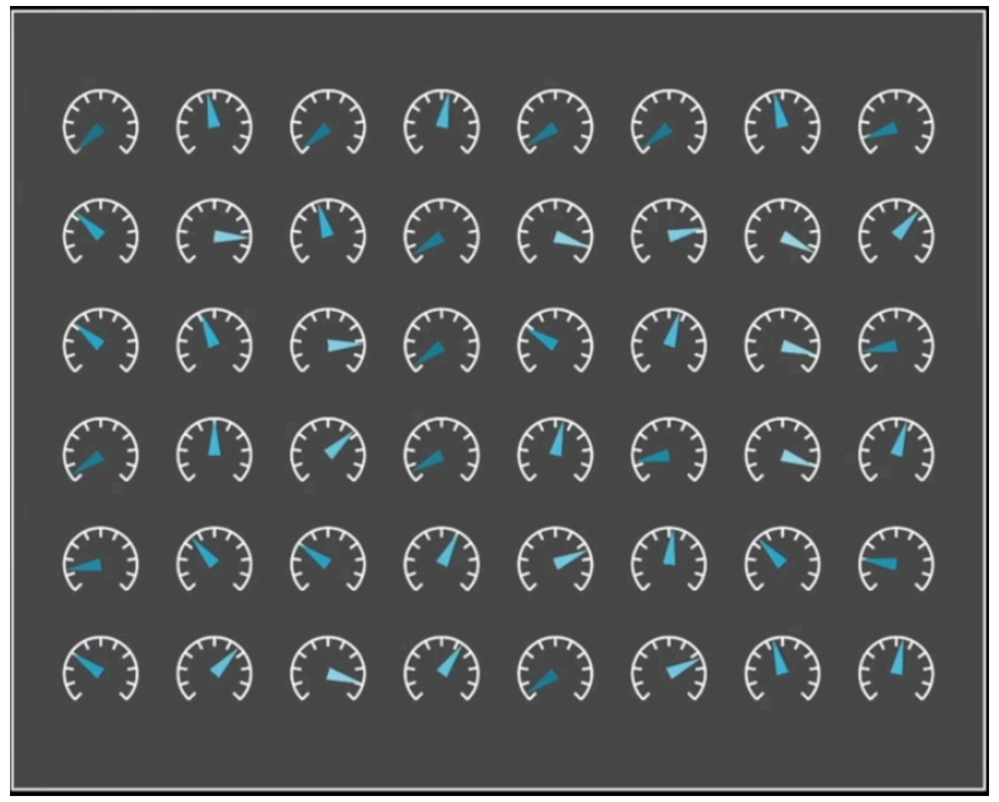
Embedding matrix

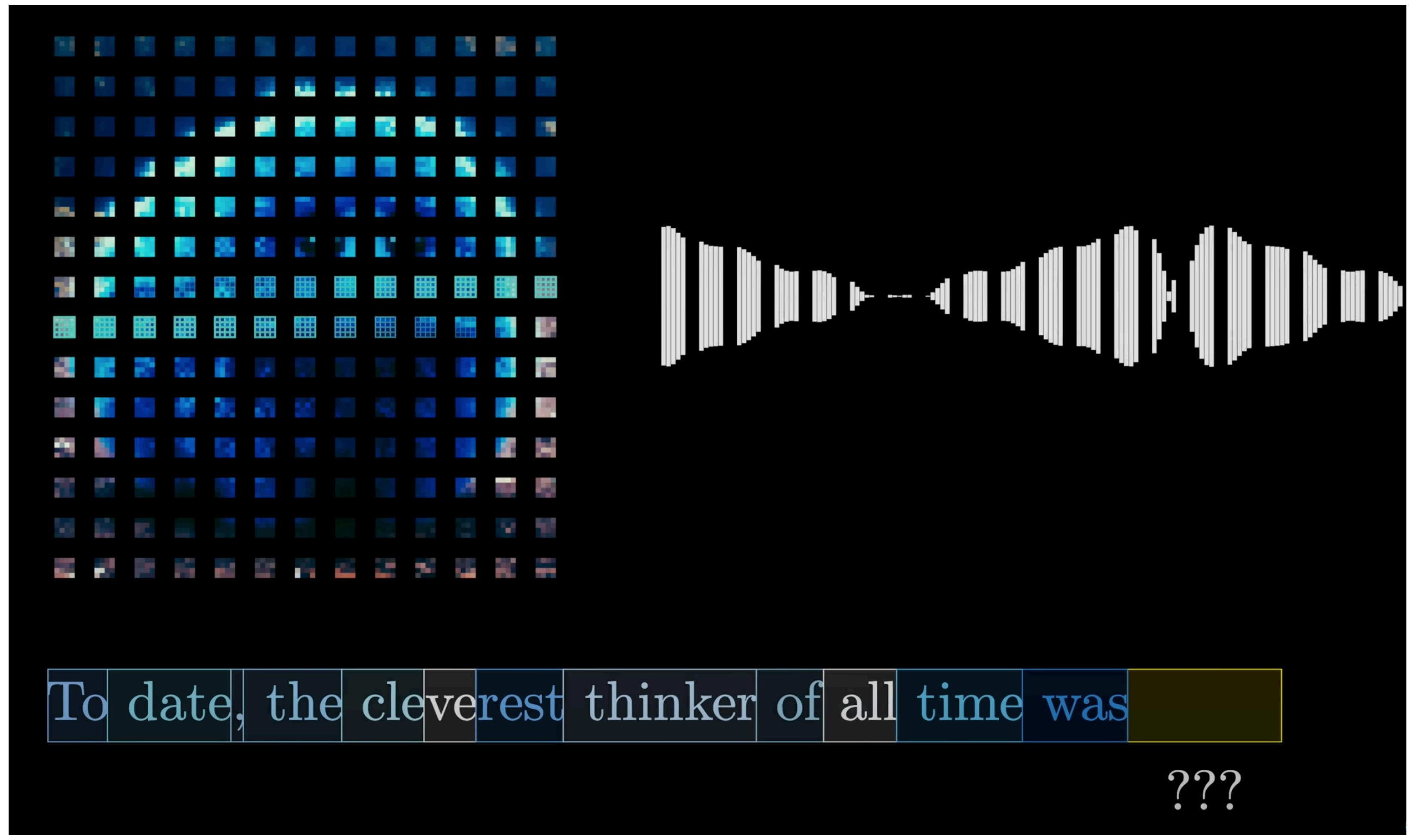


Words "Embedding" Vectors

All
data
in
deep
learning
must
be
represented
as
vectors

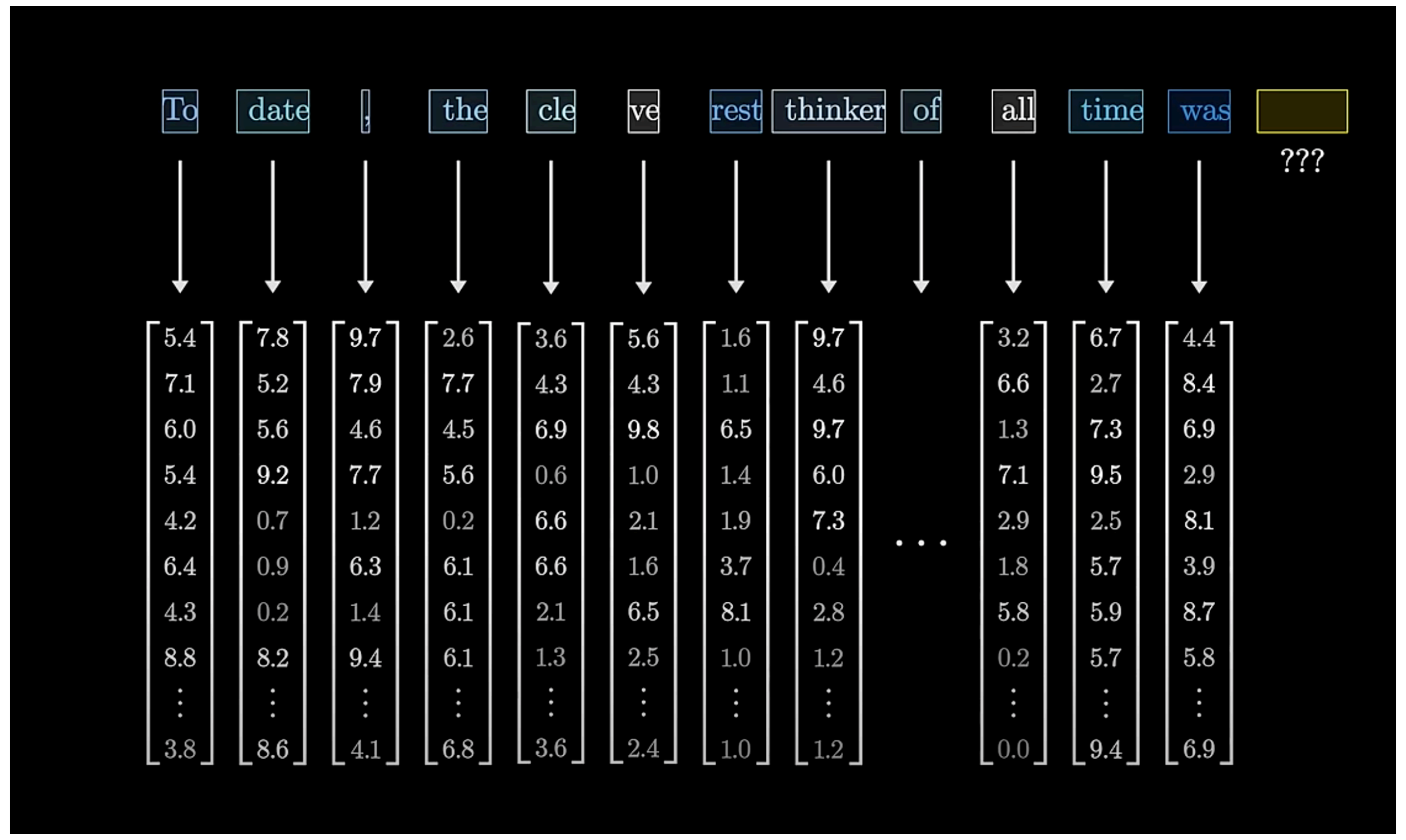
Weights/parameters



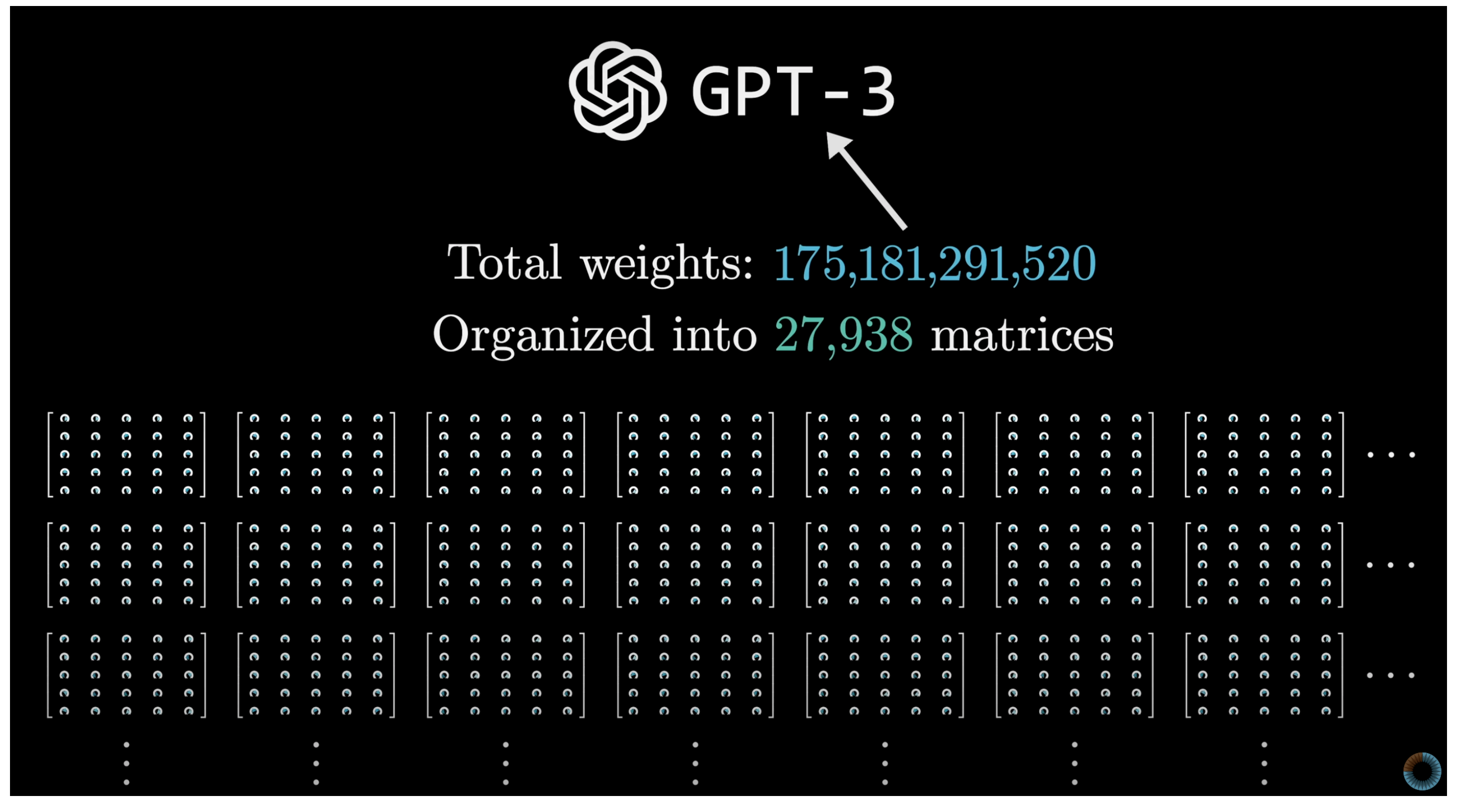


To date, the cleverest thinker of all time was

???



$$\begin{array}{ccccccccccccc}
\text{To} & \text{date} & , & \text{the} & \text{cle} & \text{ve} & \text{rest} & \text{thinker} & \text{of} & \text{all} & \text{time} & \text{was} & ??? \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \\
5.4 & 7.8 & 9.7 & 2.6 & 3.6 & 5.6 & 1.6 & 9.7 & & 3.2 & 6.7 & 4.4 & \\
7.1 & 5.2 & 7.9 & 7.7 & 4.3 & 4.3 & 1.1 & 4.6 & & 6.6 & 2.7 & 8.4 & \\
6.0 & 5.6 & 4.6 & 4.5 & 6.9 & 9.8 & 6.5 & 9.7 & & 1.3 & 7.3 & 6.9 & \\
5.4 & 9.2 & 7.7 & 5.6 & 0.6 & 1.0 & 1.4 & 6.0 & & 7.1 & 9.5 & 2.9 & \\
4.2 & 0.7 & 1.2 & 0.2 & 6.6 & 2.1 & 1.9 & 7.3 & \cdots & 2.9 & 2.5 & 8.1 & \\
6.4 & 0.9 & 6.3 & 6.1 & 6.6 & 1.6 & 3.7 & 0.4 & & 1.8 & 5.7 & 3.9 & \\
4.3 & 0.2 & 1.4 & 6.1 & 2.1 & 6.5 & 8.1 & 2.8 & & 5.8 & 5.9 & 8.7 & \\
8.8 & 8.2 & 9.4 & 6.1 & 1.3 & 2.5 & 1.0 & 1.2 & & 0.2 & 5.7 & 5.8 & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \\
3.8 & 8.6 & 4.1 & 6.8 & 3.6 & 2.4 & 1.0 & 1.2 & & 0.0 & 9.4 & 6.9 &
\end{array}$$



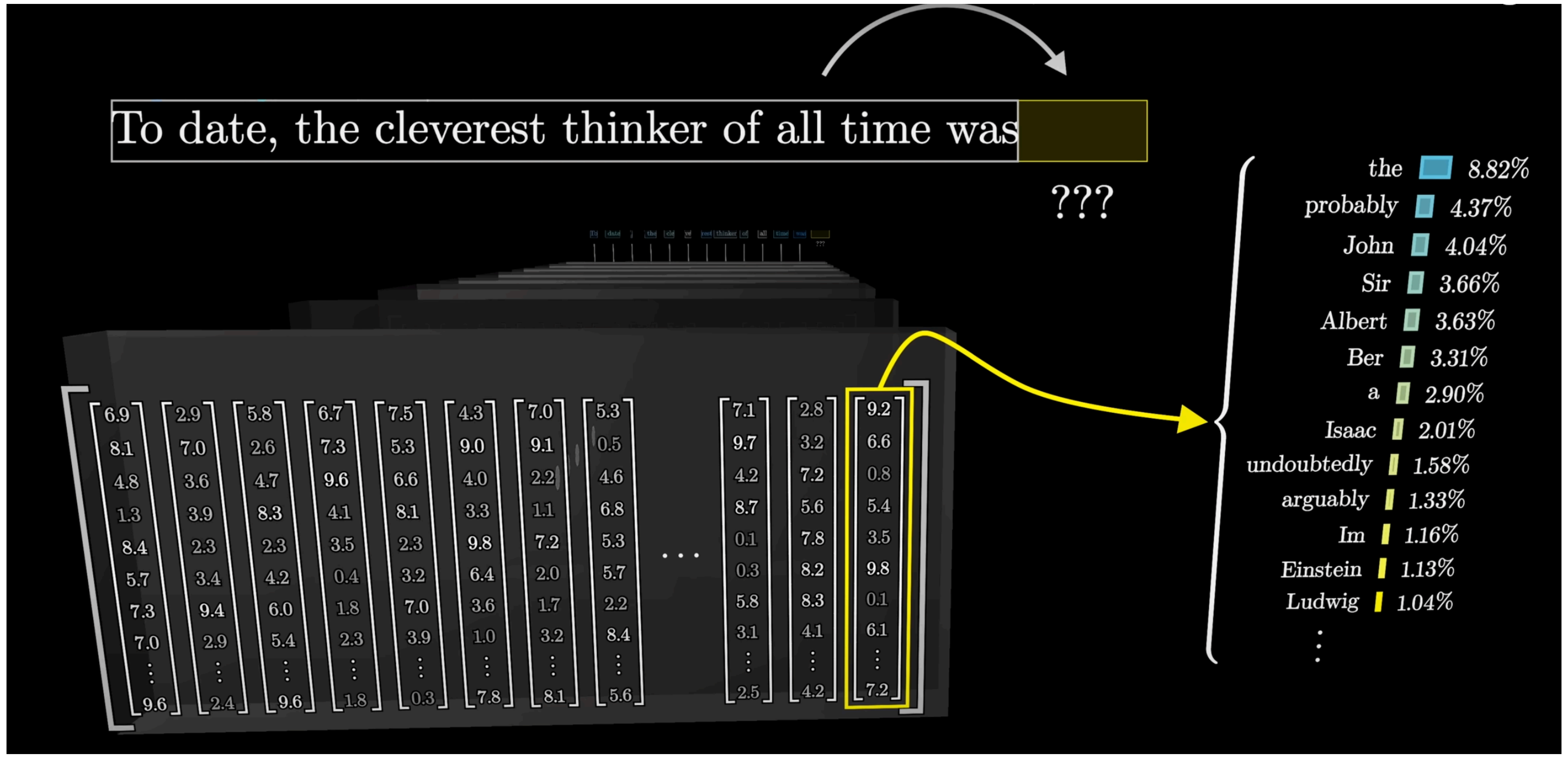
GPT-3

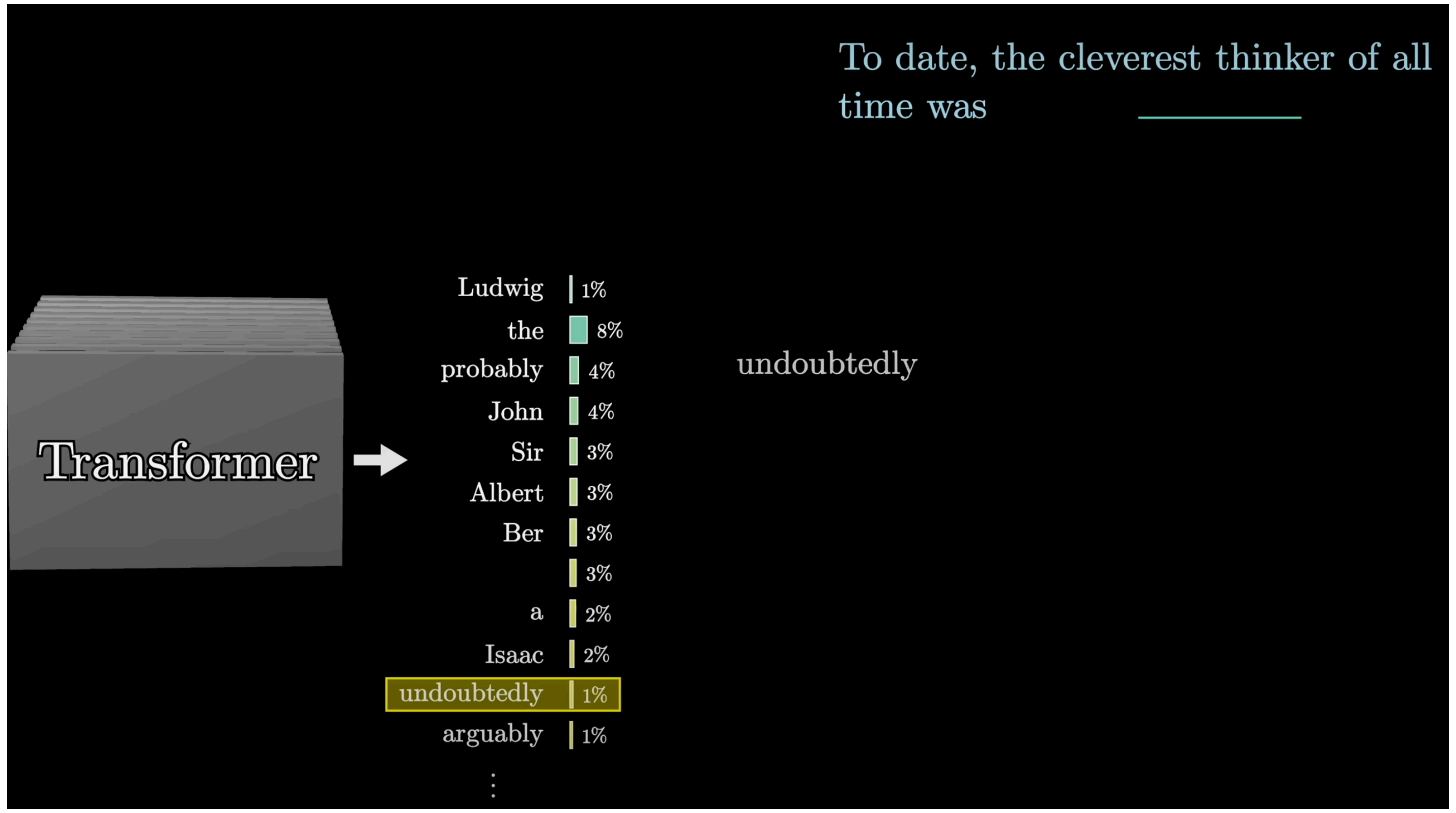
Total weights: 175,181,291,520

Organized into 27,938 matrices

3Blue1Brown: https://www.youtube.com/watch?v=wjZofJX0v4M

- Produces a probability distribution over all possible tokens that might come next





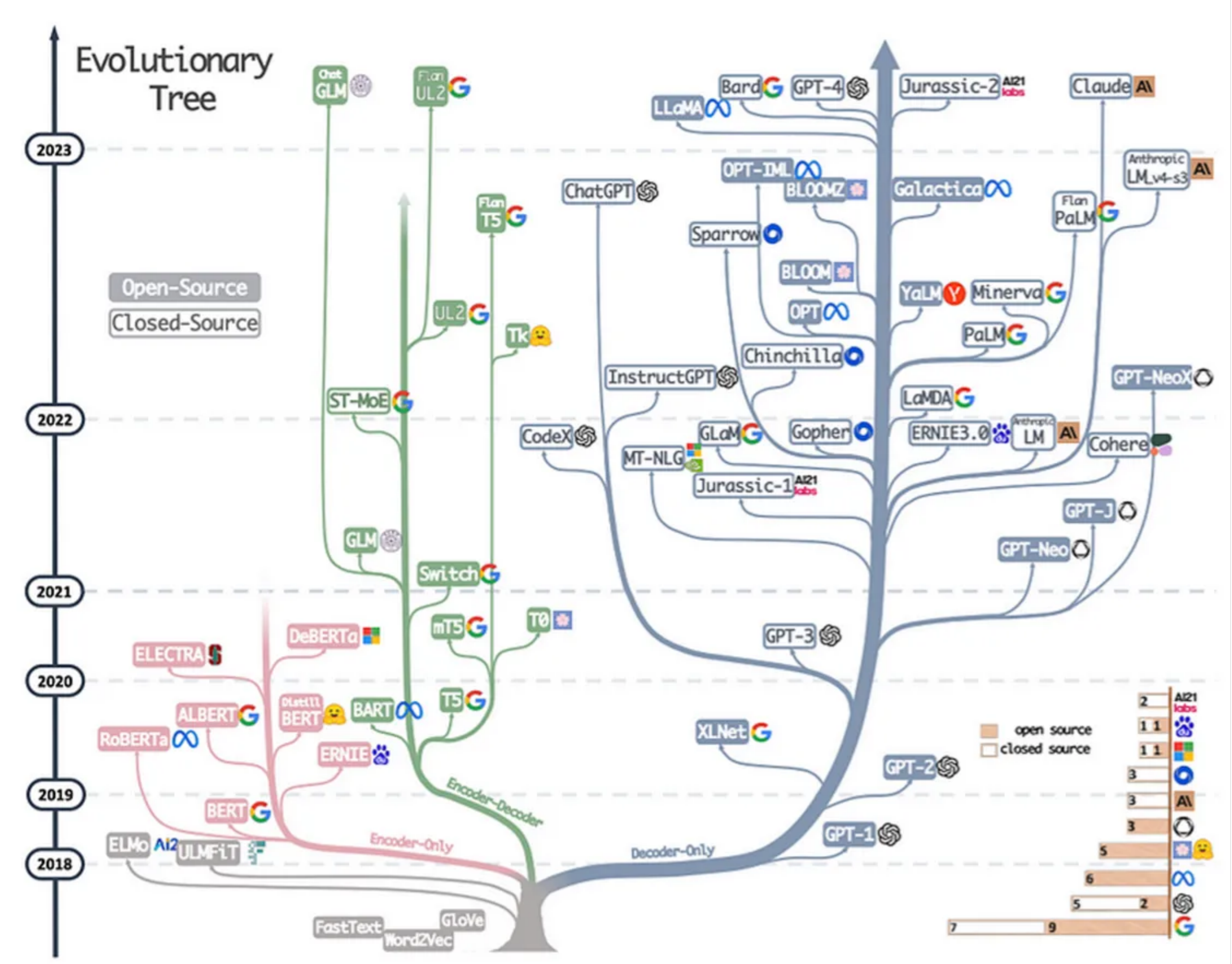
# Models

- Foundational models

  - any model that is trained on **broad data** (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a **wide range** of downstream tasks;
  - e.g.: for images DALL-E and Flamingo ,
  for music MusicGen,
  for robotic control RT-2
  for language Large Language Models (LLM)
  - e.g. OpenAI's "GPT-n" series, Google's BERT

  - foundation models are being built for astronomy, radiology, genomics, music, coding, times-series forecasting, mathematics.

* Source: https://en.wikipedia.org/wiki/Foundation_model



Evolutionary Tree of Large Language Models: From Word2Vec to GPT-4. Image credit: Yang, Jingfeng et. al

# Energy Consumption

- Hardware:

  CPU, GPU, TPU, IBMs NorthPole chip (256 cores, each of which contains its own memory)

- Model architecture: Neural Networks, Transformer, iterations -> training duration

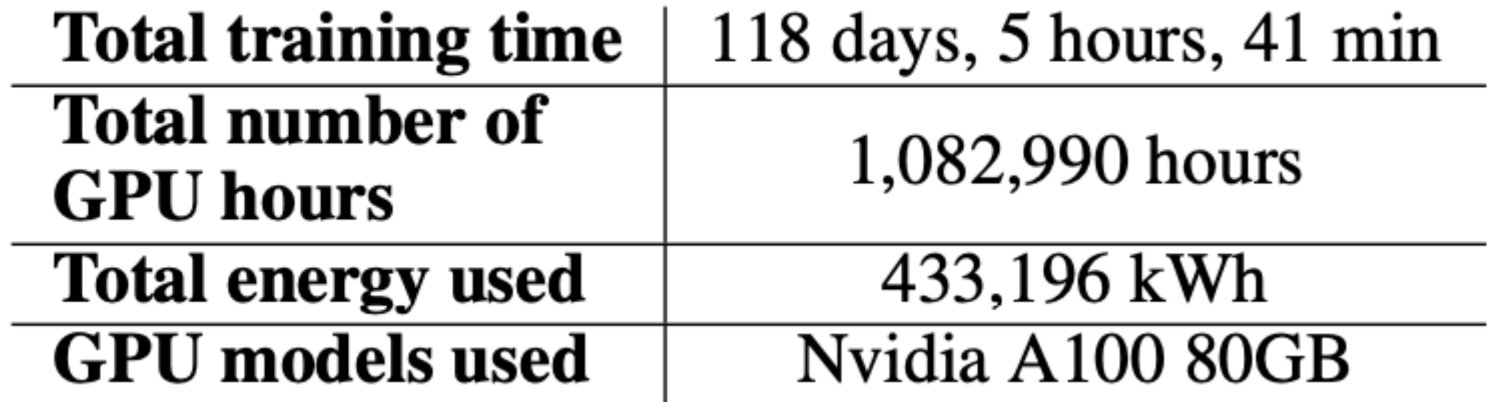
- Data: collection, storing, processing

- Location:  energy source type (fossil fuel, renewable, nuclear) for the data centres,

- Water

# Energy Consumption

- the carbon footprint of BLOOM, a 176-billion parameter language model, across its life cycle[1]



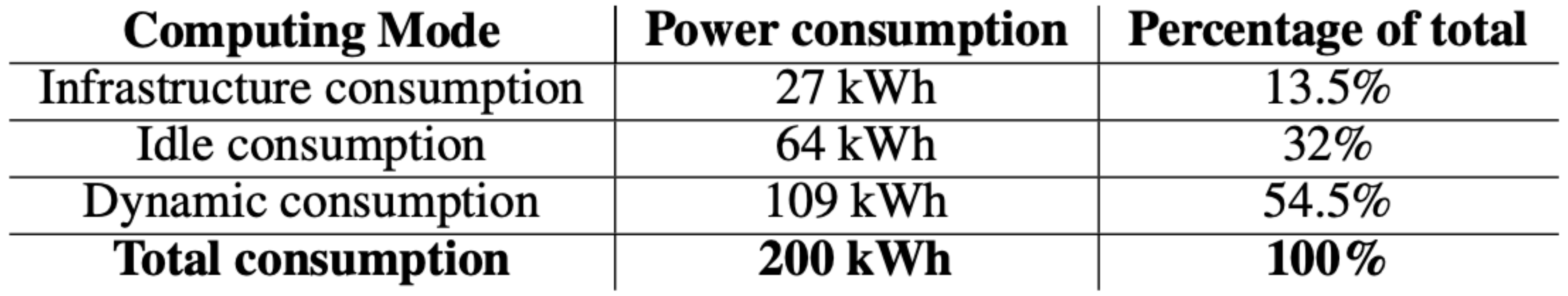
| **Total training time** | 118 days, 5 hours, 41 min |
|---|---|
| **Total number of GPU hours** | 1,082,990 hours |
| **Total energy used** | 433,196 kWh |
| **GPU models used** | Nvidia A100 80GB |

Key statistics about BLOOM model training

- the BLOOM model required a total of 1.08 million GPU hours on a hardware partition constituted of Nvidia A100 SXM4 GPUs with 80GB of memory, which have a thermal design power (TDP) of 400W

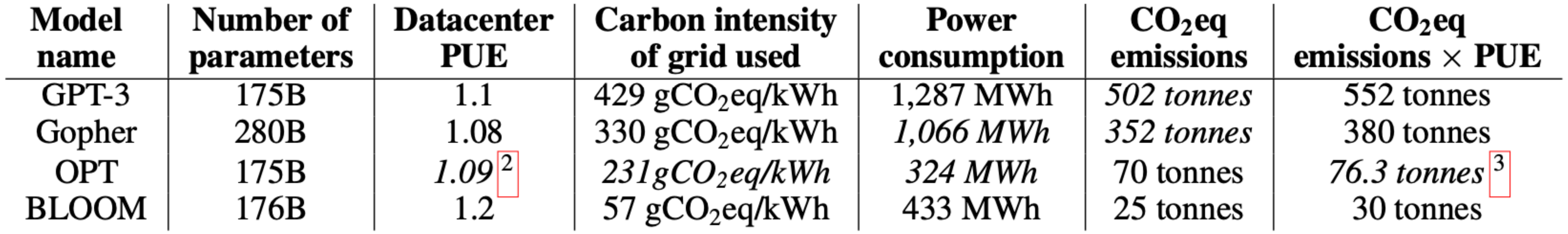
- During training: energy consumption of CPUs ~ 40 < GPUs - typically not as solicited during the model training process; GPUs however ~100% utilisation

- GPT-4: 1.7 trillion parameters , 13 trillion tokens (word snippets), 100 days, 25.000 NVIDIA A100 GPUs -> an estimated 50 GWh of energy usage during training



| Computing Mode | Power consumption | Percentage of total |
|---|---|---|
| Infrastructure consumption | 27 kWh | 13.5% |
| Idle consumption | 64 kWh | 32% |
| Dynamic consumption | 109 kWh | 54.5% |
| **Total consumption** | **200 kWh** | **100%** |

Idle Power Consumption



| Model name | Number of parameters | Datacenter PUE | Carbon intensity of grid used | Power consumption | $CO_2$eq emissions | $CO_2$eq emissions × PUE |
|---|---|---|---|---|---|---|
| GPT-3 | 175B | 1.1 | 429 $gCO_2eq/kWh$ | 1,287 MWh | *502 tonnes* | 552 tonnes |
| Gopher | 280B | 1.08 | 330 $gCO_2eq/kWh$ | *1,066 MWh* | *352 tonnes* | 380 tonnes |
| OPT | 175B | *1.09*[2] | *231 $gCO_2eq/kWh$* | *324 MWh* | 70 tonnes | *76.3 tonnes*[3] |
| BLOOM | 176B | 1.2 | 57 $gCO_2eq/kWh$ | 433 MWh | 25 tonnes | 30 tonnes |

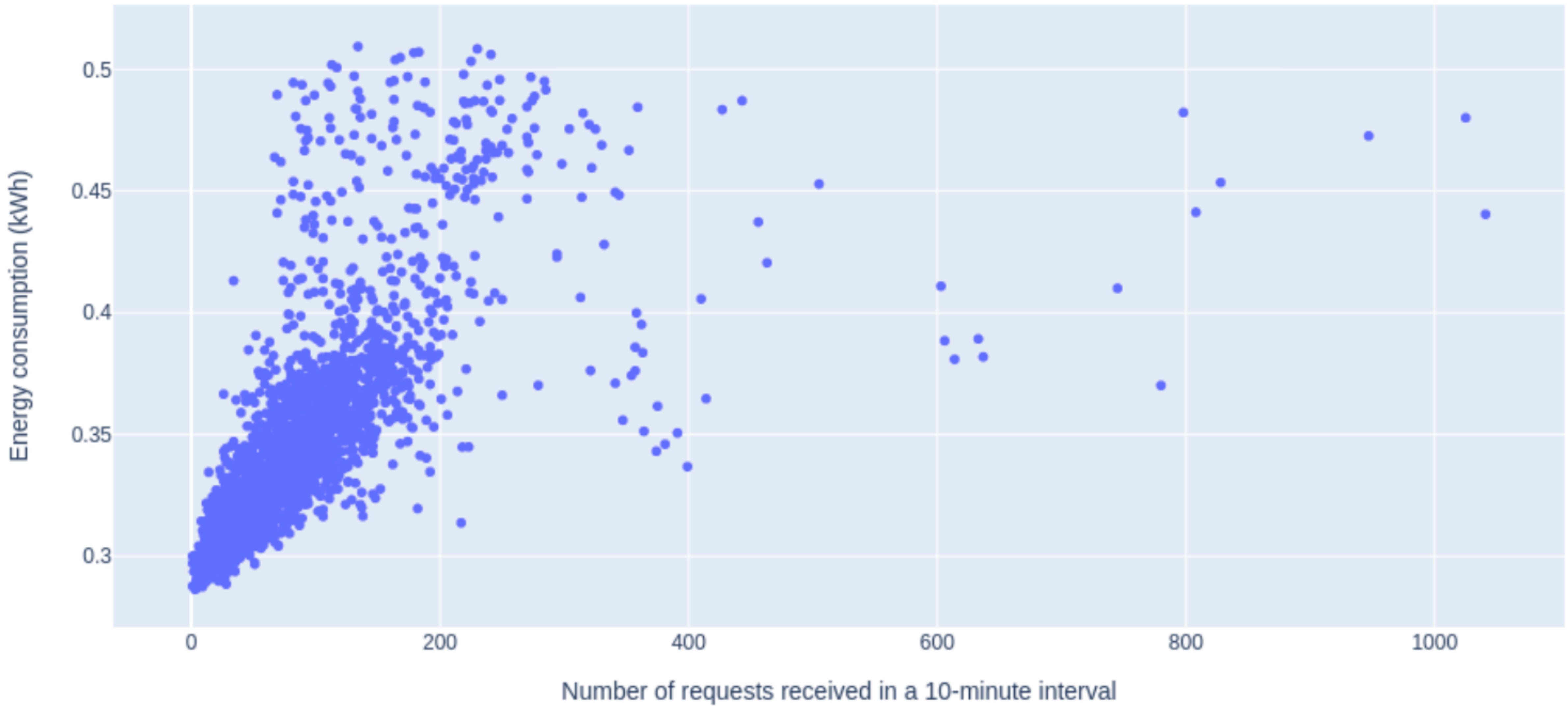
Comparison of carbon emissions between BLOOM and similar LLMs. Numbers in italics have been inferred based on data provided in the papers describing the models

- Alexandra Sasha Luccioni, Sylvain Viguier, Anne-Laure Ligozat. 2022. Estimating The Carbon Footprint Of Bloom, A 176B Parameter Language Model, (Https://Doi.Org/10.48550/Arxiv.2211.02001)

# Deployment and Inference Energy Consumption

- 78-171W per GPU, which is significantly less than the TDP of this type of GPUs (400W) [1]

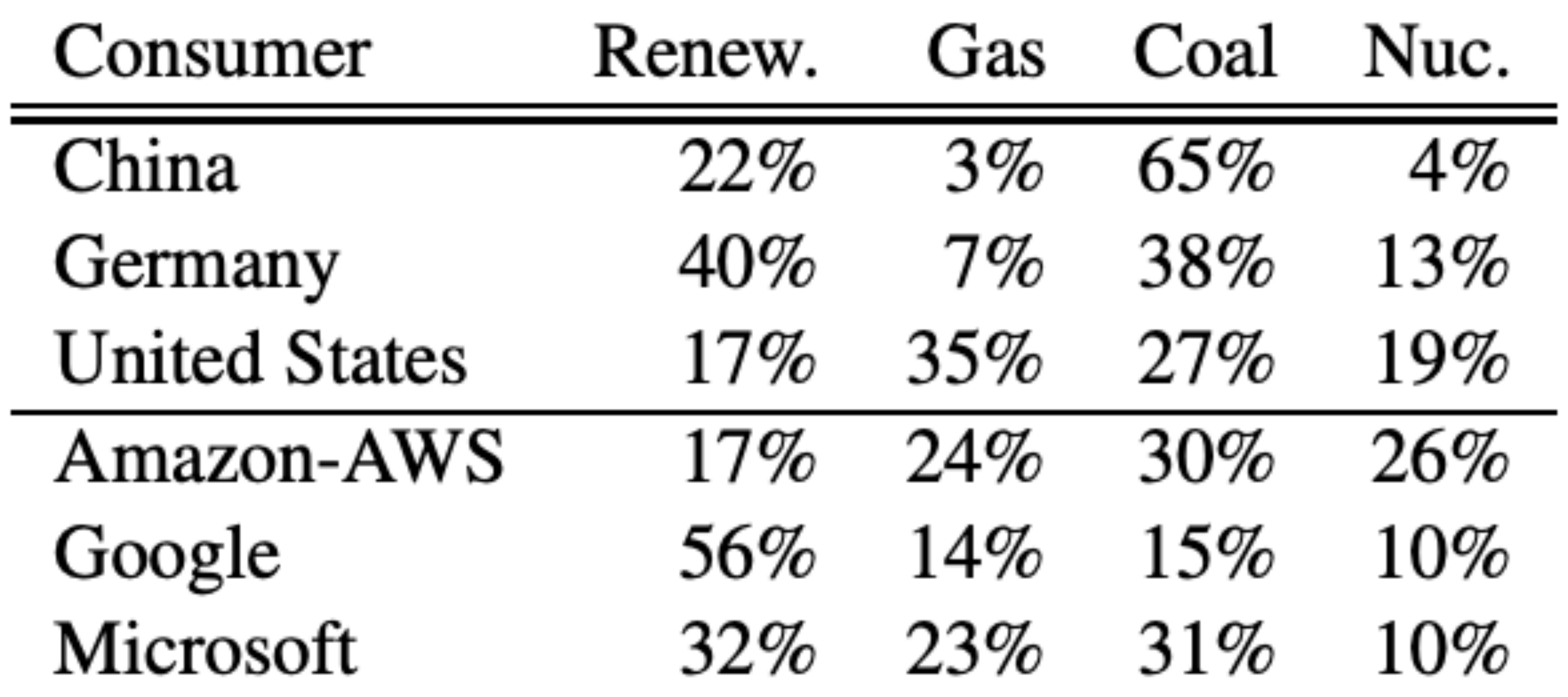
- RAM 2% , CPU (18.5 kWh) , GPU 75.3% of the total measured consumption - on GCP



The quantity of energy used by the GCP instance (on the y axis) versus the number of requests received by the instance in a 10 minute interval (on the x axis). It can be seen that even when zero requests are received by the instance in this time span (bottom left of the graph), the energy consumption remains at approximately 0.28 kWh.

- even when there are almost no incoming requests during a 10 minute interval, there is still ~0.28kWh of energy that is consumed during this interval, which represents the energy consumption of the model when it is not responding to any user requests-

- 242 kWh/day; 88300 kWh/year - just deployed model



| Consumer | Renew. | Gas | Coal | Nuc. |
|---|---|---|---|---|
| China | 22% | 3% | 65% | 4% |
| Germany | 40% | 7% | 38% | 13% |
| United States | 17% | 35% | 27% | 19% |
| Amazon-AWS | 17% | 24% | 30% | 26% |
| Google | 56% | 14% | 15% | 10% |
| Microsoft | 32% | 23% | 31% | 10% |

Percent energy sourced from: Renewable (e.g. hydro, solar, wind), natural gas, coal and nuclear for the top 3 cloud compute providers (Cook et al., 2017), compared to the United States,4 China5 and Germany (Burger, 2019).

- annual energy consumption in 2020 was 15.4 TWh for Google and 10.8 TWh for Microsoft [2]

# References

- Alexandra Sasha Luccioni, Sylvain Viguier, Anne-Laure Ligozat. 2022. Estimating The Carbon Footprint Of Bloom, A 176B Parameter Language Model, (Https://Doi.Org/10.48550/Arxiv.2211.02001)

- David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, And Jeff Dean. The Carbon Footprint Of Machine Learning Training Will Plateau, Then Shrink

- The Carbon Footprint Of Machine Learning Training Will Plateau, Gary Cook, Jude Lee, Tamina Tsai, Ada Kongn, John Deans, Brian Johnson, Elizabeth Jardim, And Brian Johnson. 2017. Clicking Clean: Who Is Winning The Race To Build A Green Internet? Technical Report, Greenpeace.

- Bruno Burger. 2019. Net Public Electricity Generation In Germany In 2018. Technical Report, Fraunhofer Institute For Solar Energy Systems Ise.

- Bommasani, Rishi; et al. (18 August 2021). On the Opportunities and Risks of Foundation Models (Report). arXiv:2108.07258.

- JINGFENG YANG, HONGYE JIN, RUIXIANG TANG,  XIAOTIAN HAN, QIZHANG FENG, HAOMING JIANG, BING YIN, XIA HU. 2023. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. https://arxiv.org/pdf/2304.13712

- Sarah Gao, Andrew Kean Gao, 2023.  On the Origin of LLMs: An Evolutionary Tree and Graph for 15,821 Large Language Models. https://arxiv.org/abs/2307.09793