Introduction to Large Language Models

Changho Suh

January 29, 2024

Natural language processing (NLP)

The demand for NLP is growing at a phenomenal rate.

Applications:

speech recognition, question answering,
machine translation, grammar correction,
text summarization, image captioning, etc.

One killer app that has received particular attention:

Machine translation

Performance measure for machine translation

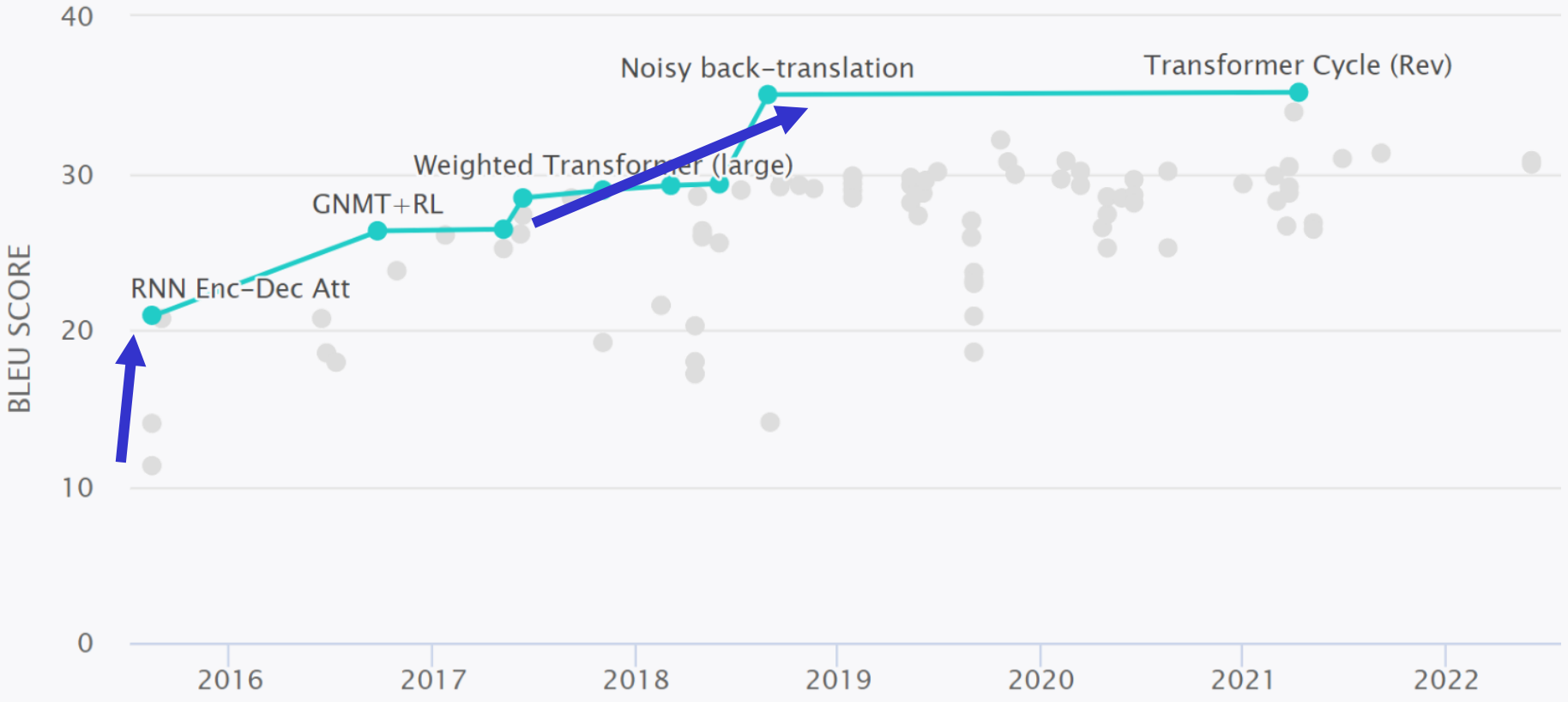
BLEU score (**Bi**-**L**ingual **E**valuation **U**nderstudy):

A number between 0 and 1 that measures the similarity of the machine-translated text to a set of high quality reference translations.

A benchmark dataset:

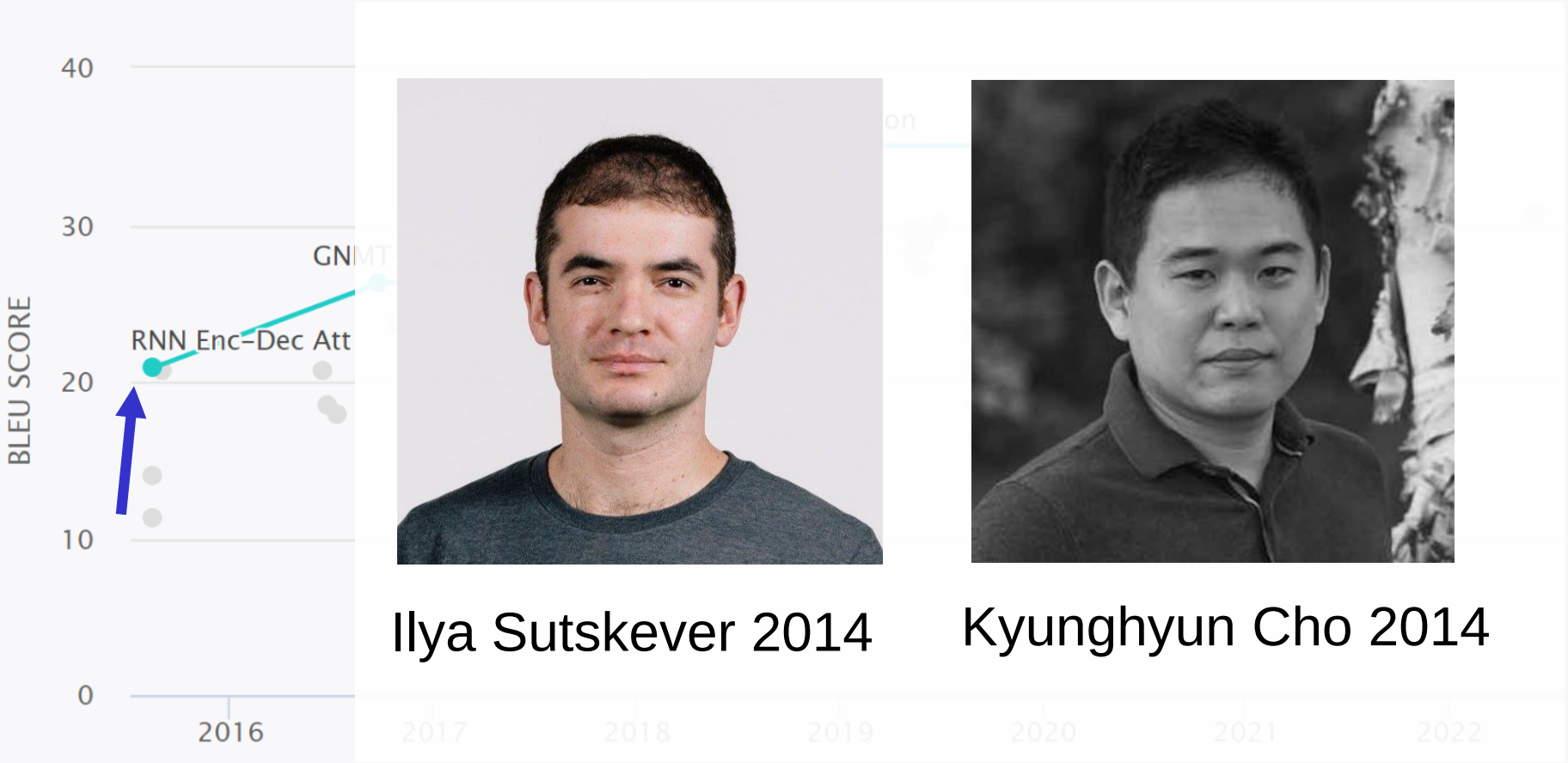
WMT dataset (4M sentences)
(**W**orkshop on Statistical **M**achine **T**ranslation)

Two breakthroughs



Source: https://paperswithcode.com/sota/machine-translation-on-wmt2014-english-german

RNN Encoder-Decoder Attention



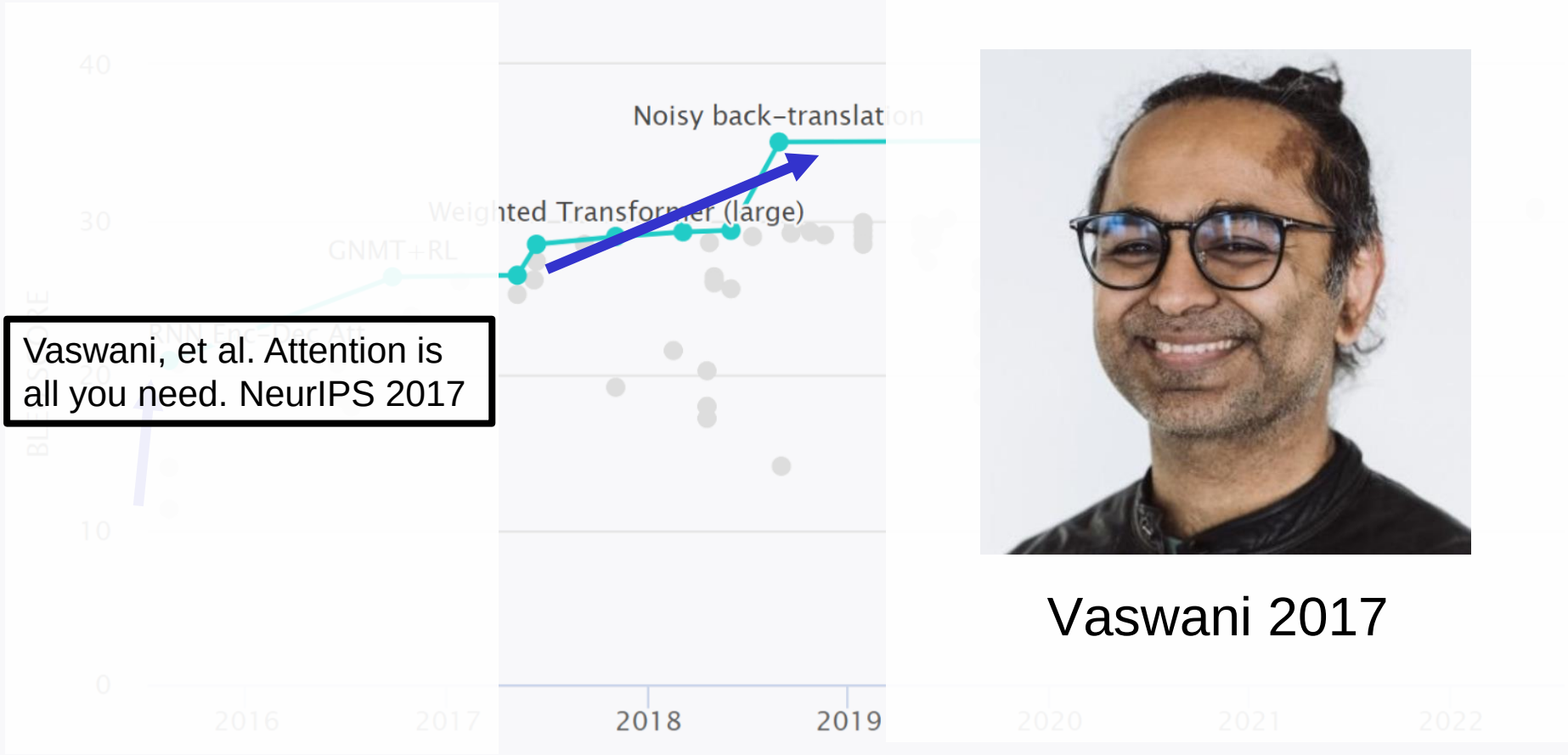
Ilya Sutskever 2014

Kyunghyun Cho 2014

Sutskever, et al. Sequence to Sequence Learning with Neural Networks. NeurIPS 2014

Cho, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. EMNLP 2014

Transformer



Vaswani, et al. Attention is all you need. NeurIPS 2017

Vaswani 2017

Turns out: Forms the basis of LLMs (e.g., ChatGPT).

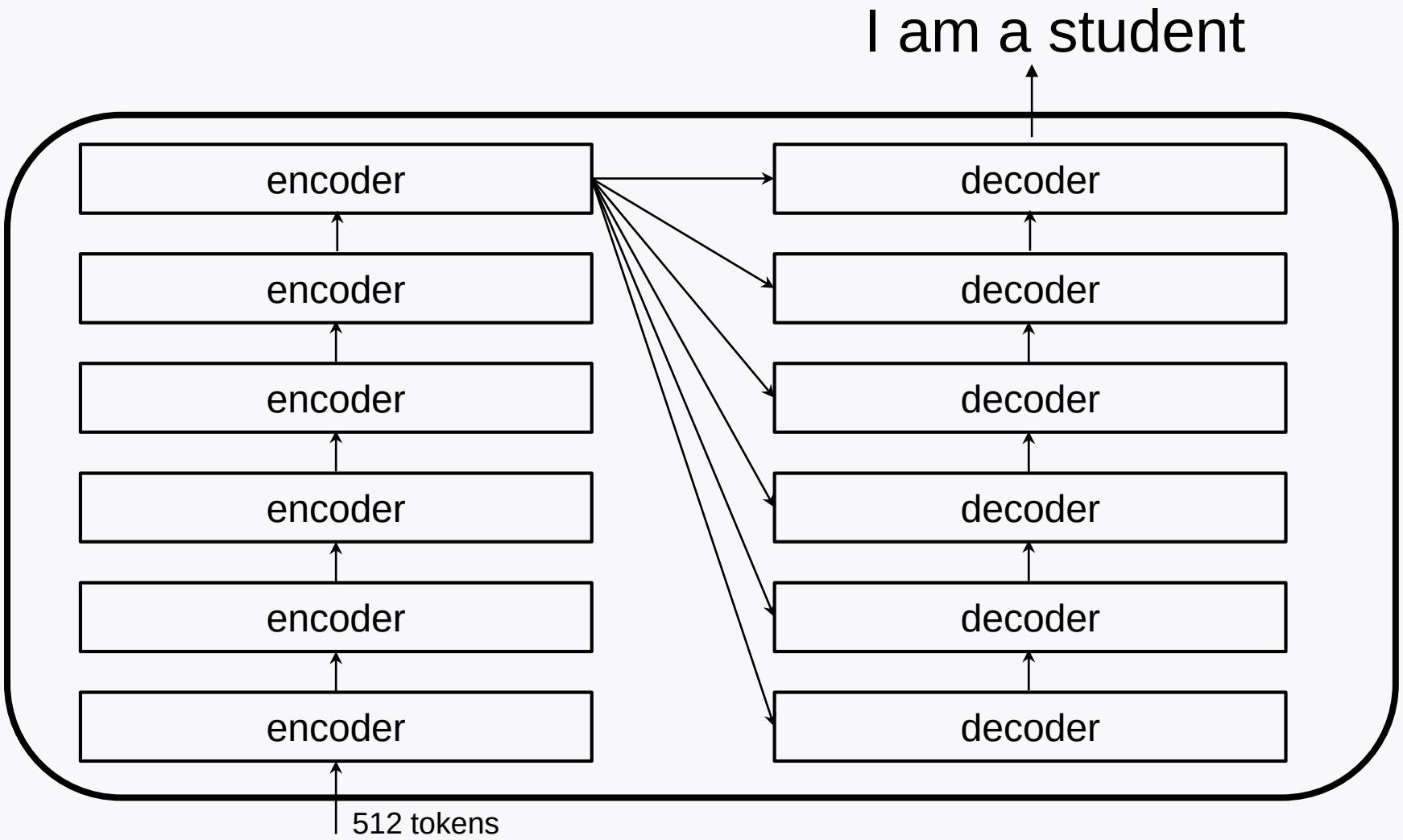
Outline

1. Study the Transformer architecture.

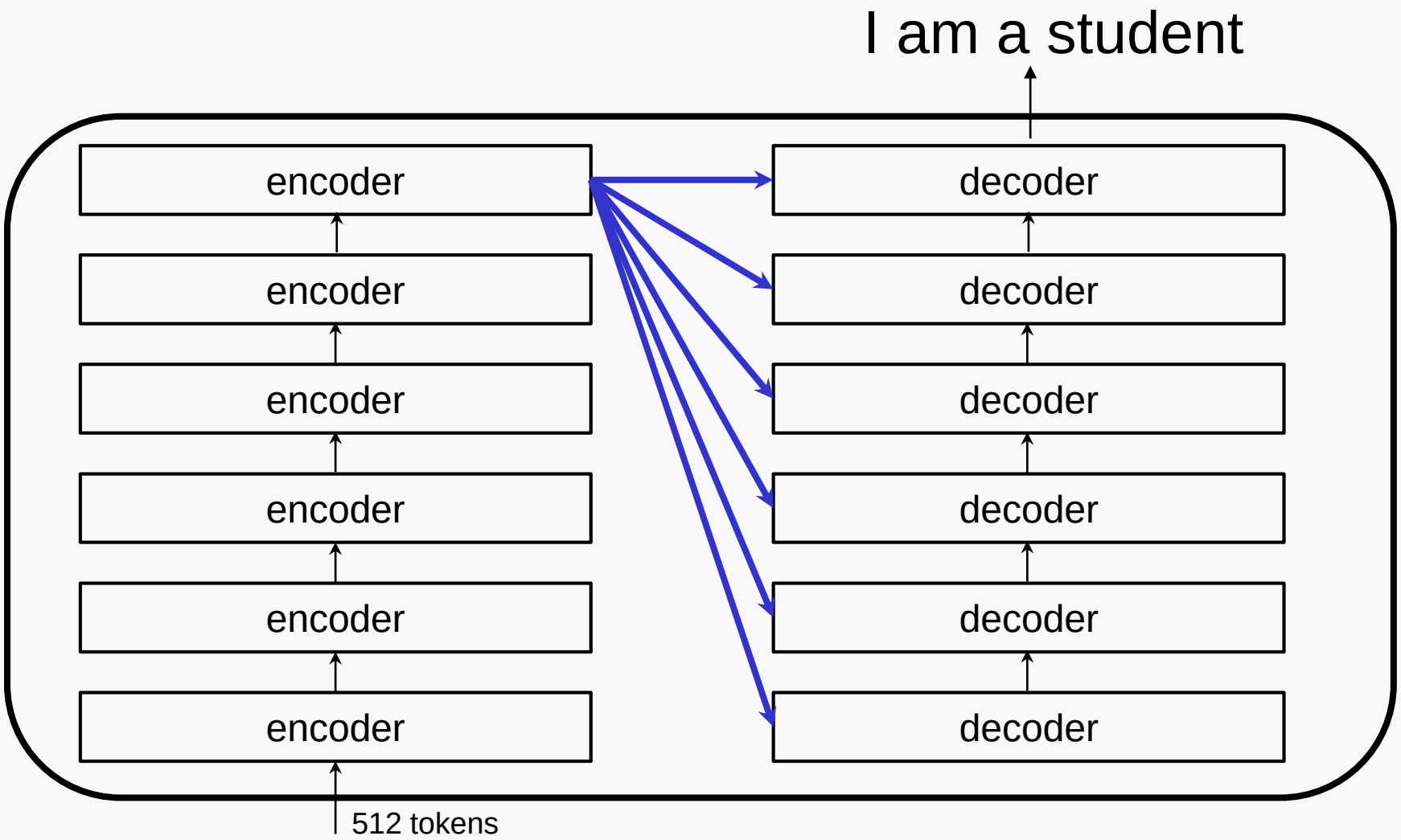
2. Explore OpenAI's LLMs (**GPT** series) based on the Transformer **decoder**.

3. Explore Google's LLMs (**BERT** and **RoBERTa**) based on the Transformer **encoder**.

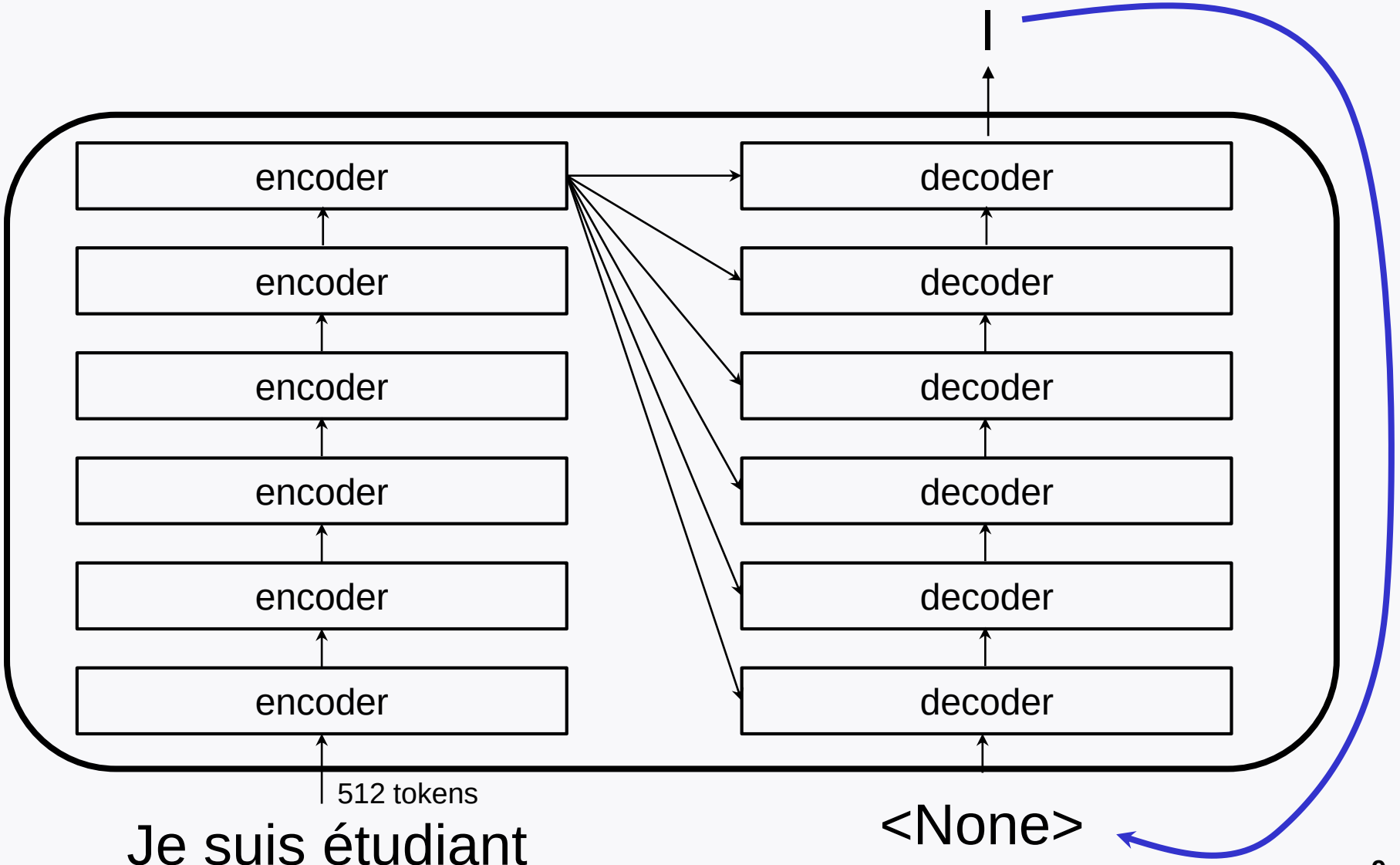
Transformer: A high-level architecture



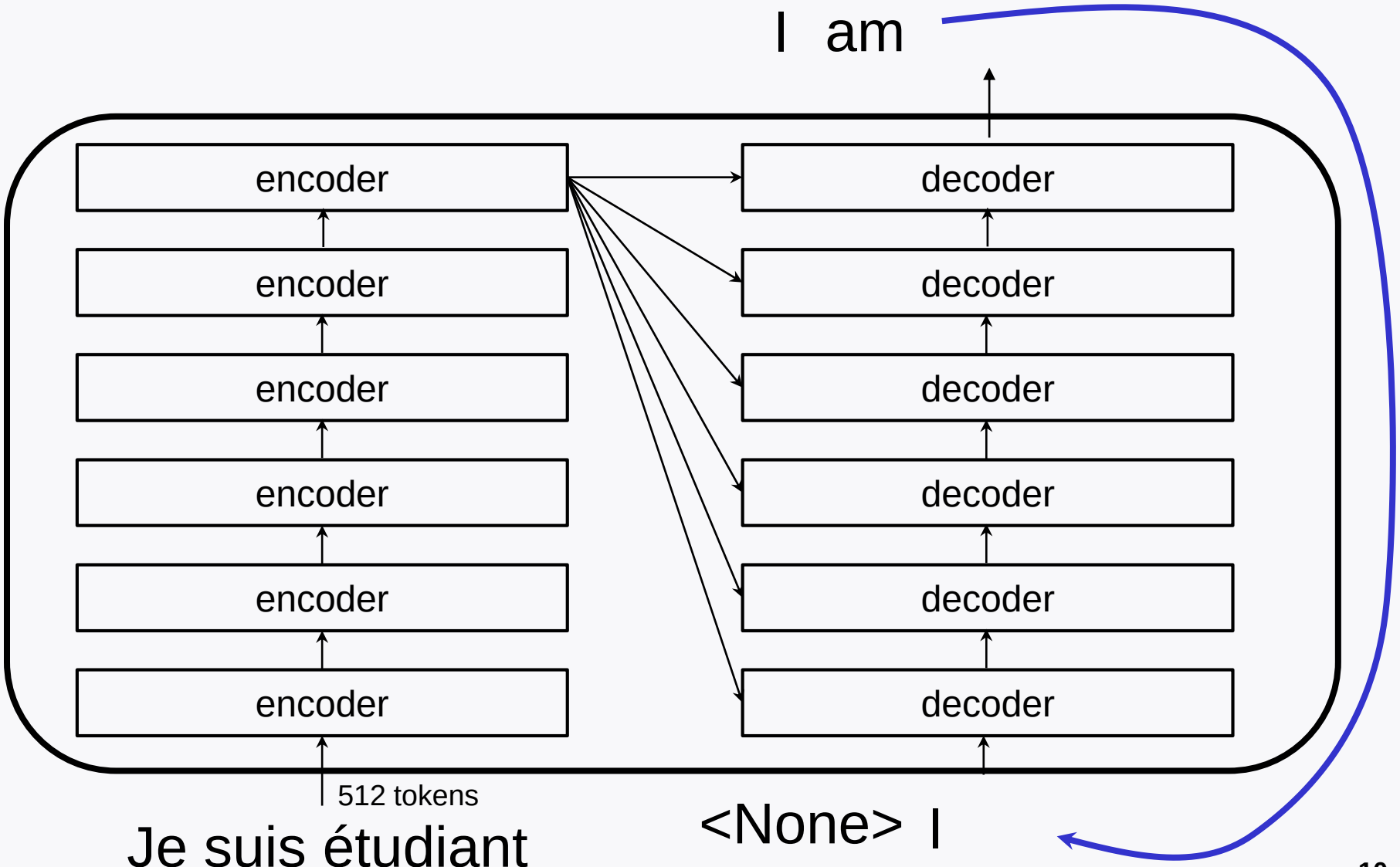
Feature #1: Encoder-decoder attention



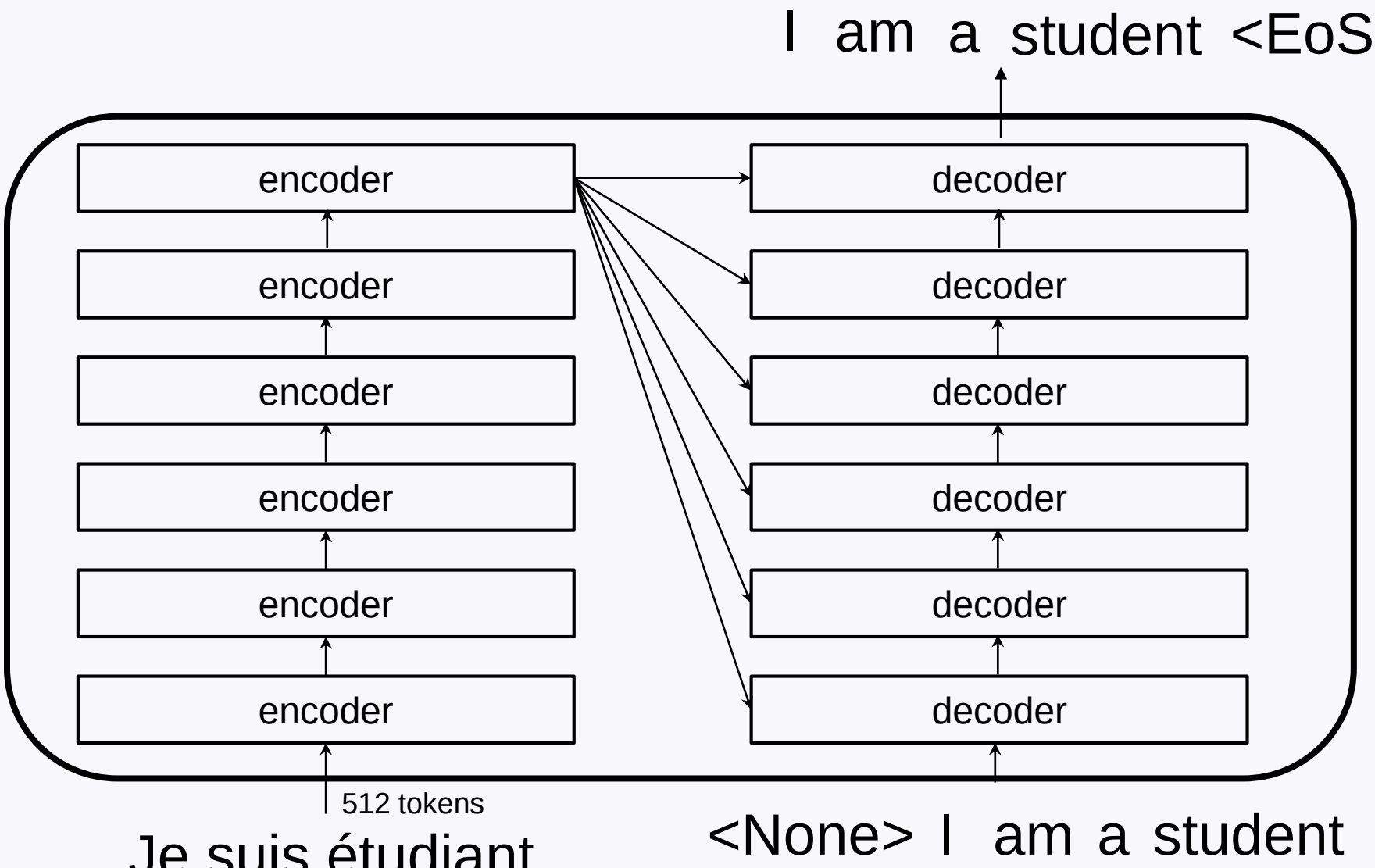
Feature #2: Recursion in decoders



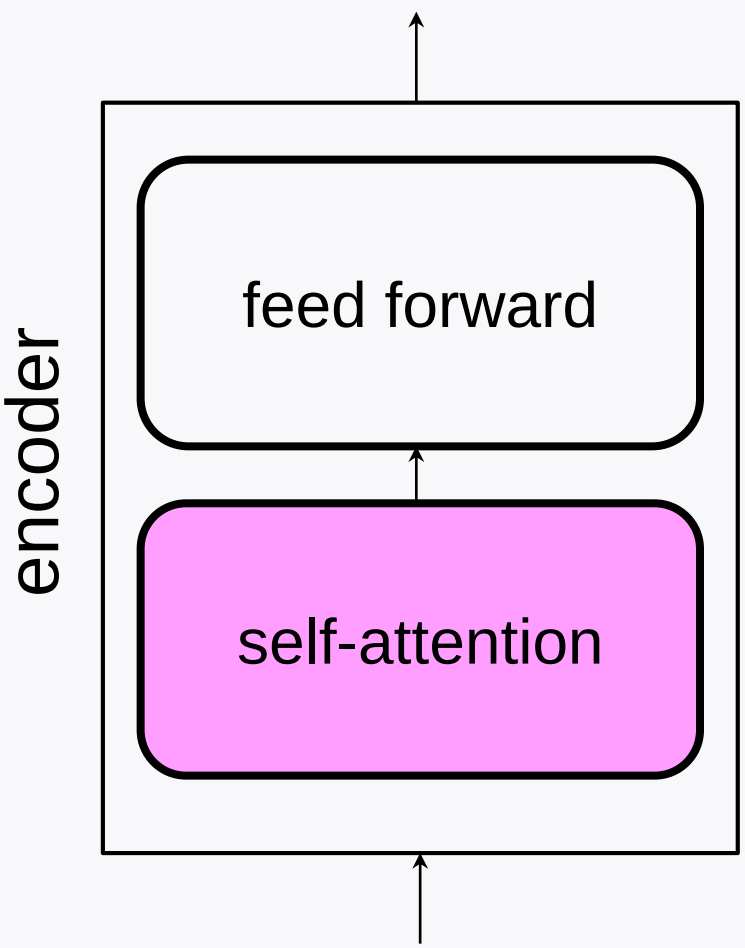
Feature #2: Recursion in decoders

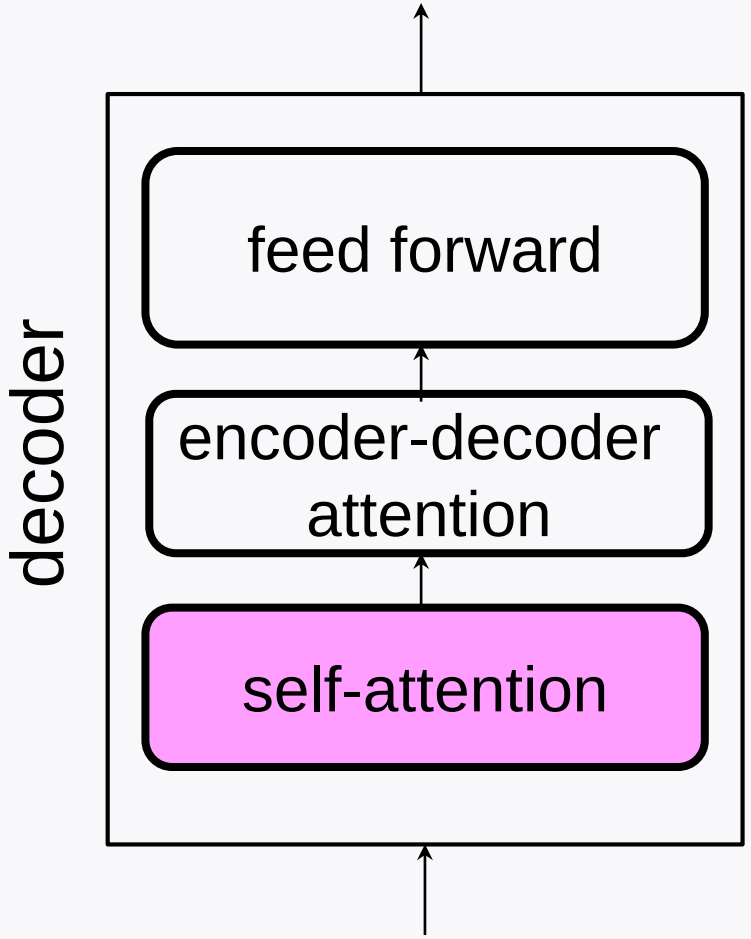


Feature #2: Recursion in decoders

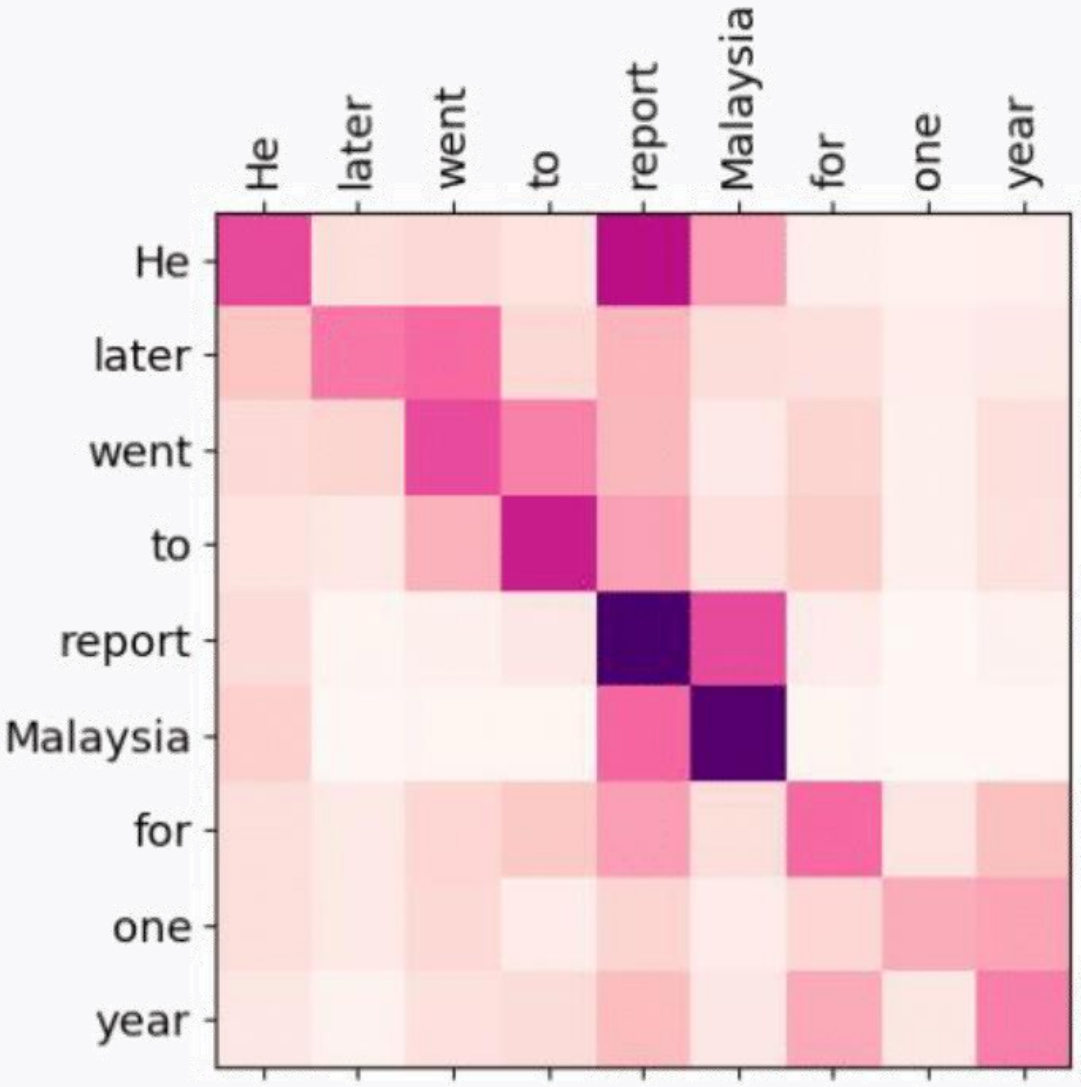


Feature #3: Attention layer

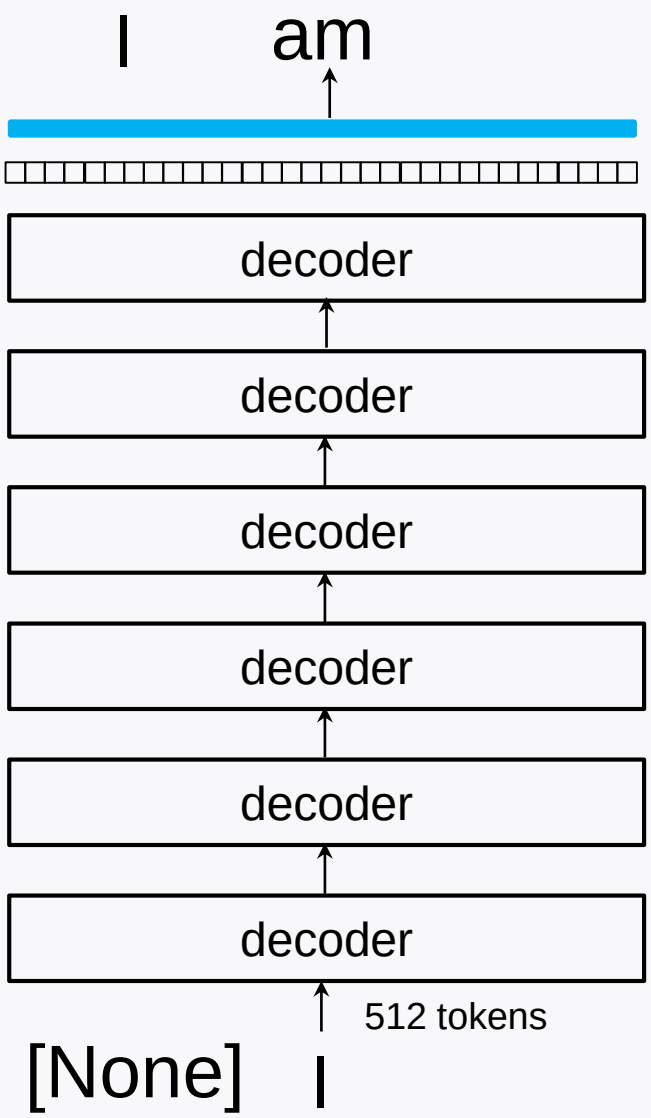




Feature #3: Attention layer (visualization)

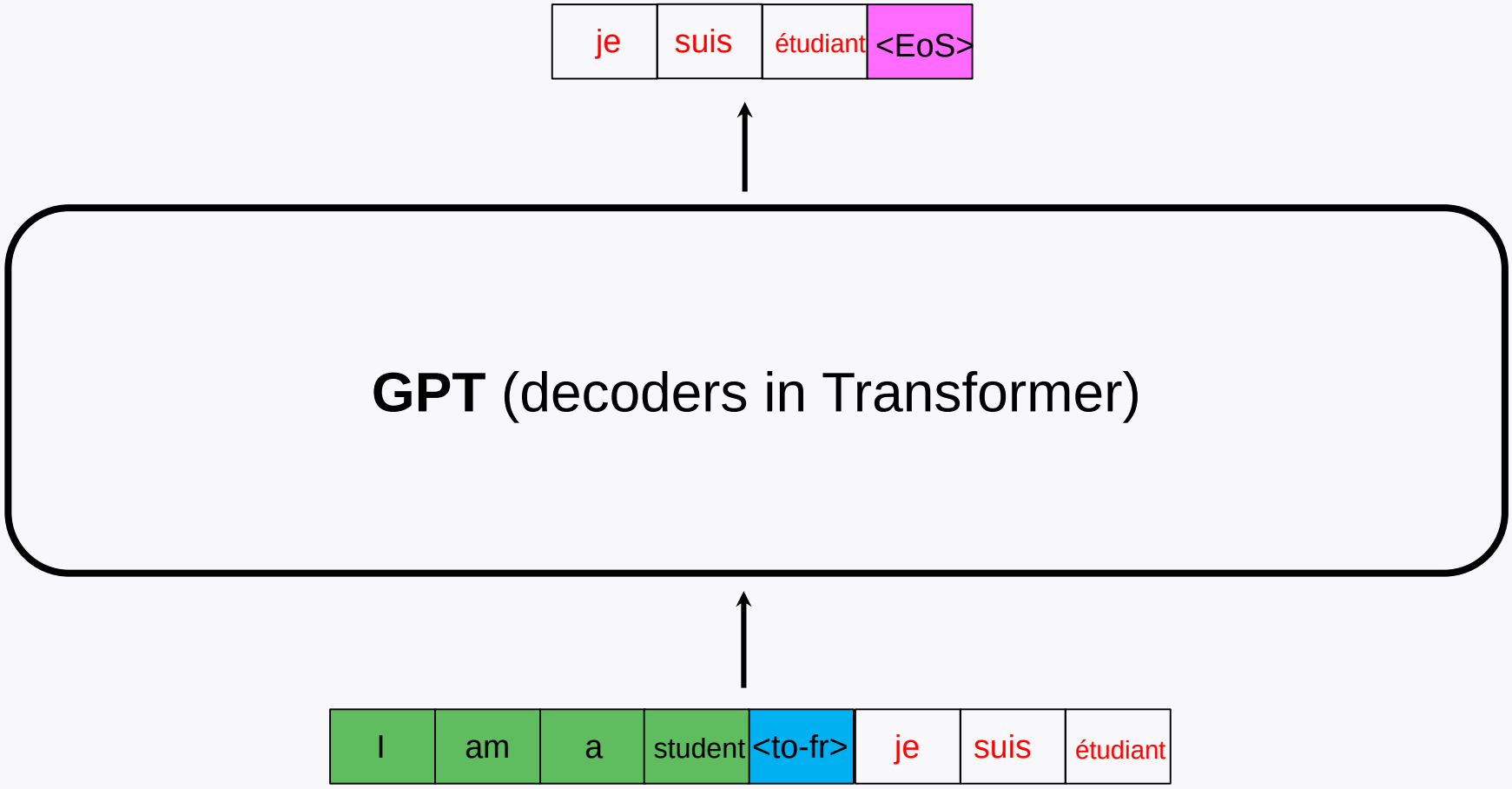


Feature #4: Dense & softmax layers in dec.



neurons = vocabulary size

GPT: Machine translation (How it works)



GPT: Machine translation (training dataset)

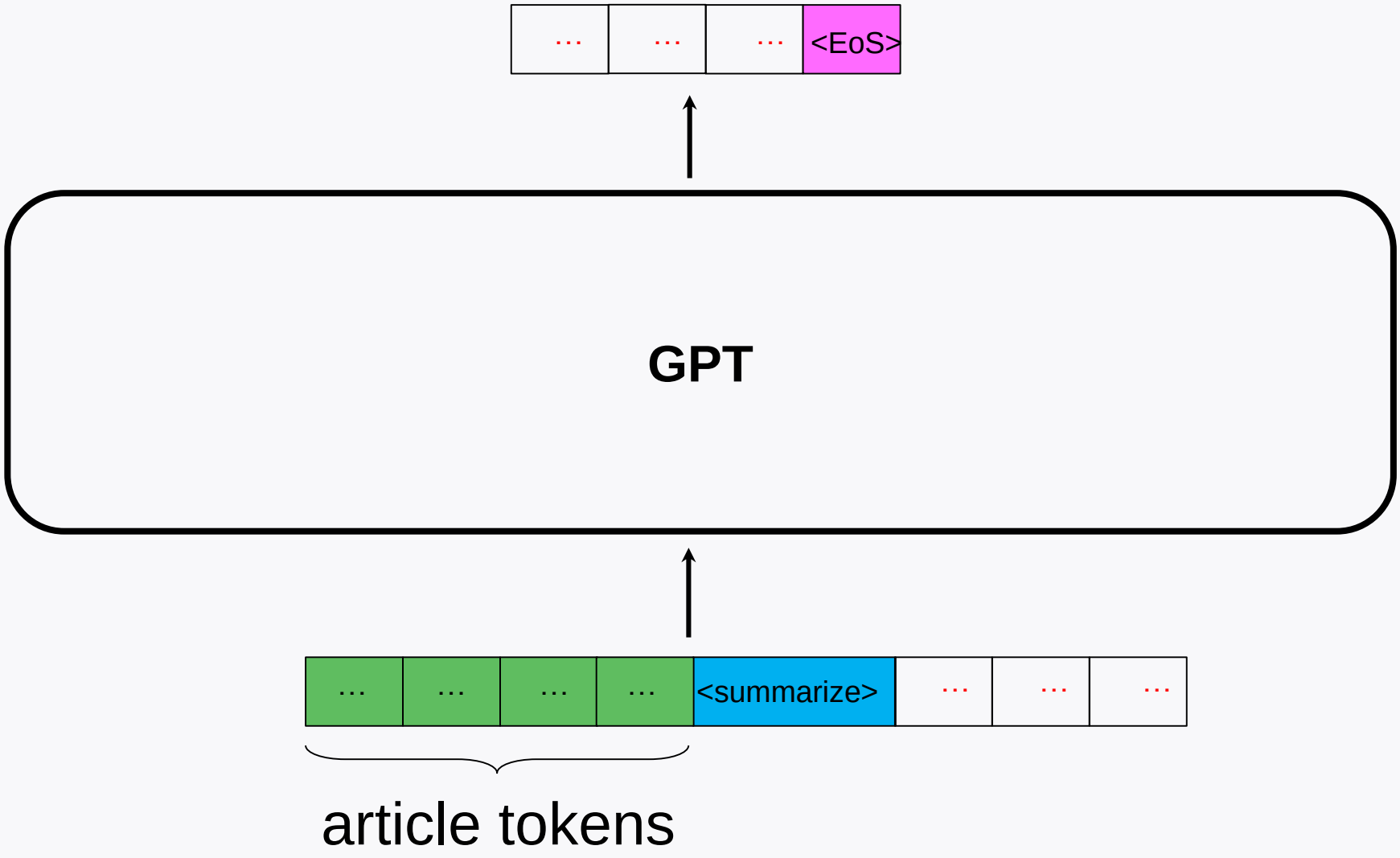
Original:

I	am	a	student	<to-fr>	je	suis	étudiant
let	them	eat	cake	<to-fr>	Qu'ils	mangent	de
good	morning	<to-fr>	Bonjour				

Manipulated:

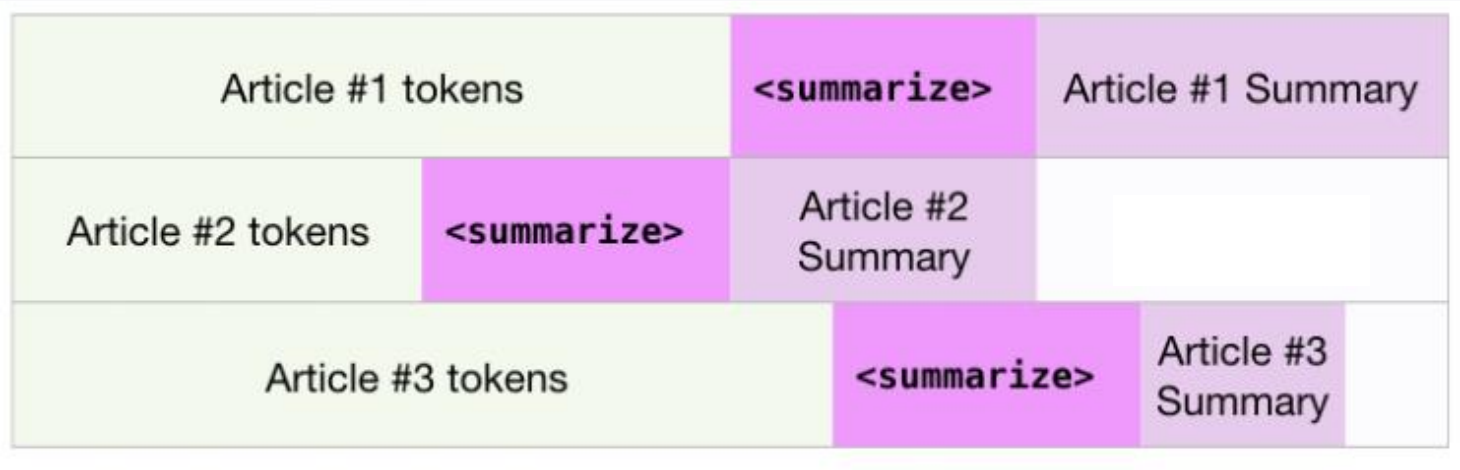
I	am	a	student	<to-fr>				→	je
I	am	a	student	<to-fr>	je			→	suis
I	am	a	student	<to-fr>	je	suis		→	étudiant
I	am	a	student	<to-fr>	je	suis	étudiant	→	<EoS>

GPT: Text summarization

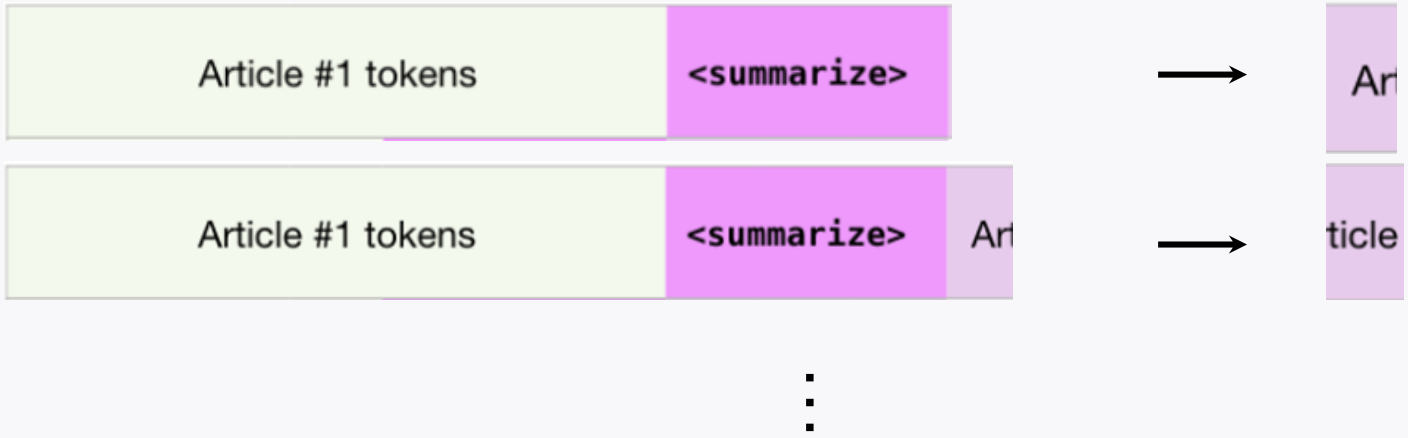


GPT: Text summarization (training dataset)

Original:



Manipulated:



GPT series

GPT (2018): 110 M parameters
512 tokens

GPT-2 (2019): 117M ~ **1542M** parameters
768 ~ 1600 tokens

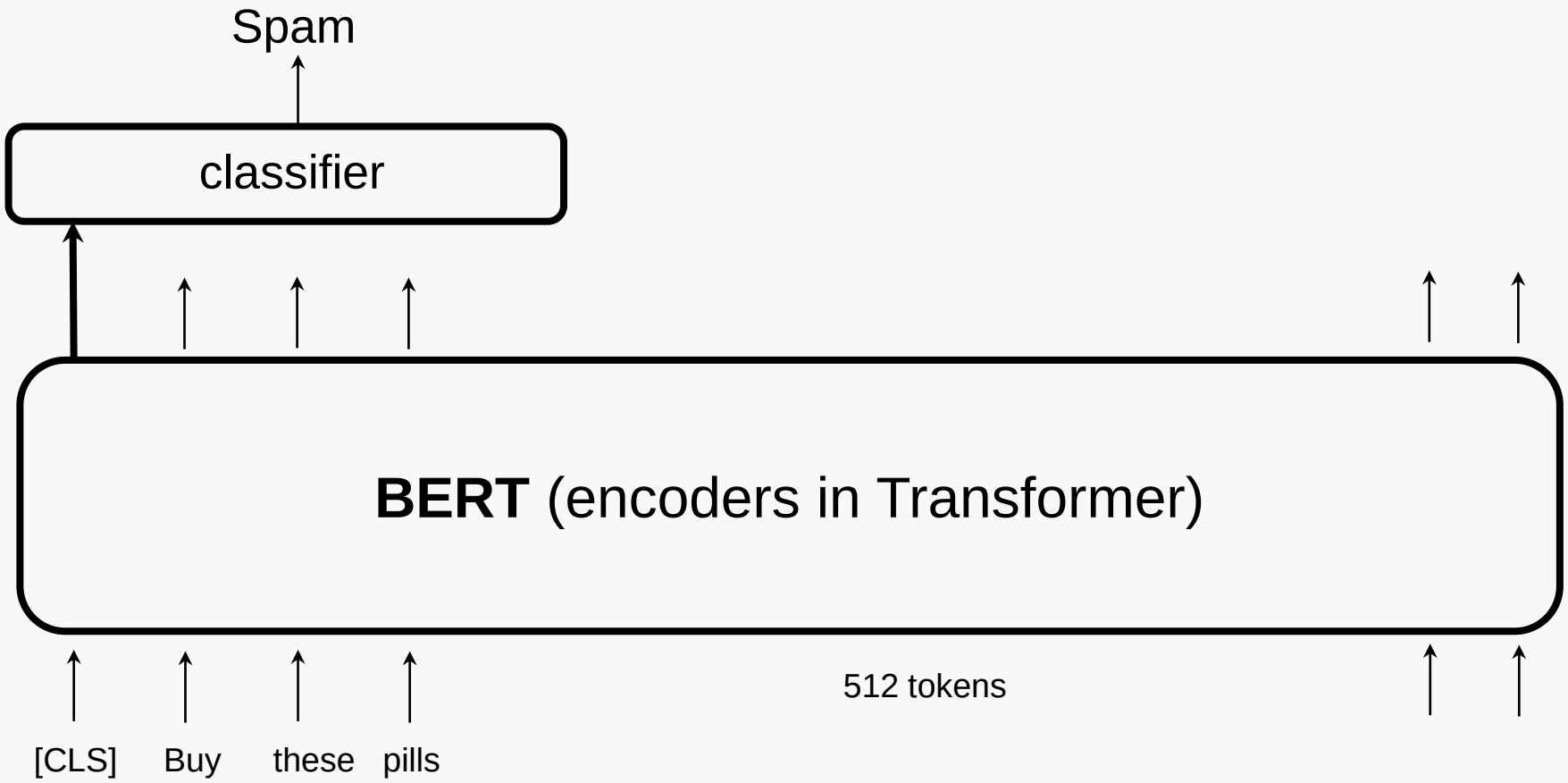
GPT-3 (2020): 125M ~ **175B** parameters
768 ~ 12288 tokens

GPT-3.5=ChatGPT (2022): Instructed GPT-3

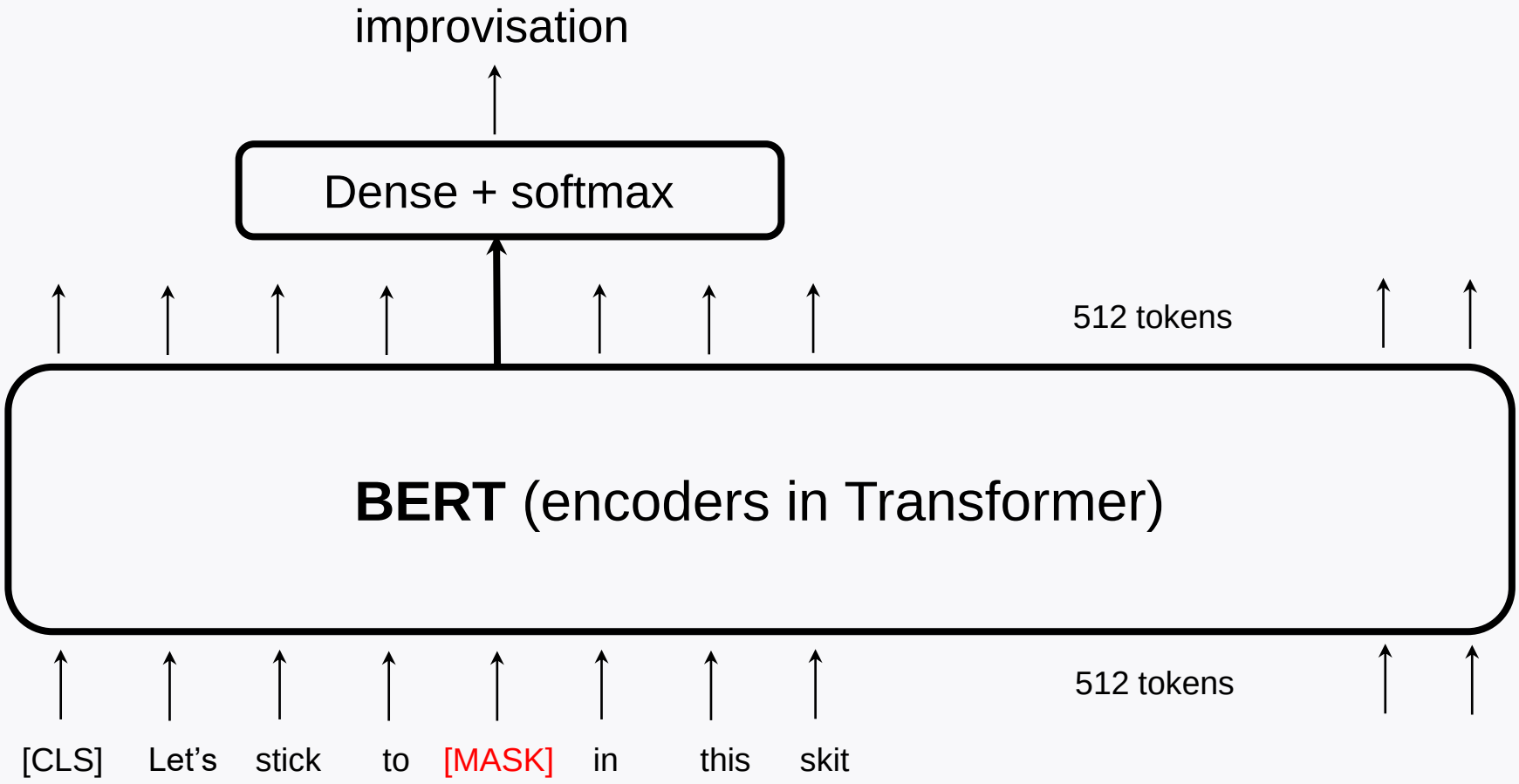
GPT-4 (2023): ~**1.8T** parameters
~ 25000 tokens

BERT: Classification

Bidirectional **E**ncoder **R**epresentations from **T**ransformers



BERT: Word prediction



data: Let's stick to improvisation in this skit

BERT and RoBERTa

BERT (2018): 4.4M ~ 340M parameters
512 tokens

RoBERTa (2019): 125M ~ 355M parameters
512 tokens

Robustly **o**ptimized **BERT a**pproach

Turns out:

Forms the basis of Google's LLMs (e.g., BARD).