ChatGPT: Journey Through an LLM

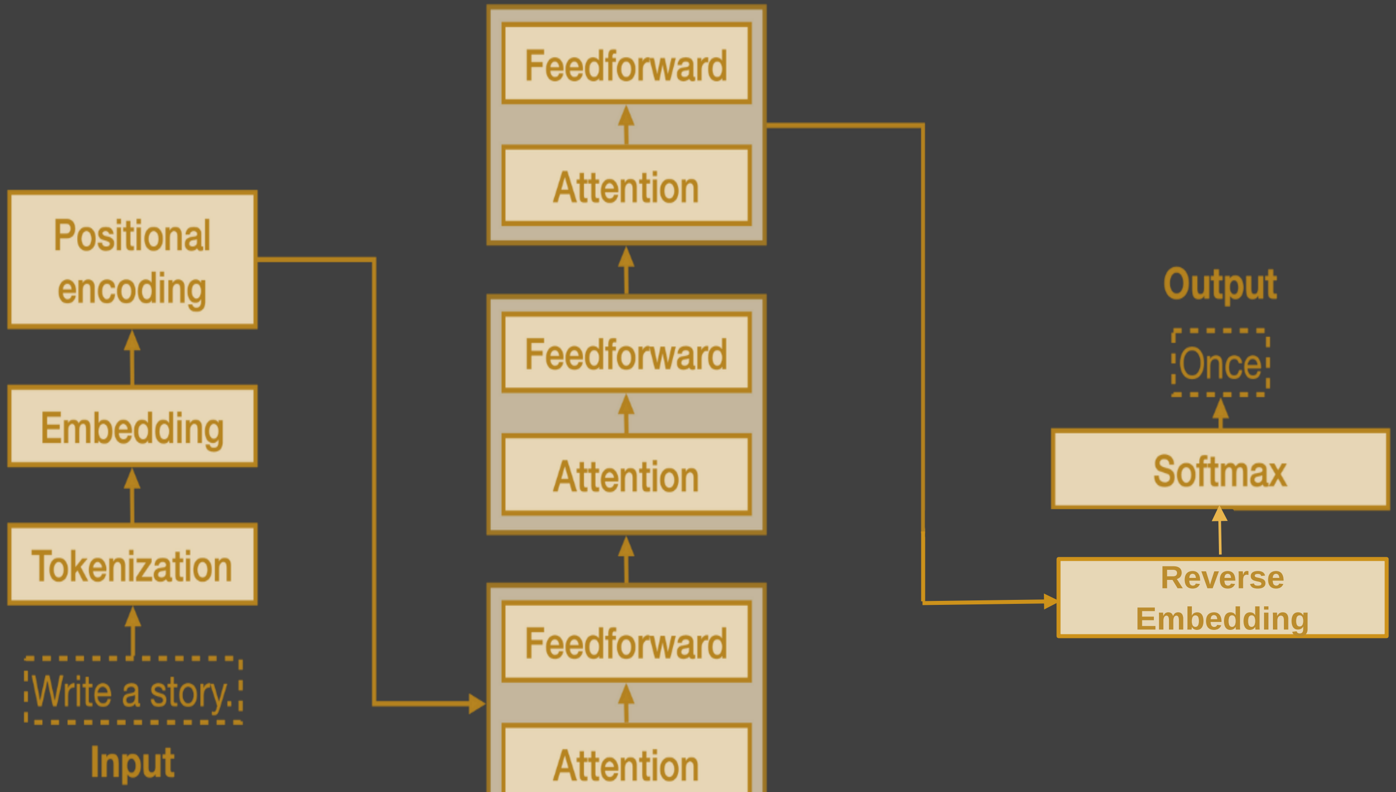
SAM SCOTT, MOHAWK COLLEGE, JUNE 2023

Reminder: The Core Task of ChatGPT

Given a text **prompt**, predict the natural language **token** (word) that comes next.

ChatGPT is a **Large Language Model** powered by a deep **Artificial Neural Network** architecture called a **Transformer**.

The Journey …



Positional encoding

Embedding

Tokenization

Write a story.

Input

Feedforward

Attention

Feedforward

Attention

Feedforward

Attention

Output

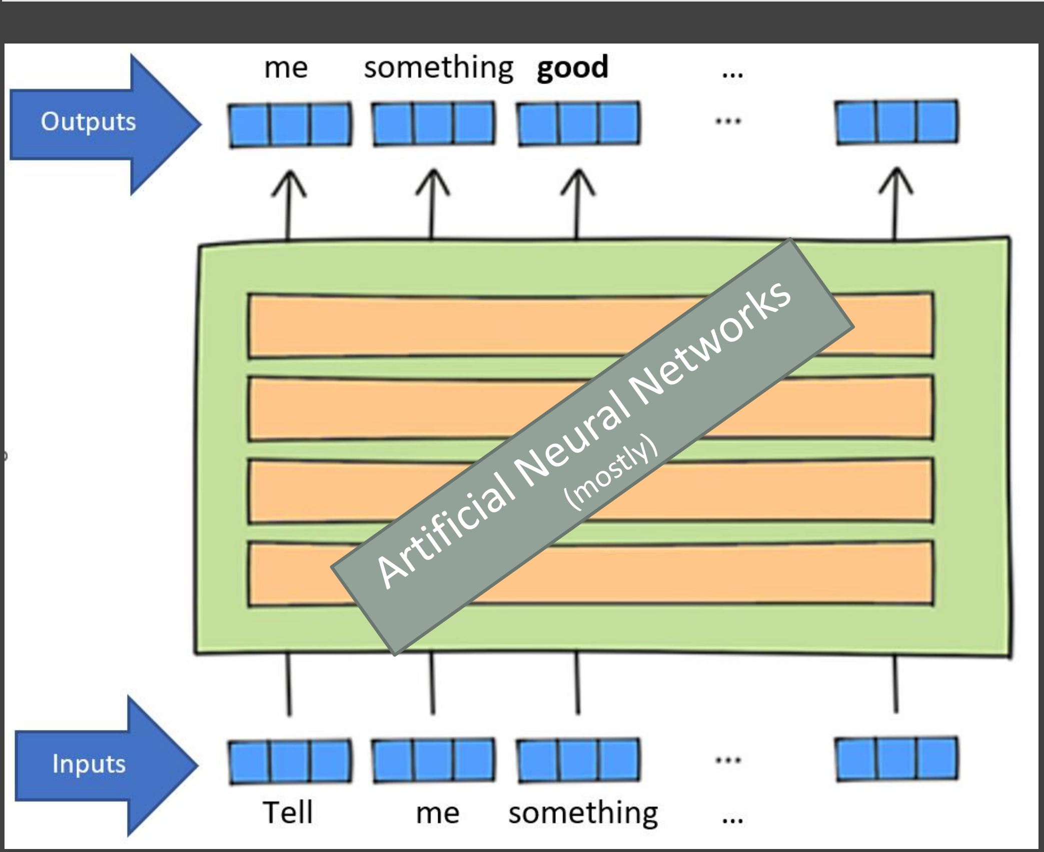
Once

Softmax

Reverse Embedding

Adapted from https://txt.cohere.com/what-are-transformer-models

A 1000-foot View



Adapted from https://www.lavivienpost.com/how-chatgpt-works-architecture-illustrated/

Tokens are converted to **vectors**
(lists of numbers)

All the tokens are fed in at the same time.

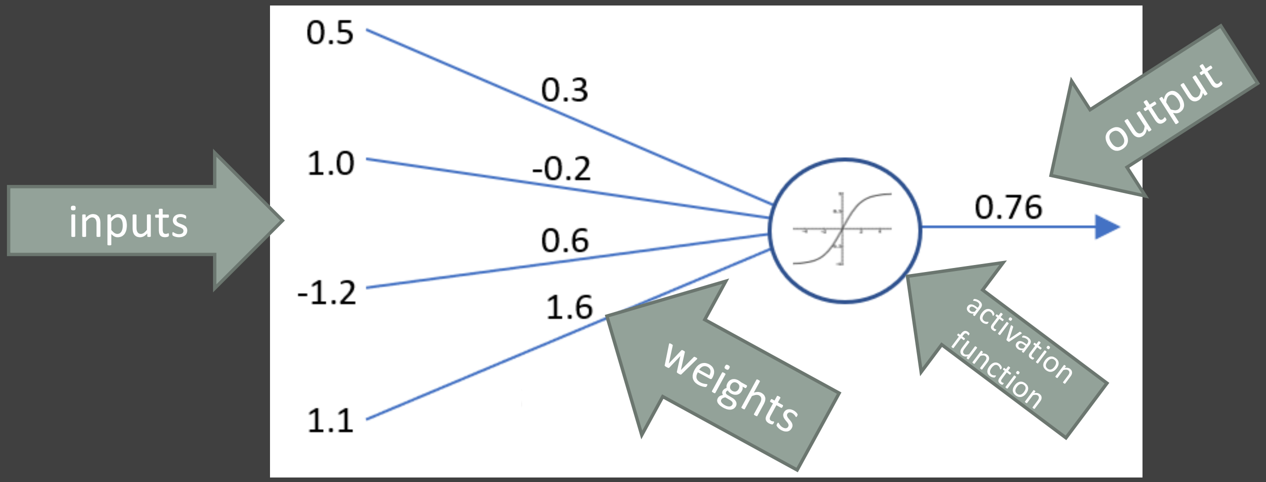
2048 x 50000 = 102 400 000 inputs & outputs

What is an Artificial Neural Network?

It’s a bunch of simple calculation devices (artificial neurons) all hooked up together.

It’s a way of turning one set of numbers into another set of numbers.

Here’s an artificial neuron



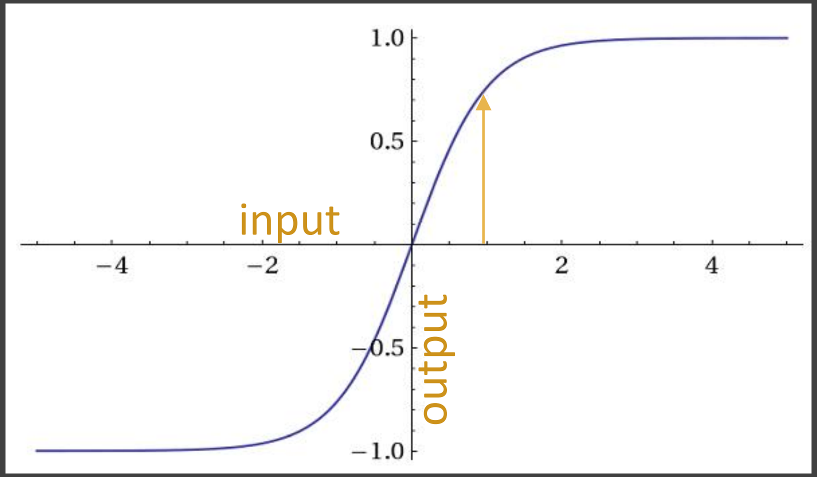
Input = 0.5 x 0.3 + 1.0 x (-0.2) + (-1.2) x 0.6 + 1.1 x 1.6 = 0.99

What is an Artificial Neural Network?

It’s a bunch of simple calculation devices (artificial neurons) all hooked up together.

It’s a way of turning one set of numbers into another set of numbers.

Here’s its Activation Function



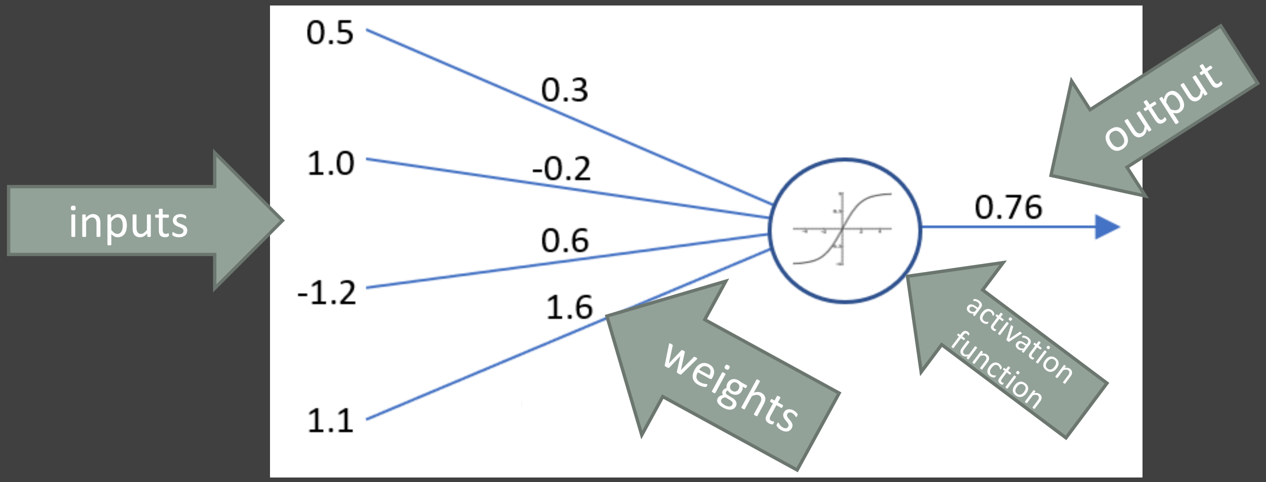
Output = tanh(0.99) = 0.76

What is an Artificial Neural Network?

It’s a bunch of simple calculation devices (artificial neurons) all hooked up together.

It’s a way of turning one set of numbers into another set of numbers.

Here’s an artificial neuron



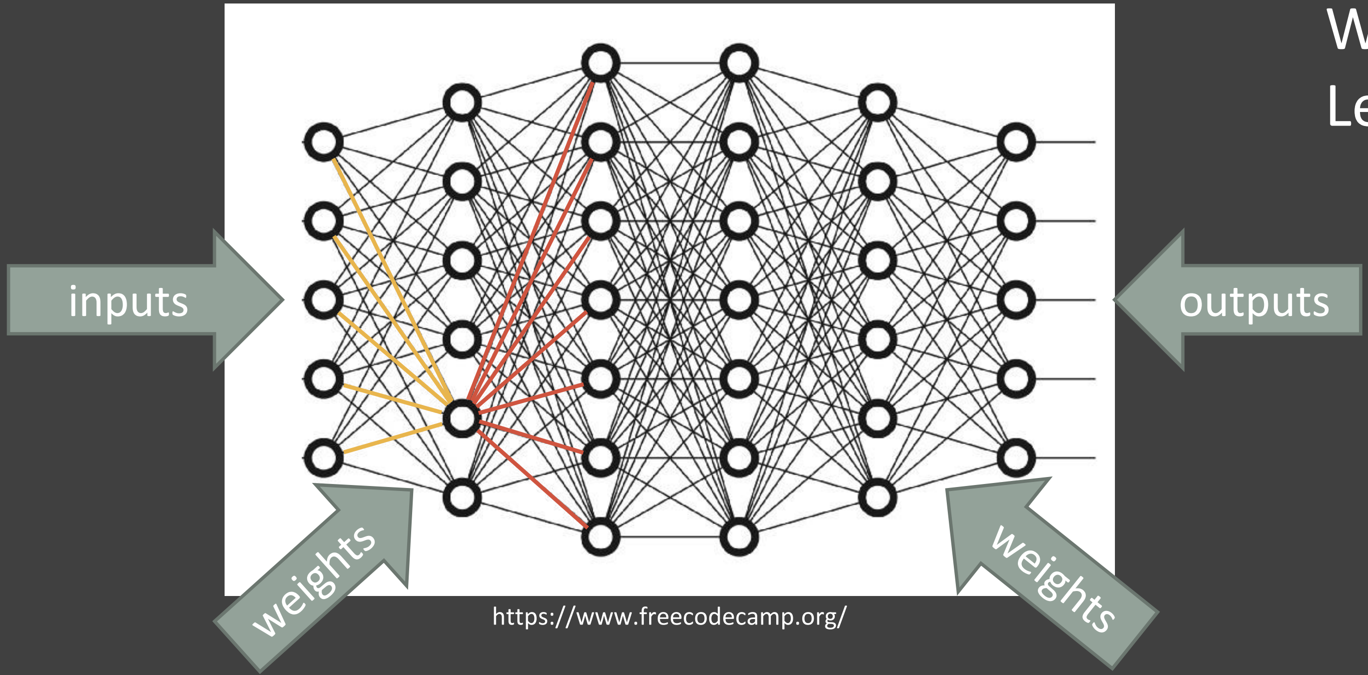
Input = 0.5 x 0.3 + 1.0 x (-0.2) + (-1.2) x 0.6
+ 1.1 x 1.6 = 0.99

What is an Artificial Neural Network?

It’s a bunch of simple calculation devices (artificial neurons) all hooked up together.

It’s a way of turning one set of numbers into another set of numbers.

Here’s an artificial neural network



https://www.freecodecamp.org/

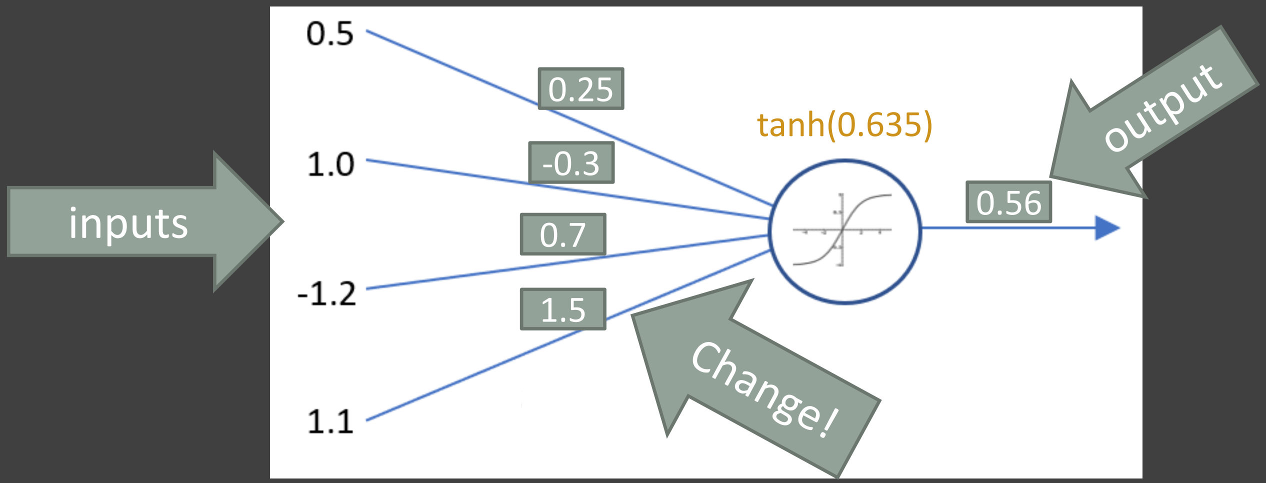
Weights = Parameters (175 Billion)
Learning = Adjusting Parameters

What is an Artificial Neural Network?

It’s a bunch of simple calculation devices (artificial neurons) all hooked up together.

It’s a way of turning one set of numbers into another set of numbers.

Training an artificial neuron



Input = 0.5 x 0.25 + 1.0 x (-0.3) + (-1.2) x 0.7 + 1.1 x 1.5 = 0.635

I got 0.76 but I wanted 0.

So, change the weights a little bit.

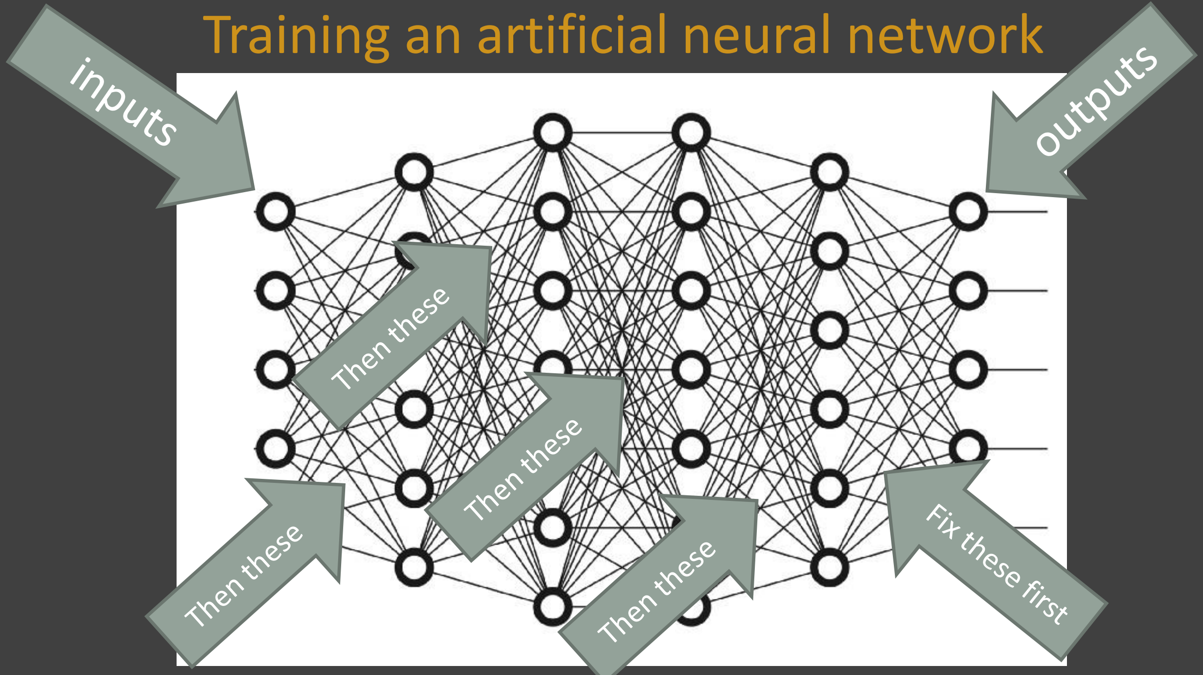
(Note the 5th bias weight is not shown)

What is an Artificial Neural Network?

It’s a bunch of simple calculation devices (artificial neurons) all hooked up together.

It’s a way of turning one set of numbers into another set of numbers.

Training an artificial neural network



https://www.freecodecamp.org/

First , fix the output layer weights

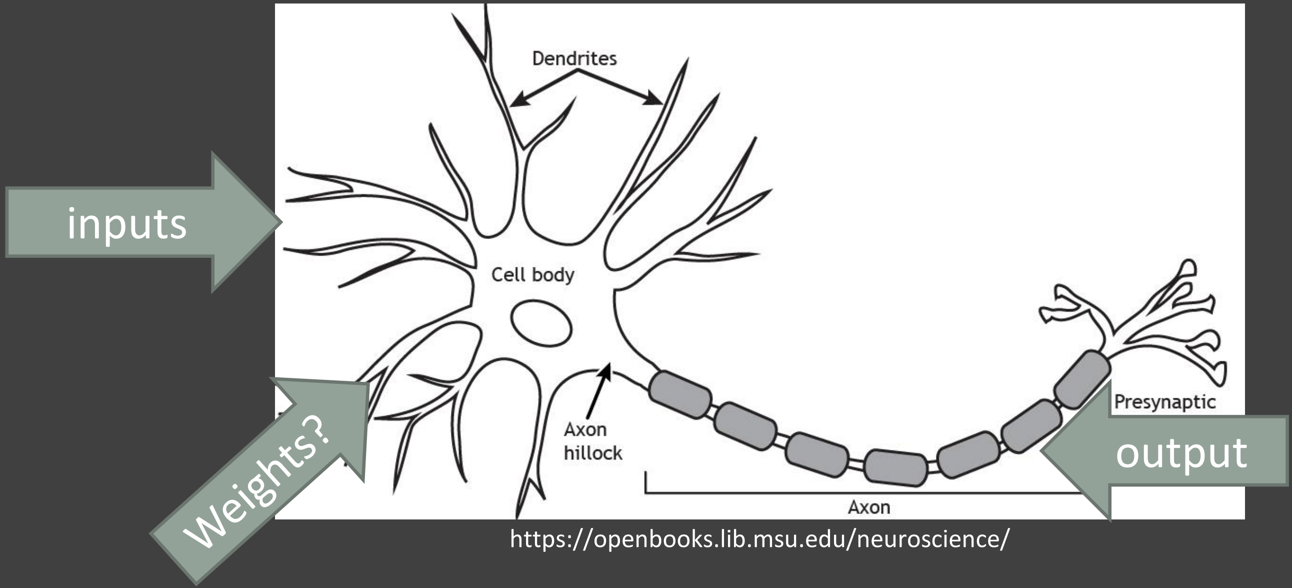
“Backpropagate” to fix the rest

What is an Artificial Neural Network?

It’s a bunch of simple calculation devices (artificial neurons) all hooked up together.

It’s a way of turning one set of numbers into another set of numbers.

Here’s a real neuron (artist’s depiction)



inputs

Weights?

output

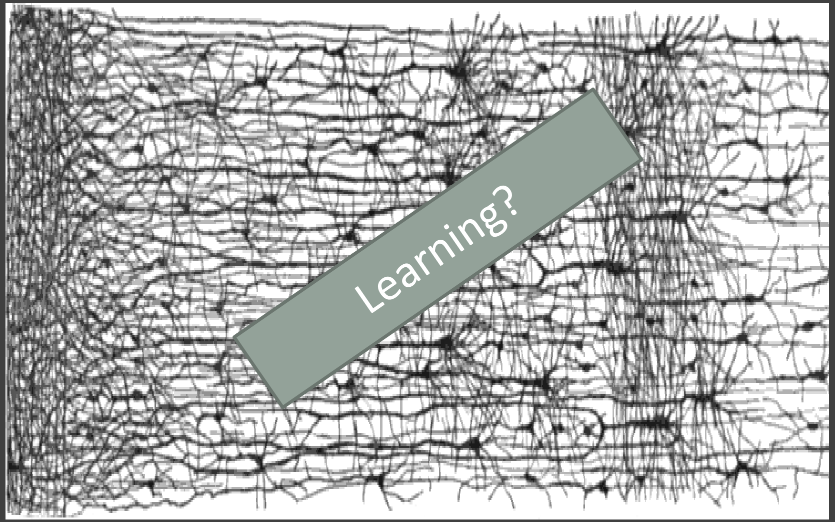
https://openbooks.lib.msu.edu/neuroscience/

What is an Artificial Neural Network?

It’s a bunch of simple calculation devices (artificial neurons) all hooked up together.

It’s a way of turning one set of numbers into another set of numbers.

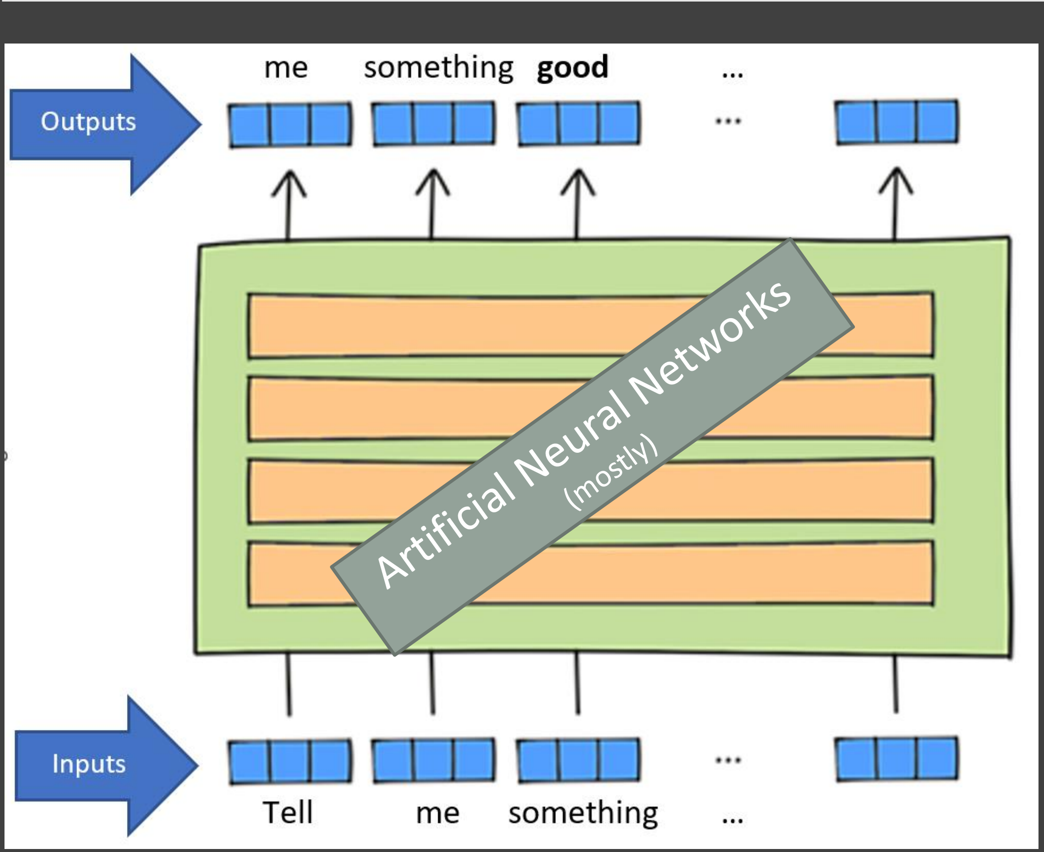
Here’s a real neural network (detail)



Learning?

https://www.oreilly.com/

A 1000-foot View of a GPT Model



Adapted from https://www.lavivienpost.com/how-chatgpt-works-architecture-illustrated/

Tokens are converted to **vectors**
(lists of numbers)

All the tokens are fed in at the same time.

2048 x 50000 = 102 400 000 inputs & outputs

Starting the Journey

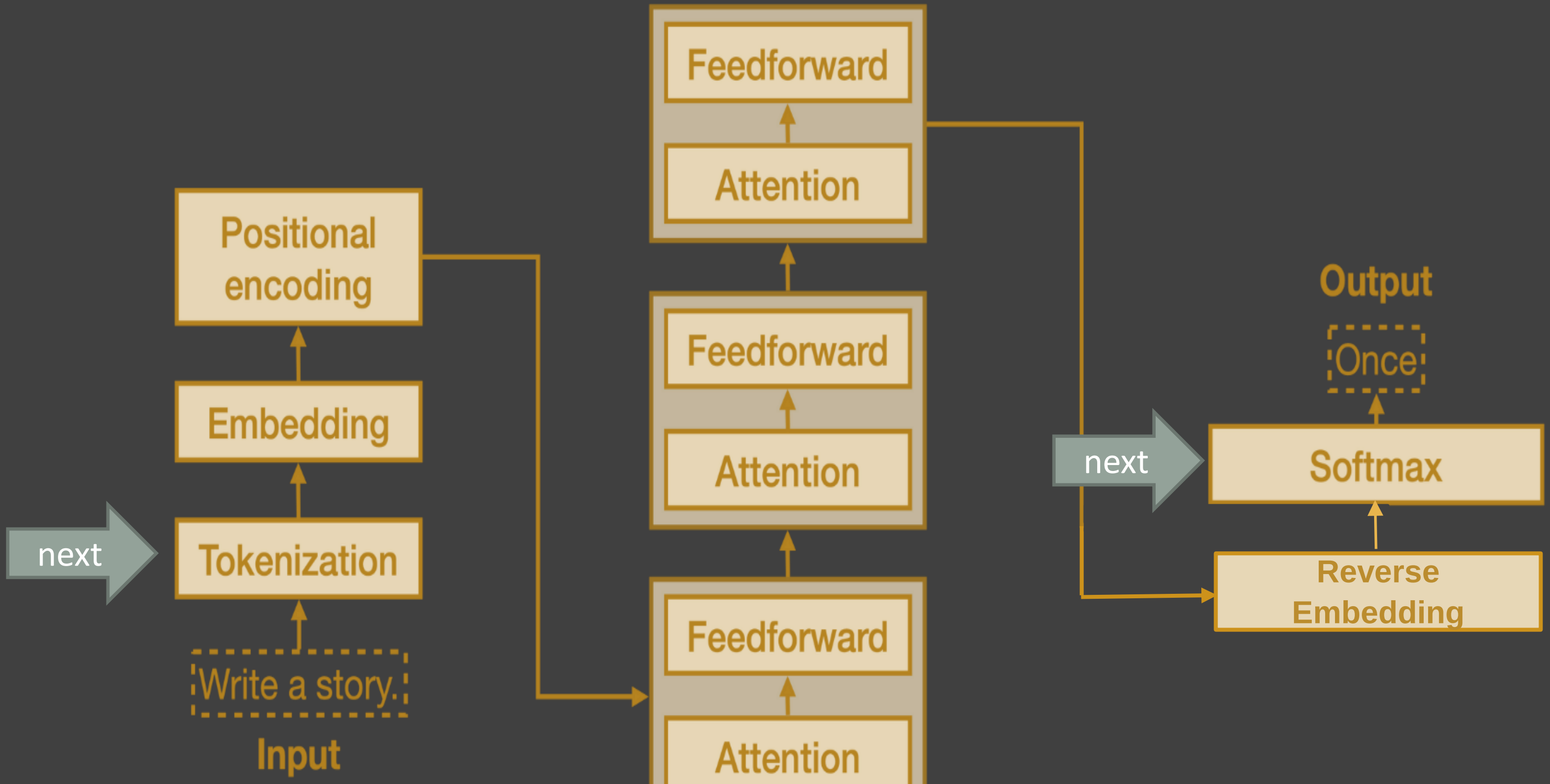
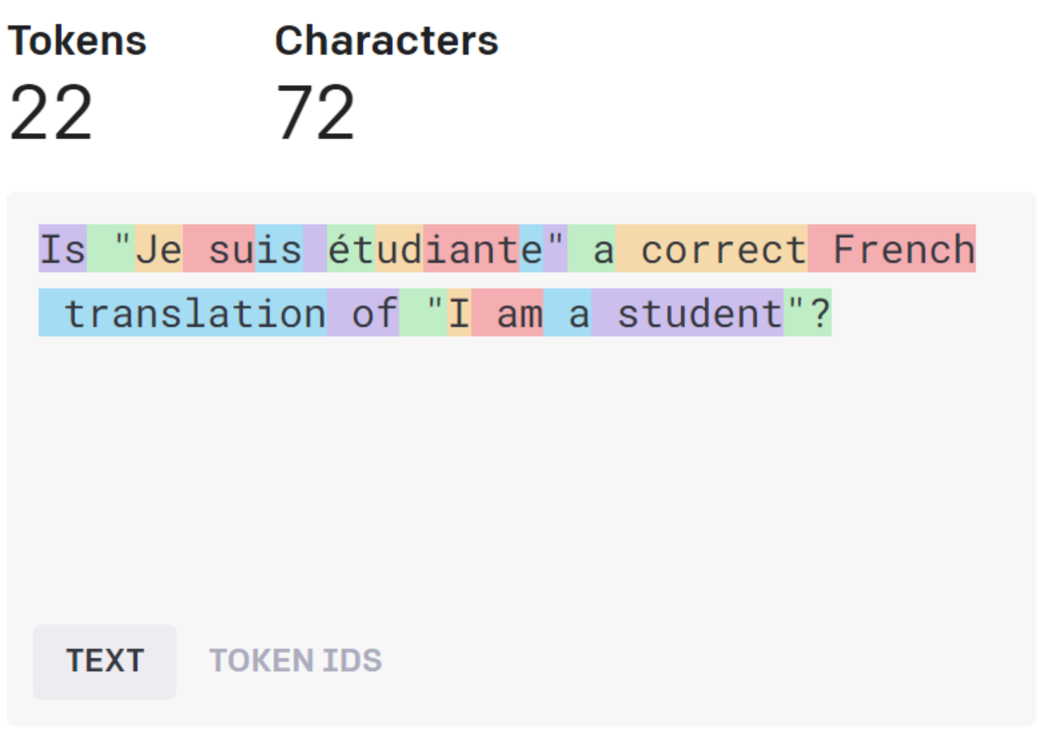


Diagram adapted from https://txt.cohere.com/what-are-transformer-models

Tokenization: Text to Tokens



Tokens
22

Characters
72

Is "Je suis étudiante" a correct French translation of "I am a student"?

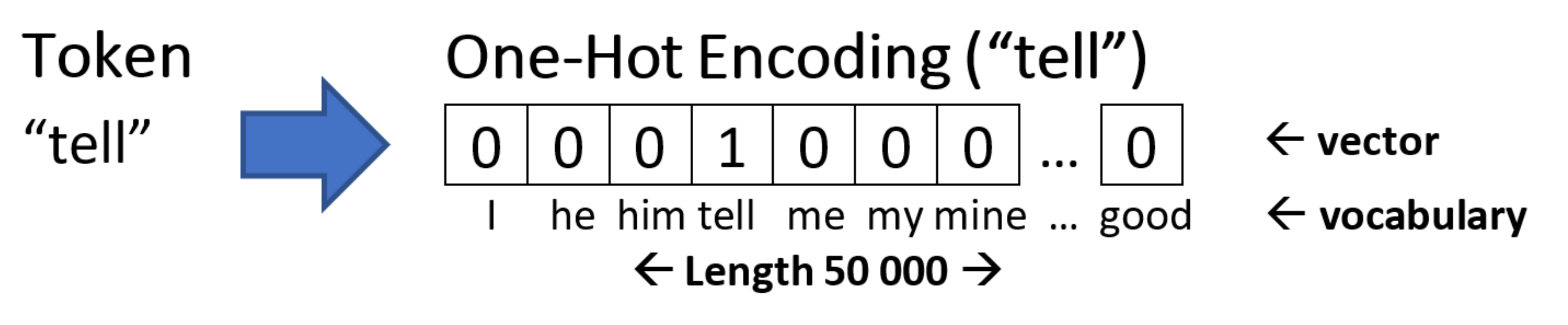
https://platform.openai.com/tokenizer

Token: word, part of word, punctuation, etc…

Vocabulary: set of tokens the LLM “knows”

Tokenization: splitting up input into tokens

Tokens to One-hot Encodings



Token

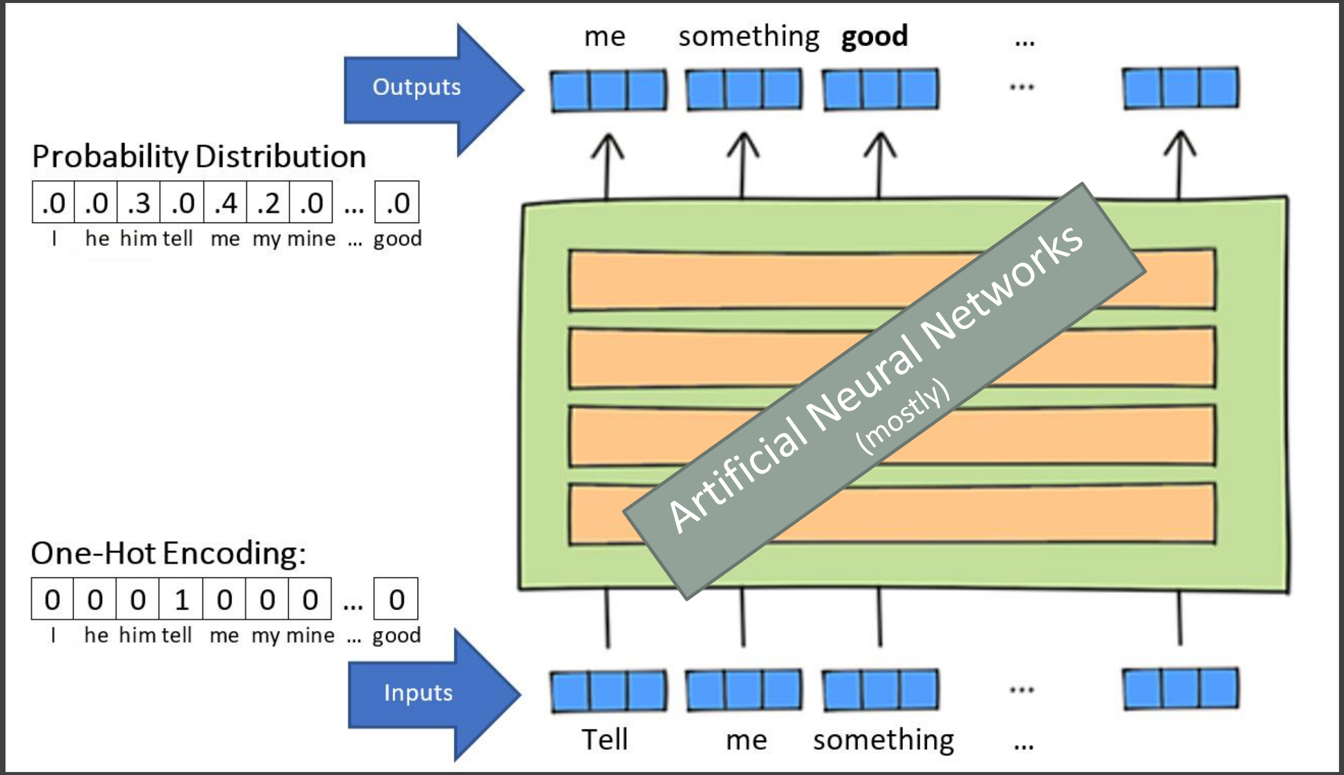
“tell”

One-Hot Encoding (“tell”)

0	0	0	1	0	0	0	…	0	← **vector**
I	he	him	tell	me	my	mine	…	good	← **vocabulary**

← Length 50 000 →

Back to the 1000-foot View



Probability Distribution

I	he	him	tell	me	my	mine	...	good
.0	.0	.3	.0	.4	.2	.0	...	.0

One-Hot Encoding:

I	he	him	tell	me	my	mine	...	good
0	0	0	1	0	0	0	...	0

Adapted from https://www.lavivienpost.com/how-chatgpt-works-architecture-illustrated/

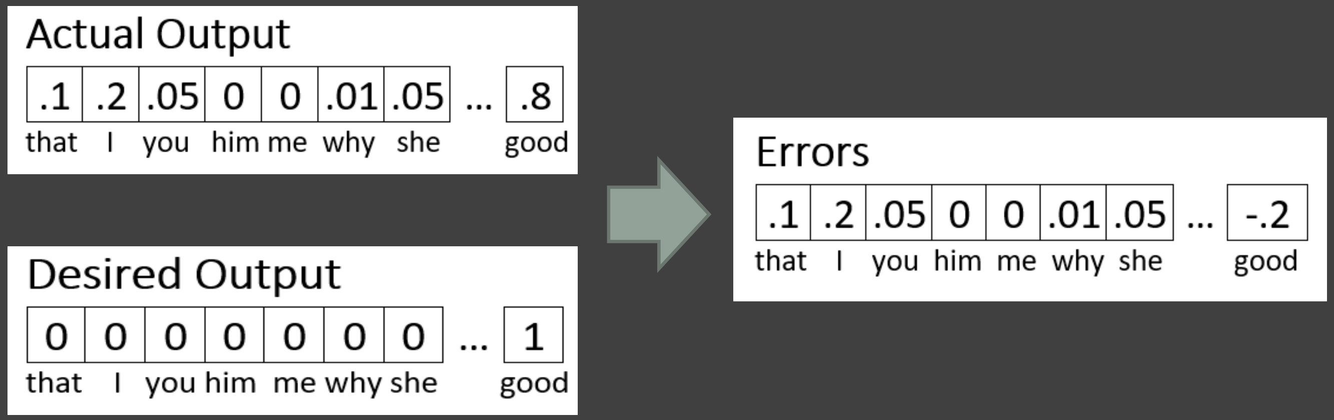
ChatGPT predicts a next token for *every* token in the prompt.

But the process is messy, so we get a probability distribution.

Usually, we're only interested in the *final* token prediction.

A Note on Training

In training, we can compare the probability distribution we get for each word to the one-hot encoding for the *actual* next word.



Actual Output

.1	.2	.05	0	0	.01	.05	…	.8
that	I	you	him	me	why	she		good

Desired Output

0	0	0	0	0	0	0	…	1
that	I	you	him	me	why	she		good

Errors

.1	.2	.05	0	0	.01	.05	…	-.2
that	I	you	him	me	why	she		good

This “Error Signal” is then used to tune the weights (parameters) of the artificial neurons that produced the output.

The Journey So Far

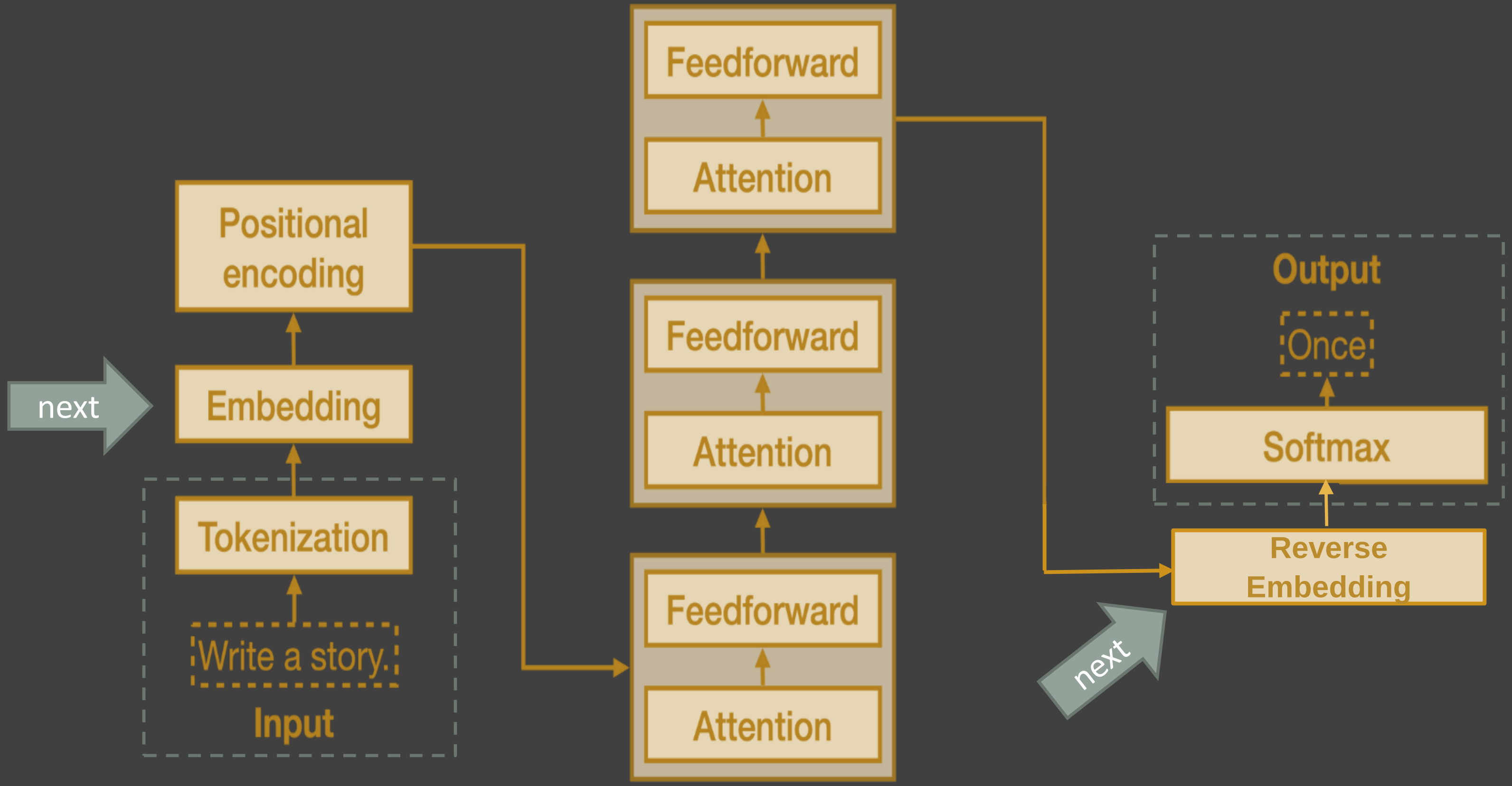
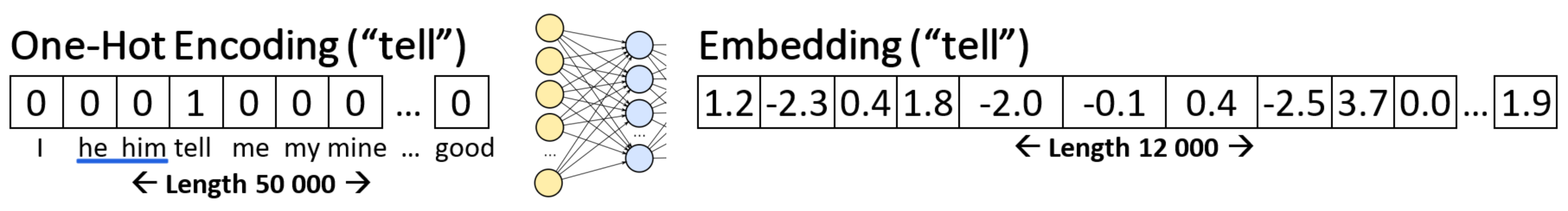


Diagram adapted from https://txt.cohere.com/what-are-transformer-models

One-hot Encodings → Embeddings

Embedding: “small”, information-rich token vectors.



One-Hot Encoding (“tell”)

0	0	0	1	0	0	0	…	0
I	he	him	tell	me	my	mine	…	good

← Length 50 000 →

Embedding (“tell”)

1.2	-2.3	0.4	1.8	-2.0	-0.1	0.4	-2.5	3.7	0.0	…	1.9

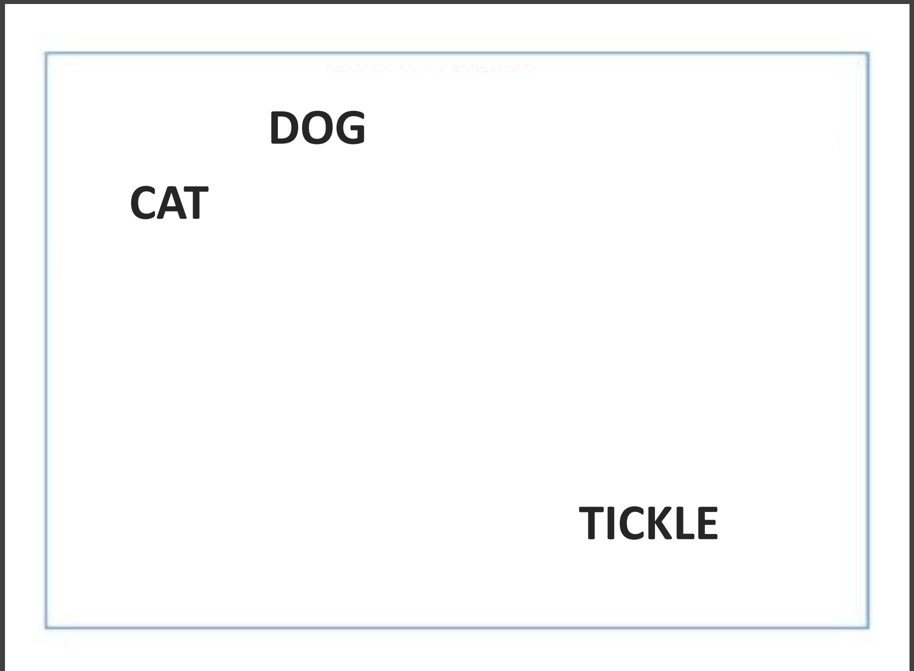
← Length 12 000 →

Embeddings are computed by an Artificial Neural Network.

What’s in an Embedding?

Embeddings encode distribution patterns for words.

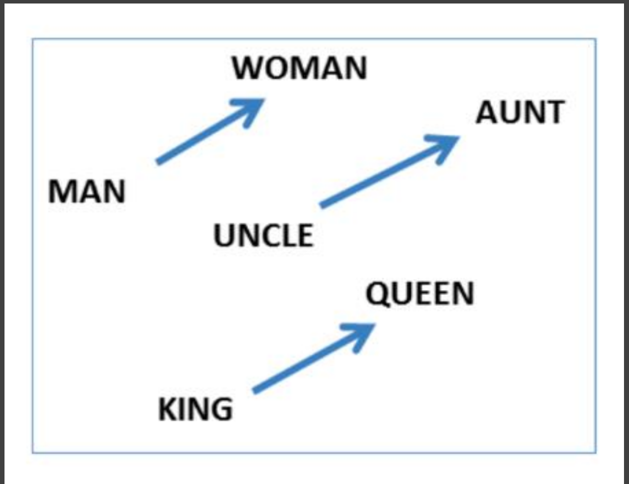
E.g., the embeddings for “cat” and “dog” will be more similar than “cat” and “tickle”.

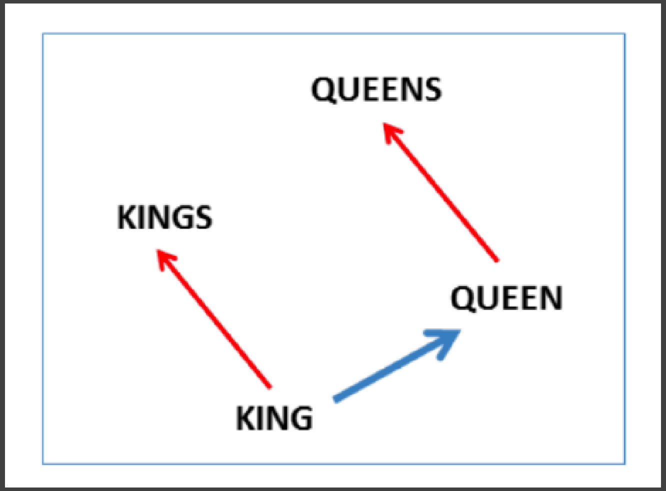


What’s in an Embedding?

Embeddings are very hard to interpret.

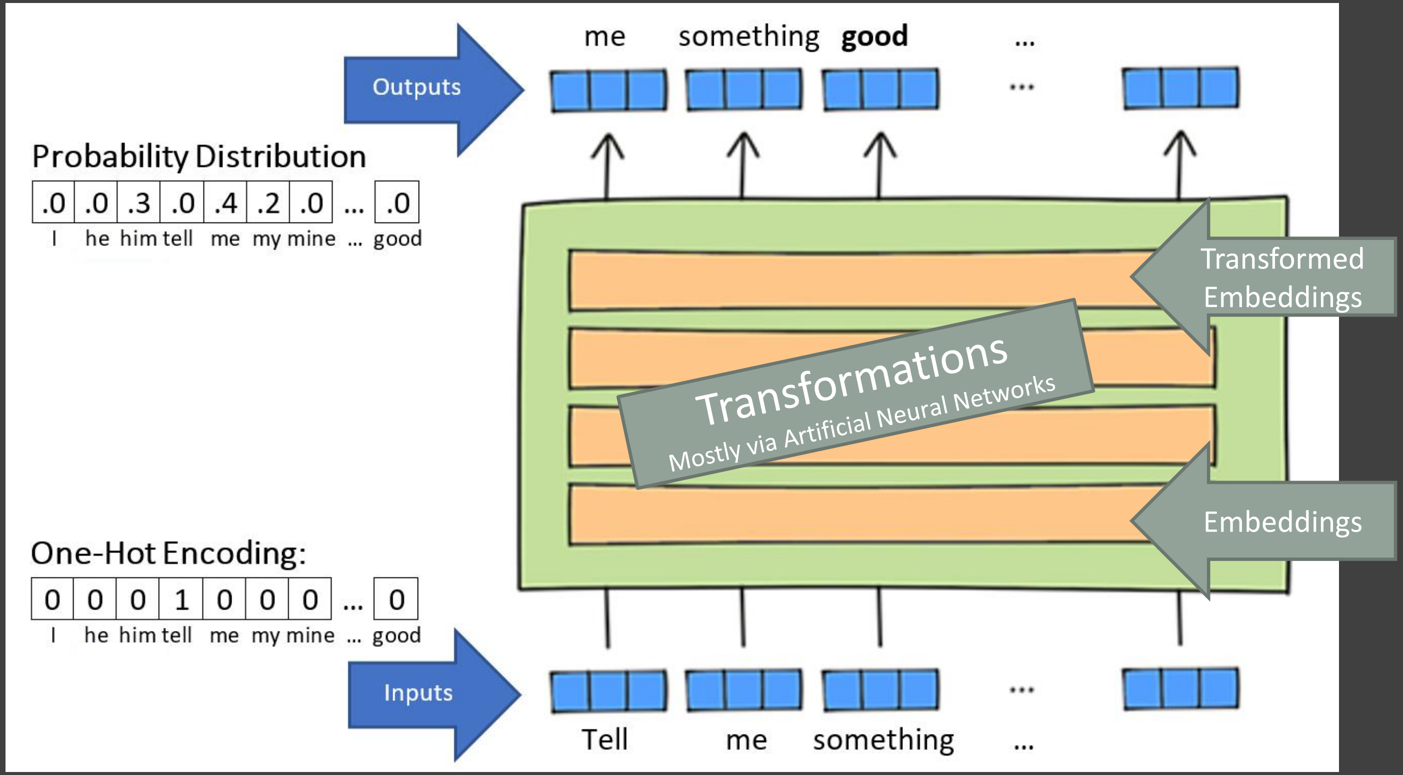
But researchers have shown consistent relationships between embeddings.





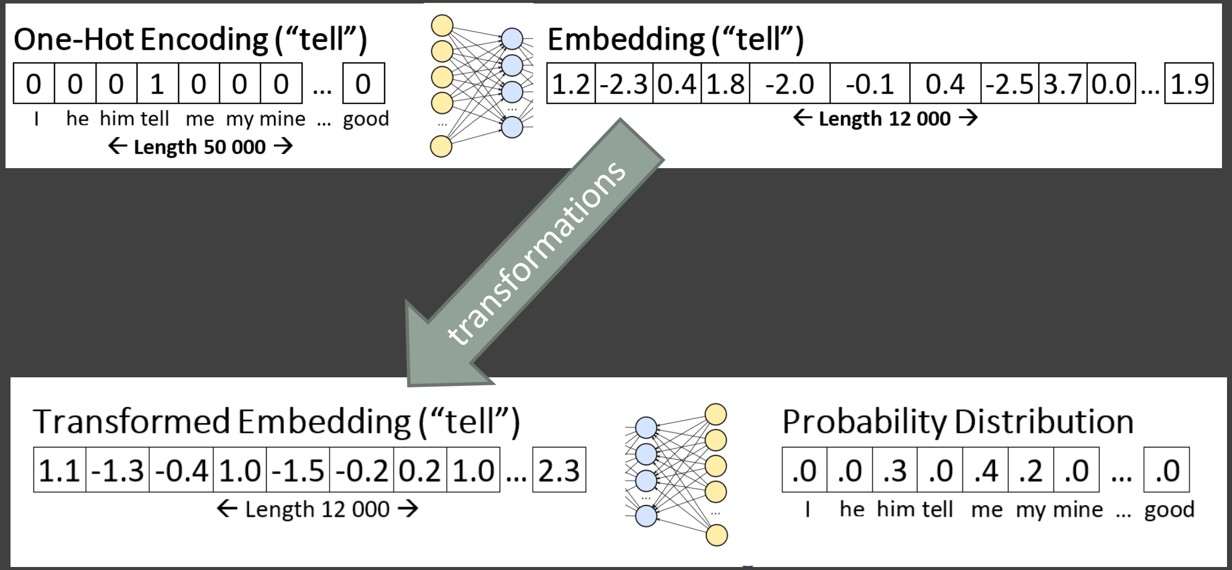
Embeddings are what allows the transformer to generalize.

Embeddings get Transformed



Adapted from https://www.lavivienpost.com/how-chatgpt-works-architecture-illustrated/

Transformed Embeddings are Decoded



One-Hot Encoding (“tell”)

I	he	him	tell	me	my	mine	...	good
0	0	0	1	0	0	0	...	0

← Length 50 000 →

Embedding (“tell”)

1.2	-2.3	0.4	1.8	-2.0	-0.1	0.4	-2.5	3.7	0.0	...	1.9

← Length 12 000 →

transformations

Transformed Embedding (“tell”)

1.1	-1.3	-0.4	1.0	-1.5	-0.2	0.2	1.0	...	2.3

← Length 12 000 →

Probability Distribution

I	he	him	tell	me	my	mine	...	good
.0	.0	.3	.0	.4	.2	.0	...	.0

The Journey So Far

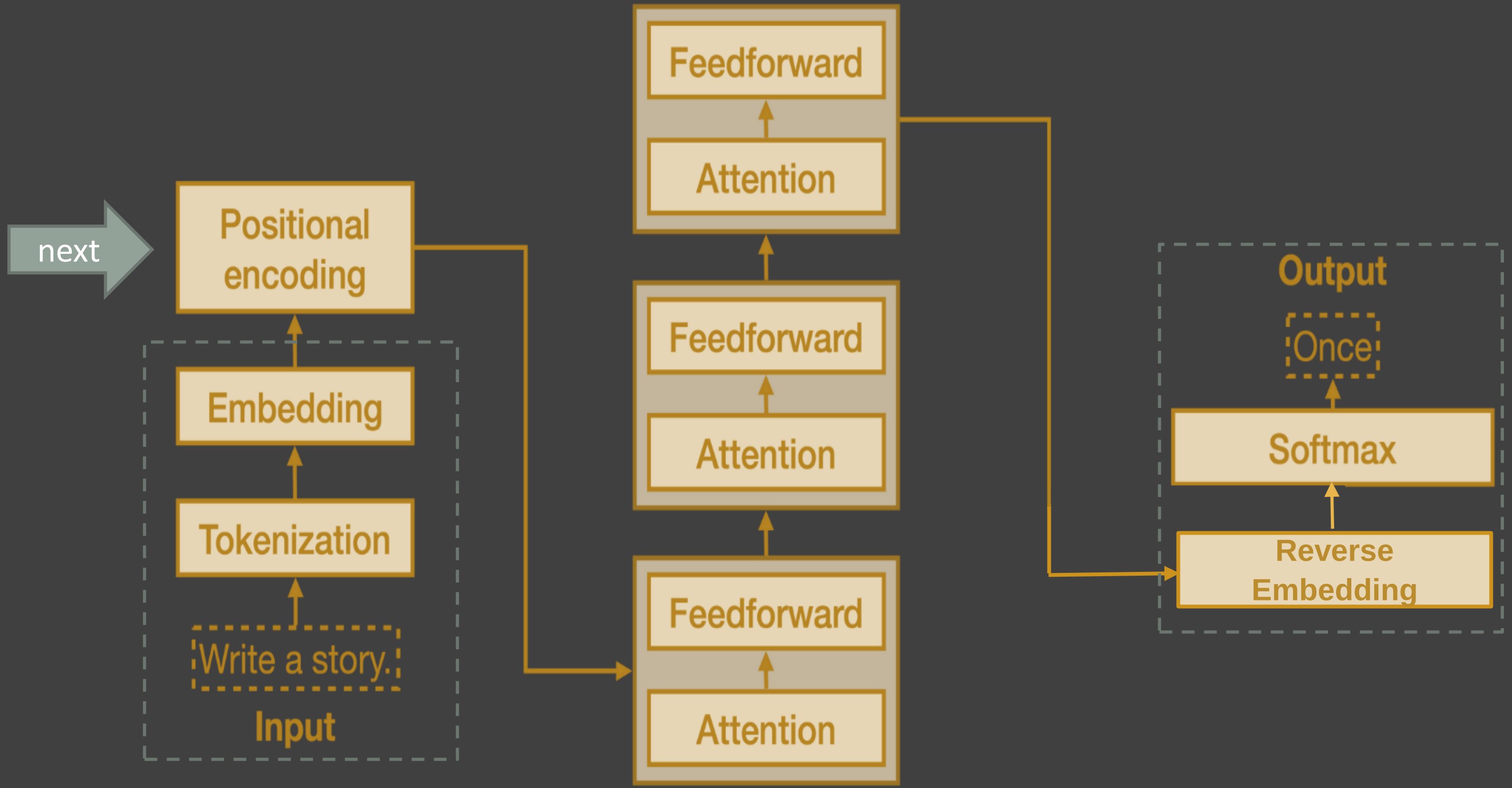


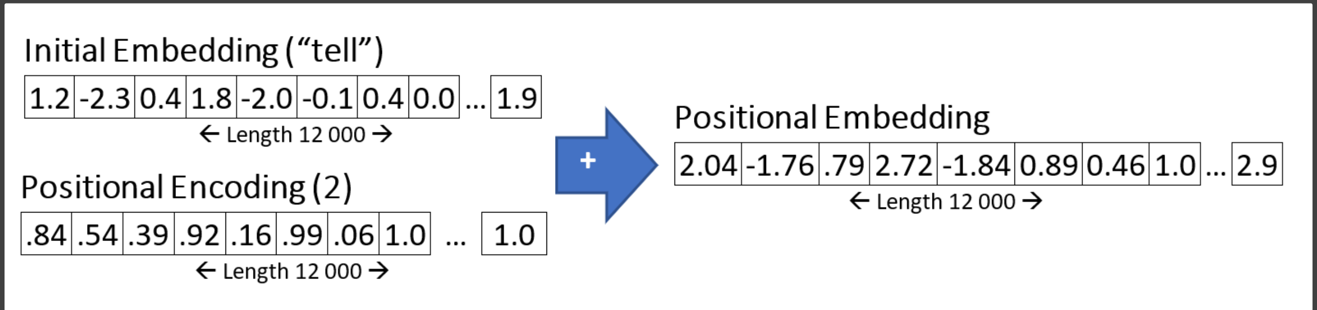
Diagram adapted from https://txt.cohere.com/what-are-transformer-models

Positional Encoding

Word position is important!

“Please **tell** the poker players.” vs “Every poker player has a **tell**.”

Every embedding has a **positional encoding** added to it.
Positional encoding is different for each position.



Initial Embedding (“tell”)

1.2	-2.3	0.4	1.8	-2.0	-0.1	0.4	0.0	…	1.9

← Length 12 000 →

Positional Encoding (2)

.84	.54	.39	.92	.16	.99	.06	1.0	…	1.0

← Length 12 000 →

Positional Embedding

2.04	-1.76	.79	2.72	-1.84	0.89	0.46	1.0	…	2.9

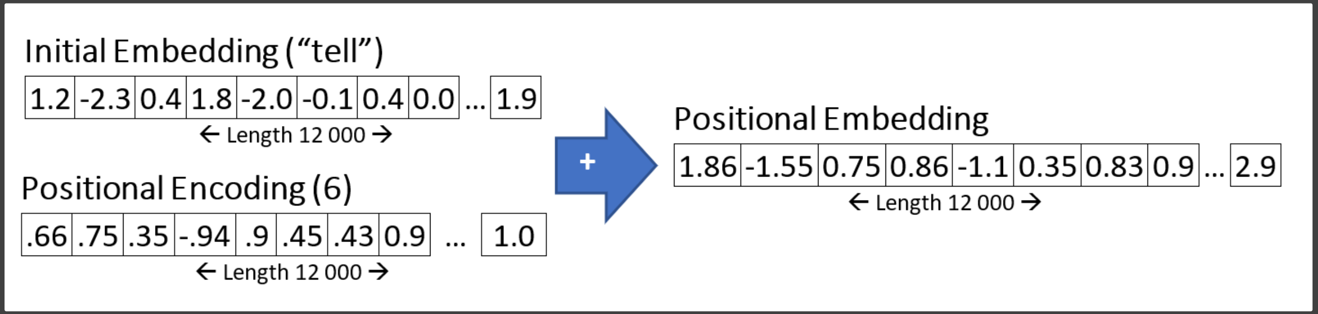
← Length 12 000 →

Positional Encoding

Word position is important!

“Please **tell** the poker players.” vs “Every poker player has a **tell**.”

Every embedding has a **positional encoding** added to it.
Positional encoding is different for each position.



Initial Embedding (“tell”)

1.2	-2.3	0.4	1.8	-2.0	-0.1	0.4	0.0	…	1.9

← Length 12 000 →

Positional Encoding (6)

.66	.75	.35	-.94	.9	.45	.43	0.9	…	1.0

← Length 12 000 →

Positional Embedding

1.86	-1.55	0.75	0.86	-1.1	0.35	0.83	0.9	…	2.9

← Length 12 000 →

Why Does this Work?

Somebody had a hunch and it worked out.

The Journey So Far

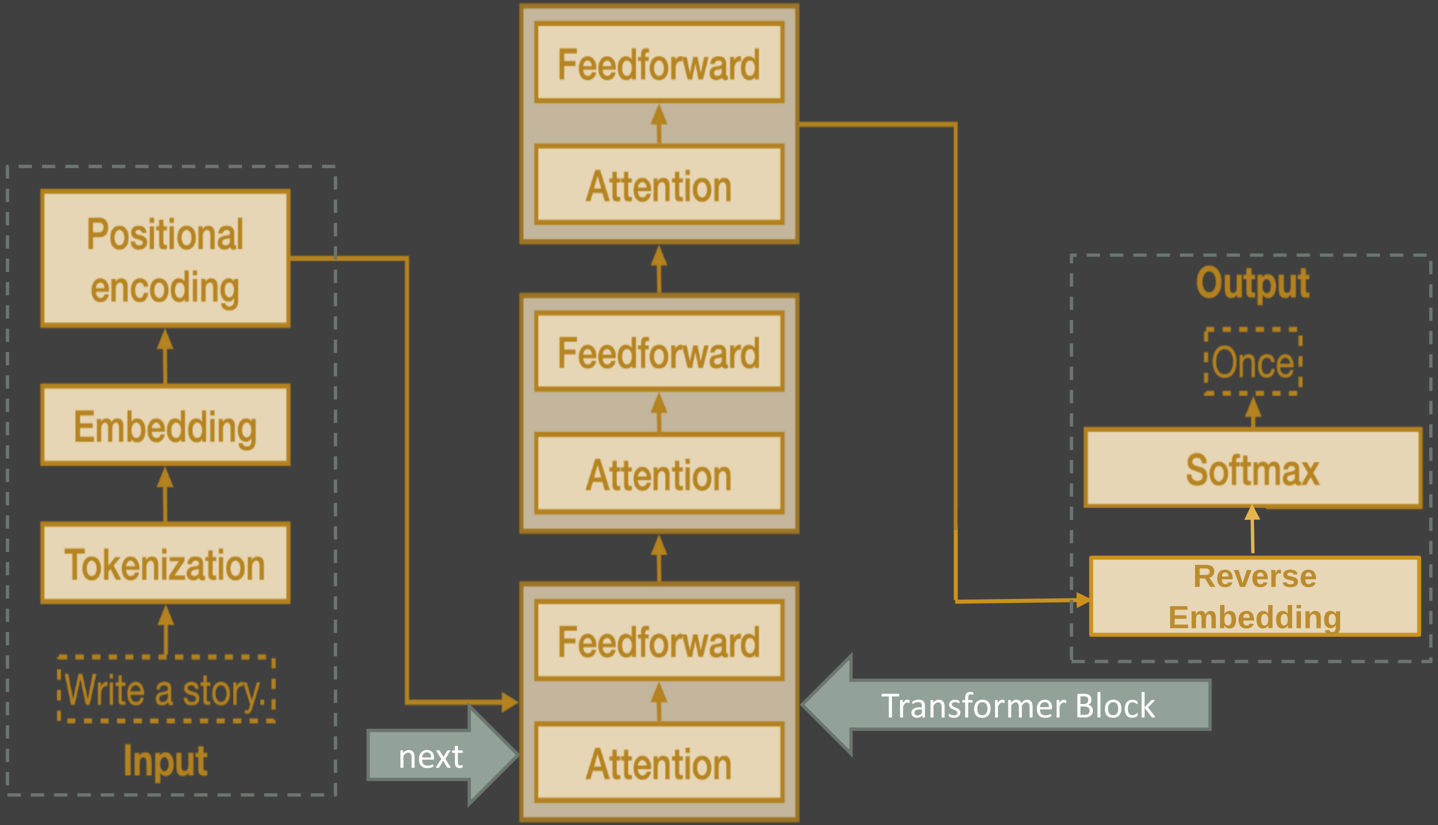


Diagram adapted from https://txt.cohere.com/what-are-transformer-models

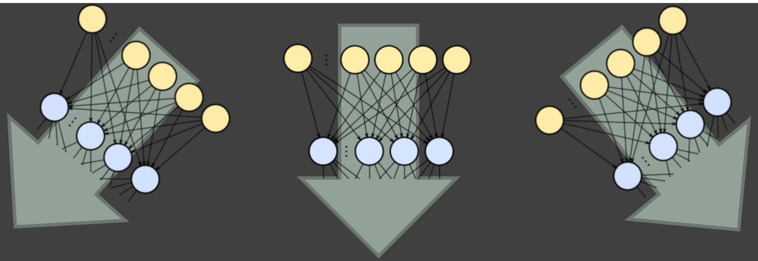
The Attention Mechanism

Positional Embedding

1.86	-1.55	0.75	0.86	-1.1	0.35	0.83	0.9	…	2.9

← Length 12 000 →

96 x



Query

0.2	-1.5	3.4	…	0.9

← Length 128 →

Key

-0.2	1.3	0.1	…	1.1

← Length 128 →

Value

3.1	1.5	2.0	…	-0.1

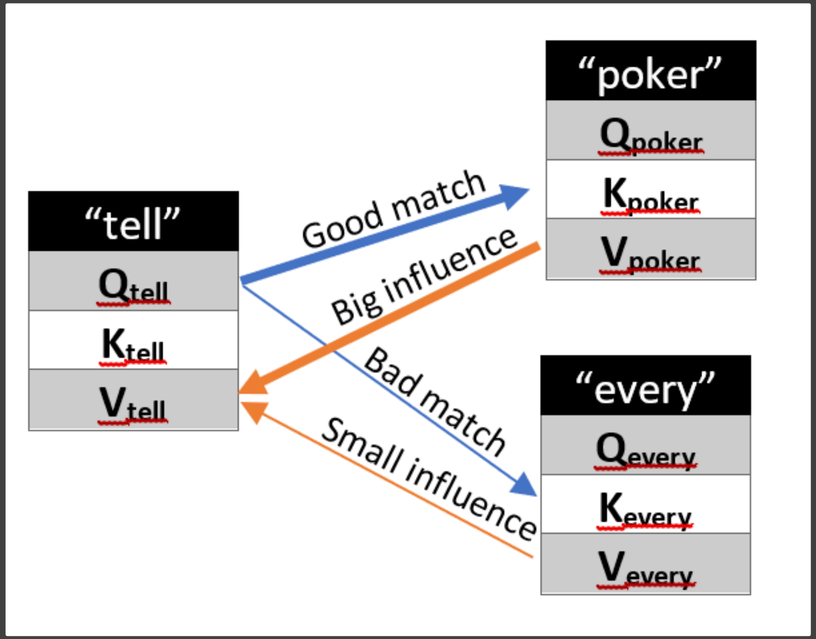
← Length 128 →

Relevant Context Words Combined

Tokens: Every poker player has a tell .

Query is compared to each preceding token's **Key**.

The **Values** of good matches are combined.



X 96

Final Values are Concatenated

Value$_1$, Value$_2$, Value$_3$, …, Value$_{96}$ = New Embedding Vector!

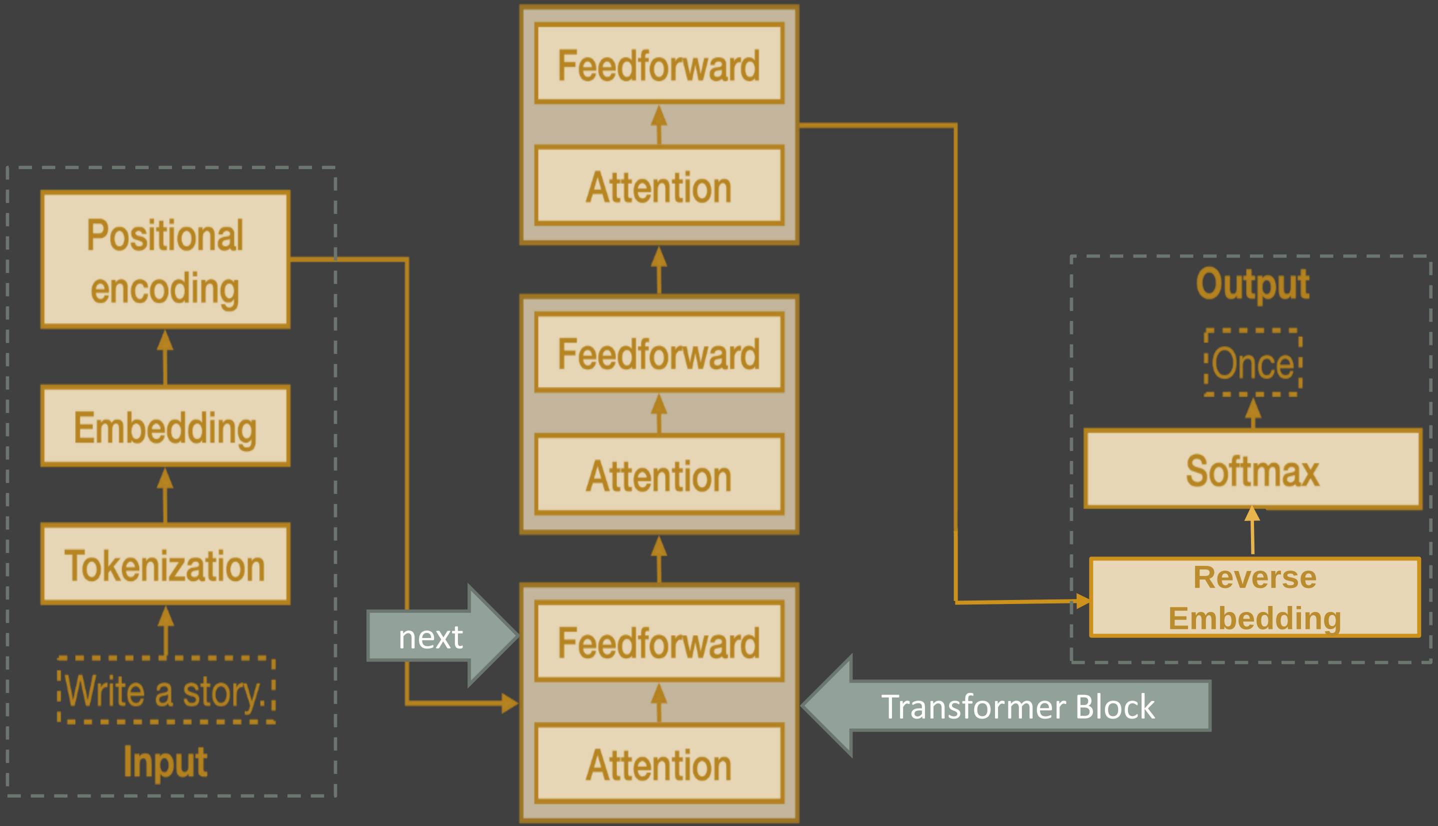
“Tell” → 96 combined **value** vectors

→ **value** vector length = 128

→ 96 x 128 ≅ 12 000

Now “tell” has a new, contextualized, embedding

The Journey So Far



Positional encoding

Embedding

Tokenization

Write a story.

Input

next

Feedforward

Attention

Feedforward

Attention

Feedforward

Attention

Transformer Block

Output

Once

Softmax

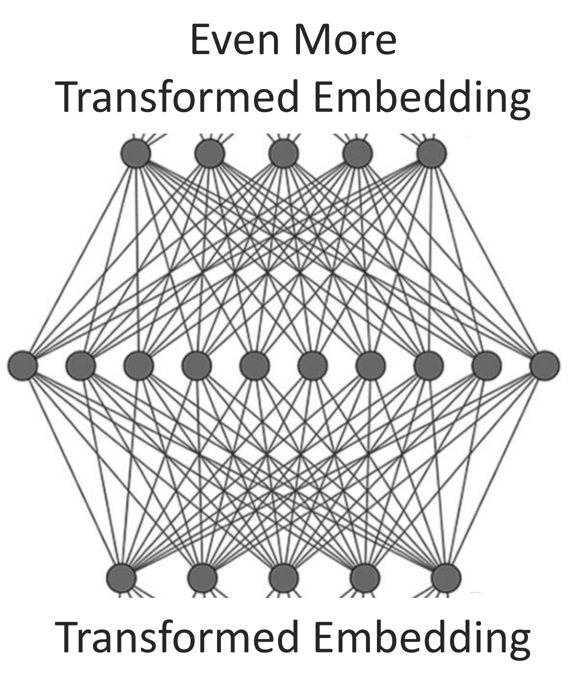
Reverse Embedding

Diagram adapted from https://txt.cohere.com/what-are-transformer-models

The Feedforward Layer

Feedforward = Neural Network Layer

Transforms the embeddings again but doesn’t change their length.



The Transformer Blocks are Repeated

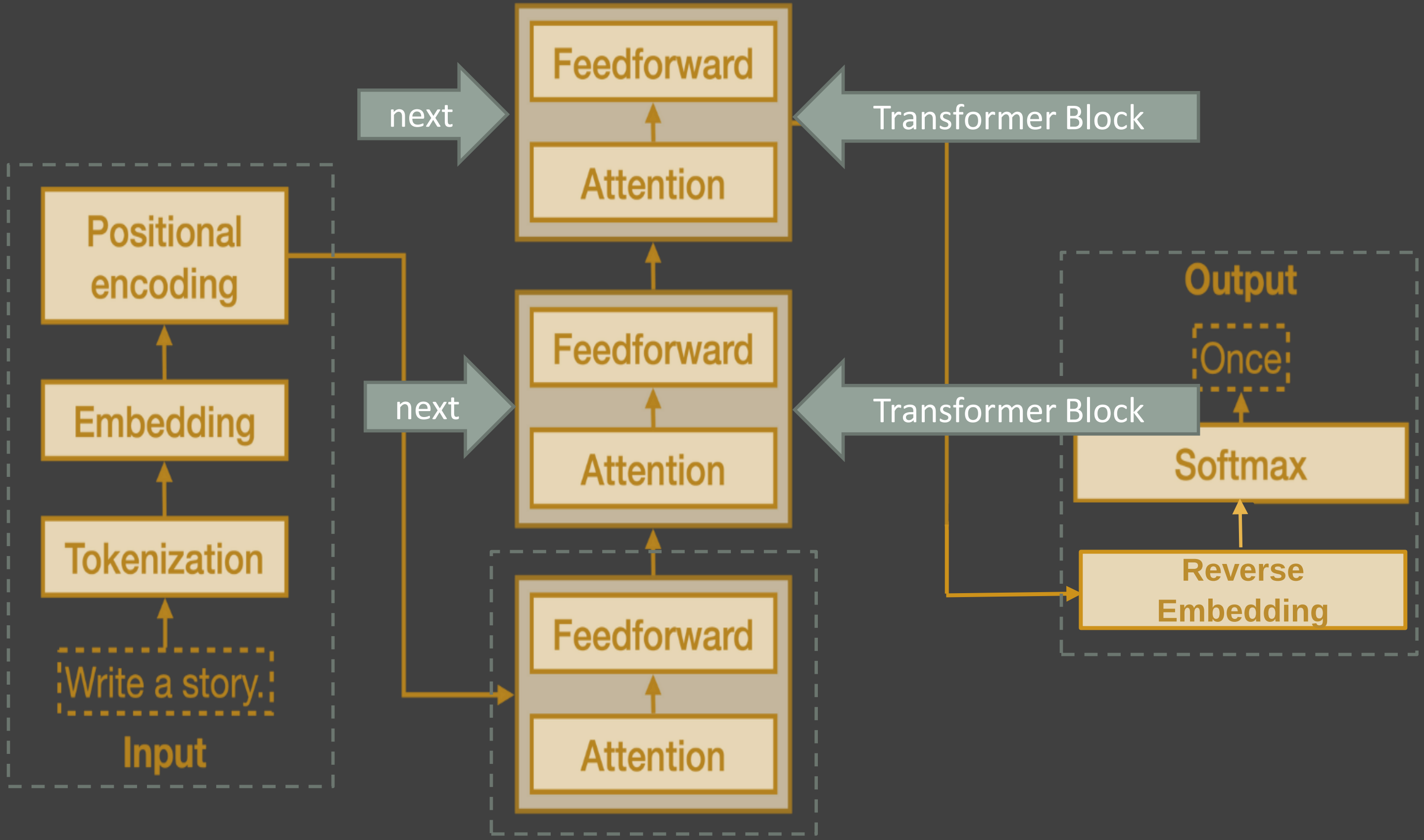


Diagram adapted from https://txt.cohere.com/what-are-transformer-models

And That’s the Whole Journey!

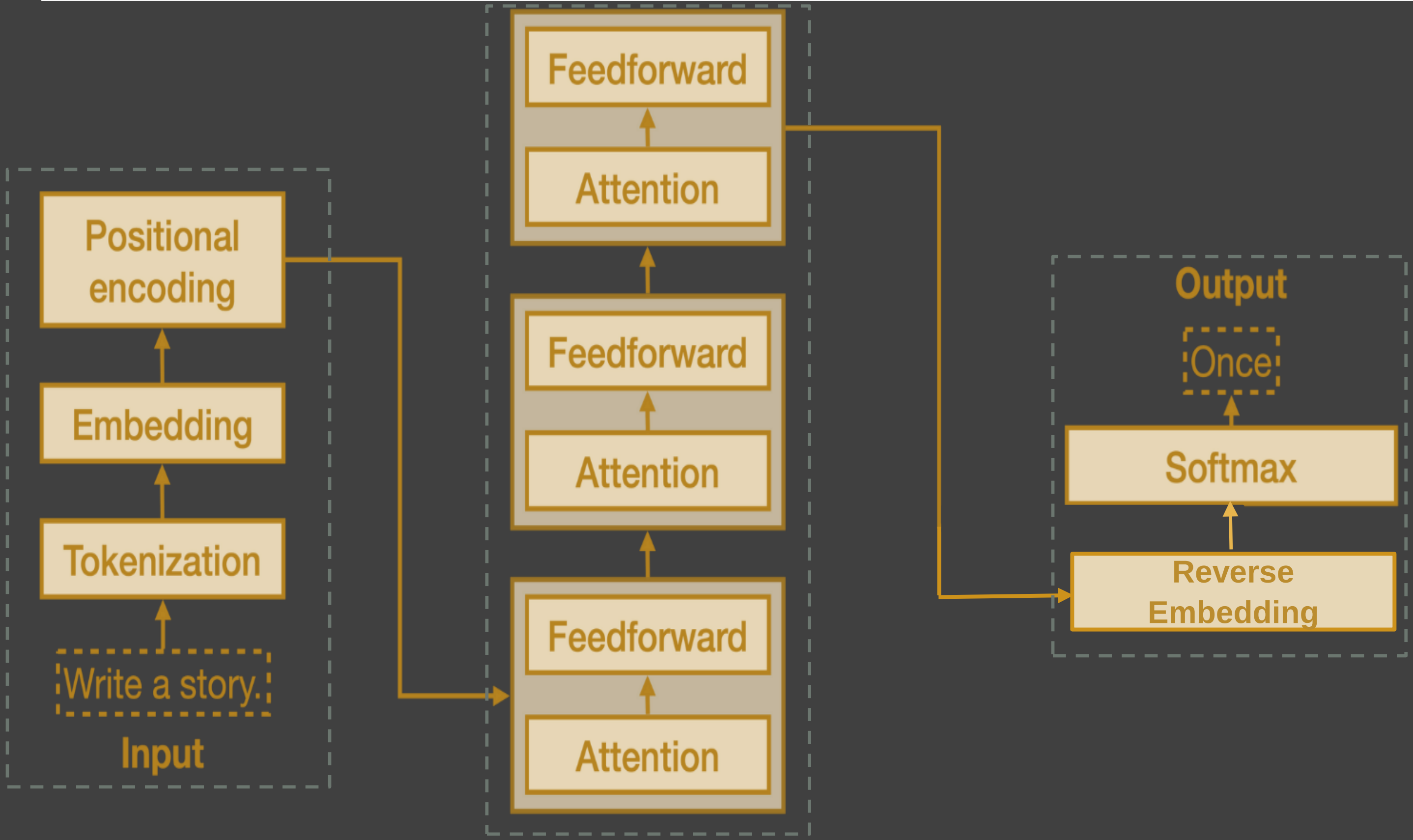
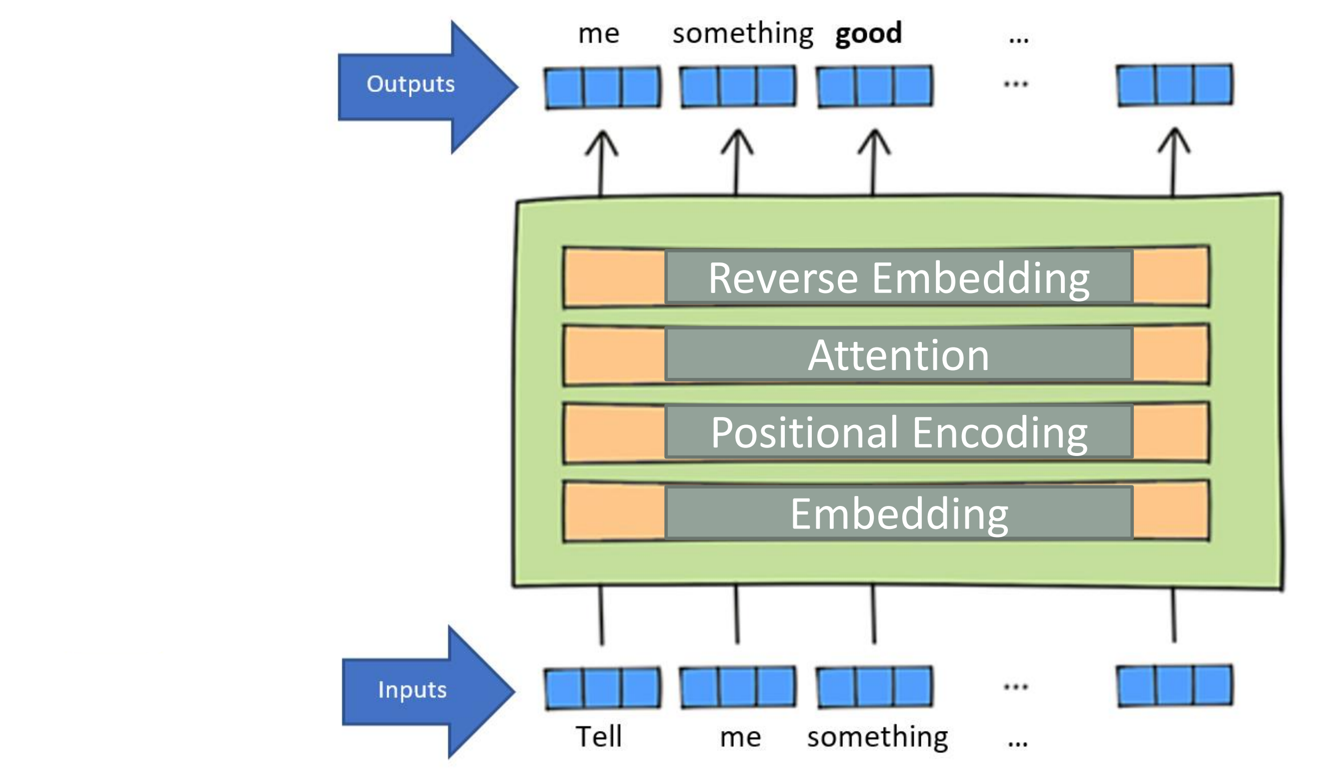


Diagram adapted from https://txt.cohere.com/what-are-transformer-models

Recap from 1000 Feet



Adapted from https://www.lavivienpost.com/how-chatgpt-works-architecture-illustrated/

What We Know and Don’t Know

ChatGPT is predicting the next token based on a generalized context.

It’s processes thousands of words in parallel.

Tokens are represented as vectors of numbers.

Each vector gets transformed into a prediction.

Sometimes ChatGPT “hallucinates”. Can this be solved?

Is ChatGPT thinking or reasoning?

What would William of Okham say?

Up Next…

→Training and Fine-Tuning

→The Discovery of Prompt Engineering

→Social, Ethical, Cognitive Implications