AI Camp

TRANSFORMERS AND ATTENTION

LLMs

PRESENTED BY

BOUKHARI AIMEN

school of ai Algiers

AI Camp

THE ENCODER-DECODER FRAMEWORK

- LSTMs and RNNs were state-of-the-art in NLP
- Feedback loop allows information propagation across steps, ideal for sequential data like text.
- Outputs a vector (hidden state) at each step, retaining information from previous steps.

school of ai Algiers

AI Camp

THE ENCODER-DECODER FRAMEWORK

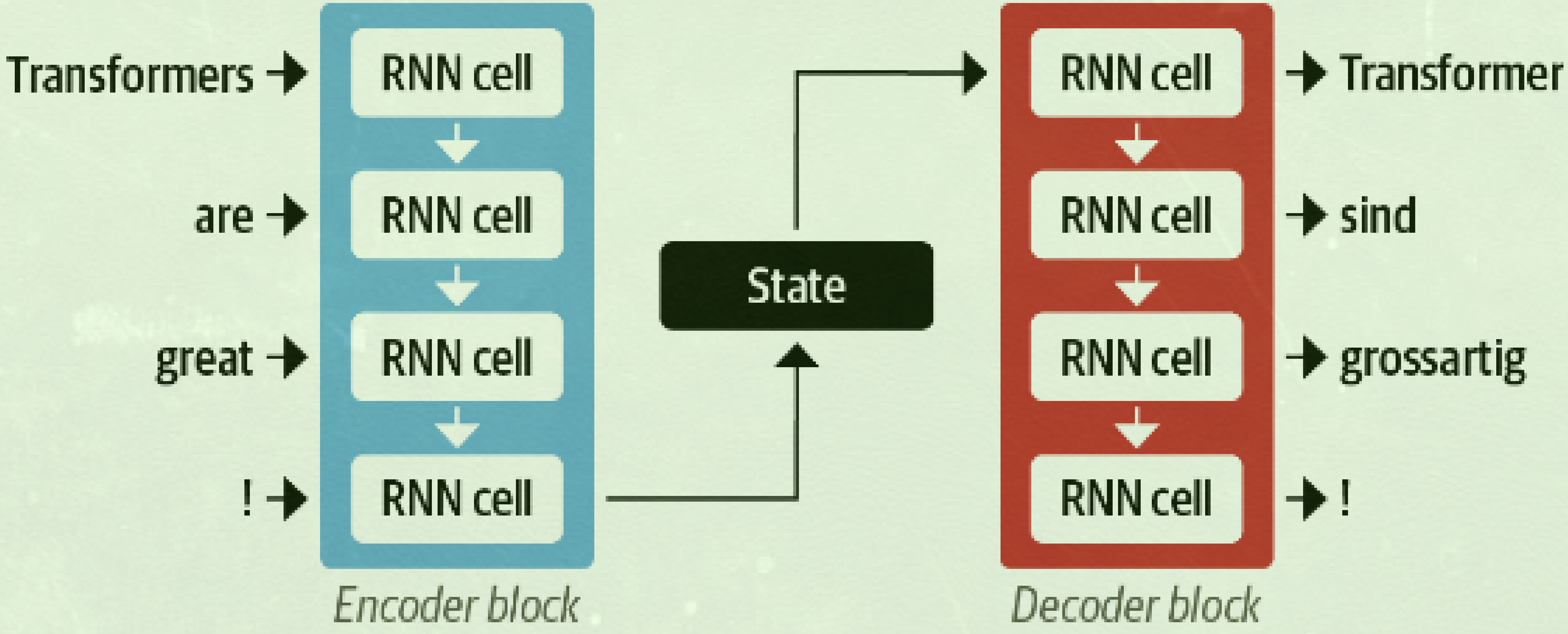
RNNs in Machine Translation

- **Encoder:** Encodes input sequence into a numerical representation (last hidden state).
- **Decoder**: Generates output sequence from the encoded representation.

school of ai Algiers

AI Camp

THE ENCODER-DECODER FRAMEWORK



school of ai Algiers

AI Camp

BUT !!!!

Information Bottleneck:

- Final hidden state must represent the entire input sequence.
- Difficult for long sequences; early information can get lost.

school of ai Algiers

AI Camp

ATTENTION MECHANISMS

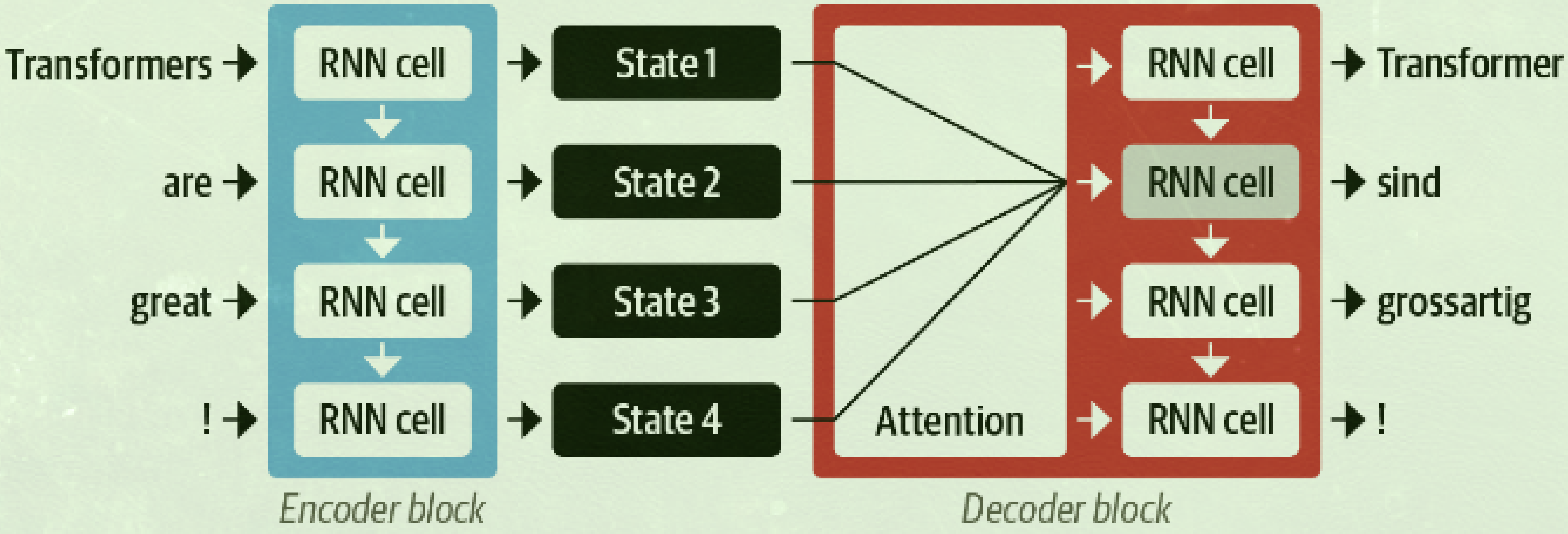
Key Ideas

- Decoder assigns weights to each encoder state at every decoding timestep.

school of ai Algiers

AI Camp

ATTENTION MECHANISMS



school of ai Algiers

AI Camp

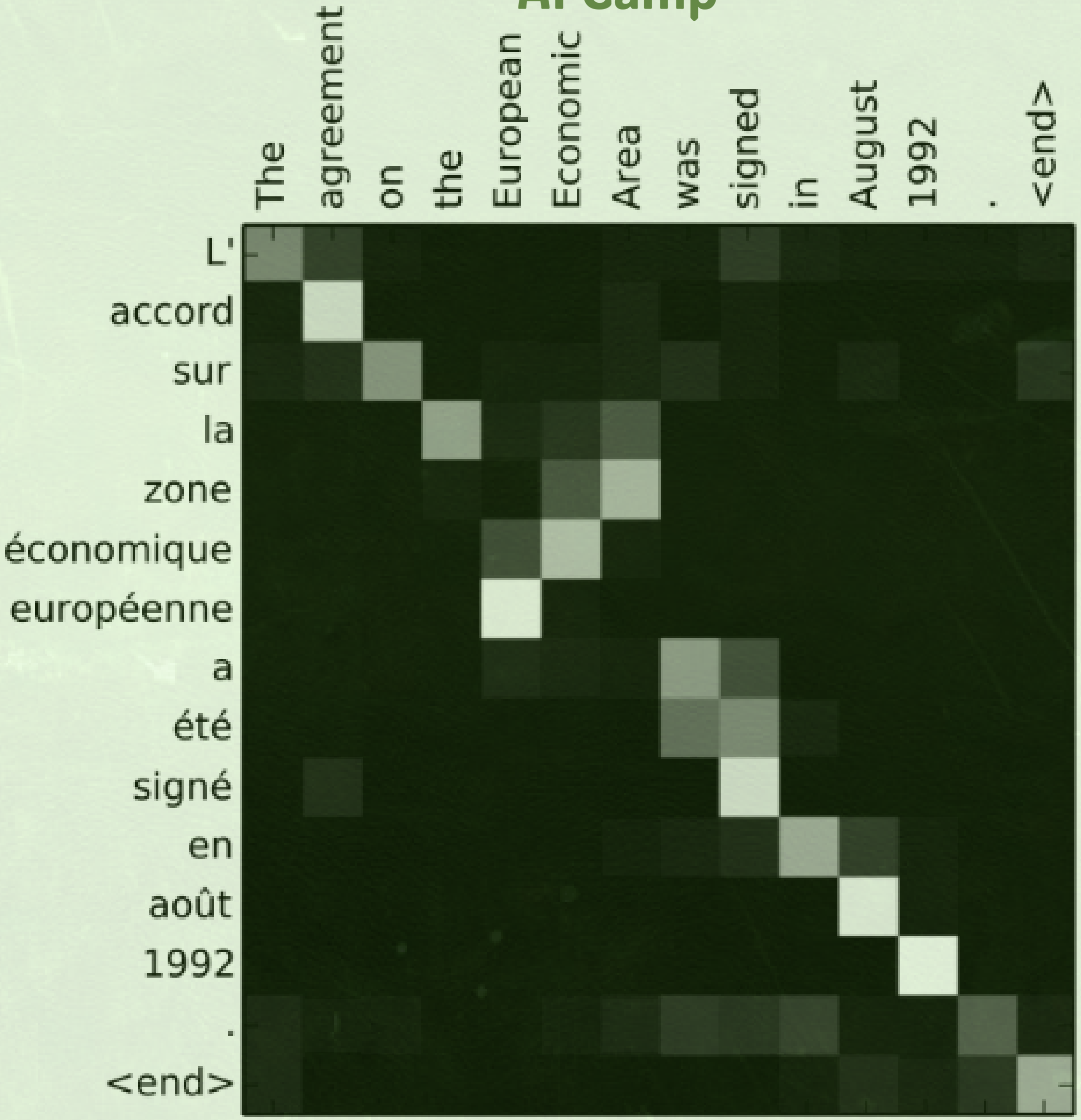
ATTENTION MECHANISMS

Key Ideas

- Focuses on the most relevant input tokens for generating each output token.

school of ai Algiers

AI Camp

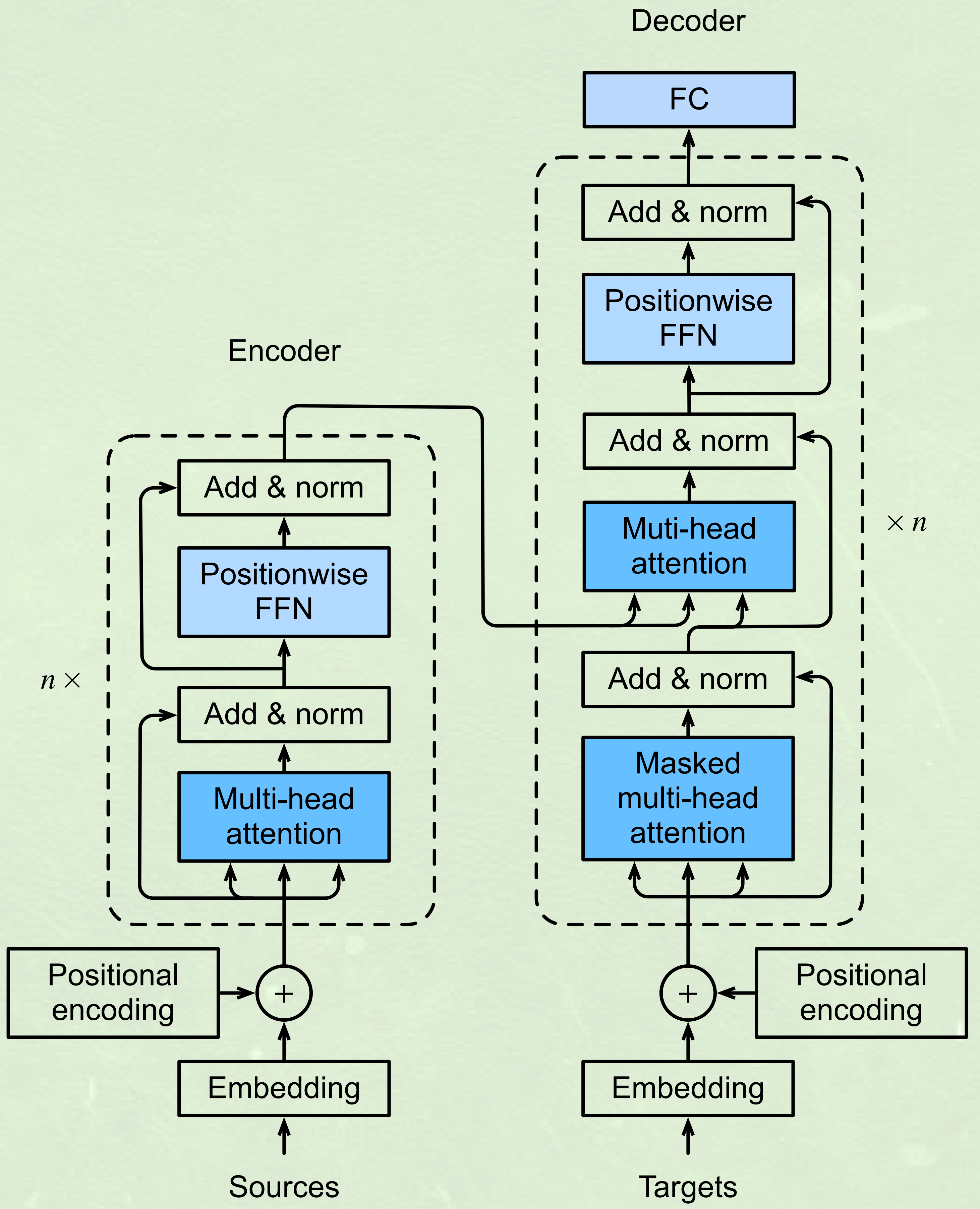


school of ai Algiers

TRANSFORMER ANATOMY

- The original Transformer is based on the encoder-decoder

AI Camp



school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

Transformer architecture types

01
ENCODER-ONLY MODELS

02
ENCODER-DECODER MODELS

03
DECODER-ONLY MODELS

AI Camp

TRANSFORMER ANATOMY

Encoder

school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

What is input embedding ?

- An embedding aims to capture the semantic meaning of the token - similar tokens have similar embeddings.

INPUT TOKENS → **EMBEDDINGS**

school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

What is input embedding ?

Original sentence (tokens)	YOUR	CAT
Input IDs (position in the vocabulary)	105	6587
Embedding (vector of size 512)	952.207	171.411
	5450.840	3276.350
	1853.448	9192.819
	...	...
	1.658	3633.421
	2671.529	8390.473

school of ai Algiers

AI Camp

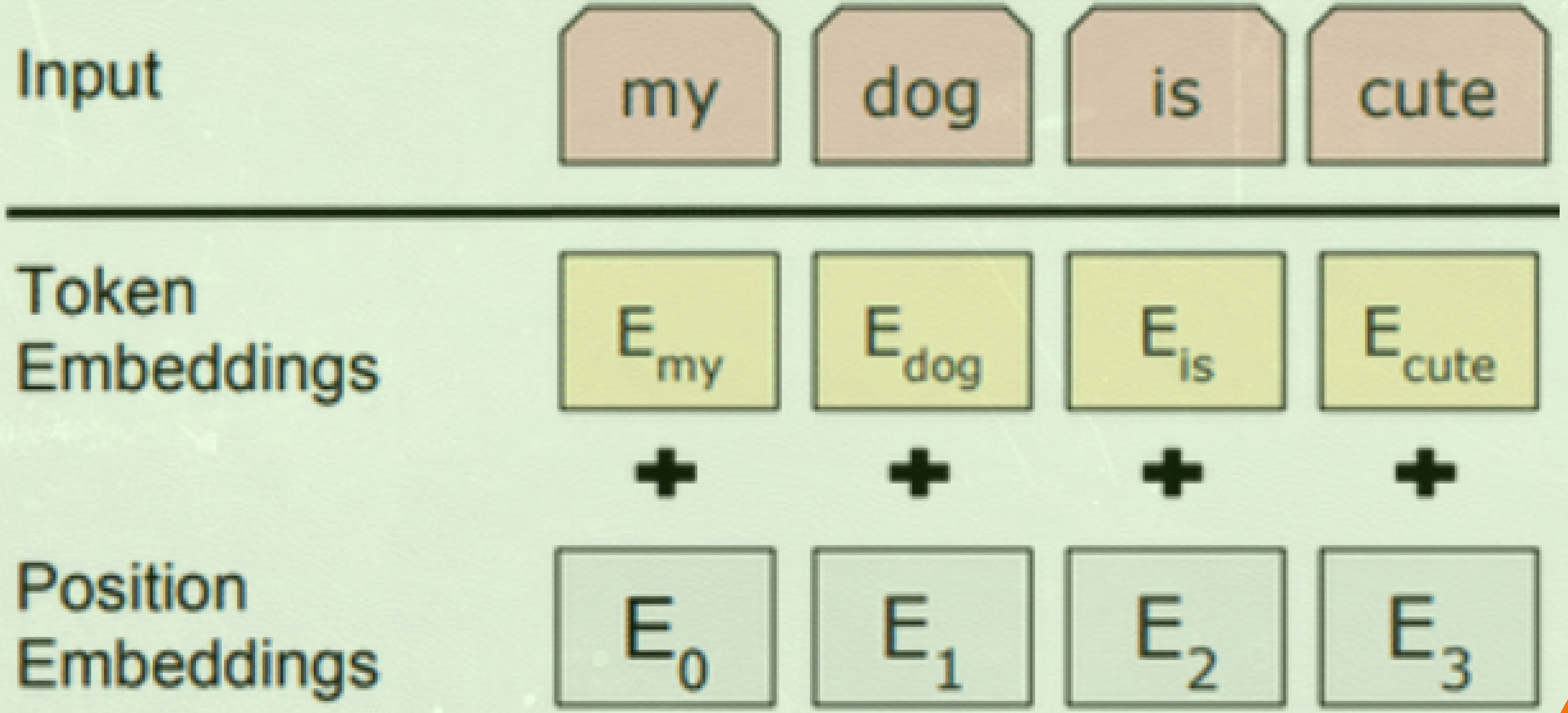
TRANSFORMER ANATOMY

Positional embeding

- We want each word to carry some information about its position in the sentence
- We want out model to treat words that appear close to each others as “close”

AI Camp

TRANSFORMER ANATOMY



school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

How positional embeding work ?

Techniques

- Absolute positional representations
  - sinusoidal
  - learned position
- Relative positional representations

school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

How positional embeding work ?

- One of the techniques they used Cos/Sin :

$$PE_{(pos\ 2i)} = sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

$$PE_{(pos\ 2i+1)} = cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

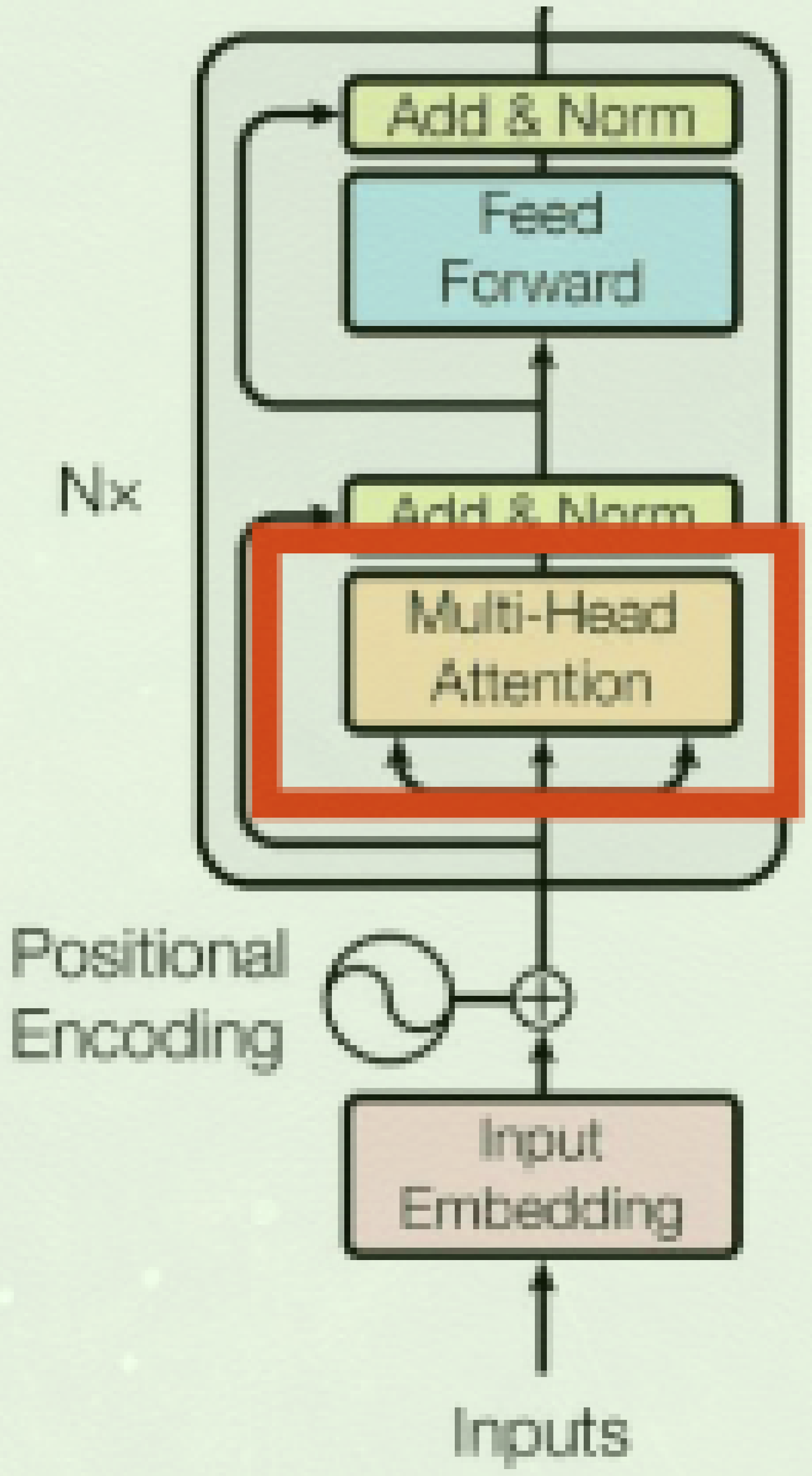
How positional embeding work ?

- Why cos/sin ? :

chose this function because we hypothesized it would allow the model to easily learn to attend by relative positions, since for any fixed offset k, PE_{pos+k} can be represented as a linear function of PE_{pos}.

school of ai Algiers

AI Camp



school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

What is Self-Attention?

- Self-Attention allows the model to relate words to each other
- Instead of using a fixed embedding for each token, the whole sequence is used to compute a weighted average for each token embedding.

school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

What is Self-Attention?

weighted average

$$x'_i = \sum_{j=1}^{n} w_{ji} x_j$$

school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

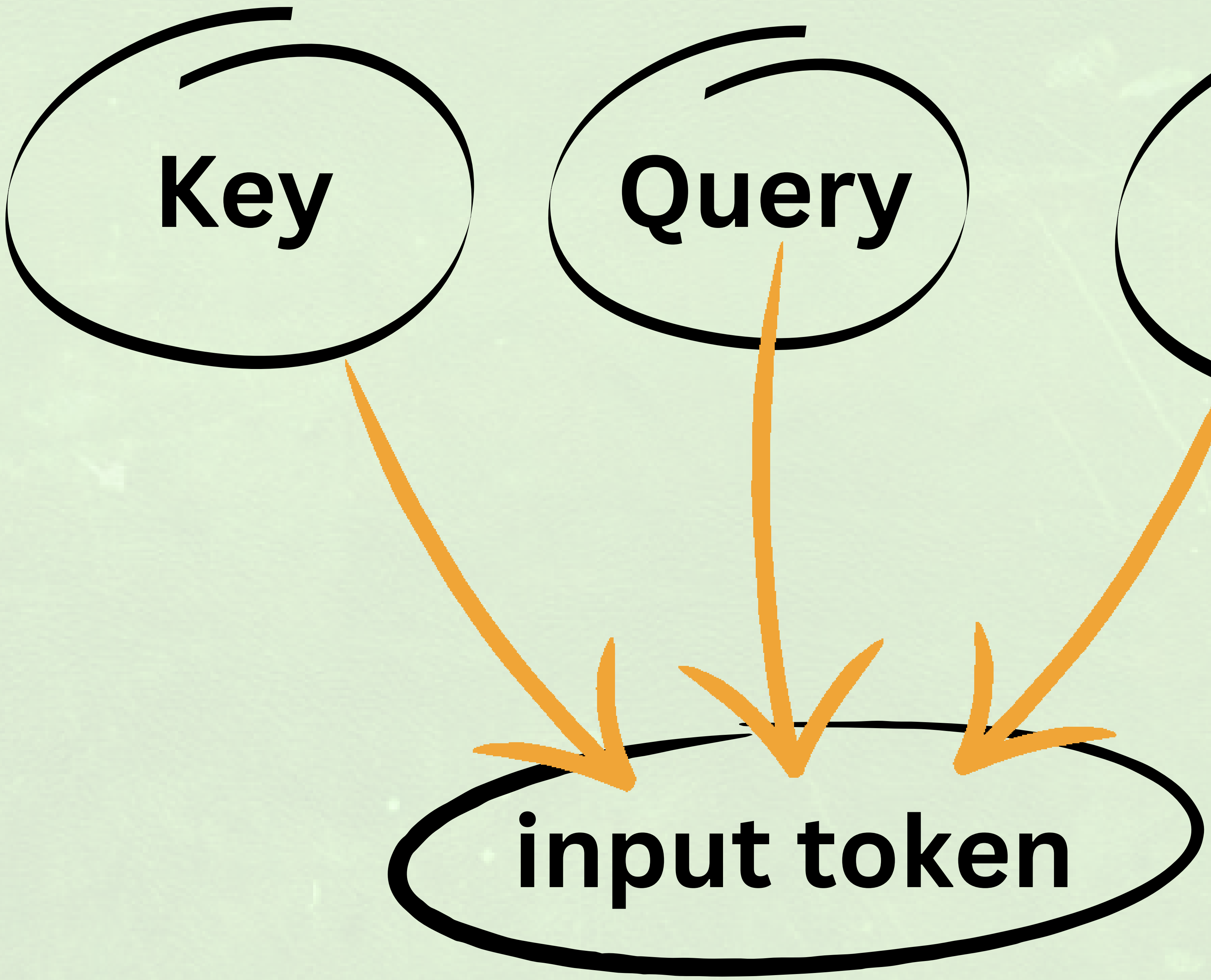
	YOUR	CAT	IS	A	LOVELY	CAT	Σ
YOUR	0.268	0.119	0.134	0.148	0.179	0.152	1
CAT	0.124	0.278	0.201	0.128	0.154	0.115	1
IS	0.147	0.132	0.262	0.097	0.218	0.145	1
A	0.210	0.128	0.206	0.212	0.119	0.125	1
LOVELY	0.146	0.158	0.152	0.143	0.227	0.174	1
CAT	0.195	0.114	0.203	0.103	0.157	0.229	1

school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

- Step 1: Project Token Embeddings:



school of ai Algiers

TRANSFORMER ANATOMY

- **Step 2: Compute Attention Scores:**
  - The similarity between the **query** and **key** vectors is computed using **the dot product**

Note :

Queries and keys that are more similar have a larger dot product

AI Camp

$$\text{softmax}\left(\frac{\underset{(6,\ 512)}{Q} \times \underset{(512,\ 6)}{K^T}}{\sqrt{512}}\right)$$

school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

- step3 : update Token Embeddings

$$x'_i = \sum_{j=1}^{n} w_{ji} x_j \longrightarrow \text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

school of ai Algiers

AI Camp

	YOUR	CAT	IS	A	LOVELY	CAT
YOUR	0.268	0.119	0.134	0.148	0.179	0.152
CAT	0.124	0.278	0.201	0.128	0.154	0.115
IS	0.147	0.132	0.262	0.097	0.218	0.145
A	0.210	0.128	0.206	0.212	0.119	0.125
LOVELY	0.146	0.158	0.152	0.143	0.227	0.174
CAT	0.195	0.114	0.203	0.103	0.157	0.229

(6, 6)

X

V

(6, 512)

=

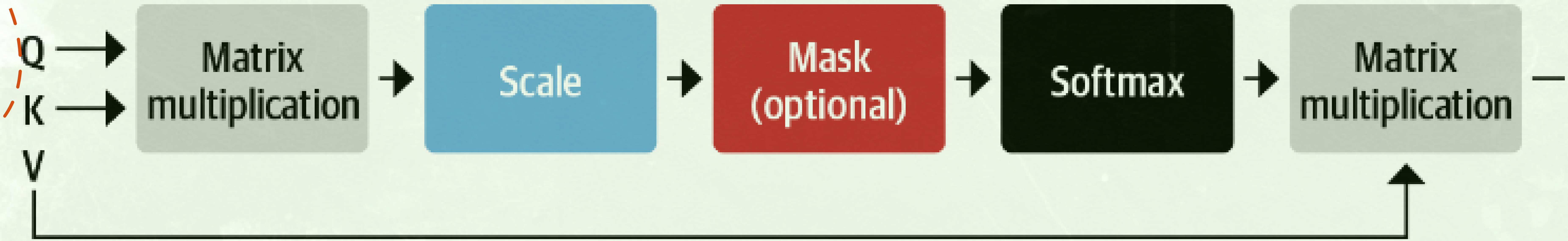
Attention

(6, 512)

school of ai Algiers

AI Camp

self-attention process



The final dimension of the encoder output is : n*d (original dimension)

school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

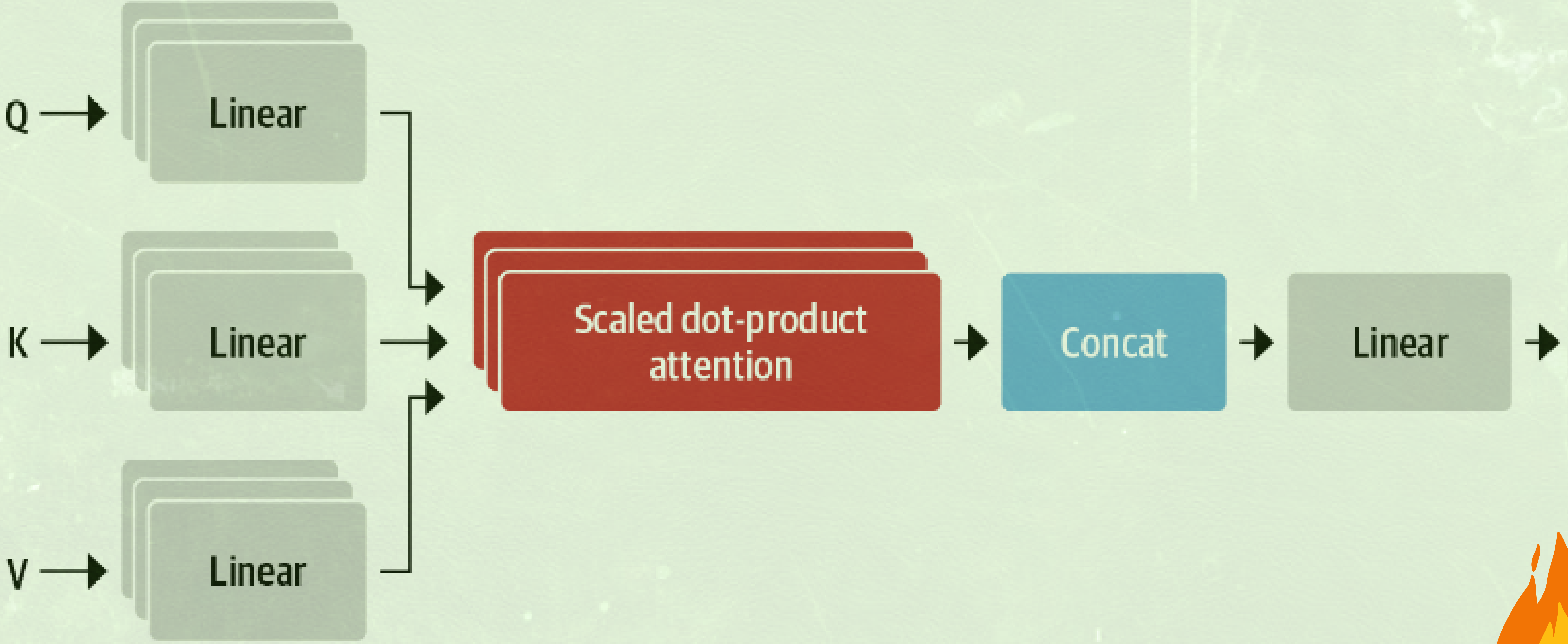
but why Multi-head Attention ?

- Instead of a single set of linear projections, multiple sets (called attention heads) are used.
- Multi-head attention allows the model to capture various semantic relations in parallel.

school of ai Algiers

AI Camp

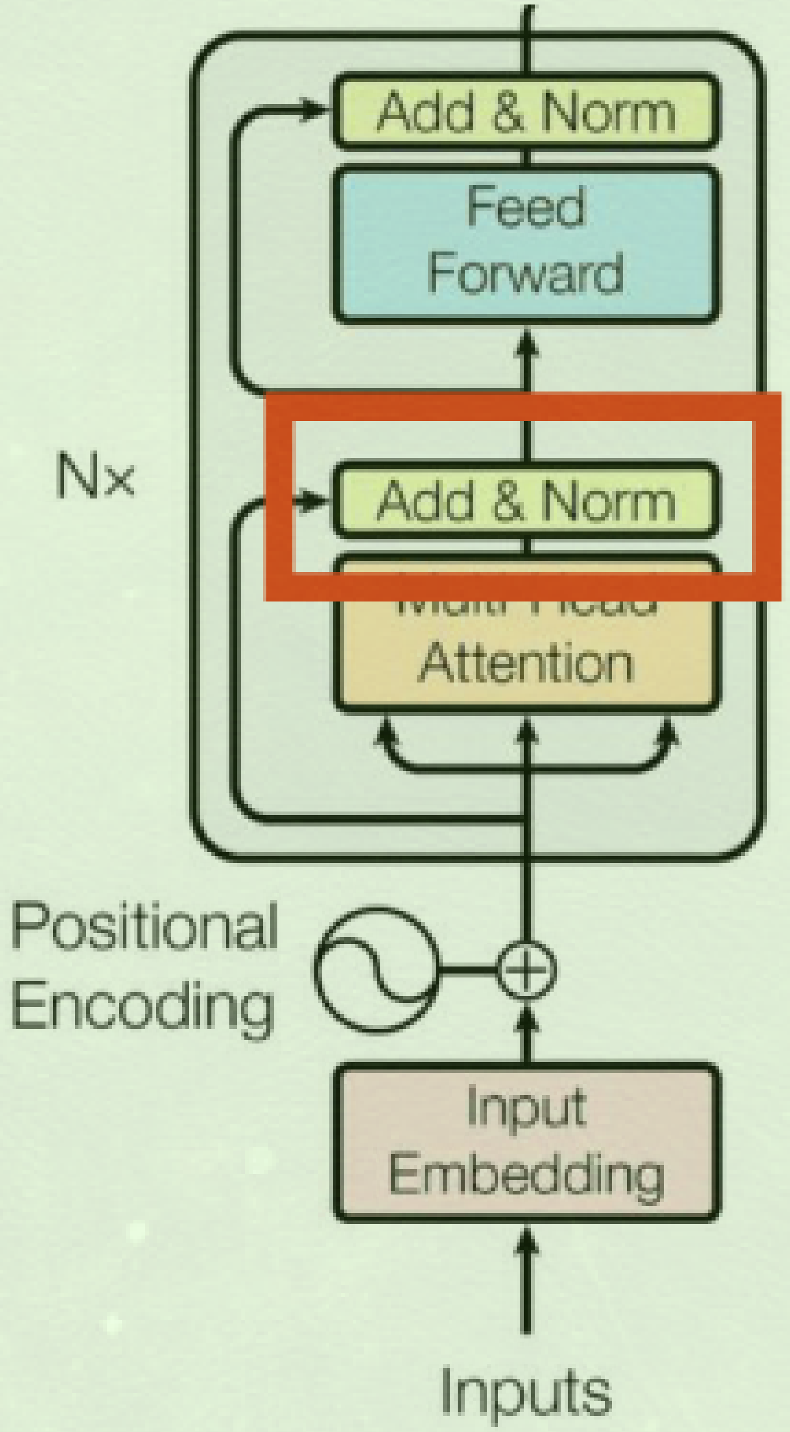
TRANSFORMER ANATOMY



school of ai Algiers

AI Camp

TRANSFORMER ANATOMY



school of ai
Algiers

AI Camp

TRANSFORMER ANATOMY

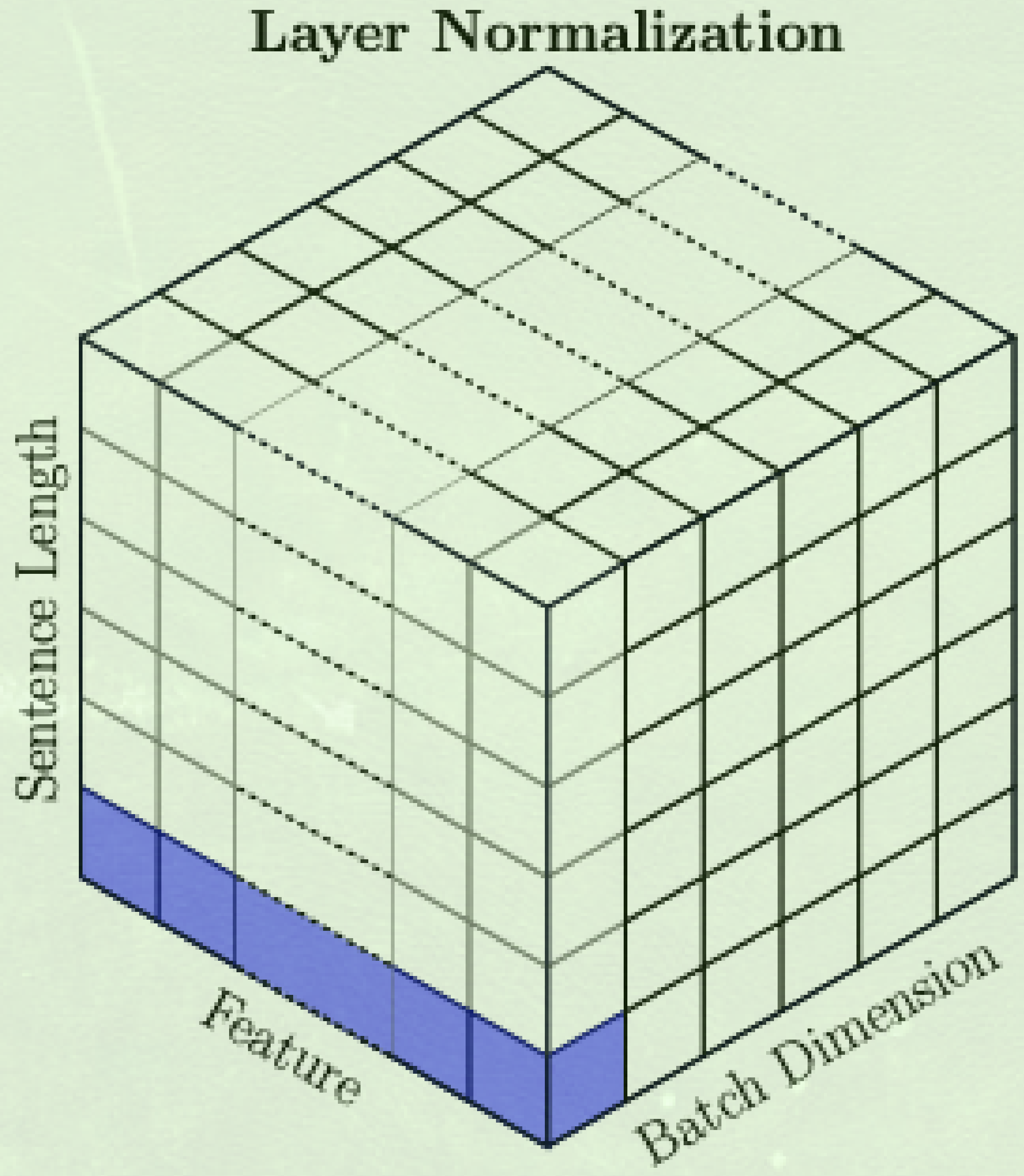
What is layer normalization?

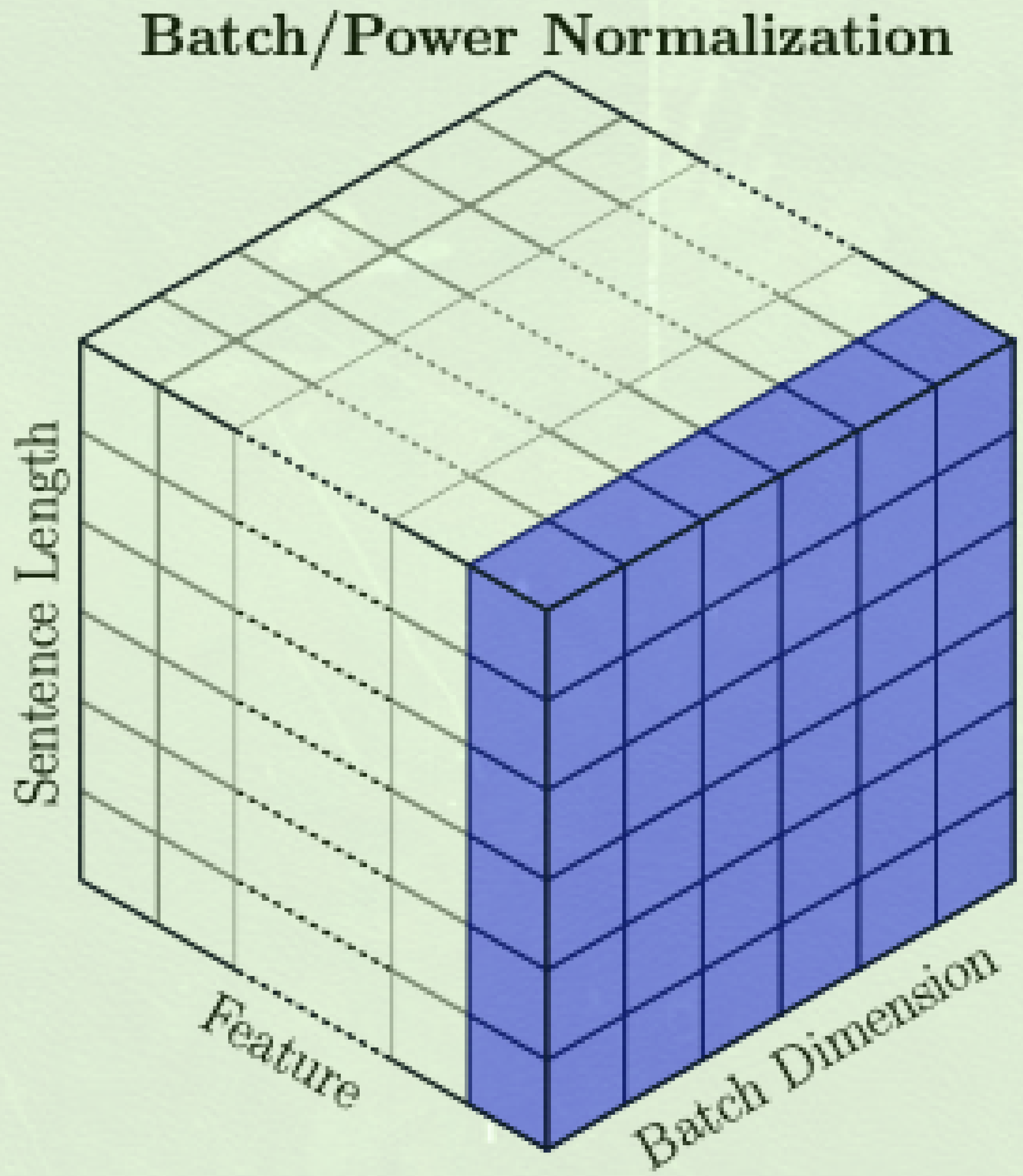
- Normalizes each input in the batch to have zero mean and unit variance.
- Improves training stability and convergence.

school of ai Algiers

AI Camp

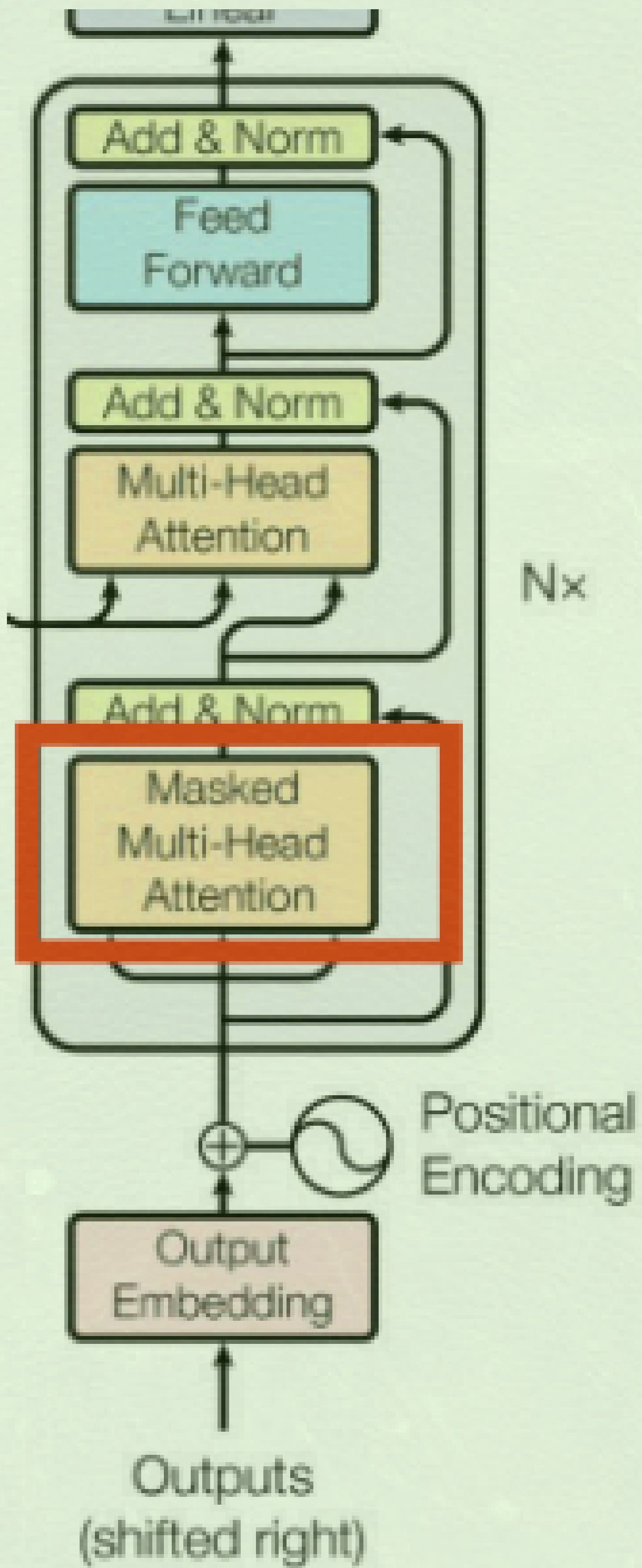
TRANSFORMER ANATOMY





school of ai Algiers

AI Camp



school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

What is Masked Multi-Head Attention?

- **Our goal is to make the model causal:**

it means the output at a certain position can only depend on the words on the previous positions. The model must not be able to see future words

school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

Attention weight between the tokens corresponding to “Life” and “short”

	Life	is	short	eat	desert	first
Life	0.17	0.13	0.18	0.16	0.15	0.18
is	0.03	0.68	0.02	0.08	0.14	0.02
short	0.19	0.06	0.25	0.14	0.11	0.23
eat	0.15	0.21	0.14	0.16	0.17	0.14
desert	0.13	0.27	0.11	0.16	0.18	0.12
first	0.19	0.02	0.31	0.11	0.07	0.27

Attention weights calculated in the previous “Self-Attention” section

In the row for corresponding to “Life”, mask out all words that come after “Life”

	Life	is	short	eat	desert	first
Life	0.17	0.13	0.18	0.16	0.15	0.18
is	0.03	0.68	0.02	0.08	0.14	0.02
short	0.19	0.06	0.25	0.14	0.11	0.23
eat	0.15	0.21	0.14	0.16	0.17	0.14
desert	0.13	0.27	0.11	0.16	0.18	0.12
first	0.19	0.02	0.31	0.11	0.07	0.27

school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

Training a transformers ?

school of ai Algiers

AI Camp

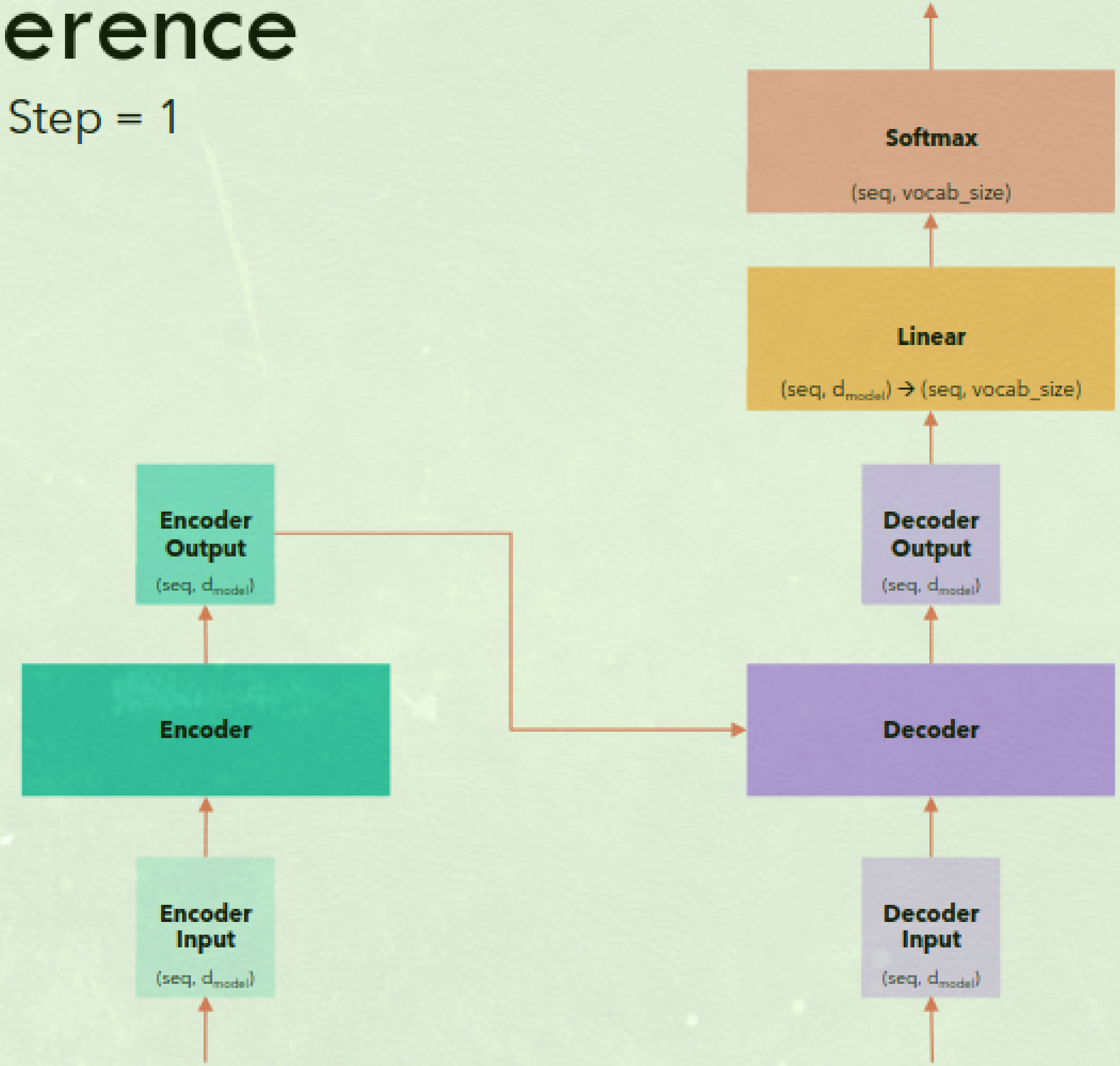
TRANSFORMER ANATOMY

Inference (Prediction) ?

school of ai Algiers

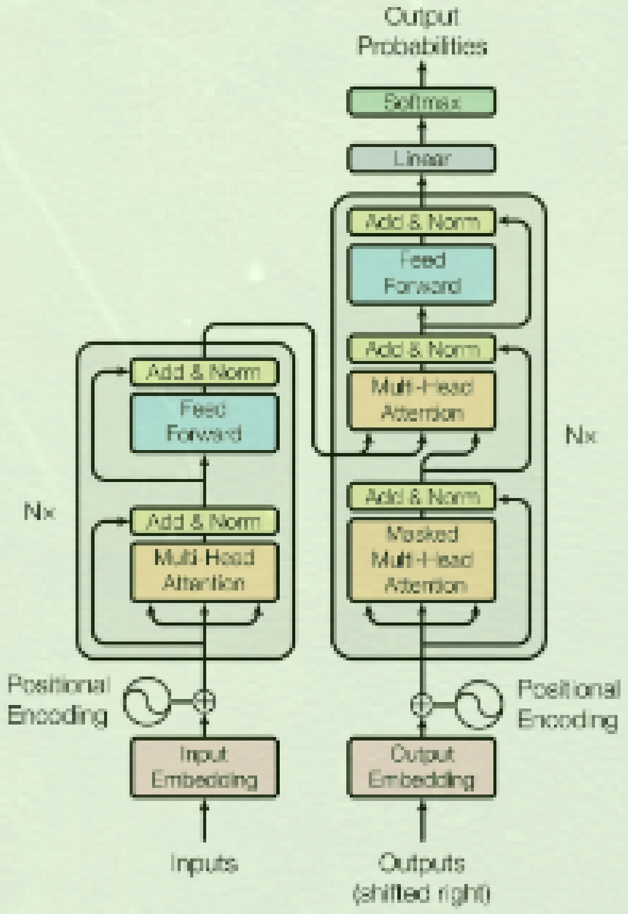
Inference

Time Step = 1



We select a token from the vocabulary corresponding to the position of the token with the maximum value.

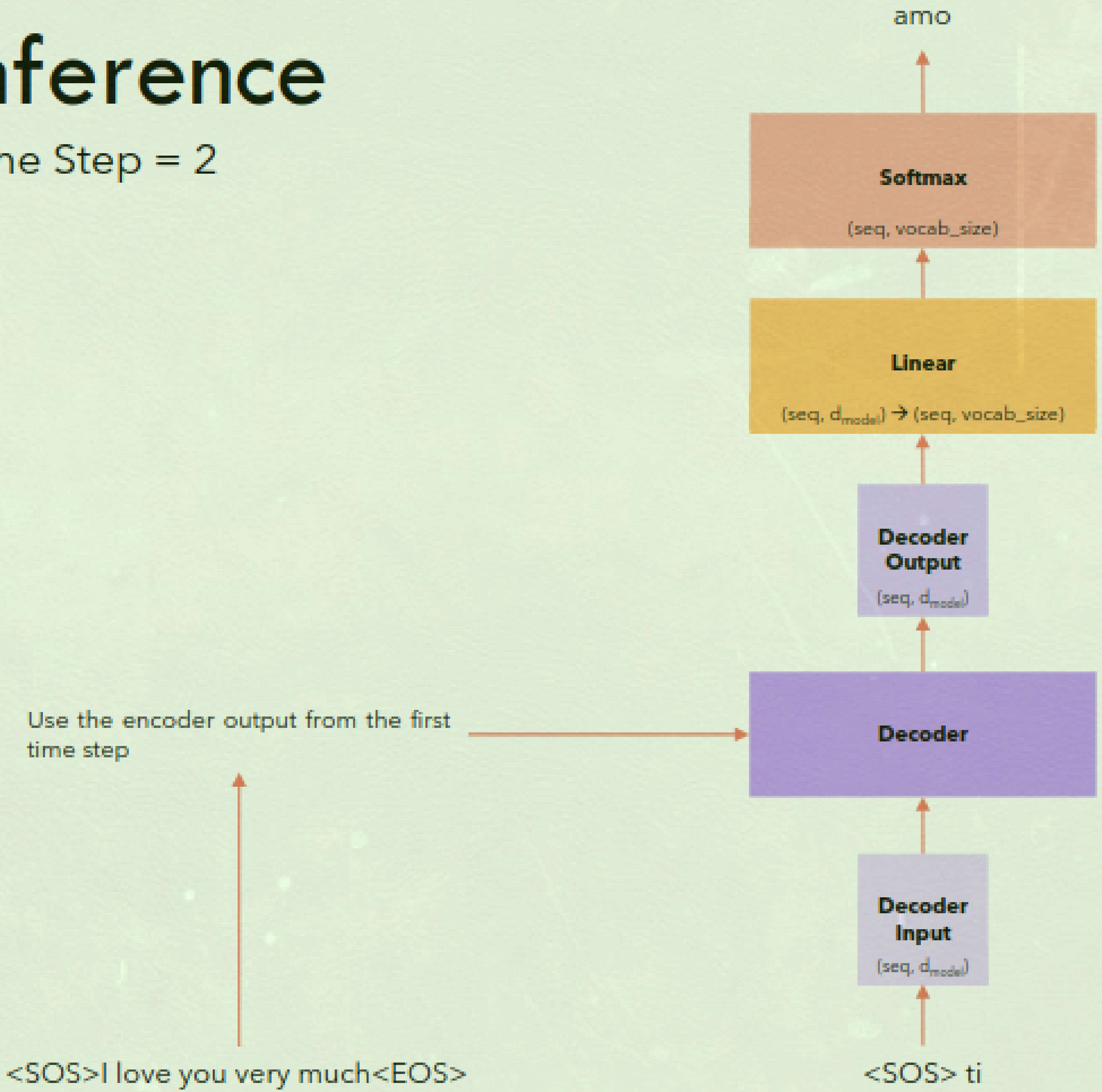
The output of the last layer is commonly known as **logits**



* Both sequences will have same length thanks to padding

Inference

Time Step = 2



AI Camp

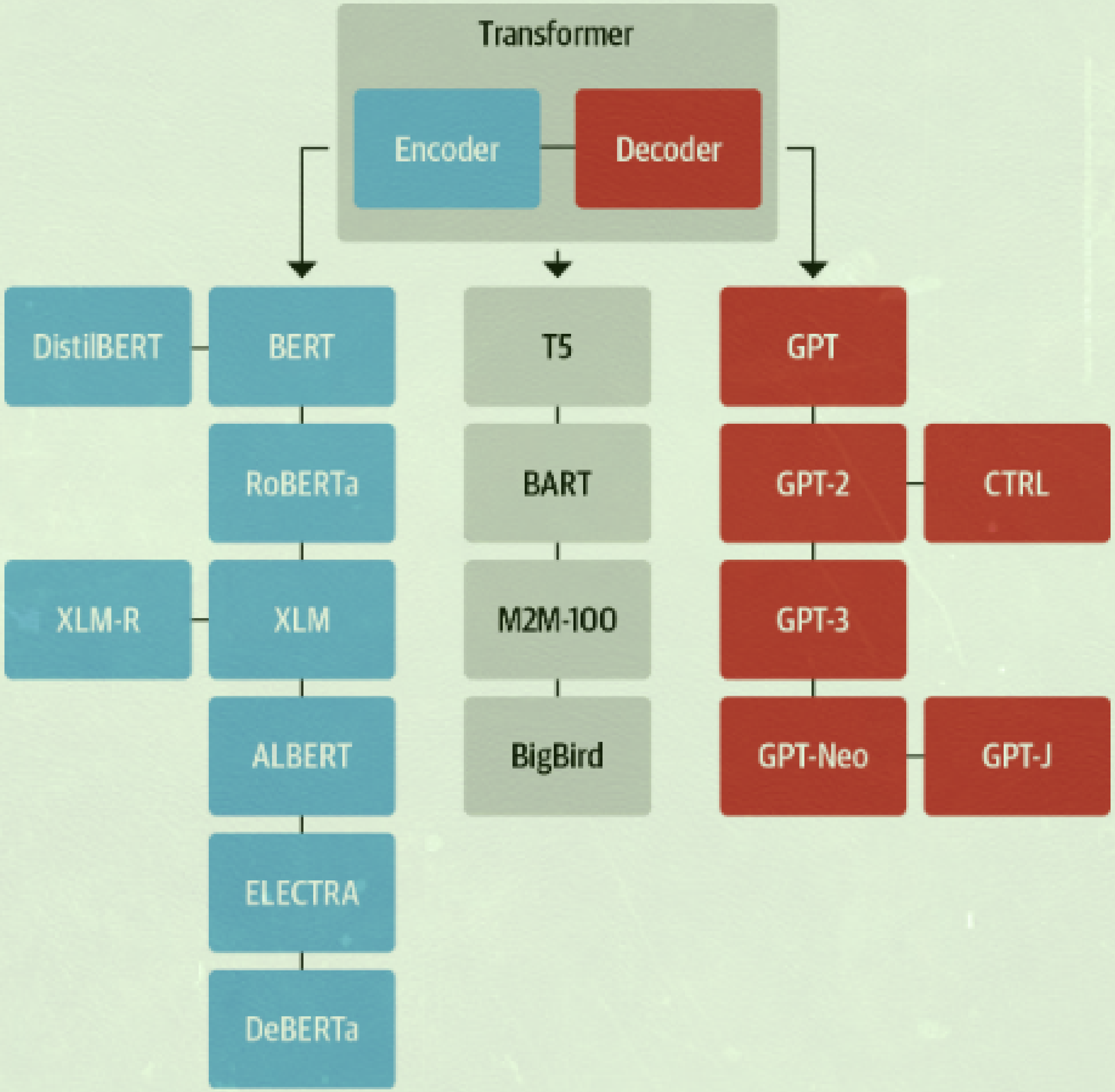
TRANSFORMER ANATOMY

Inference strategy

- Beam-search
- greedy search

school of ai Algiers

AI Camp



school of ai Algiers

AI Camp

Bidirectional Encoder Representations from Transformers

school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

Bert Information

- BERT stands for **Bidirectional Encoder Representations from Transformers**
- Unlike common language models, BERT has been trained using the **left context and the right context**

school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

Bert Information

- BERT has not been trained on the Next Token Prediction task, but rather, on the **Masked Language Model and Next Sentence Prediction task**

school of ai Algiers

AI Camp

TRANSFORMER ANATOMY

Pre-Trained BERT

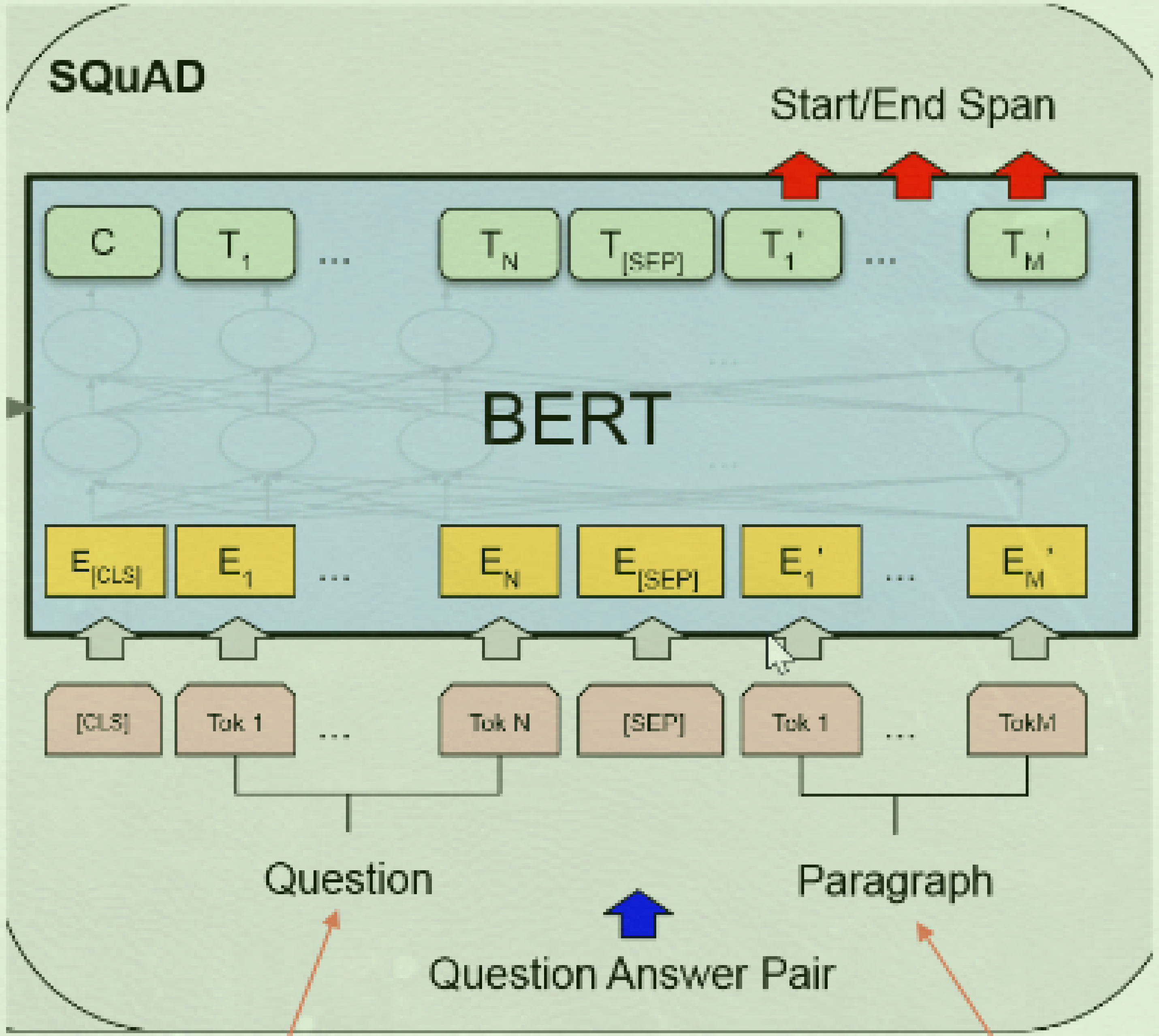
Fine Tune on text classification

Fine Tune on question answering (QA)

school of ai Algiers

AI Camp

TRANSFORMER ANATOMY



Sentence A

Sentence B

school of ai Algiers

AI Camp

school of ai Algiers

AI Camp

THANK YOU!

school of ai Algiers