SAISON 23/24

Formation

Introduction au Deep Learning

Séquence n°9

“Attention is All You Need : Transformers”

FIDLE

Formation

Introduction au Deep Learning

Questions and answers :

https://fidle.cnrs.fr/q2a

Accompanied by :
AI Support (dream) Team of IDRIS

Directed by :
Agathe, Baptiste et Yanis - UGA/DAPI
Léo, Thibaut, Kamel - IDRIS

FIDLE

Formation

Introduction au Deep Learning

FIDLE https://fidle.cnrs.fr/listeinfo

Fidle information list

Agoria http://fidle.cnrs.fr/agoria

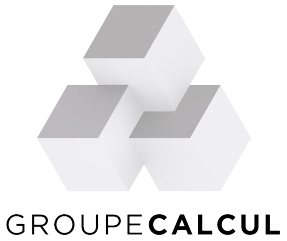
AI exchange list

Formation

Introduction au Deep Learning

https://listes.services.cnrs.fr/wws/info/devlog

List of ESR* « Software developers » group



https://listes.math.cnrs.fr/wws/info/calcul

List of ESR* « Calcul » group

(*) ESR is Enseignement Supérieur et Recherche, french universities and public academic research organizations

Abonnez-vous !

https ://fidle.cnrs.fr/youtube
https://www.youtube.com/@CNRS-FIDLE

Panoram'IA

Le magazine de actualité en IA

Vendredi 16 février, 10h00

http://www.idris.fr/panoramia.html
https://www.youtube.com/@IDRISCNRS

FIDLE

SAISON 23/24

Bases, Concepts et Enjeux

1 ✓
History and Fundamental Concepts

2 ✓ New !
Data, models and representation's hell
Data and models

3 ✓ New !
Demonstration Illustration
LLM / Text to Image

4 ✓
AI, Law, Society and Ethics

L'IA comme un outil,

5 ✓ New !
Mathematics, gradients everywhere !

6 ✓ New !
Learning methodology

7 ✓
Convolutional models
CNN

8 ✓
Sparse (text) and sequences data
Embedding, RNN

9
«Attention is All You Need»
Transformers

10
Graph Neural Network
GNN

11
Autoencoder networks
AE

12
Variational Antoencoder
VAE

13
Generative Adversarial Networks
GAN

14
Diffusion Model
Text to image

15
Deep Reinforcement Learning
RL

16
Physics-Informed Neural Networks
PINNS

Acteur de l'IA

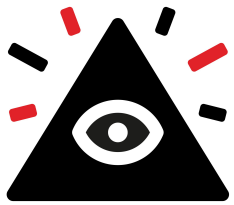
17
Learning faster and cheaper,
Eco-Friendly

18
Jean-Zay
GPU acceleration

19 New !
New models
VLM, SM, Multimodal, ...

20 New !
Case Study
Experience feedback

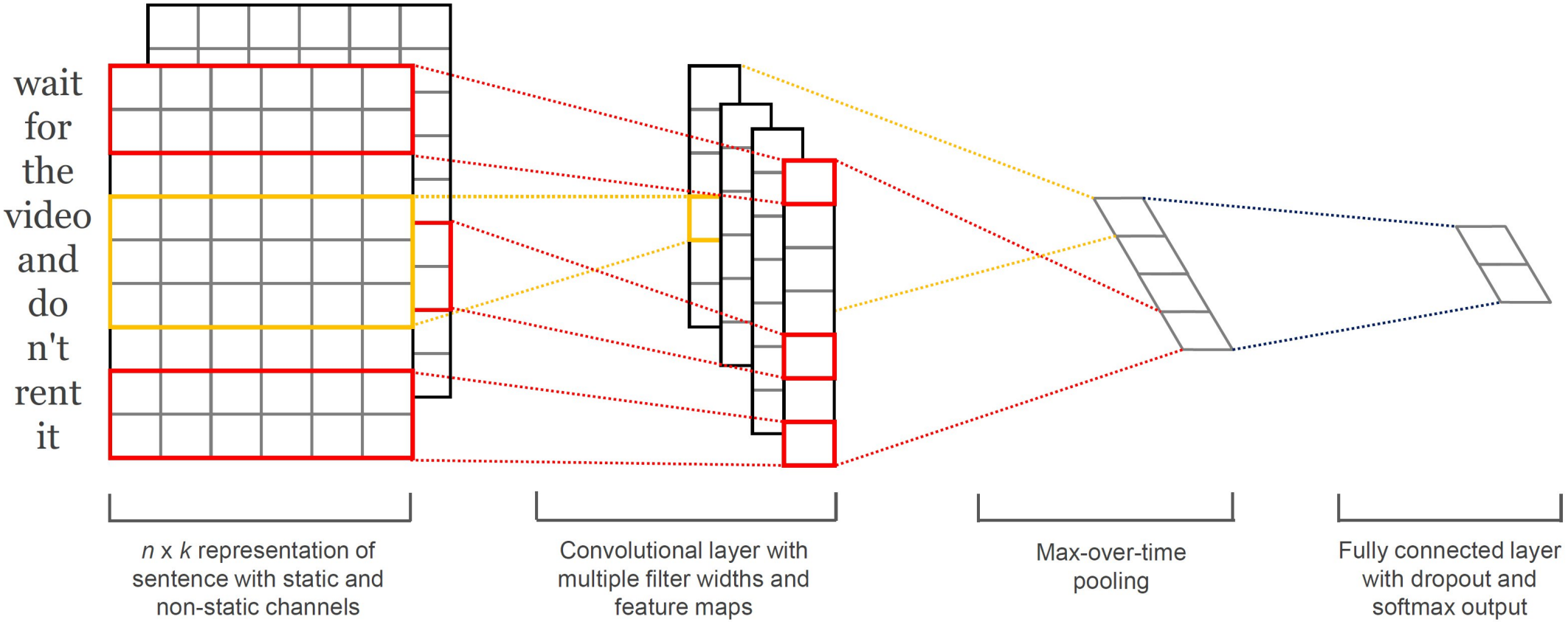
9



«Attention is All You Need»
Transformers

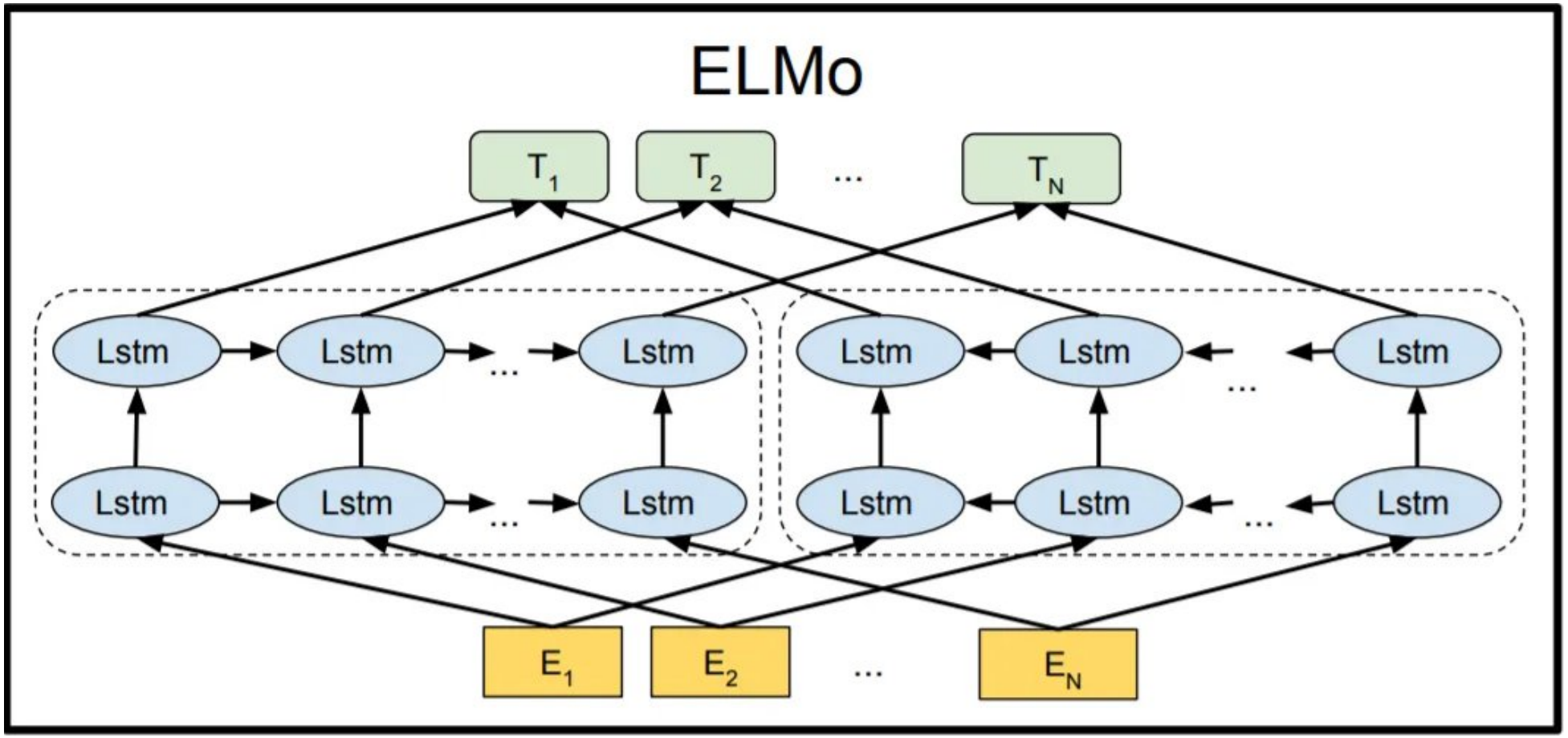
1. What is a **Transformer**? The magic of the **Attention Mechanism**
2. The different **Transformers architectures**
3. **Pre-training** and **Foundation Models**
4. **Specialization** of **Foundation Models** (especially LLM)
5. **Example: IMDB Reviews**

NLP back in the days



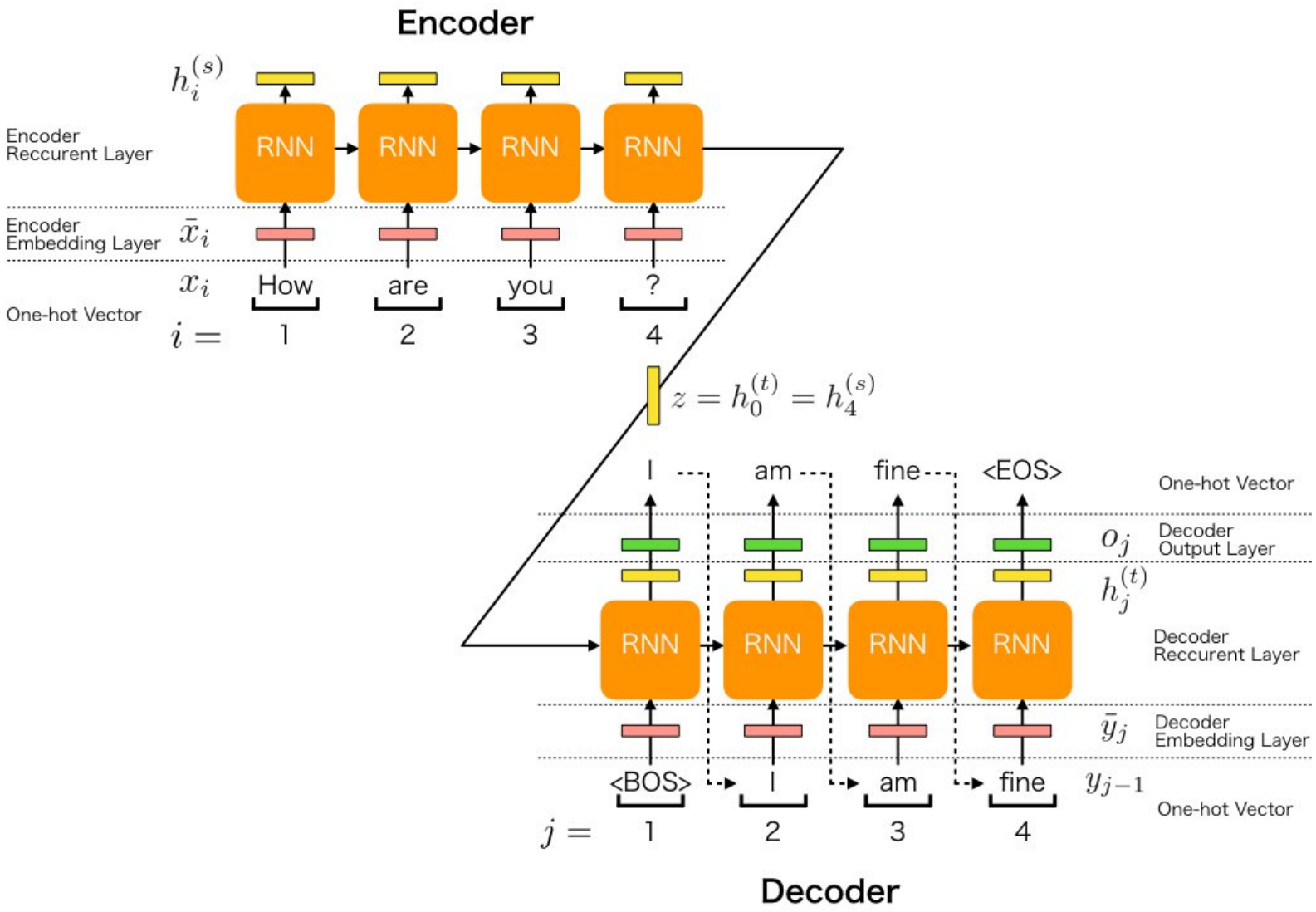
Convolutional Neural Networks for Sentence Classification, Yoon Kim, 2014, https://arxiv.org/pdf/1408.5882.pdf

NLP back in the days



https://medium.com/saarthi-ai/elmo-for-contextual-word-embedding-for-text-classification-24c9693b0045

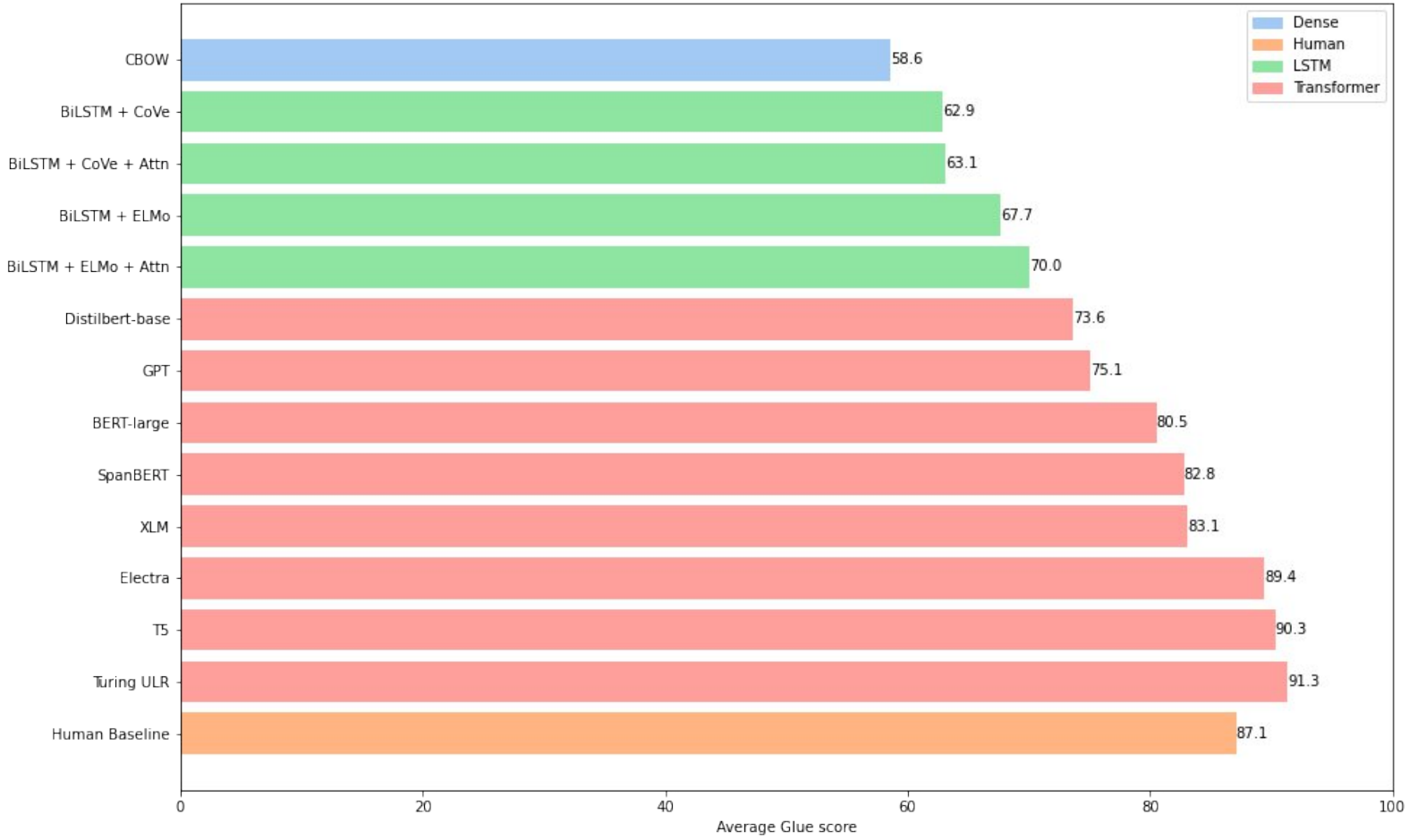
NLP back in the days



What do we want ?

- Process sequences (ideally the entire sentence)
- Easy to distribute on multiple GPUs
- Faster training than with RNN
- Initially for NLP tasks
- Allows to train huge models on gigantic datasets
- Allows for a pretraining session to pool trainings (at least partially) for multiple tasks

The King is dead. Long live the King!

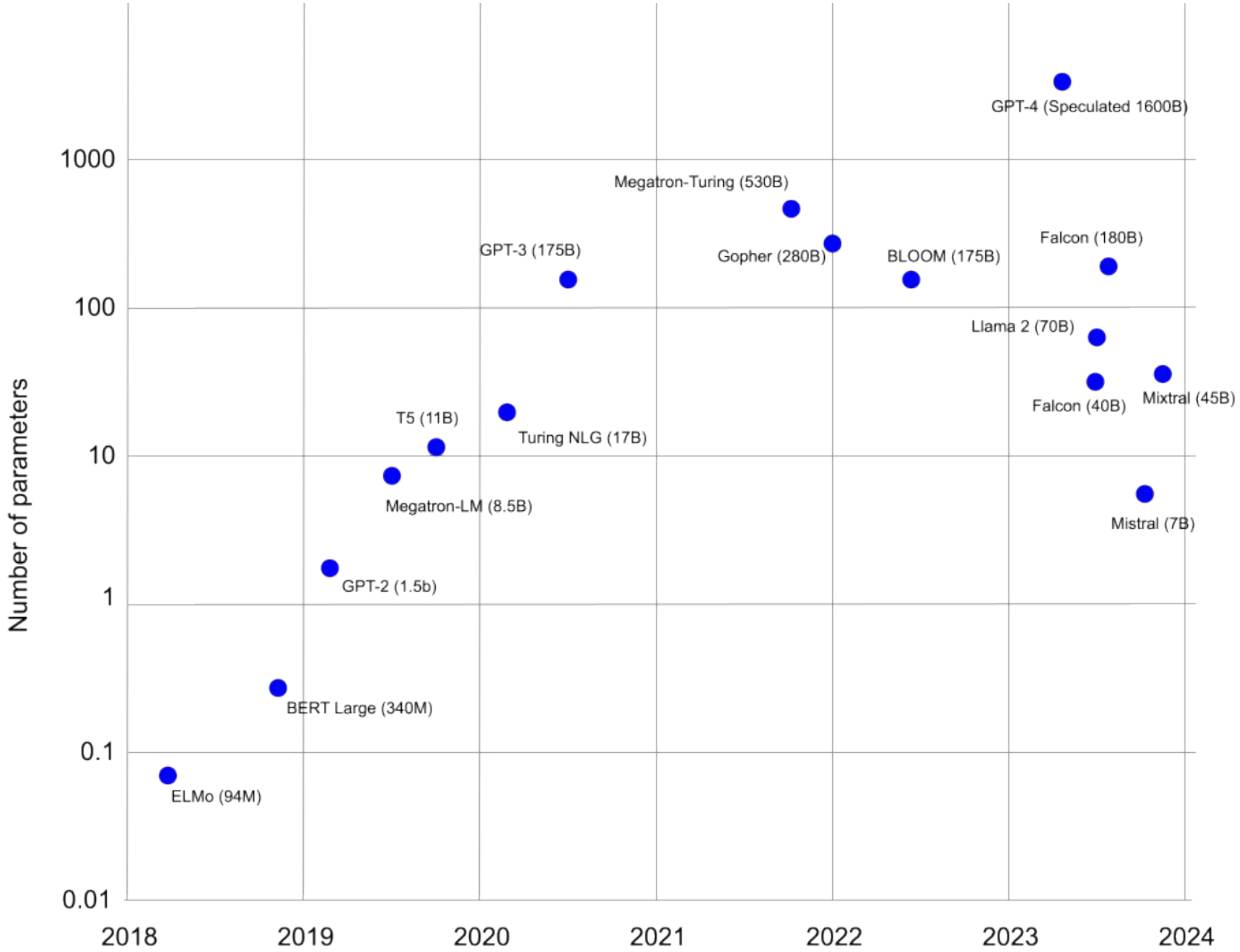


This article is more than **5 months old**

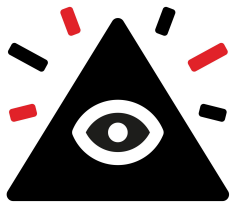
ChatGPT better than undergraduates at solving SAT problems, study suggests

Researchers at UCLA found GPT-3 solved 80% of reasoning problems correctly compared with 60% of humans

Size evolution of transformers



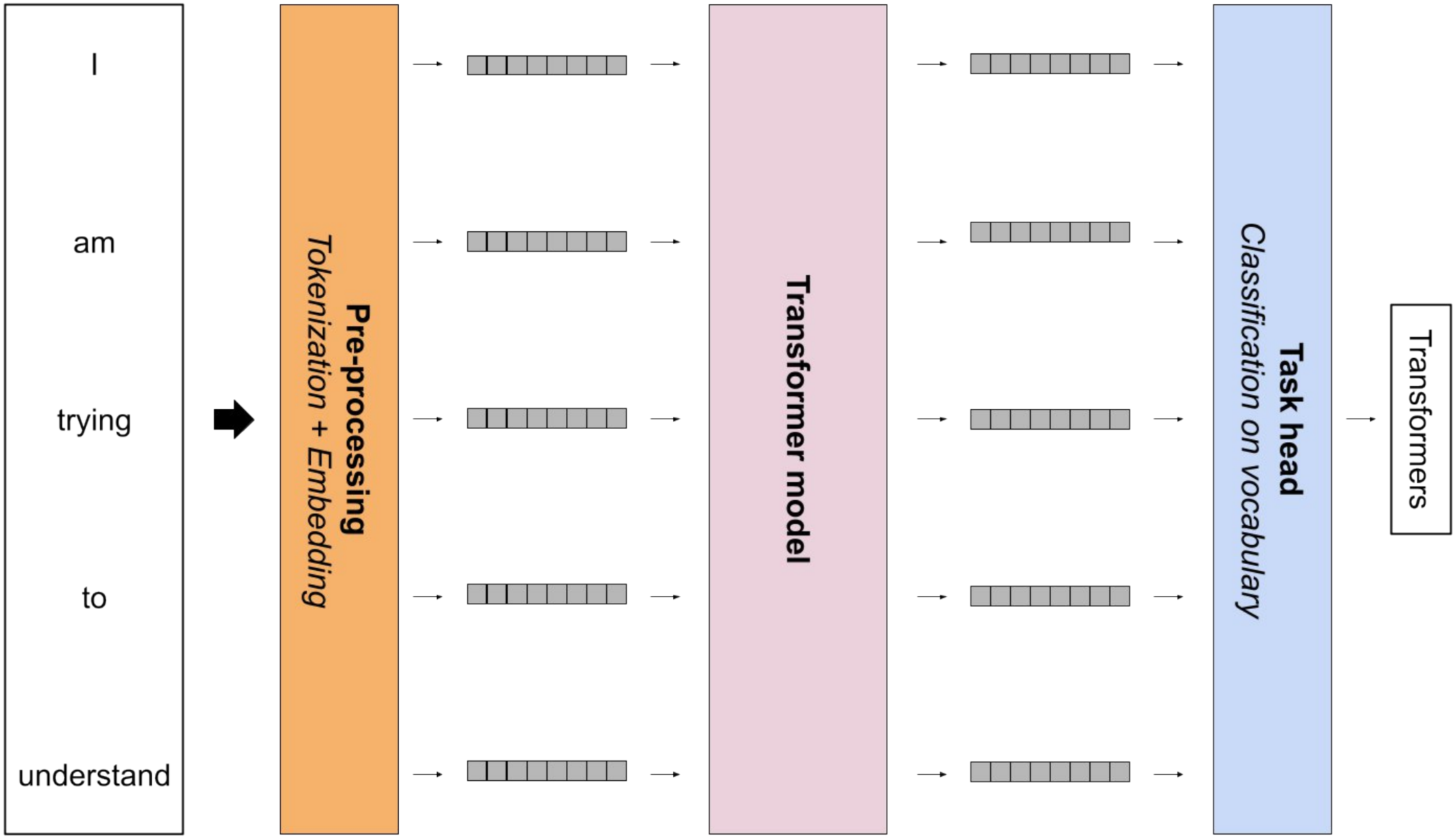
9



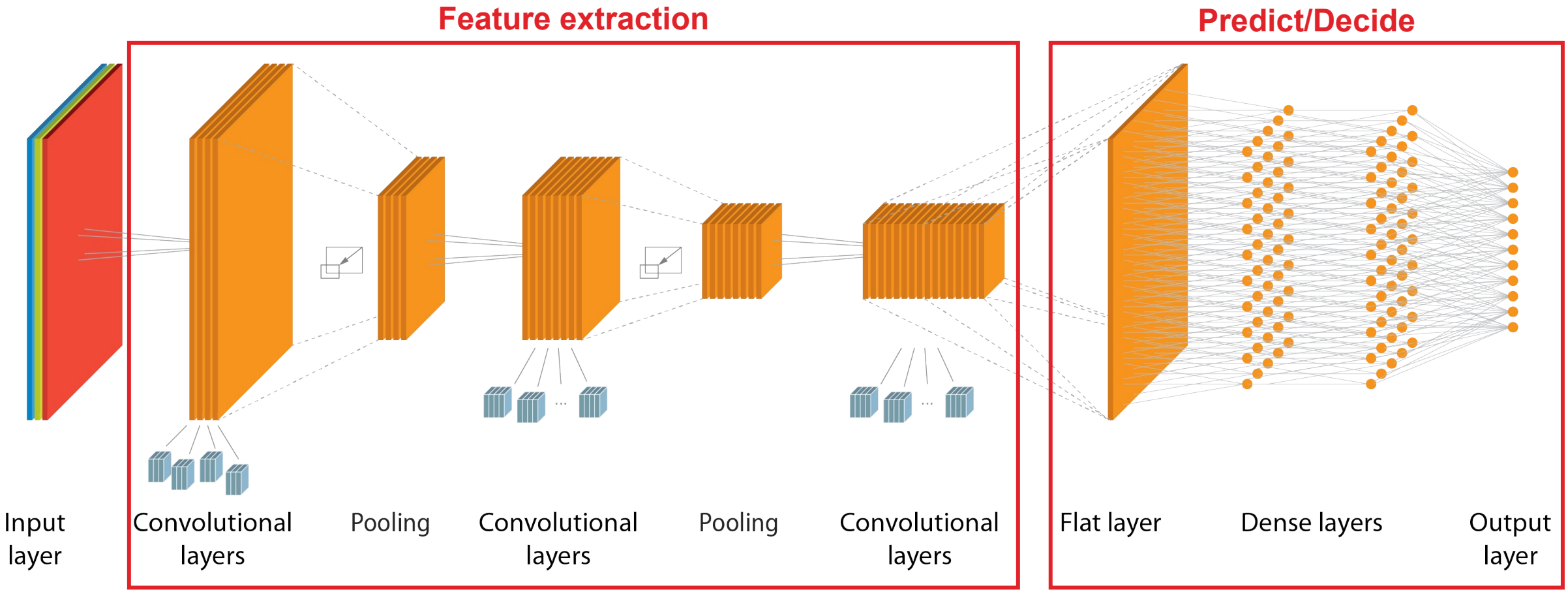
«Attention is All You Need»
Transformers

1. What is a **Transformer**? The magic of the **Attention Mechanism**
2. The different **Transformers architectures**
3. **Pre-training** and **Foundation Models**
4. **Specialization** of **Foundation Models** (especially LLM)
5. **Example: IMDB Reviews**

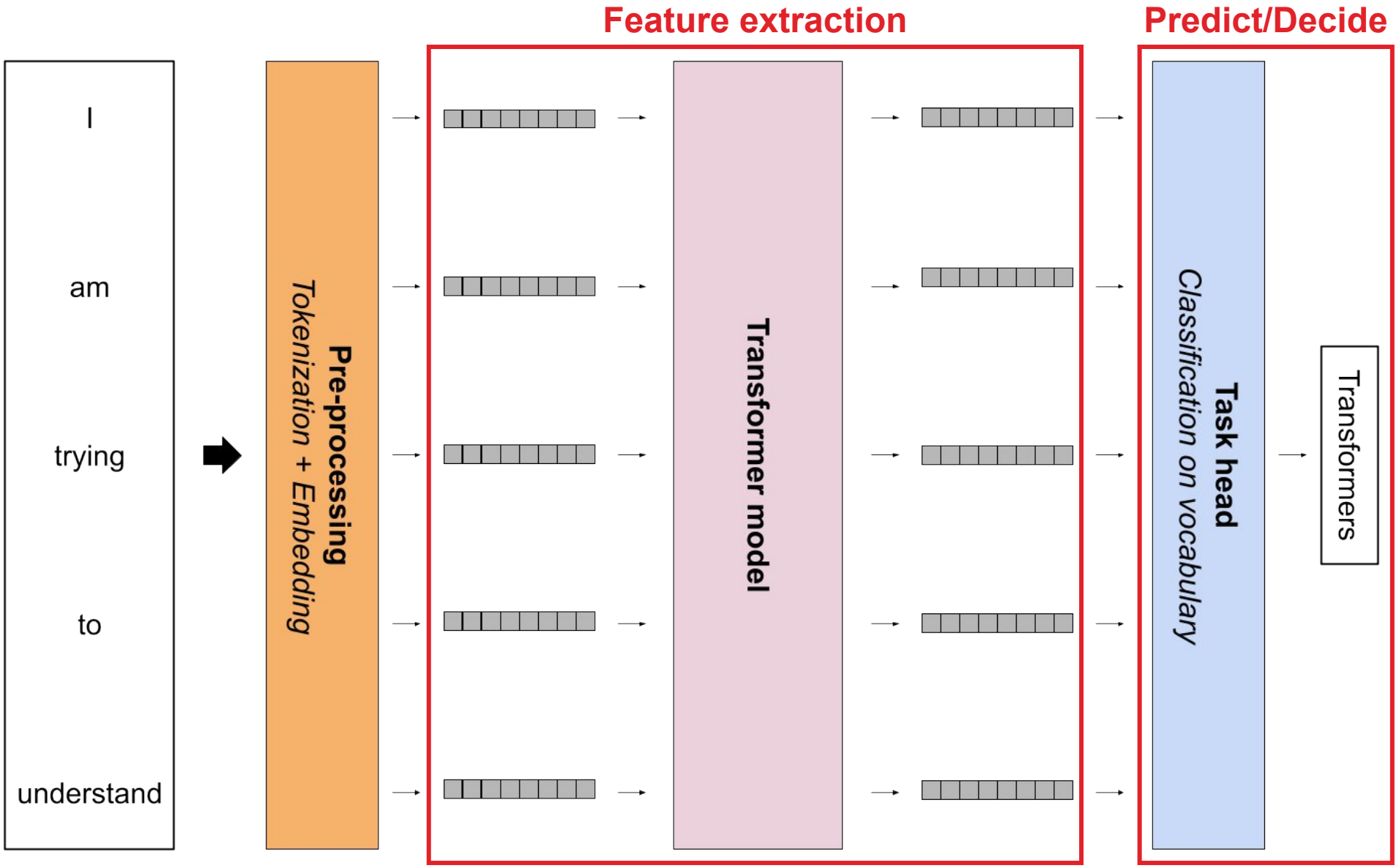
Example of NLP system



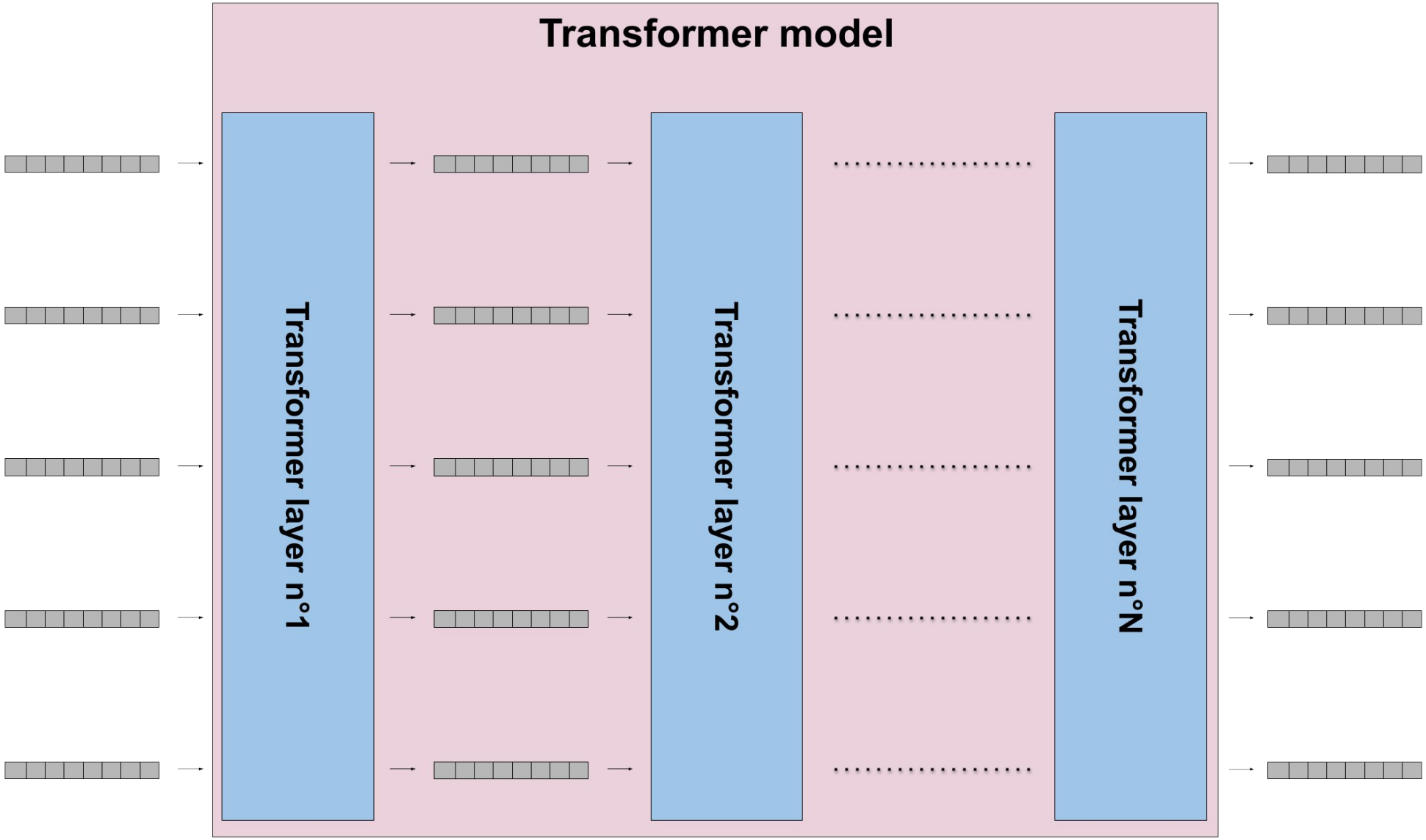
Feature extraction reminder



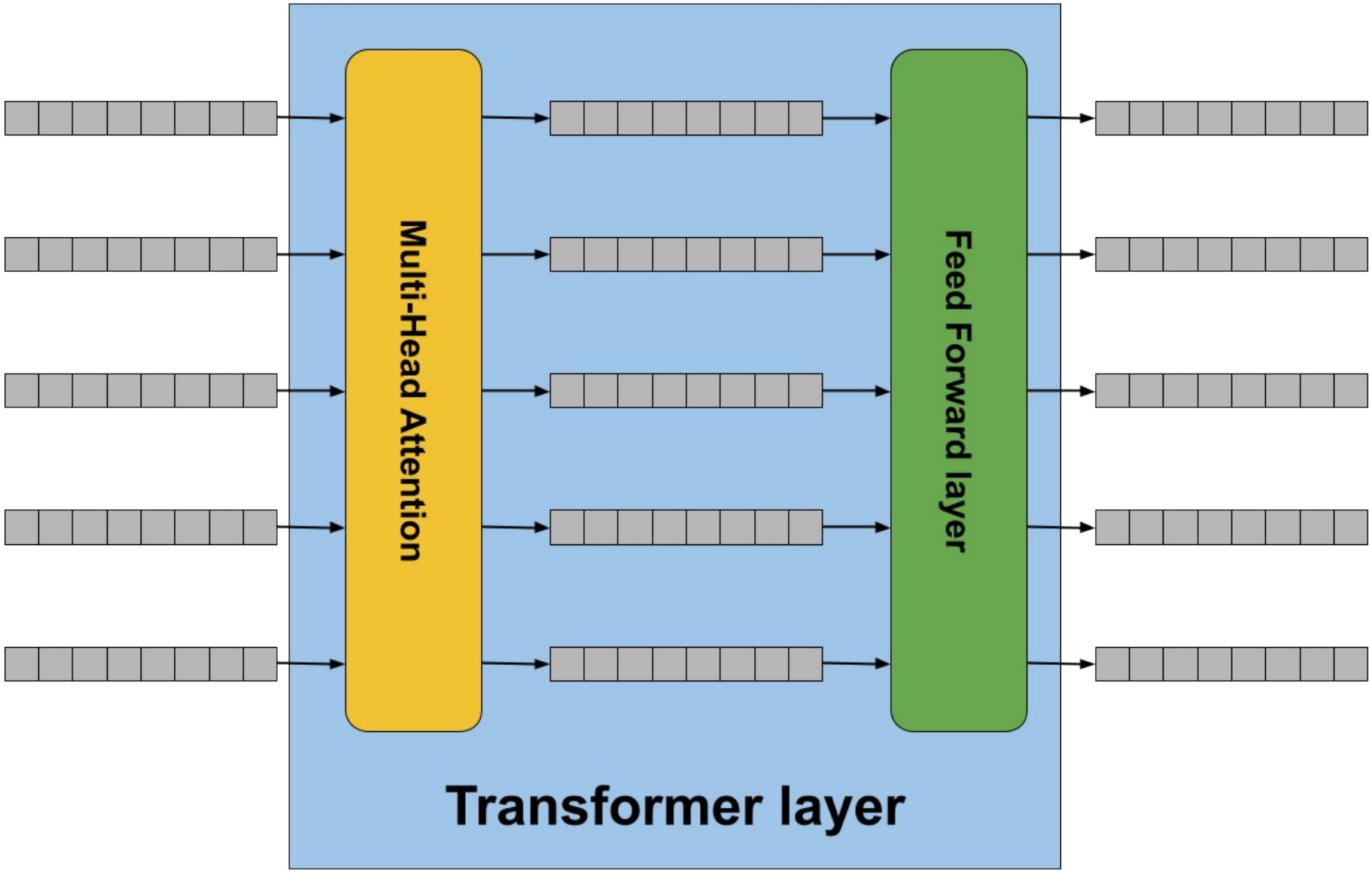
Feature extraction with Transformers



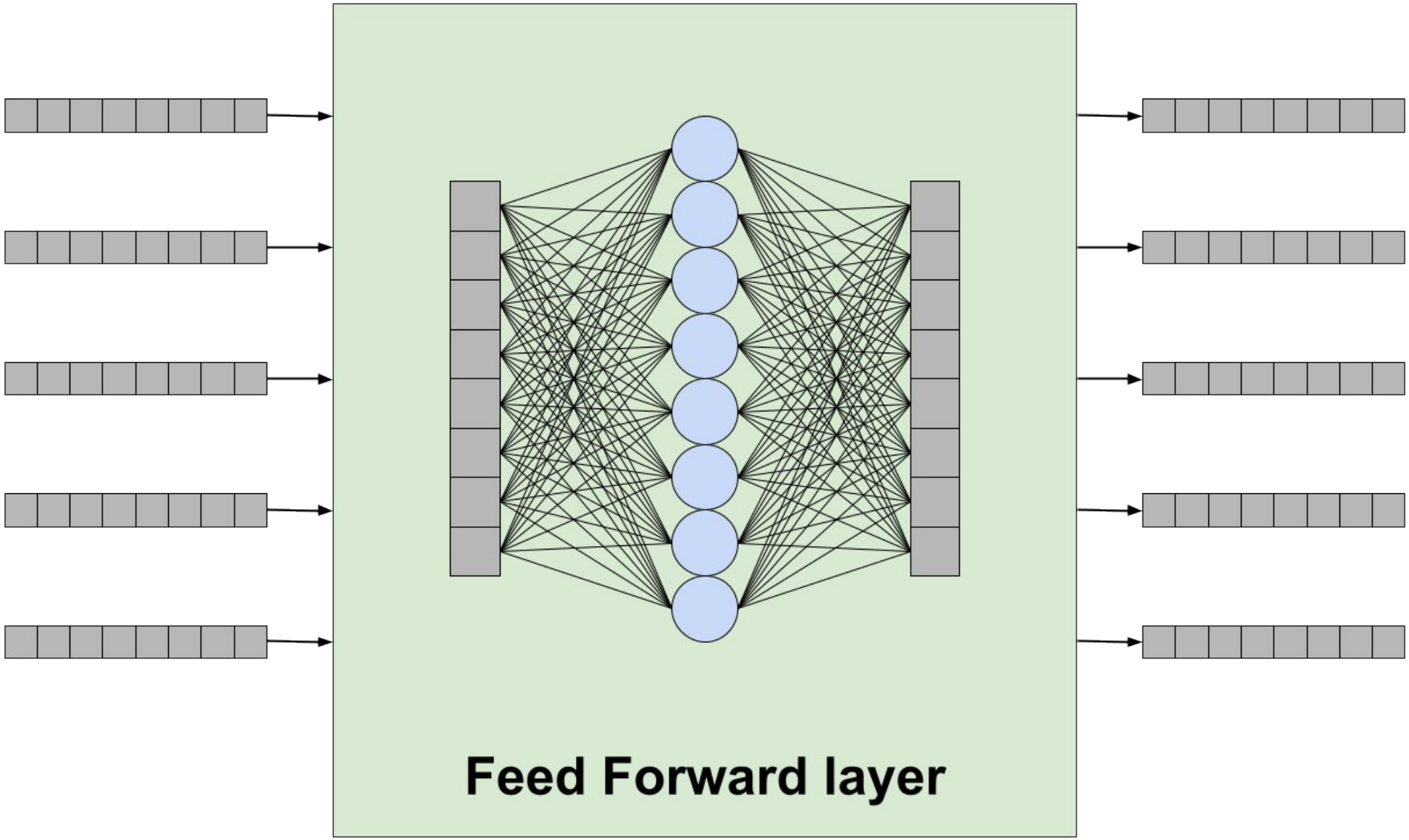
Vanilla Transformer architecture



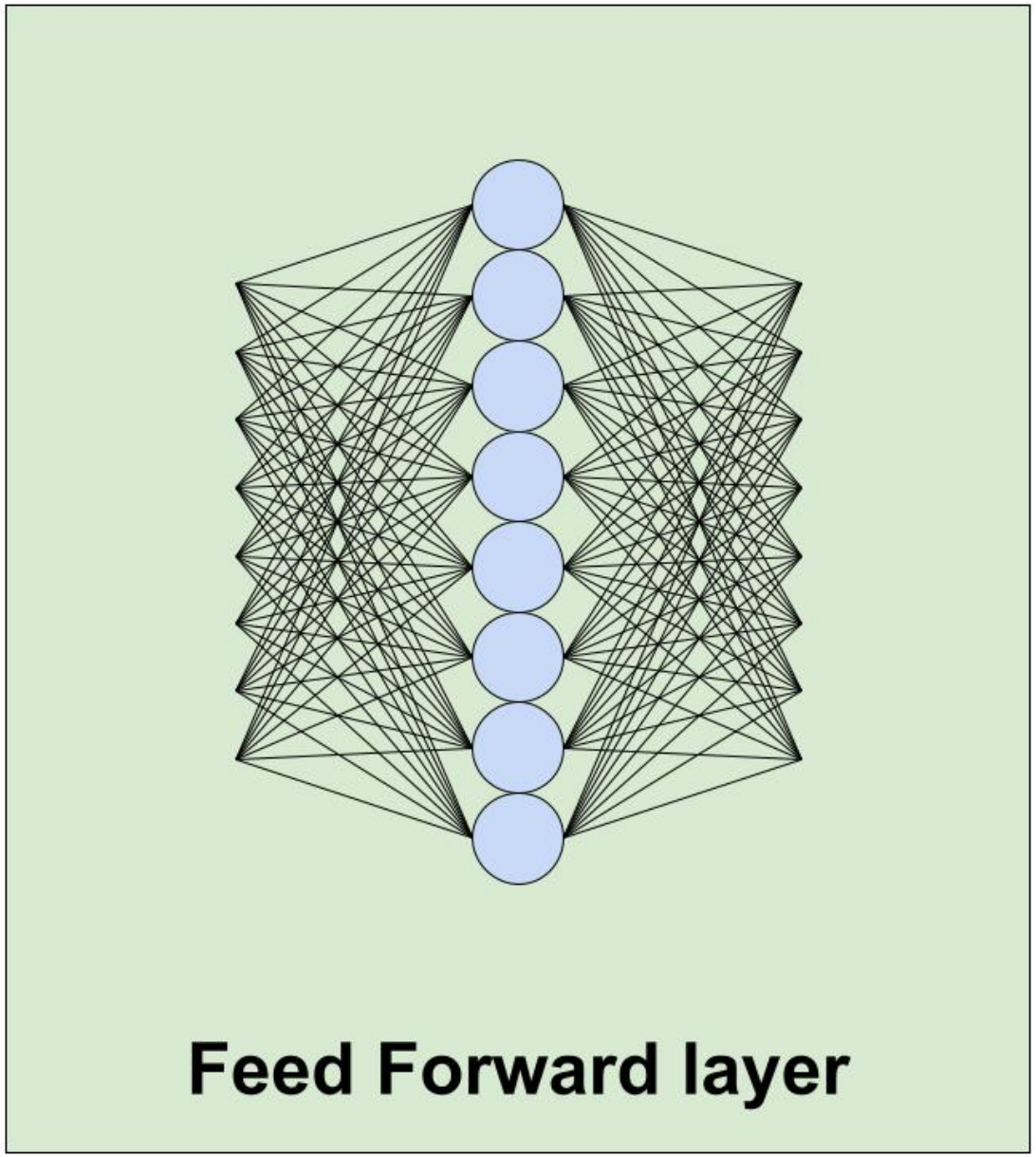
Transformer layer simplified



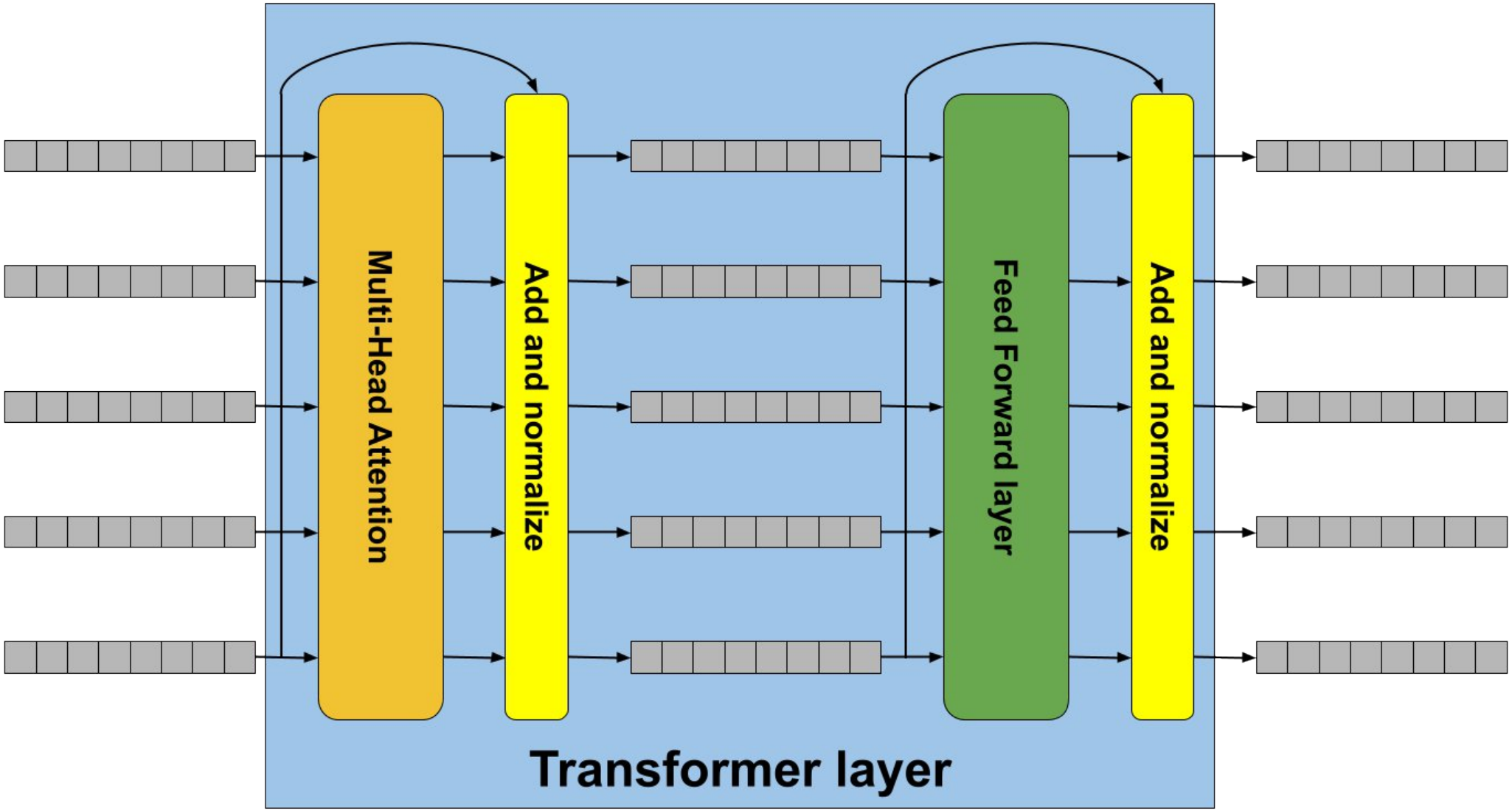
Feed Forward layer



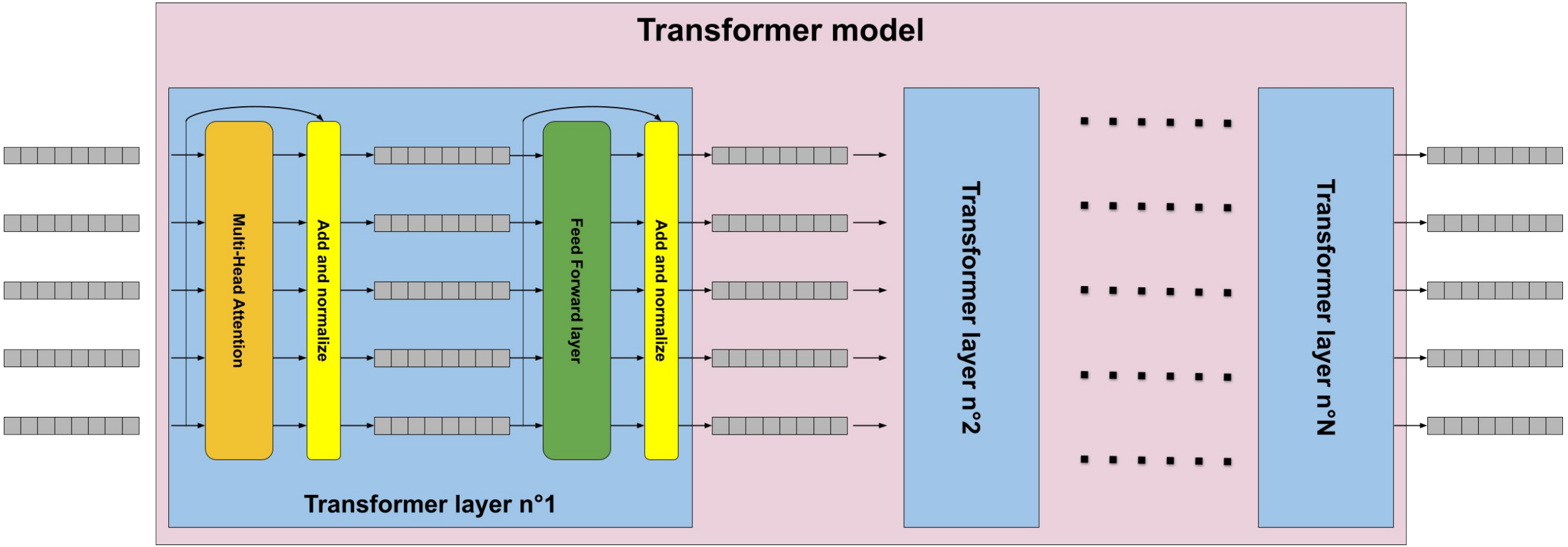
Feed Forward layer (animated)



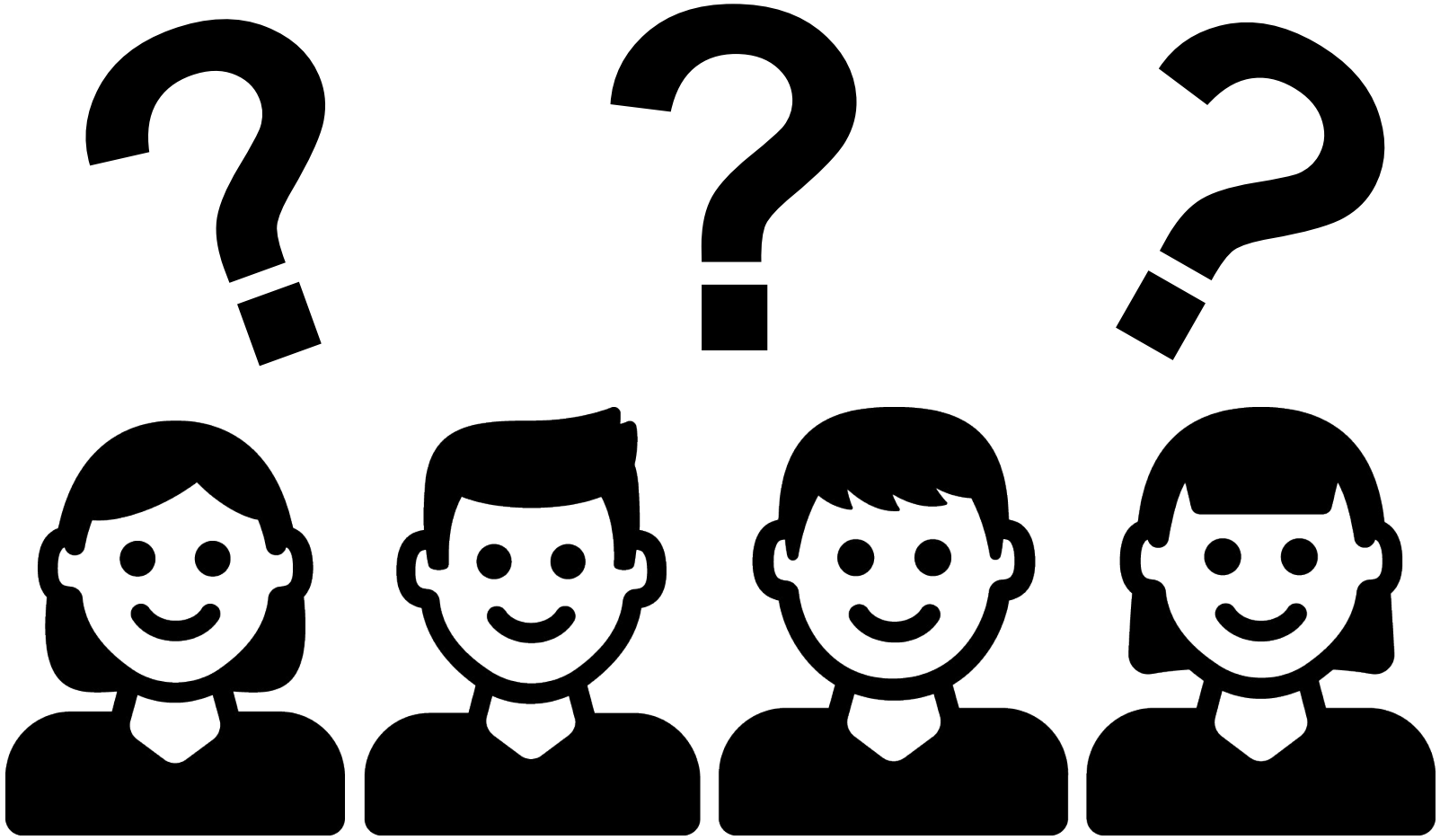
Transformer layer



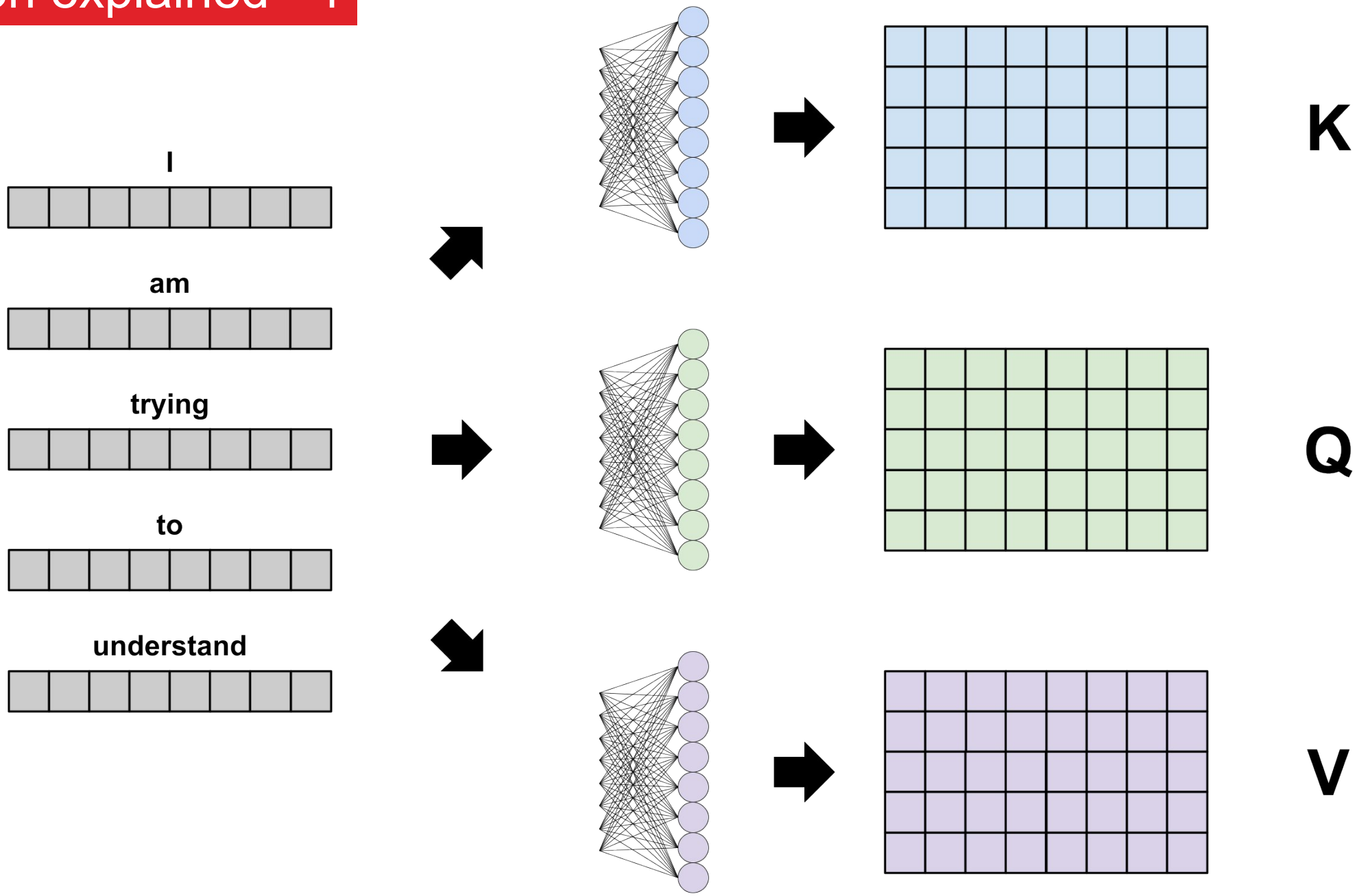
Vanilla Transformer architecture summary



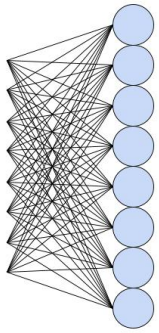
Question break #1

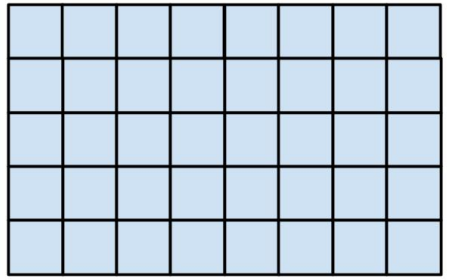


Attention explained - 1



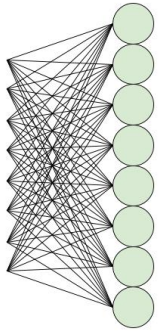
Attention explained - 1 (animated)

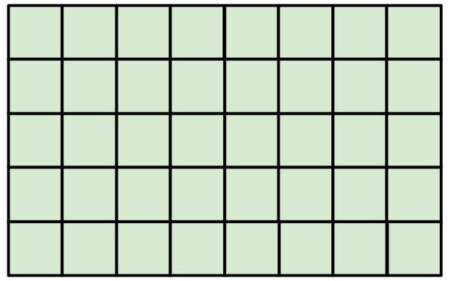




understand

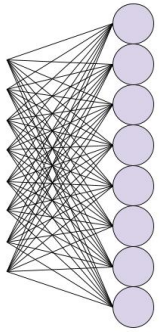
K

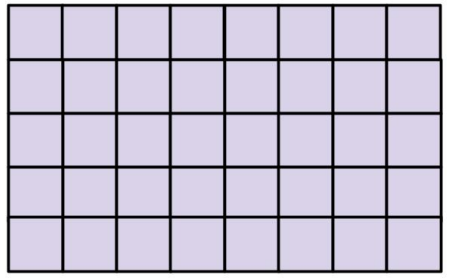




understand

Q

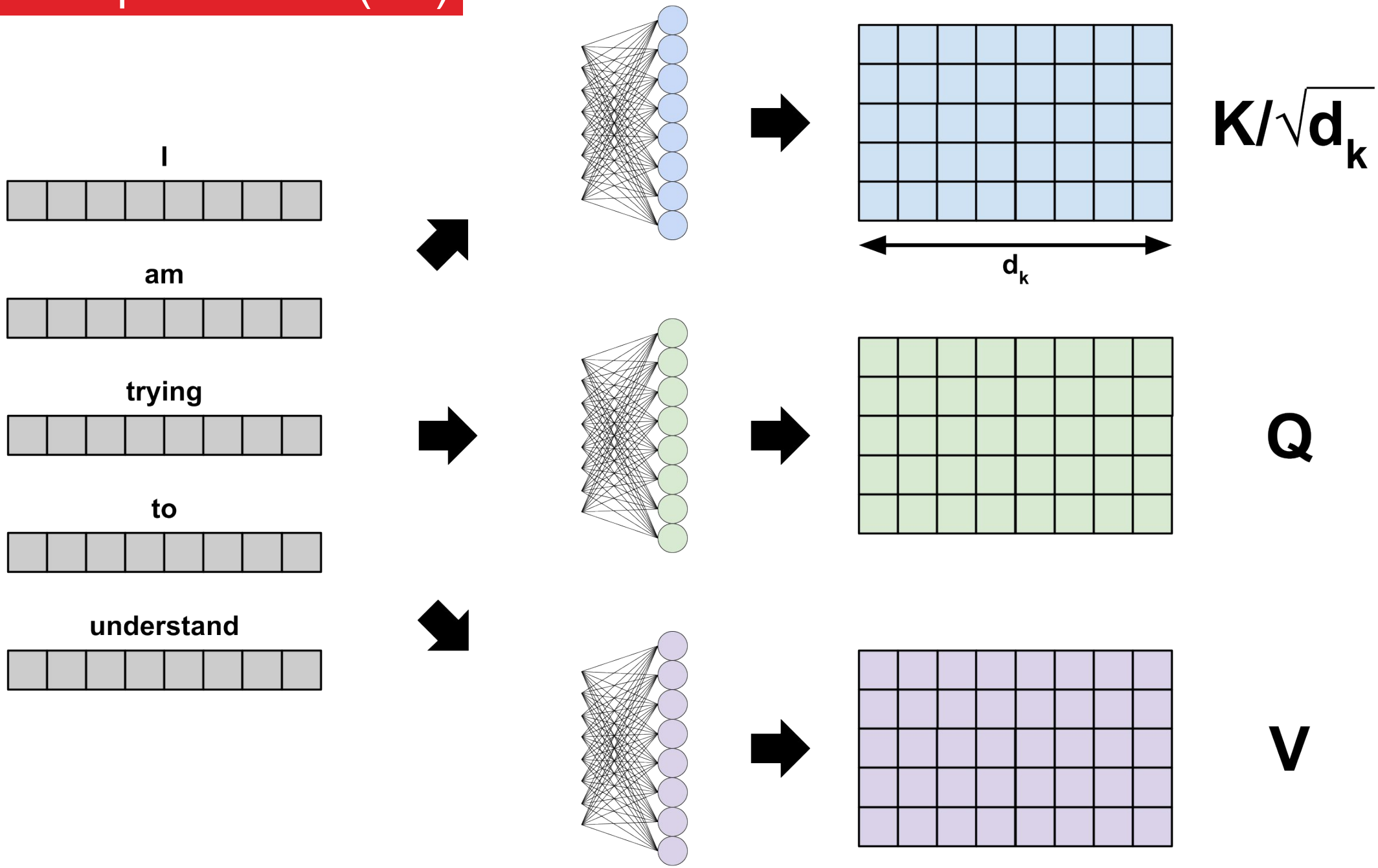




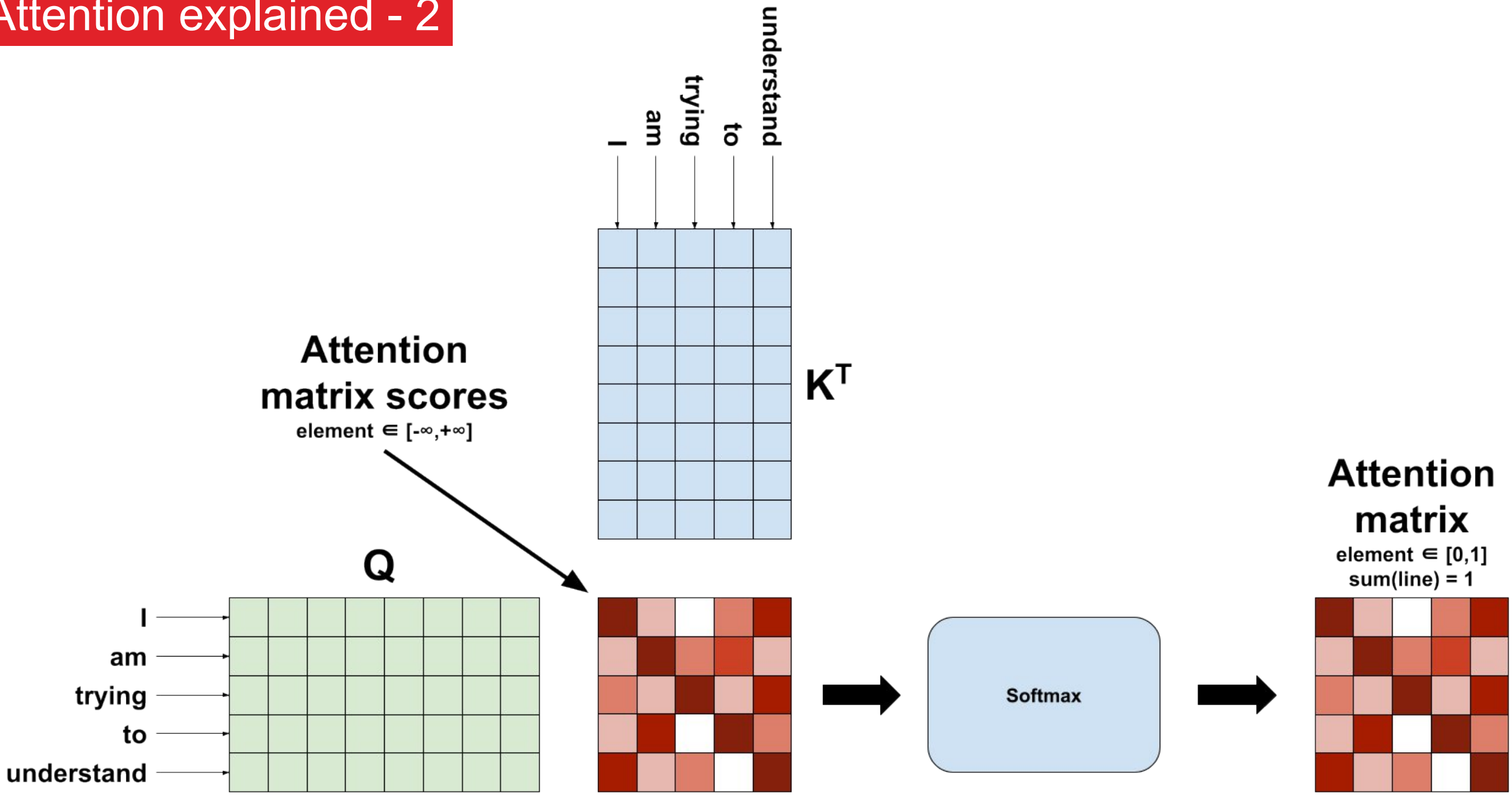
understand

V

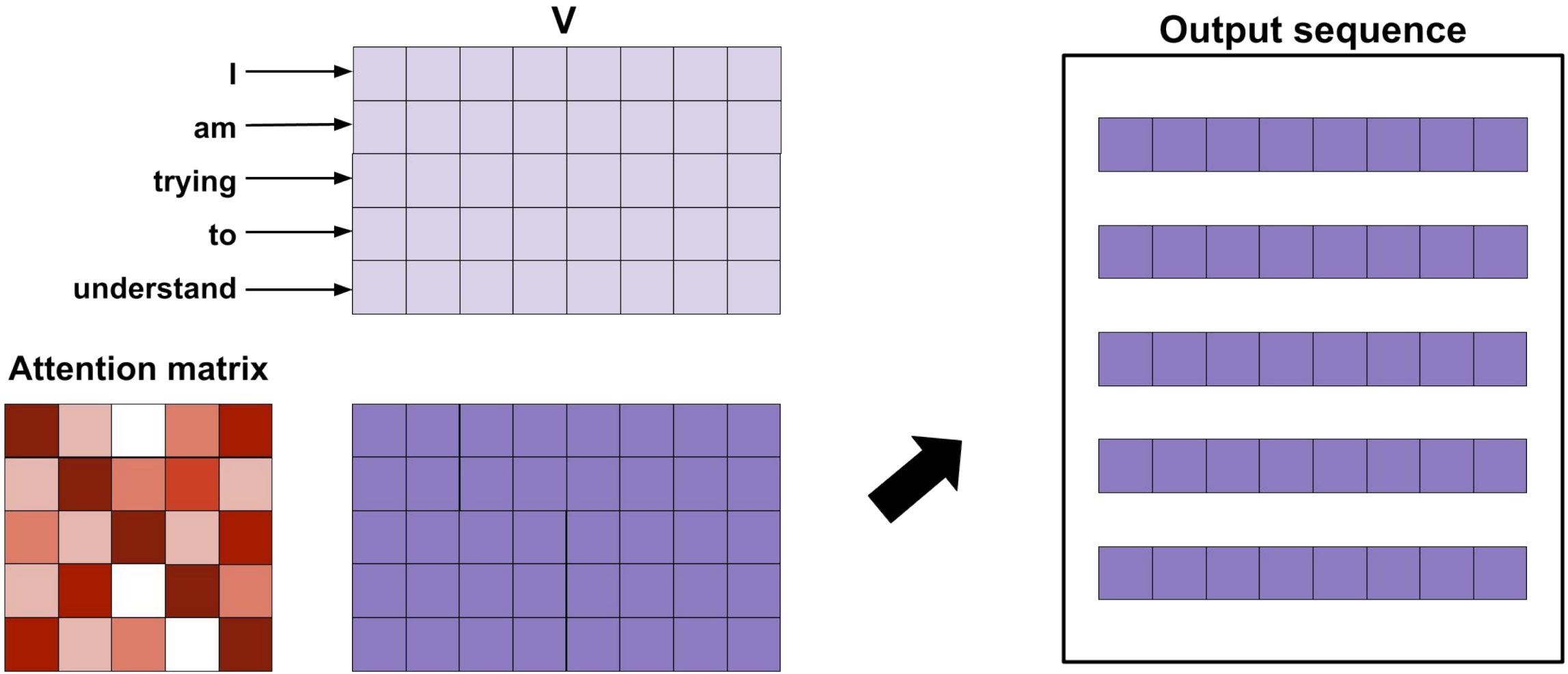
Attention explained - 1 (full)



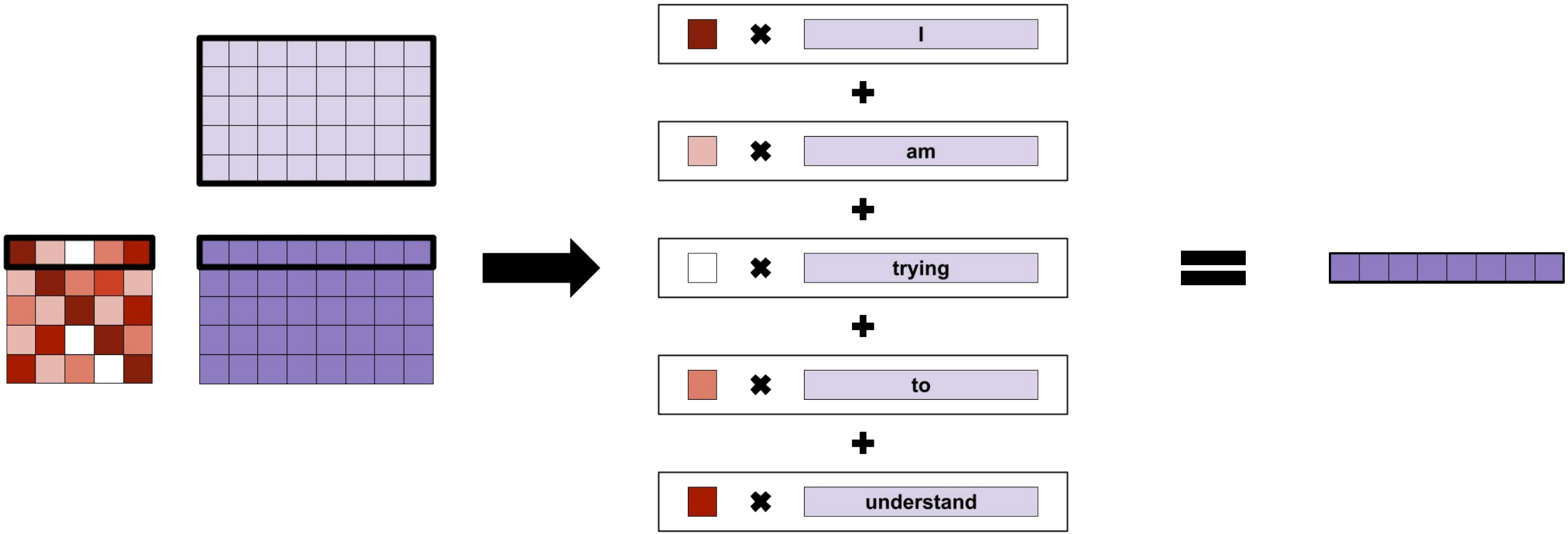
Attention explained - 2



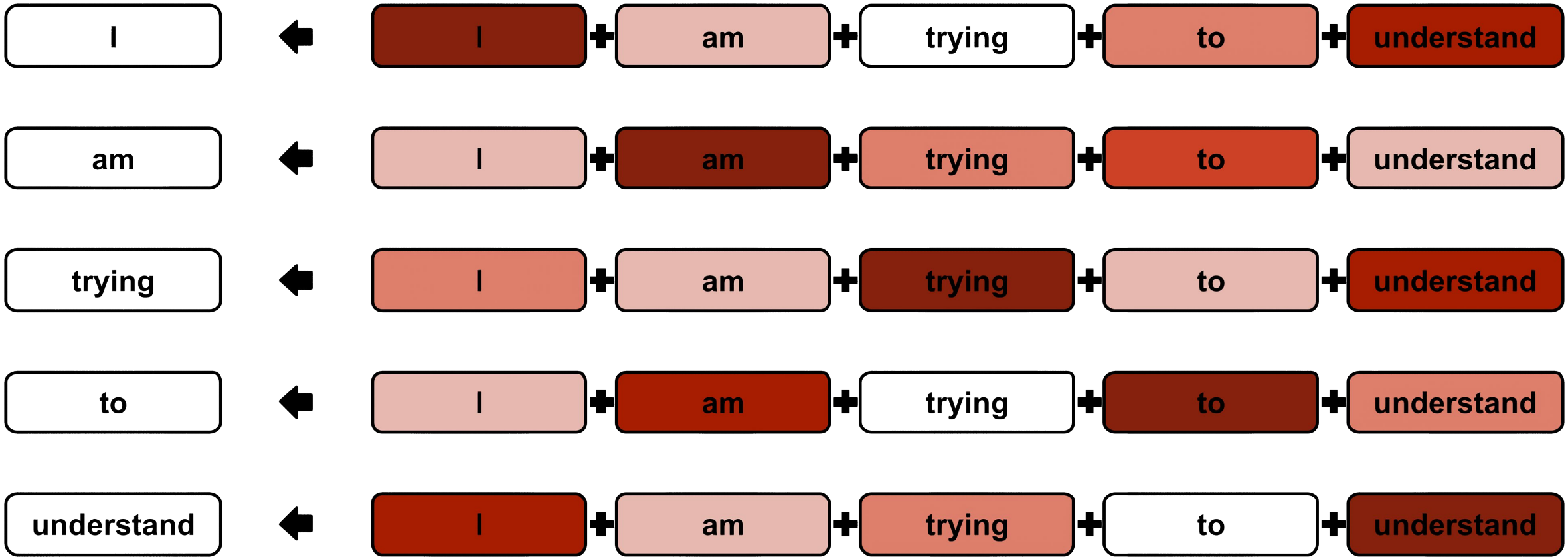
Attention explained - 3



Attention explained - 3 (detailed)



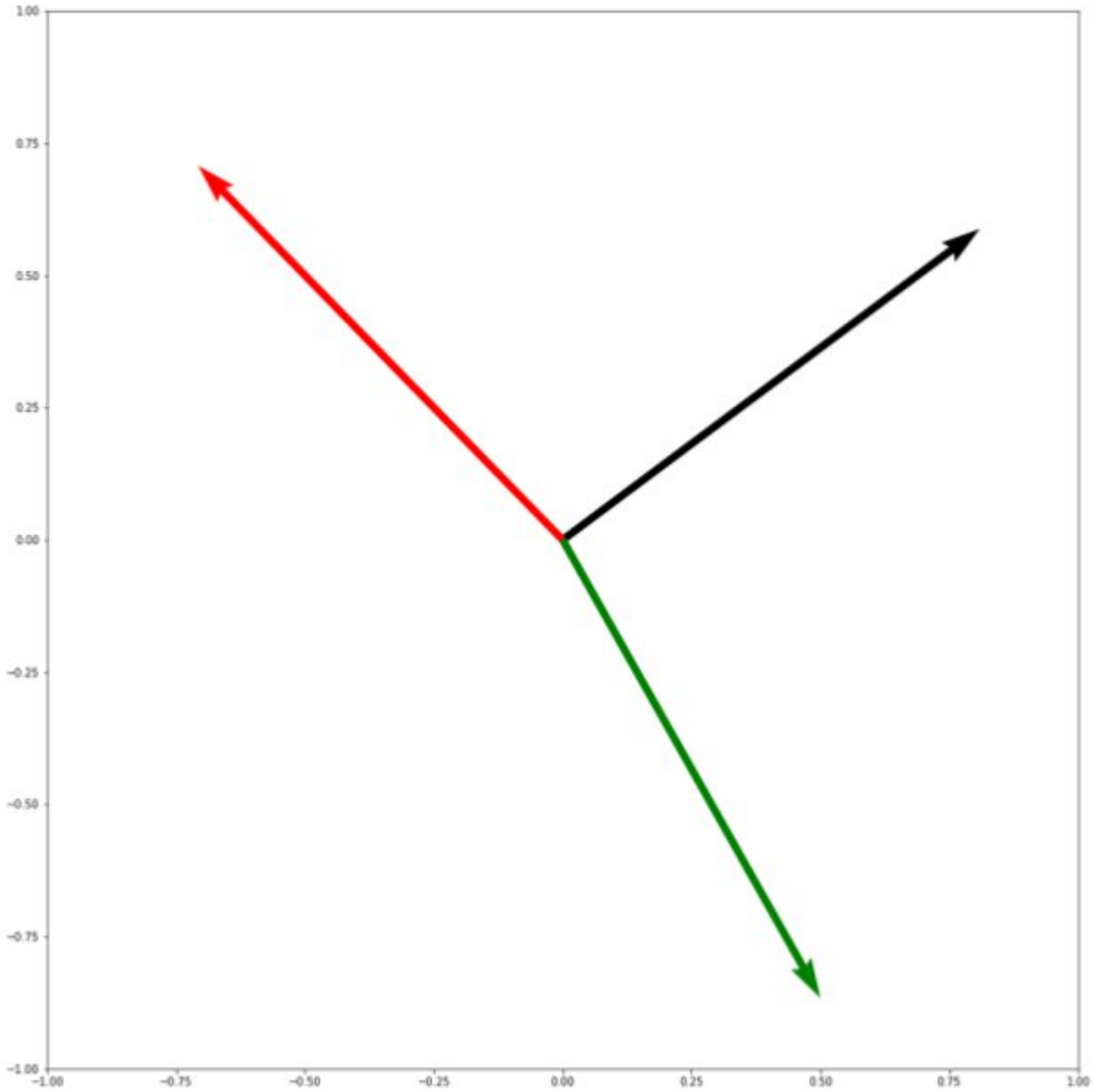
Intuition behind the attention mechanism



Intuition behind the attention mechanism - 0

The big dog

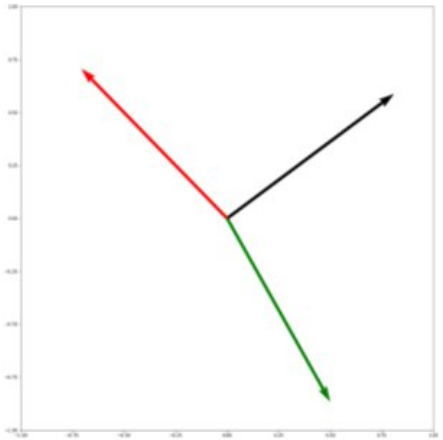
The : (0.50, -0.87)
big : (-0.70, 0.70)
dog : (0.81, 0.59)

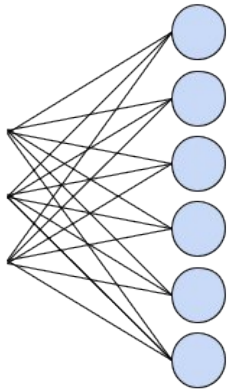


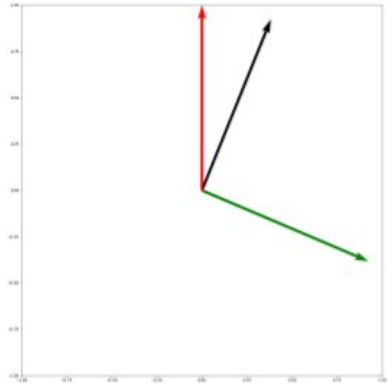
Intuition behind the attention mechanism - 1

0.50	-0.87
-0.70	0.70
0.81	0.59

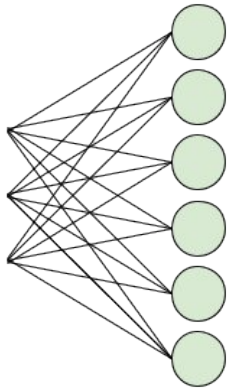
=

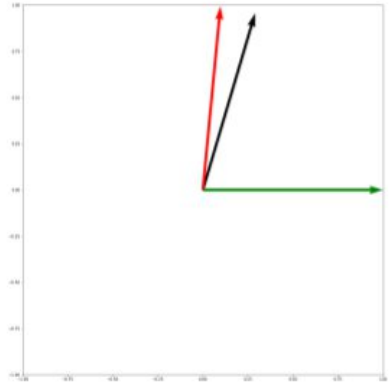




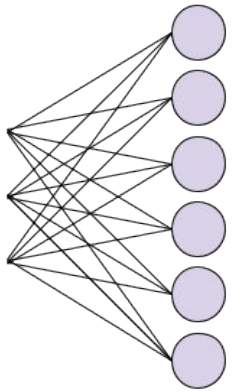


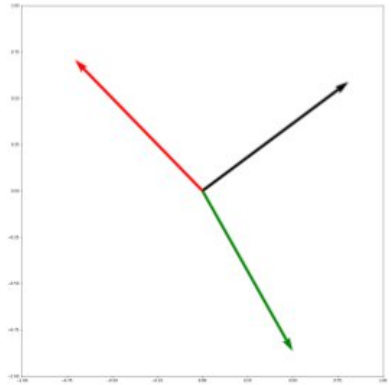
K





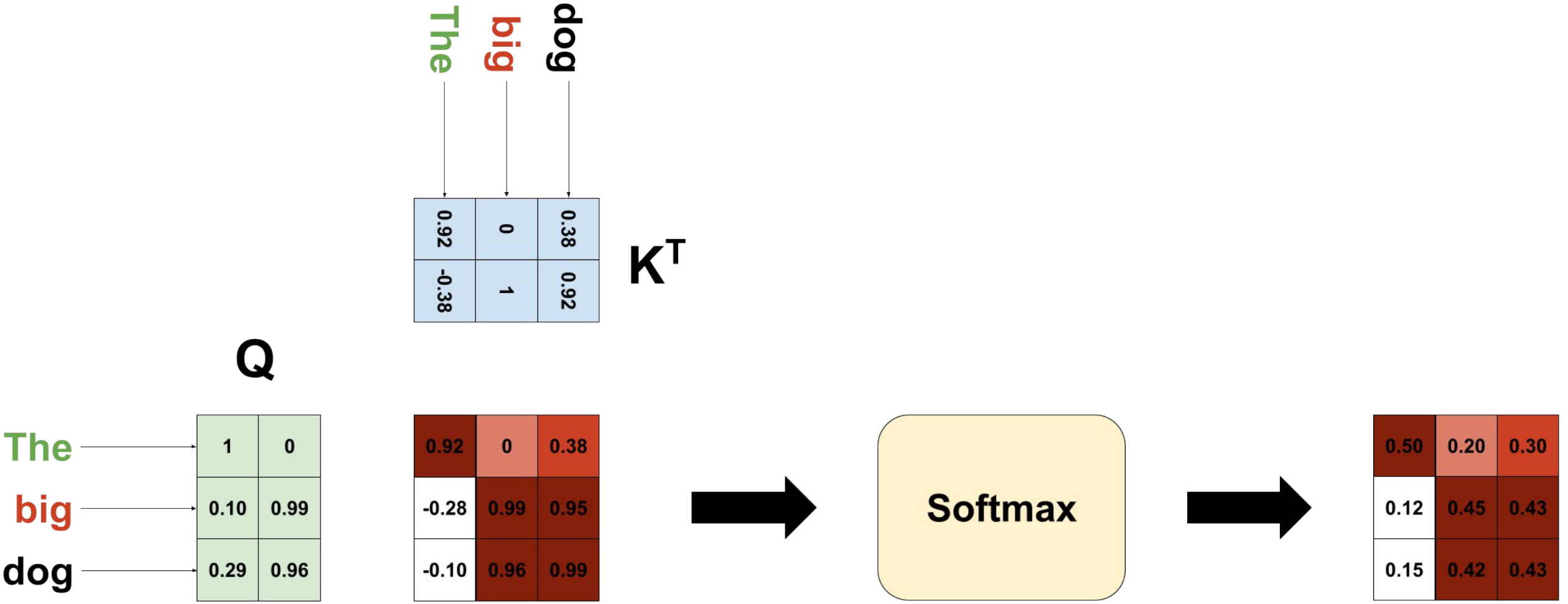
Q



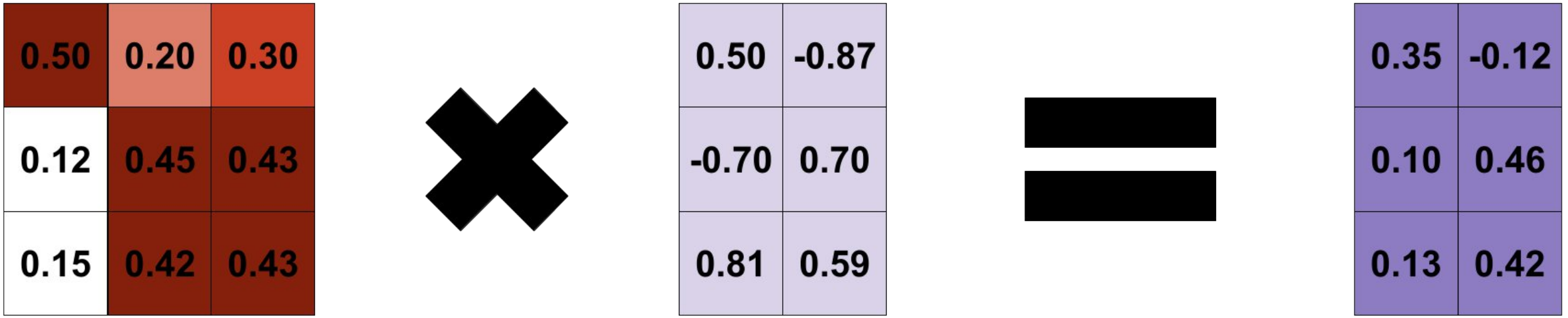


V

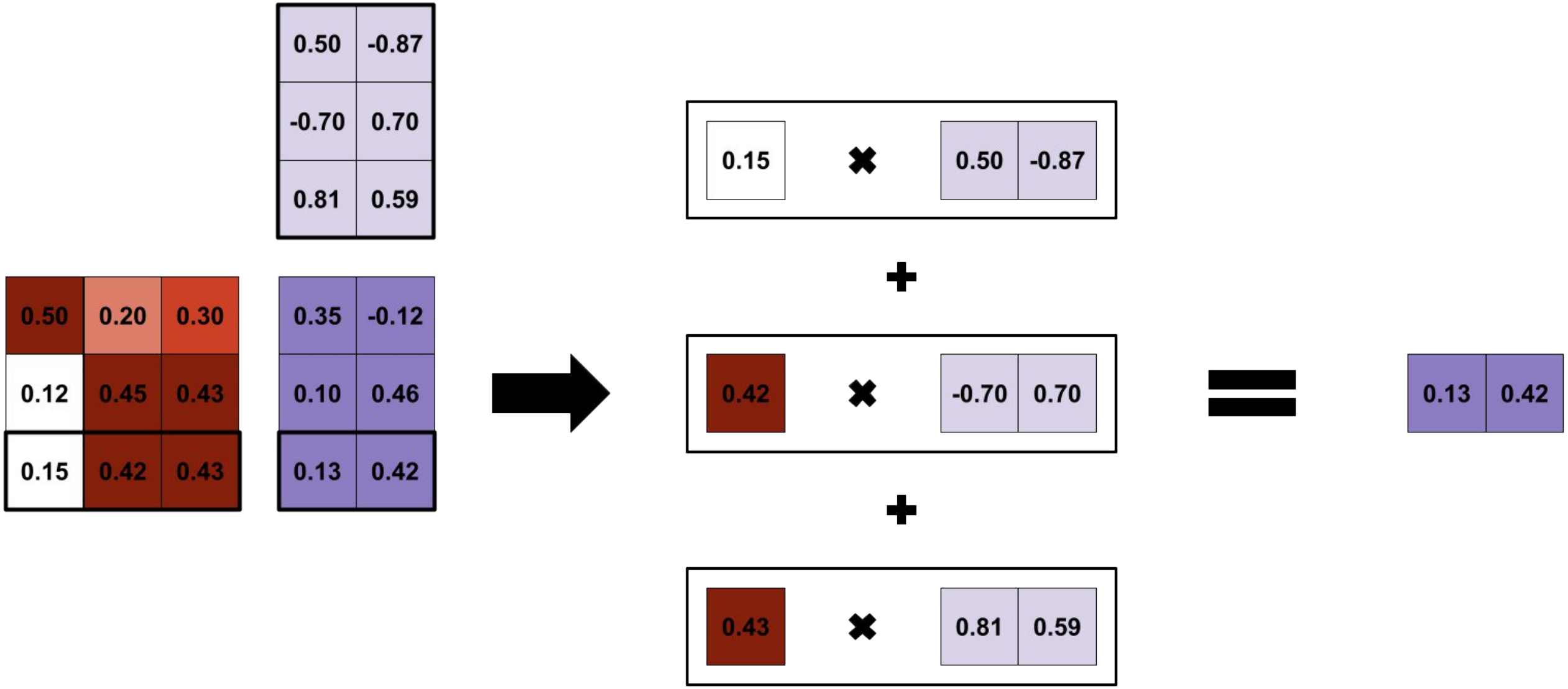
Intuition behind the attention mechanism - 2



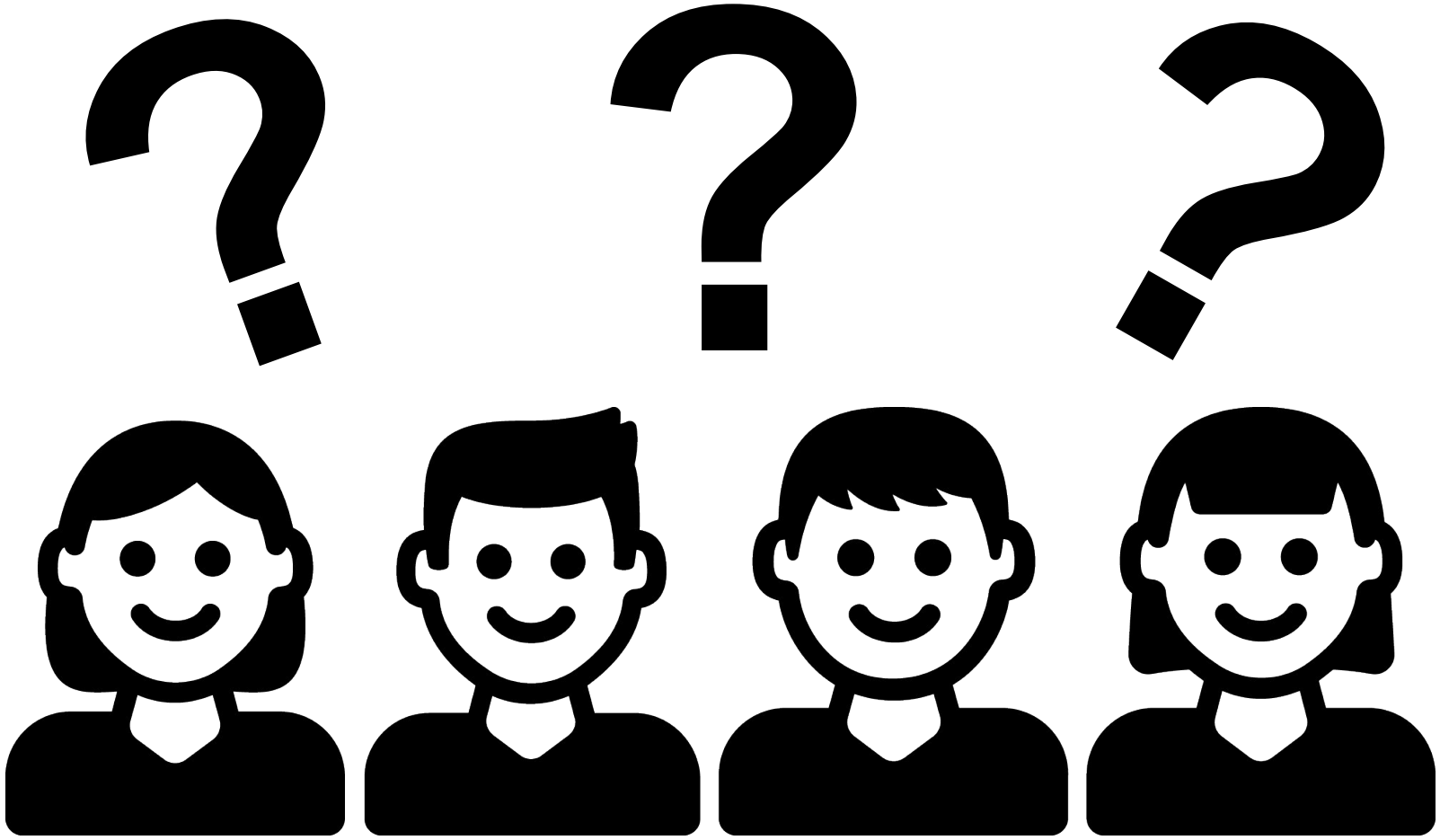
Intuition behind the attention mechanism - 3



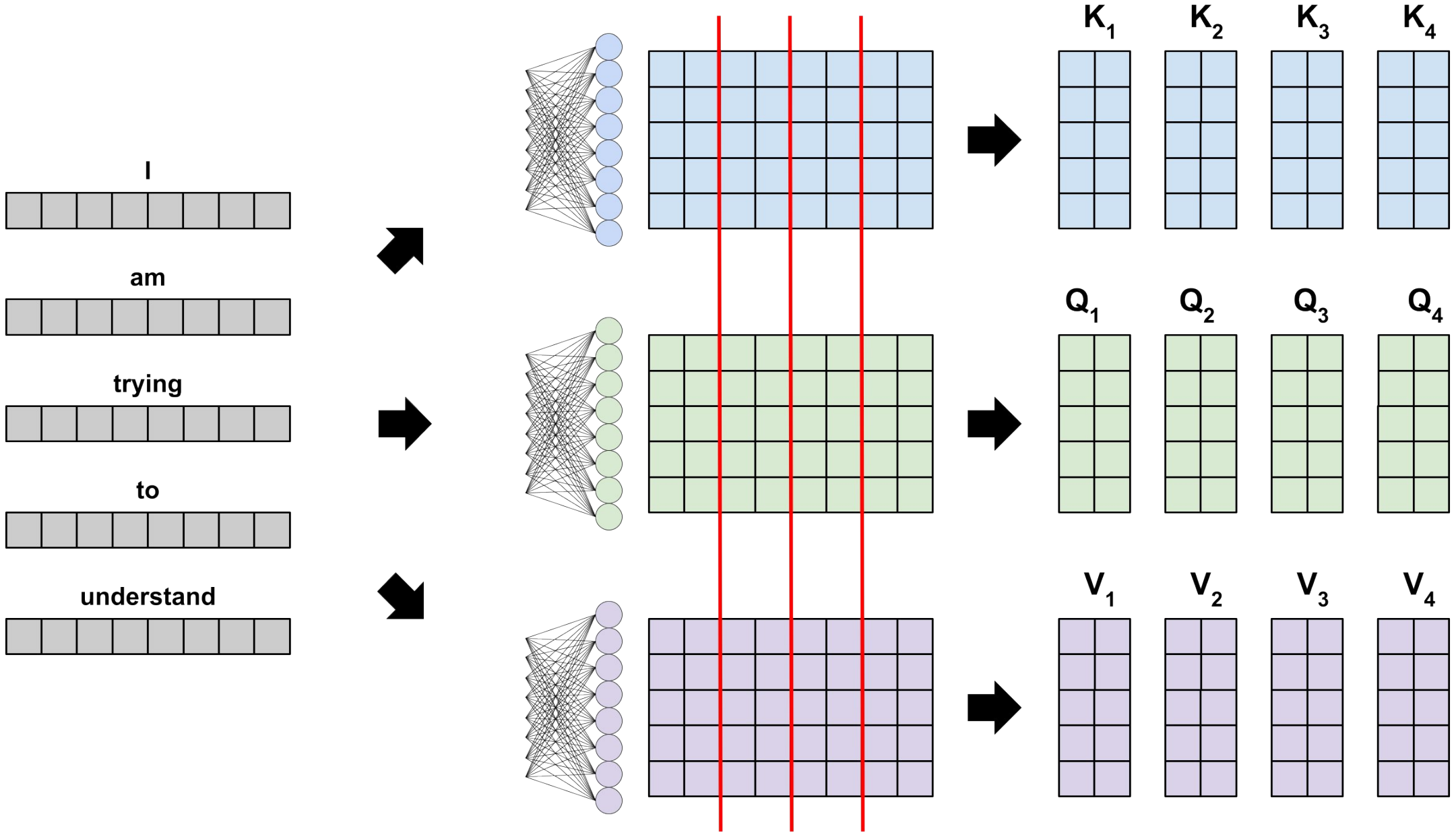
Intuition behind the attention mechanism - 3 (detailed)



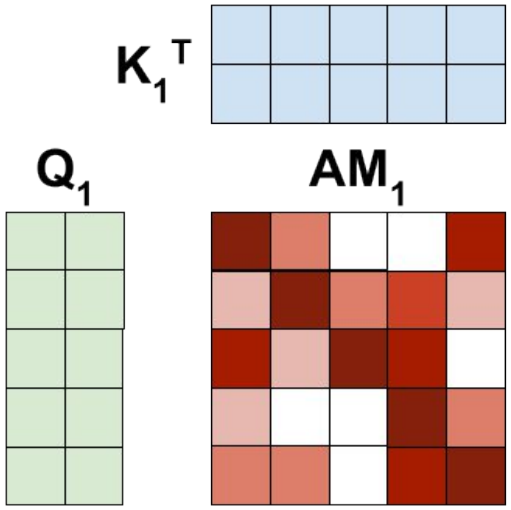
Question break #2

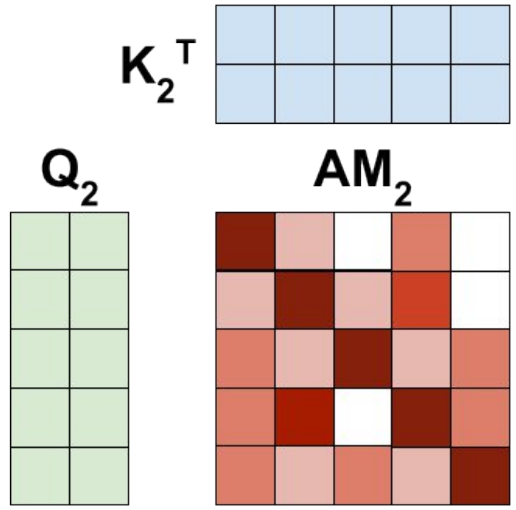


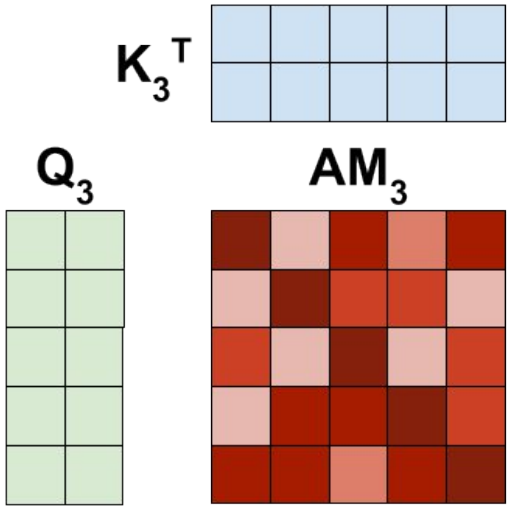
Multi-head Attention explained - 1

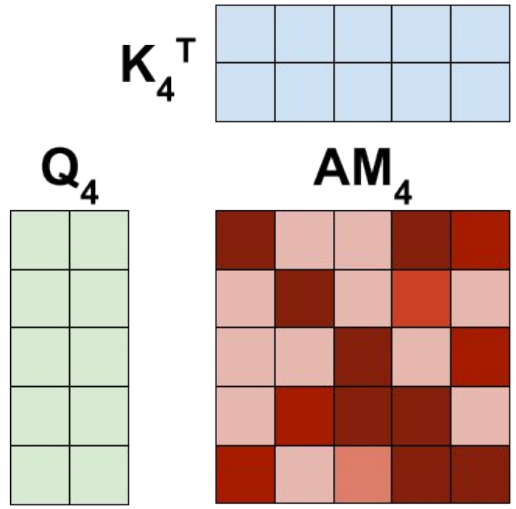


Multi-head Attention explained - 2

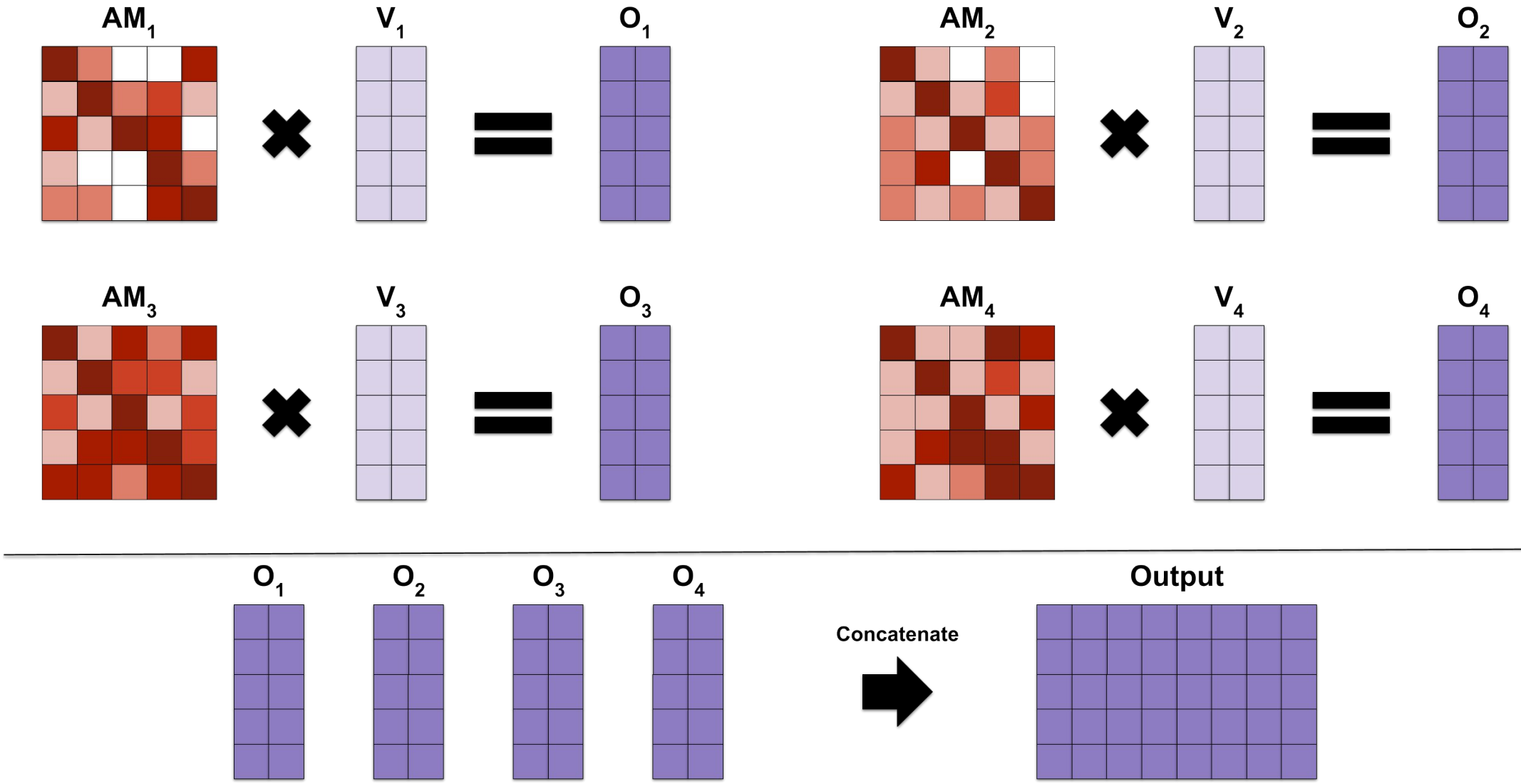




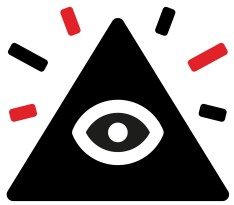




Multi-head Attention explained - 3



9



«Attention is All You Need»
Transformers

1. What is a **Transformer**? The magic of the **Attention Mechanism**
2. The different **Transformers architectures**
3. **Pre-training** and **Foundation Models**
4. **Specialization** of **Foundation Models** (especially LLM)
5. **Example: IMDB Reviews**

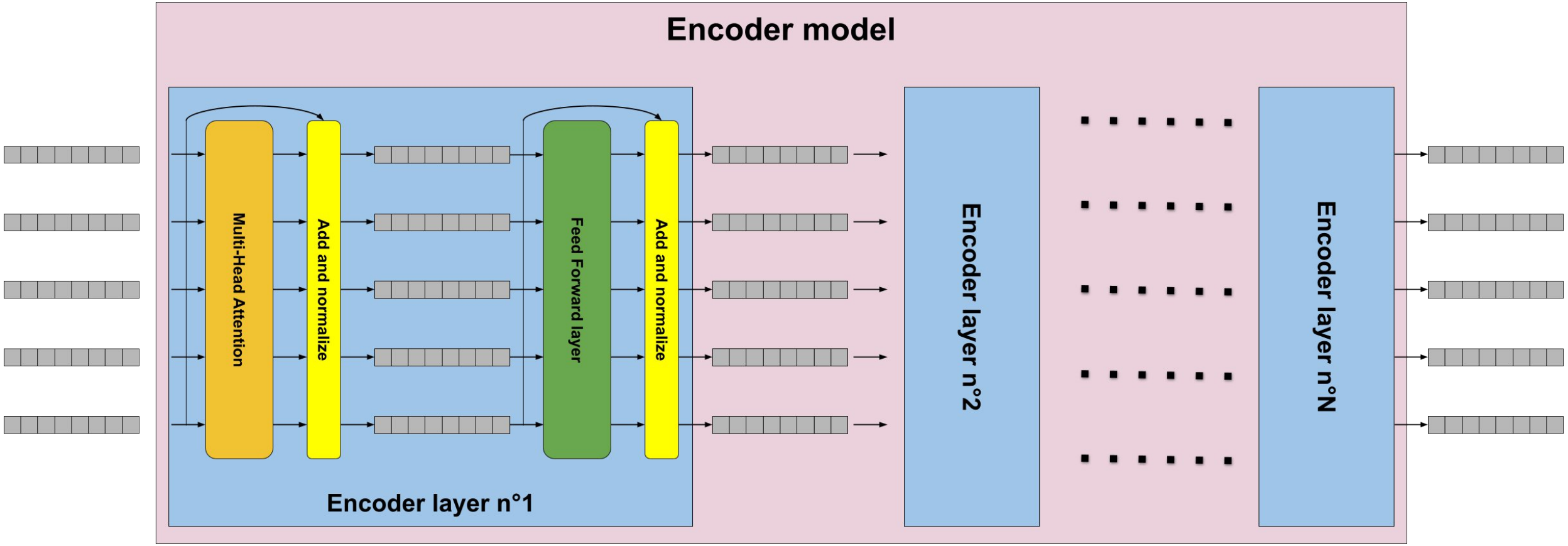
Encoder-Decoder

Encoder

Decoder

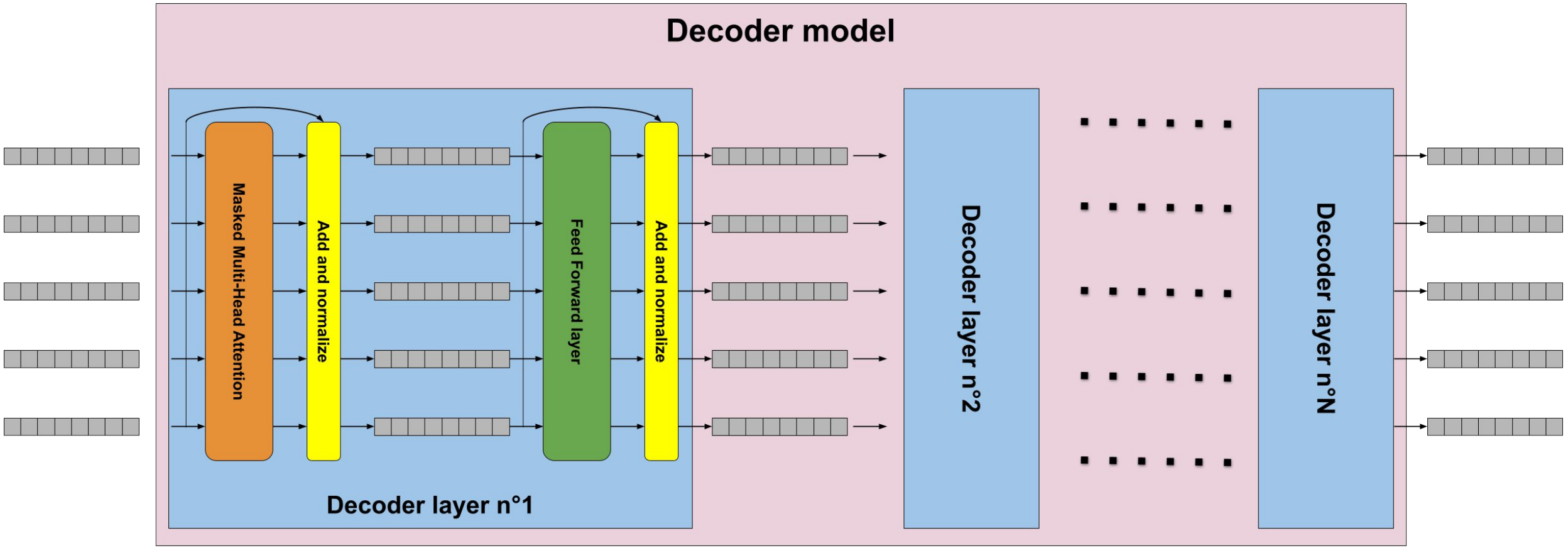
Encoder architecture

BERT / Encoder / Auto-encoding



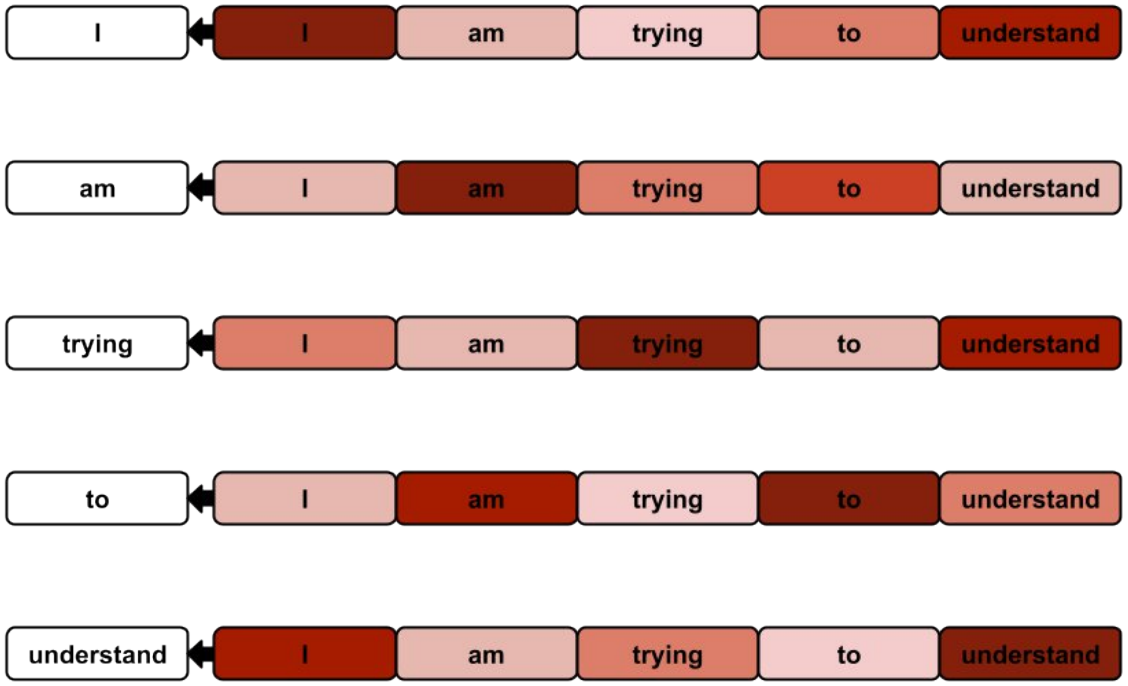
Decoder architecture

GPT / Decoder / Auto-regressive



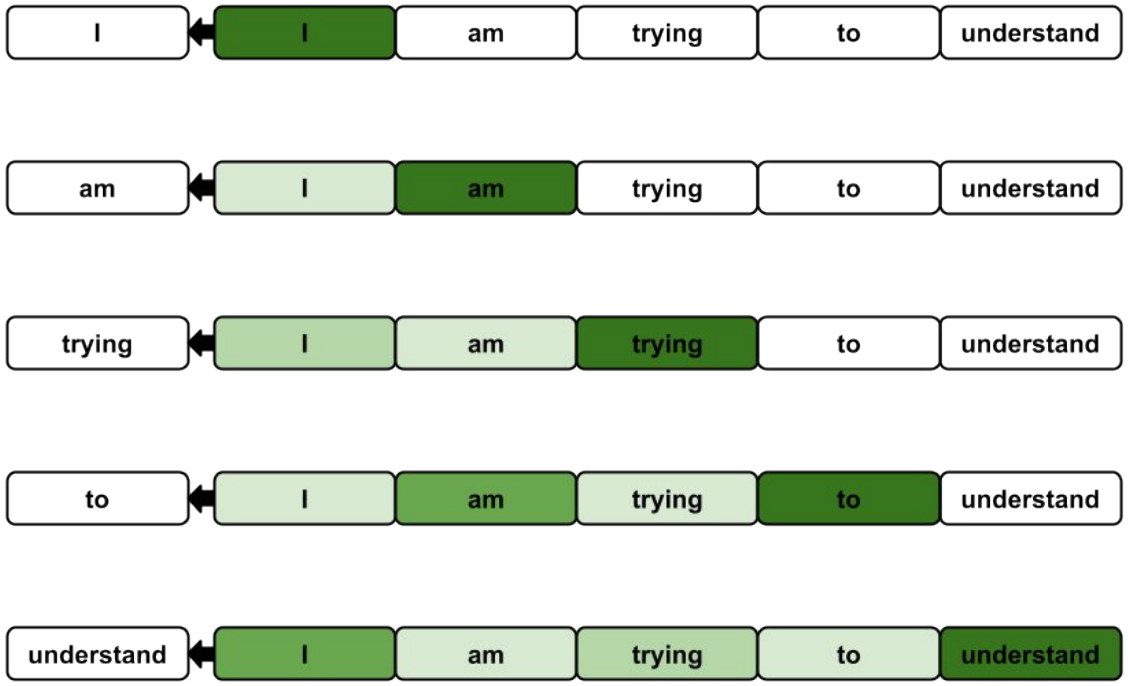
Bidirectional vs Unidirectional attention

Bidirectional attention for Encoder

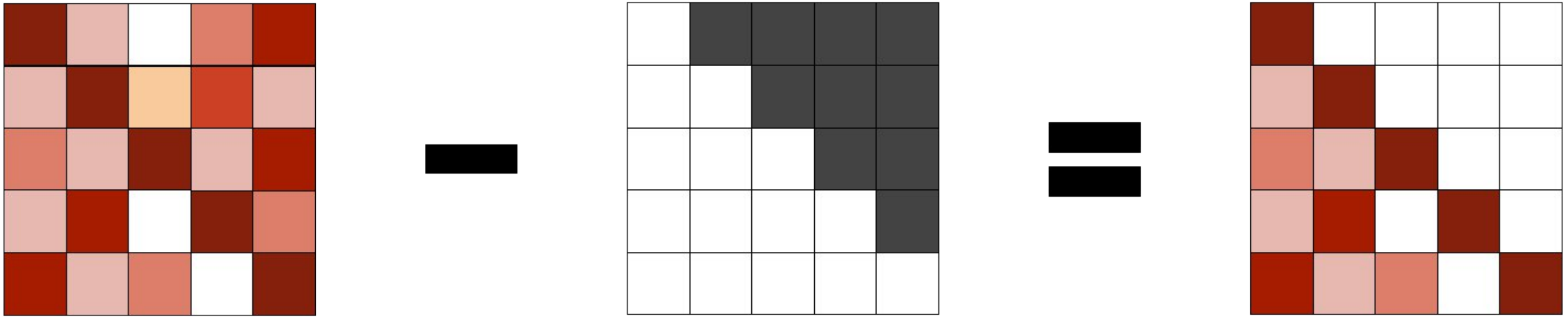


VS

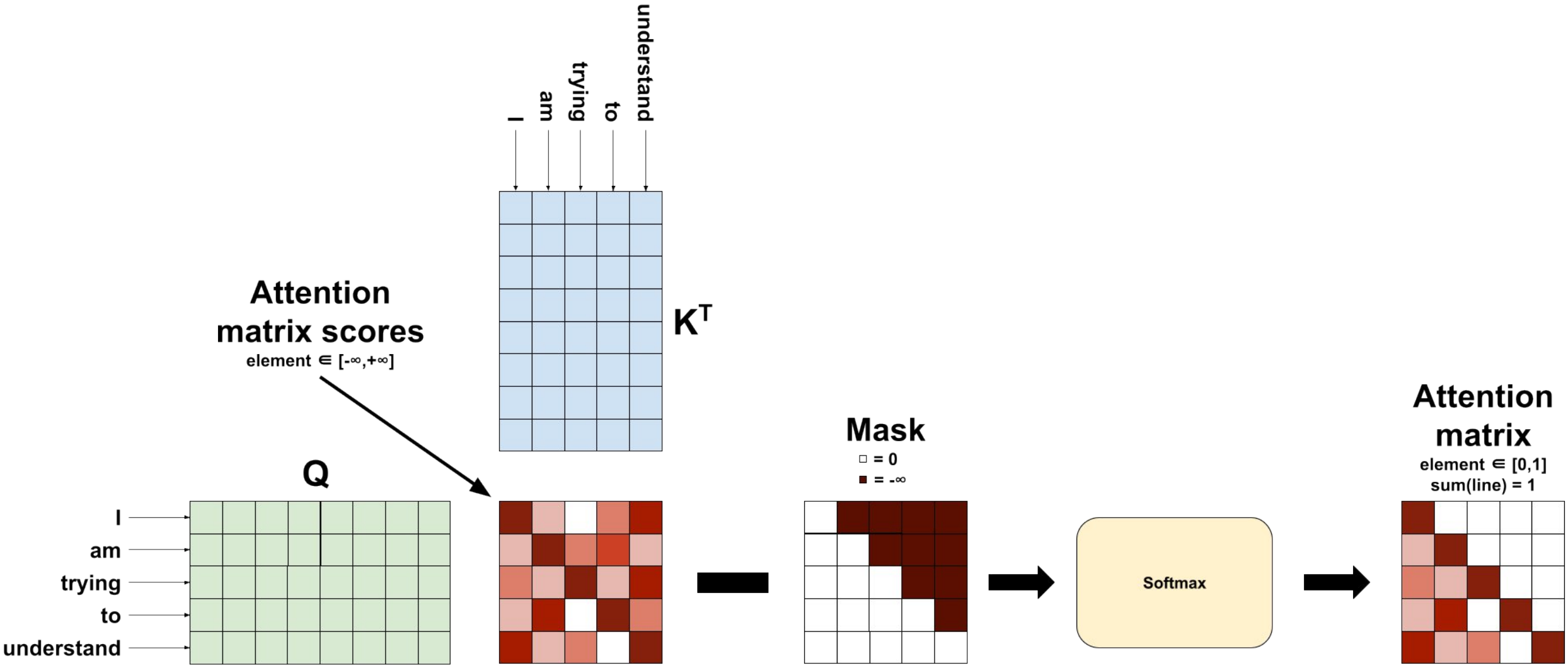
Unidirectional attention for Decoder



Mask attention

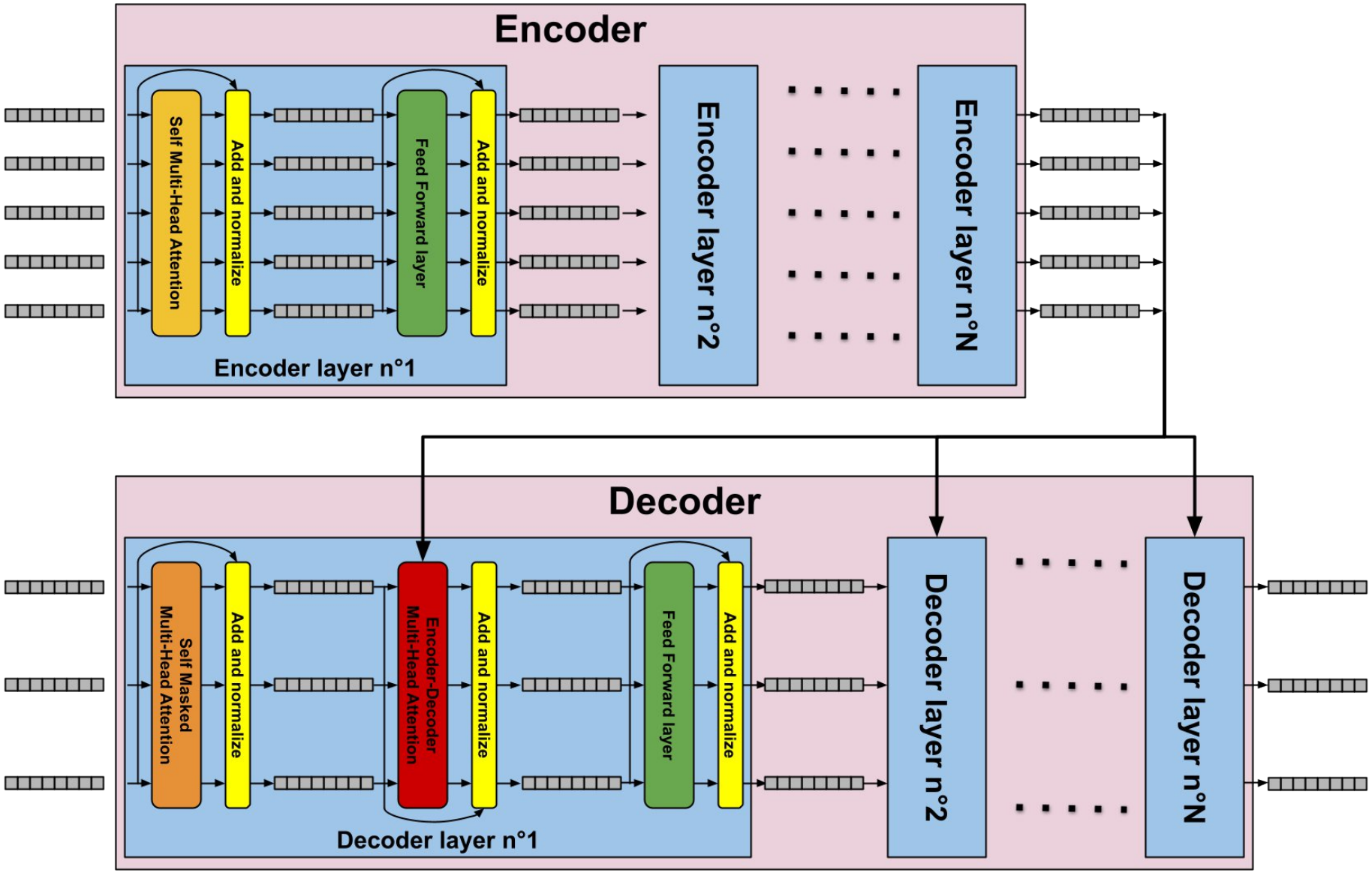


Unidirectional attention detailed

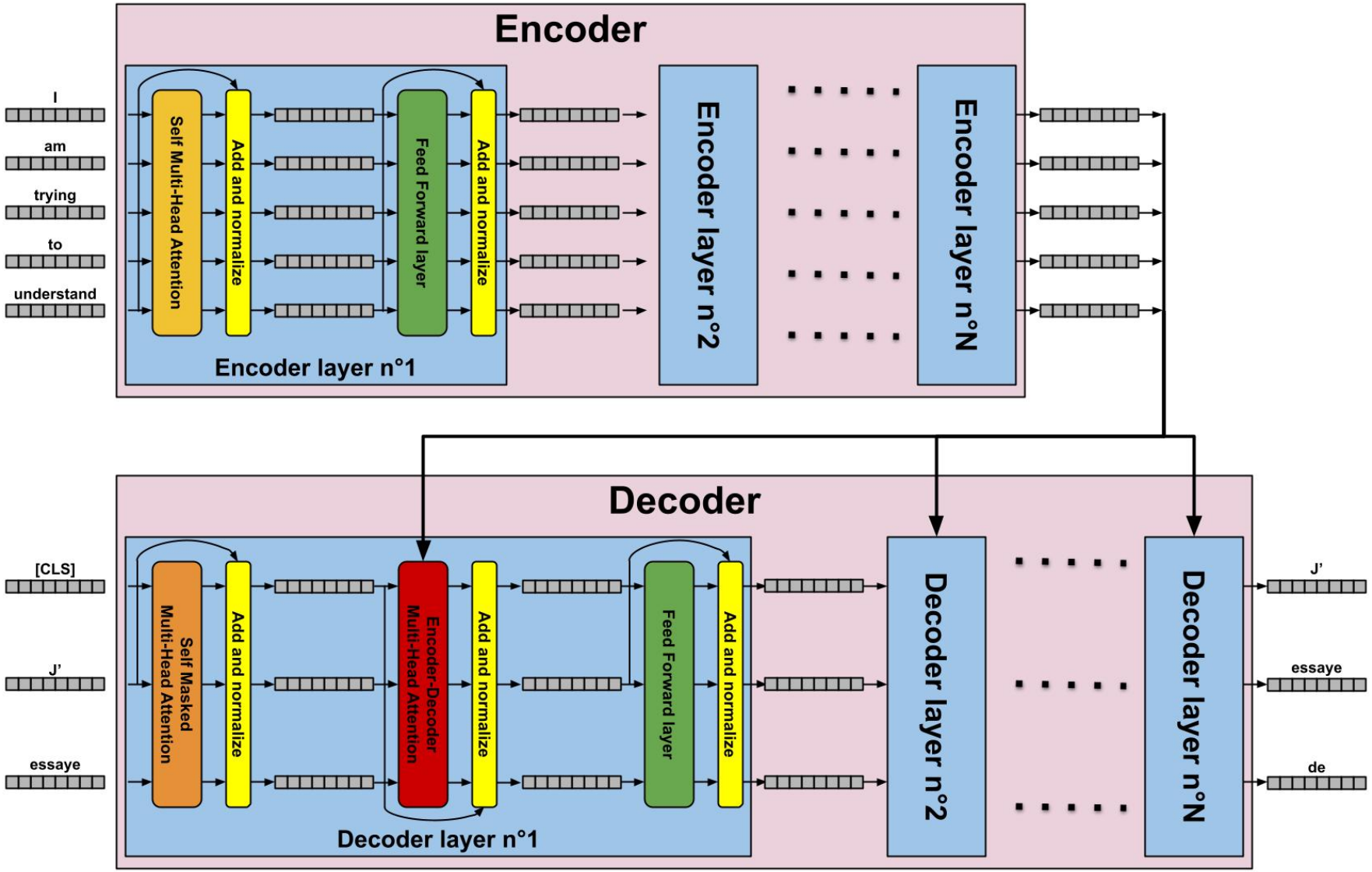


Encoder-Decoder architecture

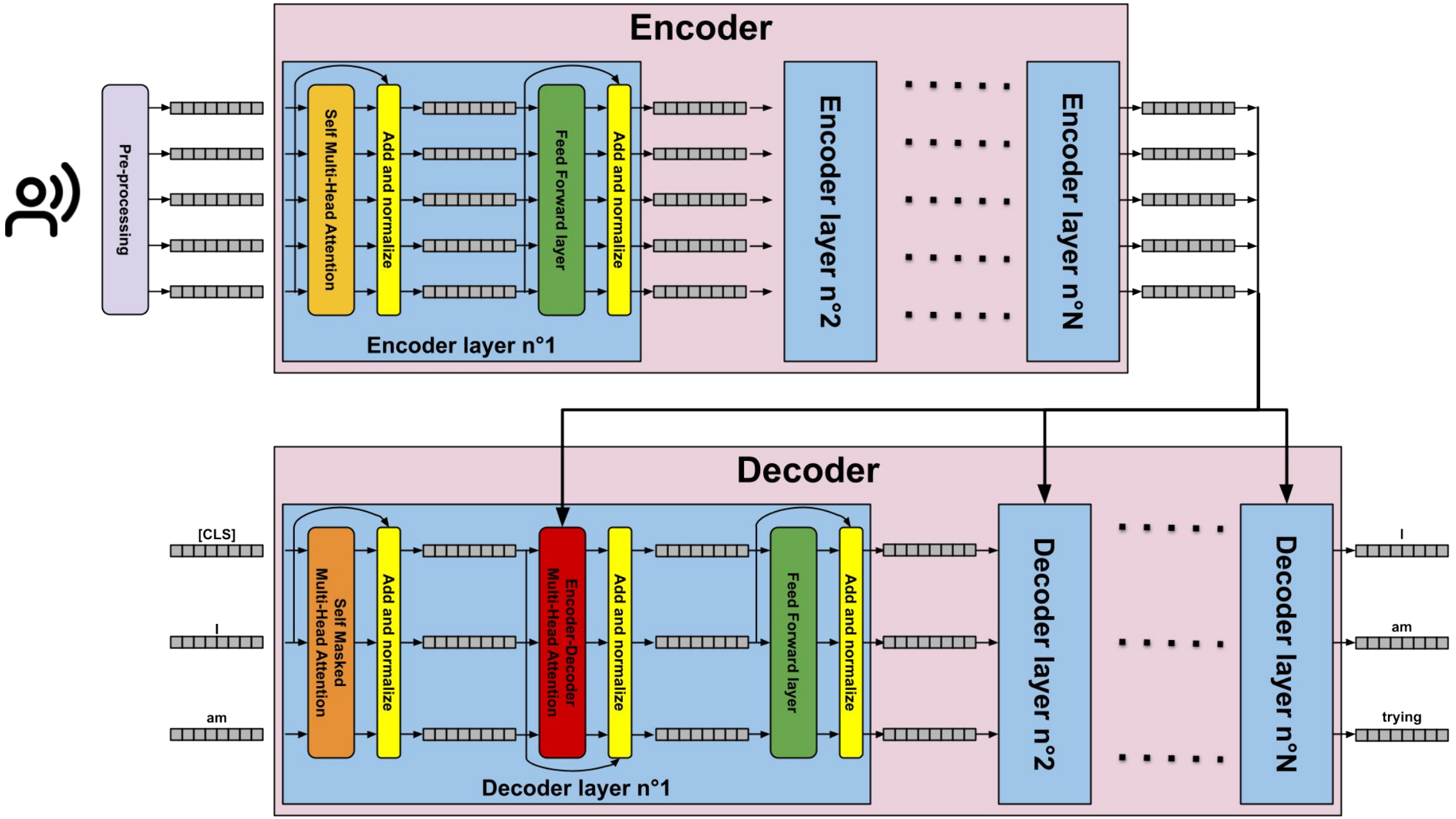
T5 / Encoder-Decoder / Sequences-to-Sequences



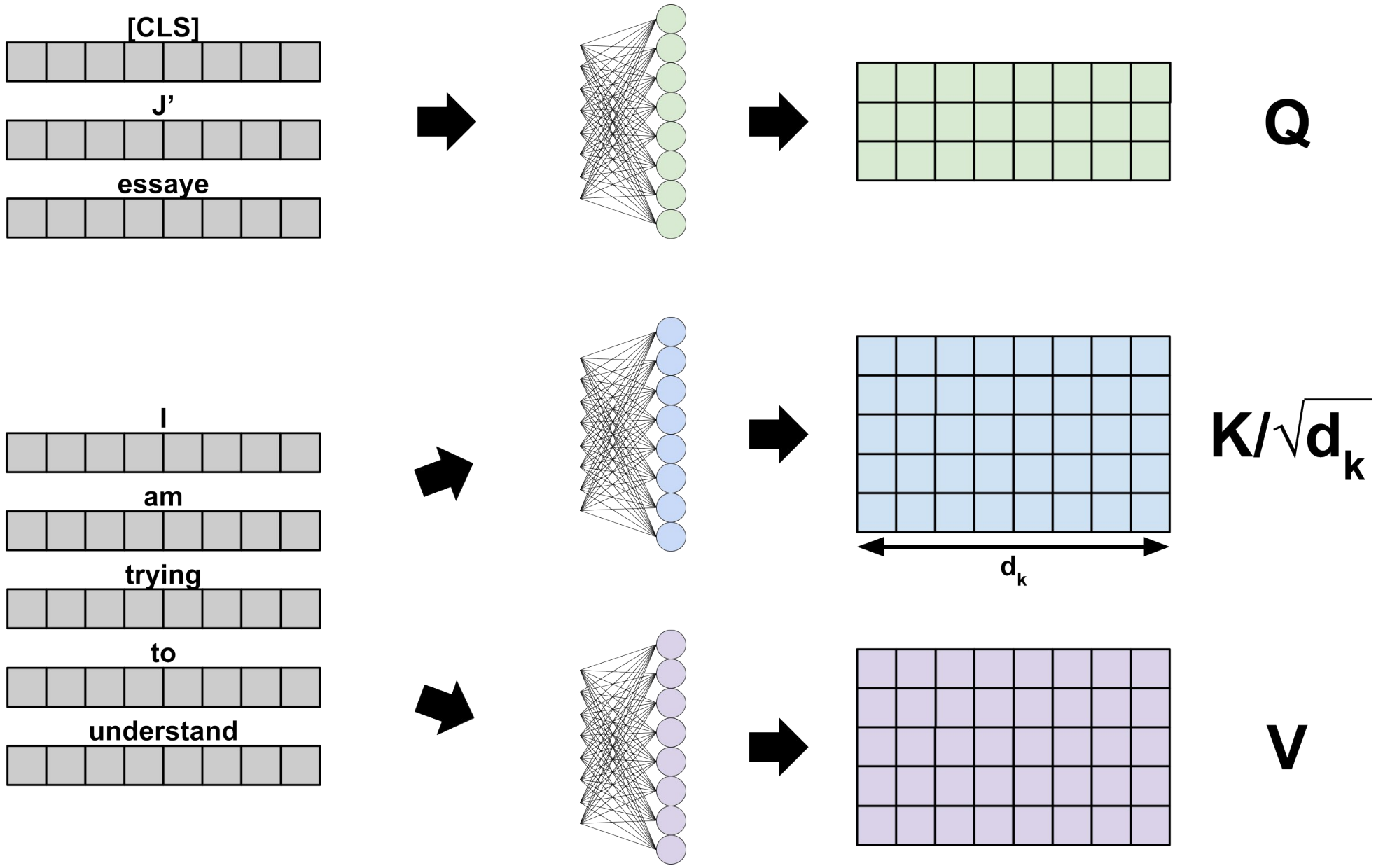
Encoder-Decoder translation example



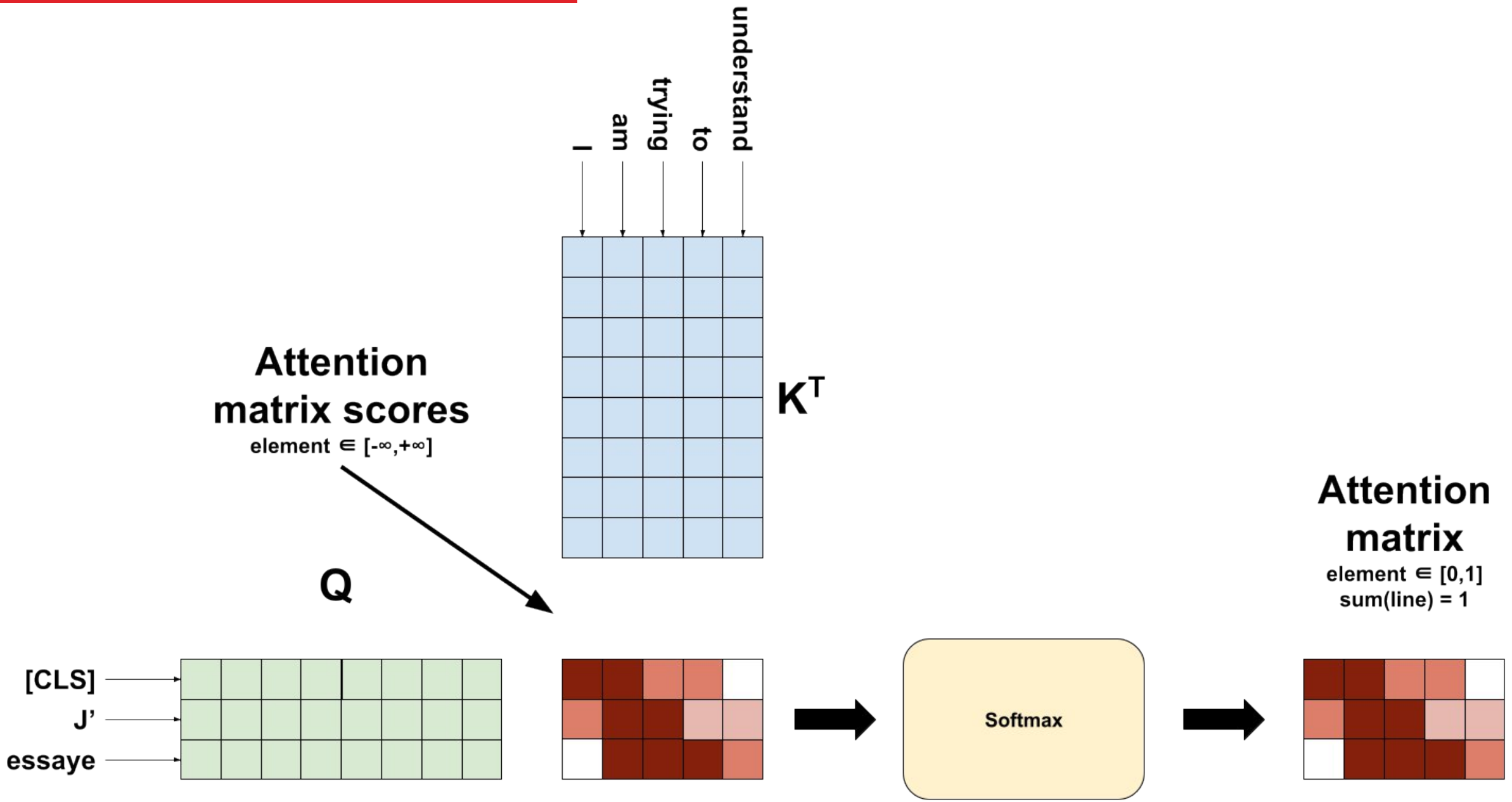
Encoder-Decoder whisper example



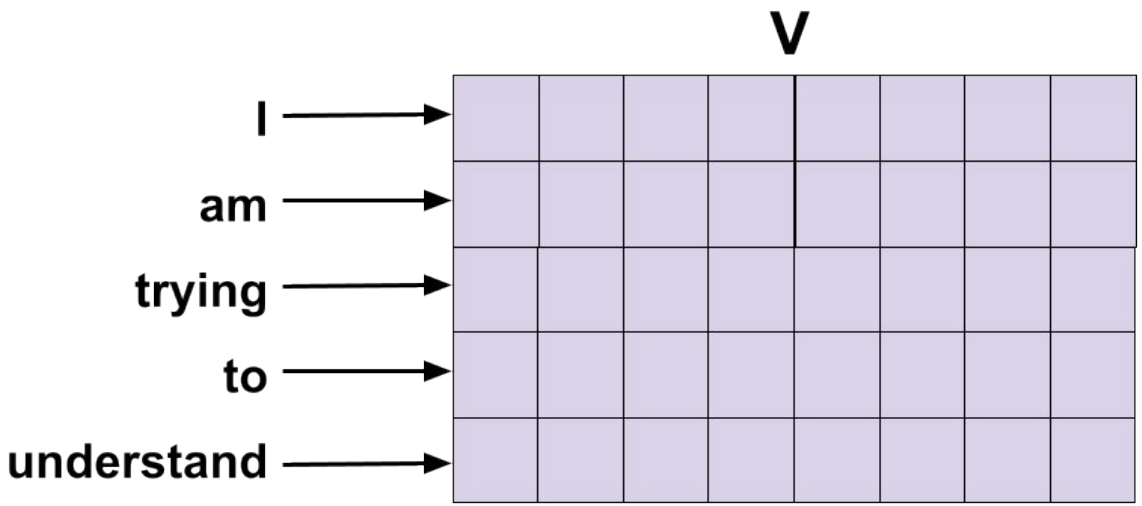
Encoder-Decoder attention - 1



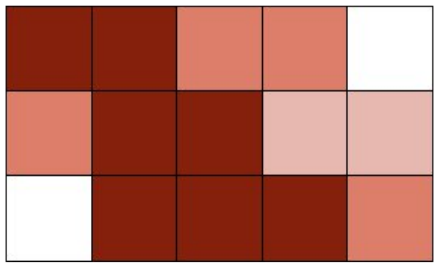
Encoder-Decoder attention - 2

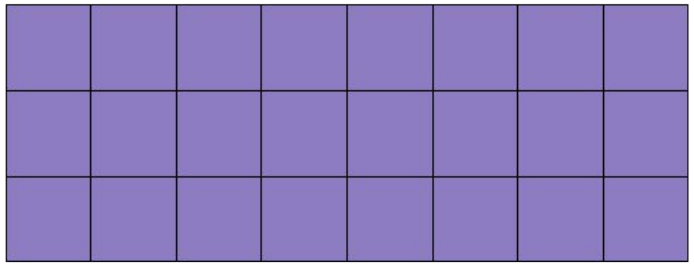


Encoder-Decoder attention - 3

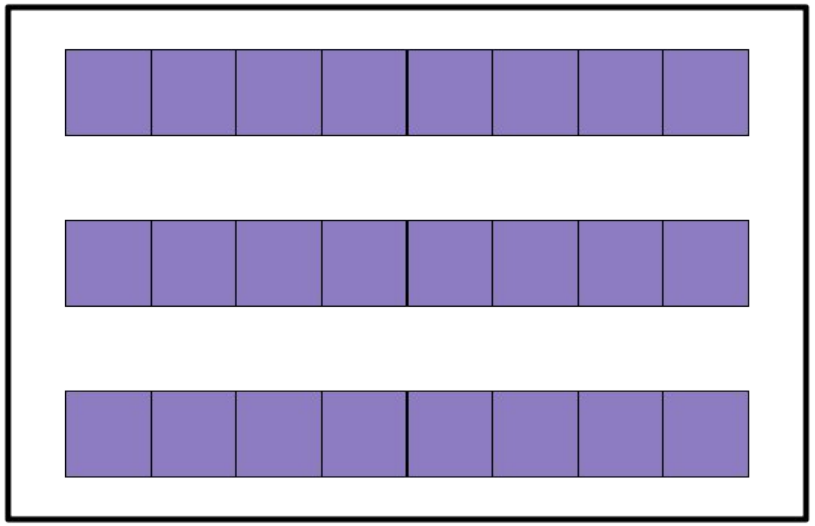


Attention matrix

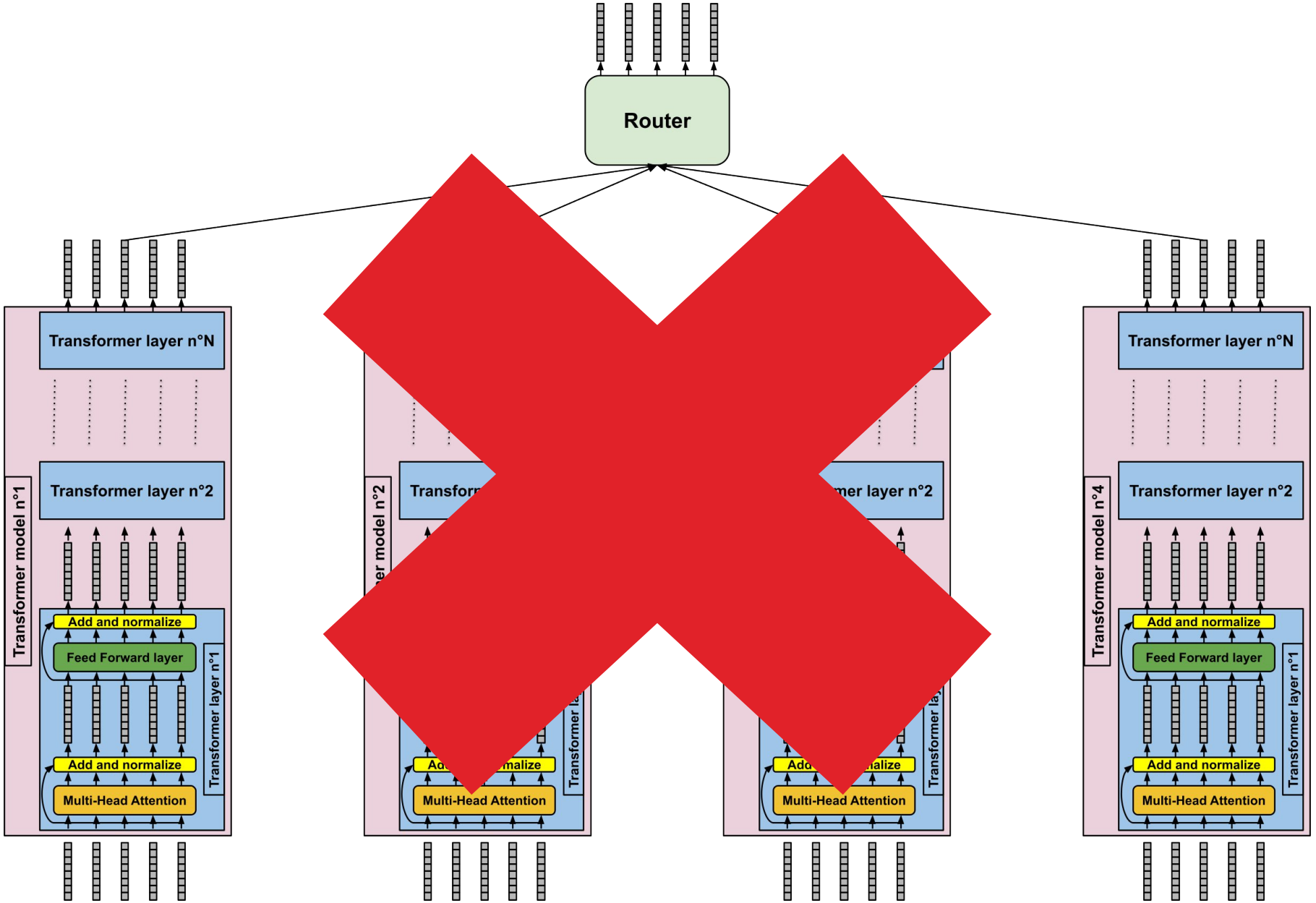




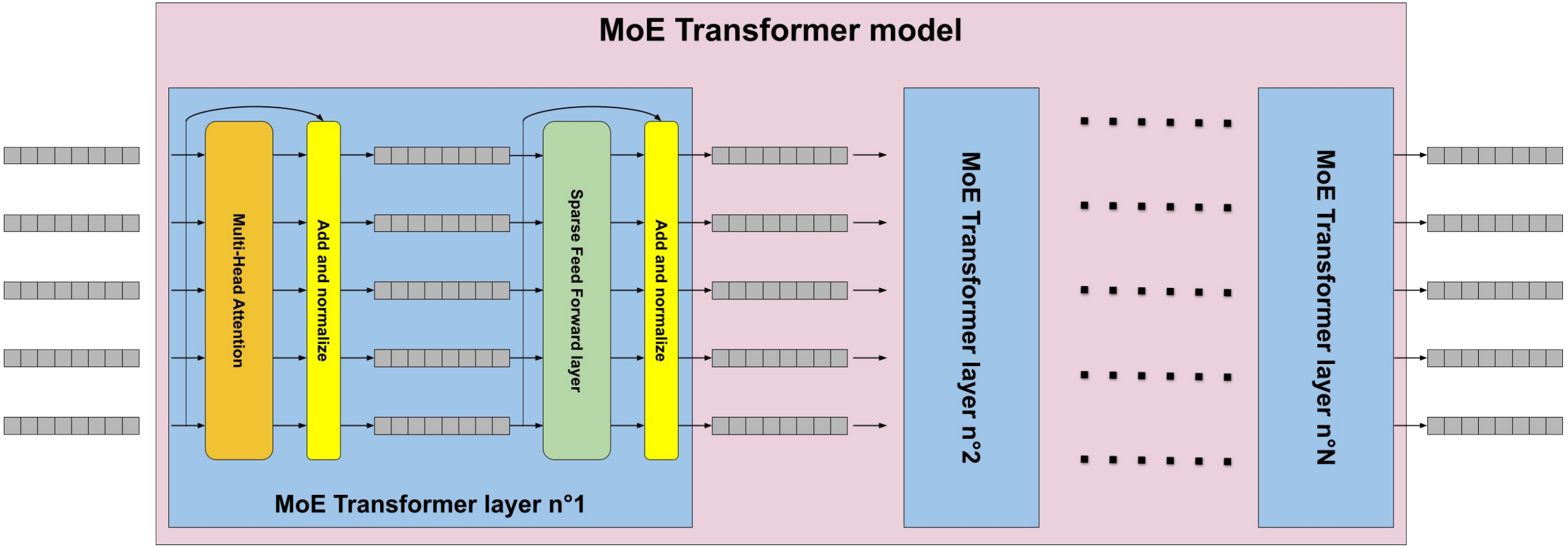
Output sequence



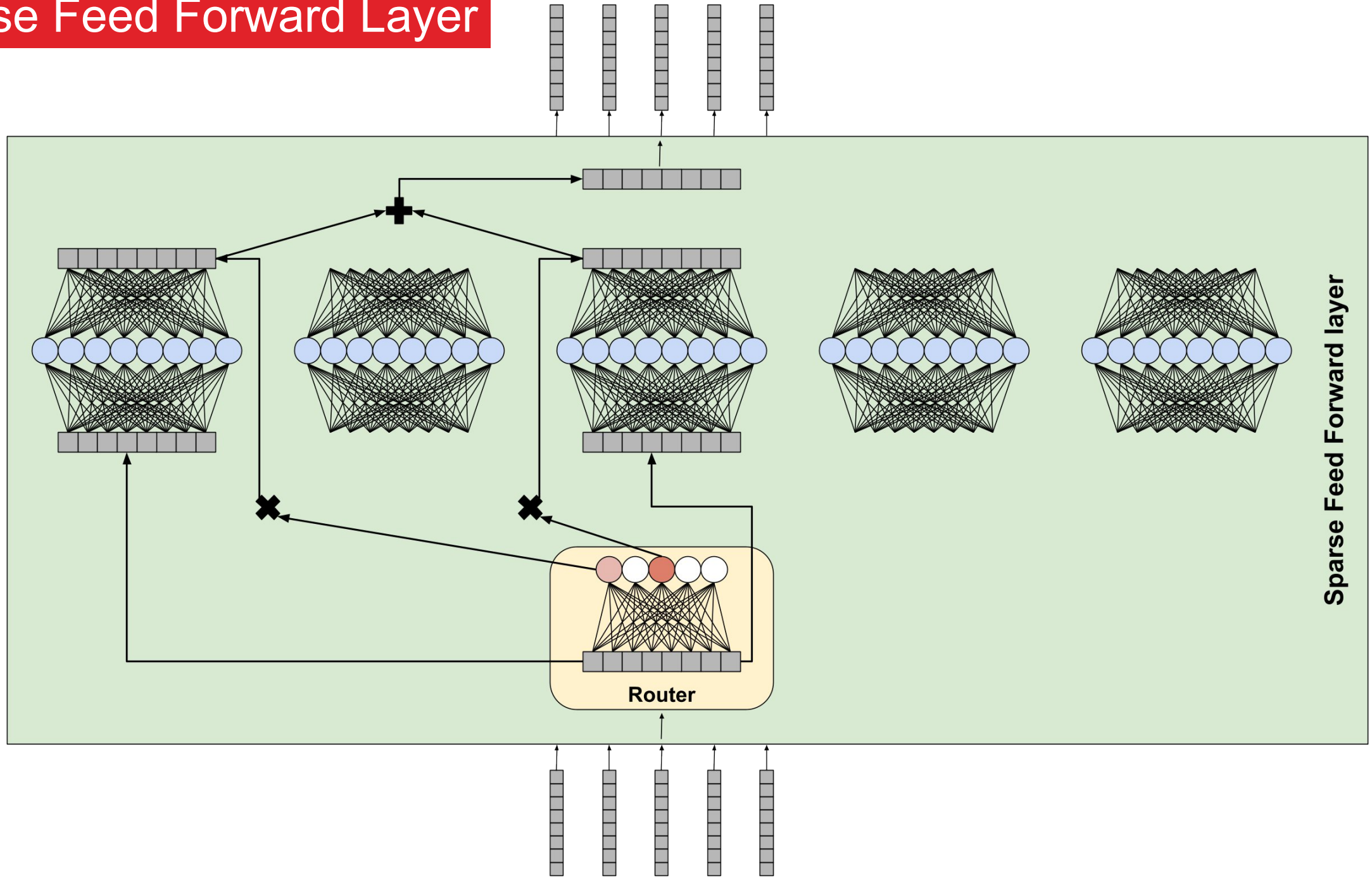
Misunderstanding about Mixture of Experts



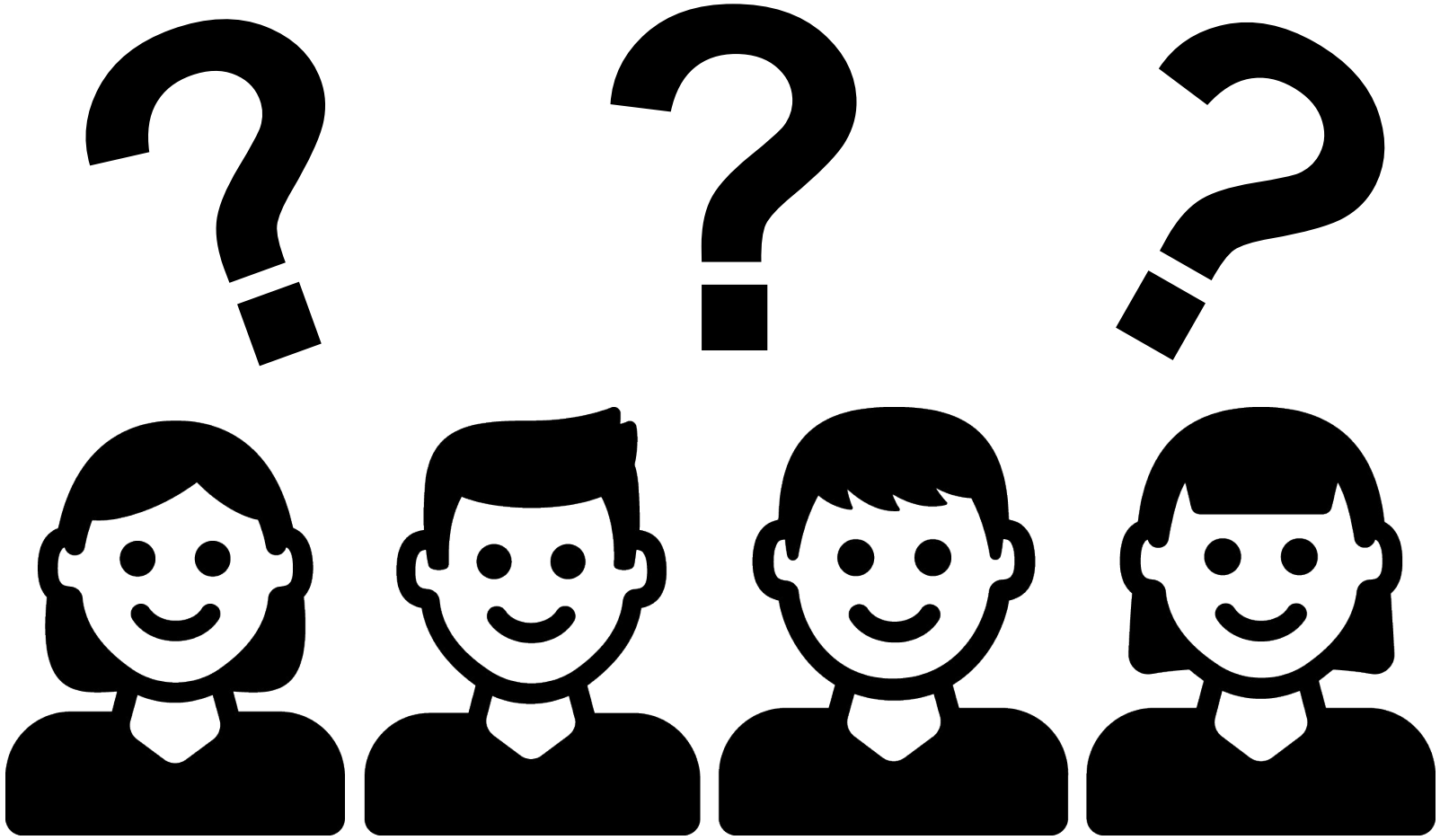
Mixture of Expert Transformer Architecture



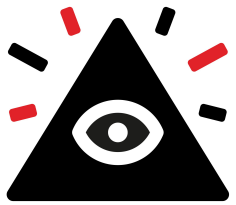
Sparse Feed Forward Layer



Question break #3



9



«Attention is All You Need»
Transformers

1. What is a **Transformer**? The magic of the **Attention Mechanism**
2. The different **Transformers architectures**
3. **Pre-training** and **Foundation Models**
4. **Specialization** of **Foundation Models** (especially LLM)
5. **Example: IMDB Reviews**

Training a language model

I want to make a chatbot to help people better understand the civil code.

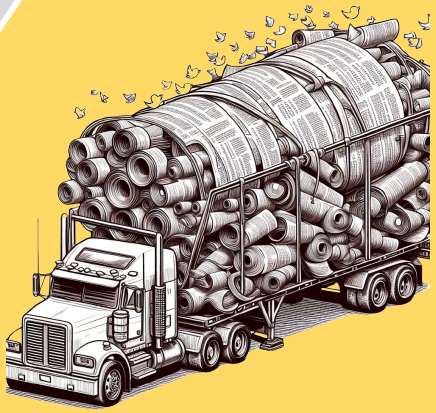
I want to make a bot which can filter out respectful comments for a real reddit experience.

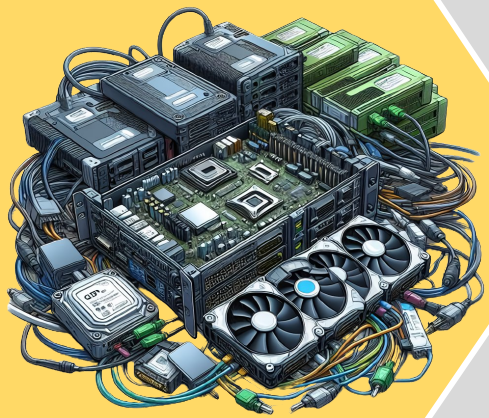
Training a language model

What do you need to train a large language model ?

A truckload of data

Transformers are ravenous. You need to feed them with a substantial and hard-to-come-by dataset.





A mighty compute infrastructure

GPUs with high throughput and large VRAM can execute this training in a reasonable amount of time.

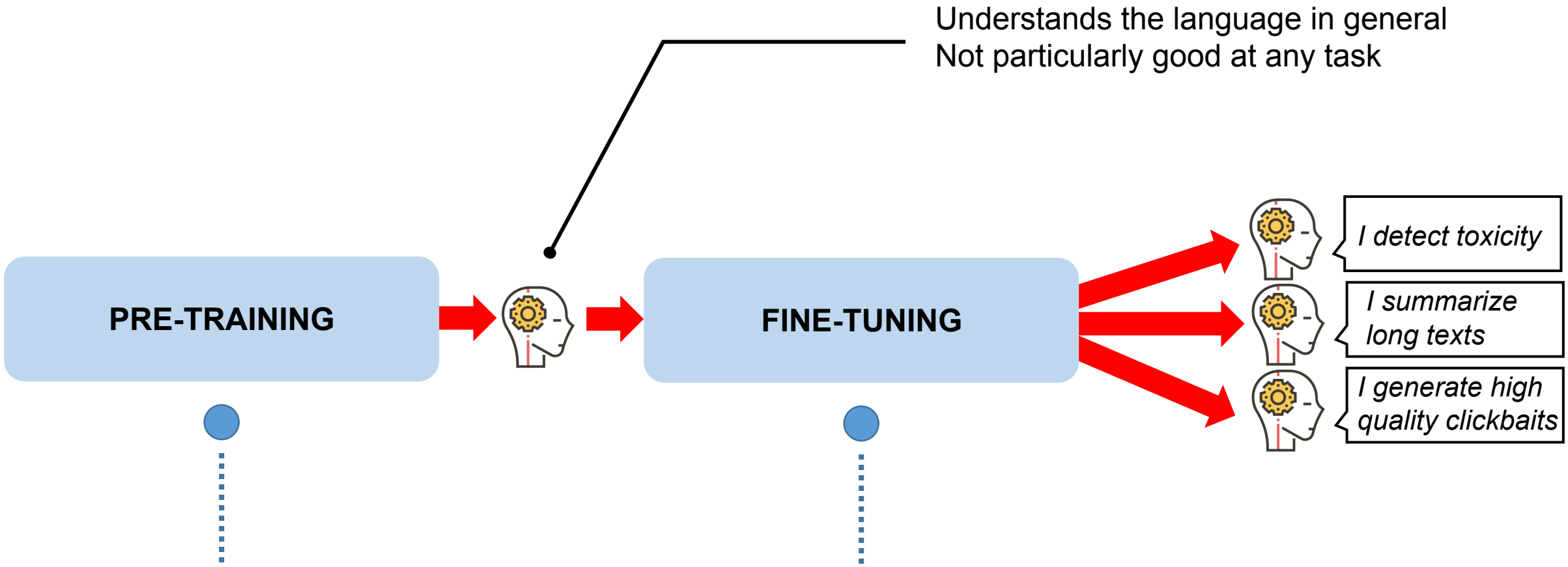
A copious amount of electricity

Storage, memory and compute power consume a lot of electricity.

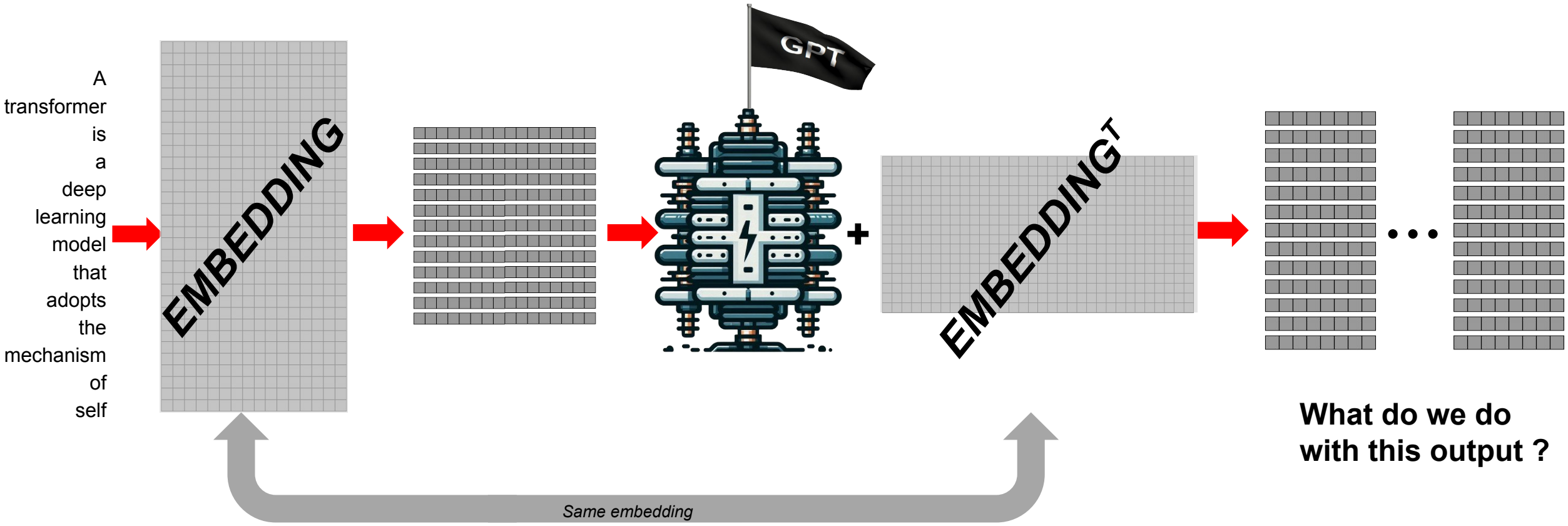
An abundance of manpower

Cleaning the dataset, making experiments, monitoring SOTA advancements is a lot of work.

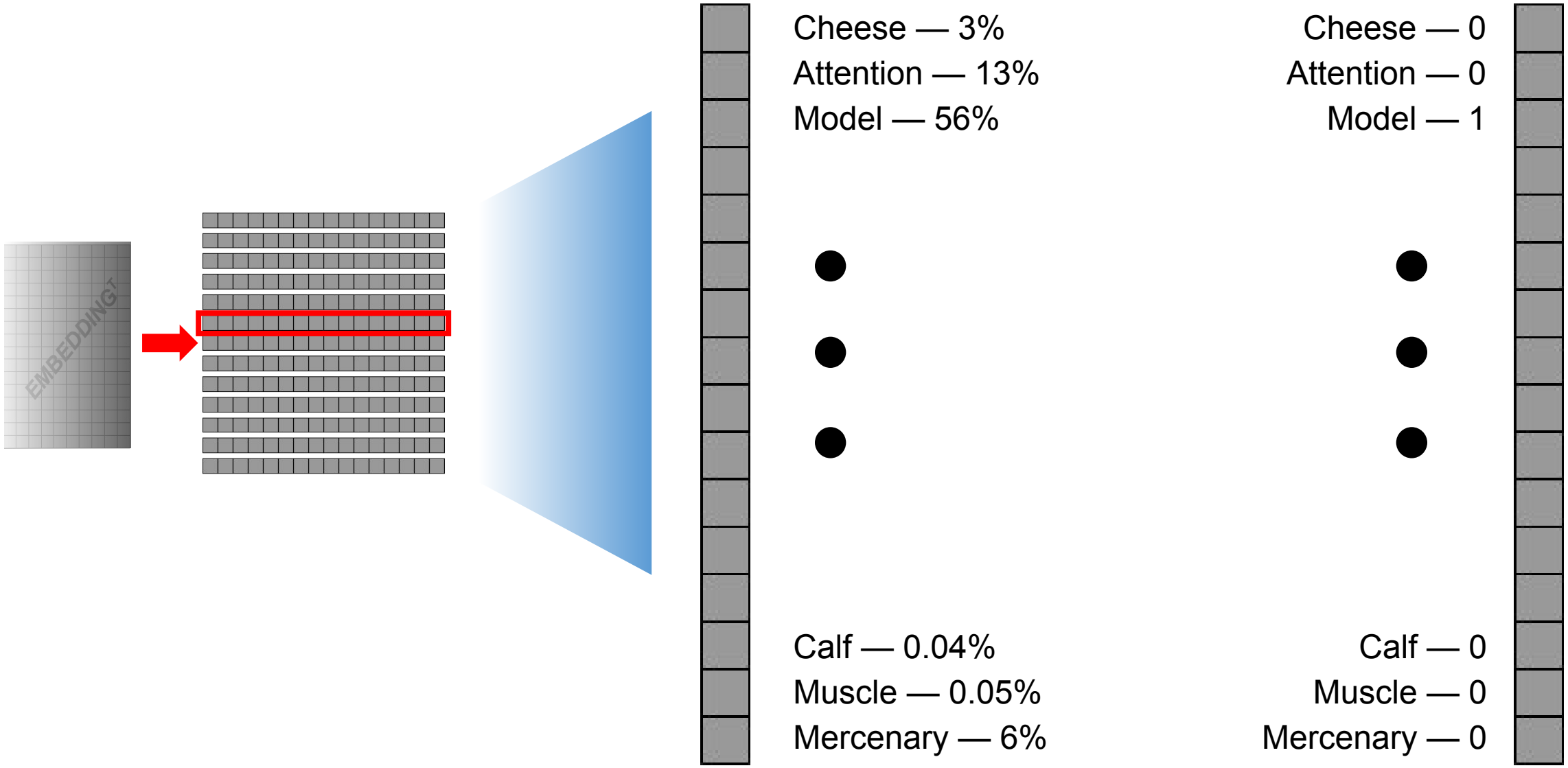
Training a language model



Pretraining a GPT-style transformer



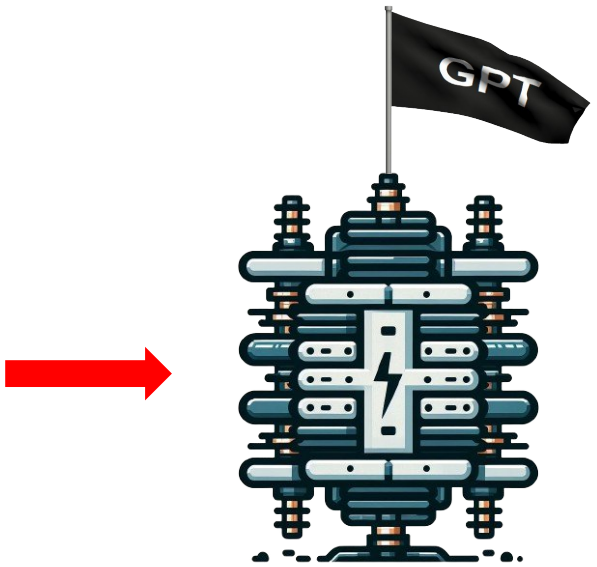
Pretraining a GPT-style transformer



Pretraining a GPT-style transformer

Input

A
transformer
is
a
deep
learning
model
that
adopts
the
mechanism
of
self

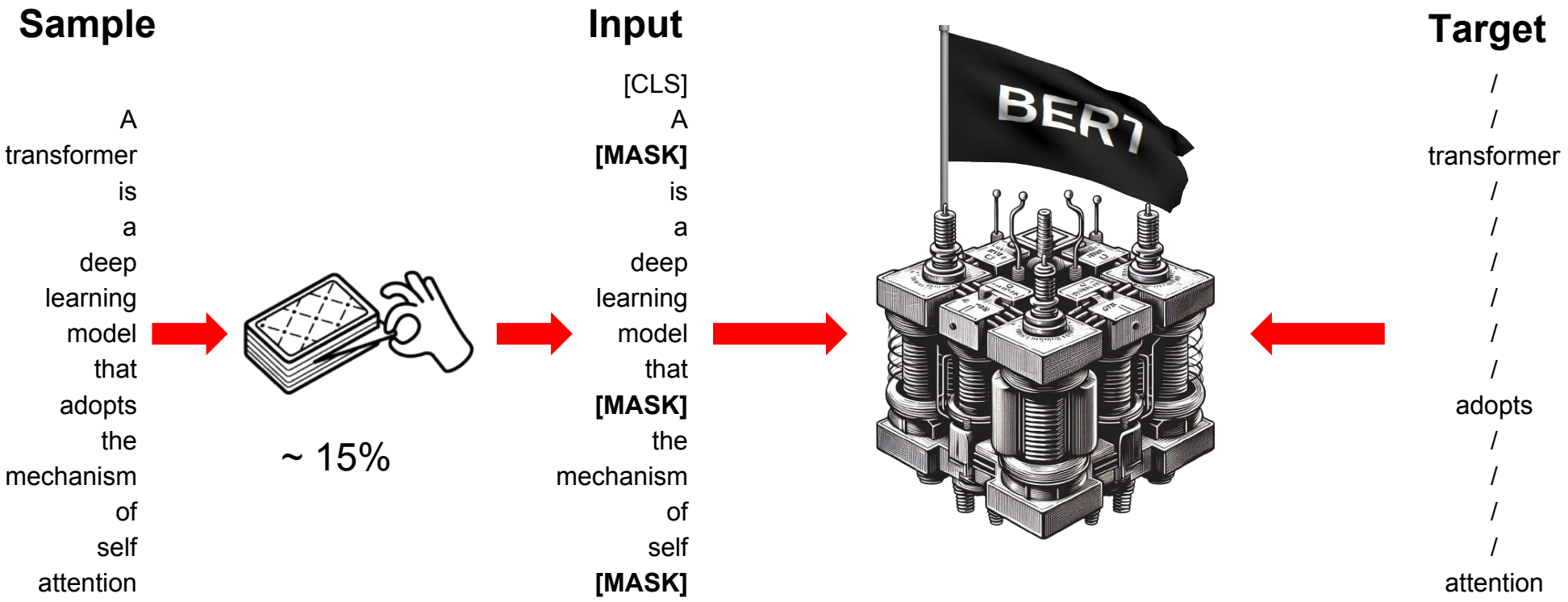


Target

transformer
is
a
deep
learning
model
that
adopts
the
mechanism
of
self
attention

Next word prediction

Pretraining a BERT-style transformer

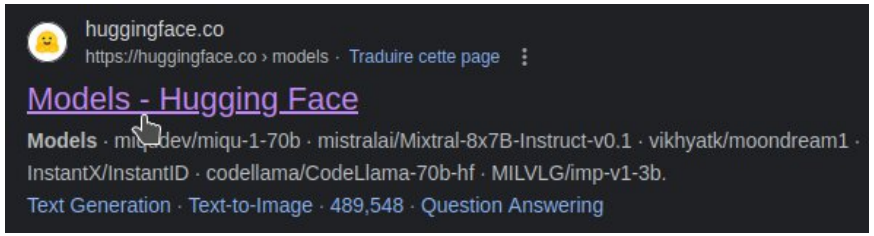


Masked words prediction

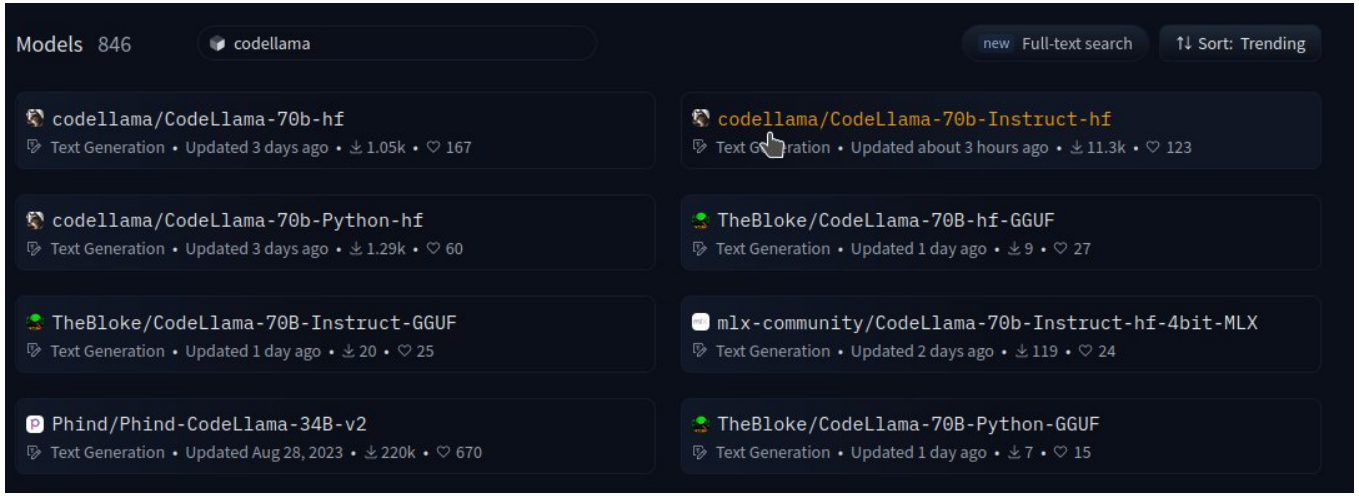
Finding a pretrained model

The largest free pretrained transformer models database

1 — Go to https://huggingface.co/



2 — Look for a model



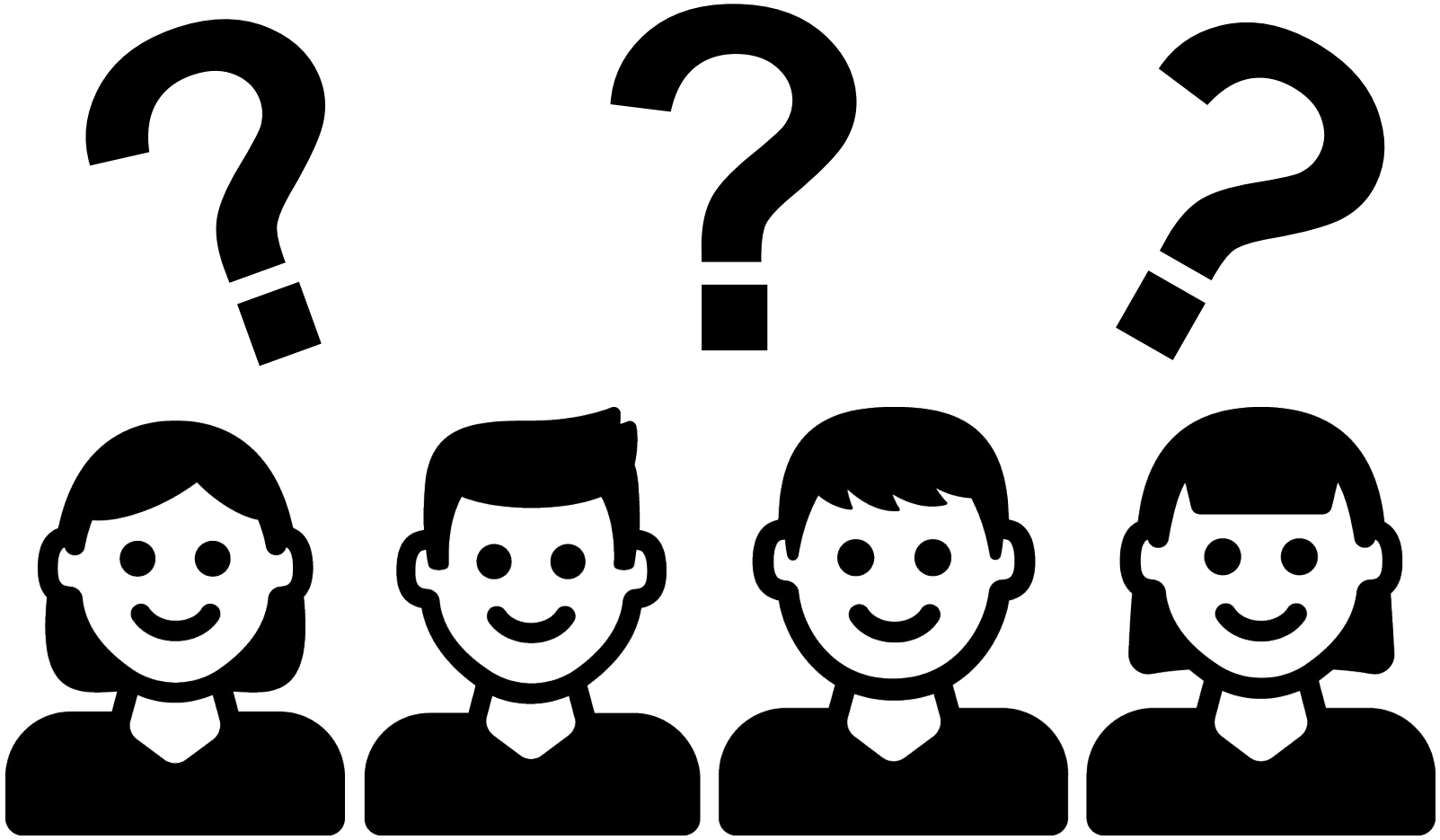
3 — Download the model

```
ncassereau@jean-zay:/fidle/transformers$ git clone https://huggingface.co/codellama/CodeLlama-70b-Instruct-hf
```

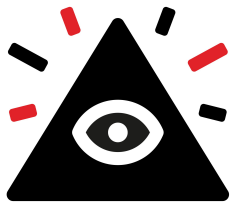
4 — Enjoy

```python
from transformers import AutoTokenizer, AutoModelForCausalLM

path = "/fidle/transformers/CodeLlama-70b-Instruct-hf"
tokenizer = AutoTokenizer.from_pretrained(path)
model = AutoModelForCausalLM.from_pretrained(path)

input = tokenizer(["I love Jean Zay", "Hatim truly is the best"], padding=True)
output = model(**input)
```

Question break #4



9



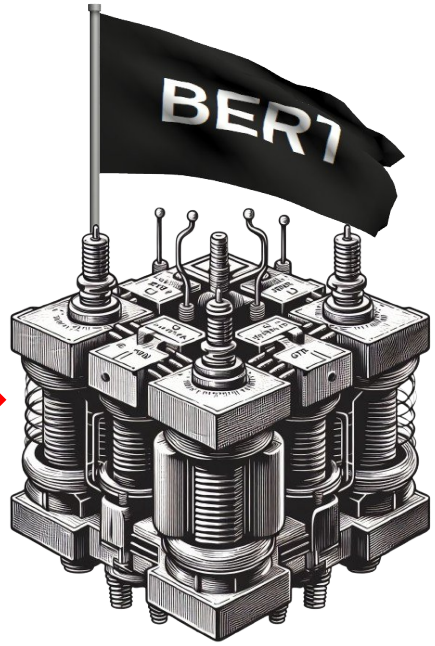
«Attention is All You Need»
Transformers

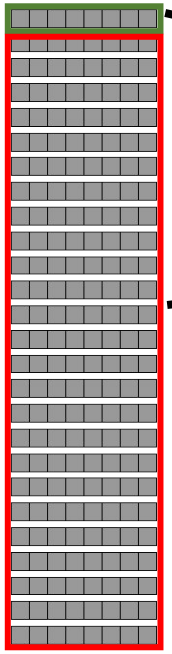
1. What is a **Transformer**? The magic of the **Attention Mechanism**
2. The different **Transformers architectures**
3. **Pre-training** and **Foundation Models**
4. **Specialization** of **Foundation Models** (especially LLM)
5. **Example: IMDB Reviews**

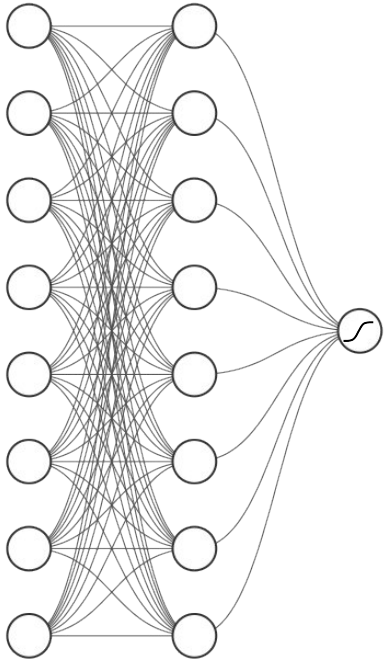
Fine-tuning of language models

Fine-tuning a BERT — Sentence classification

[CLS] Star Wars The Last Jedi is really bad!



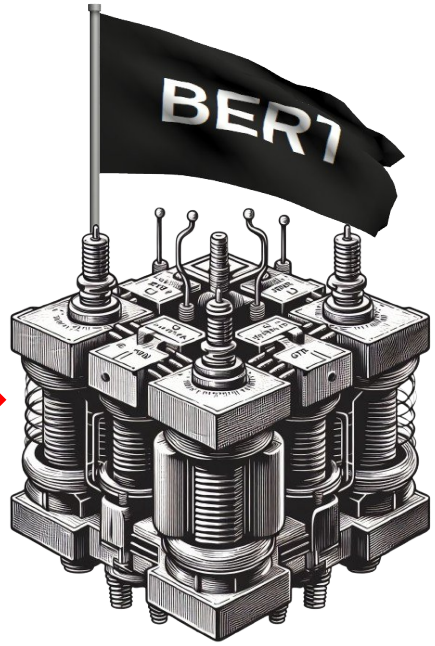


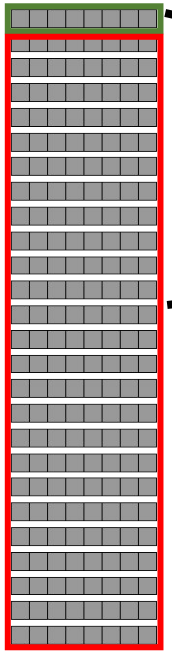


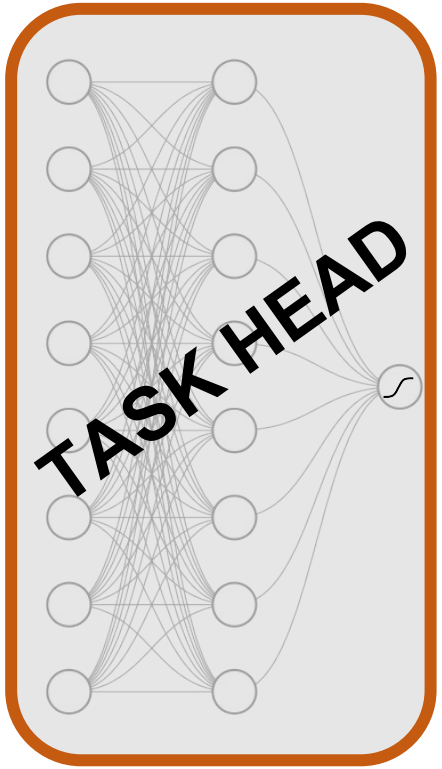
Target: 0

Fine-tuning a BERT — Sentence classification

[CLS] Star Wars The Last Jedi is really bad!



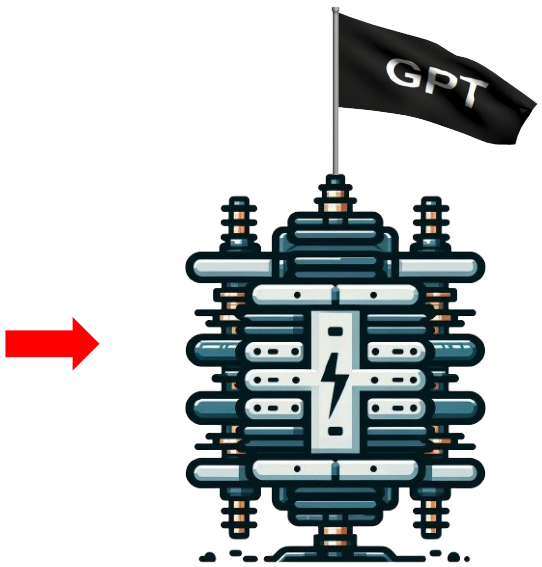


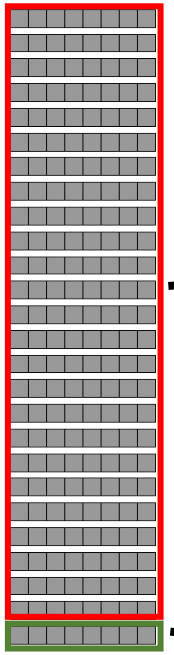


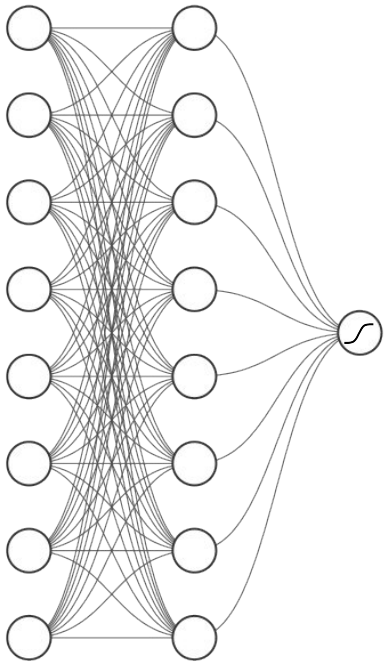
Target: 0

Fine-tuning a GPT — Sentence classification

Star Wars The Last Jedi is really bad!



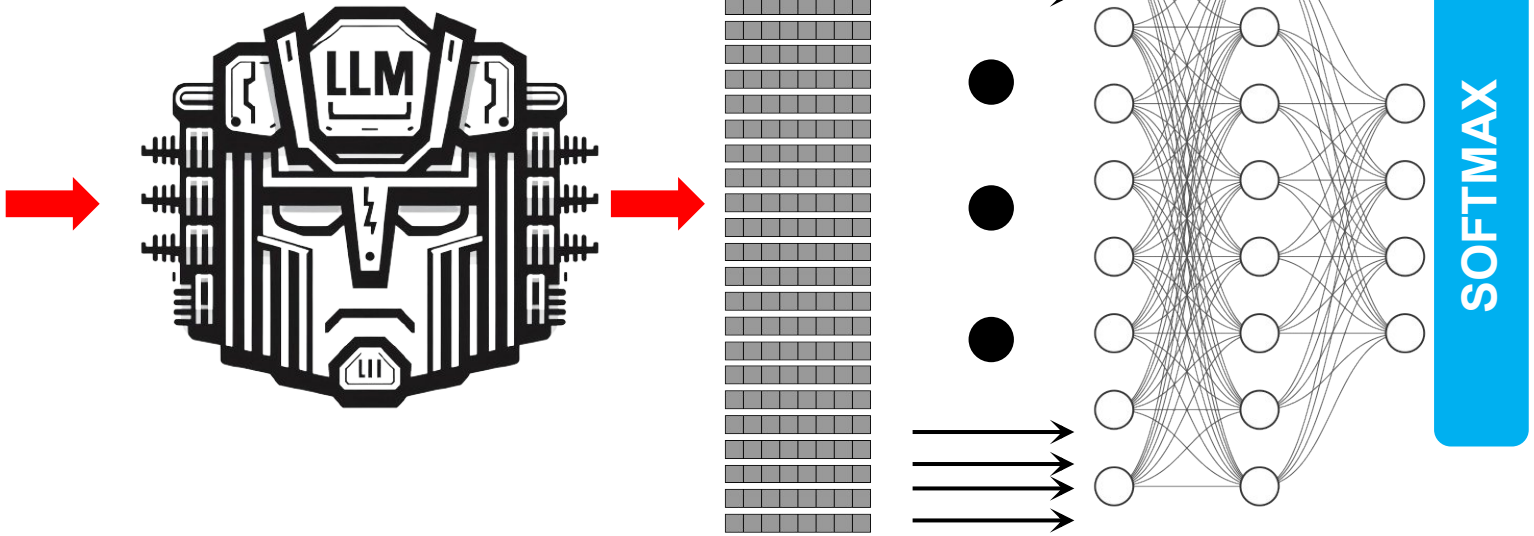




Target: 0

Fine-tuning a transformer — Named Entity Recognition

Zidane came to Paris to watch the Paris Saint-Germain



Output	Target
Person: 0.45	Person: 1
Organization: 0.12	Organization: 0
Location: 0.27	Location: 0
None: 0.16	None: 0

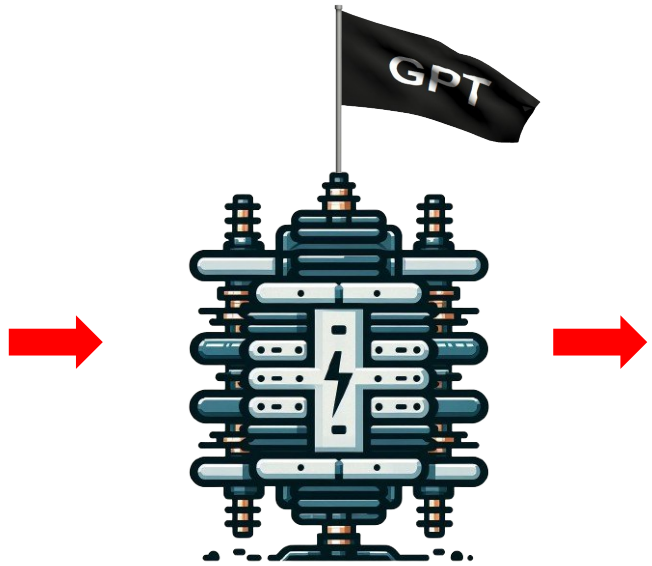
Fine-tuning a GPT with prompting

Review

This film is really trash!

Template

{{ REVIEW }} This review is (positive, negative or neutral):



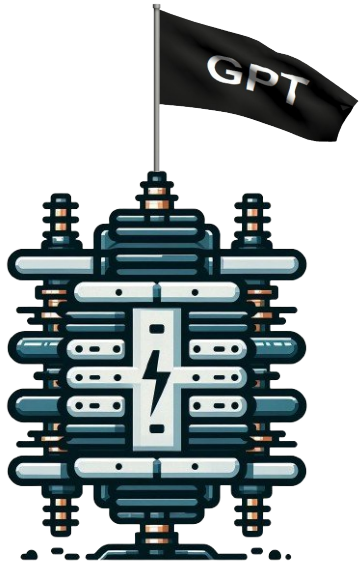
Output	**Target**
Positive: 0.18	Positive: 0
Negative: 0.44	Negative: 1
Neutral: 0.38	Neutral: 0

Fine-tuning a GPT — Example of summarization

Input

Template

{{ INPUT }} TL;DR:



GENERATION HEAD

Target

She dumped me because I took a sip before toasting.

Fine-tuning a GPT with templates

Type	Task	Input ([X])	Template	Answer ([Z])
Text CLS	Sentiment	I love this movie.	[X] The movie is [Z].	great fantastic ...
	Topics	He prompted the LM.	[X] The text is about [Z].	sports science ...
	Intention	What is taxi fare to Denver?	[X] The question is about [Z].	quantity city ...
Text-span CLS	Aspect Sentiment	Poor service but good food.	[X] What about service? [Z].	Bad Terrible ...
Text-pair CLS	NLI	[X1]: An old man with ... [X2]: A man walks ...	[X1]? [Z], [X2]	Yes No ...
Tagging	NER	[X1]: Mike went to Paris. [X2]: Paris	[X1][X2] is a [Z] entity.	organization location ...
Text Generation	Summarization	Las Vegas police ...	[X] TL;DR: [Z]	The victim ... A woman ... ...
	Translation	Je vous aime.	French: [X] English: [Z]	I love you. I fancy you. ...

Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, Pengfei Liu and al

Orienting the style with templates

System

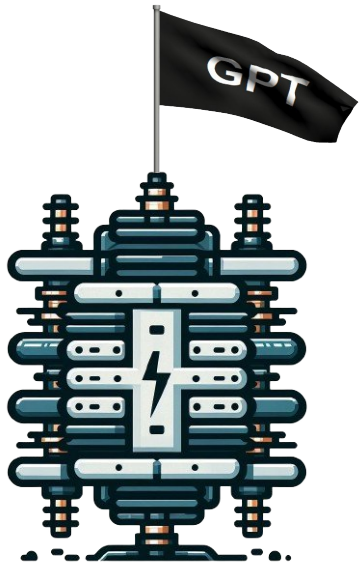
Speak as if you were a pirate.

Input

What is the Cayley-Hamilton theorem ?

Template

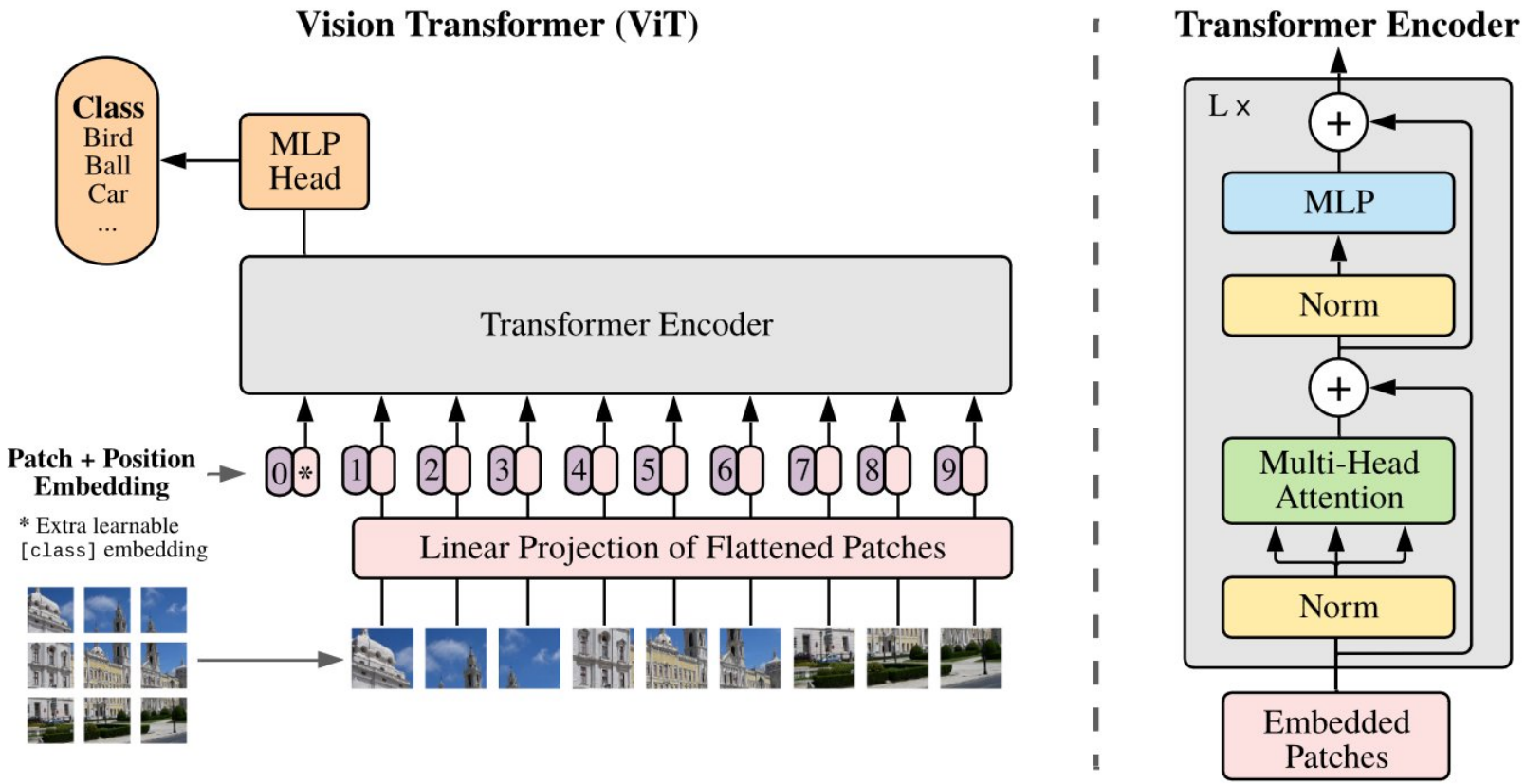
<|system|>
{{ SYSTEM }}
<|user|>
{{ INPUT }}
<|assistant|>

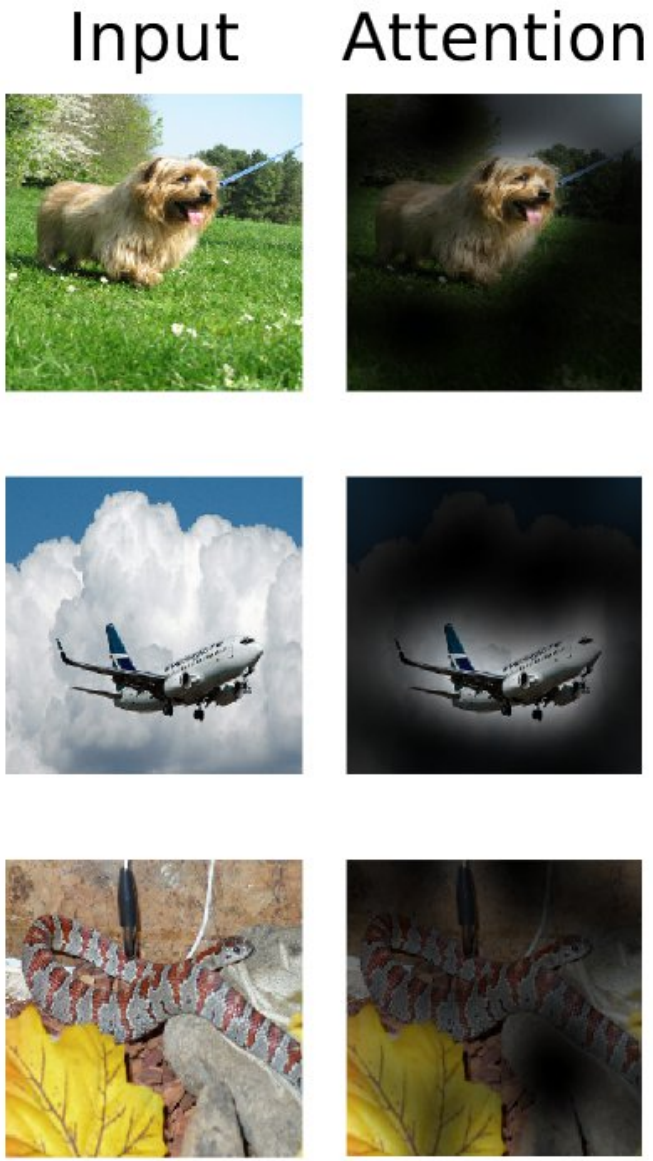


GENERATION HEAD

Ahoy, me hearties! I be hearin' 'bout a mathematical concept called the Cayley-Hamilton theorem. It's a fancy way o' sayin' that every square matrix is a root of its own characteristic equation, which is a polynomial equation that describes the matrix's properties. This theorem has important implications in linear algebra and matrix theory, as it allows for the calculation of matrix powers and inverses in certain cases. So, if you ever find yourself in a sea of matrices, remember the Cayley-Hamilton theorem and you'll be sailing smoothly!

Vision Transformer

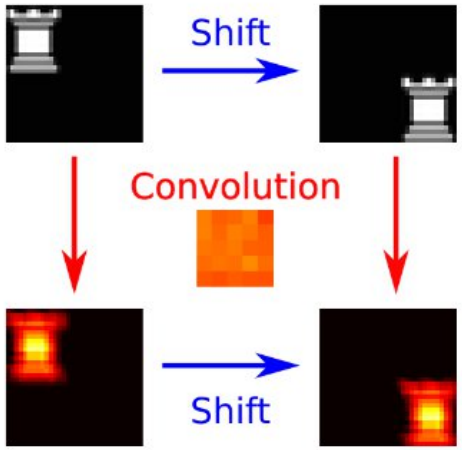




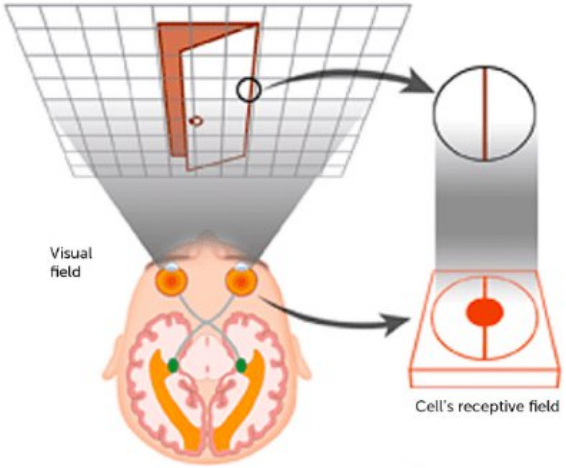
CoAtNet

"Marrying Convolution and Attention for All Data Sizes"

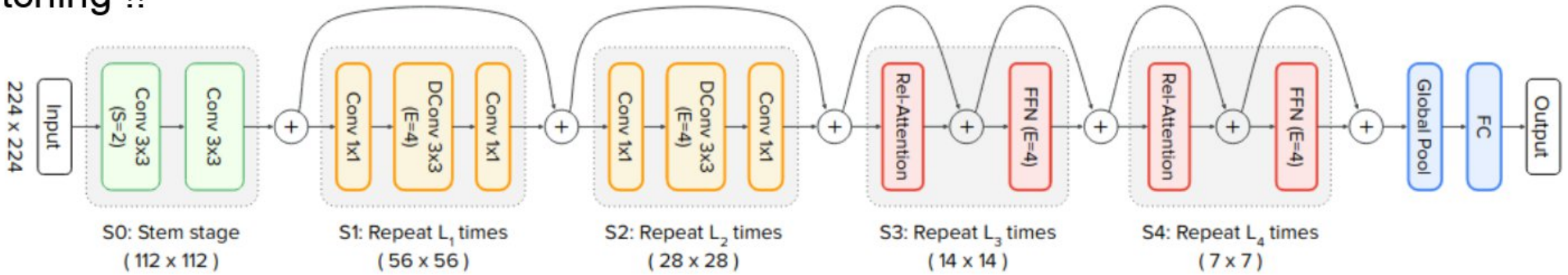
ConvNet
Translation Equivariance



Self-Attention Net
Global Receptive Field
Context comprehension



No patching !!



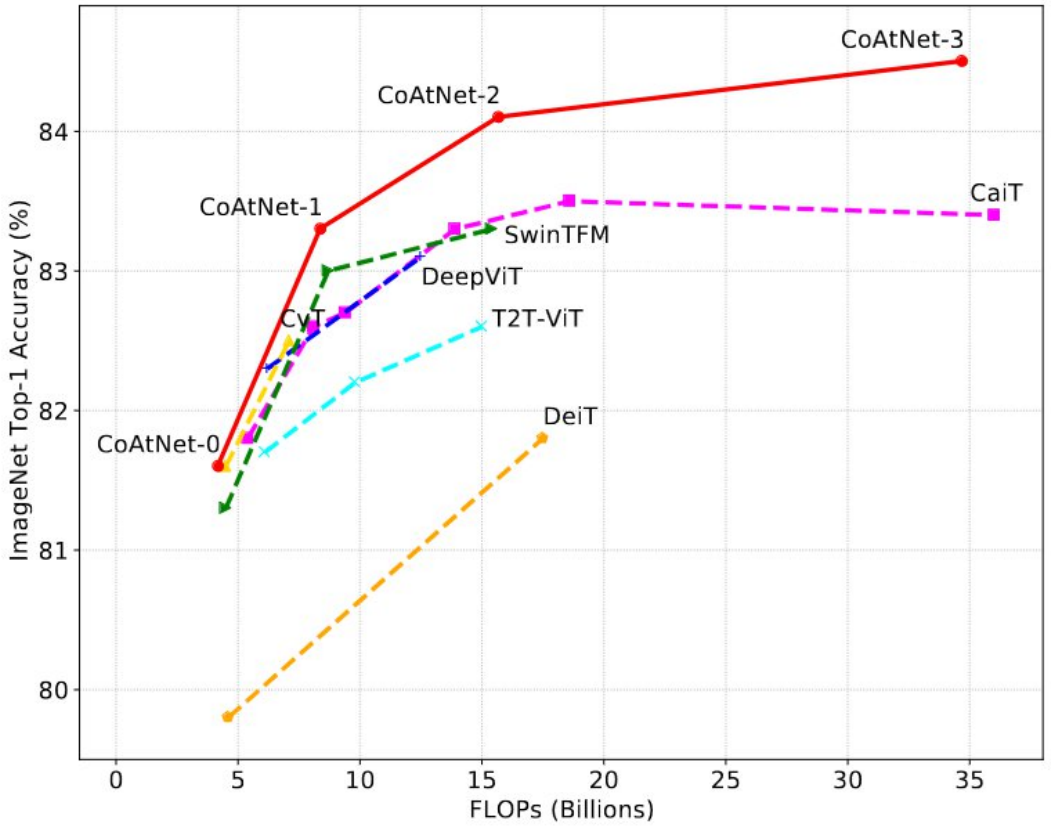


Figure 2: Accuracy-to-FLOPs scaling curve under ImageNet-1K only setting at 224x224.

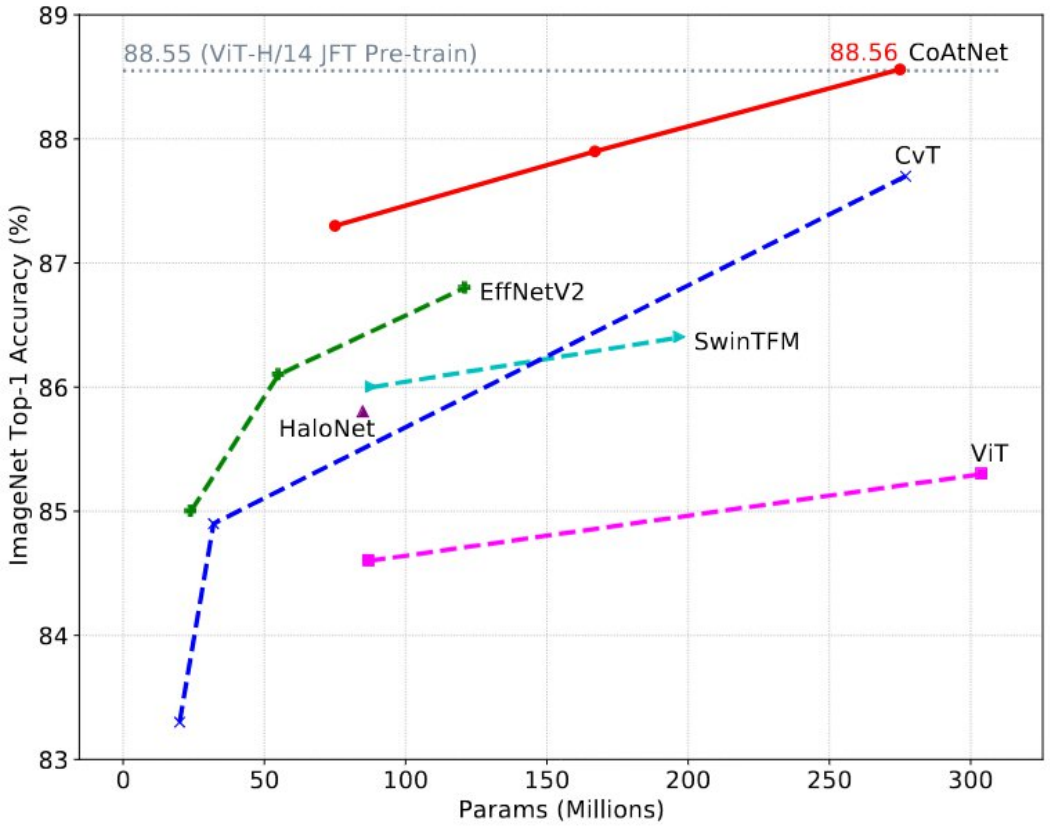
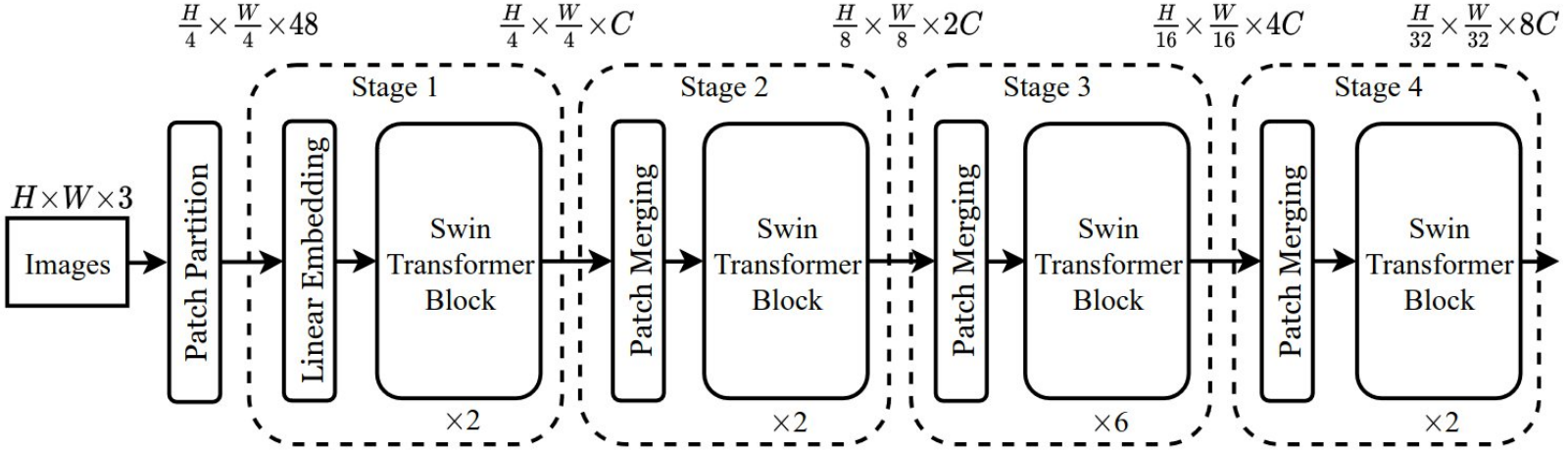
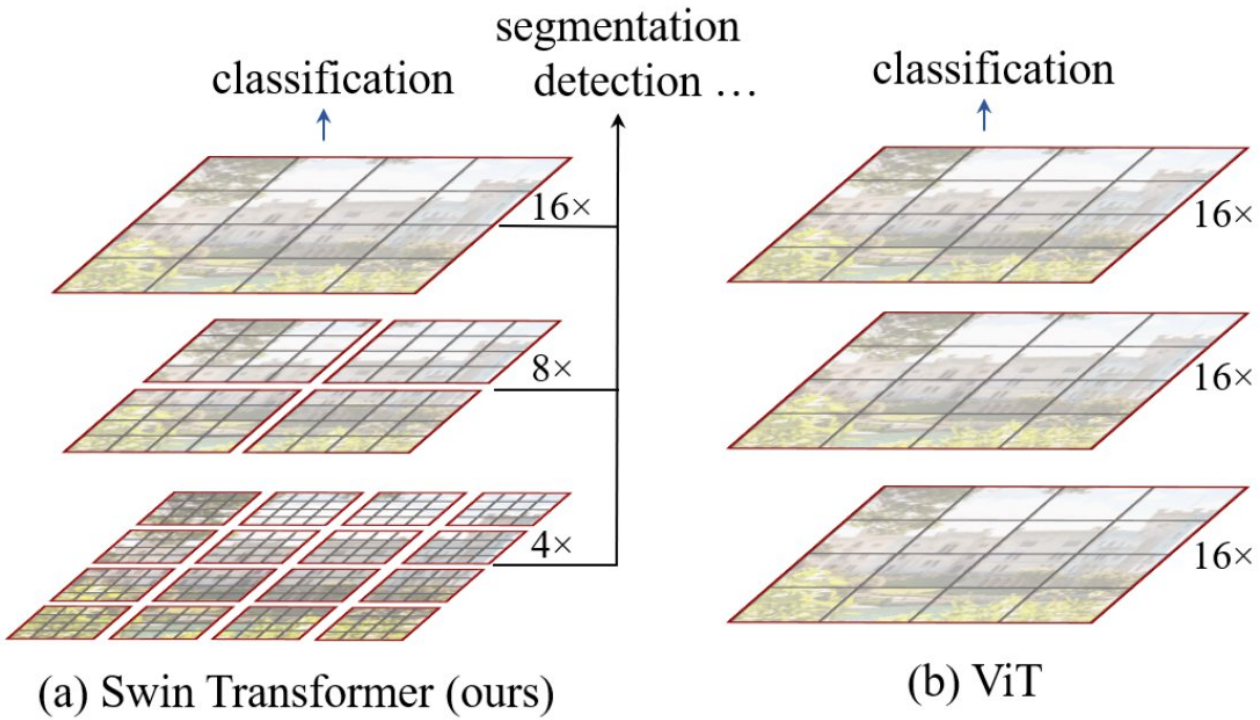


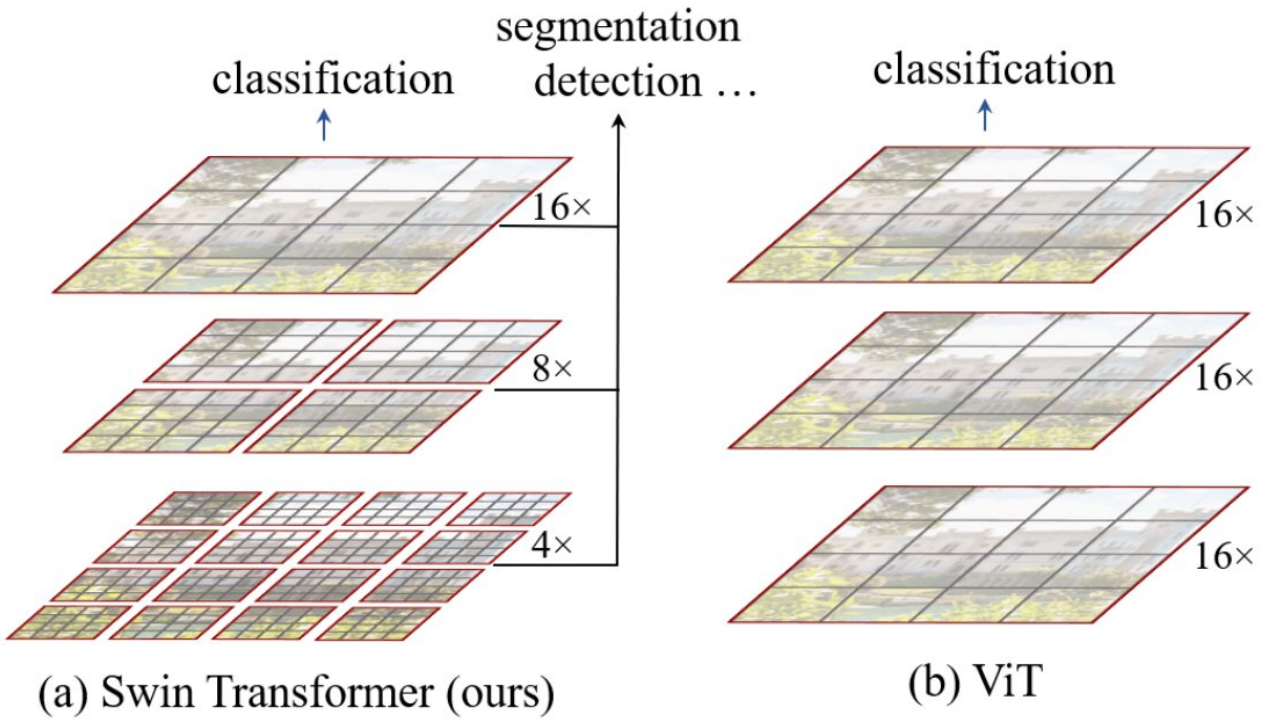
Figure 3: Accuracy-to-Params scaling curve under ImageNet-21K ⇒ ImageNet-1K setting.

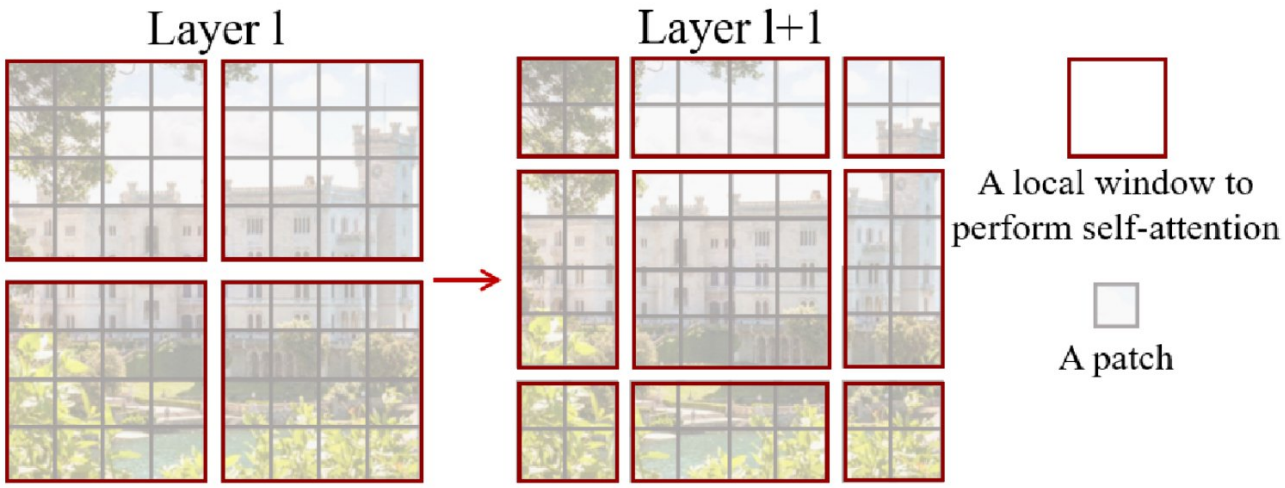
SwinTransformer



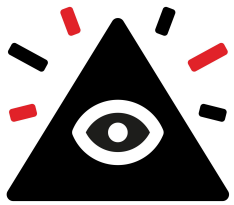


SwinTransformer





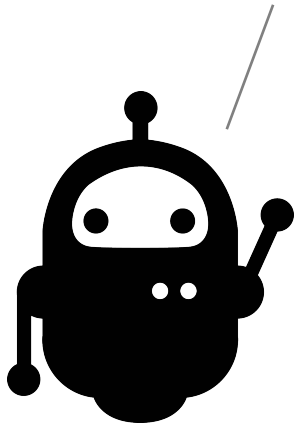
9

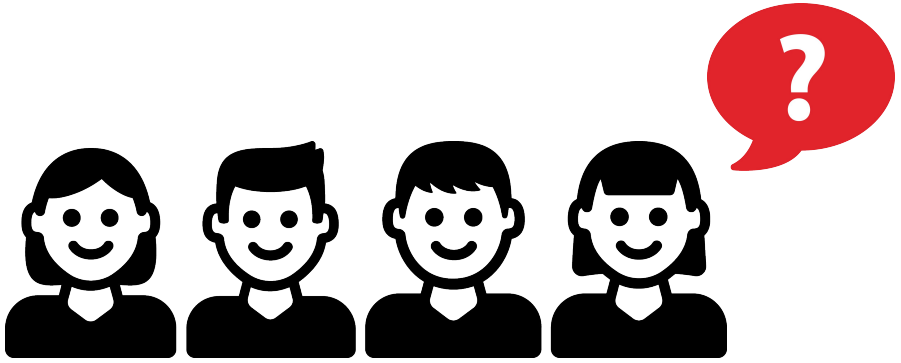


«Attention is All You Need»
Transformers

1. What is a **Transformer**? The magic of the **Attention Mechanism**
2. The different **Transformers architectures**
3. **Pre-training** and **Foundation Models**
4. **Specialization** of **Foundation Models** (especially LLM)
5. **Example: IMDB Reviews**

Merci beaucoup !!
Quelques questions ?





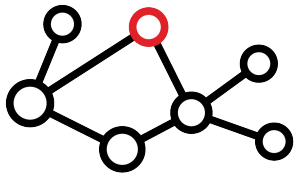
FIDLE

https://youtube.com/@CNRS-FIDLE
https://fidle.cnrs.fr

Jeudi
15
Février
à 14h00

Next, on Fidle : 6

L'IA comme un outil



Graph Neural Network
GNN

cnrs MIAI Grenoble Alpes UGA Université Grenoble Alpes EFELIA - MIAI
GENCI ANITI DEVLOG Resinfo