AI and Large Language Models

October 2024

Accelerate Programme for Scientific Discovery

ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES

Welcome!

About the course

- Introduction to LLMs and how they work
- What is available?
- How to use them

Code and slides available on:

Accelerate Science GitHub repo

ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES

Today's Schedule

- Introduction to the Accelerate Science Programme
- Introduction to LLMs
- BREAK
- Augmenting LLMs
  - Finetuning
  - Prompting
- LUNCH
- Small language models
- Quantization
- Data Ethics
- BREAK
- What's out there?
  - Popular models
  - APIs
  - No Code

ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES

Who are we?

Katie Light

Caroline Chater

Catherine Breslin

Ryan Daniels

ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES

Accelerate Science Programme

"Accelerate Science pursues research at the interface of AI and the sciences, generating new scientific insights and developing AI methods that can be deployed to advance scientific knowledge."

ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES

Accelerate Science Programme

LLM Workshop Goals

We want to:

- Support researchers across the university to use AI – that's LLMs in this group.
- Better understand the challenges that researchers face.
- Identify what training courses or software resources we might want the Accelerate Science Programme to create.
  - Including shared code
- Start to build a community of like-minded researchers across the university.

ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES

Introduction to LLMs

ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES

Introduction to LLMs

Overview

- What is a large language model?
- How do they work?
- How are they trained?

But first…

What do you know about LLMs…?

Introduction to LLMs

What do you know about LLMs…?

- How do they work?
- Can you name any?
- What can you use them for?
- What are potential problems?
- What Can YOU potentially use them for?

What is a Large Language Model?

Introduction to LLMs

First, some definitions…

- ***Model*** The thing you train and use to generate text.
- ***Architecture*** The high-level structure of the model.
- ***Parameters*** (or weights) The internal knobs and dials that get altered during the course of training.
- ***Context*** The input to the LLM.
- ***Embedding*** A high-dimensional representation of text.
- ***Token*** The smallest unit of text fed into a model.
- ***Pretraining*** The first stage of training the LLM, and usually the most expensive.
- ***Finetuning*** Training the LLM to do a particular task.
- ***Inference*** The act of calling the model for text generation.
- ***Quantization*** Reducing the size of a model.
- ***Source*** The availability of the model (can be open or closed).
- ***Modality*** Defined by the type of data the model is trained on (e.g. text or images).
- ***API*** (Application Programming Interface) The means by which you can interact with a model.

What are LLMs?

- LLMs are conditioned on a massive amount of text.
- Given some input tokens, and given what we know from the training data, what is the most likely next token?
- An analogy…

https://www.youtube.com/watch?v=zjkBMFhNj_g

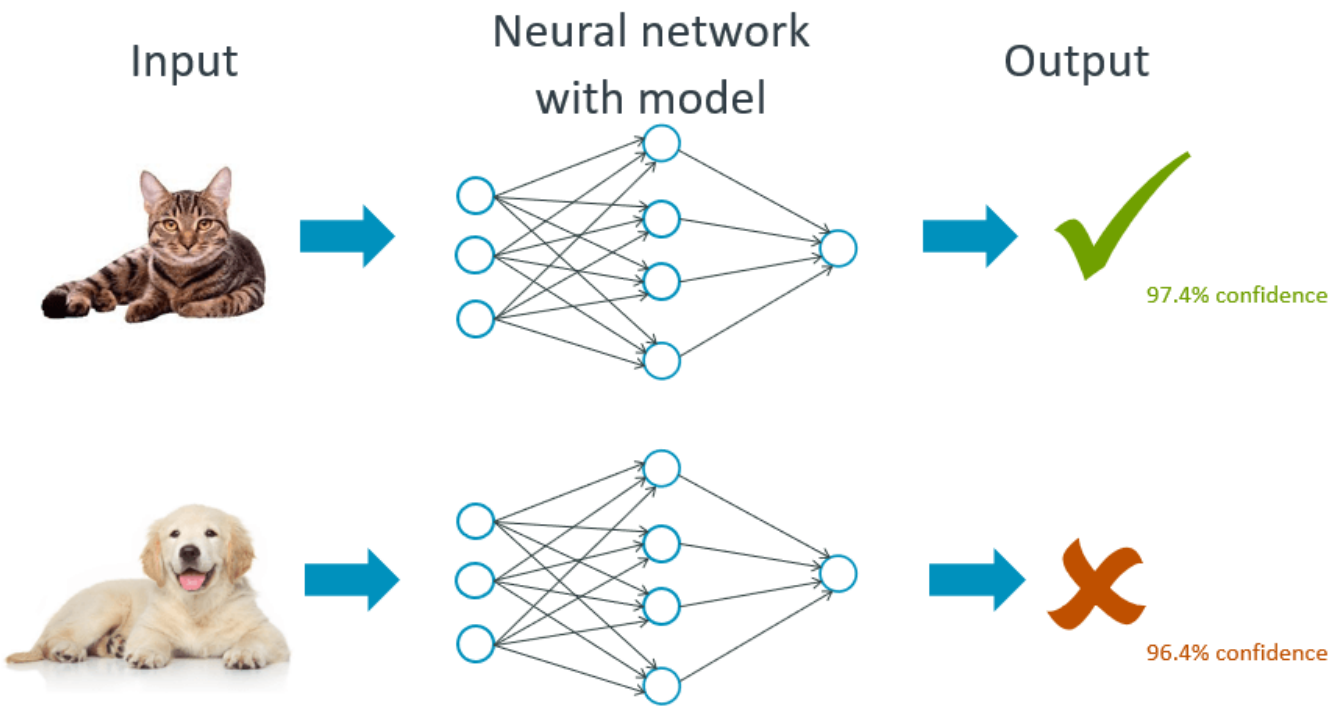
What is a large language model?

A language model is a high-parameter model that is trained on massive amounts of text with the end goal of some kind of text prediction or generation.

- Not all language models have the same architecture
- Not all language models are built for next-token prediction
  - GPT is a next token predictor
  - BERT is a masked-language model
- Not all language models are chatbots.

What is a large language model?

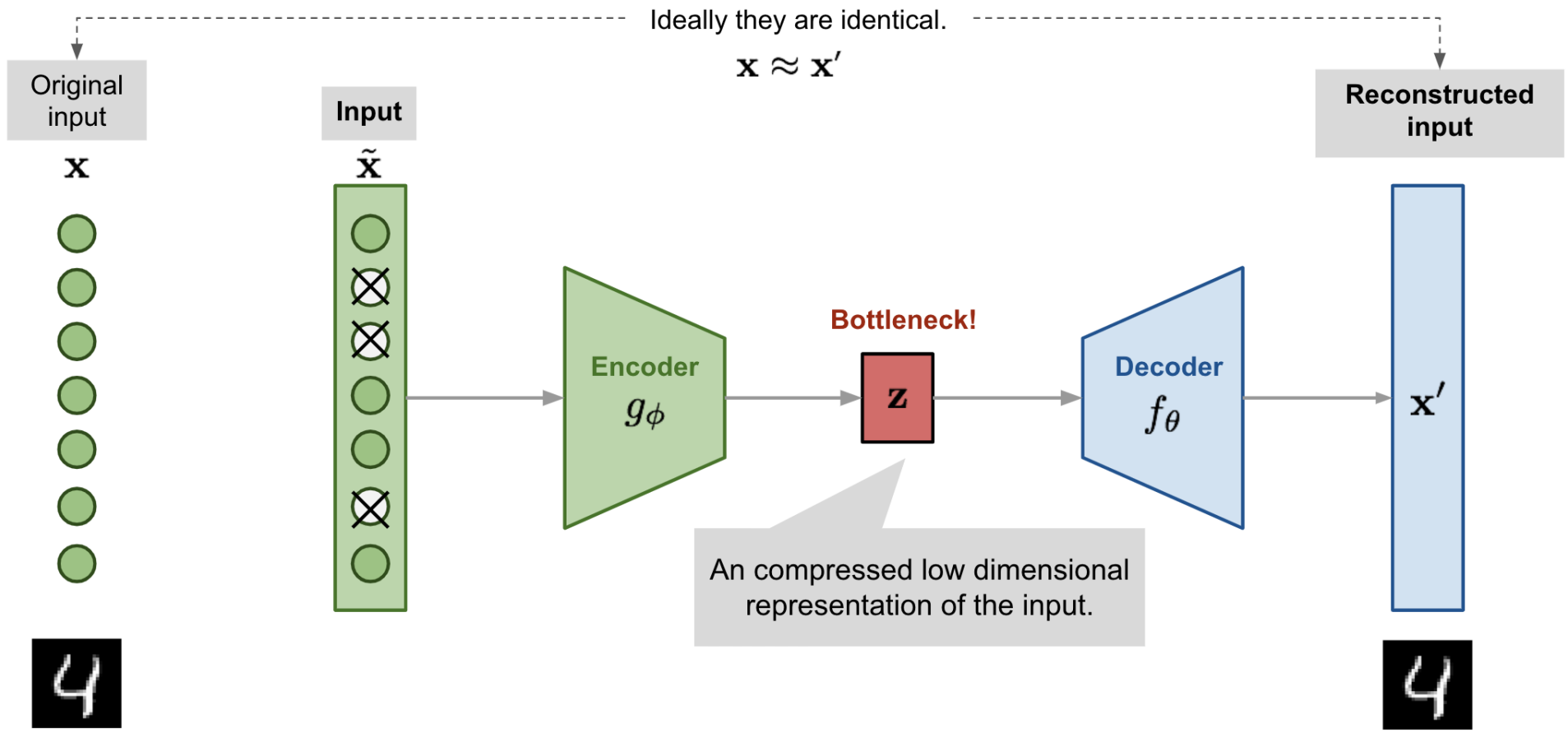
A normal neural network



1. Picture goes in
2. Pixel values propagate through the weights of the network
3. Network outputs either “cat” or “dog”
4. Update the weights of the network based on how wrong the prediction is
5. Do many times

What is a large language model?

Encoder vs Decoder



Tokenization

The machine starts from scratch

```
text = "The cat sat on the mat."
tokens = text.split()
print(tokens)


# Output: ['The', 'cat', 'sat', 'on', 'the', 'mat.']
```

ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

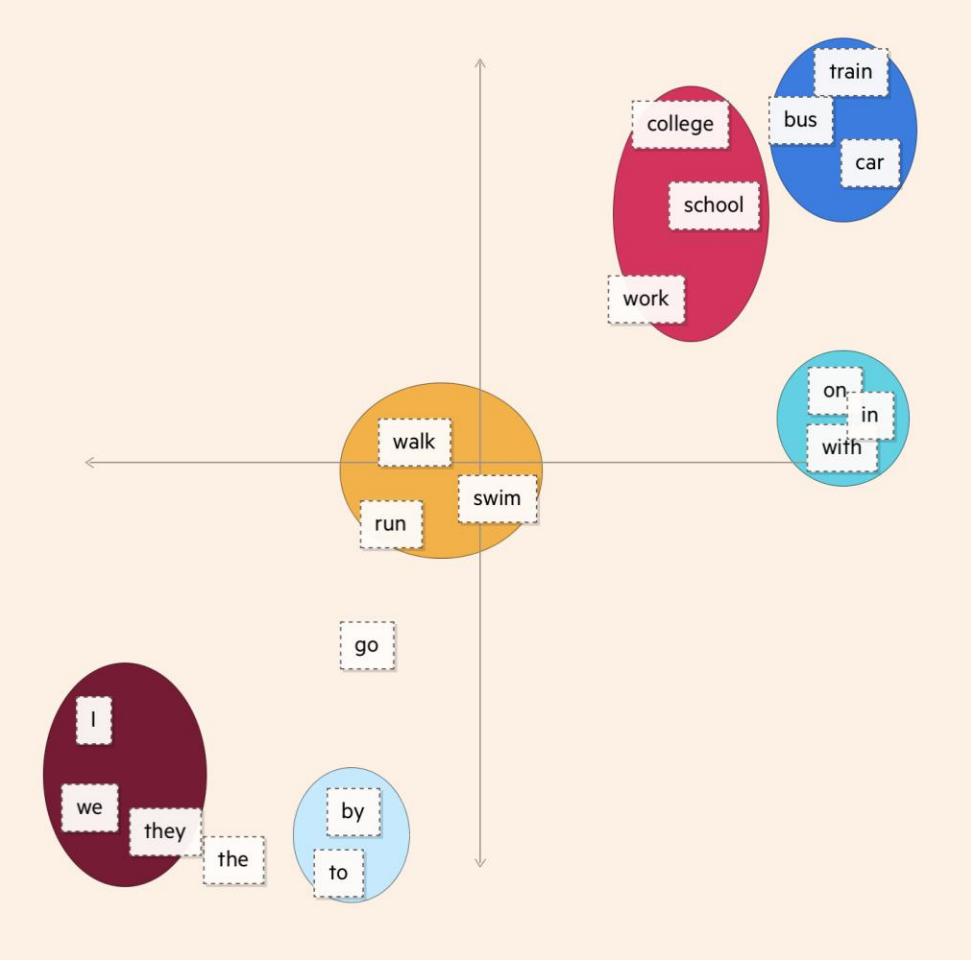
SCHMIDT FUTURES

Tokenization

The machine starts from scratch

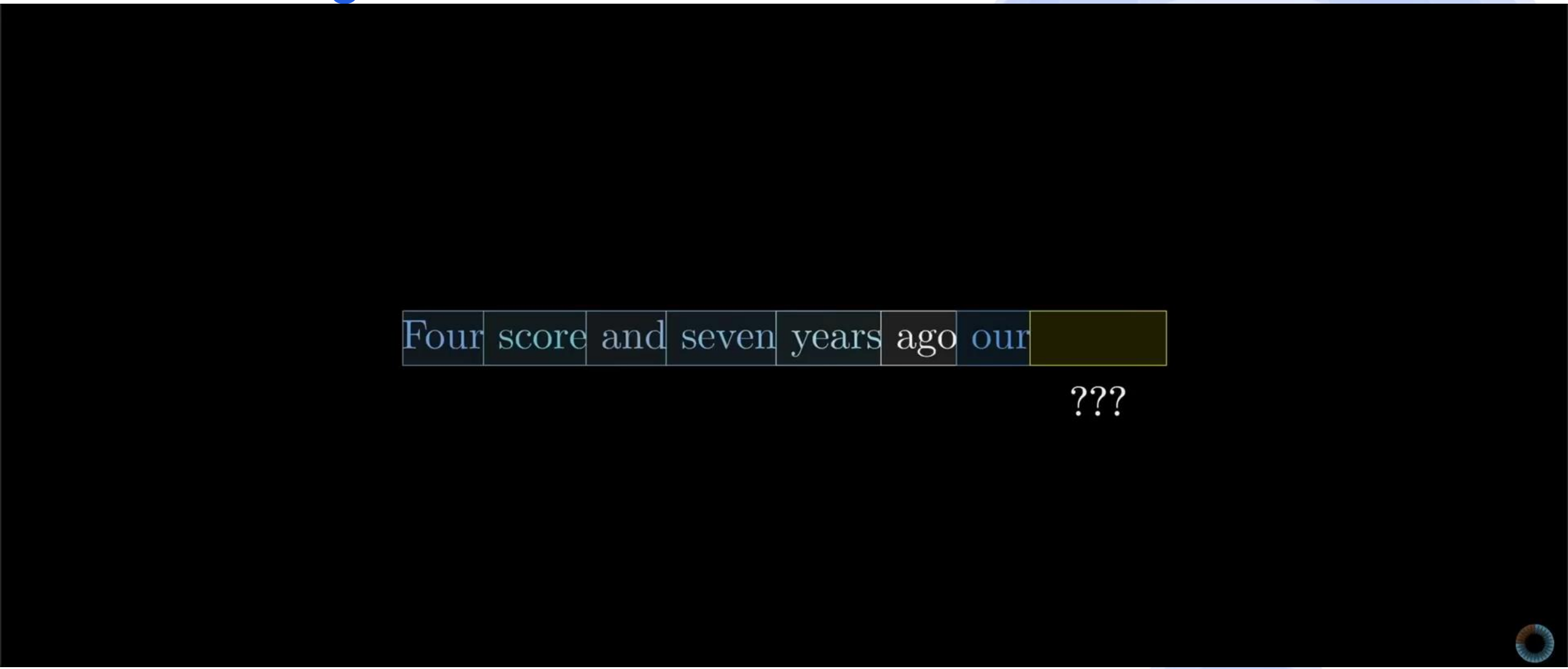
```
# Build vocabulary
vocab = sorted(set(tokens))

# Create token to index mapping
token_to_index = {token: index for index, token in enumerate(vocab)}

# Create index to token mapping
index_to_token = {index: token for token, index in token_to_index.items()}
print(token_to_index)

# Output: {'cat': 0, 'mat': 1, 'on': 2, 'sat': 3, 'the': 4}

# Convert tokens to indices
indices = [token_to_index[token] for token in tokens]
print(indices)
# Output: [4, 0, 3, 2, 4, 1]
```

Embeddings



- Embeddings are vectors (lists of numbers)
- https://ig.ft.com/generative-ai/
- There are a number of different types of embeddings:
  - Learned
  - Positional
  - Combinations

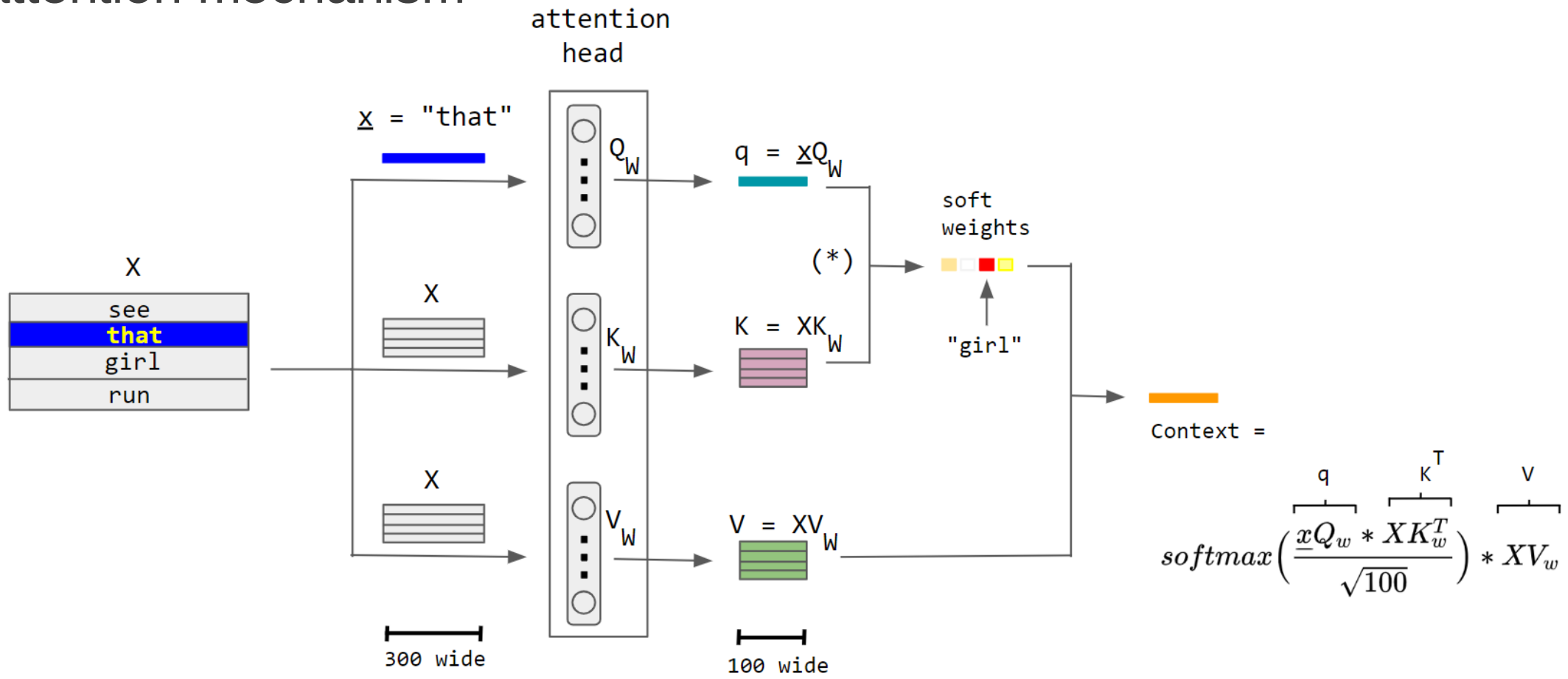
Embeddings



Shameless “borrowed” from the excellent 3Blue1Brown video series: https://www.youtube.com/watch?v=eMlx5fFNoYc

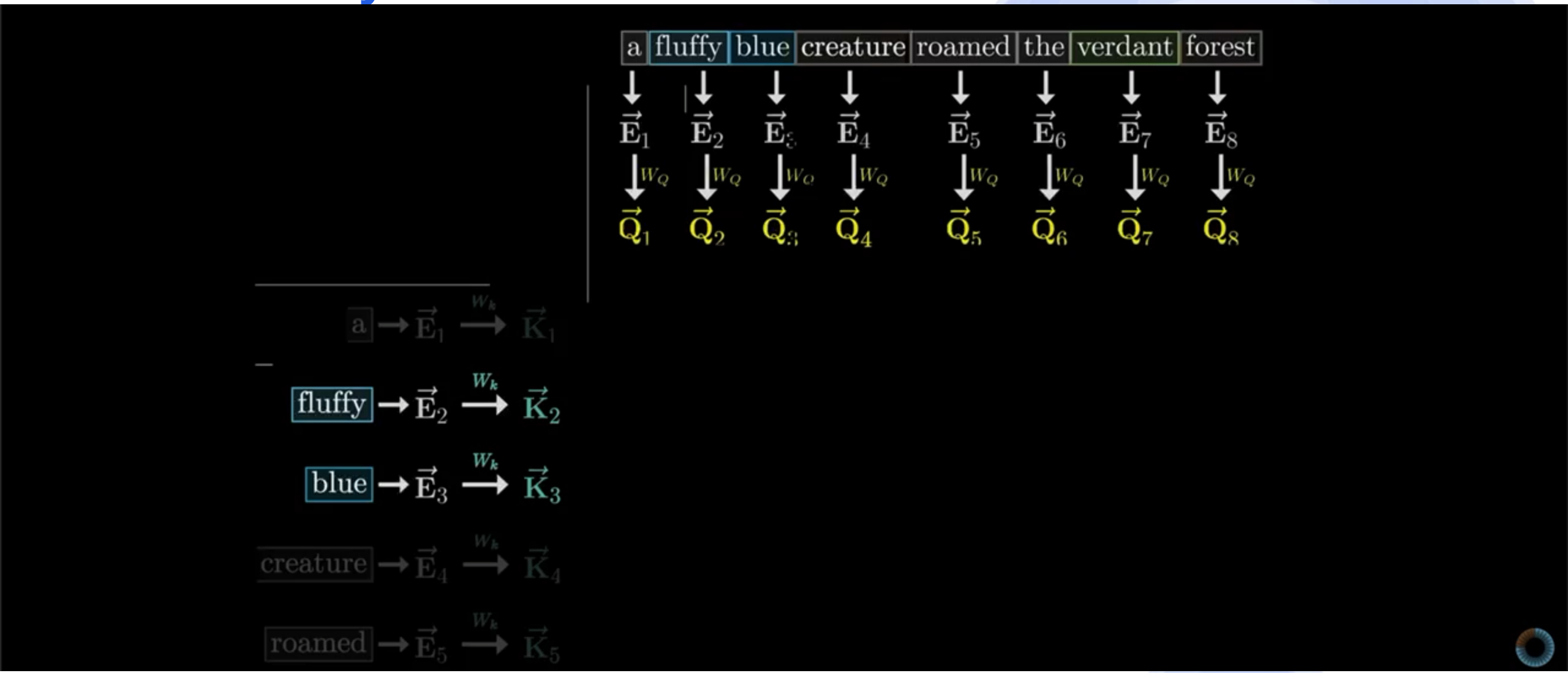
How do they work?

The attention mechanism



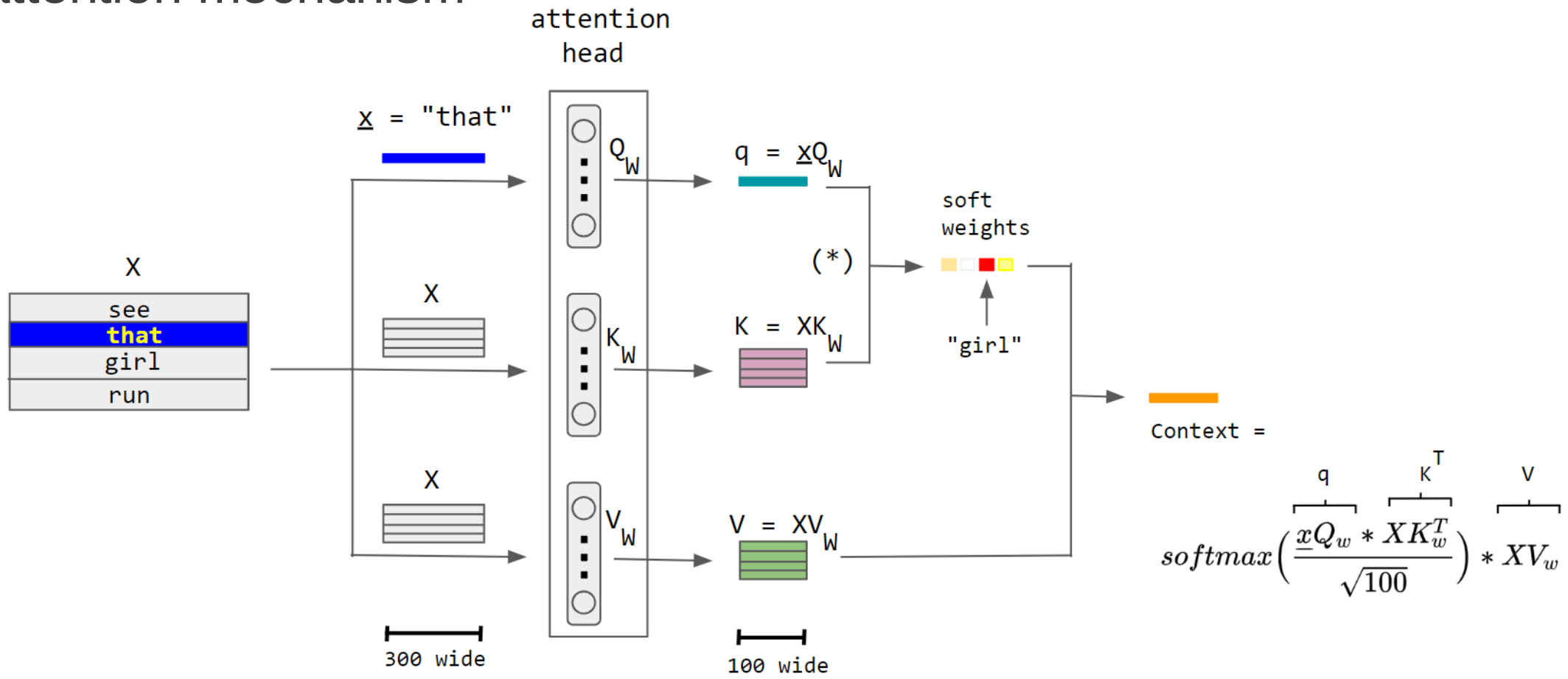
https://commons.wikimedia.org/wiki/File:Attention-qkv.png

How do they work?



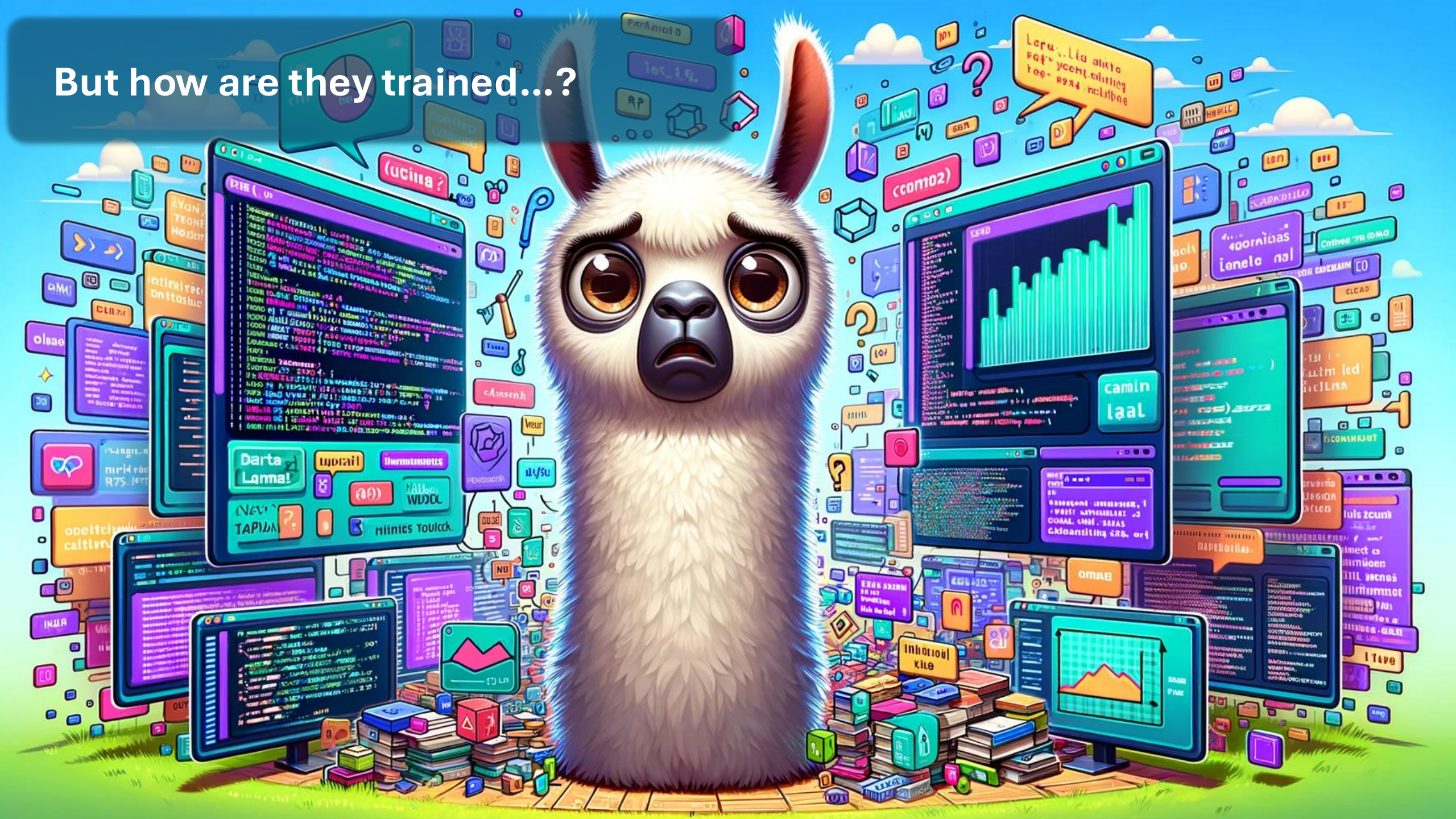
How do they work?

The attention mechanism

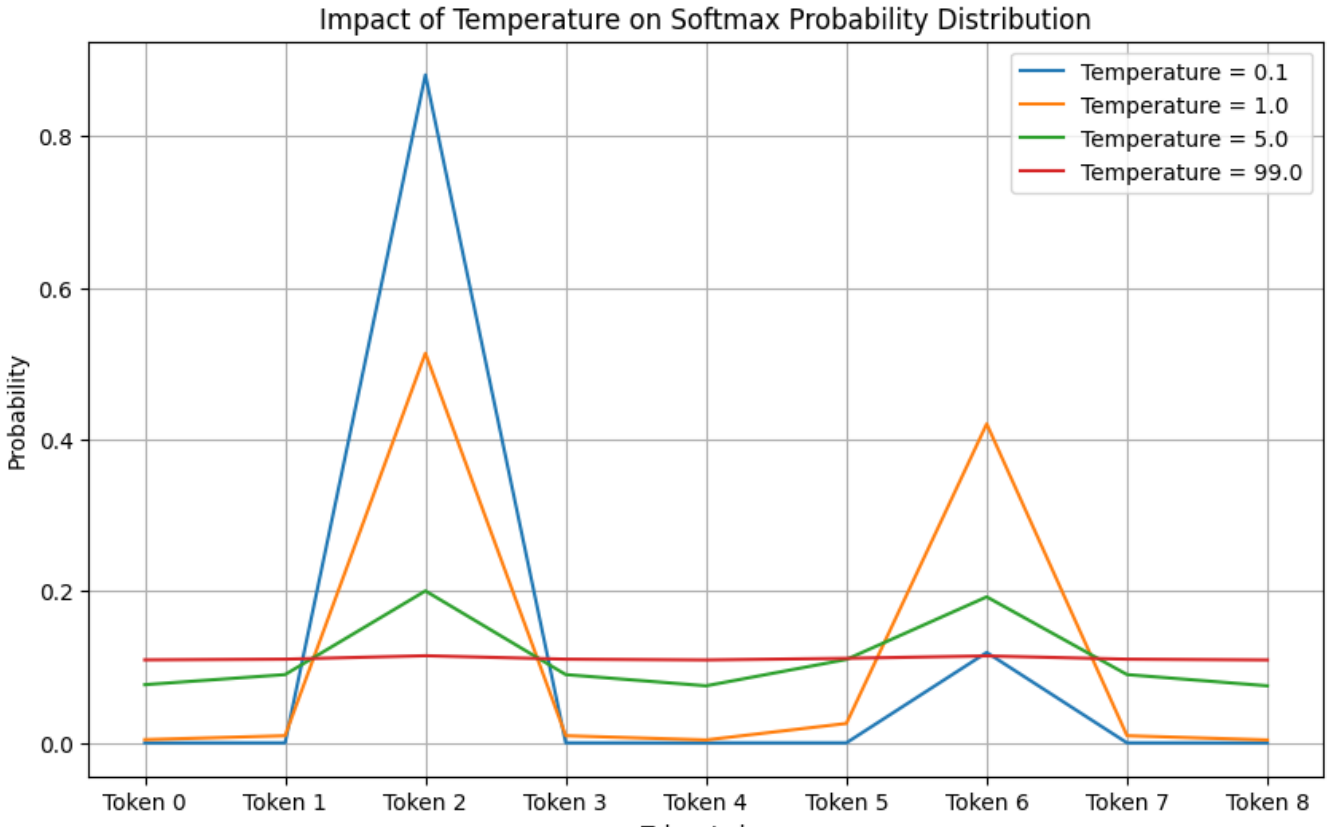


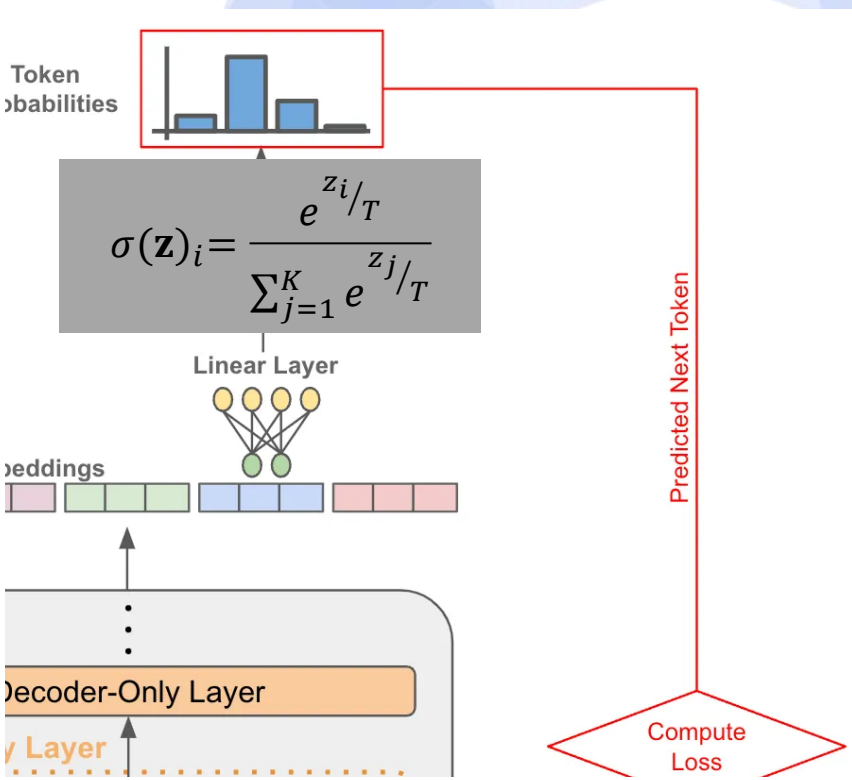
https://commons.wikimedia.org/wiki/File:Attention-qkv.png

But how are they trained...?



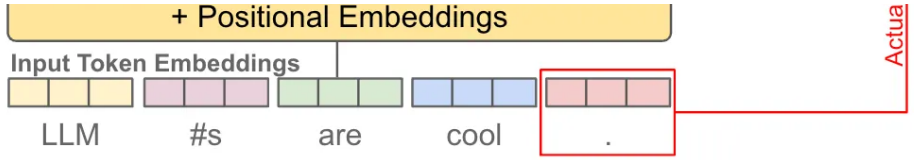
How are they trained?





 # Example logits (e.g., attention scores)

logits = np.array([0.2, 1.0, 5.0, 1.0, 0.1, 2.0, 4.8, 1.0, 0.1])

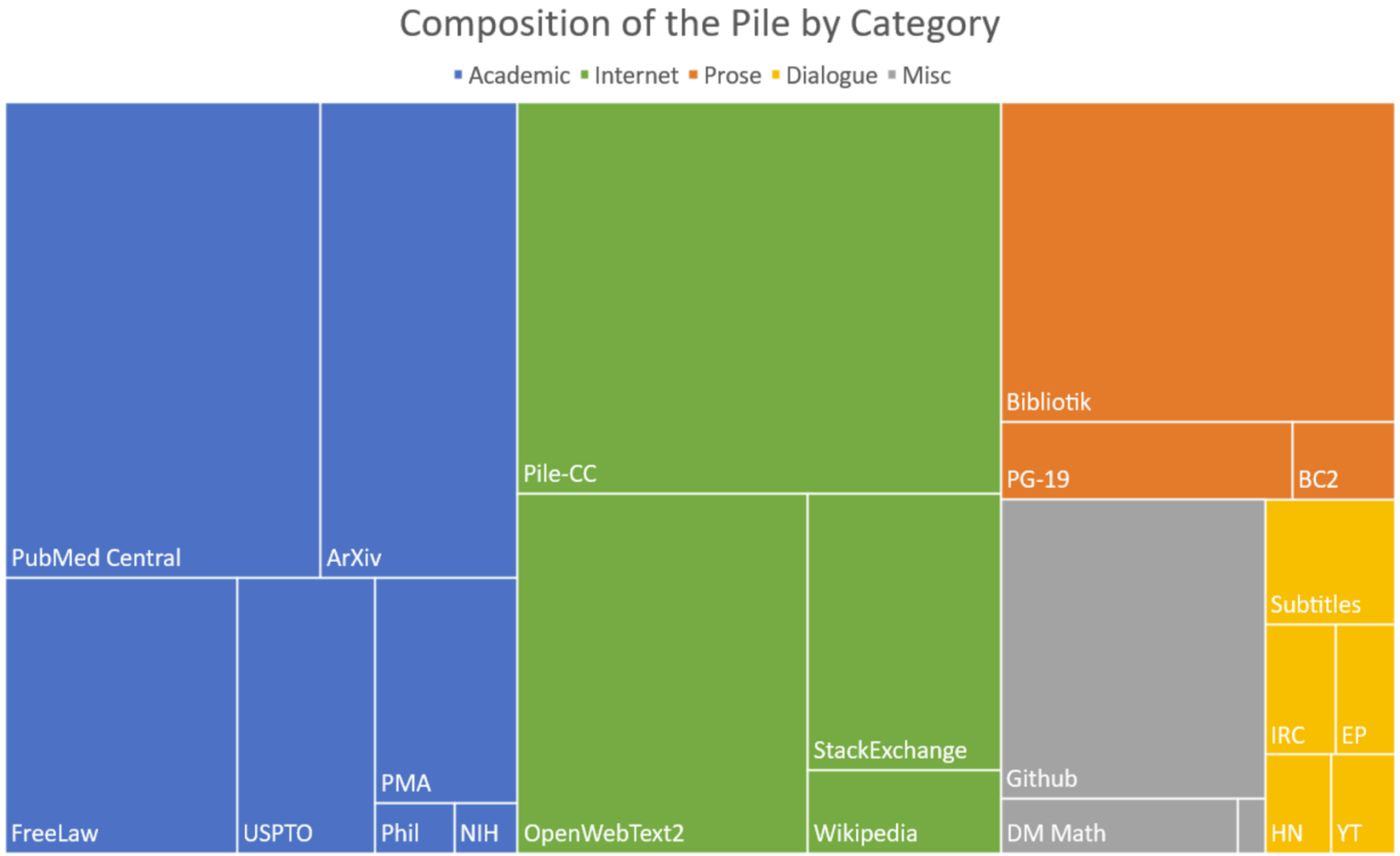


ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES

https://medium.com/@akash.kesrwani99/understanding-next-token-prediction-concept-to-code-1st-part-7054dabda347

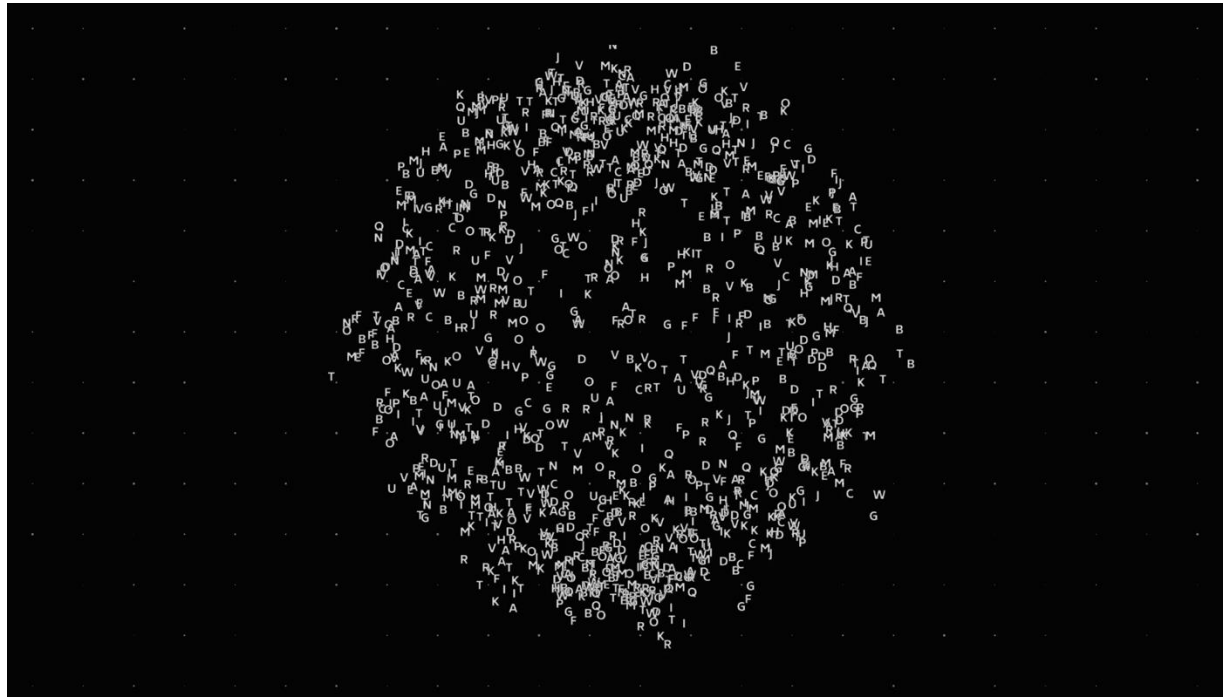
How are they trained?



ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES

LLMs as a compressed version of the internet



https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web

Some like to think of an LLM as a compressed version of the internet

One way of compressing images is to take every other pixel

When you decompress, you interpolate between the pixels

The inclination is to think that some kind of interpolation is happening in LLMs…

…but it's more complicated than this.

Scaling Laws

Chinchilla scaling laws

$$Loss \propto N^{-\alpha}$$

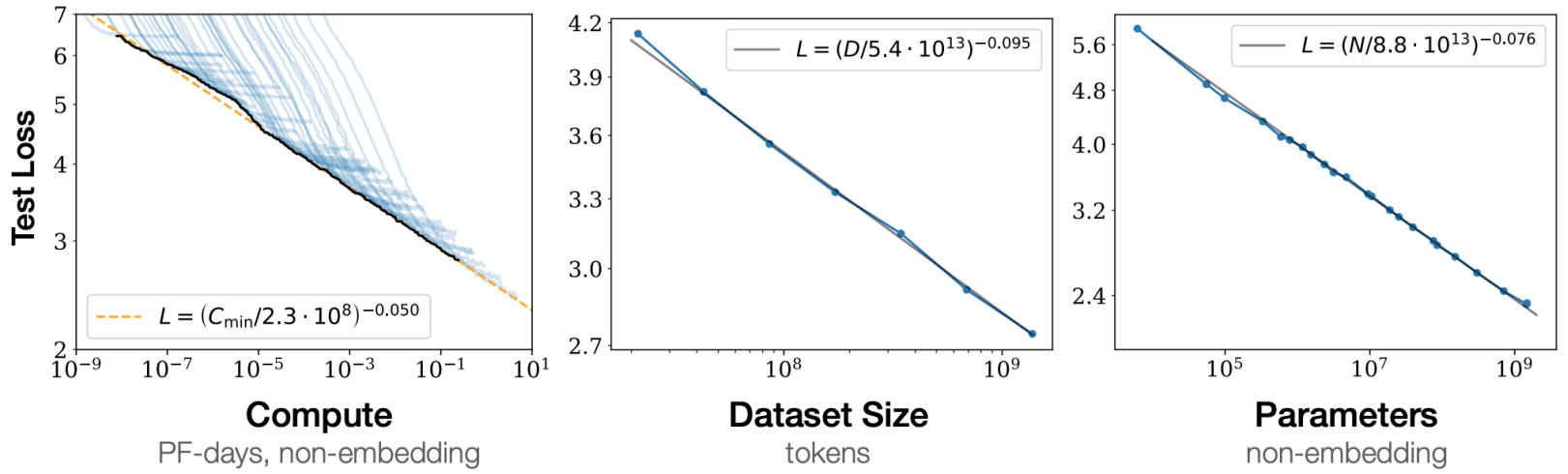
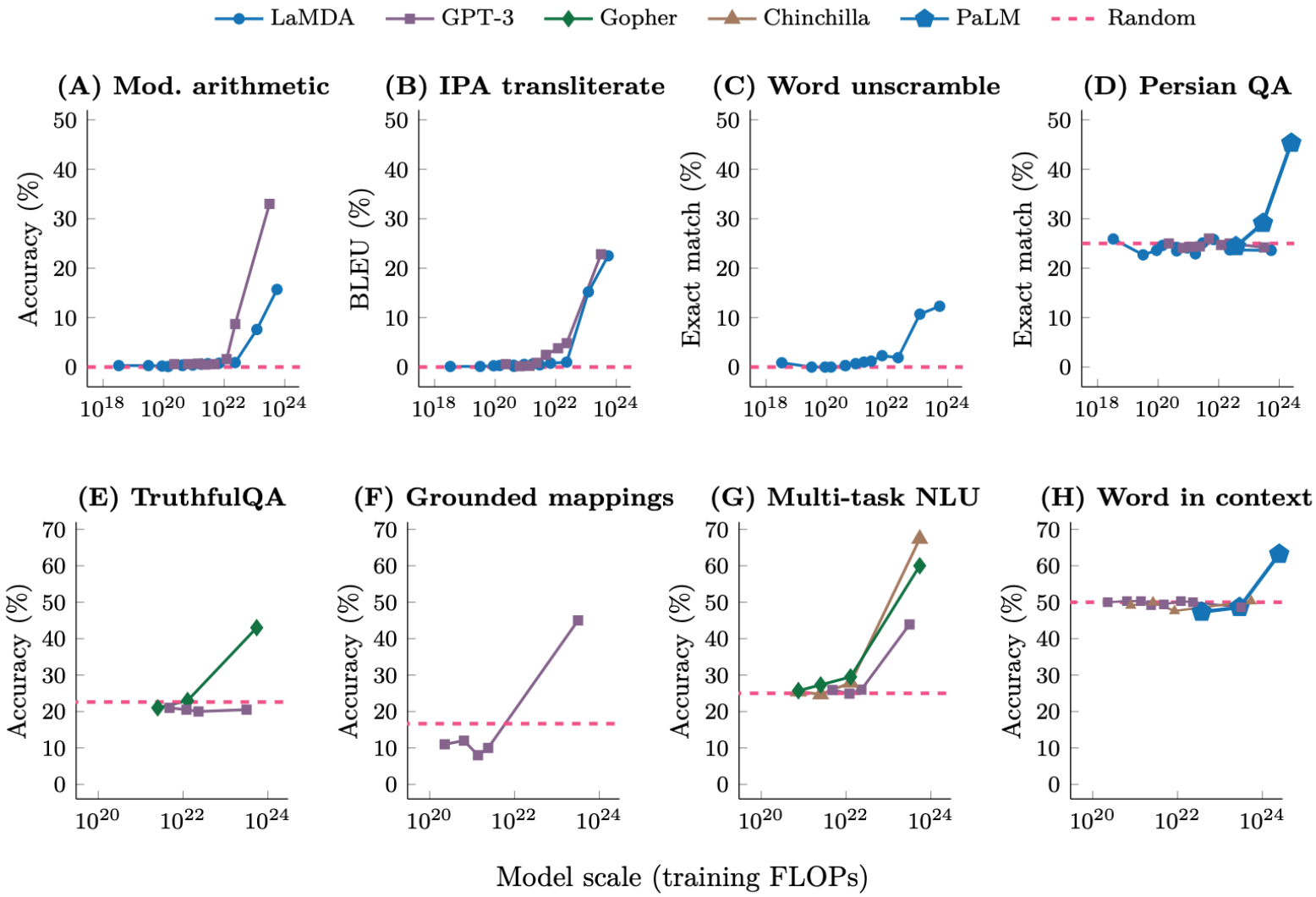


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Emergence

Maybe…



ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES

Resources

Jargon-free explanation of LLMs

Generative AI exists because of the Transformer

The dangers of Stochastic Parrots

The Pile

Survey of LLMs

Attention is all you need

Intro to Large Language Models

10 min BREAK

ACCELERATE PROGRAMME FOR SCIENTIFIC DISCOVERY

SCHMIDT FUTURES