elastic

Workshop: GenAI and LLM, searching data in natural language

Enrico Zimuel, *Tech Lead & Principal Software Engineer*

Agenda

- Brief intro to Artificial Intelligence
- Machine Learning and Neural Networks
- Large Language Model (LLM)
- Transformers architecture
- Top-k and temperature
- Retrieval Augmented Generation (RAG)
- Embedding and Vector Search
- Lab: LangChain, Llama 3.2, Elasticsearch

Image generated using dall-e-3

Artificial intelligence

- Many definitions have been proposed:
  - The ability of a digital computer to perform tasks commonly associated with intelligent beings
  - The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages
  - An umbrella term for a range of algorithm-based technologies that solve complex tasks by carrying out functions that previously required human thinking

AI examples

- February 10, 1996, Deep Blue beat Garry Kasparov in the first game of a six-game match—the first time a computer had ever beat a human in a formal chess game
- March 15, 2016, AlphaGo beat Lee Sedol 4-1 in a formal Go game
- June 2020, Google's DeepMind A.I. beats doctors in breast cancer screening trial
- June 2022, the Google LaMDA (Language Model for Dialog Applications) chatbot apparently passed the **Turing Test**. Many experts in the field, pointing out that a language model appearing to mimic human conversation does not indicate that any intelligence is present

Machine learning

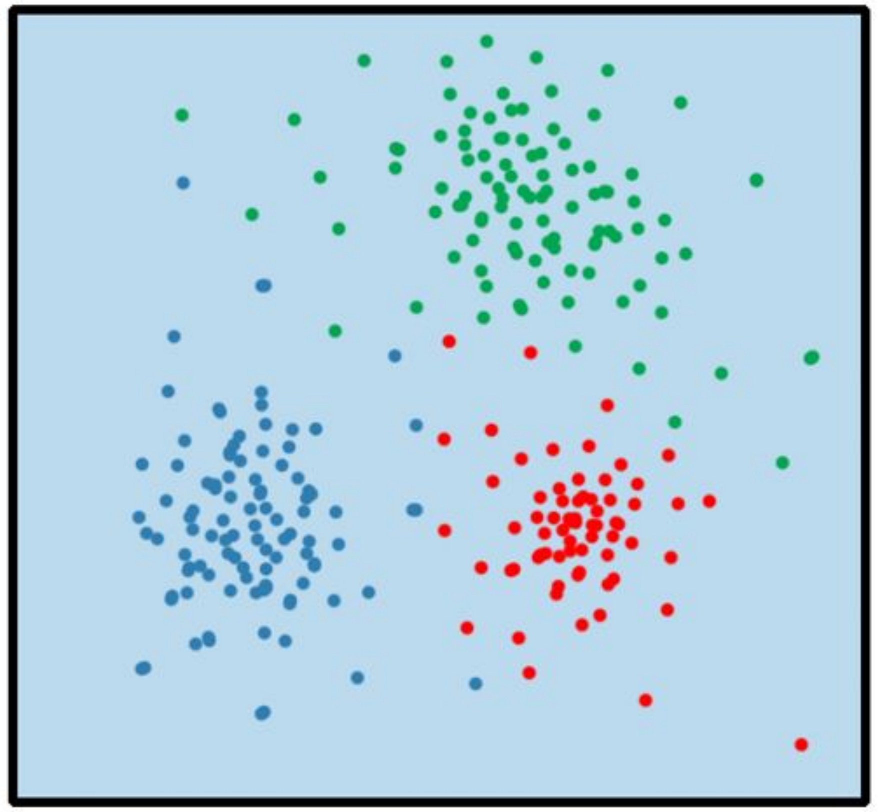
- **Machine Learning (ML)** is the use and development of computer systems that are able to **learn** and **adapt** without following explicit instructions, by using algorithms and **statistical models** to analyse and draw inferences from patterns in data
- We can have 3 types of ML:
  - **Supervised learning**: use of labeled datasets to train algorithms that to classify data or predict outcomes (eg. image and speech recognition, recommendation systems, fraud detection)
  - **Unsupervised learning**: algorithms learn patterns exclusively from unlabeled data (eg. clustering, anomaly detection, preparing data for supervised learning)
  - **Reinforcement learning:** training method based on rewarding desired behaviors and punishing undesired ones

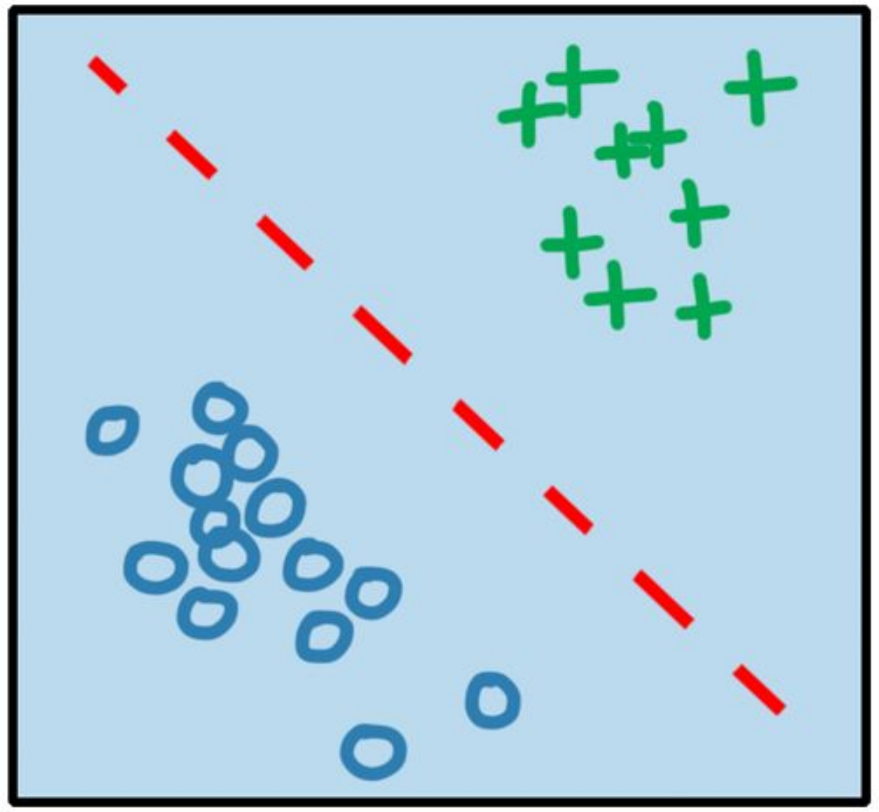
machine learning

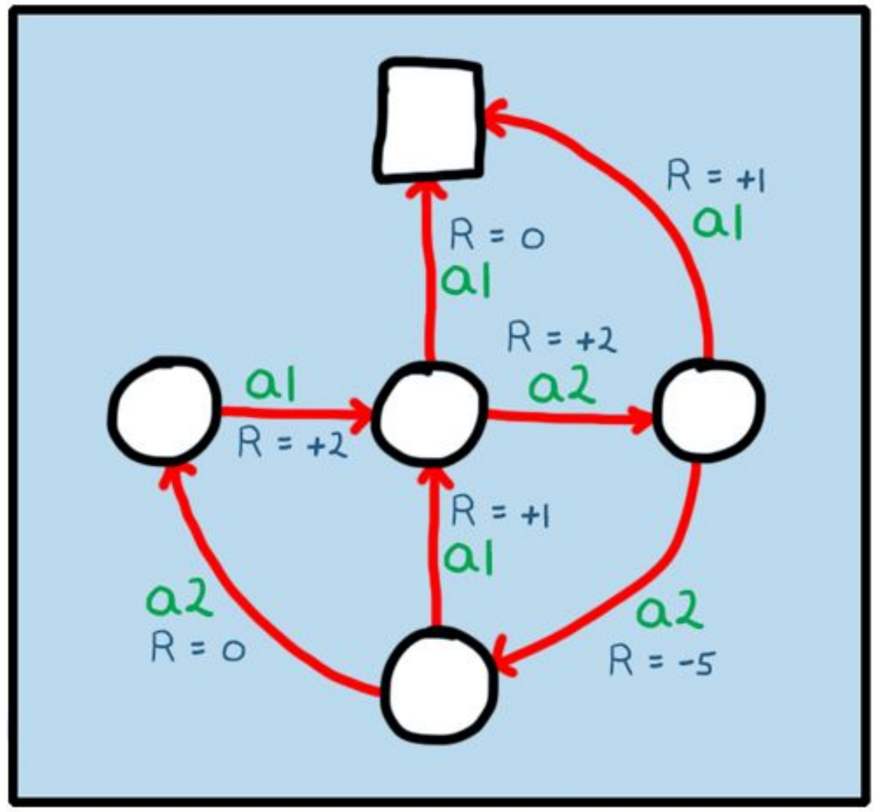
unsupervised learning

supervised learning

reinforcement learning



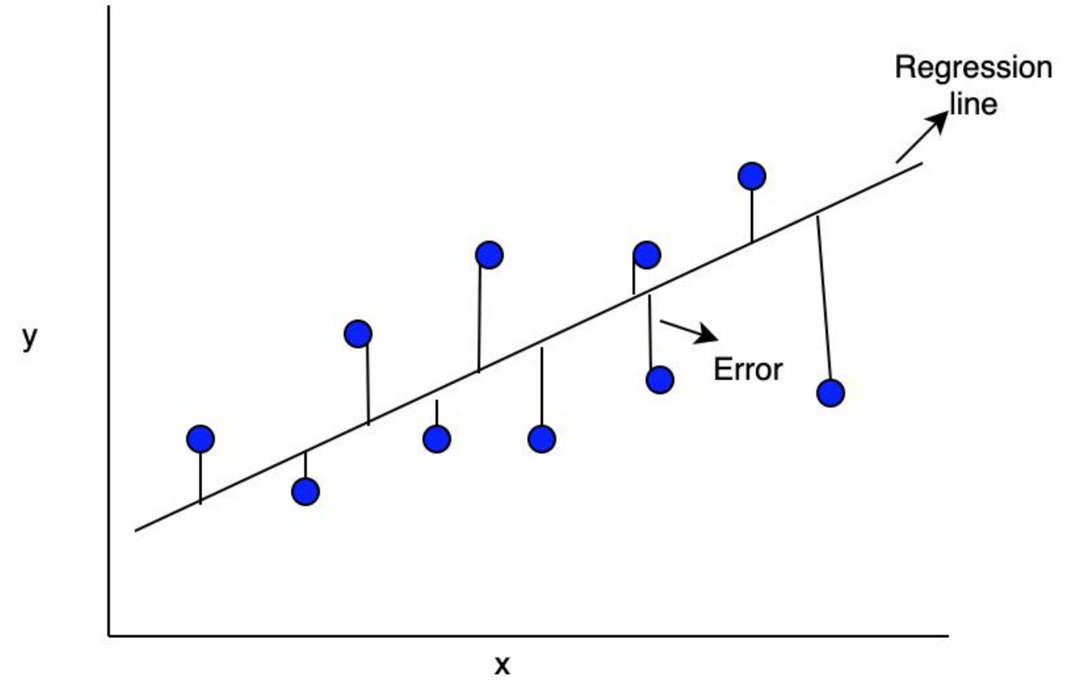


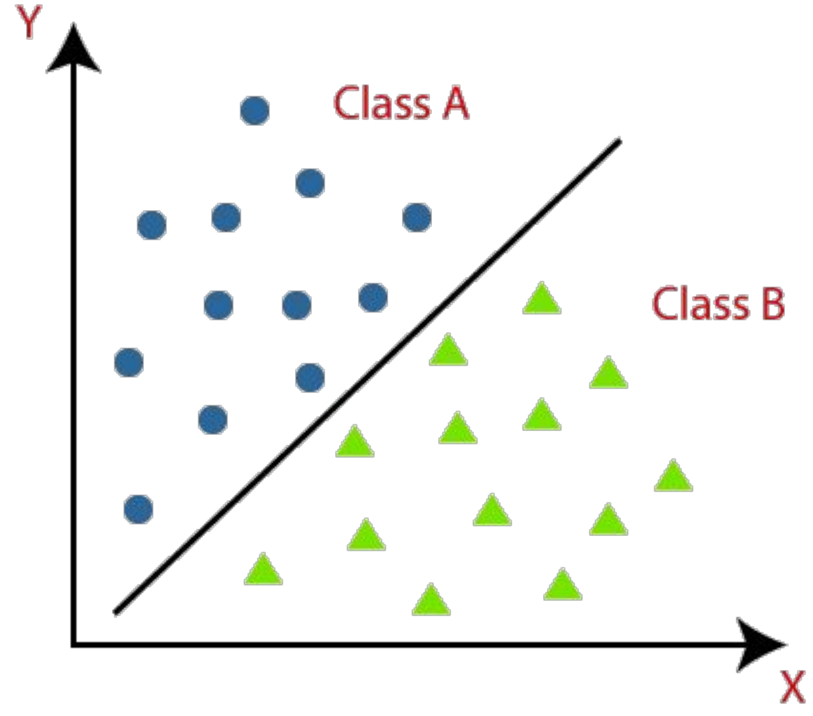


Classification and regression

- In the supervised learning we have two types of classifications:
  - **Regression**, when the target variable is continuous. Example: predict the salary of a person based on education degree, work experience and geo location

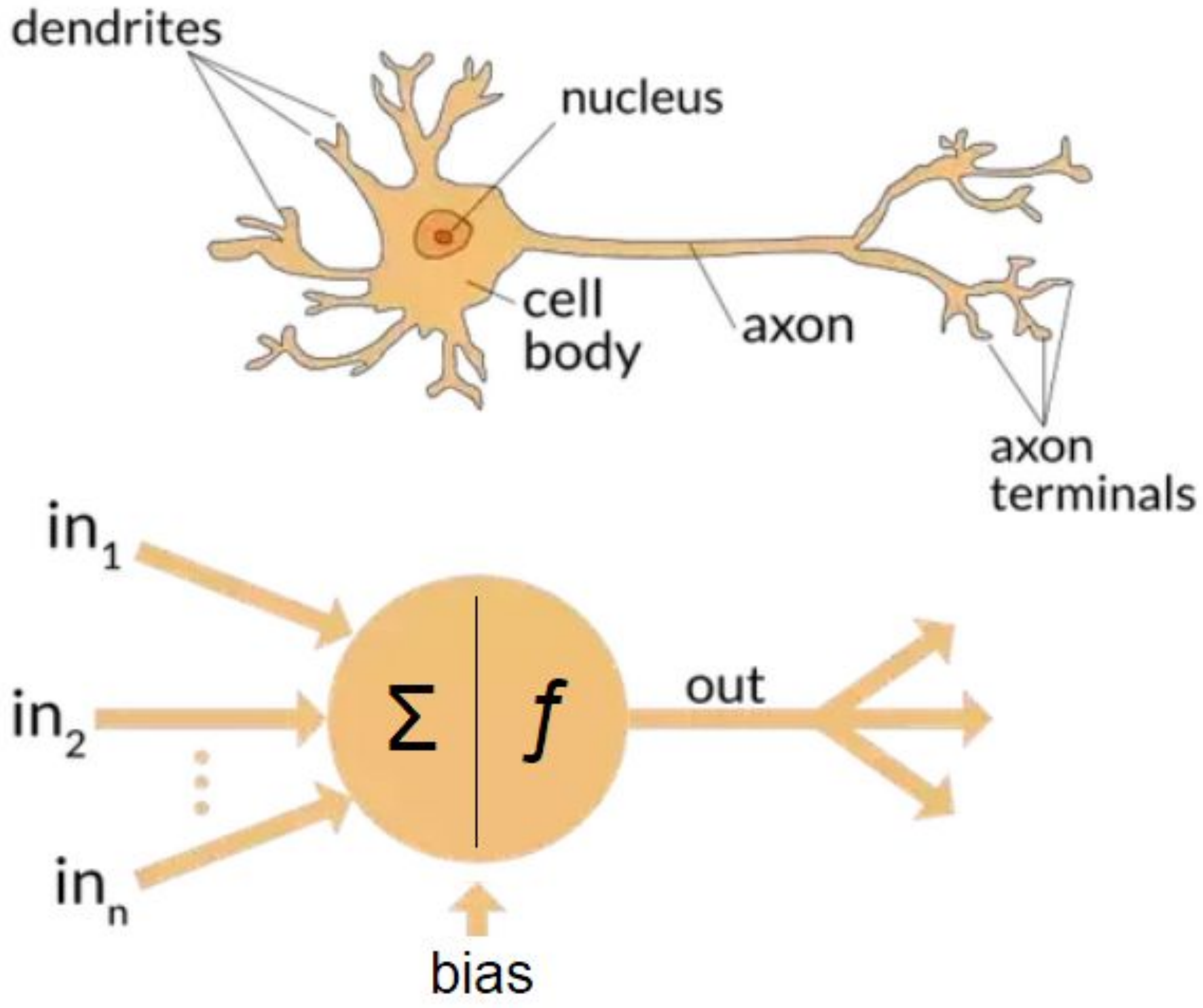
  - **Classification**, when the target variable is discrete. Example: sentiment analysis of a piece of text (e.g. comments of a product)



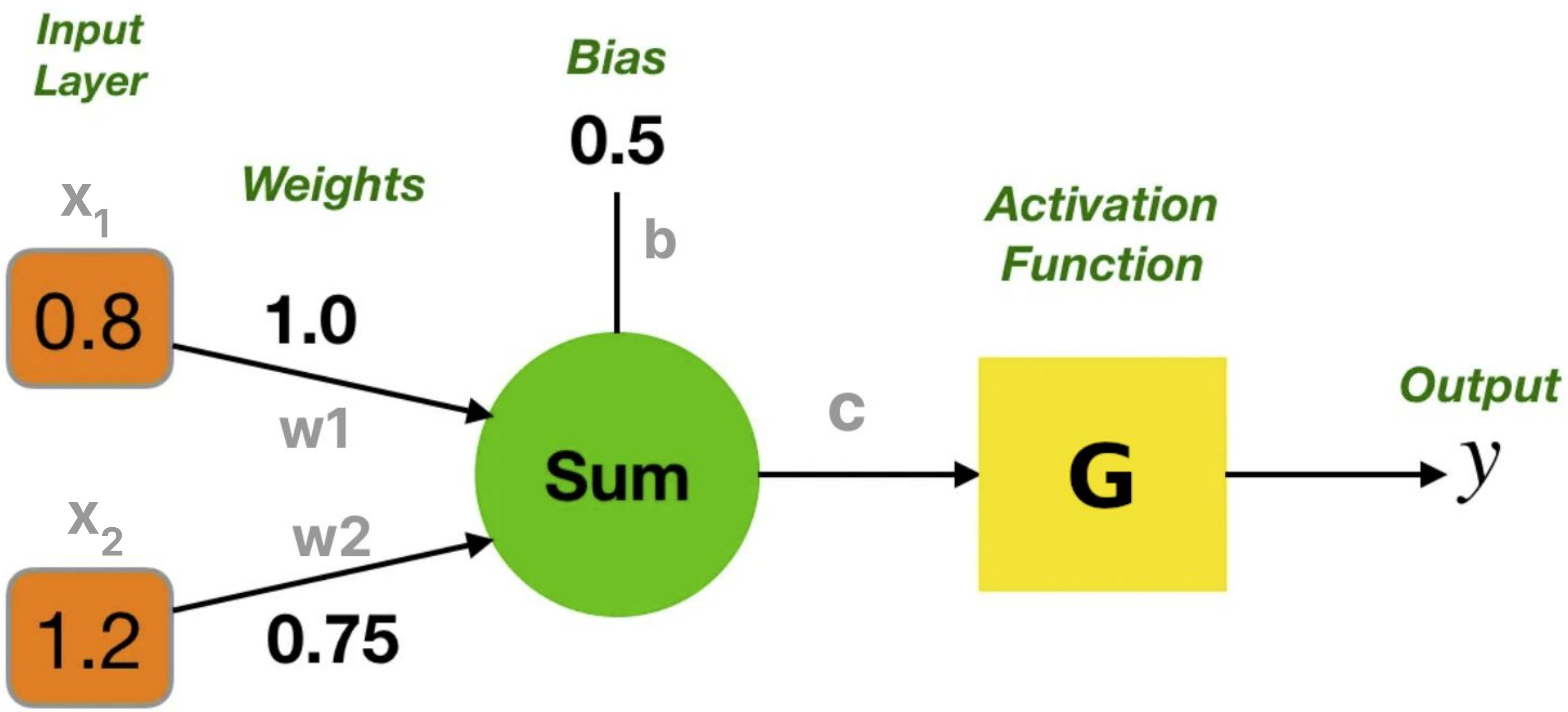


Neural Network

- A **neural network** is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain
- Collection of **nodes** (artificial neurons) with inputs and outputs. A neuron computes some non-linear function of the sum of its inputs
- The nodes are collected in **layers**
- If the number of layers is greater than 3 we say **deep learning network**

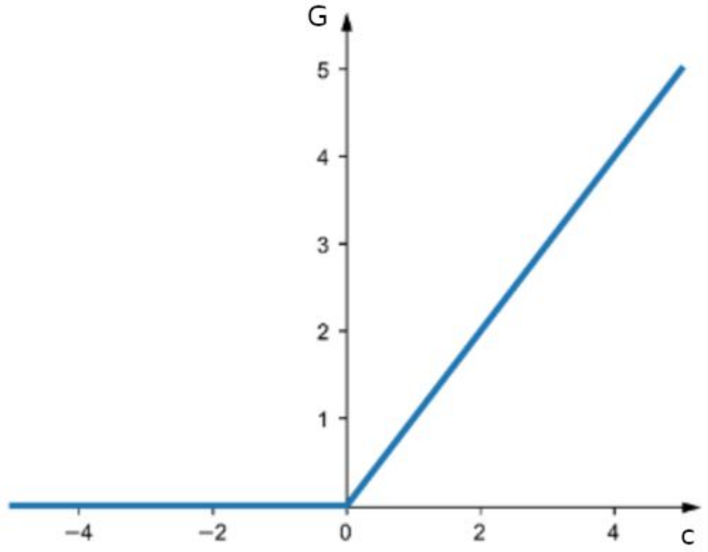


Single layer neuron



Example: ReLU activation

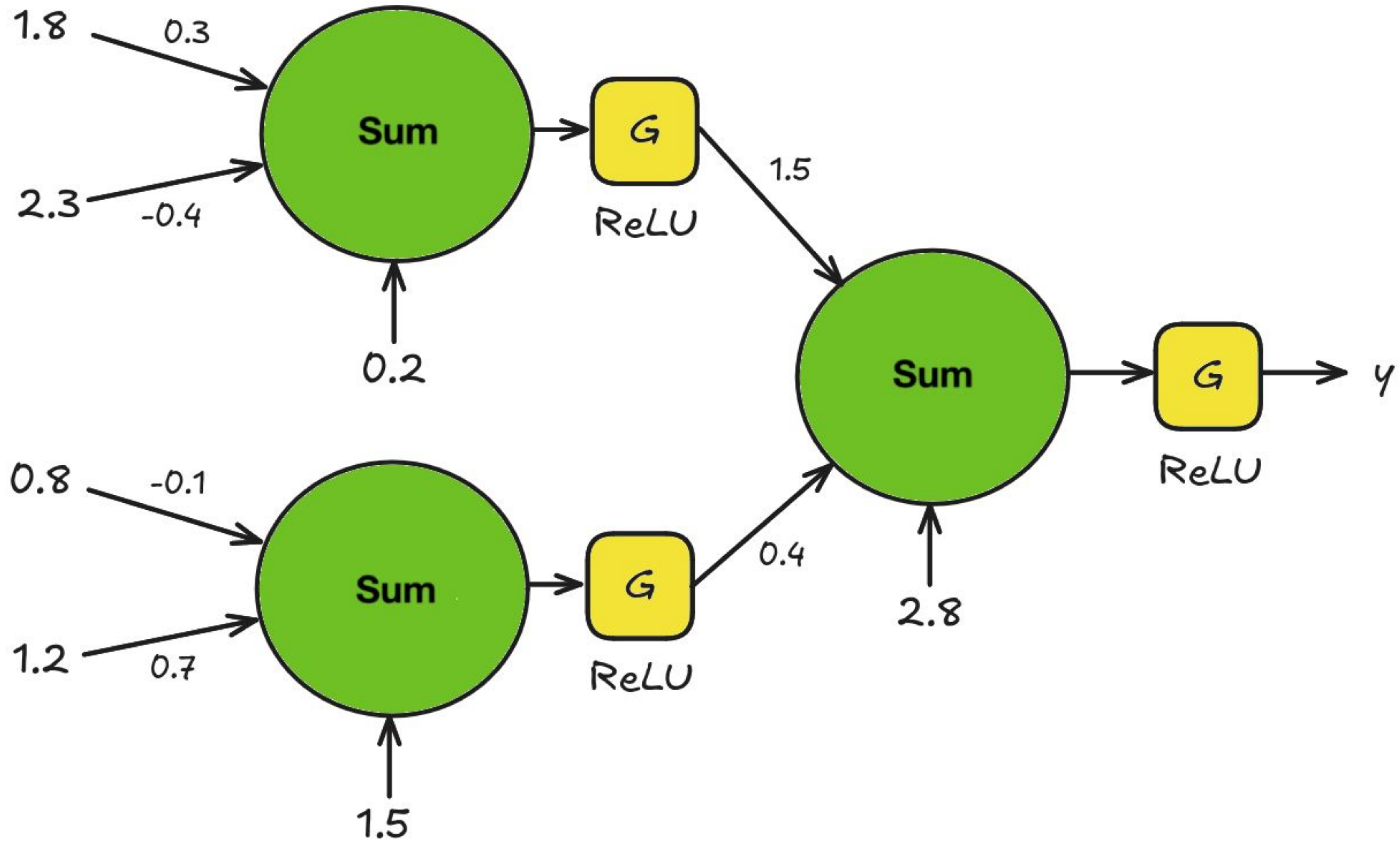
`G(c)= c≥0 ? c : 0`

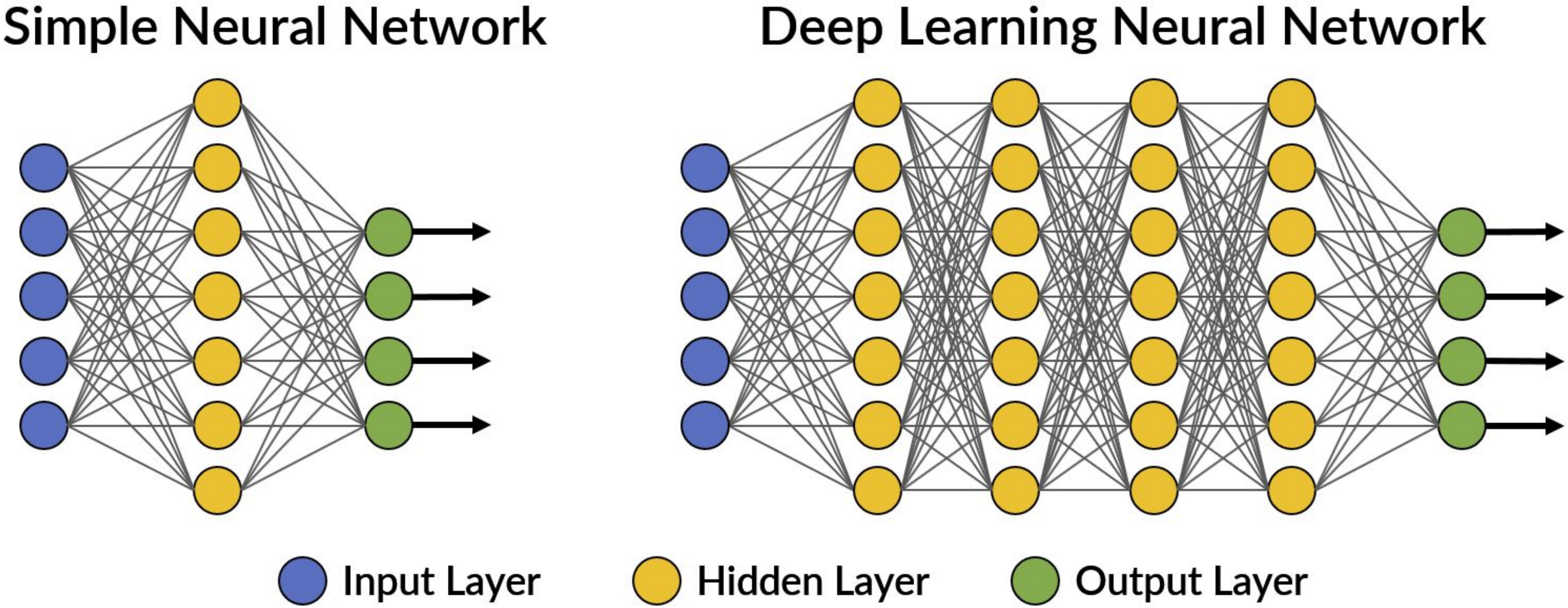


$$
\begin{aligned}
c &= x_1 * w_1 + x_2 * w_2 + b \\
&= 0.8 * 1.0 + 1.2 * 0.75 + 0.5 \\
&= 2.2
\end{aligned}
$$

$$y = G(c) = G(2.2) = 2.2$$

Exercise: what is the value of y?





Simple Neural Network

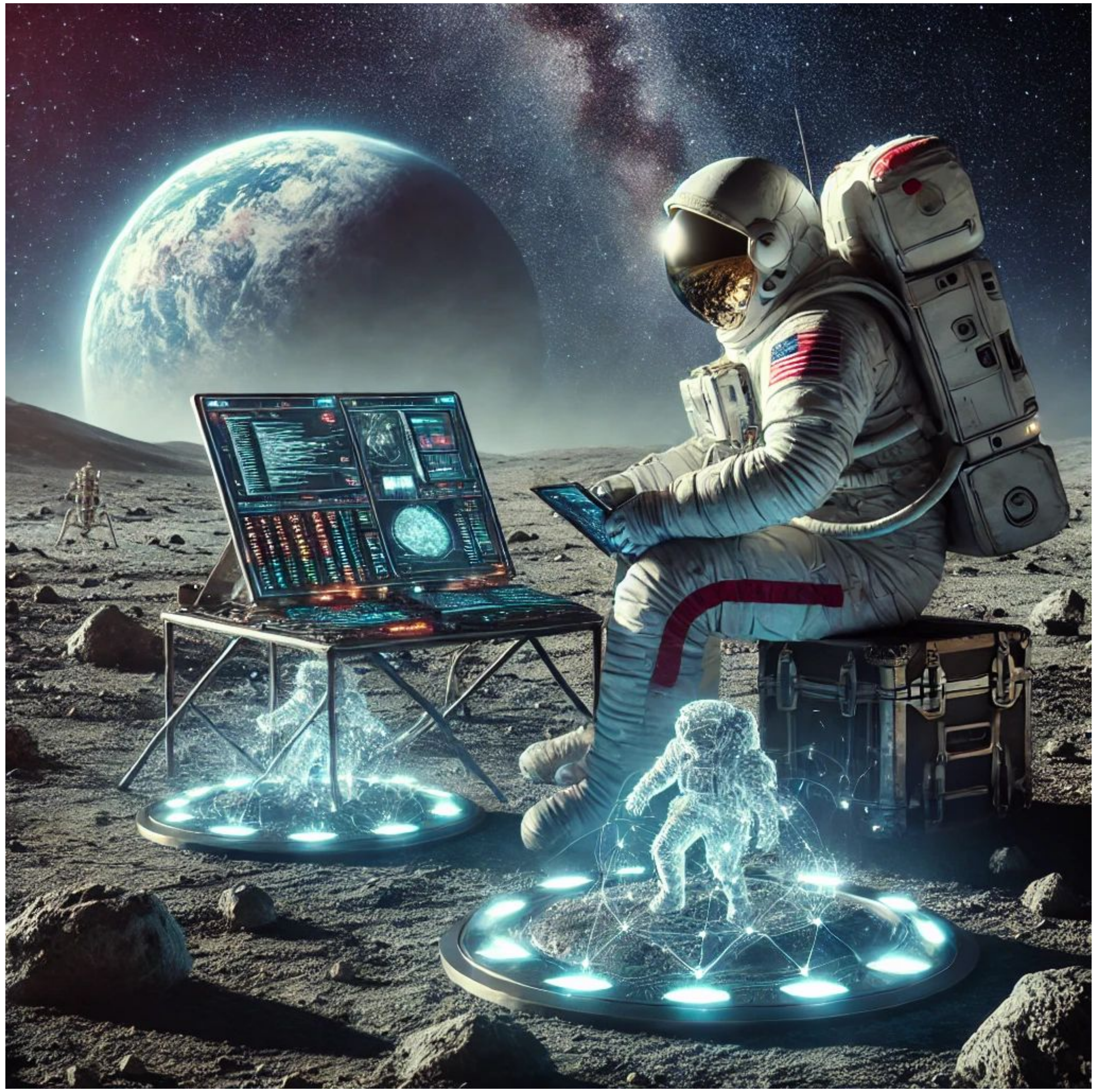
Deep Learning Neural Network

● Input Layer ● Hidden Layer ● Output Layer

Generative AI

- **Generative Artificial Intelligence** (GenAI) is artificial intelligence capable of generating text, images, or other media, using generative models
- GenAI models **learn the patterns and structure** of their input training data and then generate new data that has **similar characteristics**
- It's used in many industries, including art, writing, script writing, software development, product design, healthcare, finance, gaming, marketing, and fashion
- The GenAI market size has been valued at $36 billion in 2024, and is projected to reach $191.8 billion by 2032

Examples

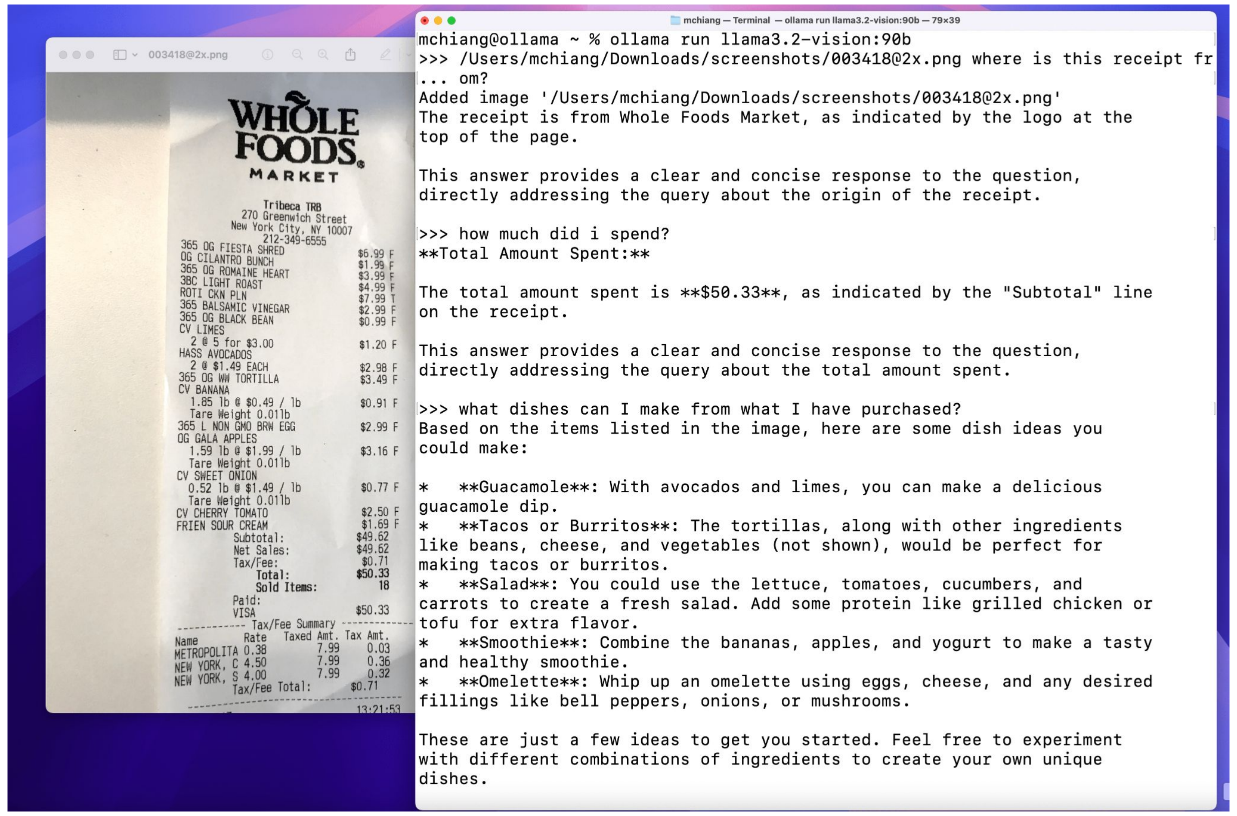


Prompt for GPT-4o
Create a picture of an astronaut programming a computer on the Moon

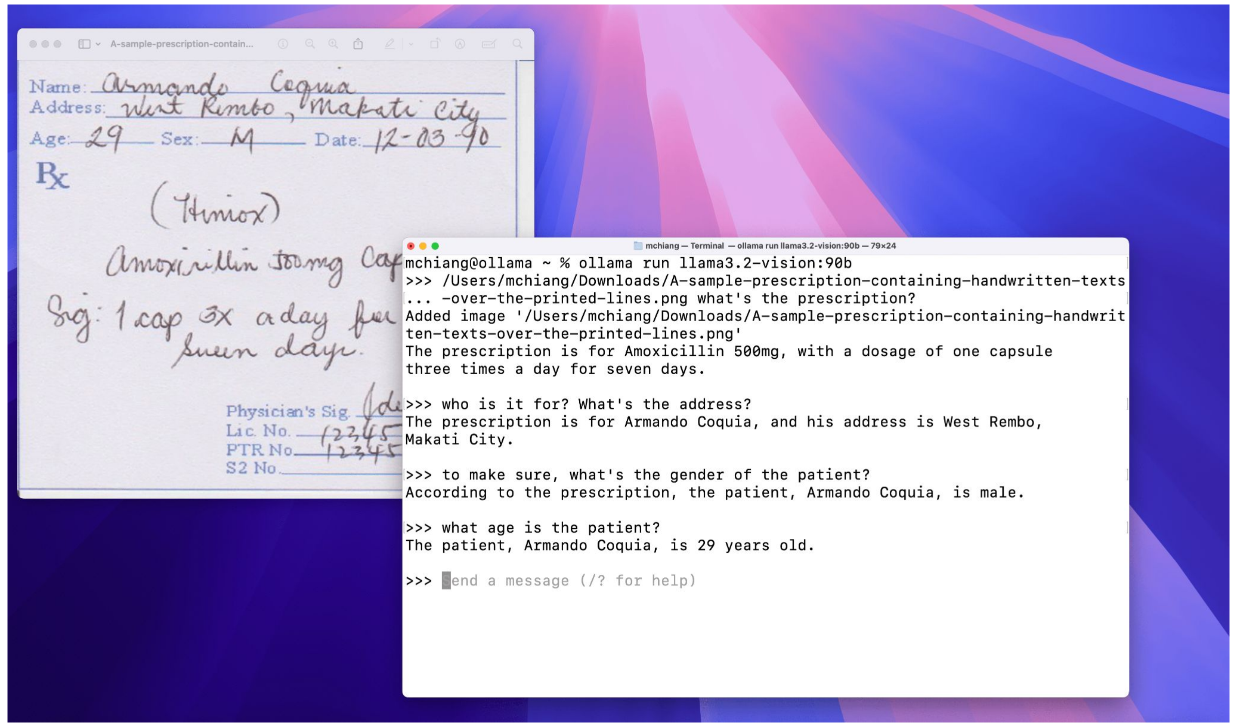
Audio file generated using tts-1

Llama3.2-vision

- A multimodal small (11B) and medium-sized (90B) vision LLM by Meta
- Support image reasoning, such as document-level understanding including charts and graphs, captioning of images, and more
- Released Sept. 25, 2024
- A quick DEMO using ollama

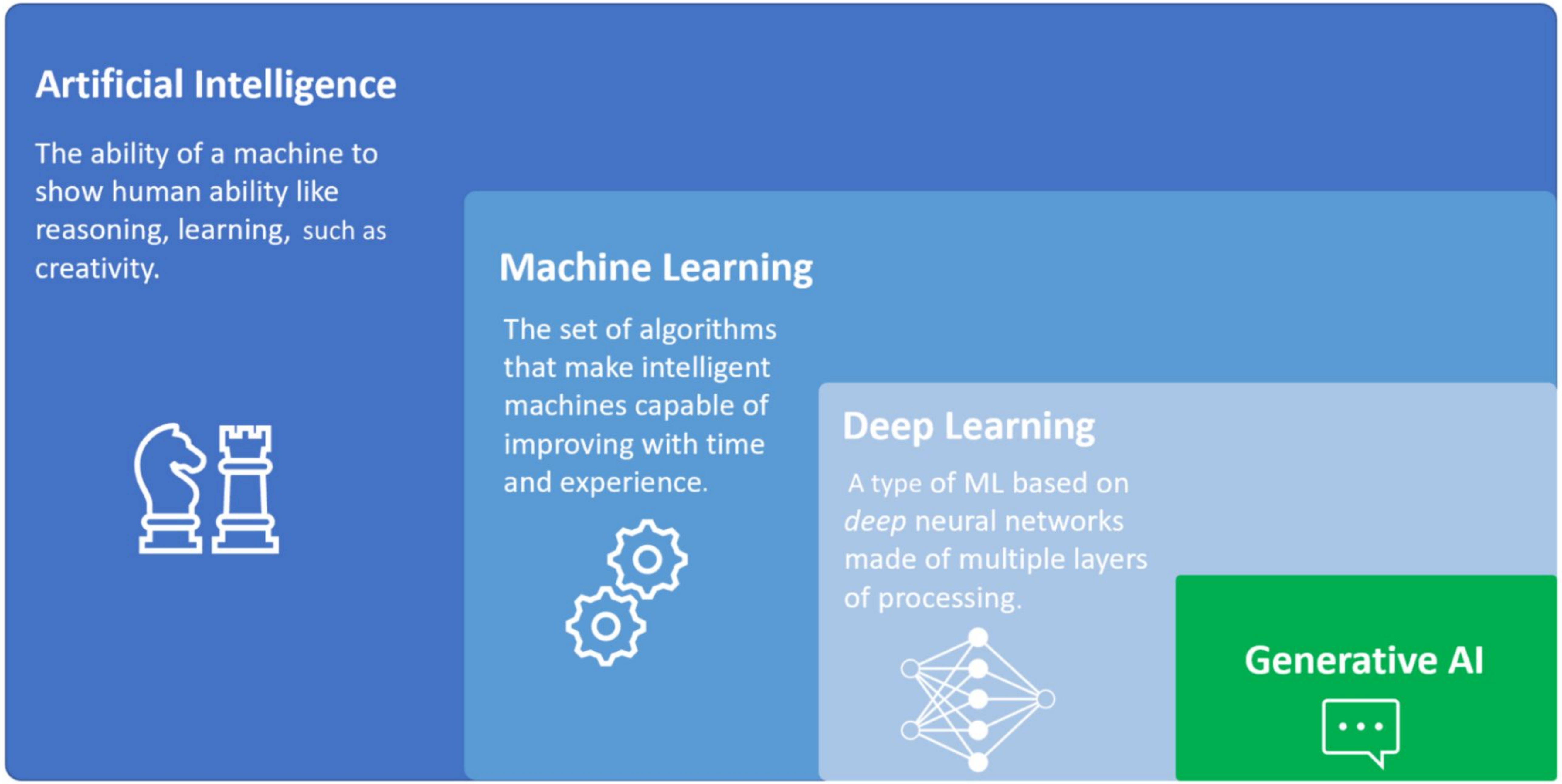


```
mchiang@ollama ~ % ollama run llama3.2-vision:90b
>>> /Users/mchiang/Downloads/screenshots/003418@2x.png where is this receipt fr
... om?
Added image '/Users/mchiang/Downloads/screenshots/003418@2x.png'
The receipt is from Whole Foods Market, as indicated by the logo at the
top of the page.

This answer provides a clear and concise response to the question,
directly addressing the query about the origin of the receipt.

>>> how much did i spend?
**Total Amount Spent:**

The total amount spent is **$50.33**, as indicated by the "Subtotal" line
on the receipt.

This answer provides a clear and concise response to the question,
directly addressing the query about the total amount spent.

>>> what dishes can I make from what I have purchased?
Based on the items listed in the image, here are some dish ideas you
could make:

*   **Guacamole**: With avocados and limes, you can make a delicious
guacamole dip.
*   **Tacos or Burritos**: The tortillas, along with other ingredients
like beans, cheese, and vegetables (not shown), would be perfect for
making tacos or burritos.
*   **Salad**: You could use the lettuce, tomatoes, cucumbers, and
carrots to create a fresh salad. Add some protein like grilled chicken or
tofu for extra flavor.
*   **Smoothie**: Combine the bananas, apples, and yogurt to make a tasty
and healthy smoothie.
*   **Omelette**: Whip up an omelette using eggs, cheese, and any desired
fillings like bell peppers, onions, or mushrooms.

These are just a few ideas to get you started. Feel free to experiment
with different combinations of ingredients to create your own unique
dishes.
```



```
mchiang@ollama ~ % ollama run llama3.2-vision:90b
>>> /Users/mchiang/Downloads/A-sample-prescription-containing-handwritten-texts
... -over-the-printed-lines.png what's the prescription?
Added image '/Users/mchiang/Downloads/A-sample-prescription-containing-handwrit
ten-texts-over-the-printed-lines.png'
The prescription is for Amoxicillin 500mg, with a dosage of one capsule
three times a day for seven days.

>>> who is it for? What's the address?
The prescription is for Armando Coquia, and his address is West Rembo,
Makati City.

>>> to make sure, what's the gender of the patient?
According to the prescription, the patient, Armando Coquia, is male.

>>> what age is the patient?
The patient, Armando Coquia, is 29 years old.

>>> Send a message (/? for help)
```

AI ⊃ ML ⊃ DL ⊃ GenAI



Artificial Intelligence

The ability of a machine to show human ability like reasoning, learning, such as creativity.

Machine Learning

The set of algorithms that make intelligent machines capable of improving with time and experience.

Deep Learning

A type of ML based on *deep* neural networks made of multiple layers of processing.

Generative AI

State of AI 2024

- stateof.ai is a very good report about AI in Research, Industry, Politics, Safety and Predictions

STATE OF AI REPORT.

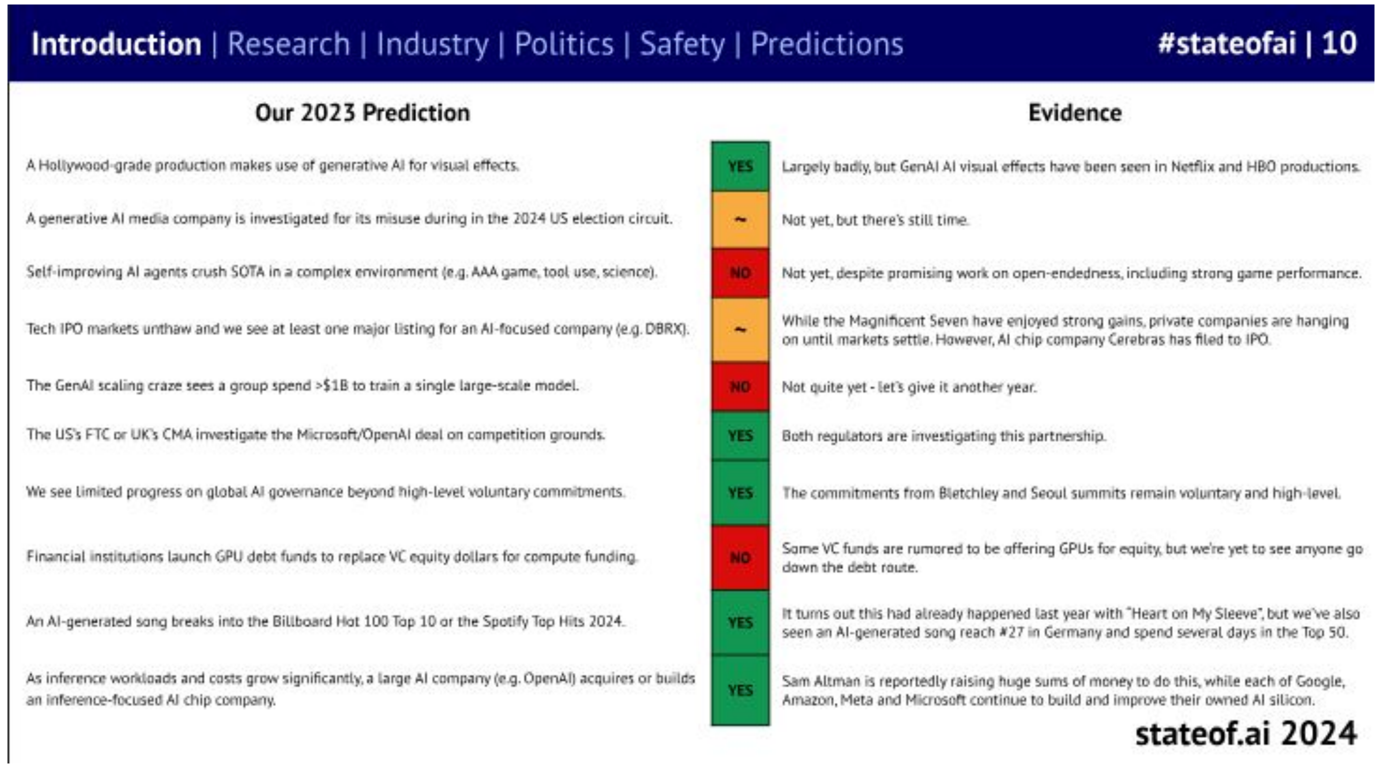
October 10, 2024

Nathan Benaich

AIR STREET CAPITAL.

stateof.ai

airstreet.com



Introduction | Research | Industry | Politics | Safety | Predictions

#stateofai | 10

Our 2023 Prediction		Evidence
A Hollywood-grade production makes use of generative AI for visual effects.	YES	Largely badly, but GenAI AI visual effects have been seen in Netflix and HBO productions.
A generative AI media company is investigated for its misuse during in the 2024 US election circuit.	~	Not yet, but there's still time.
Self-improving AI agents crush SOTA in a complex environment (e.g. AAA game, tool use, science).	NO	Not yet, despite promising work on open-endedness, including strong game performance.
Tech IPO markets unthaw and we see at least one major listing for an AI-focused company (e.g. DBRX).	~	While the Magnificent Seven have enjoyed strong gains, private companies are hanging on until markets settle. However, AI chip company Cerebras has filed to IPO.
The GenAI scaling craze sees a group spend >$1B to train a single large-scale model.	NO	Not quite yet - let's give it another year.
The US's FTC or UK's CMA investigate the Microsoft/OpenAI deal on competition grounds.	YES	Both regulators are investigating this partnership.
We see limited progress on global AI governance beyond high-level voluntary commitments.	YES	The commitments from Bletchley and Seoul summits remain voluntary and high-level.
Financial institutions launch GPU debt funds to replace VC equity dollars for compute funding.	NO	Some VC funds are rumored to be offering GPUs for equity, but we're yet to see anyone go down the debt route.
An AI-generated song breaks into the Billboard Hot 100 Top 10 or the Spotify Top Hits 2024.	YES	It turns out this had already happened last year with "Heart on My Sleeve", but we've also seen an AI-generated song reach #27 in Germany and spend several days in the Top 50.
As inference workloads and costs grow significantly, a large AI company (e.g. OpenAI) acquires or builds an inference-focused AI chip company.	YES	Sam Altman is reportedly raising huge sums of money to do this, while each of Google, Amazon, Meta and Microsoft continue to build and improve their owned AI silicon.

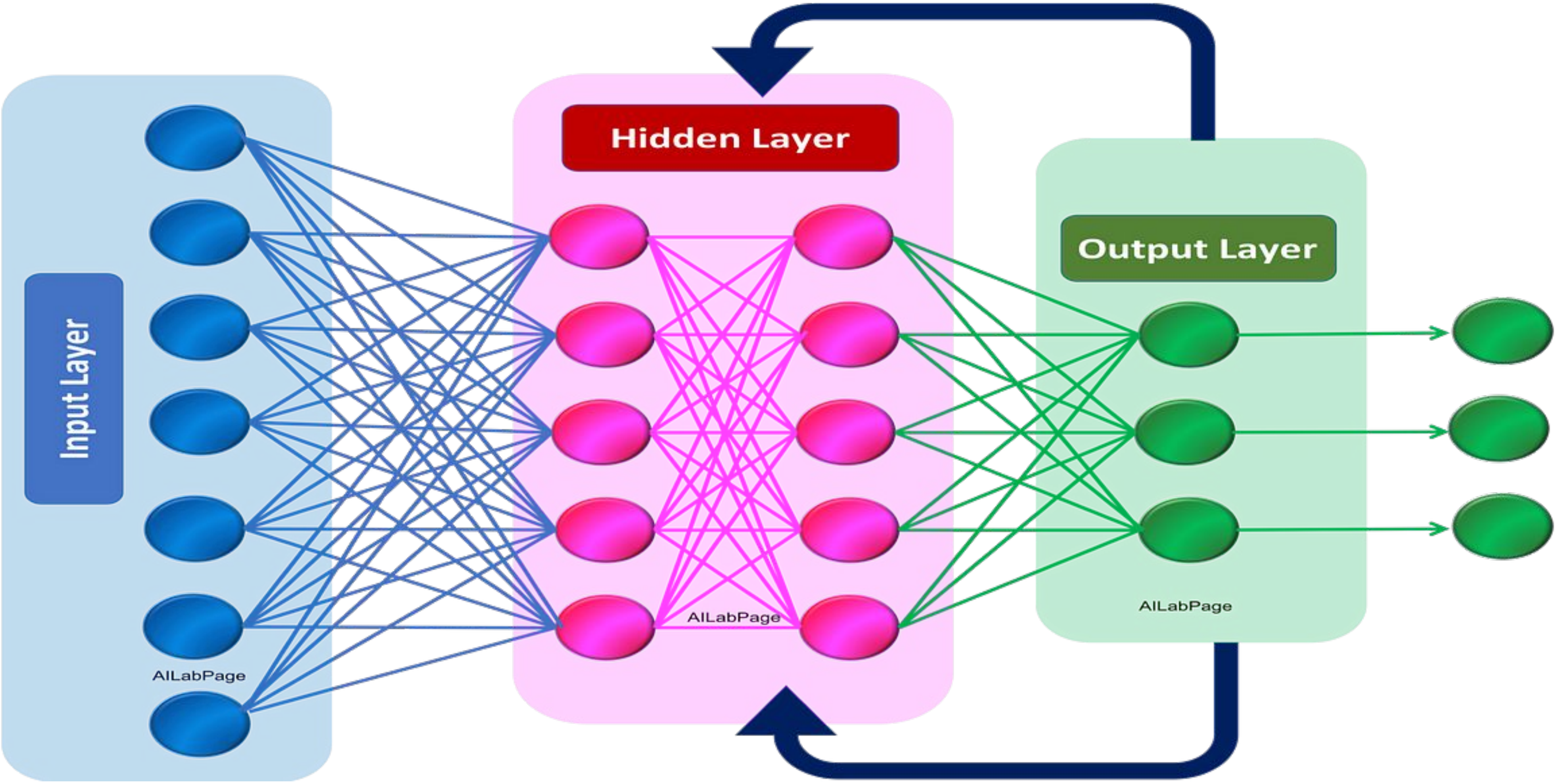
stateof.ai 2024

Large Language Model (LLM)

RNN, before LLM

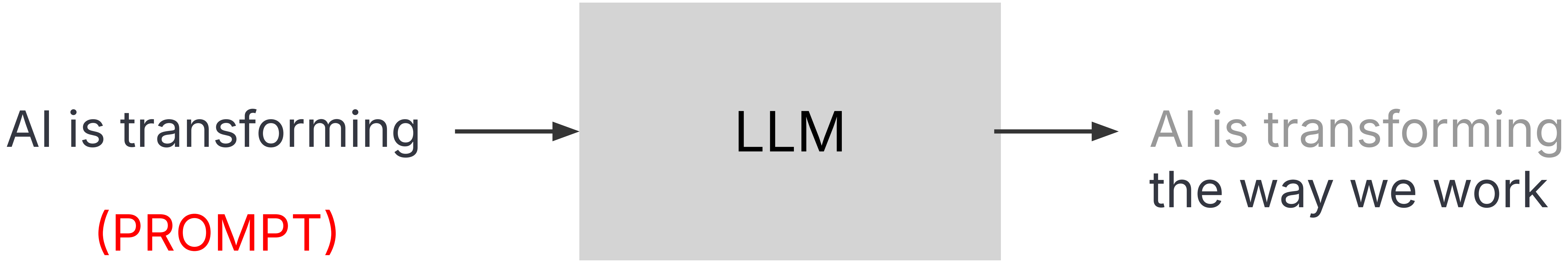
- **Recurrent Neural Networks** (RNN)
- Prediction of the next words based on the previous words
- RNN does not scale
- To complete a sentence the model needs to understand the structure of the entire sentence
- Eg. “The teacher taught the students with the book”
  - Did the teacher teach using the book?
  - Did the student have the book?
  - Or was it both?

RNN backpropagation



LLM

- **Large Language Model** (LLM) are probabilistic models that produce sentence in natural language
- These models work by completing sentences



AI is transforming

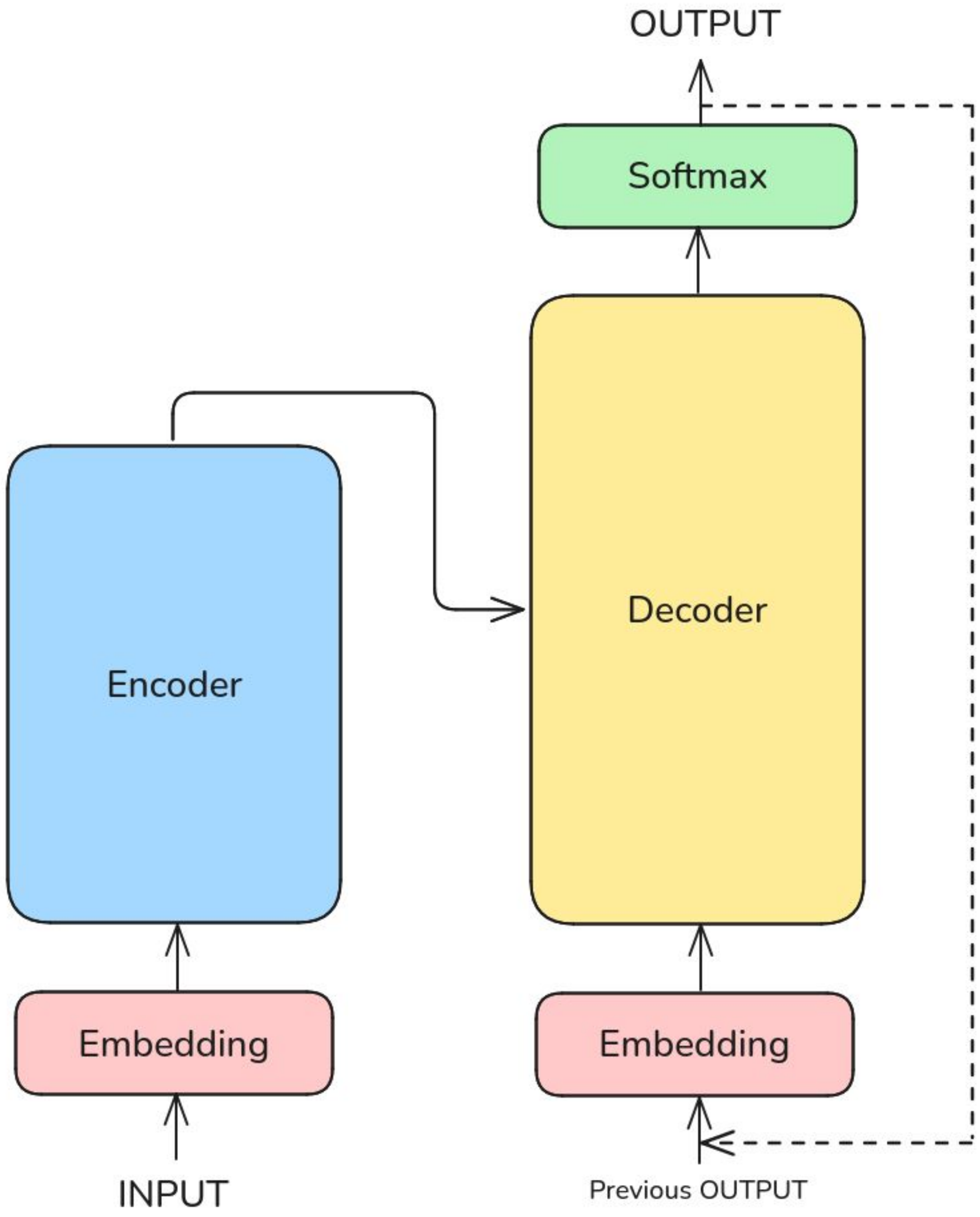
(PROMPT)

LLM

AI is transforming the way we work

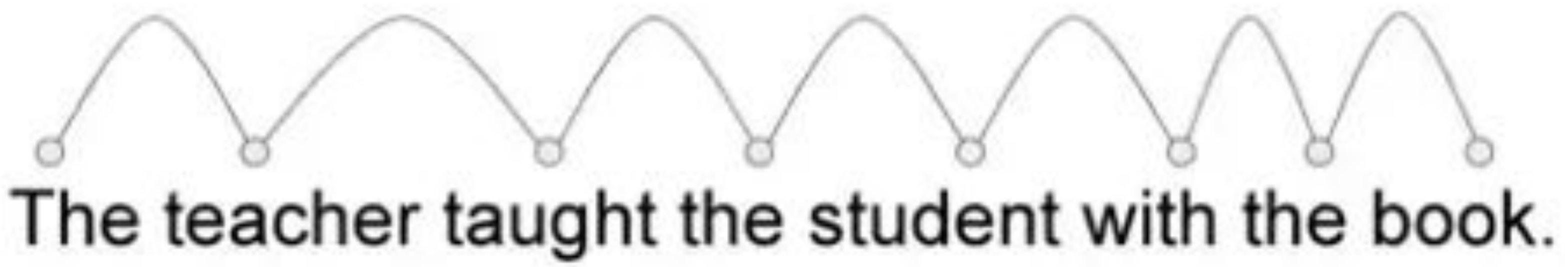
Transformer architecture

- Introduced in Attention is All You Need paper in 2017
- Basement of all LLMs
- The sentences are analyzed using a **self-attention** mechanism: each part of a sentence is evaluated in relation to every other part to understand contextual relationships and assign appropriate weights

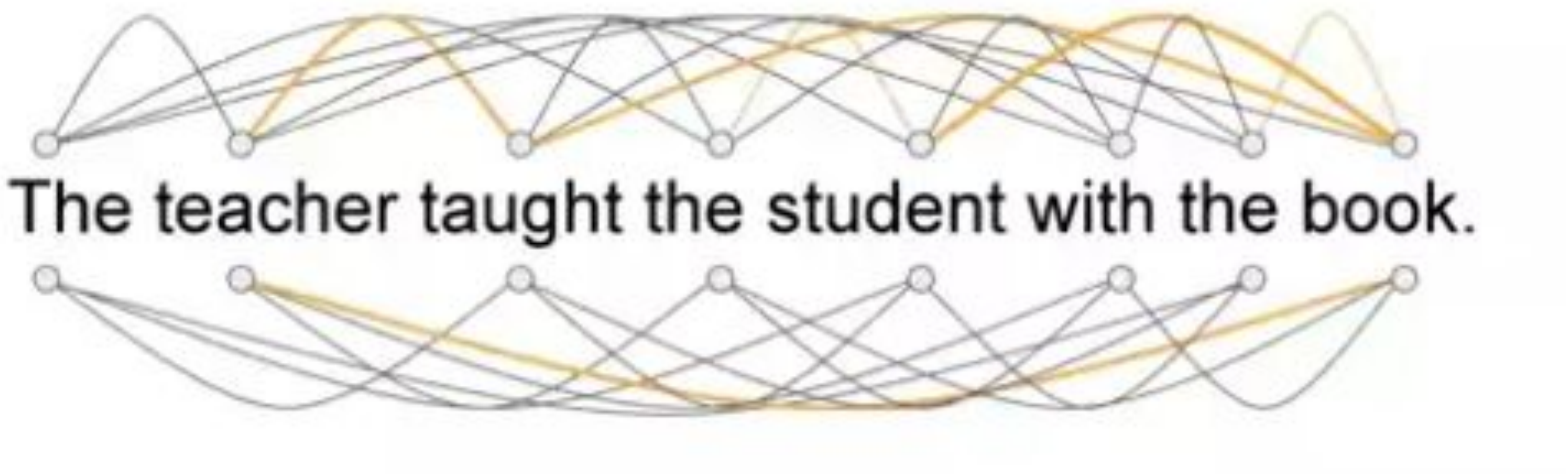


RNN vs Transformers

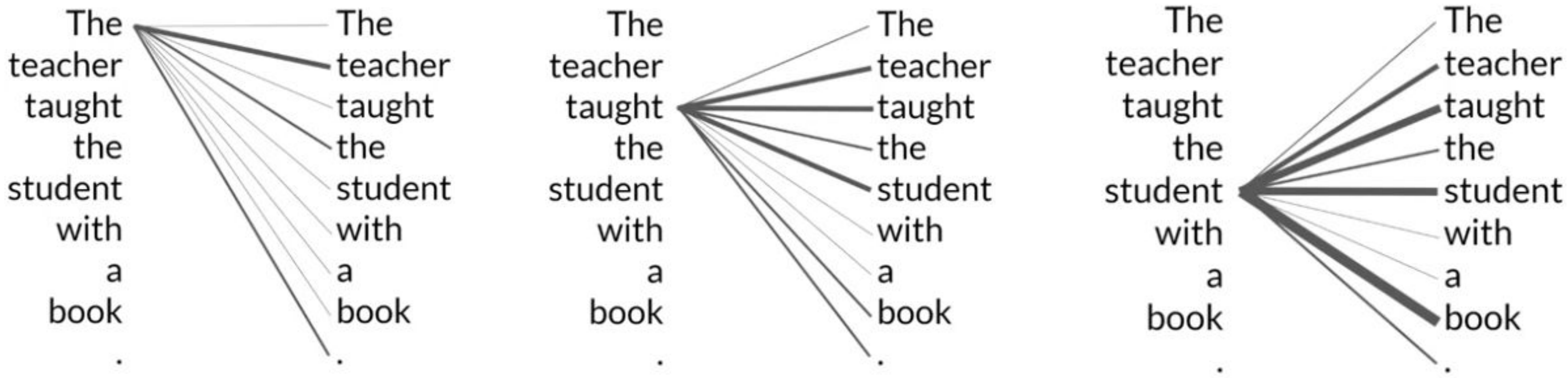
RNN



Transformers

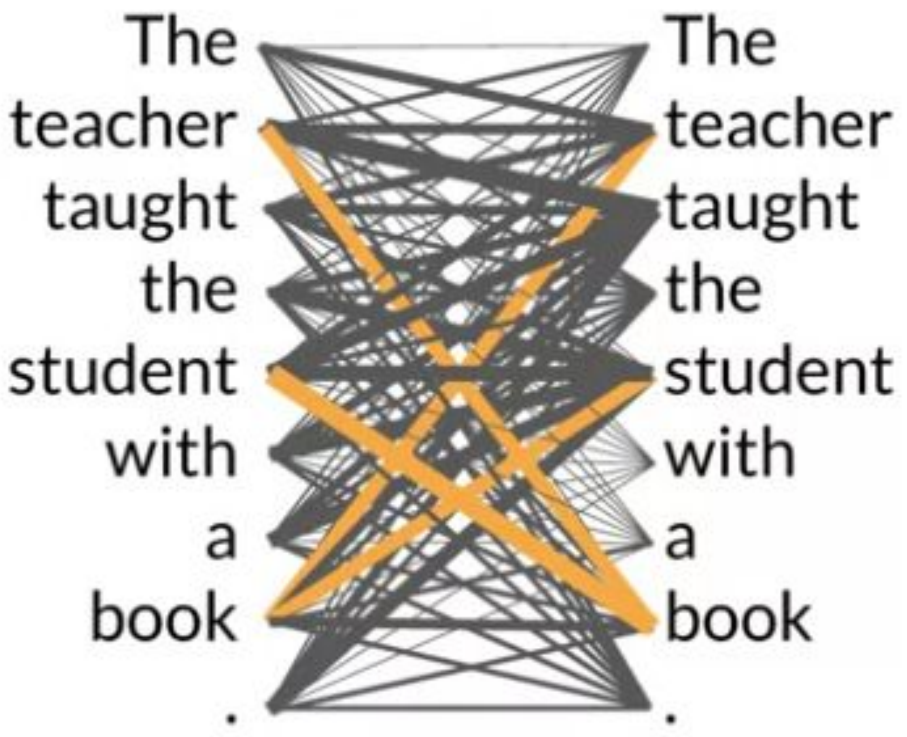


Attention map



The teacher taught the student with a book .
The teacher taught the student with a book .
The teacher taught the student with a book .
The teacher taught the student with a book .
The teacher taught the student with a book .
The teacher taught the student with a book .

eg. **book** is strongly connected with **teacher** and **student**



The teacher taught the student with a book .
The teacher taught the student with a book .

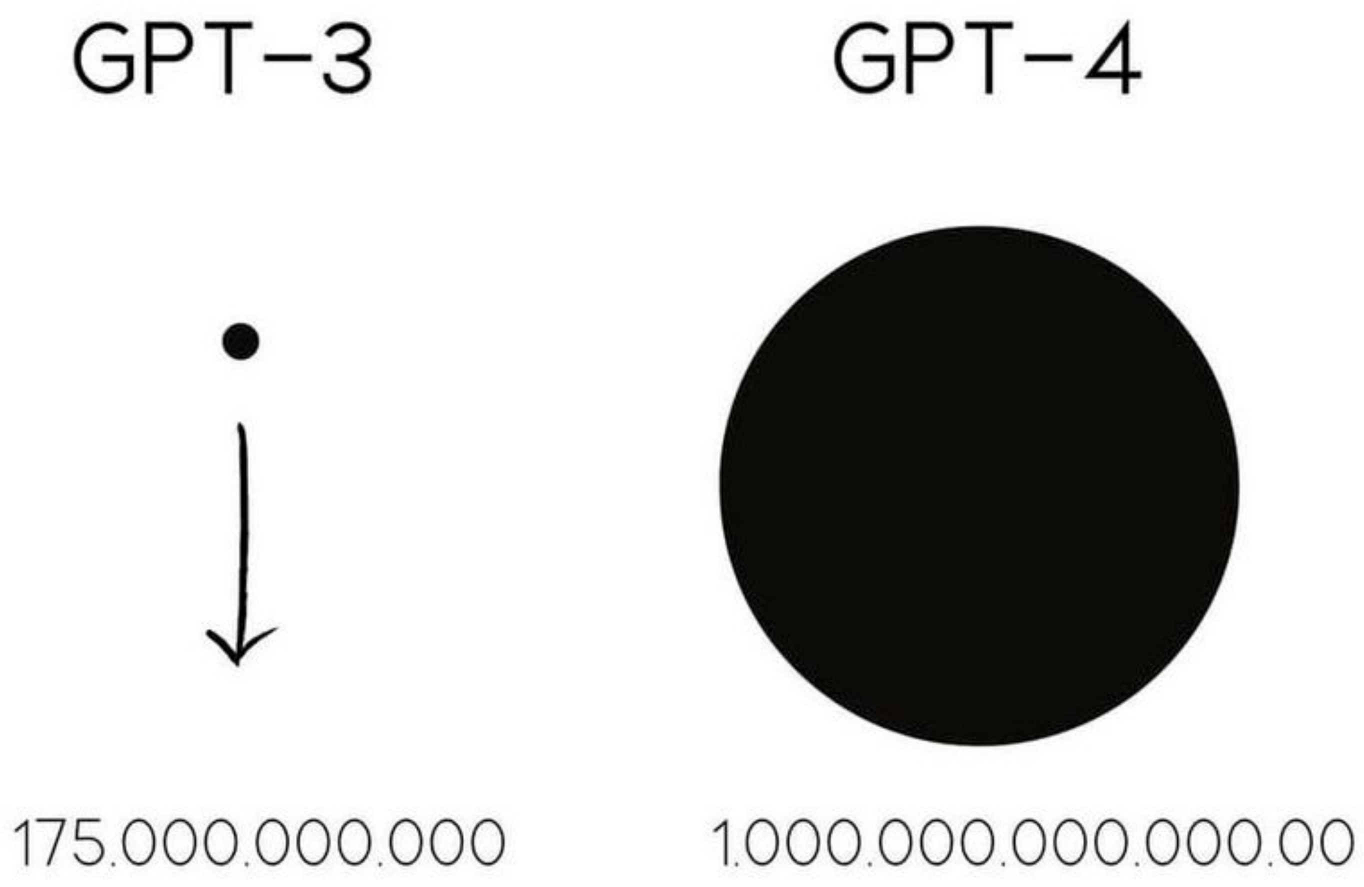
self-attention

LLM

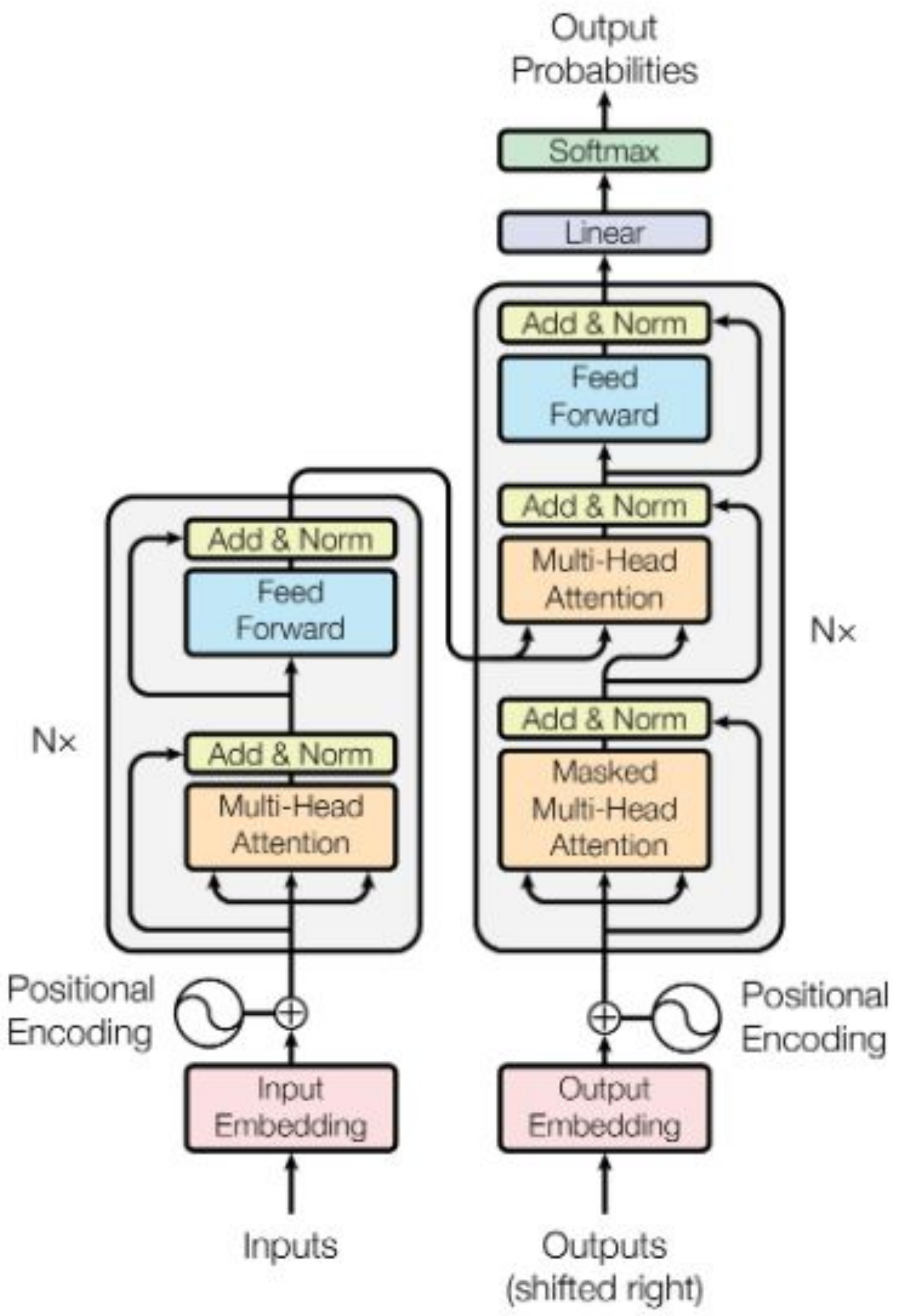
- **Large Language Model** (LLM) consisting of a neural network with many parameters (typically billions of weights or more), trained on large quantities of unlabelled text using self-supervised learning
- A message is splitted in **tokens**
- Each token is translated in a number using an operation called **embeddings**
- LLM **repeatedly predicting** the next token

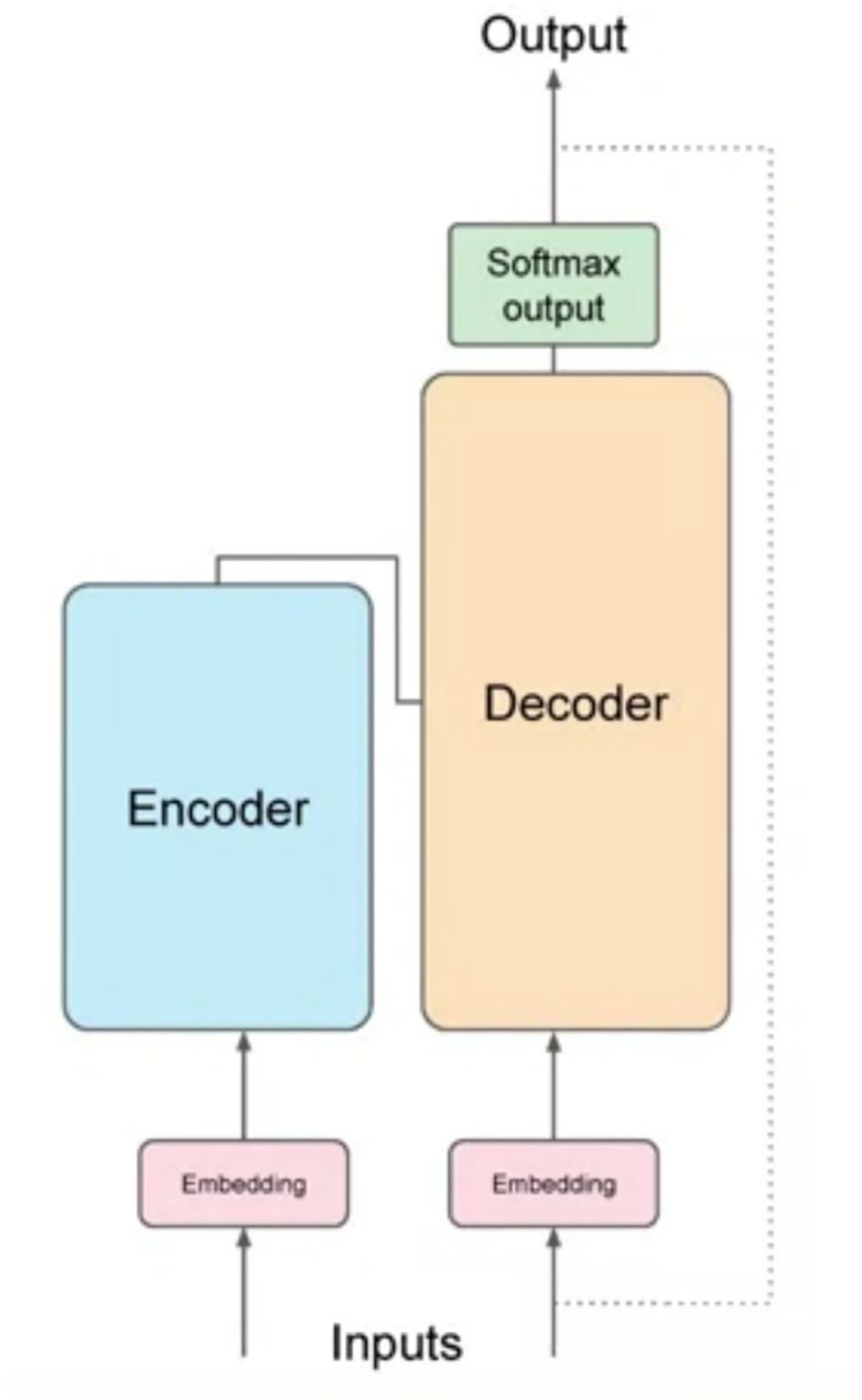
Size of GPT-4

- Around **1.76 trillion** parameters
- Neural network with **120** layers
- Process up to **25,000** words at once
- Estimated training cost is $200M using 10,000 Nvidia A100 GPU for 11 months



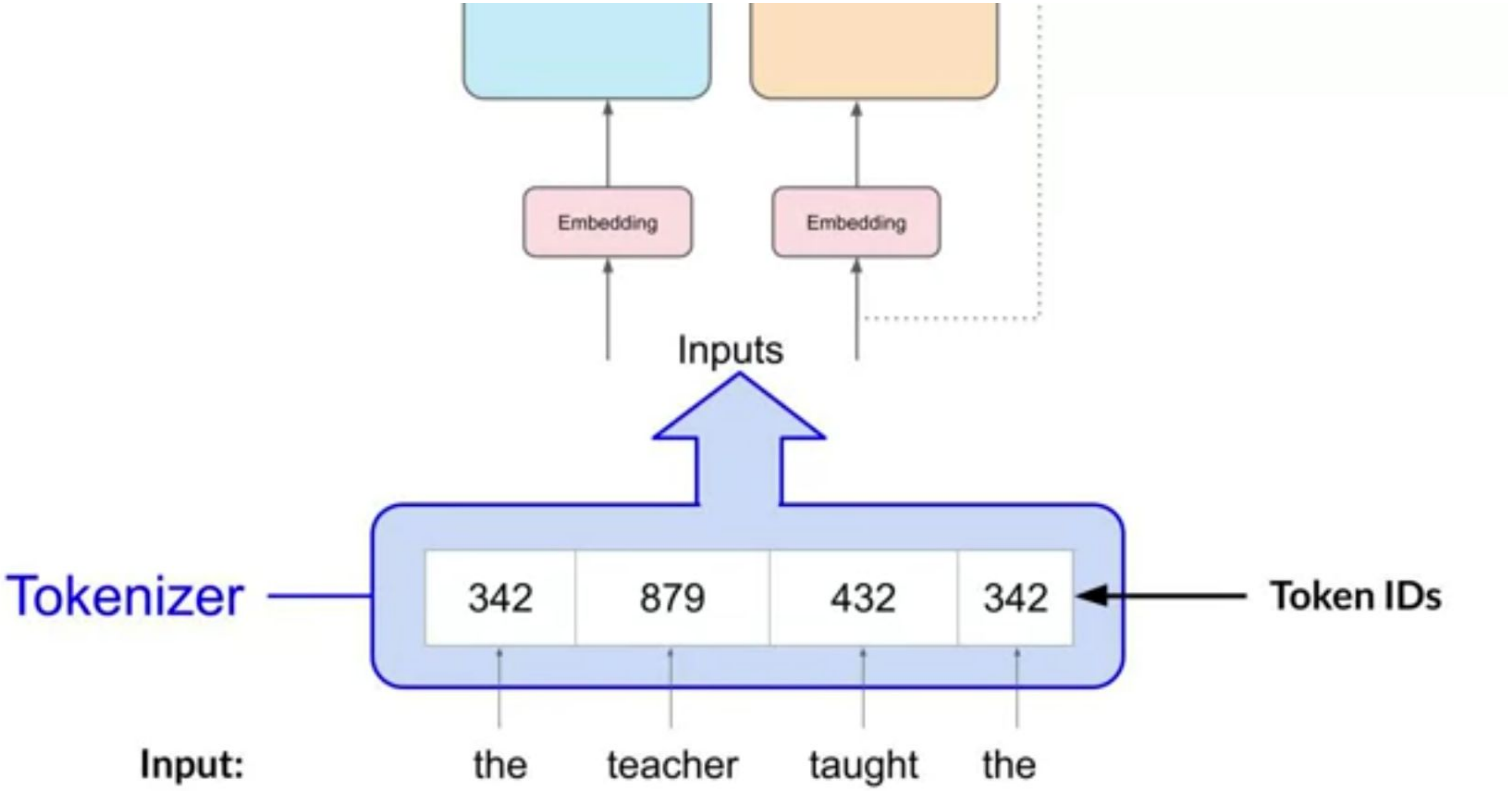
Transformers architecture



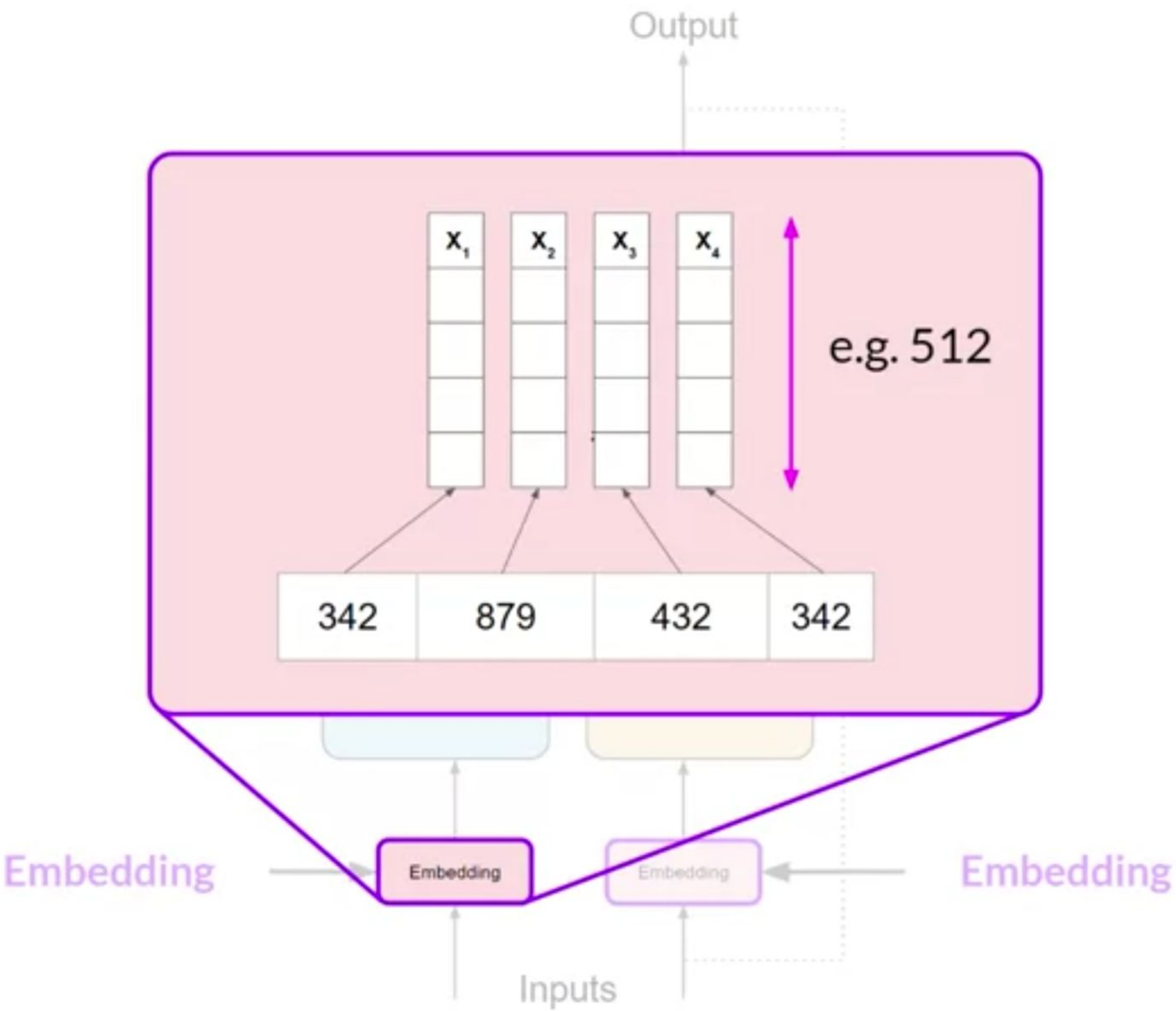


Tokenizer

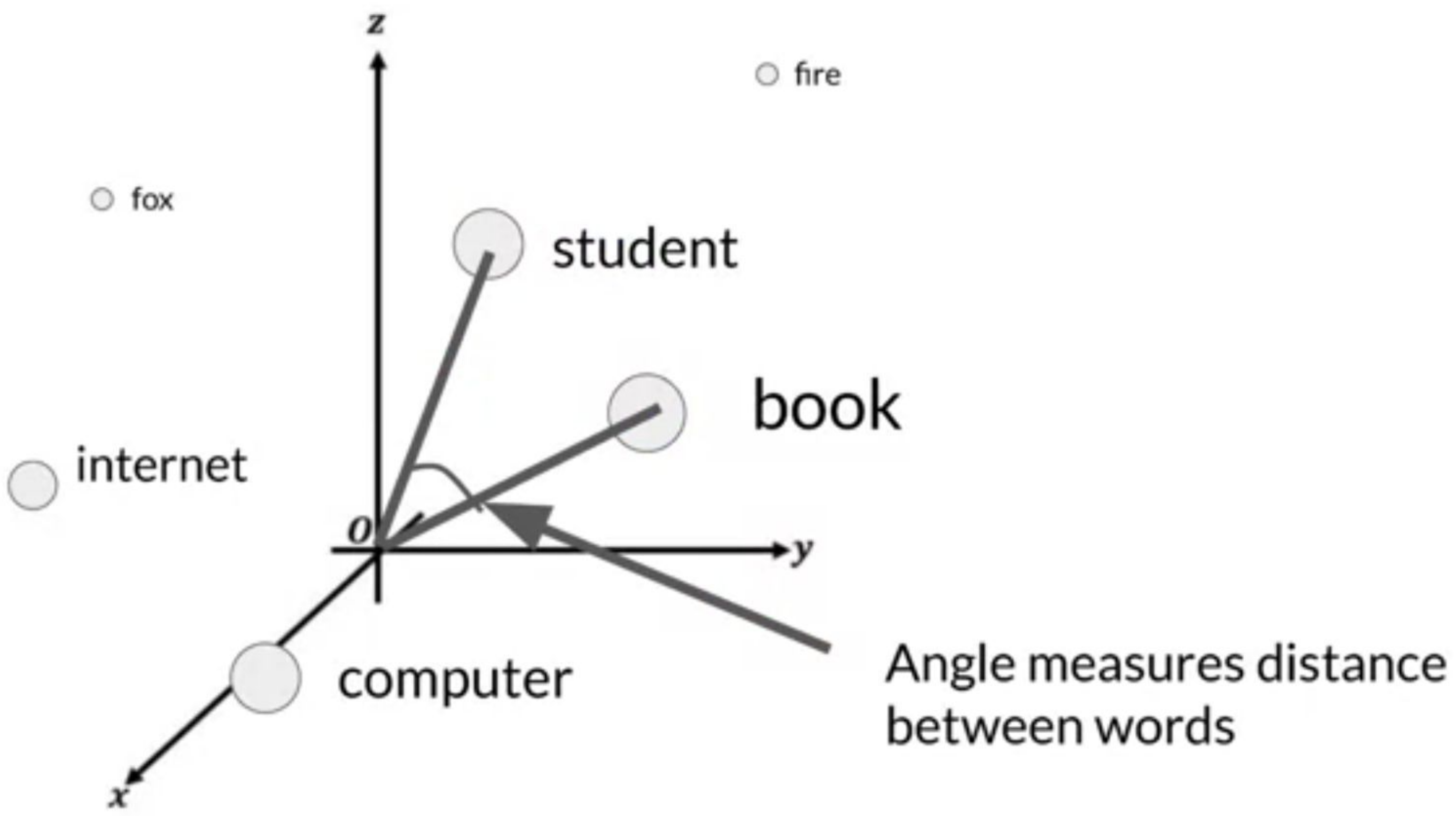
- We need to convert a sentence in numbers using a **tokenizer**



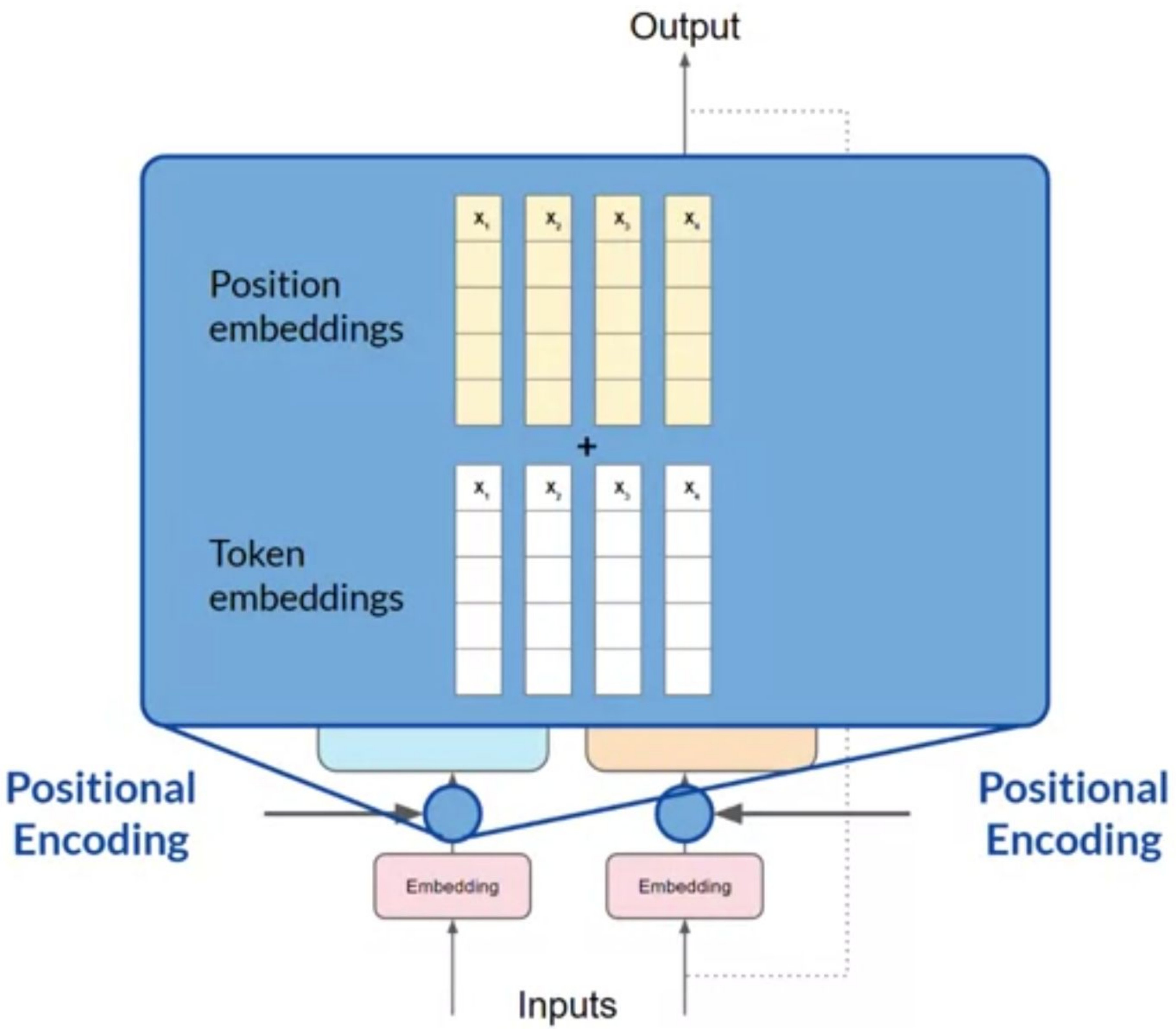
Embedding



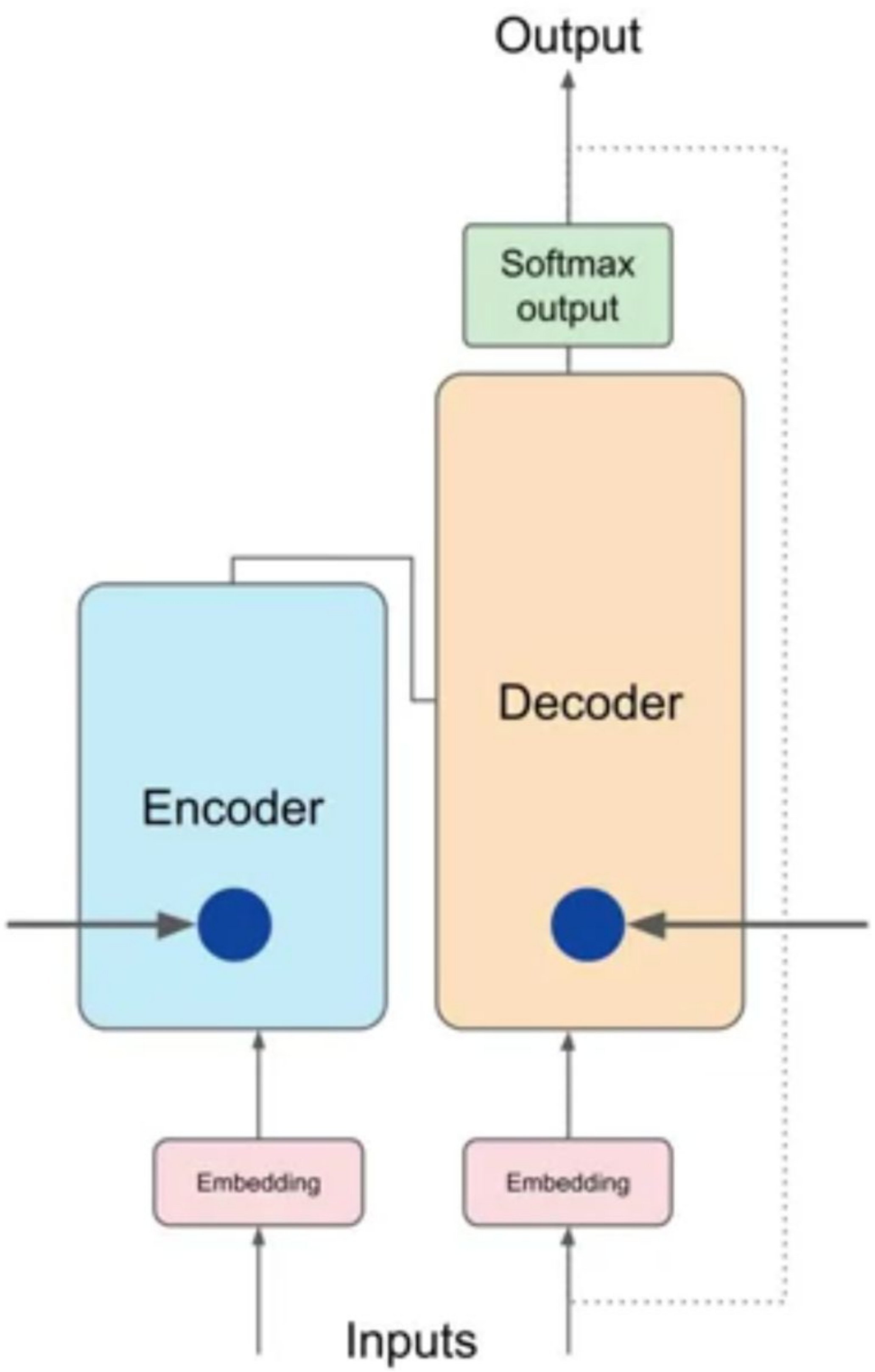
Example with 3 dimensions



Positional encoding



Self-attention



Self-attention

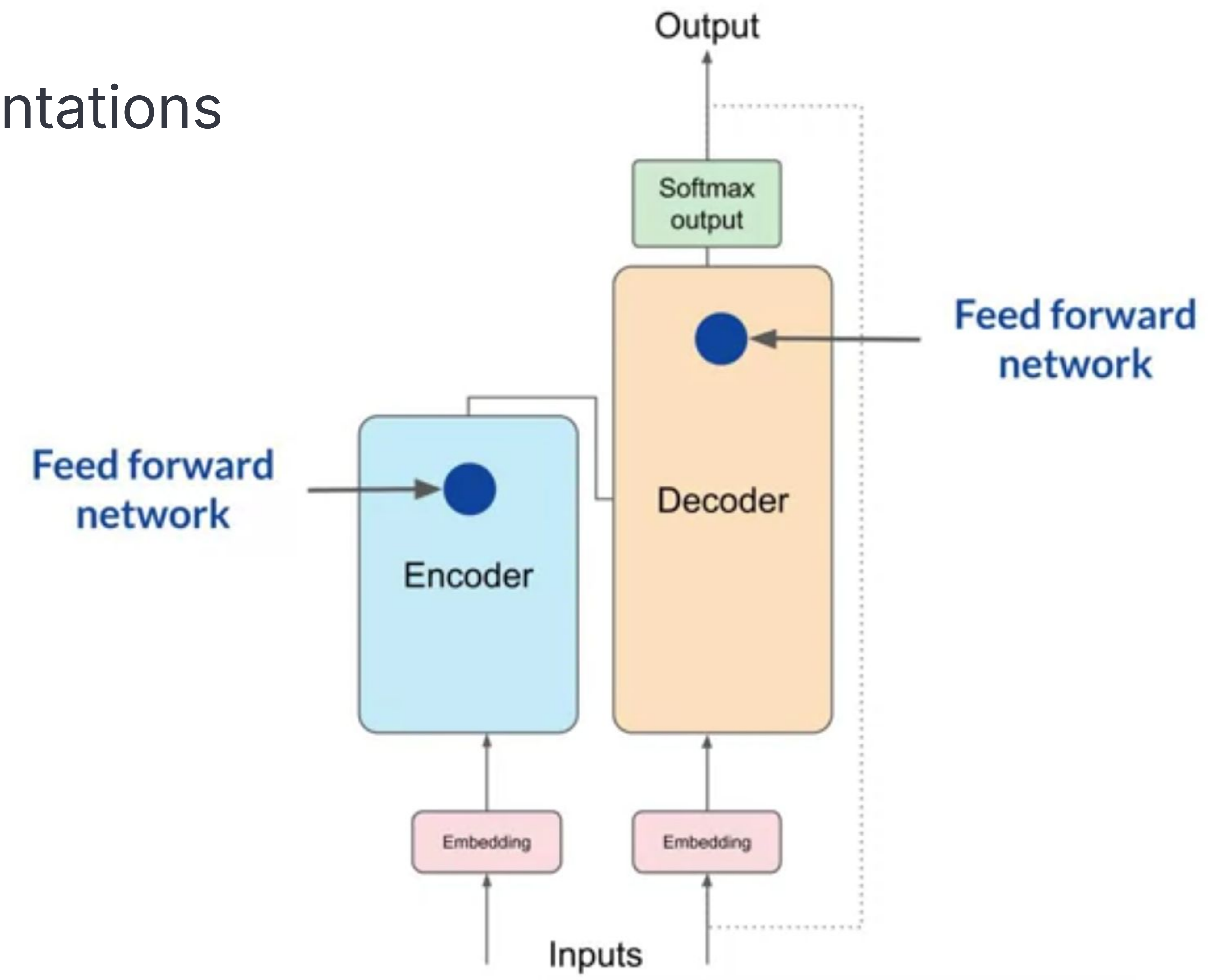
Multi-head Self-attention

Self-attention

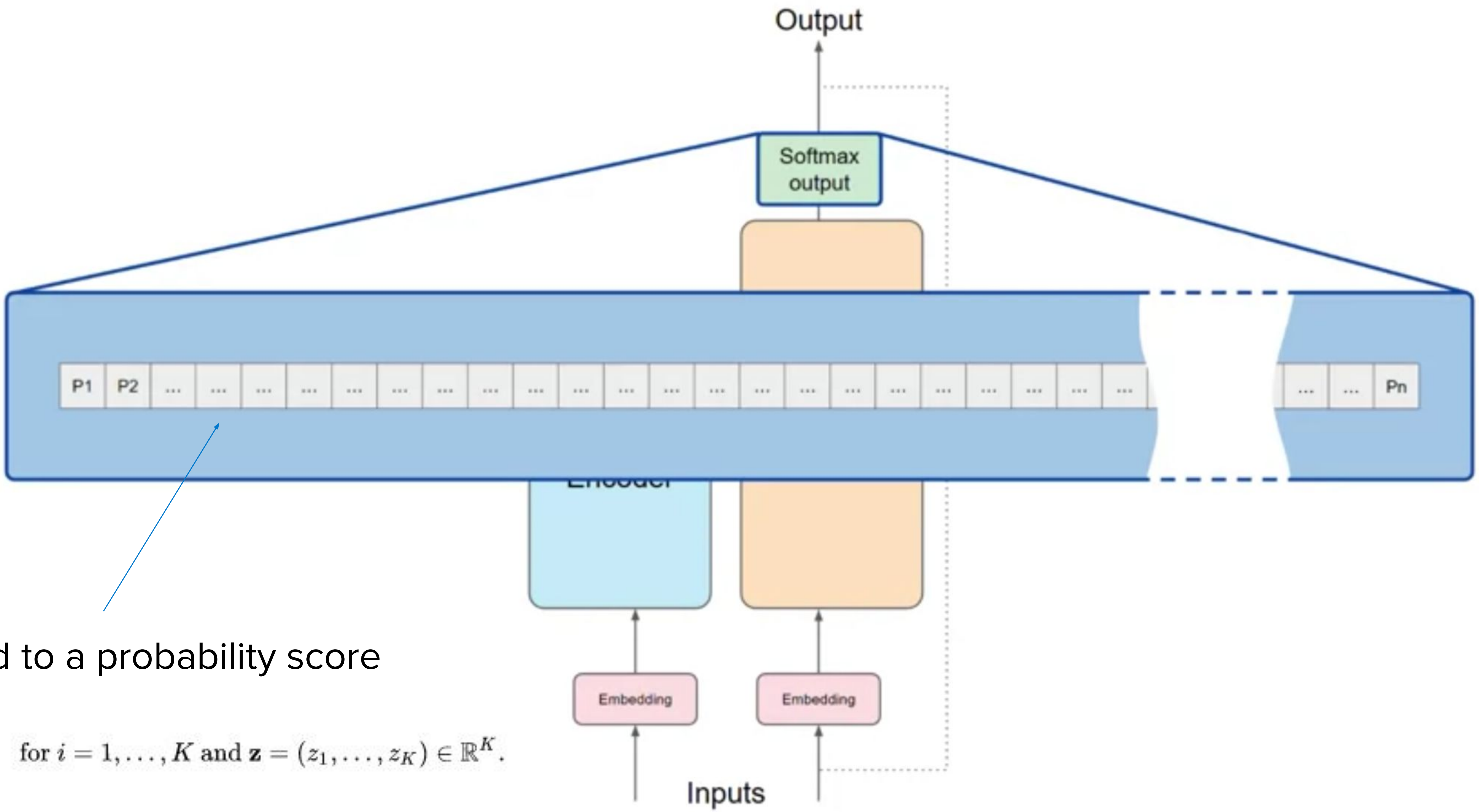
Multi-head Self-attention

Feed forward network

It helps refine the token representations learned from self-attention



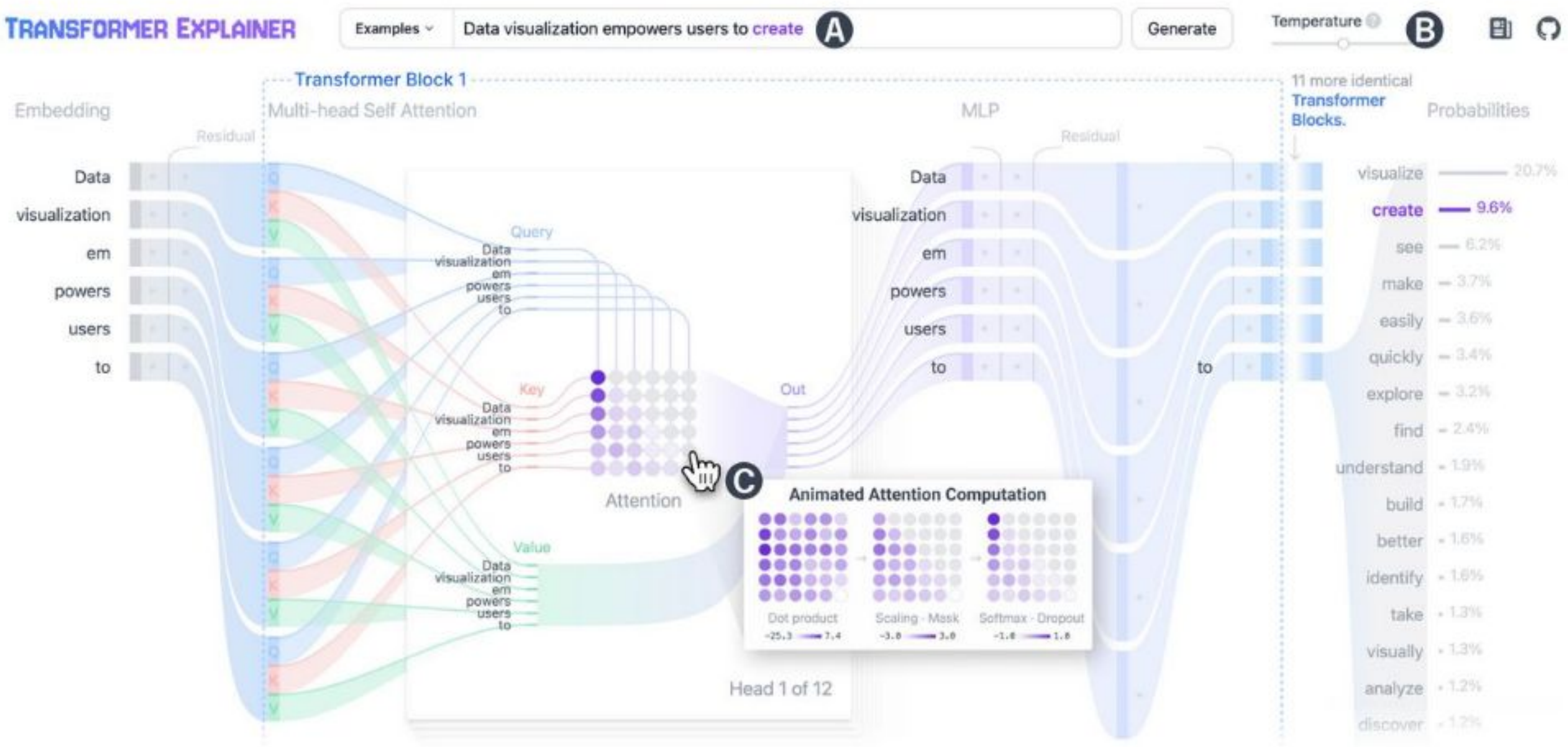
Softmax



Normalized to a probability score

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{for } i = 1, \dots, K \text{ and } \mathbf{z} = (z_1, \dots, z_K) \in \mathbb{R}^K.$$

Transformer Explainer



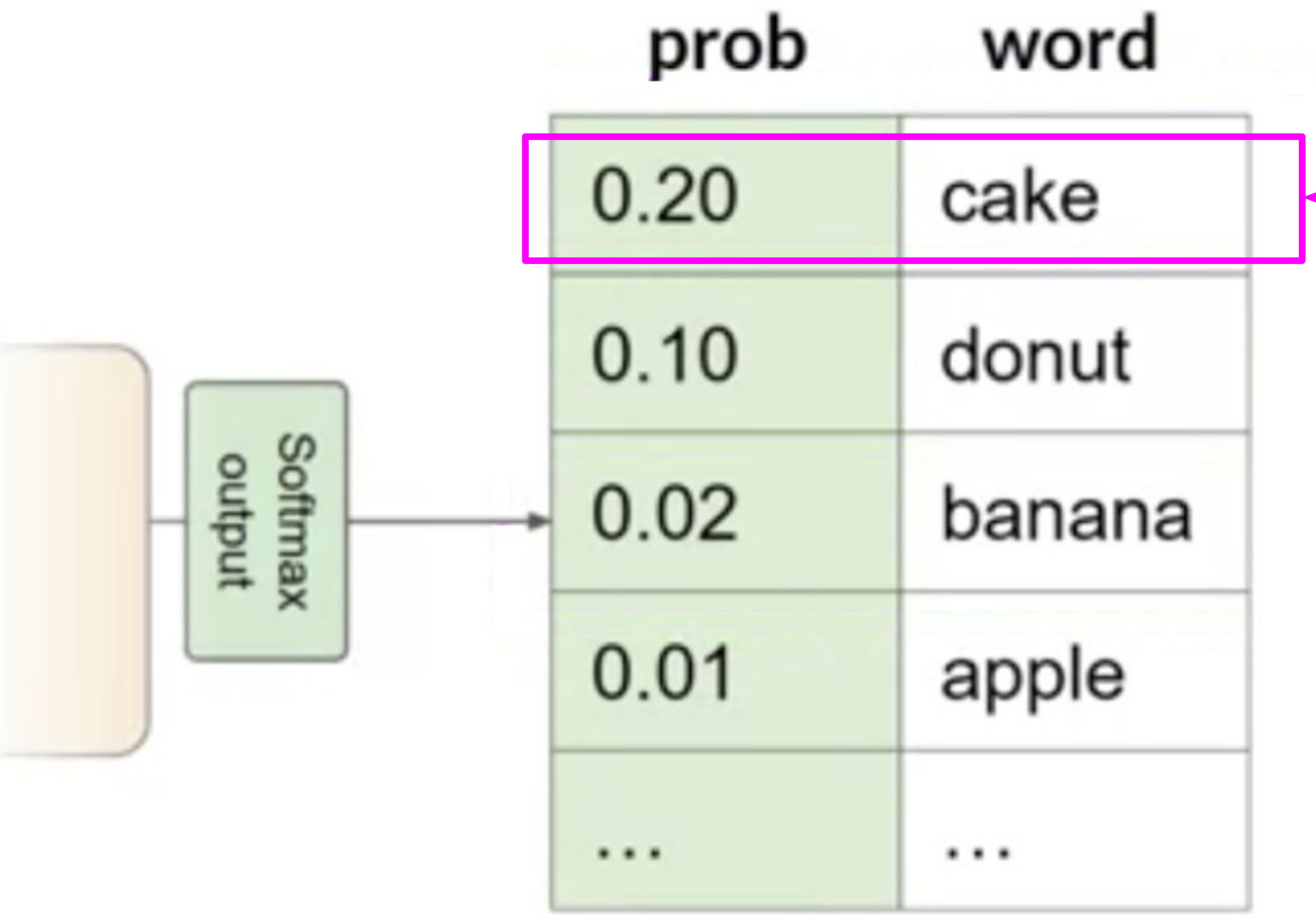
Examples ˅ Data visualization empowers users to create

Temperature

visualize	20.7%
create	9.6%
see	6.2%
make	3.7%
easily	3.6%
quickly	3.4%
explore	3.2%
find	2.4%
understand	1.9%
build	1.7%
better	1.6%
identify	1.6%
take	1.3%
visually	1.3%
analyze	1.2%
discover	1.2%

https://poloclub.github.io/transformer-explainer/

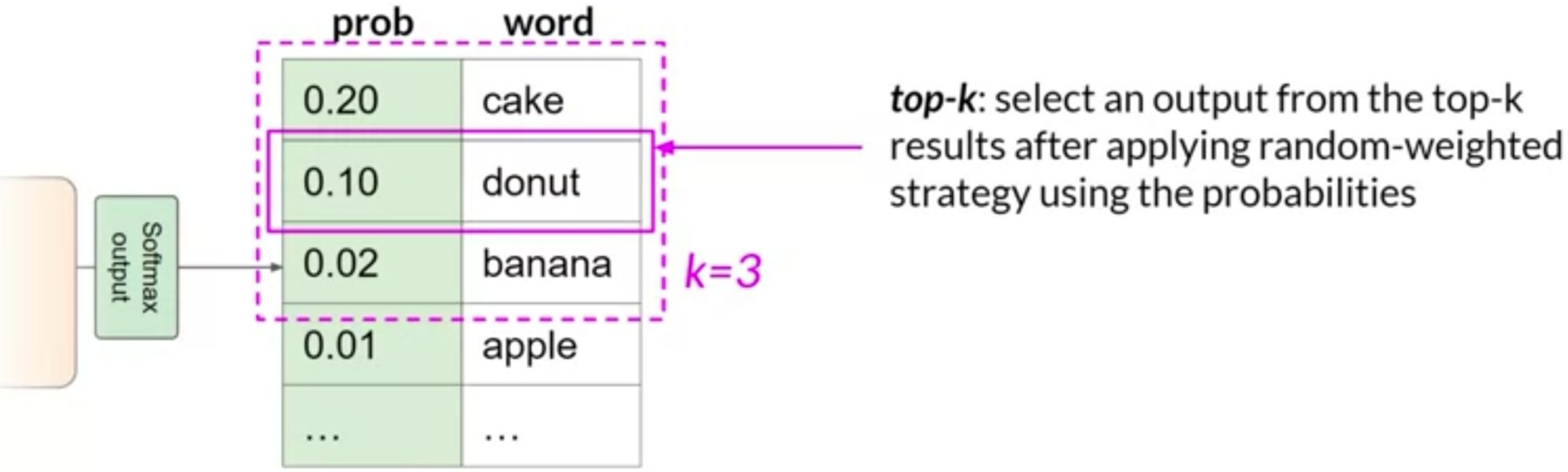
Predict the next word



prob	word
0.20	cake
0.10	donut
0.02	banana
0.01	apple
...	...

Choose the one with greatest probability (greedy algorithm)

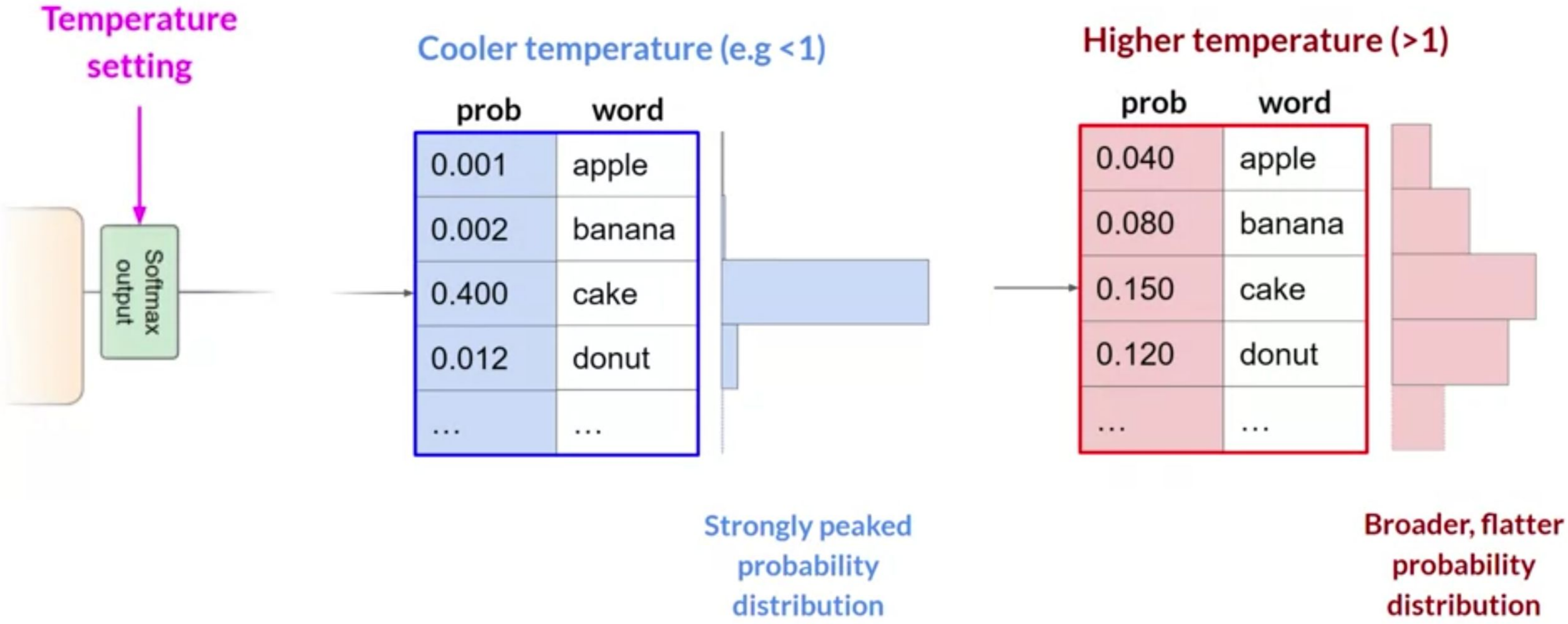
Top-k



prob	word
0.20	cake
0.10	donut
0.02	banana
0.01	apple
...	...

top-k: select an output from the top-k results after applying random-weighted strategy using the probabilities

Temperature



Cooler temperature (e.g <1)

prob	word
0.001	apple
0.002	banana
0.400	cake
0.012	donut
...	...

Strongly peaked probability distribution

Higher temperature (>1)

prob	word
0.040	apple
0.080	banana
0.150	cake
0.120	donut
...	...

Broader, flatter probability distribution

LLM visualization

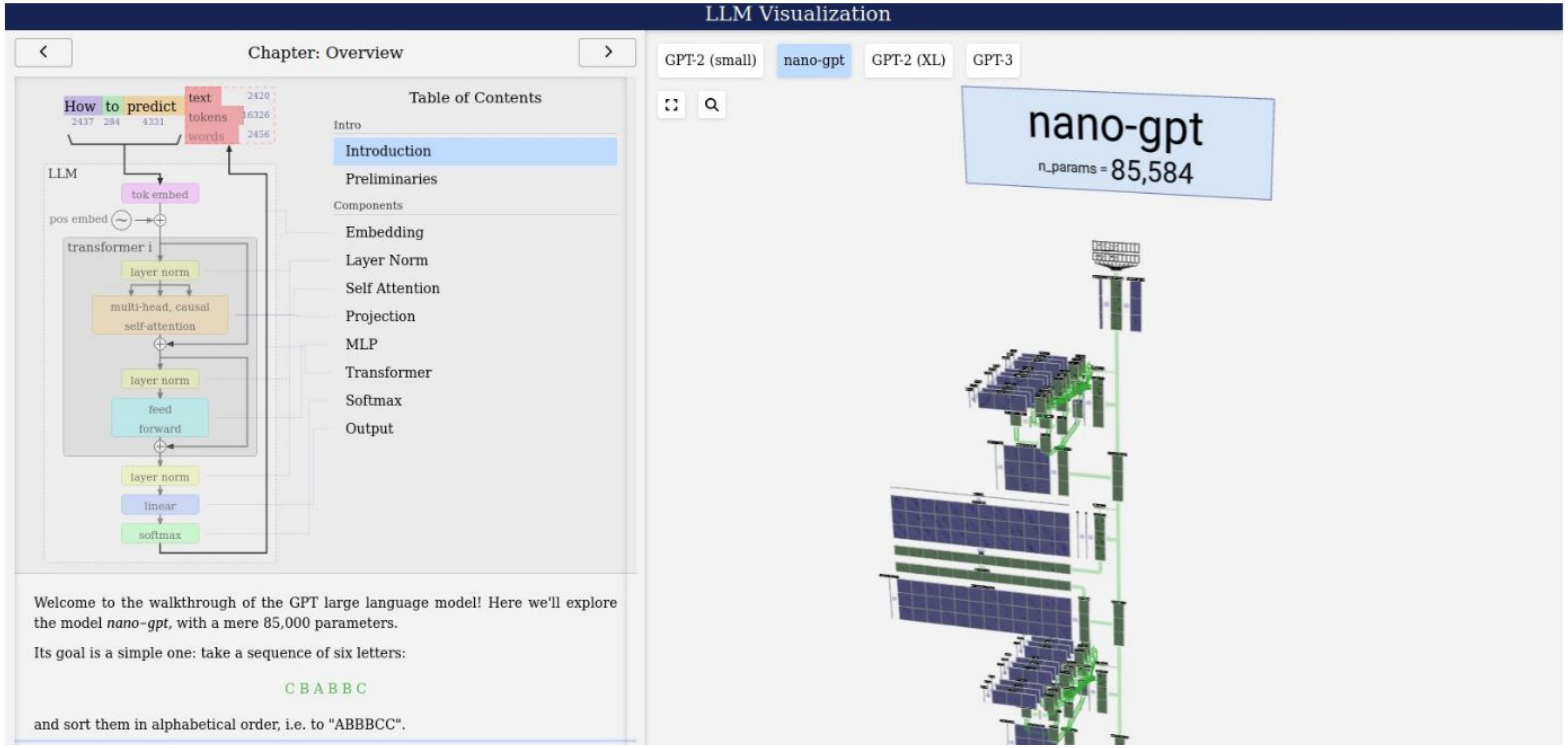


Table of Contents

Intro

- Introduction
- Preliminaries

Components

- Embedding
- Layer Norm
- Self Attention
- Projection
- MLP
- Transformer
- Softmax
- Output

Welcome to the walkthrough of the GPT large language model! Here we'll explore the model *nano-gpt*, with a mere 85,000 parameters.

Its goal is a simple one: take a sequence of six letters:

C B A B B C

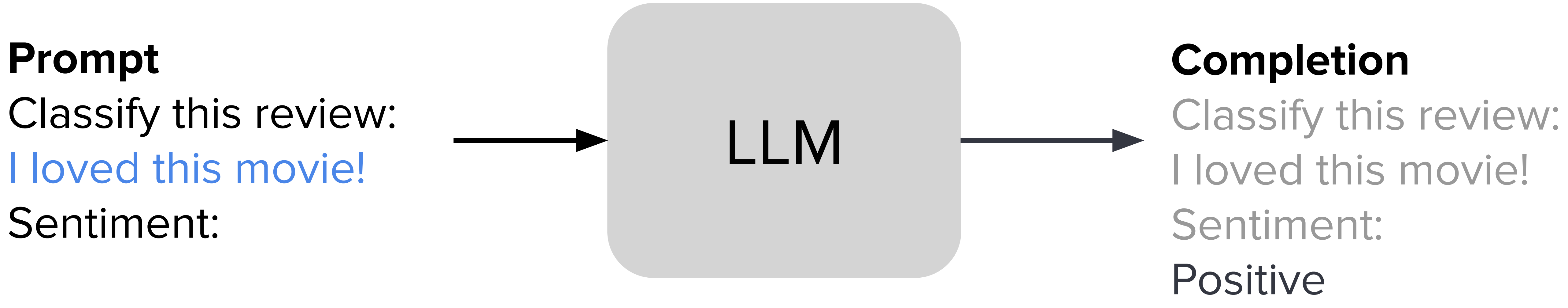
and sort them in alphabetical order, i.e. to "ABBBCC".

https://bbycroft.net/llm

Prompt engineering

- You can encounter situations where the model doesn't produce the outcome that you want on the first try
- You may have to revisit the language several times to get a good answer
- The development and improvement of the prompt is known as **prompt engineering**
- One powerful strategy is to include examples of the task that you want the model to carry out inside the prompt
- This is called **In-Context Learning (ICL)**

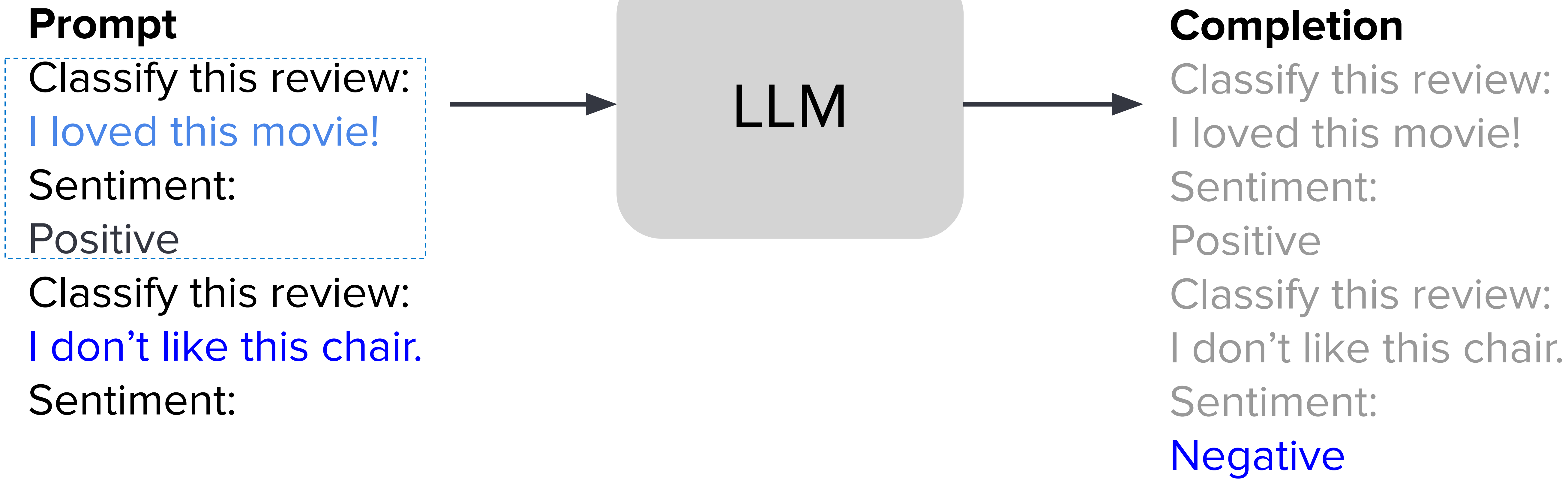
ICL - zero shot inference



Prompt
Classify this review:
I loved this movie!
Sentiment:

Completion
Classify this review:
I loved this movie!
Sentiment:
Positive

ICL - one shot inference



Prompt

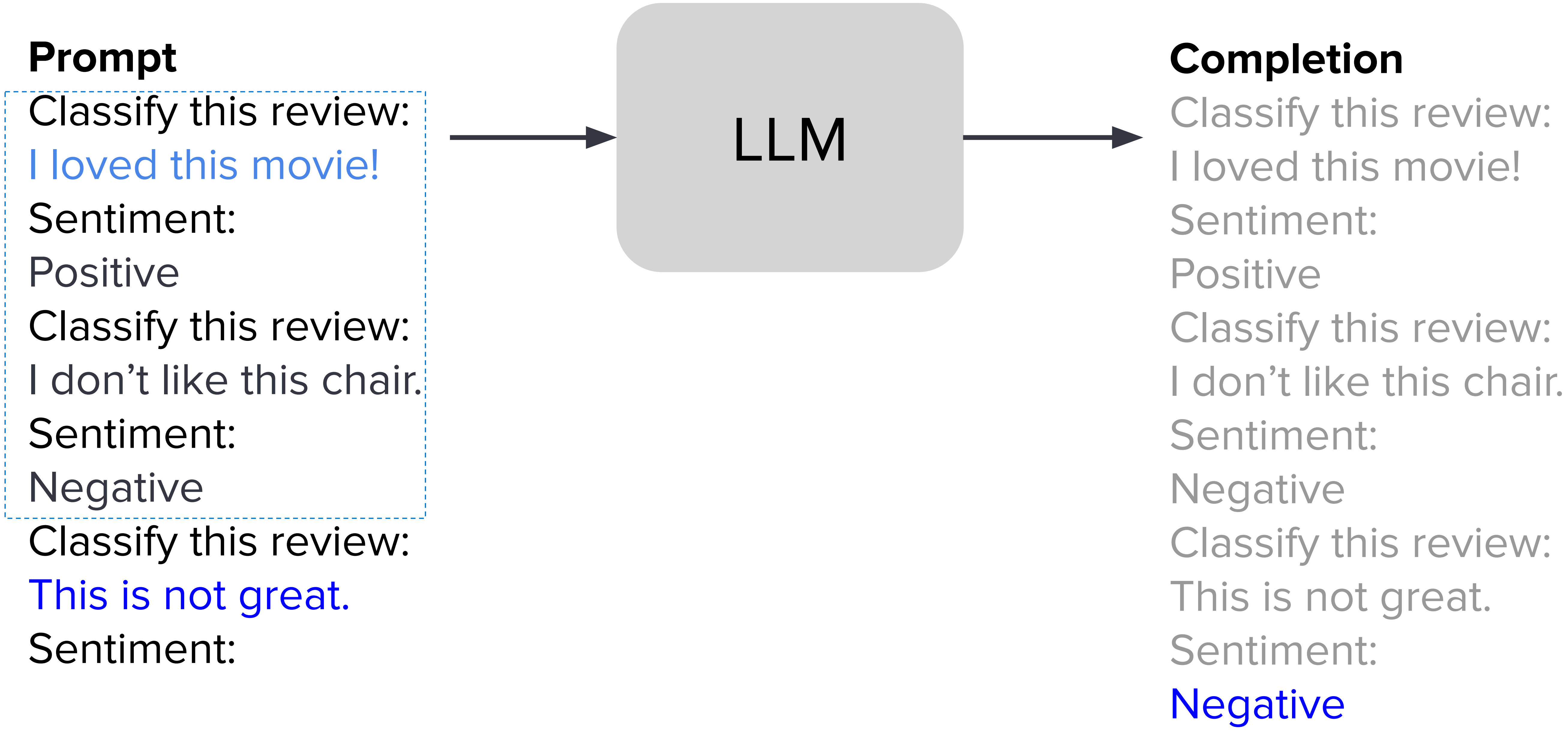
Classify this review:
I loved this movie!
Sentiment:
Positive
Classify this review:
I don’t like this chair.
Sentiment:

LLM

Completion

Classify this review:
I loved this movie!
Sentiment:
Positive
Classify this review:
I don’t like this chair.
Sentiment:
Negative

ICL - few shot inference



Prompt

Classify this review:
I loved this movie!
Sentiment:
Positive
Classify this review:
I don’t like this chair.
Sentiment:
Negative
Classify this review:
This is not great.
Sentiment:

LLM

Completion

Classify this review:
I loved this movie!
Sentiment:
Positive
Classify this review:
I don’t like this chair.
Sentiment:
Negative
Classify this review:
This is not great.
Sentiment:
Negative

Ollama

- Ollama is a software for downloading and running LLMs locally
- Llama 3, Phi 3, Mistral, Gemma, and other models
- Simple command line tool:
  - ollama pull llama3.2:3b
  - ollama run llama3.2:3b

Laboratory:
LLM examples with Ollama and OpenAI

Sorgenti Python

Retrieval-Augmented Generation (RAG)

elastic

Retrieval-Augmented Generation (RAG)

- **RAG** is a technique in natural language processing that combines information retrieval systems with **Large Language Models** (LLM) to generate more informed and accurate responses
- It is composed by the following parts:
  - **Retrieval-Augmented**
  - **Generation**

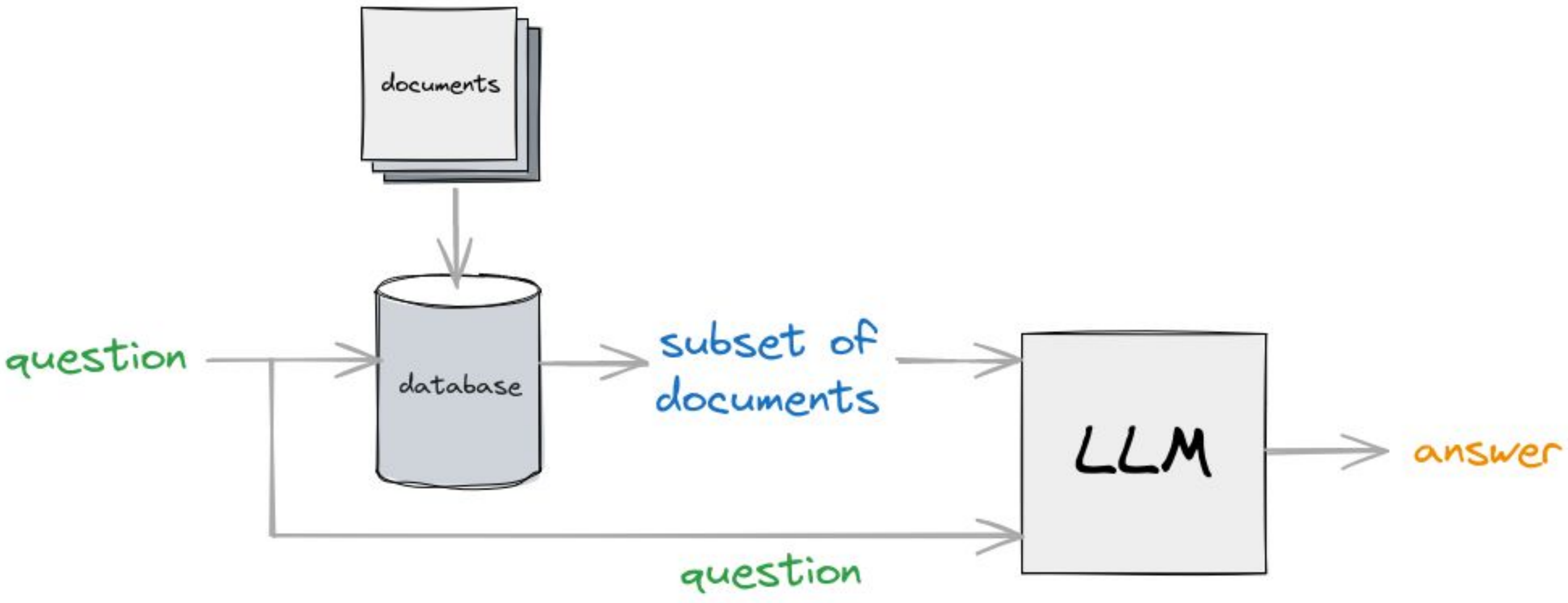
Generation

- LLMs are very powerful but have some limitations:
  - **No source** (potential hallucinations)
    - How can I verify the information coming from an LLM?
    - What sources has been used to generate the answer?
  - **Out of date**
    - An LLM is trained in a period of time
    - For update we need to retraining the model (very expensive)

Retrieval-Augmented

- We collect sets of private or public document
- We build a **retrieval system** (e.g. a database) to extract a subset of documents using a **question**
- Then we pass the **question + documents found** to an LLM as prompt with a context
- The LLM can give an answer using the updated documents

RAG architecture

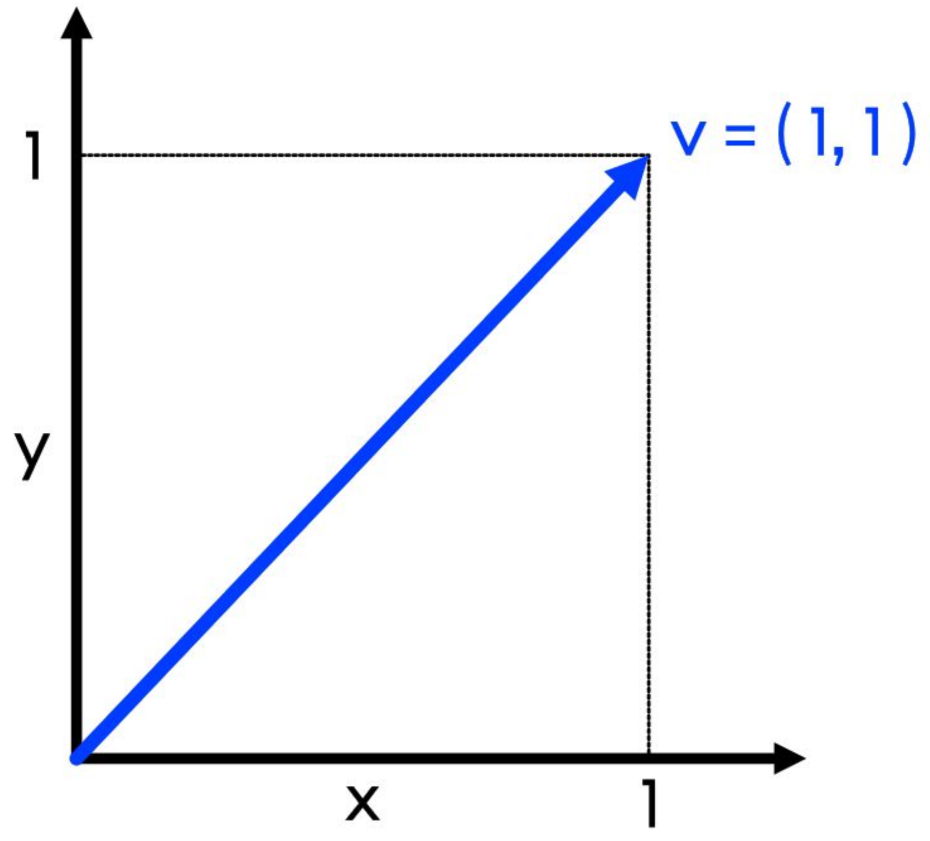


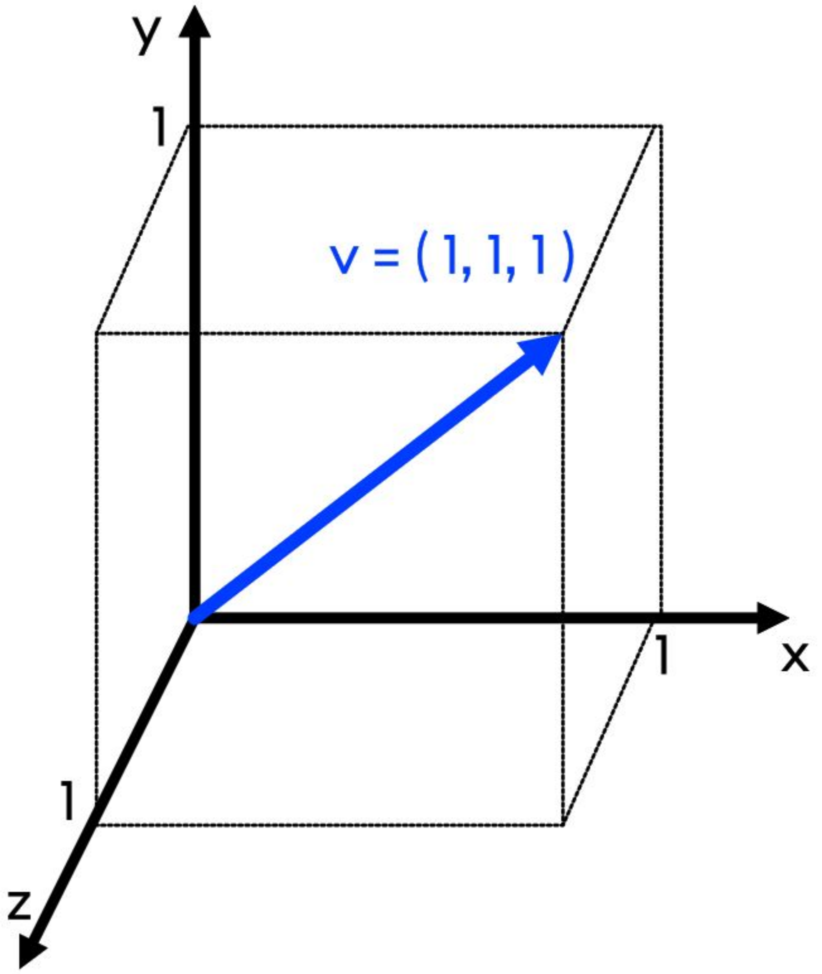
Retrieve documents from a question

- How we can retrieve documents in a database using a question?
- We need to use **semantic search**
- One solution is to use a **vector database**
- A vector database is a system that uses **vectors** (set of numbers) to retrieve information

What is a vector?

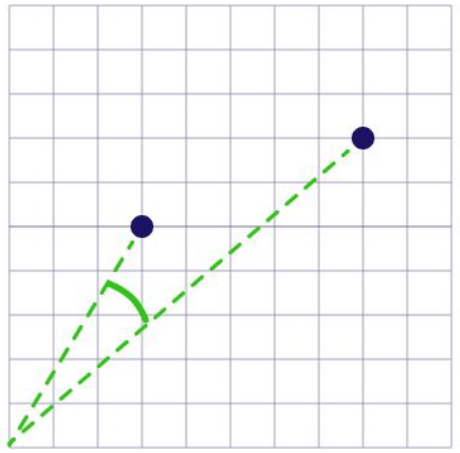
- A vector is a set of numbers
- Example: a vector of 3 elements [2, 5, -10]
- A vector can be represented in a multi-dimensional space (eg. Llama3.2 uses 3072 dimensions)





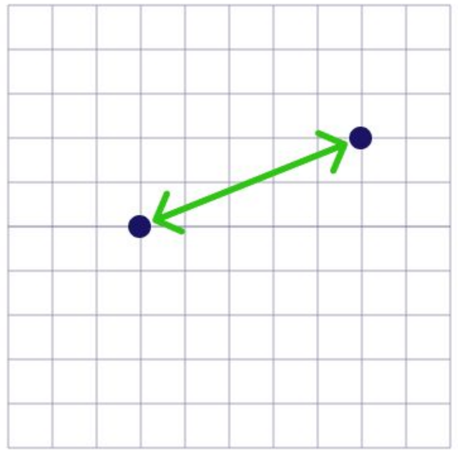
Similarity between two vectors

- Two vectors are (semantically) similar if they are close to each other
- We need to define a way to measure the similarity



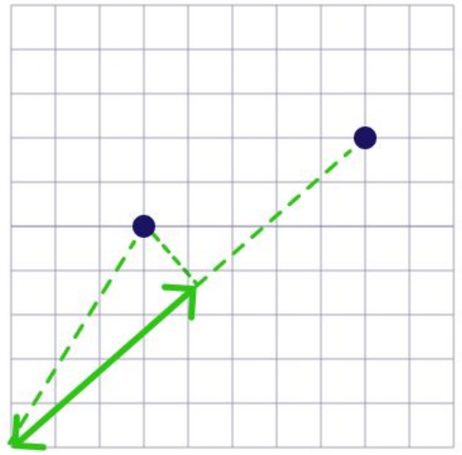
Cosine Distance

$$1 - \frac{A \cdot B}{||A|| \quad ||B||}$$



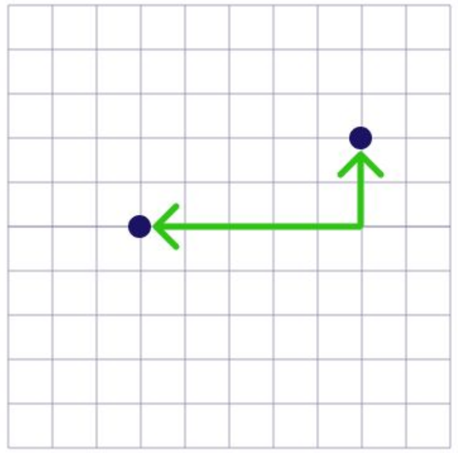
Squared Euclidean (L2 Squared)

$$\sum_{i=1}^{n} (x_i - y_i)^2$$



Dot Product

$$A \cdot B = \sum_{i=1}^{n} A_i B_i$$

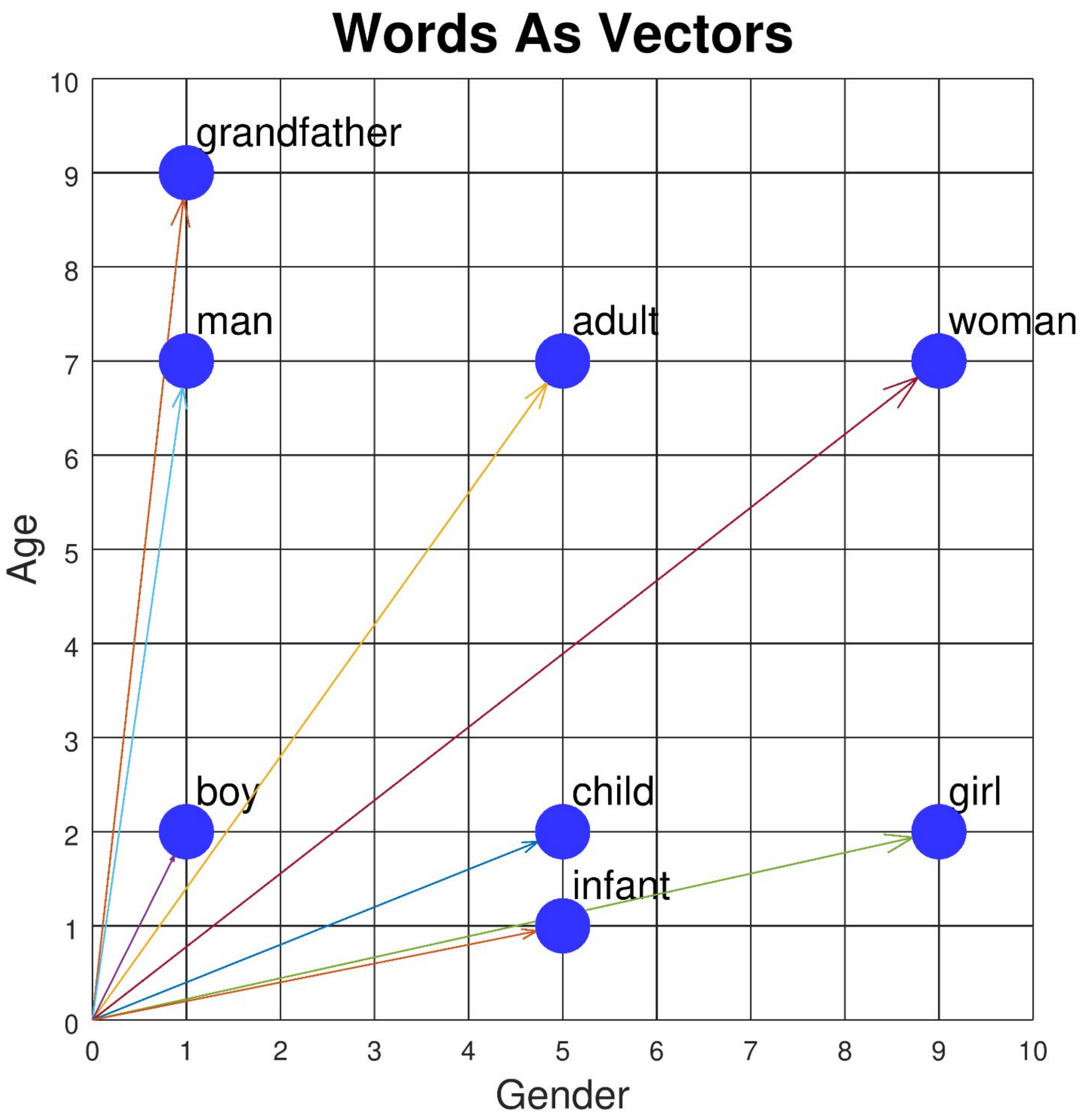


Manhattan (L1)

$$\sum_{i=1}^{n} |x_i - y_i|$$

Embedding

- Embedding is the translation of an input (document, image, sound, movie, etc) to a vector
- There are many techniques, using an LLM typically this is done by a neural network
- The goal is to group information that are semantically related to each other
- https://projector.tensorflow.org/



Vector database + LLM

- The search query (***question***) is in natural language
- We use semantic search to retrieve top-n relevant documents (***context***)
- We send the following prompt to the LLM (example):
    - *Given the following* ***{context}*** *answer to the following* ***{question}***

elastic

Split the documents in chunk

- We need to store data in the vector database using chunk of information
- We cannot use big documents since we need to pass it in the context part of the prompt for an LLM that typically has a token limit (e.g. Llama3.2 up to 128K)
- We need to split the documents in **chunk** (part of words)

Elasticsearch (vector database)

- Elasticsearch is Free and Open Source (AGPL), Distributed, RESTful Search Engine
- Distributed search and analytics engine, scalable data store and **vector database** optimized for speed and relevance on production-scale workloads.
- You can run it locally with a single command:
  - **curl -fsSL https://elastic.co/start-local | sh**

LangChain

- LangChain is an open source composable framework to build with LLMs
- Supports all the LLMs (see here)
- Integrations with many vector databases (e.g. Chroma, Elasticsearch, Milvus, Qdrant, Redis)
- Available for Python (98K ⭐) and Javascript (13K ⭐)
- MIT license
- Other interesting projects: LangGraph (MIT license) and LangSmith (commercial)

Laboratory:
RAG with LangChain + Llama 3.2 + Elasticsearch

Google Colab

Sorgenti Python

References

- What is retrieval-augmented generation? IBM research
- Ashish Vaswan et al., Attention Is All You Need, Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017)
- Albert Ziegler, John Berryman, A developer's guide to prompt engineering and LLMs, Github blog post
- Sebastian Raschka, Build a Large Language Model (From Scratch), Manning, 2024
- Elasticsearch as vector database, Elastic Search Labs
- Elasticsearch search relevance, Elastic Search Labs
- E.Zimuel, Retrieval-Augmented Generation for talking with your private data using LLM, AI Heroes 2023 conference, Turin (Italy)
- L. Gianfagna, E. Zimuel, Explainable AI (XAI) and Large Language Models (LLM): an impossible pairing?, AI Heroes 2024 conference, Turin (Italy)

Thanks!

More information: www.elastic.co

Contact information: enrico.zimuel (at) elastic.co

elastic