Transformers:
The Era of Rapid Scaling in NLP

Berkeley NLP

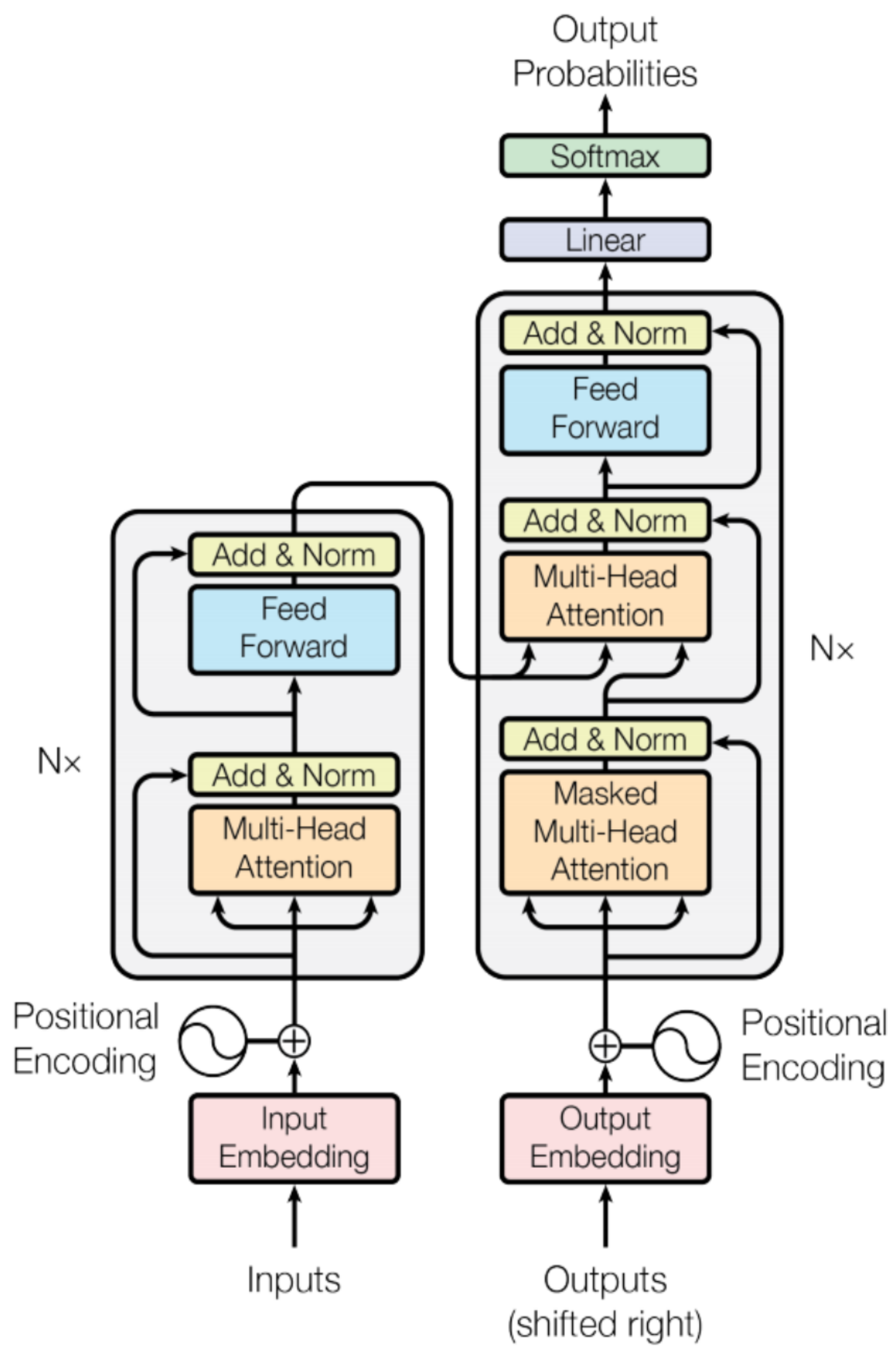
Nikita Kitaev

February 22, 2022

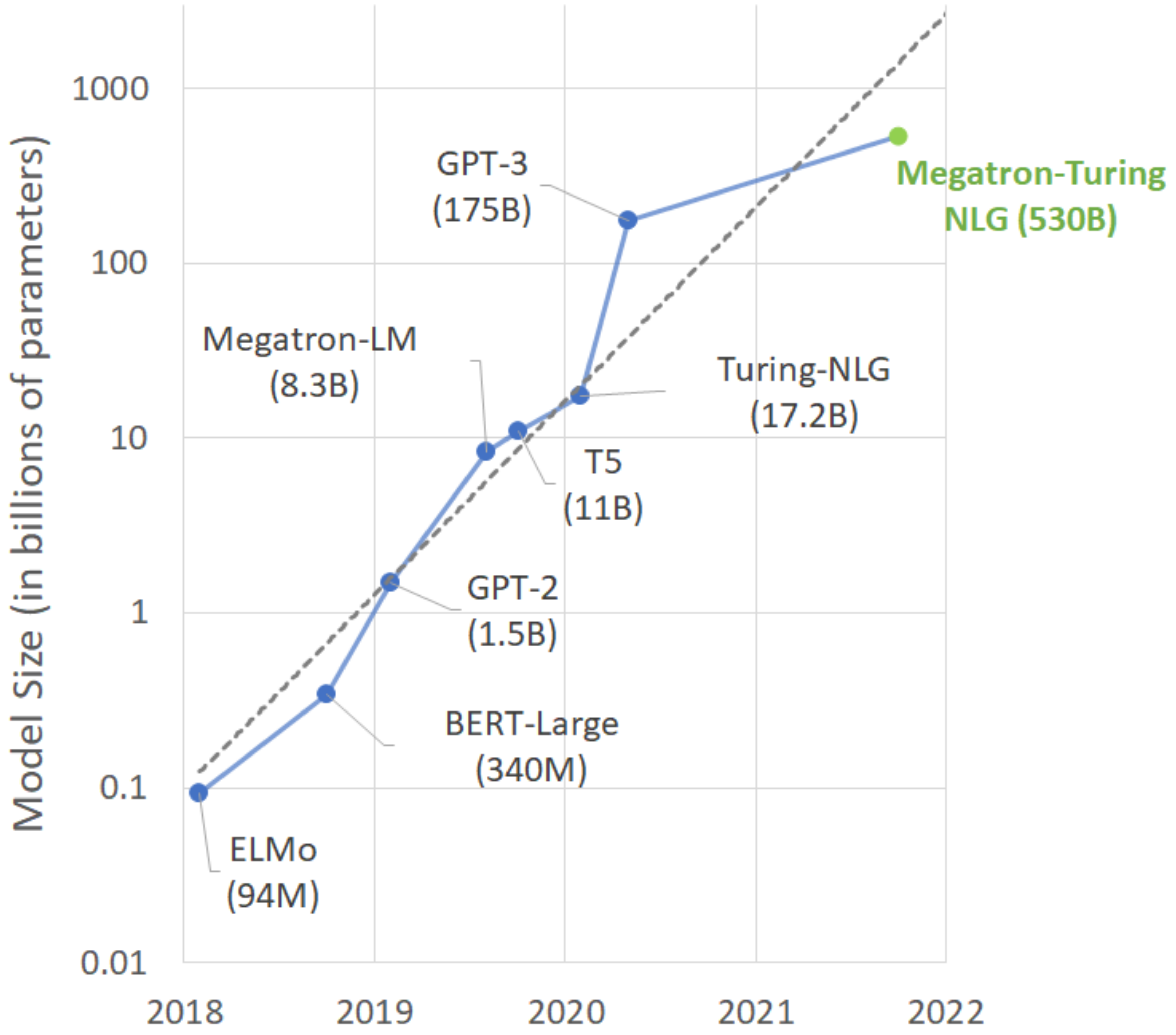
The Era of Rapid Scaling in NLP

2017: Transformer is introduced

[Vaswani+17] Attention is All You Need



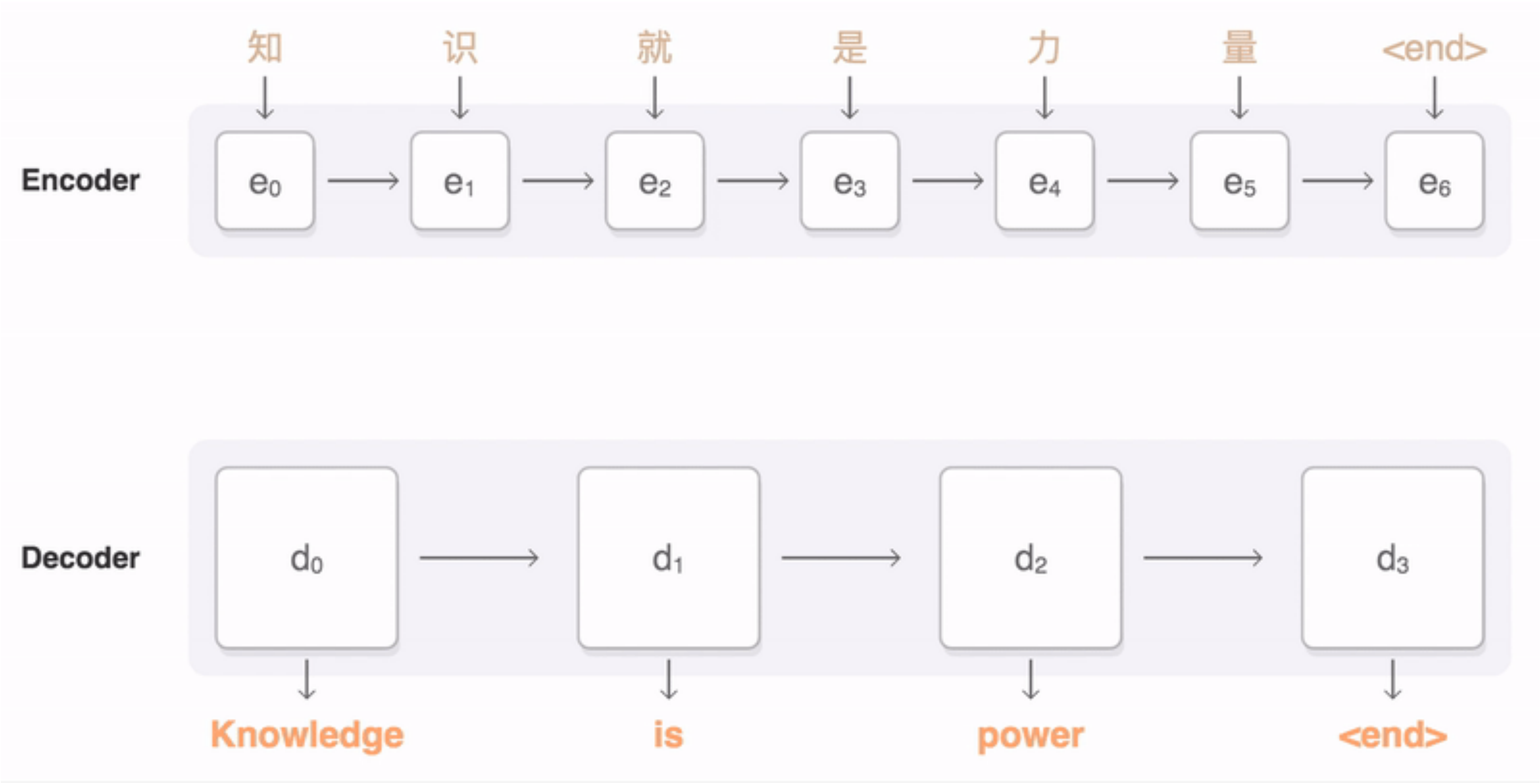
2022: Large-scale Transformer models are the dominant approach for many NLP tasks



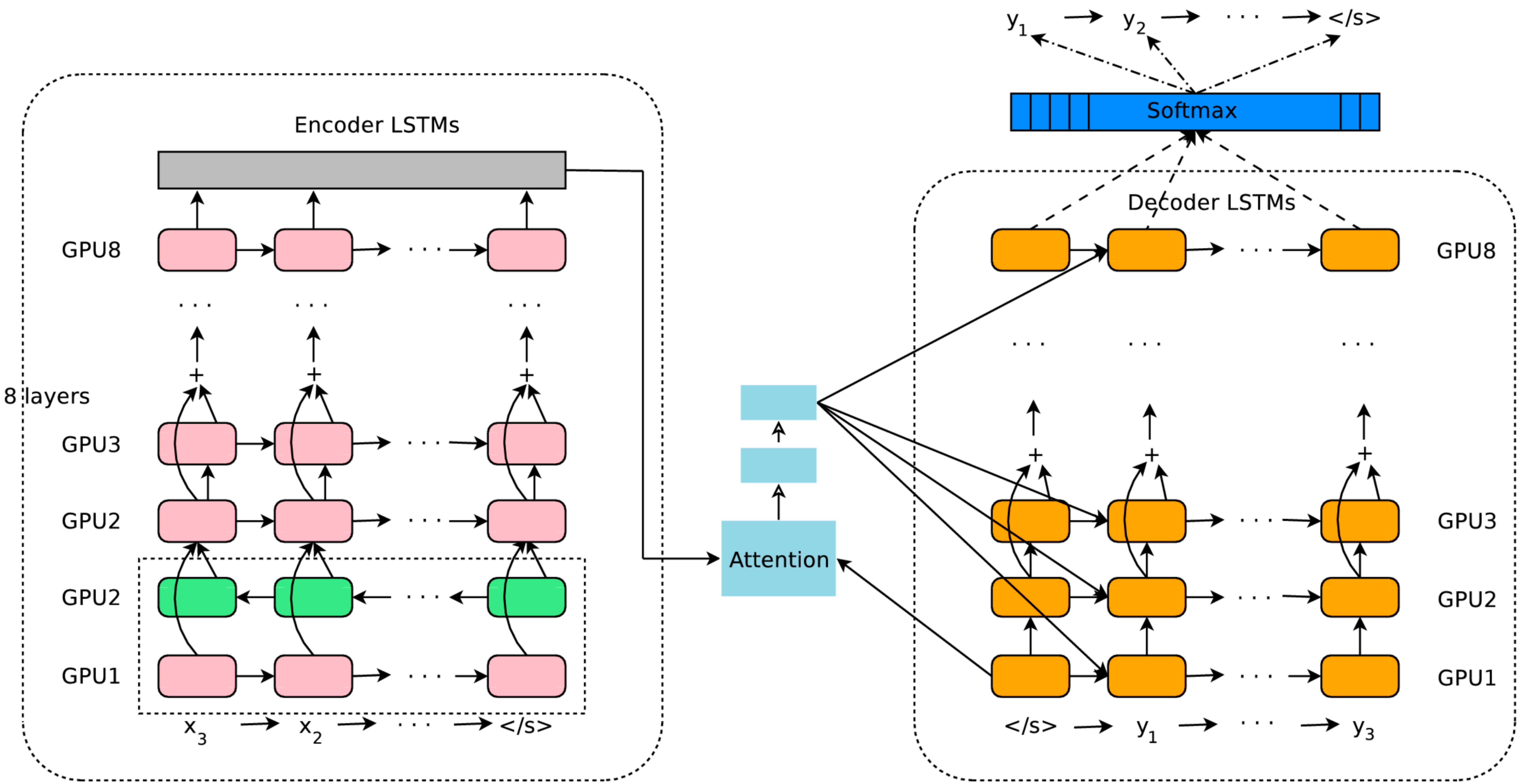
Neural MT ca. 2016

Neural Machine Translation is in production at Google

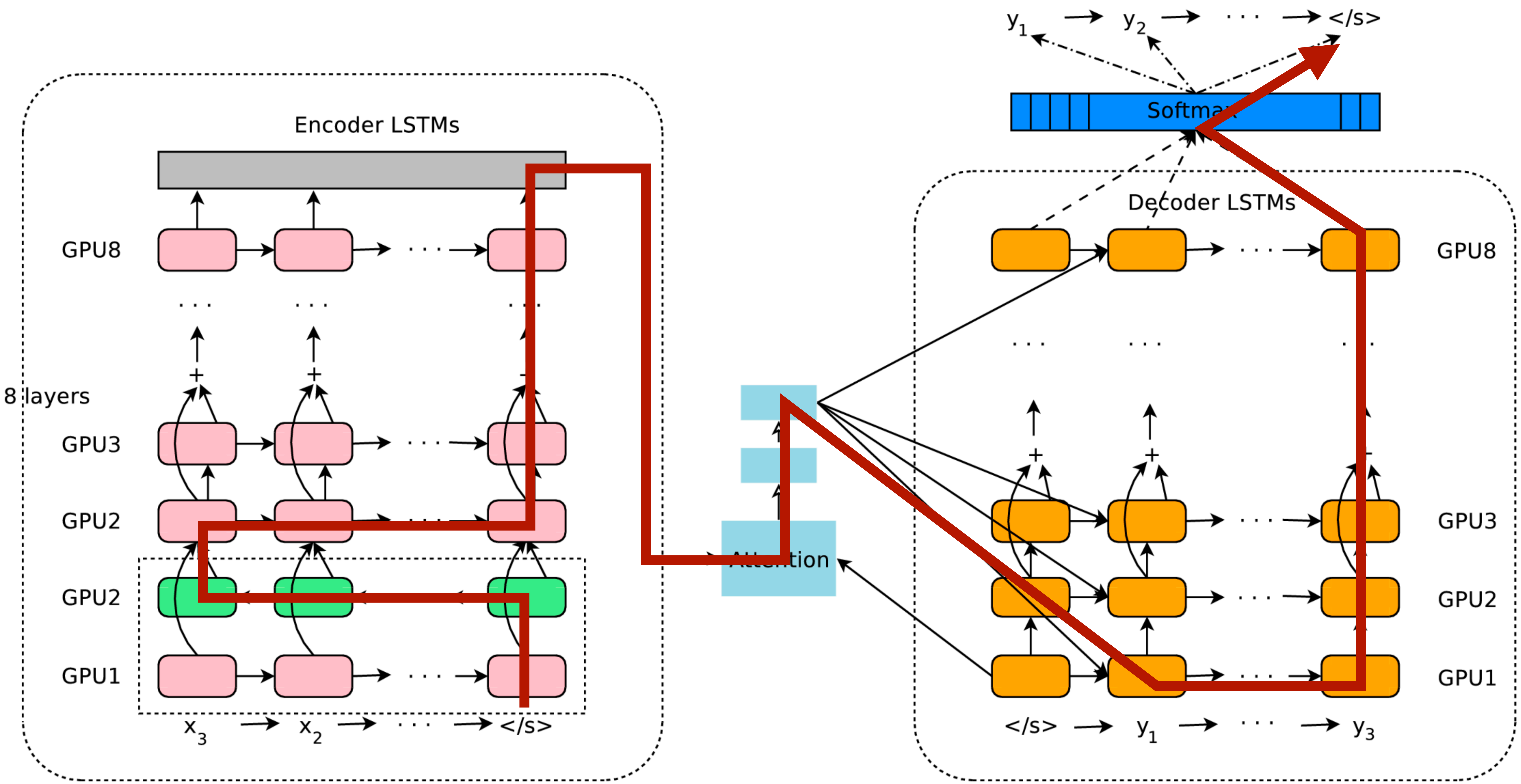
[Wu+16] Google's Neural Machine Translation System:
Bridging the Gap between Human and Machine Translation



Neural MT ca. 2016



Neural MT ca. 2016



There are computation paths through the RNN-based network that scale linearly with the sequence length, and can't be parallelized.

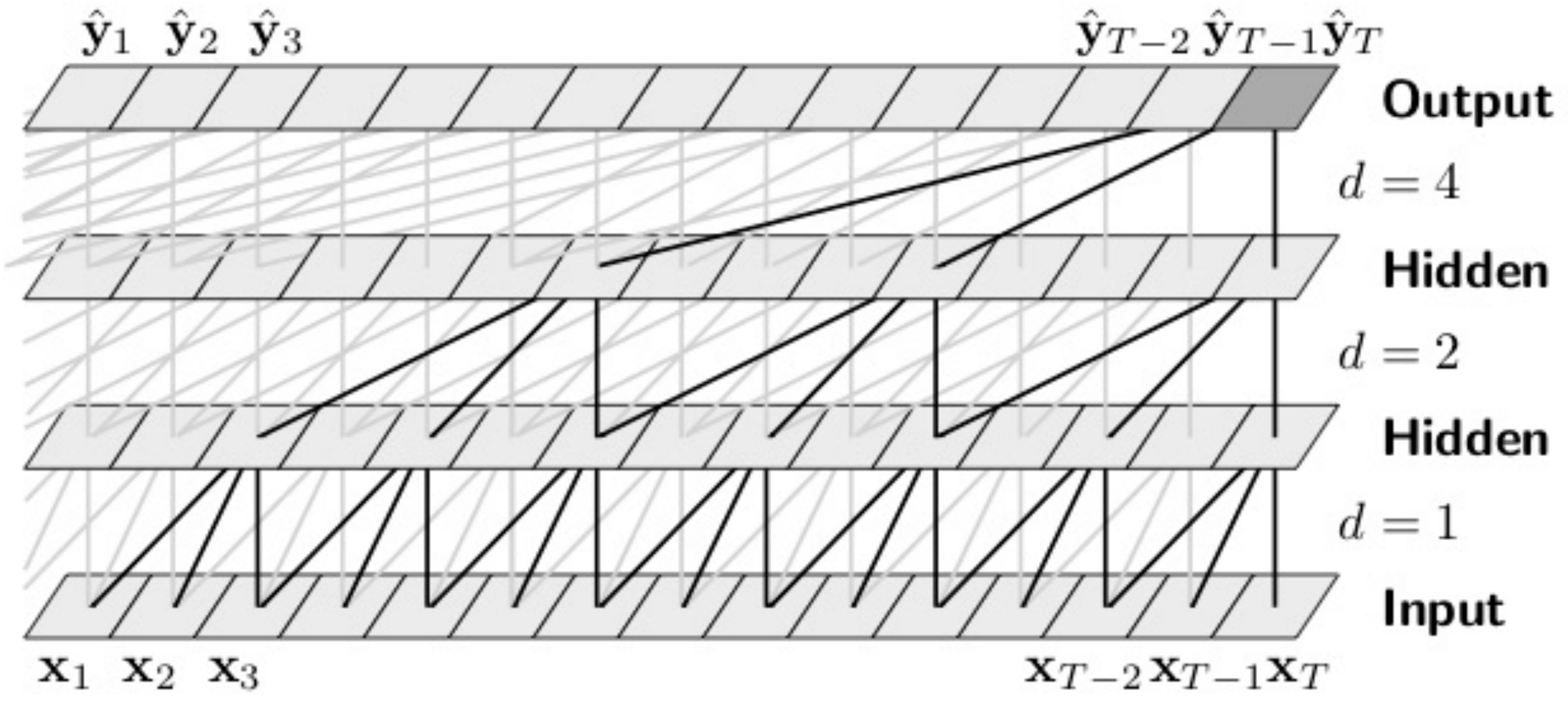
Maximum Path Length

RNN: #tokens * #layers

Maximum Path Length

RNN: #tokens * #layers

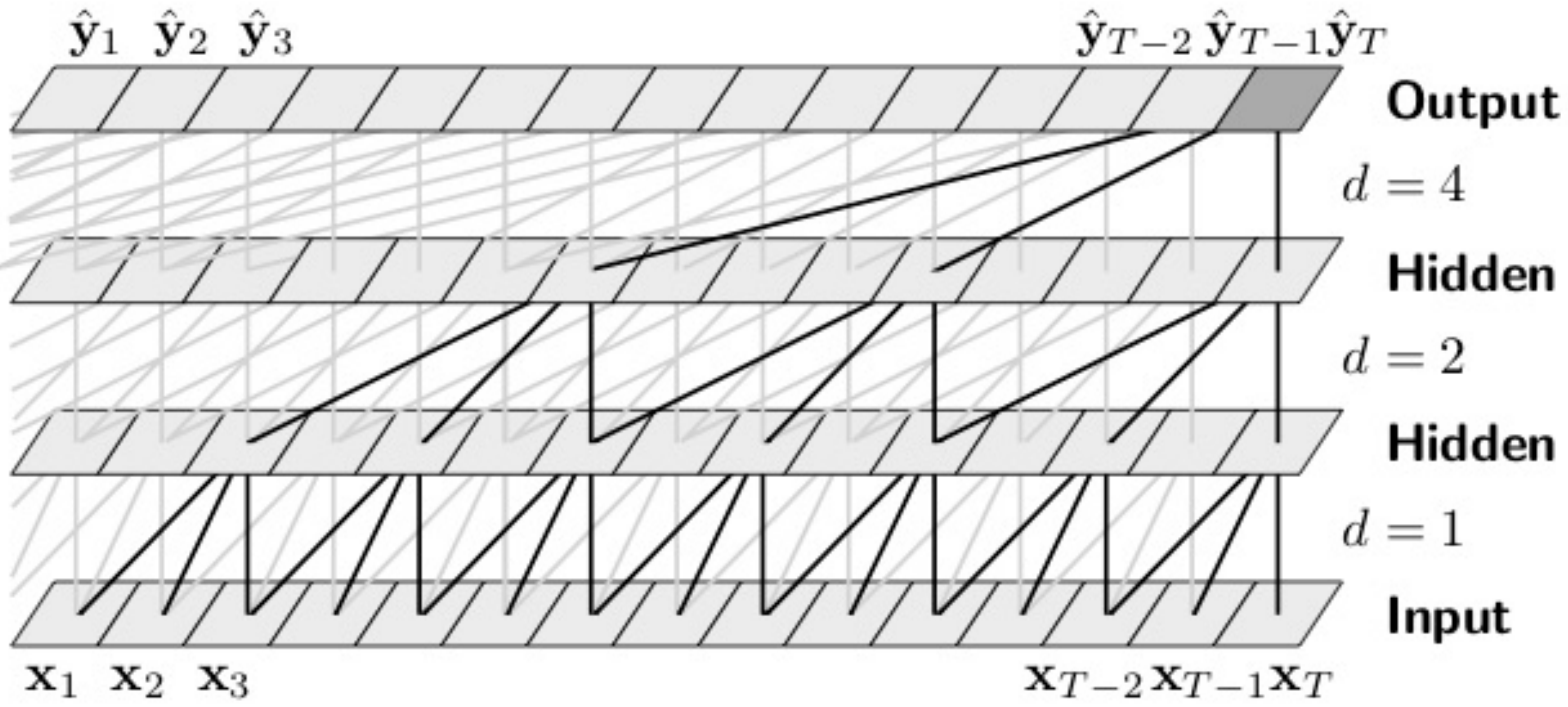
What about a Convolutional Neural Network?



Maximum Path Length

RNN: #tokens * #layers

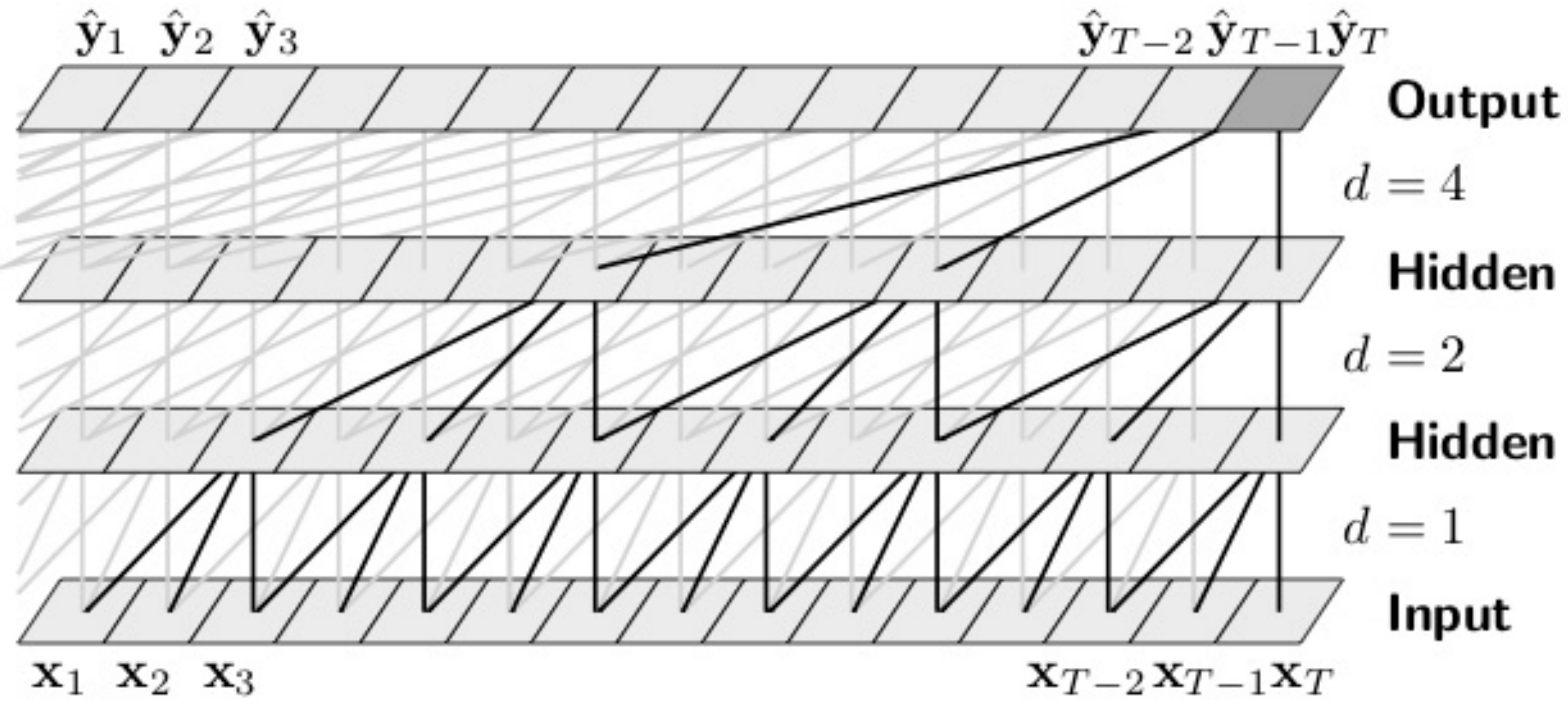
Convolutional: #layers *-- but we need to connect all tokens*



Maximum Path Length

RNN: #tokens * #layers

Convolutional: $\log_{\text{kernel size}}(\text{\#tokens})$



Maximum Path Length

RNN: #tokens * #layers

Convolutional: $\log_{\text{kernel size}}(\text{\#tokens})$

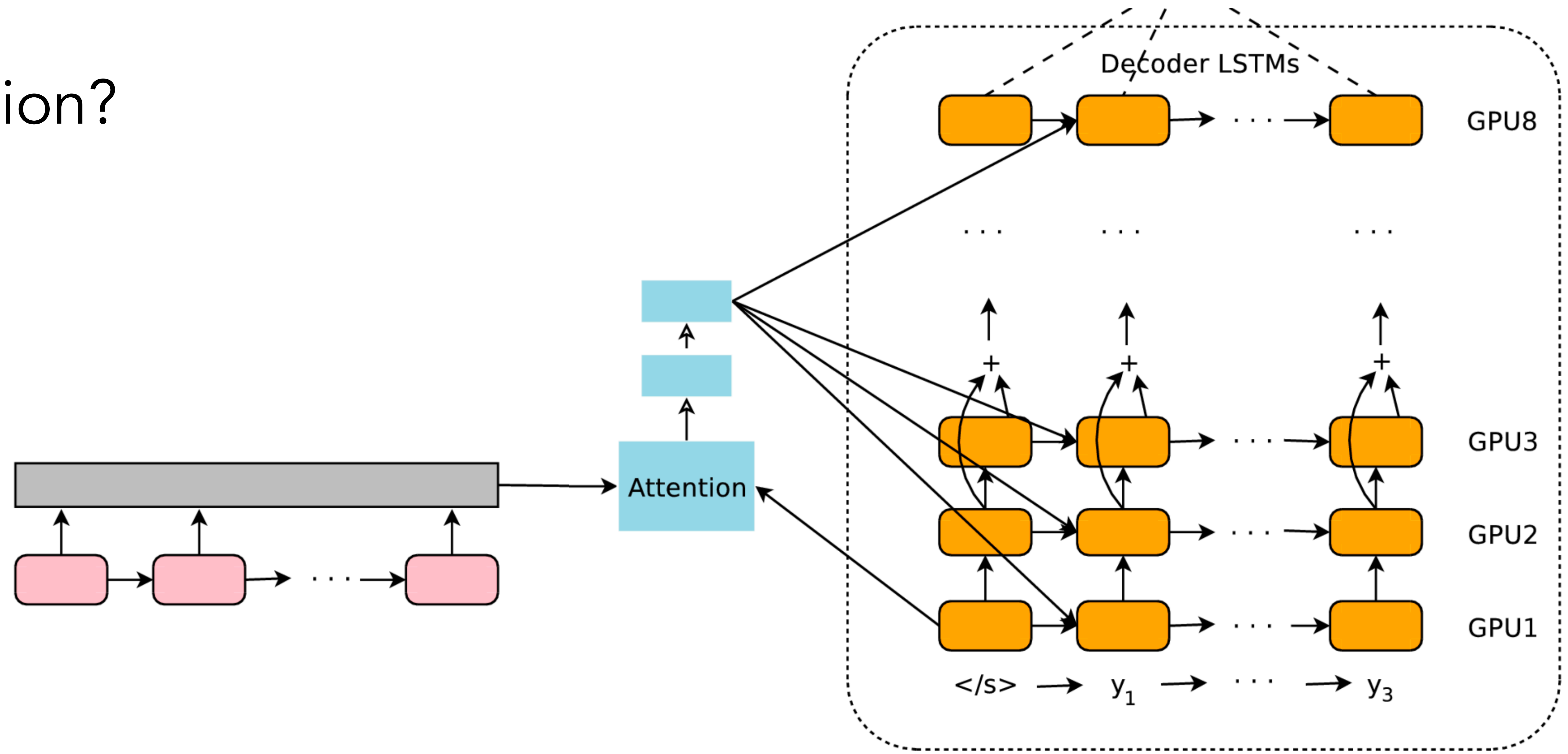
Any other alternatives?

Maximum Path Length

RNN: #tokens * #layers

Convolutional: $\log_{\text{kernel size}}(\#\text{tokens})$

How about attention?

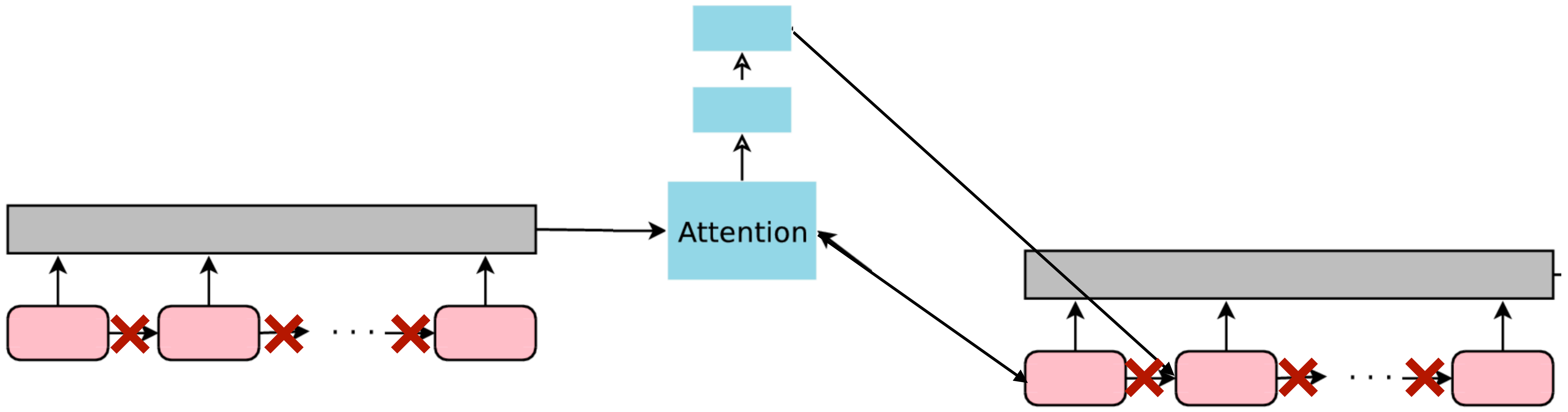


Maximum Path Length

RNN: #tokens * #layers

Convolutional: $\log_{\text{kernel size}}(\#\text{tokens})$

How about attention?

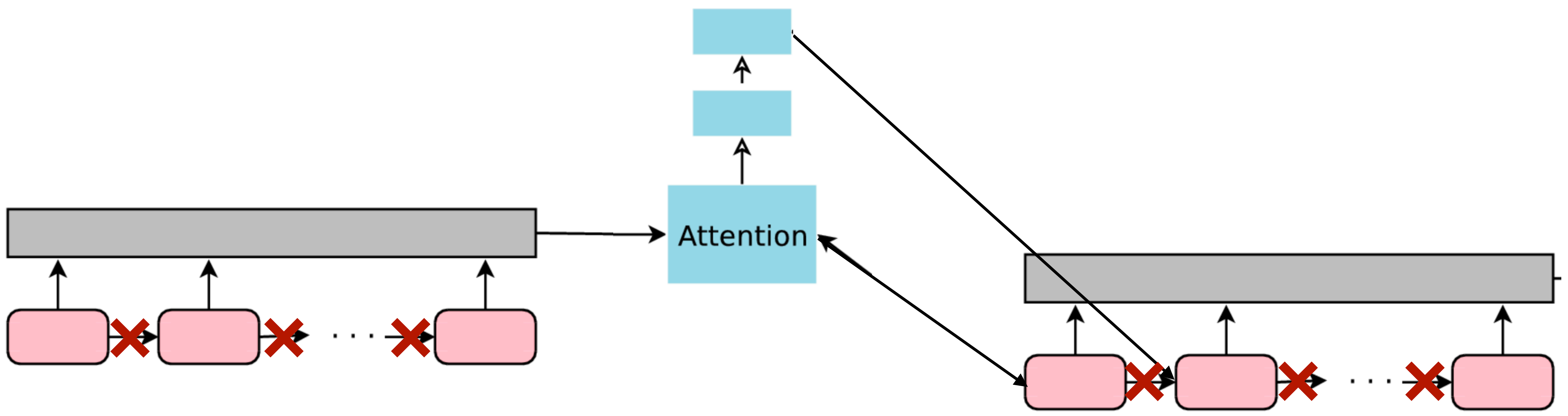


Maximum Path Length

RNN: #tokens * #layers

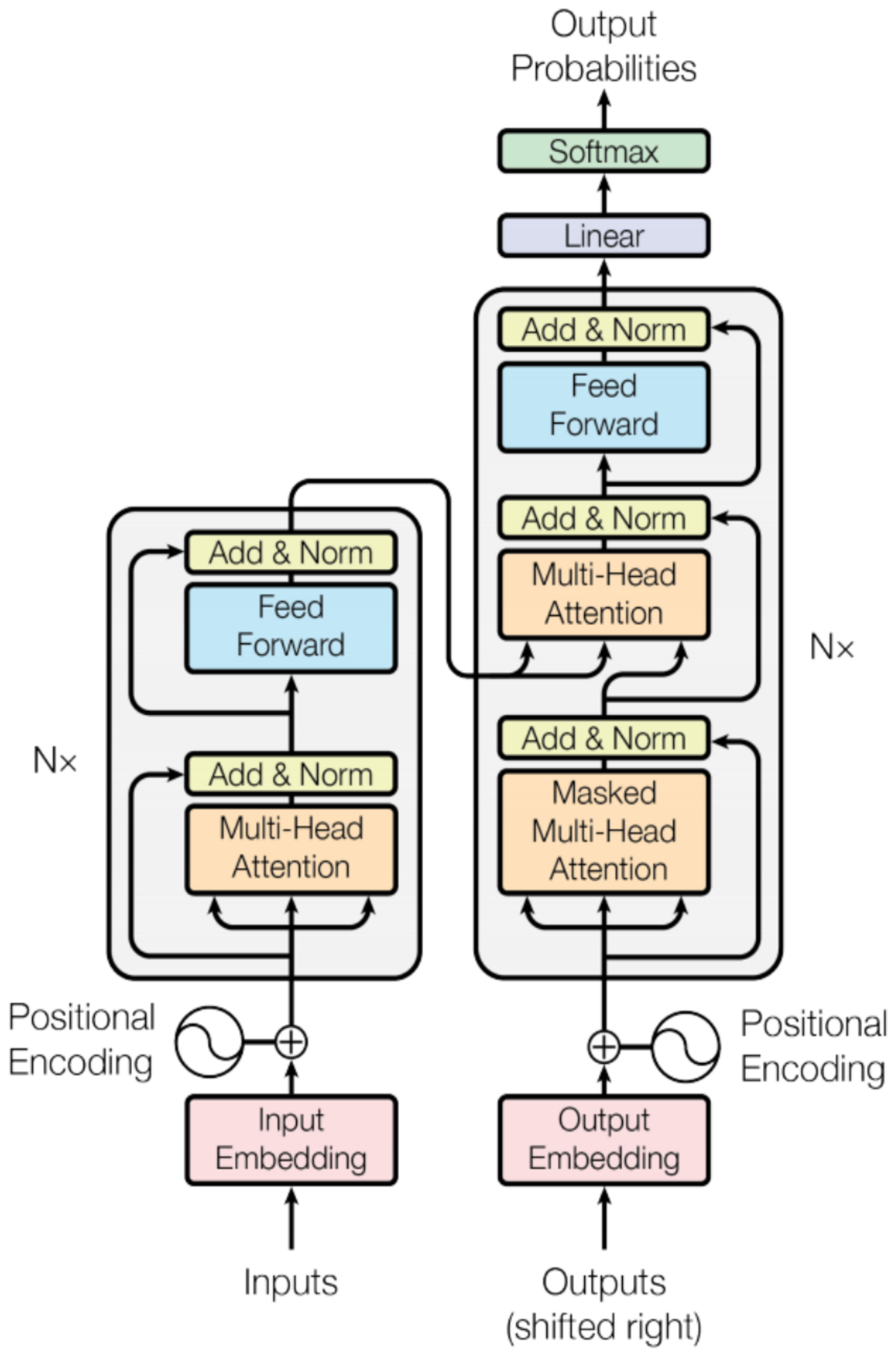
Convolutional: $\log_{\text{kernel size}}(\#\text{tokens})$

Attention: #layers

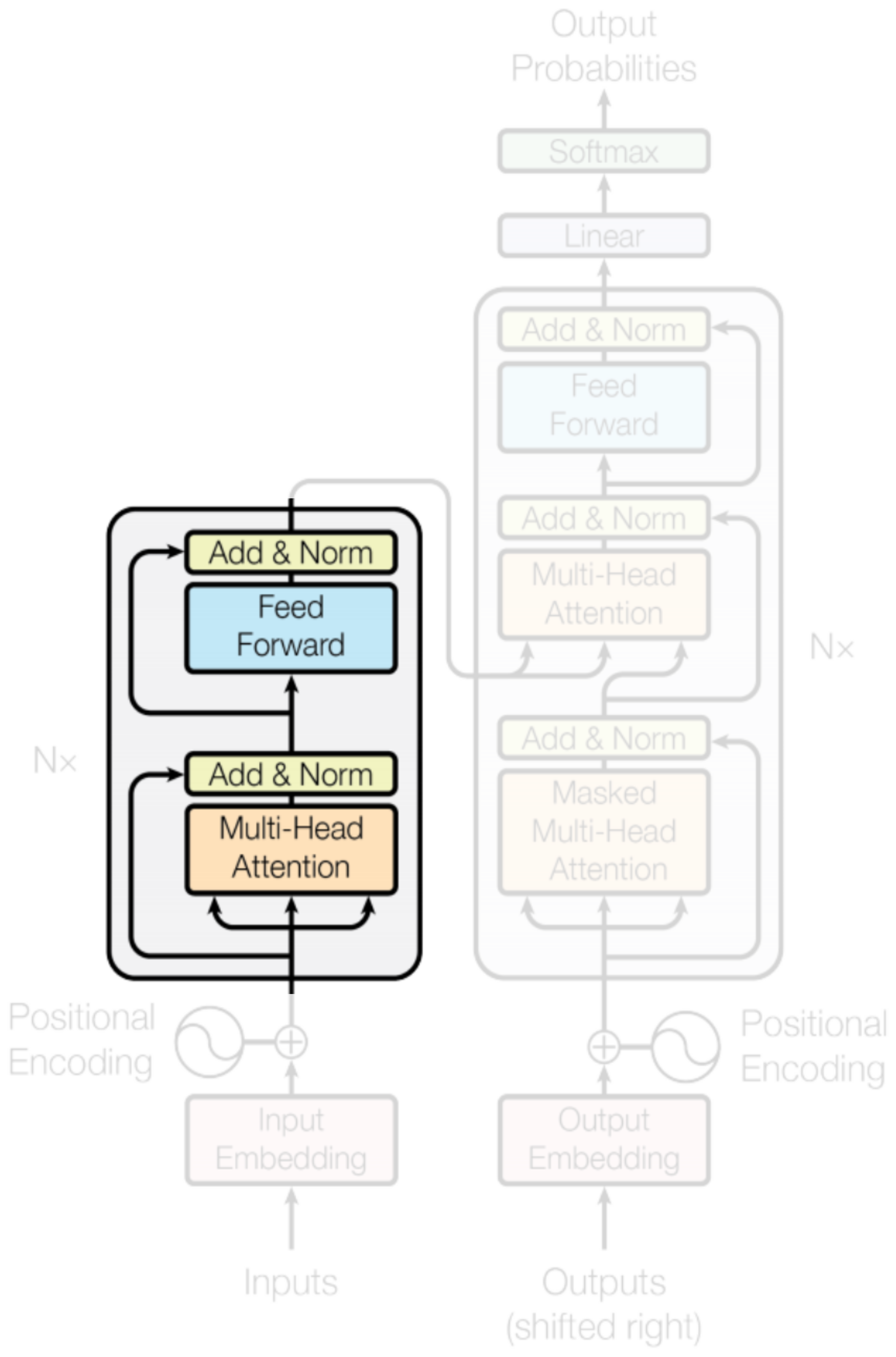


(1) Transformer Architecture

Transformer Architecture

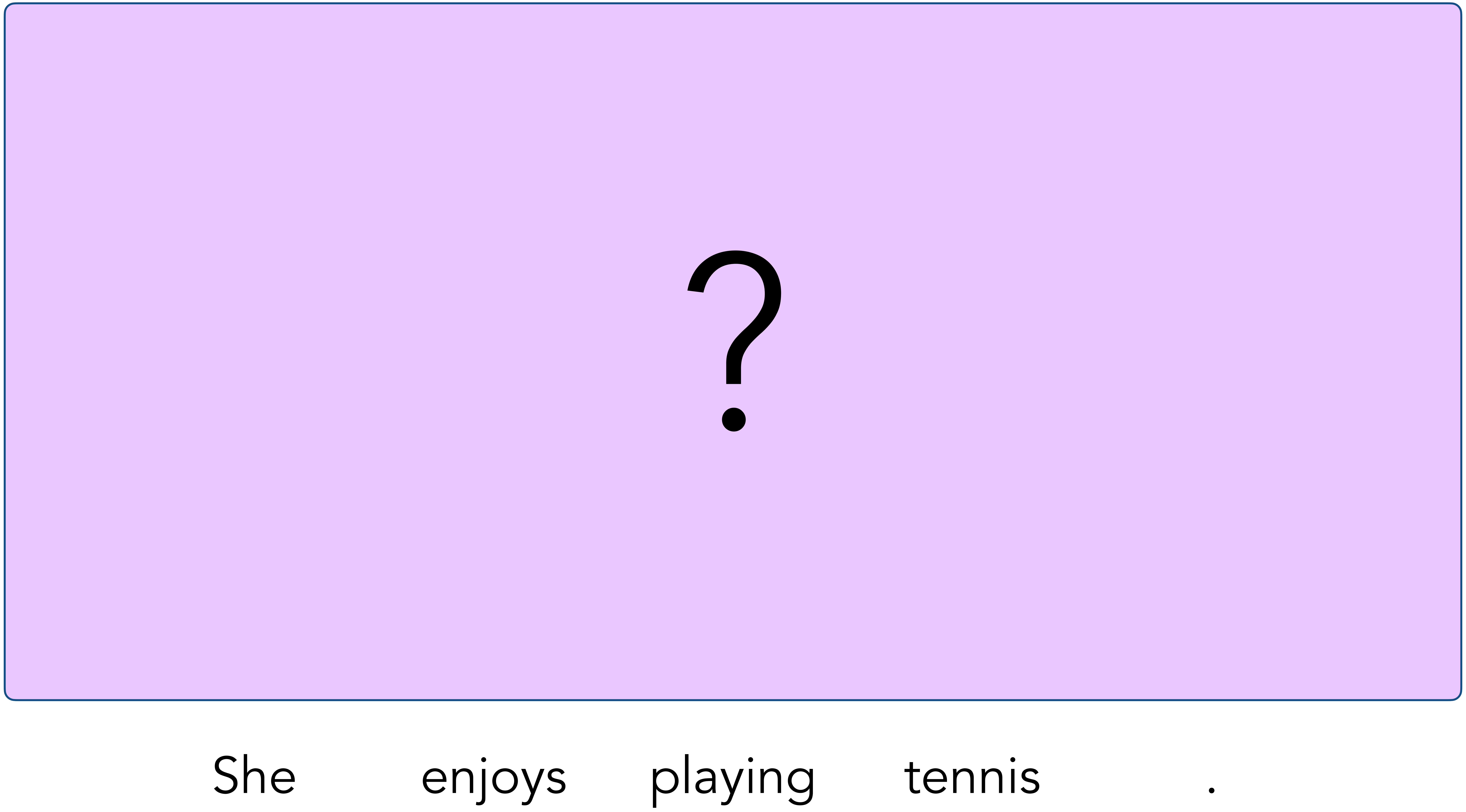


Encoder

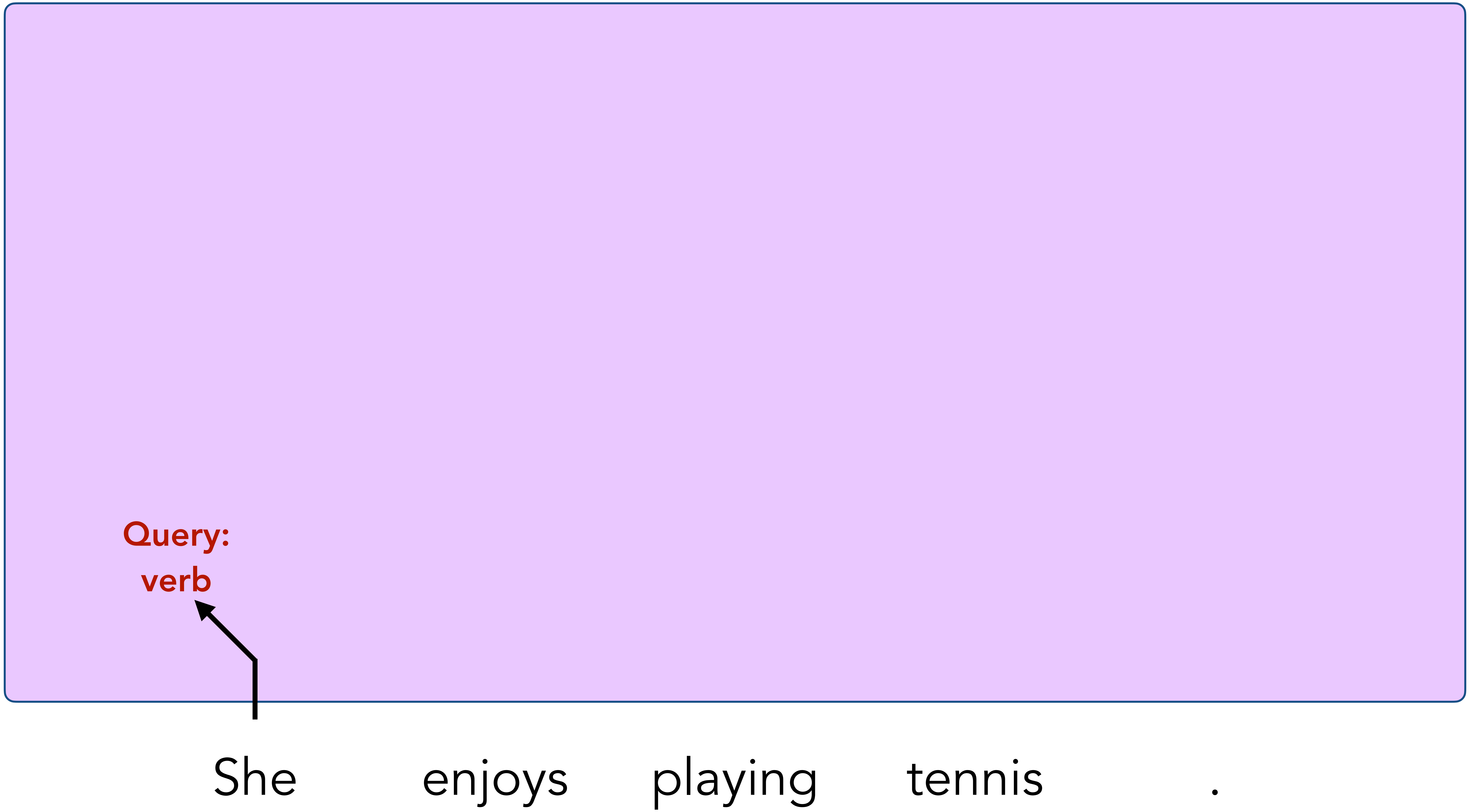


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Masked Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

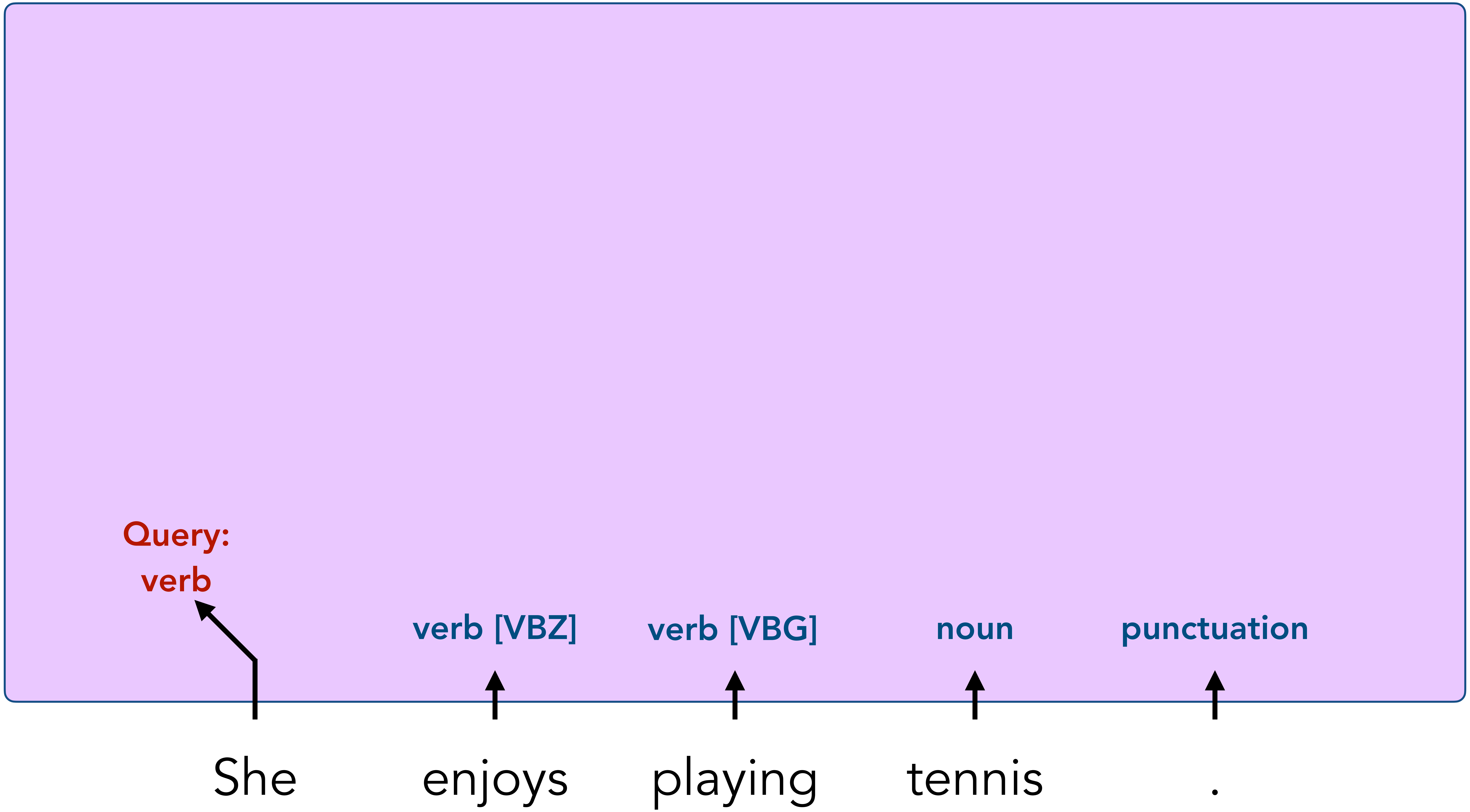
Encoder



Encoder

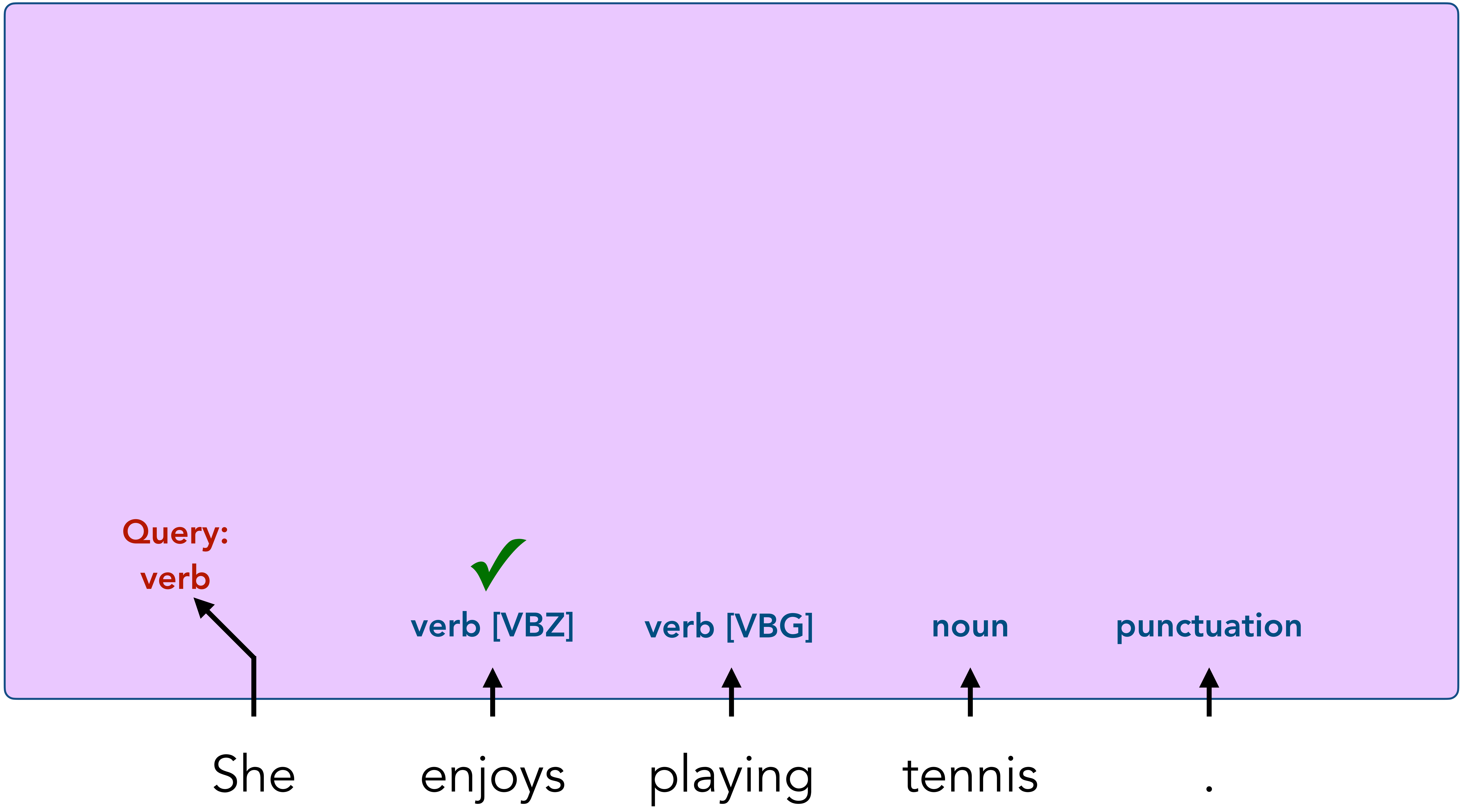


Encoder



Query:
verb
verb [VBZ]
verb [VBG]
noun
punctuation
She
enjoys
playing
tennis
.

Encoder



Query:
verb
verb [VBZ]
verb [VBG]
noun
punctuation
She
enjoys
playing
tennis
.

Encoder


Query:
verb
✓
✗
verb [VBZ]
verb [VBG]
noun
punctuation
She
enjoys
playing
tennis
.

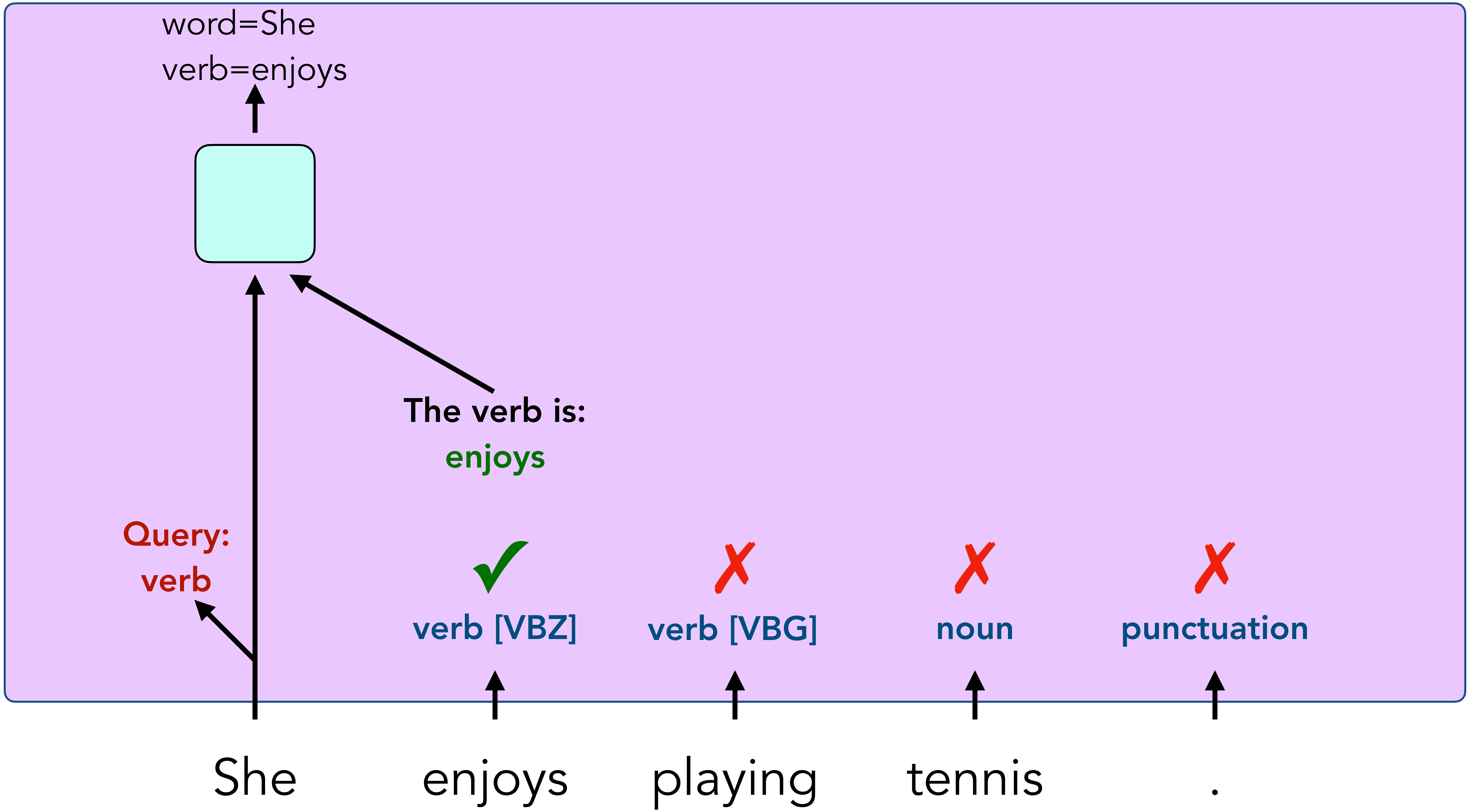
Encoder

Query:
verb
verb [VBZ]
verb [VBG]
noun
punctuation
She
enjoys
playing
tennis
.

Encoder

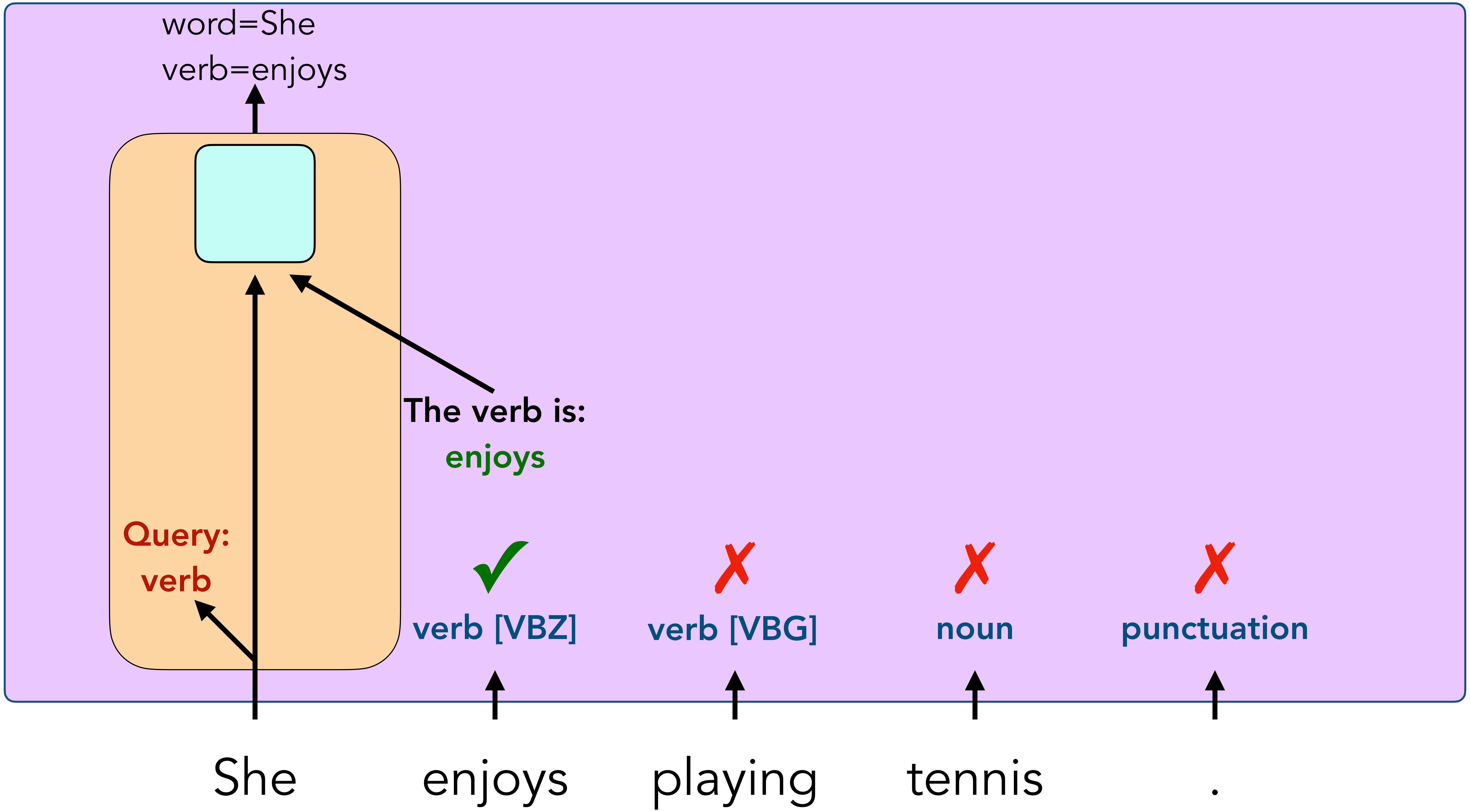
The verb is:
enjoys
Query:
verb
✓
✗
✗
✗
verb [VBZ]
verb [VBG]
noun
punctuation
She
enjoys
playing
tennis
.

Encoder



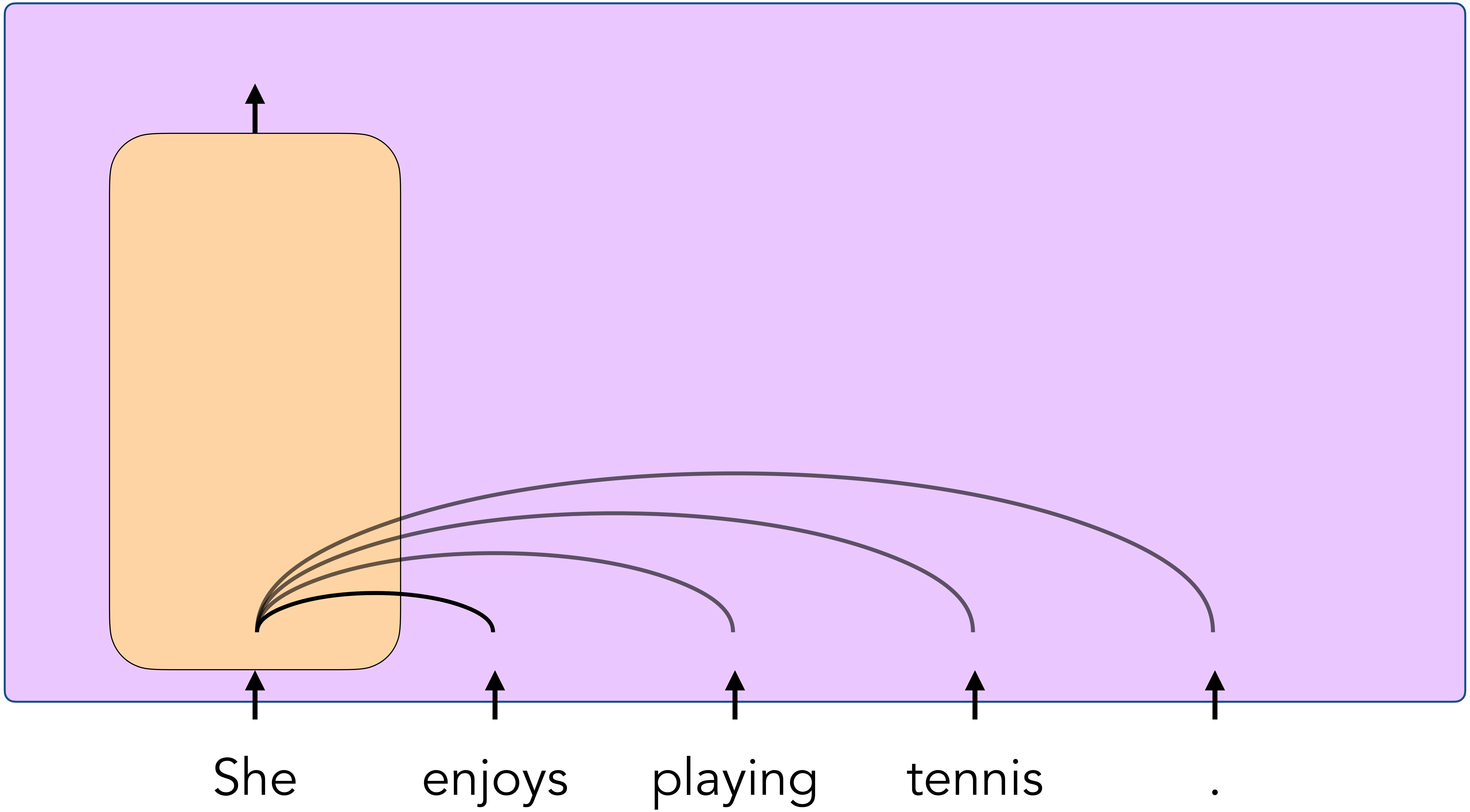
word=She
verb=enjoys
The verb is:
enjoys
Query:
verb
✓
✗
✗
✗
verb [VBZ]
verb [VBG]
noun
punctuation
She
enjoys
playing
tennis
.

Encoder

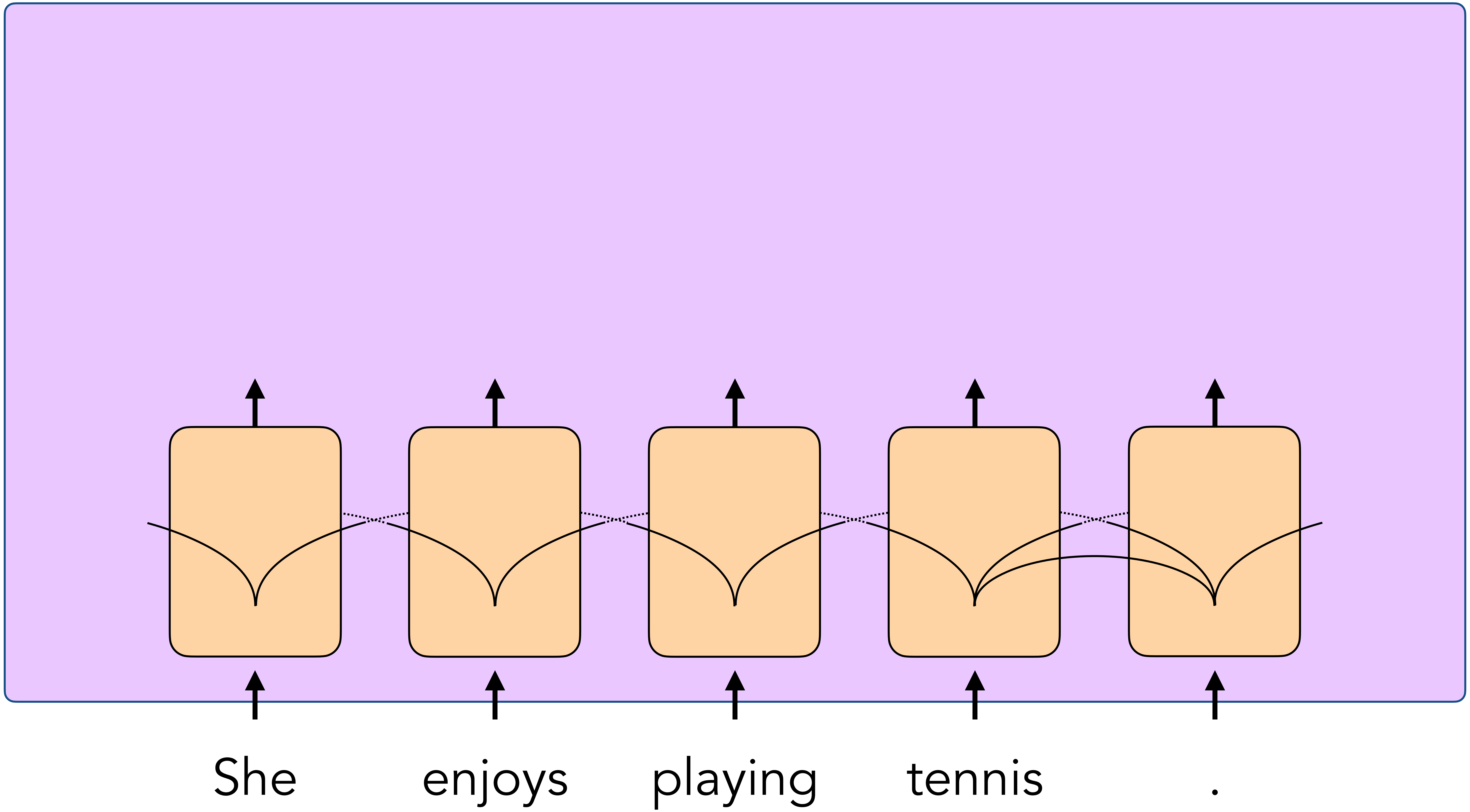


word=She
verb=enjoys
The verb is:
enjoys
Query:
verb
✓
✗
✗
✗
verb [VBZ]
verb [VBG]
noun
punctuation
She
enjoys
playing
tennis
.

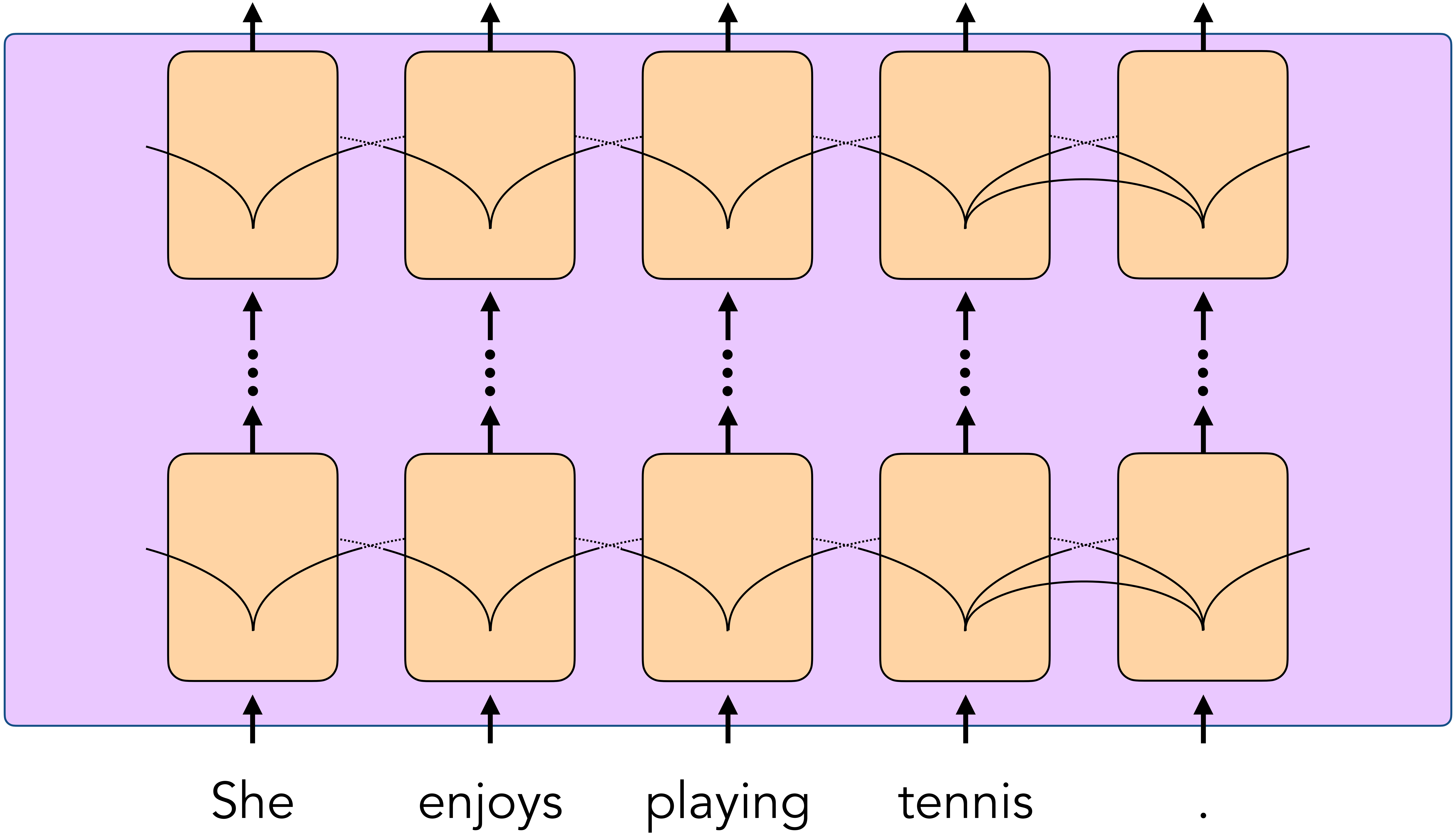
Encoder



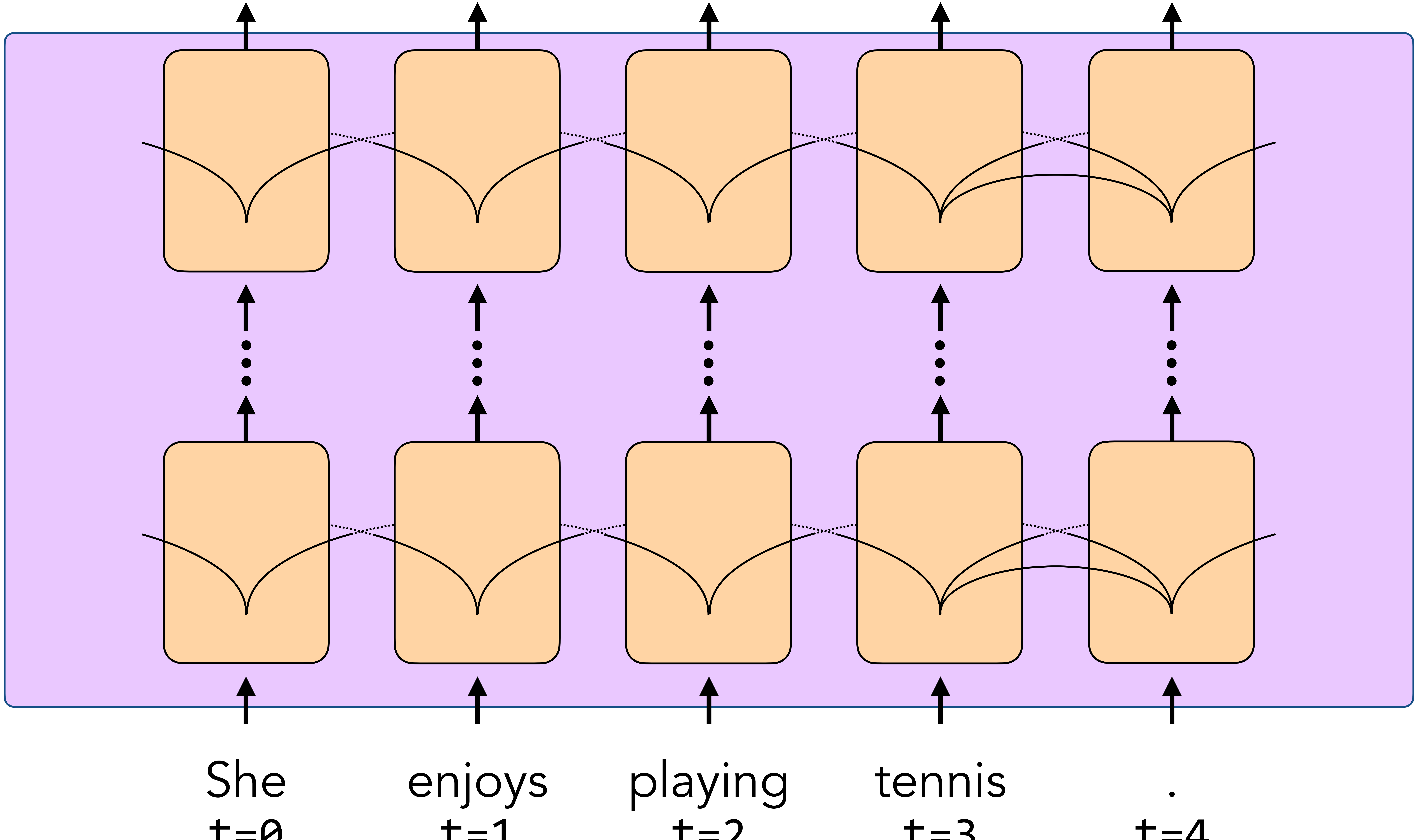
Encoder



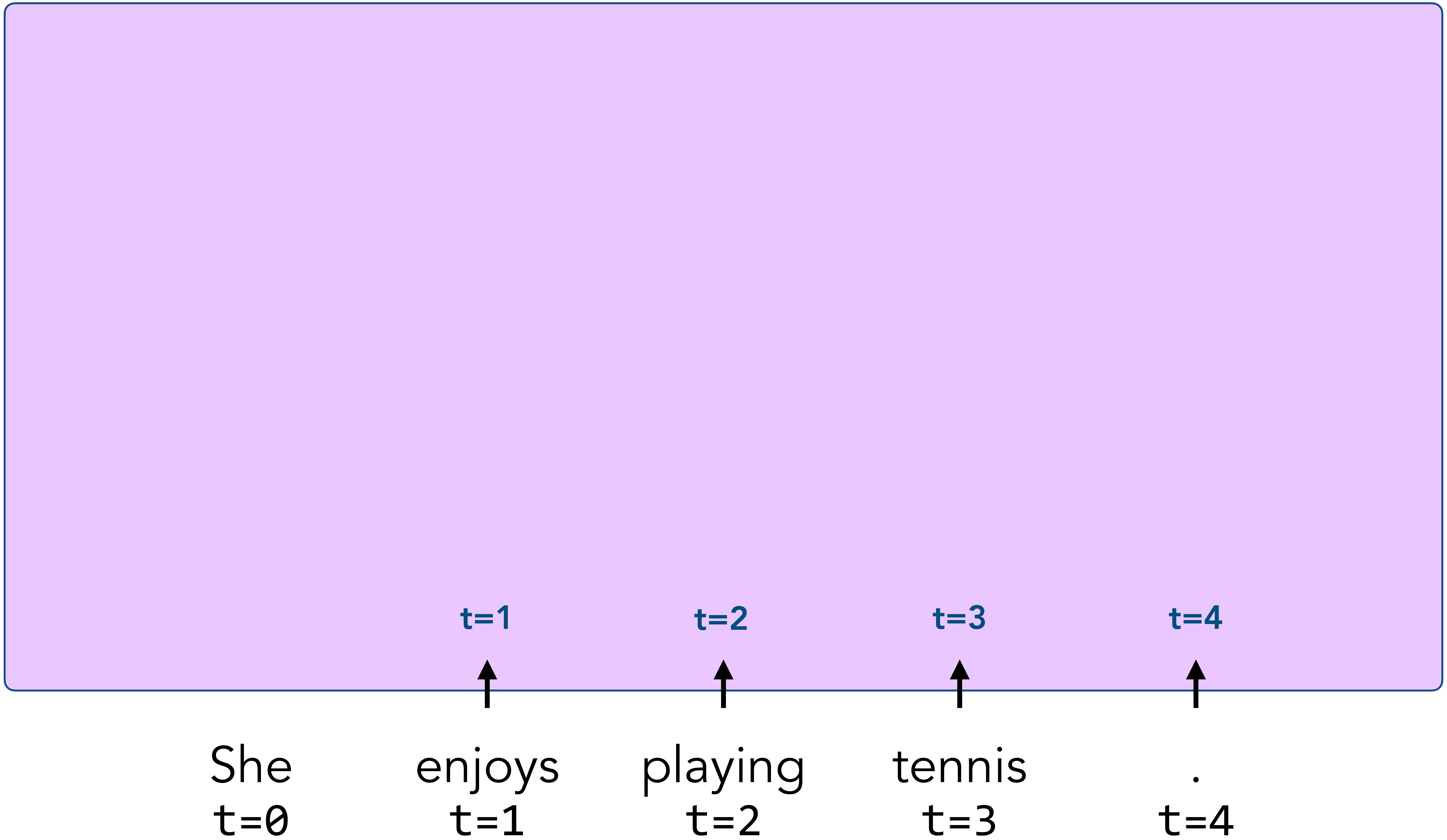
Encoder



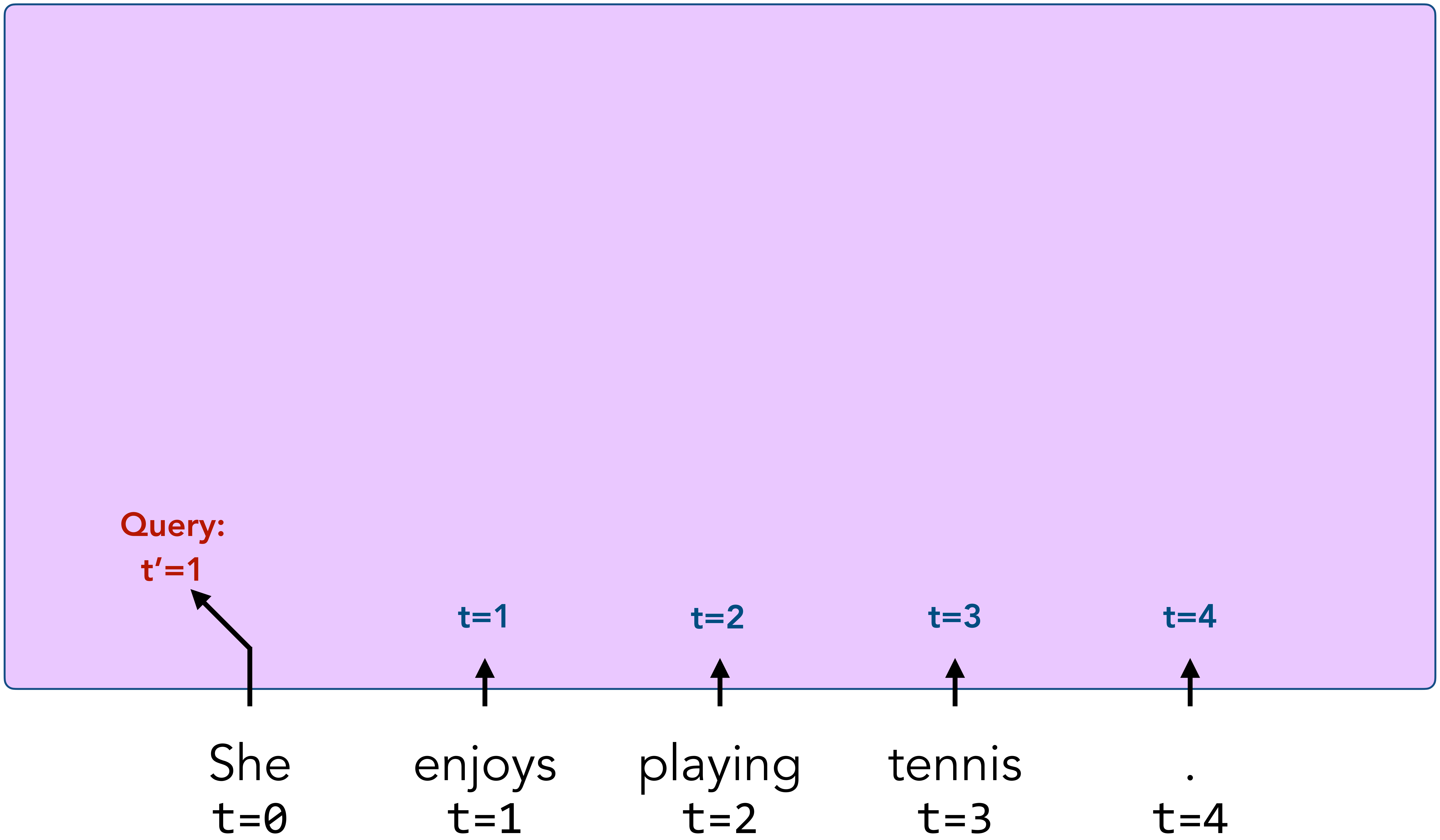
Encoder



Position-Based Attention

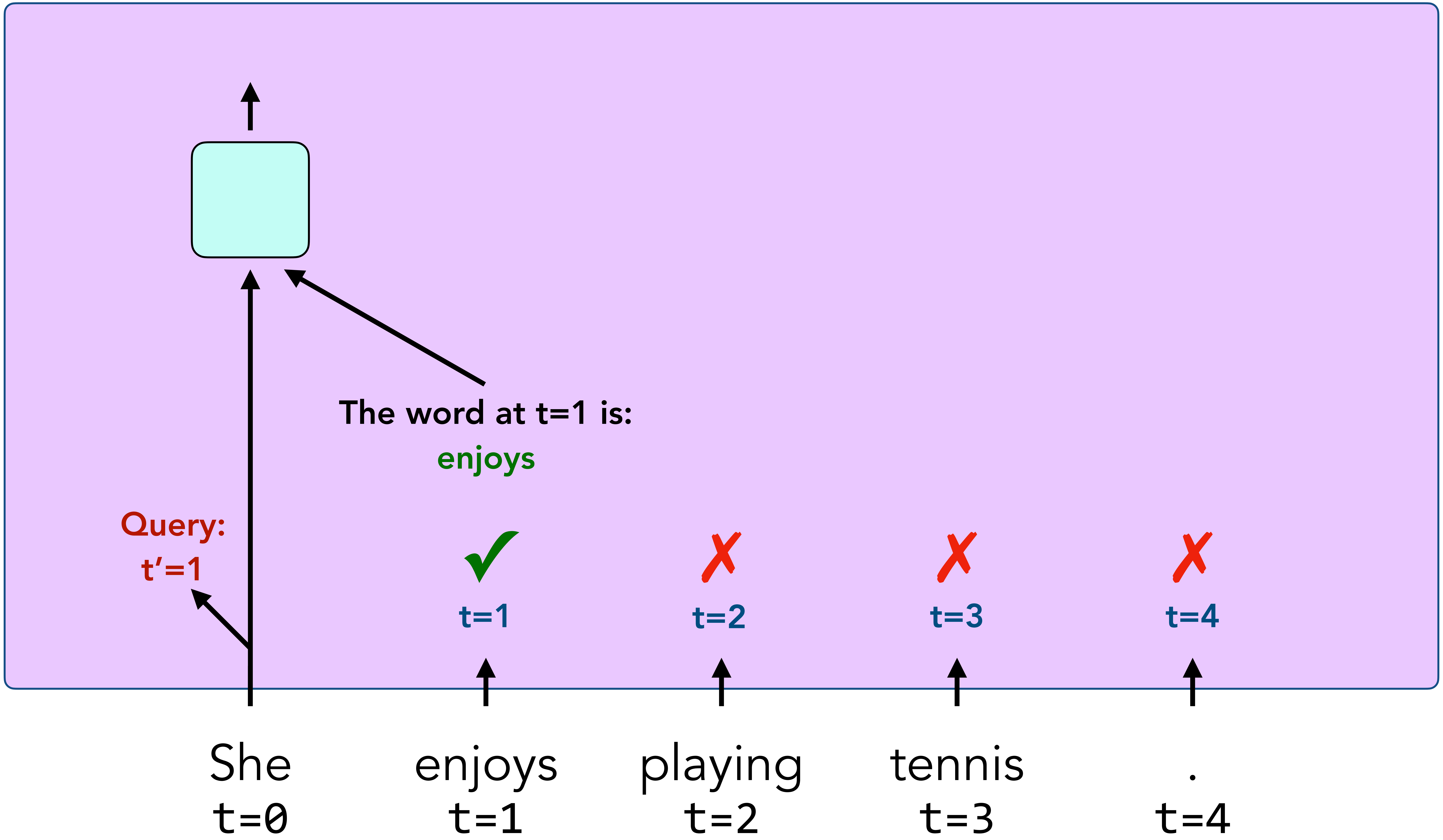


Position-Based Attention

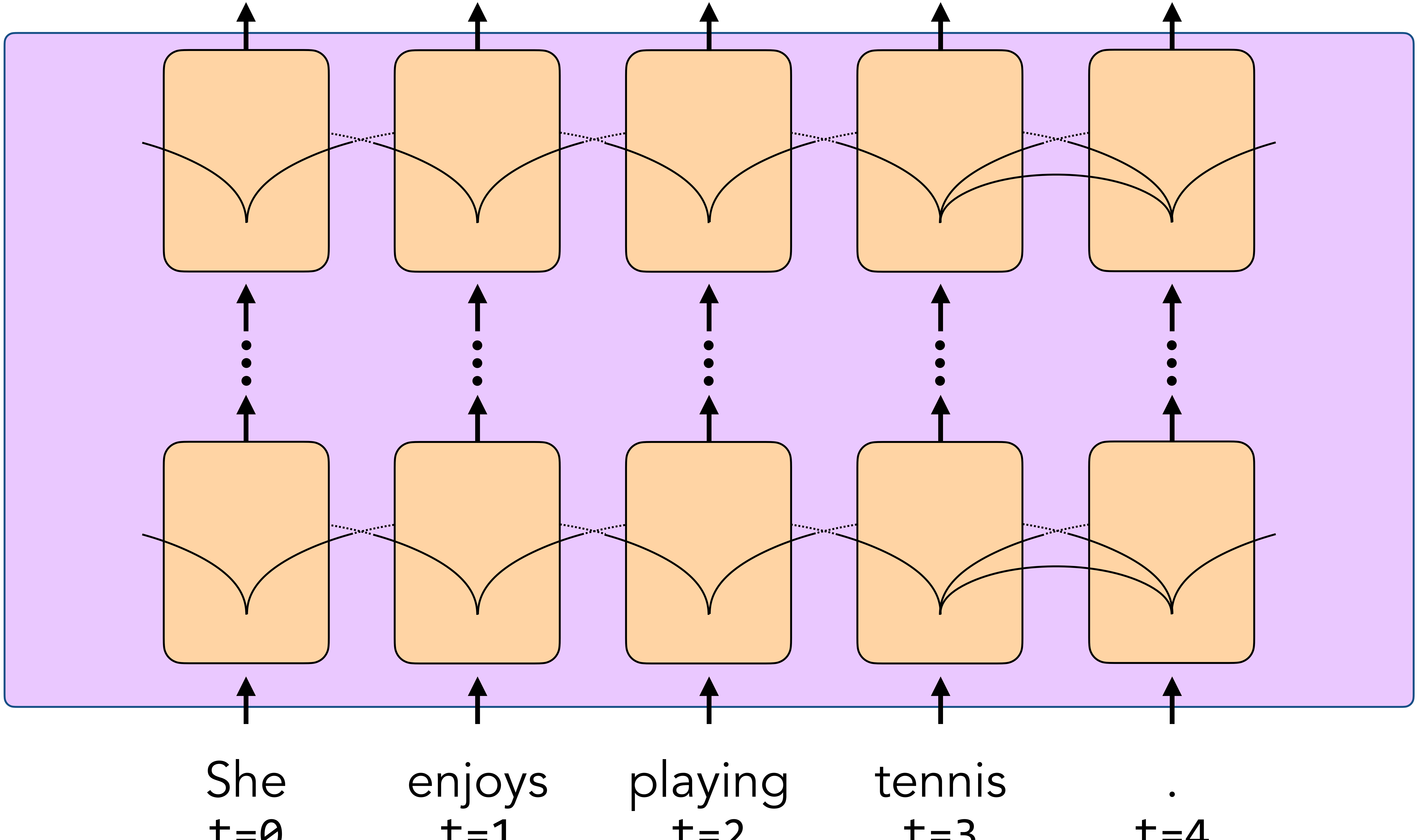


Position-Based Attention

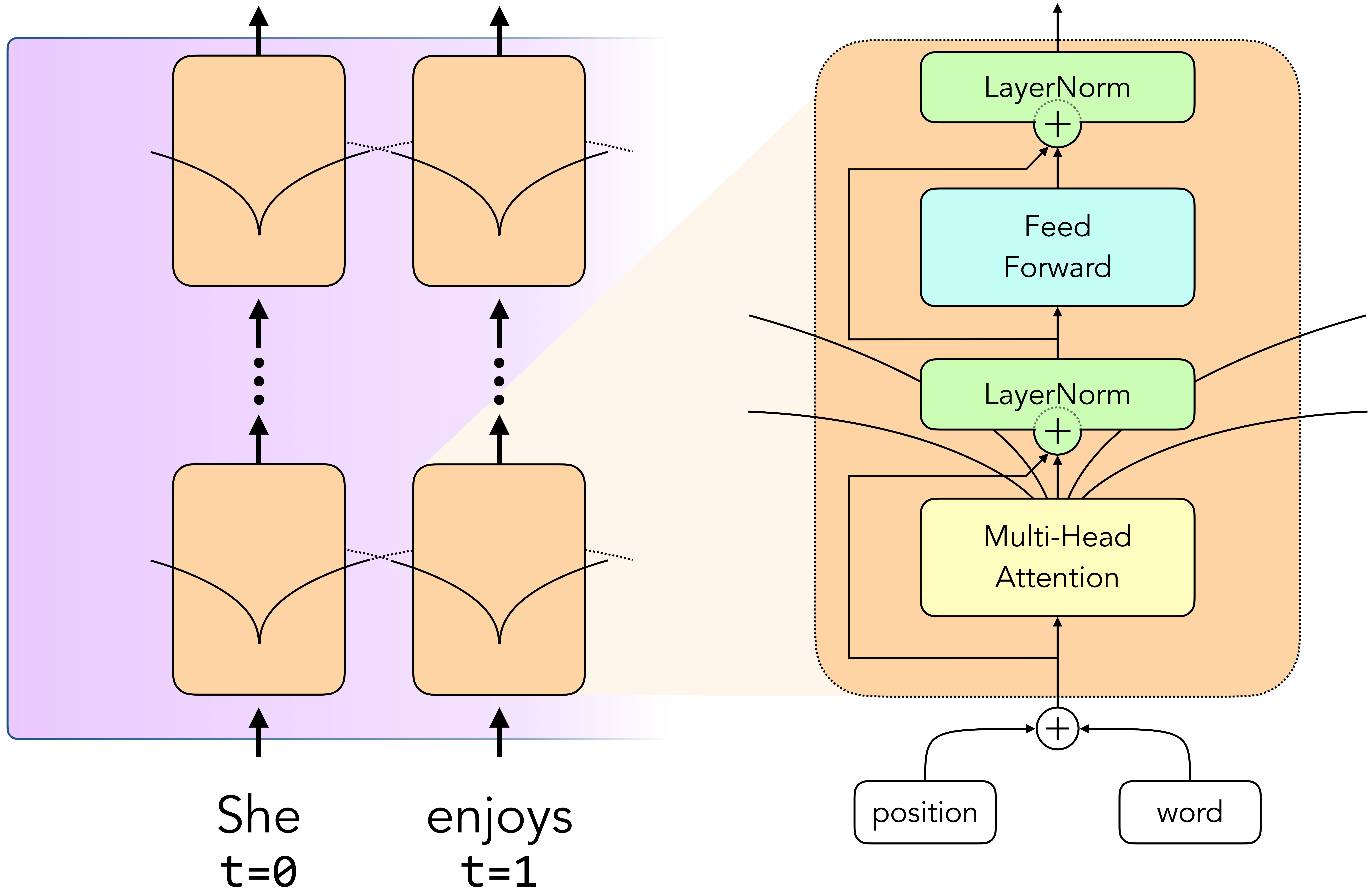
Position-Based Attention



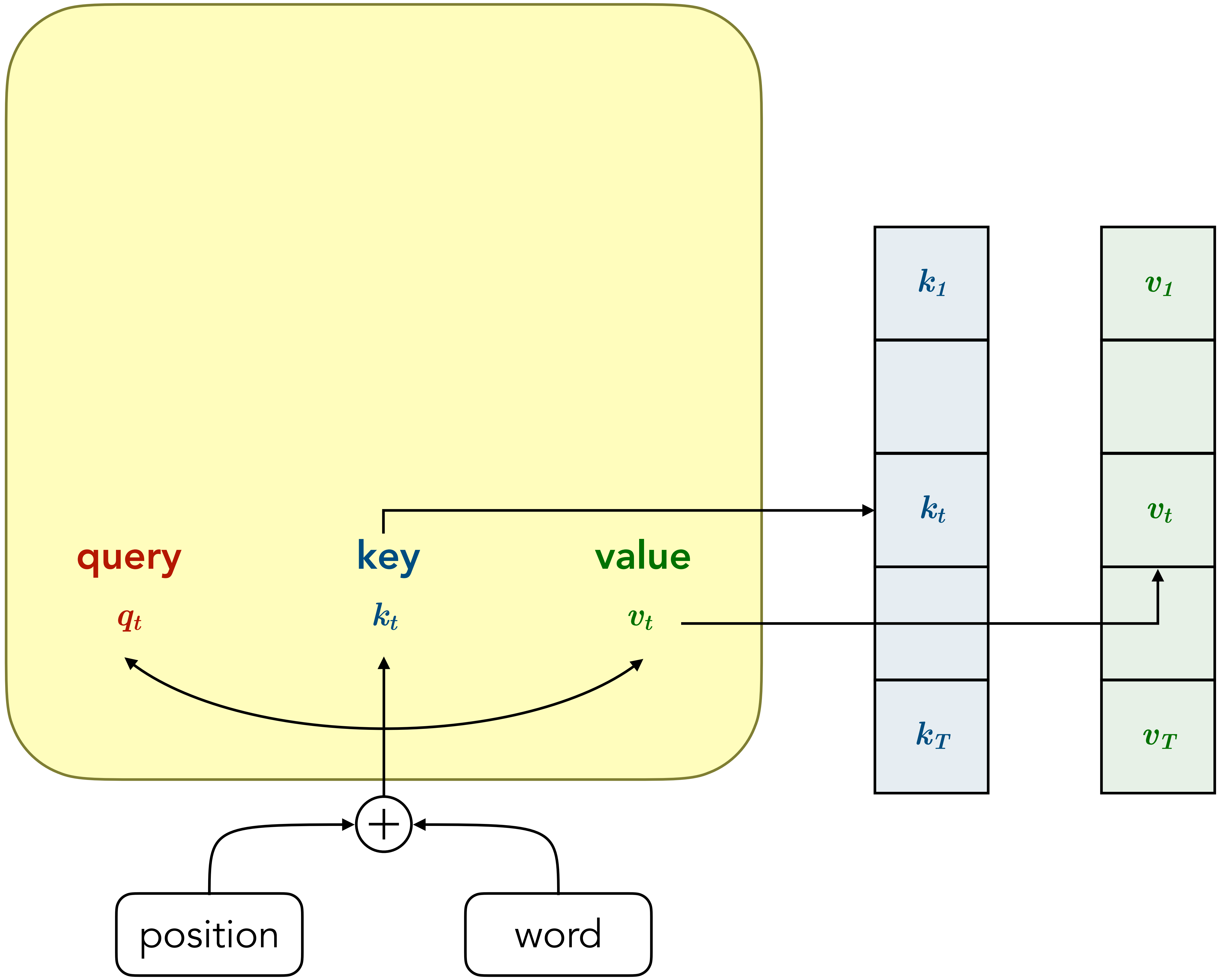
Encoder



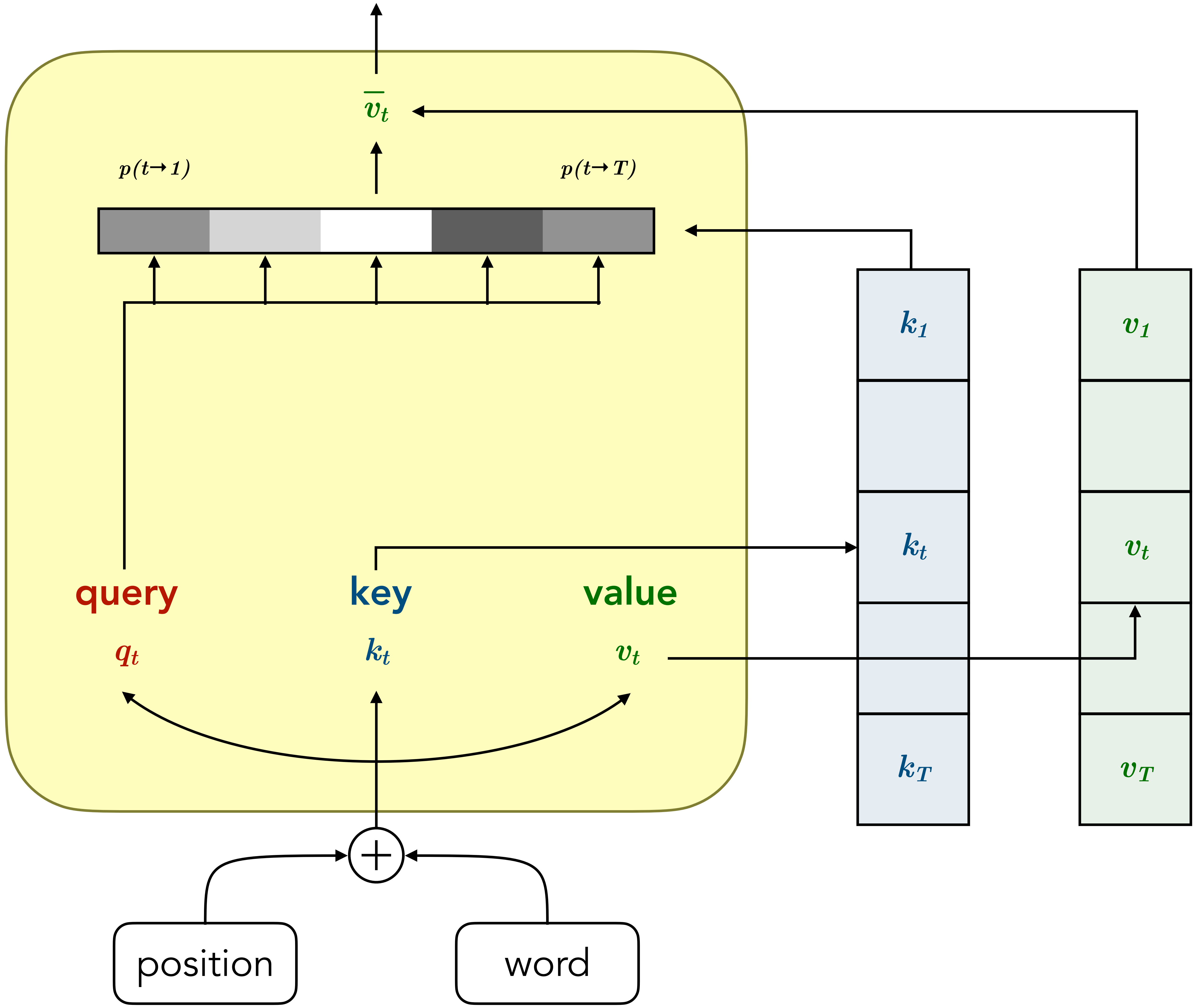
Encoder


LayerNorm
Feed
Forward
LayerNorm
Multi-Head
Attention
position
word
She
t=0
enjoys
t=1

Self-Attention

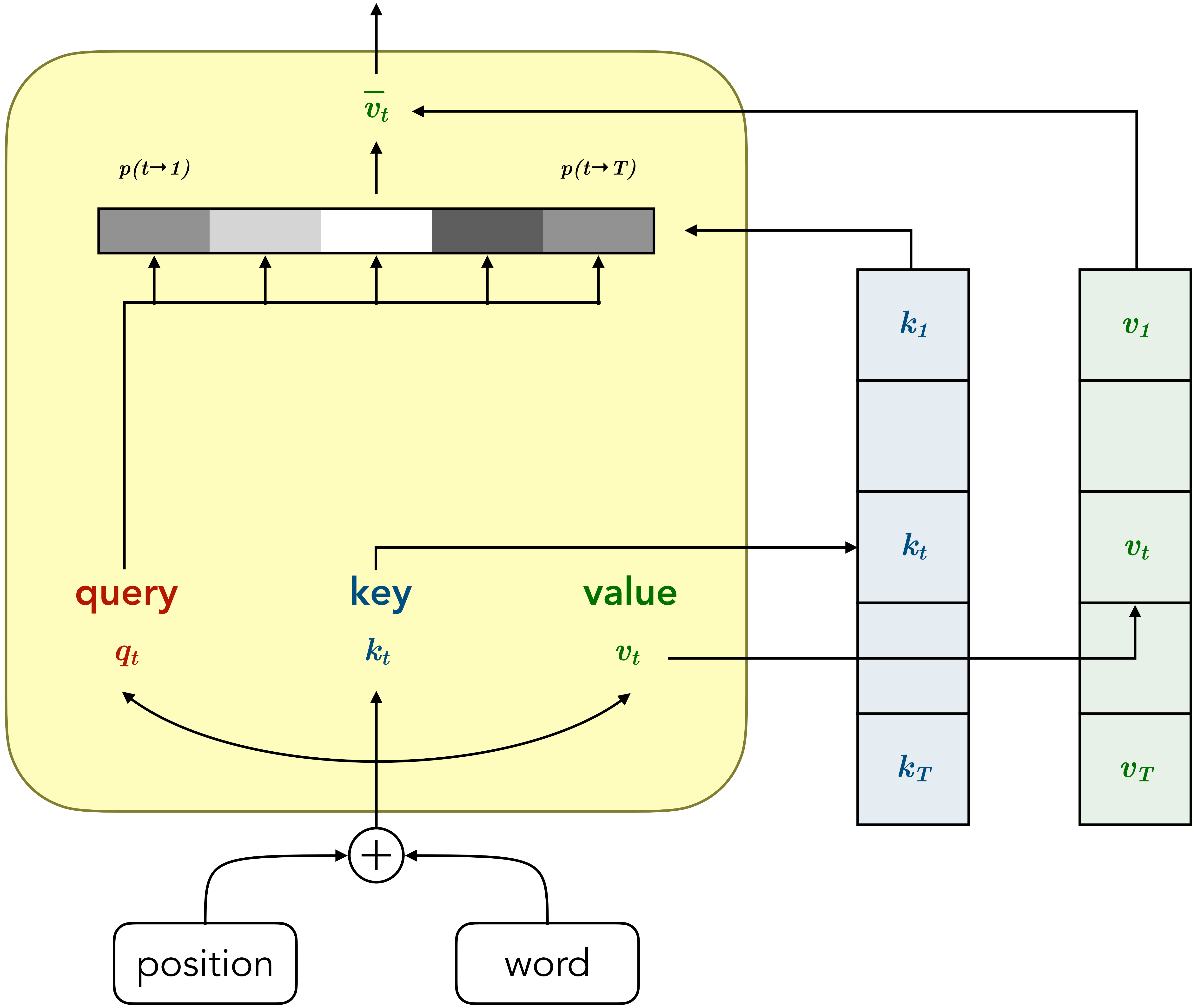


Self-Attention



Self-Attention

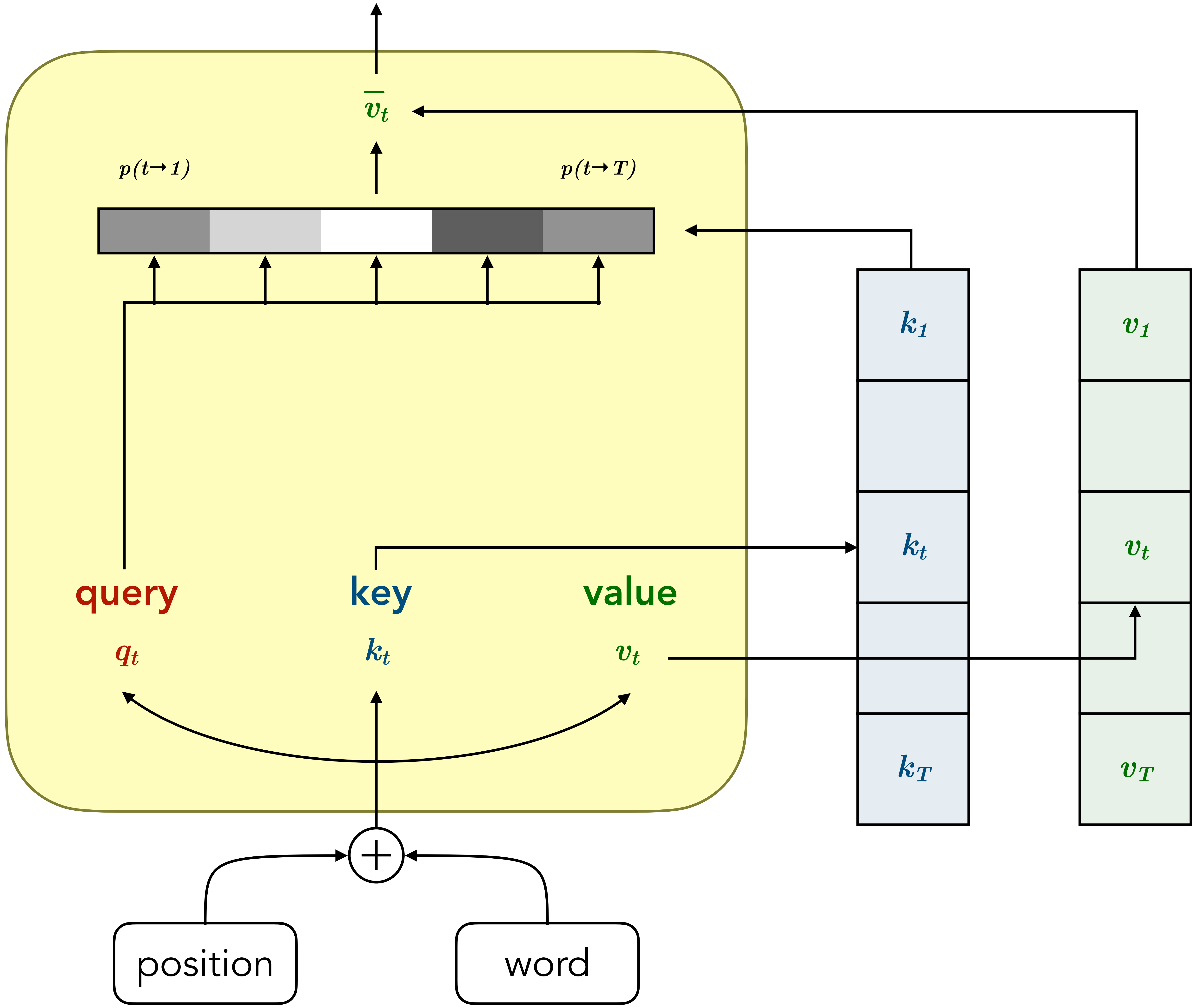
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d_k}}V)$$



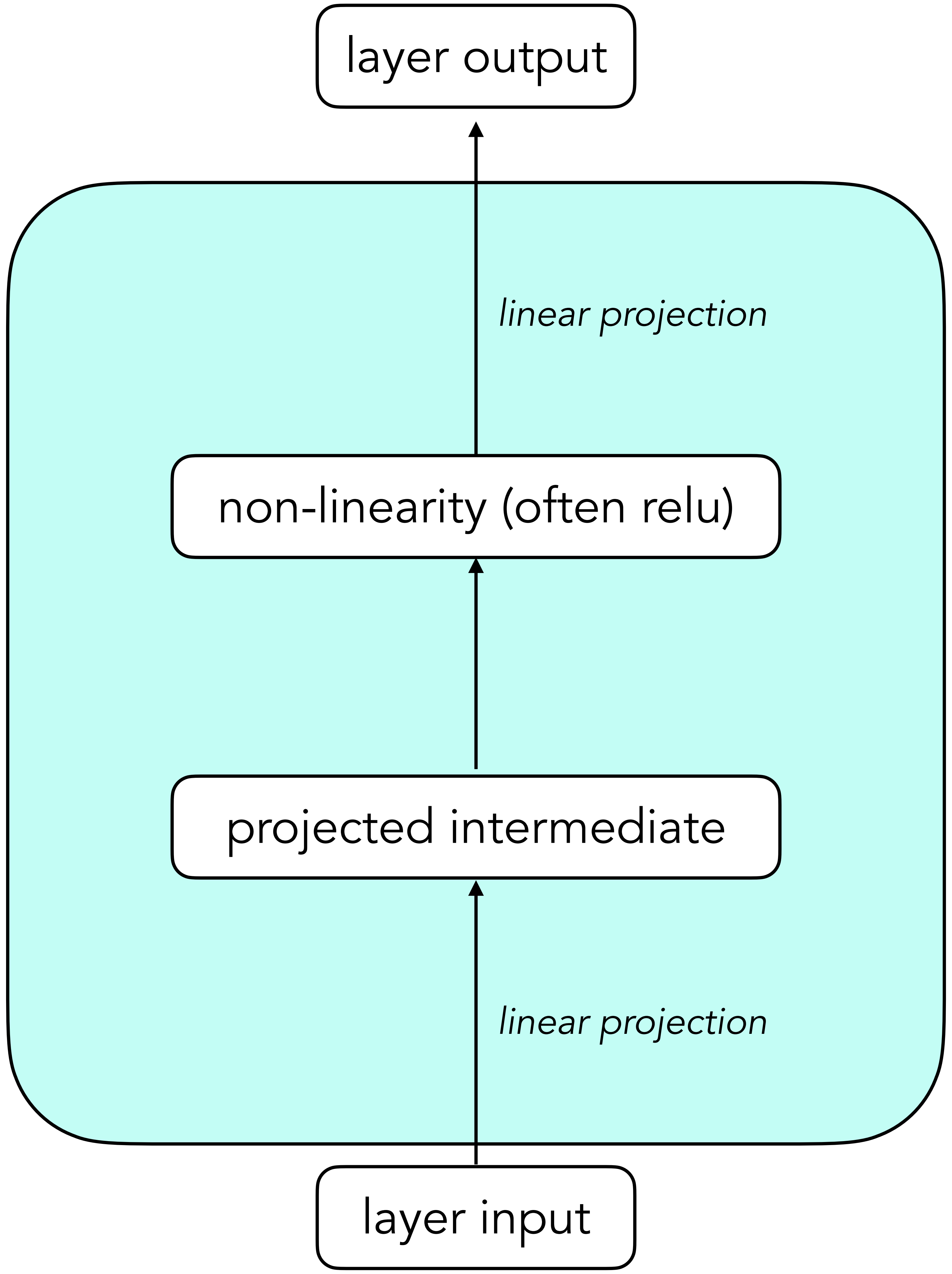
Self-Attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d_k}}V)$$

$$\text{MultiHead}(X) = \sum_{i=0}^{h} \text{Attention}(XW_i^Q, XW_i^K, XW_i^V)W_i^O$$

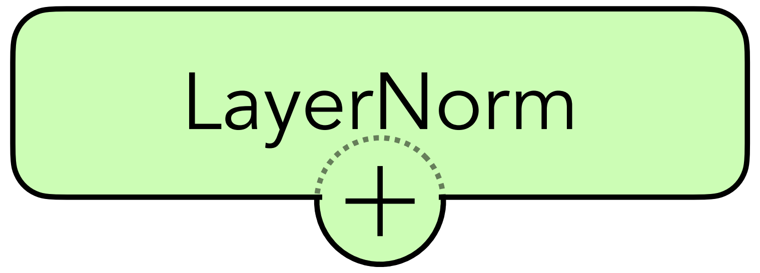


Feed-Forward



$$\text{FeedForward}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Add & Norm



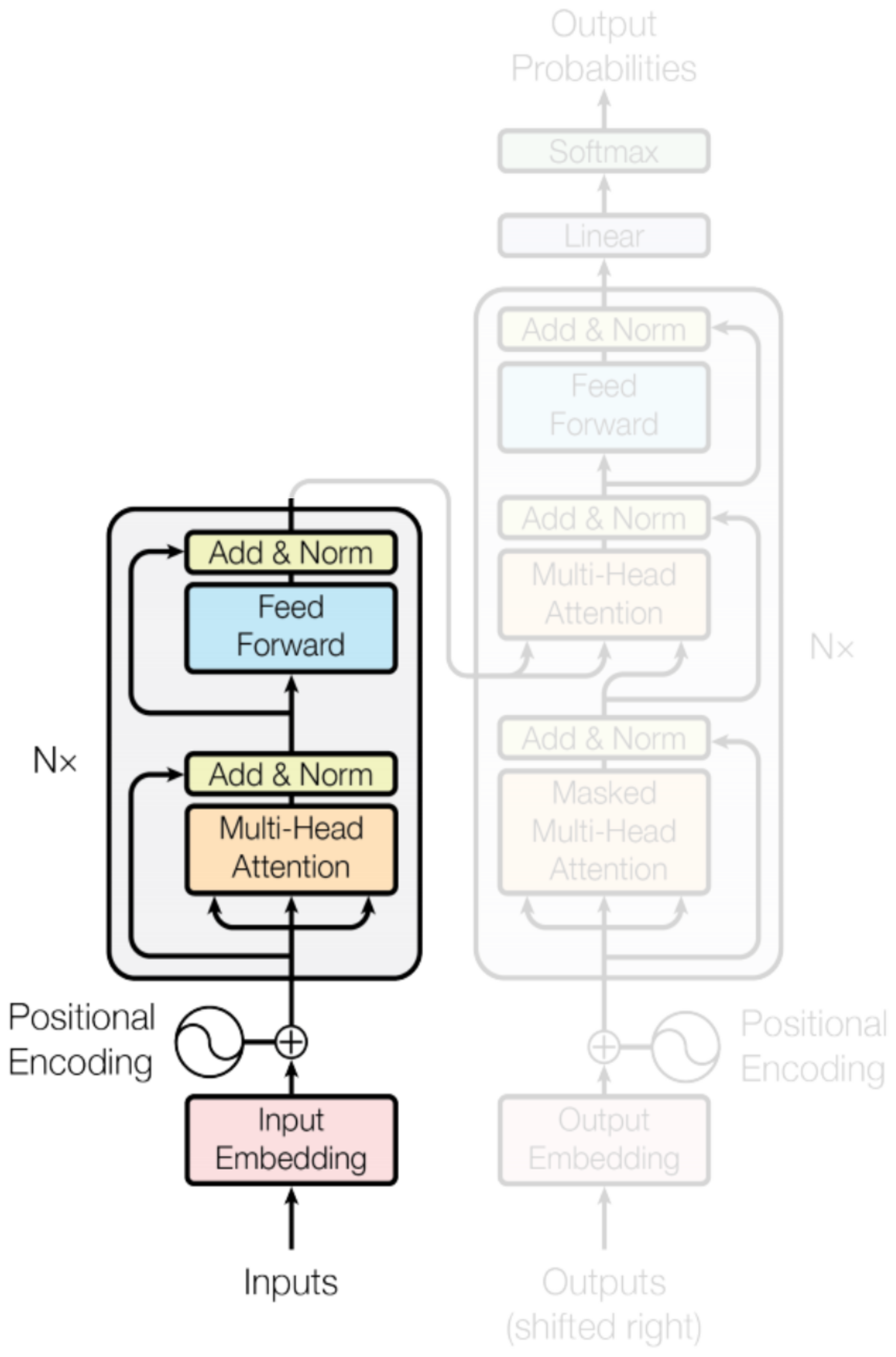
Layer Normalization [Ba+16]

improves stability of neuron activations

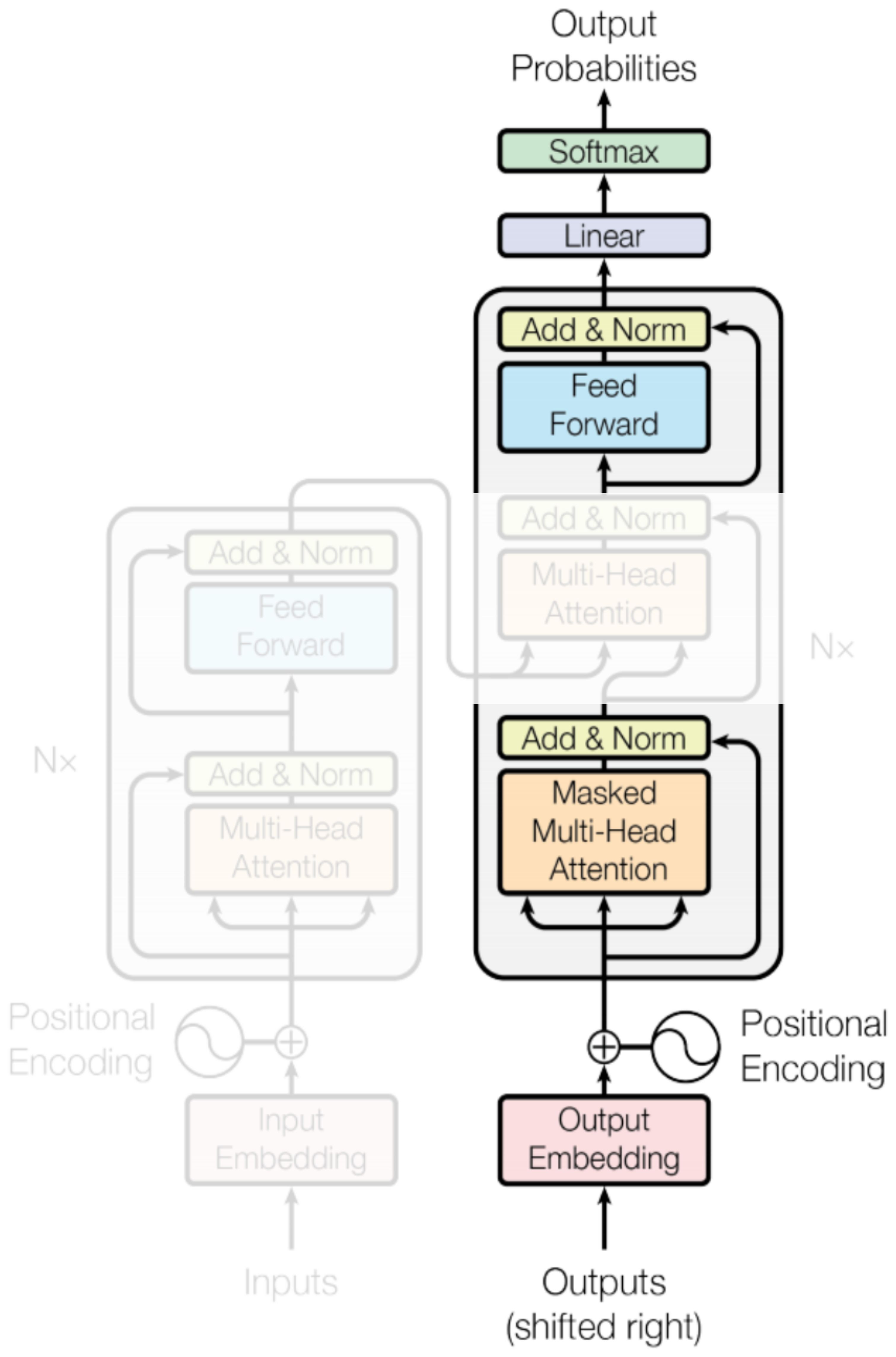
Residual Connections

useful across a variety of neural network architecture types, not just in NLP

Encoder

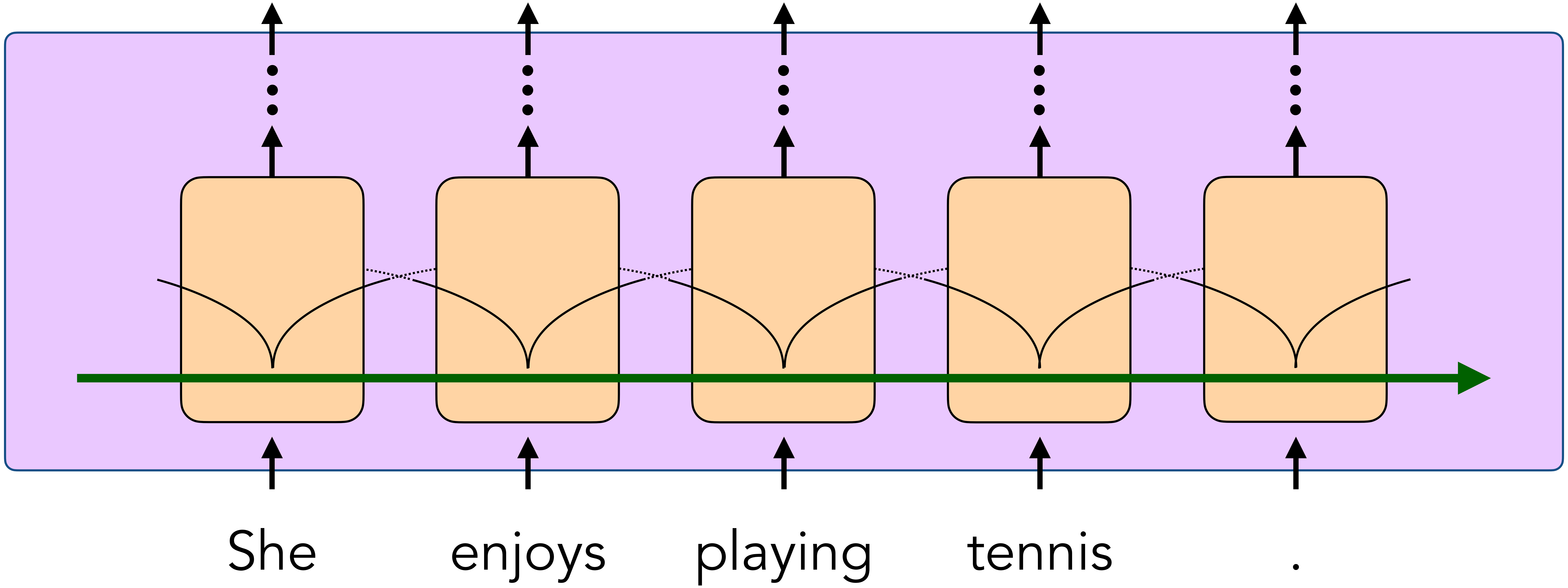

Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Add & Norm
Masked
Multi-Head
Attention
Positional
Encoding
Output
Embedding
Outputs
(shifted right)
Add & Norm
Feed
Forward
Nx
Add & Norm
Multi-Head
Attention
Positional
Encoding
Input
Embedding
Inputs

Decoder

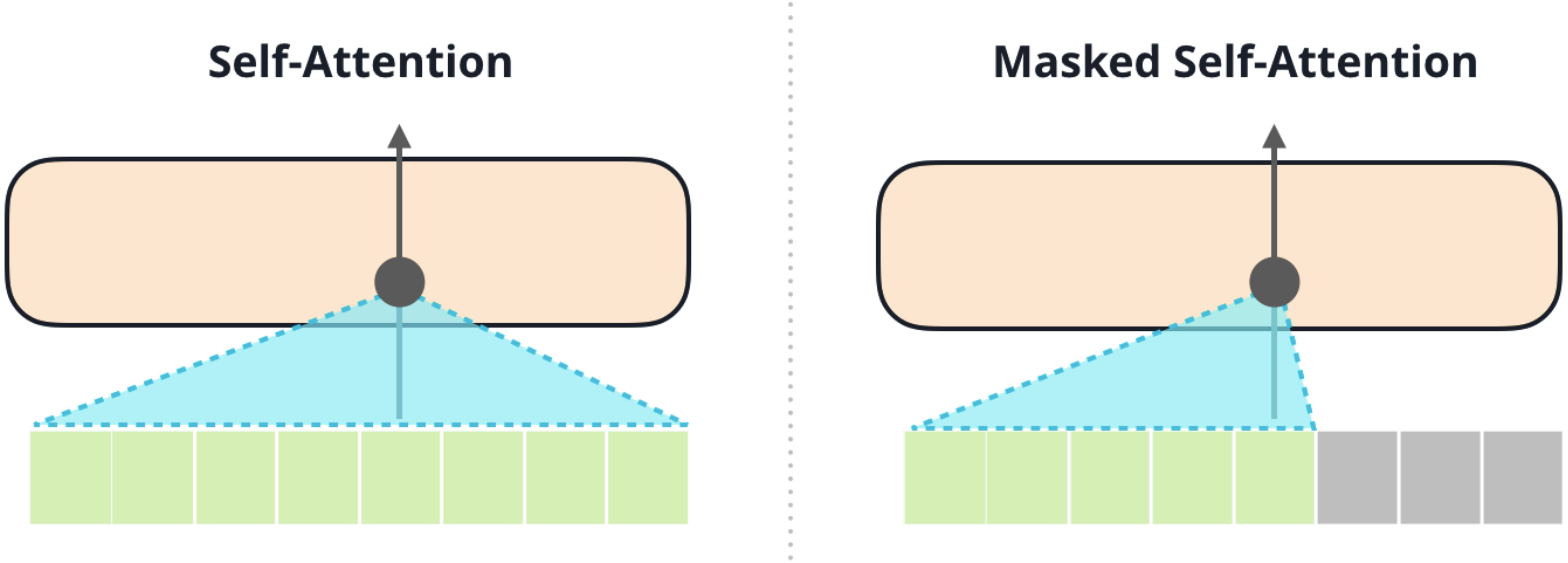


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Masked Multi-Head Attention
Positional Encoding
Output Embedding
Outputs (shifted right)
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Inputs

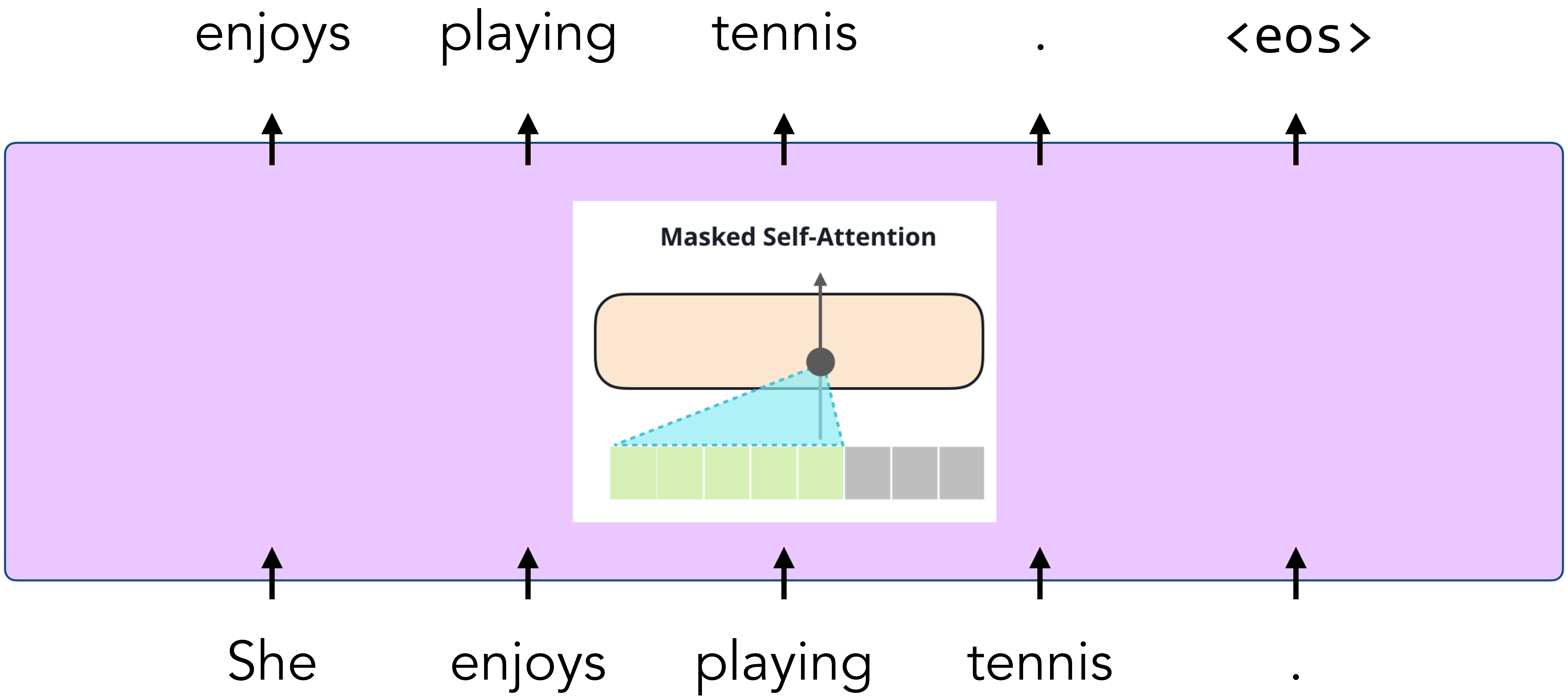
Decoder



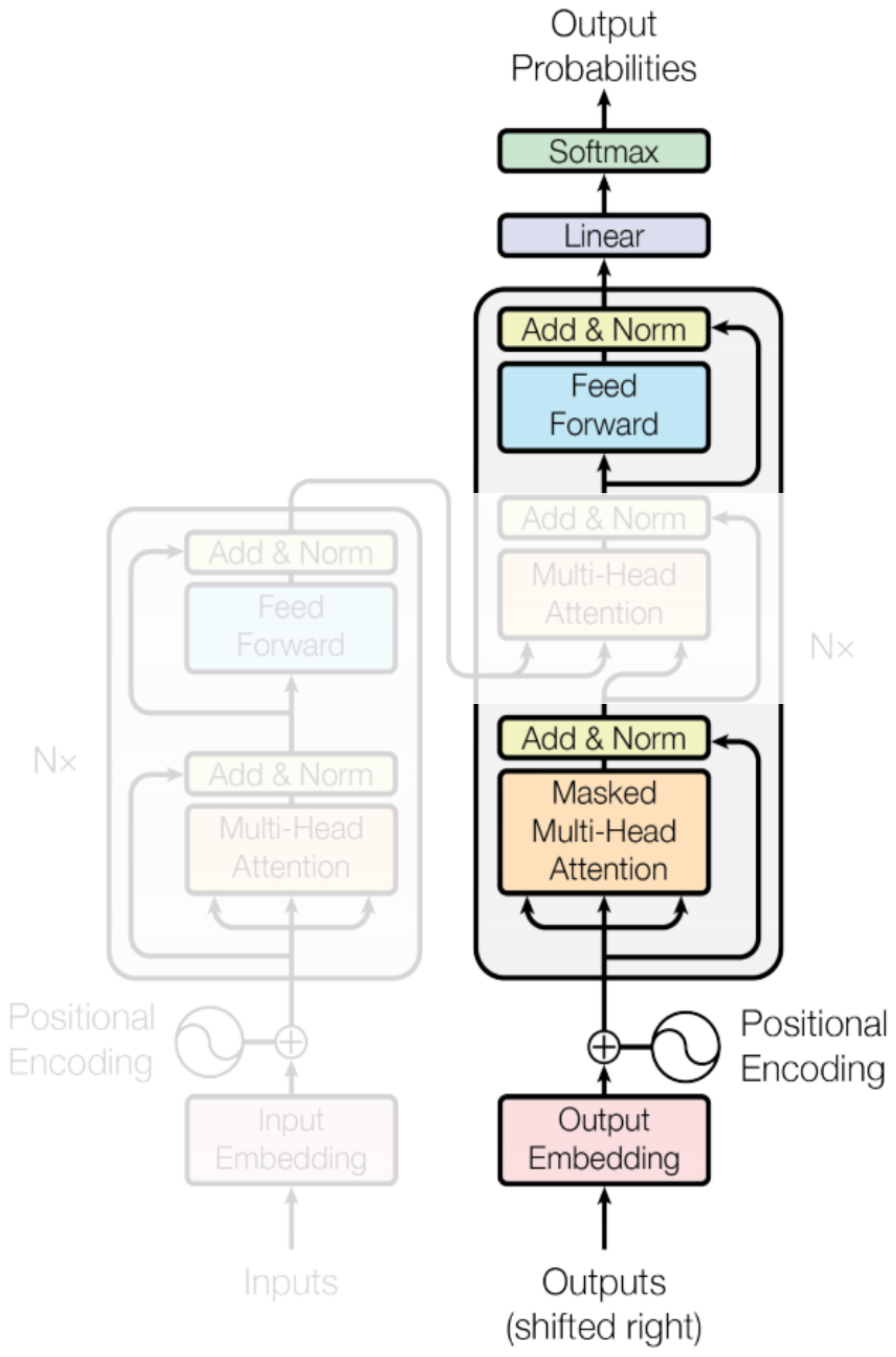
Encoder vs. Decoder



Decoder-Only Transformer Model

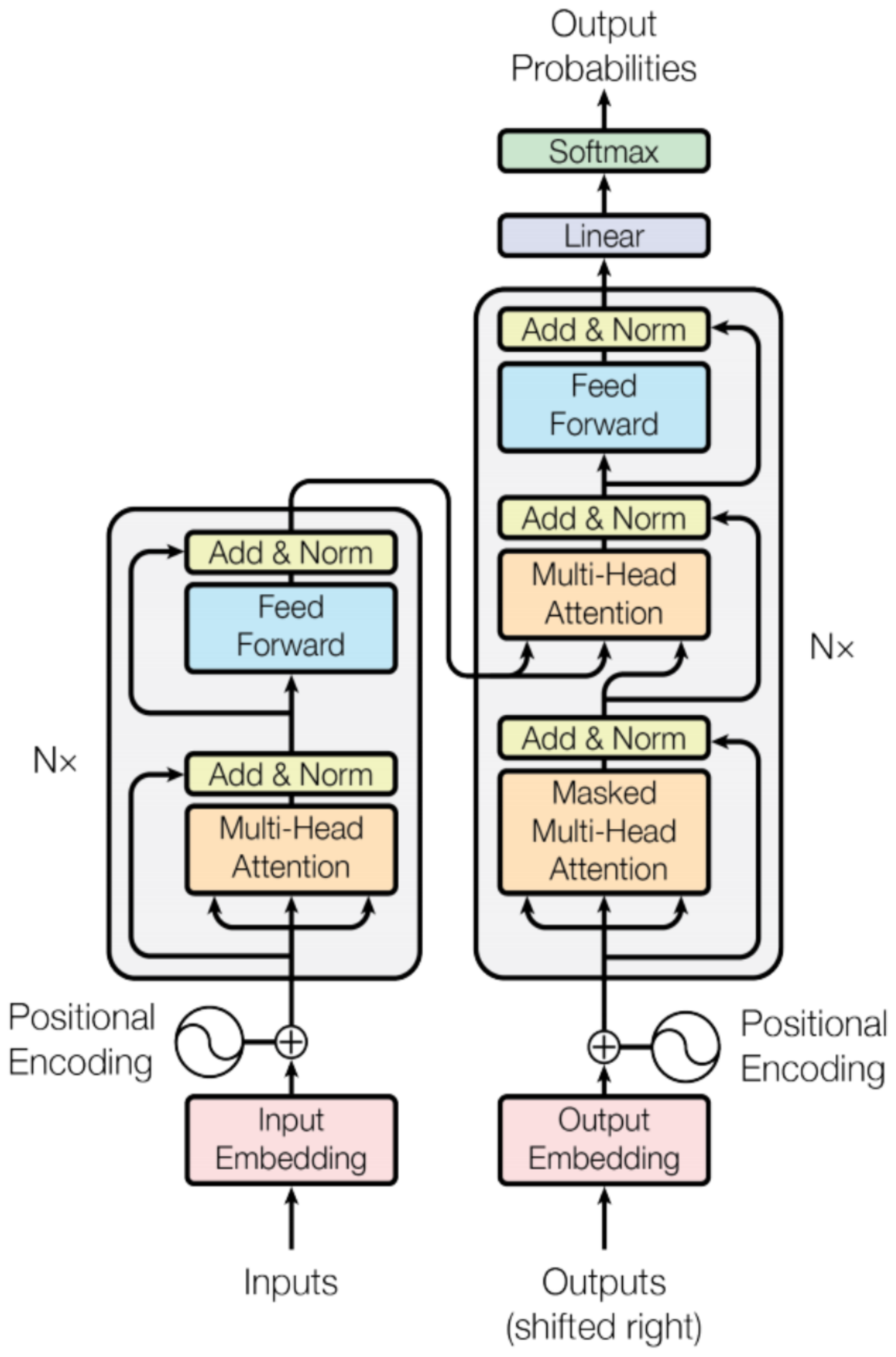


Decoder



Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Add & Norm
Masked
Multi-Head
Attention
Positional
Encoding
Output
Embedding
Outputs
(shifted right)
Add & Norm
Feed
Forward
Nx
Add & Norm
Multi-Head
Attention
Positional
Encoding
Input
Embedding
Inputs

Encoder-Decoder

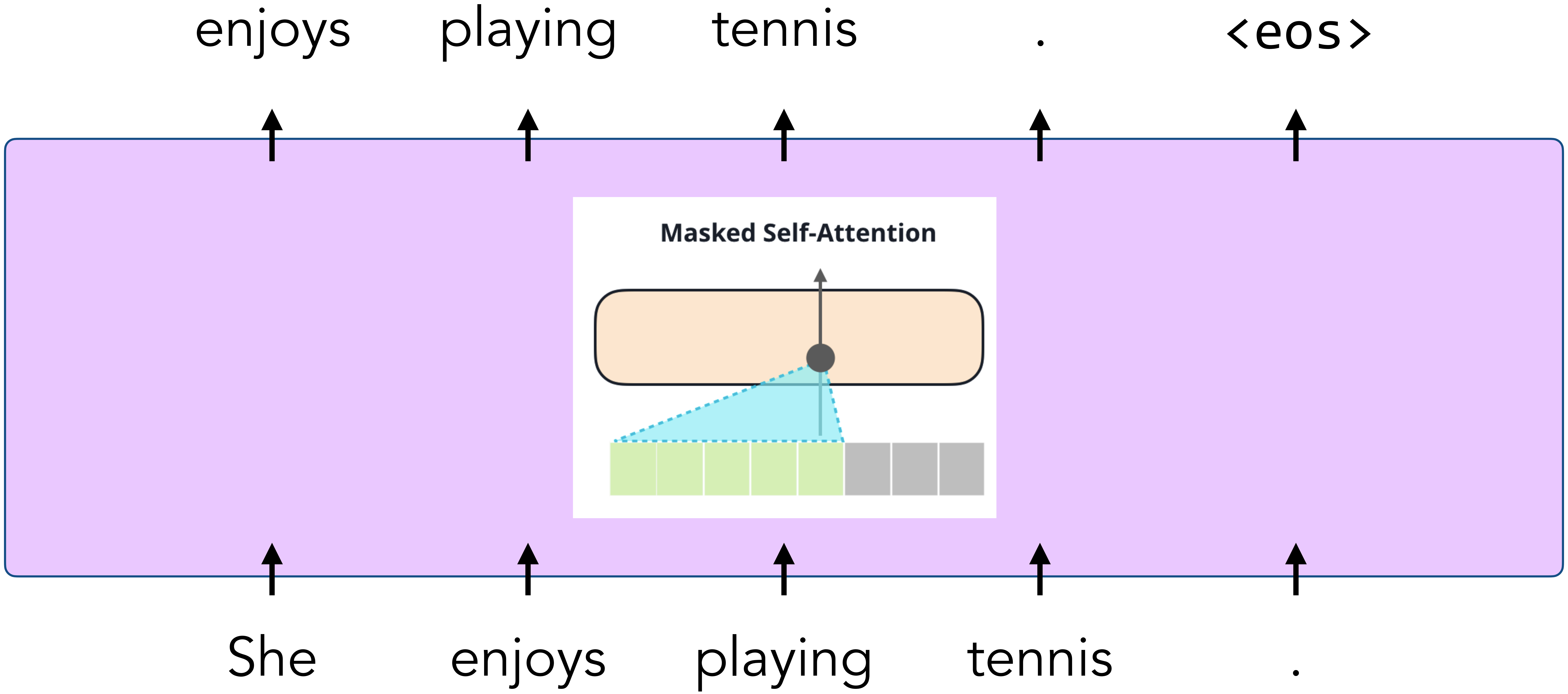


Transformer MT Results

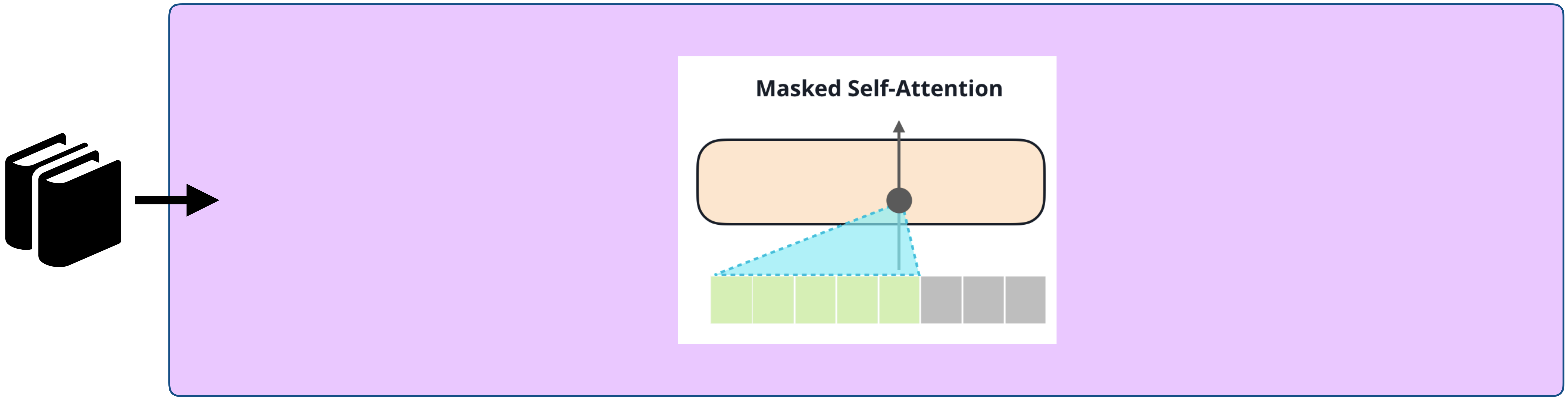
Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	~~**28.4**~~ **29.1**	~~**41.8**~~ **41.8**	$2.3 \cdot 10^{19}$	

(2) Pre-Training

Transformer Language Model



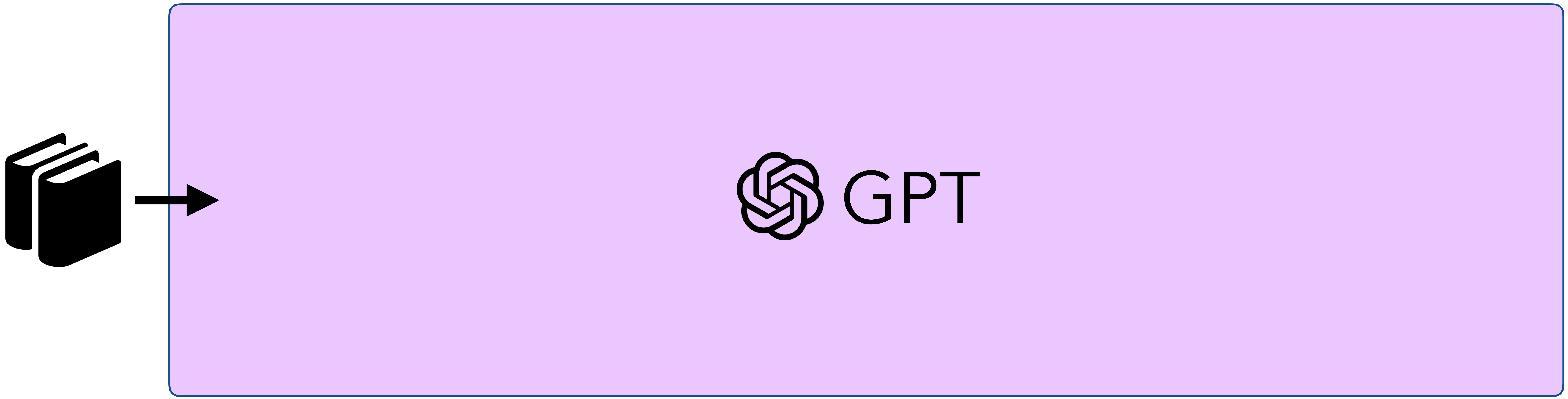
Pre-Training with LMs



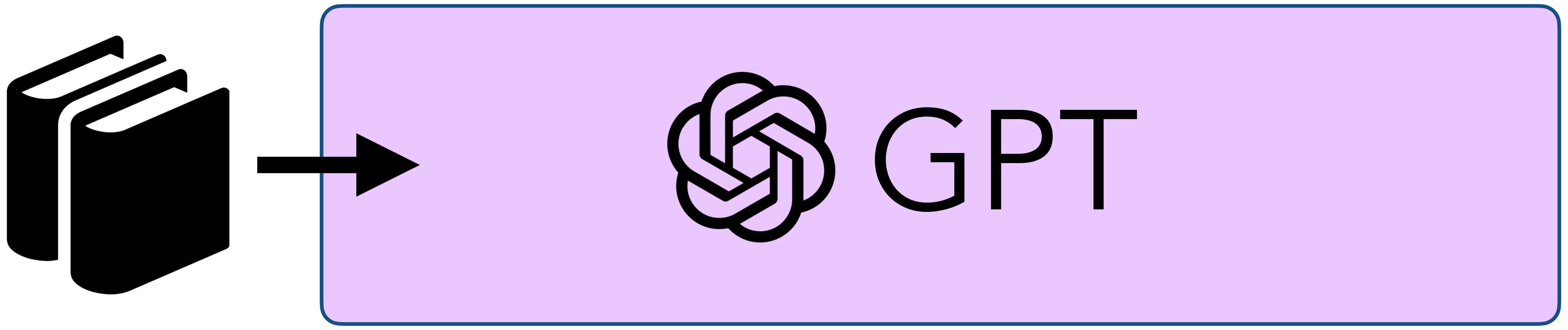
Pre-Training with LMs

Representative Model: GPT

(GPT = Generative Pre-Training)



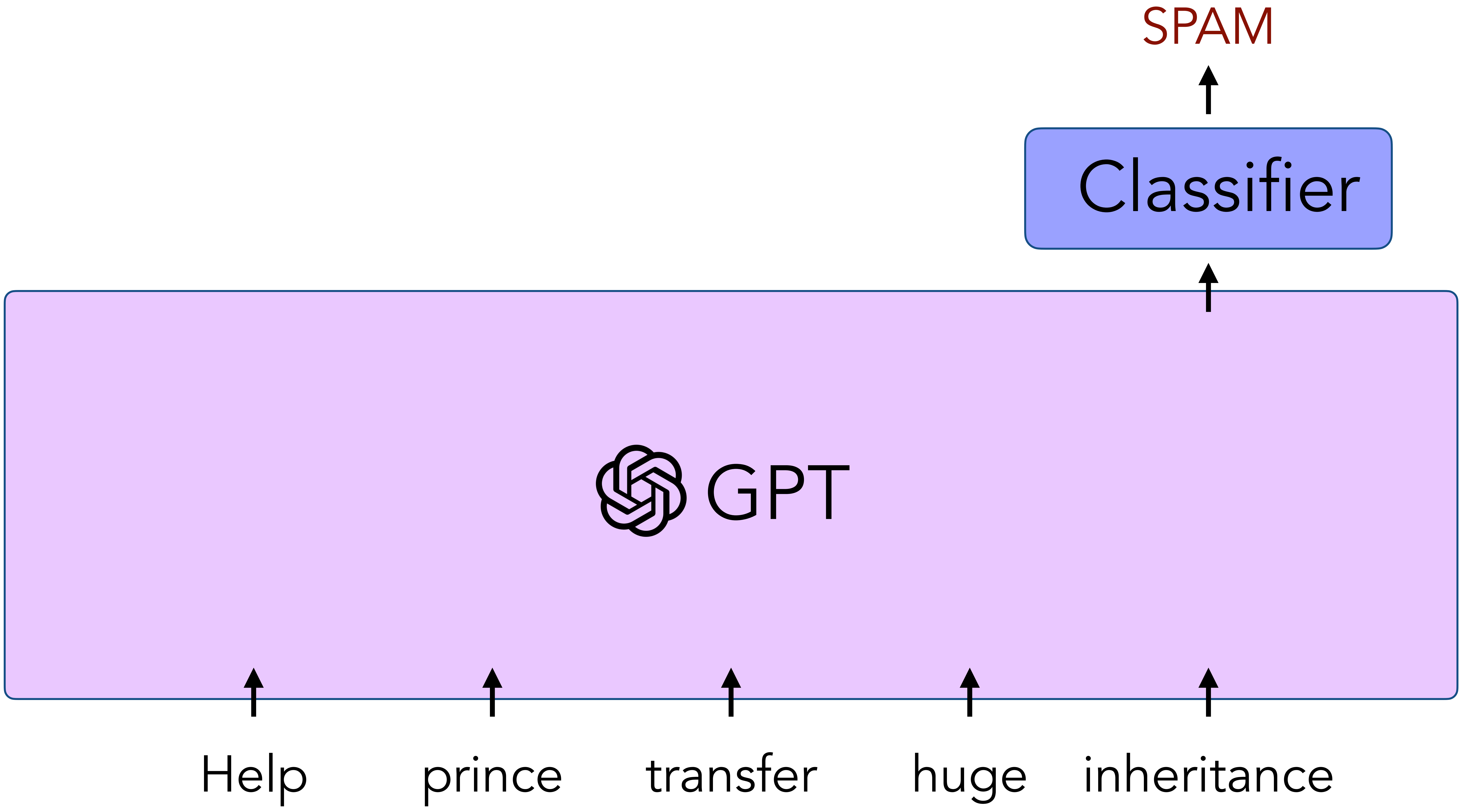
Pre-Training with LMs



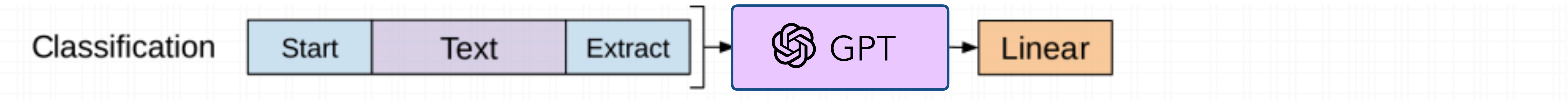
pre-training

Pre-Training with LMs

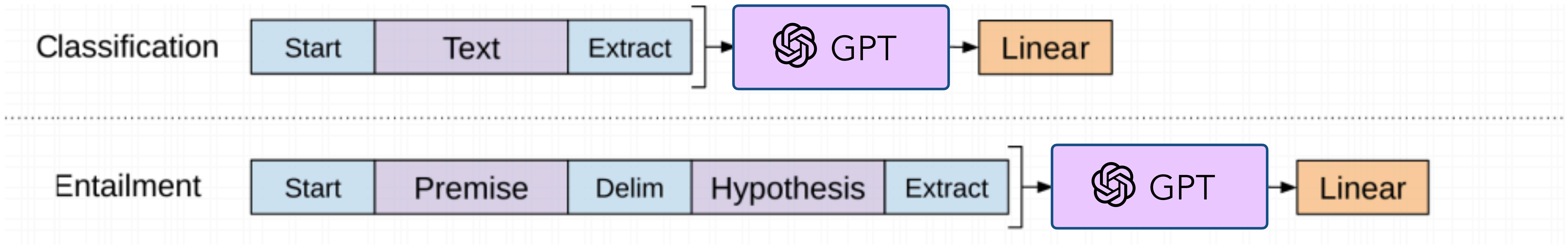
Fine-tuning with LMs



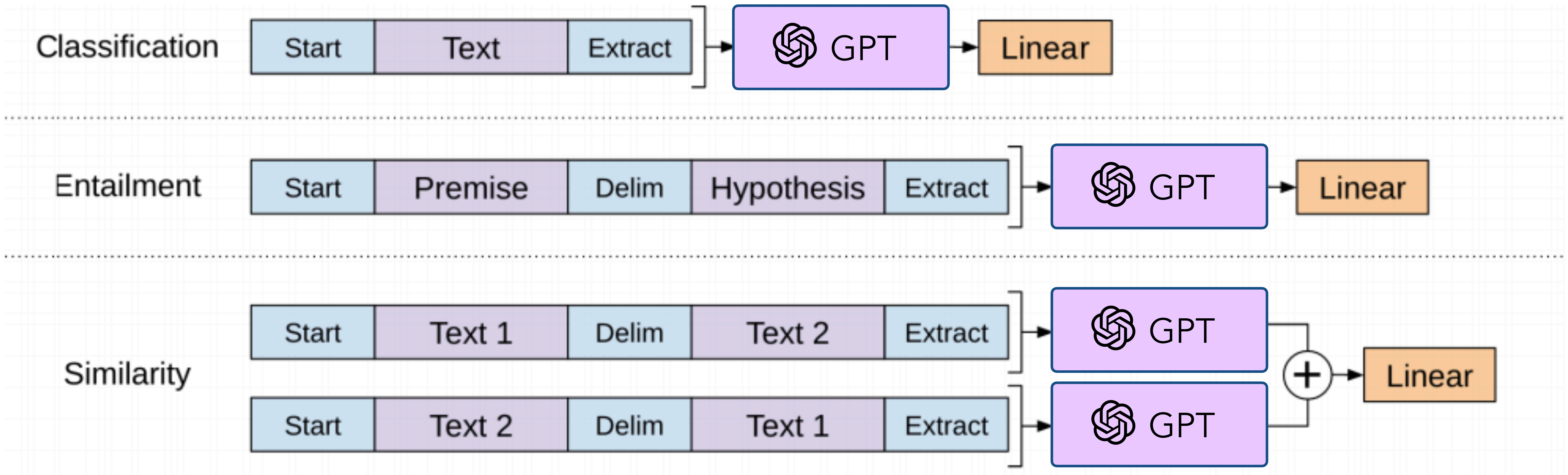
Fine-tuning with LMs



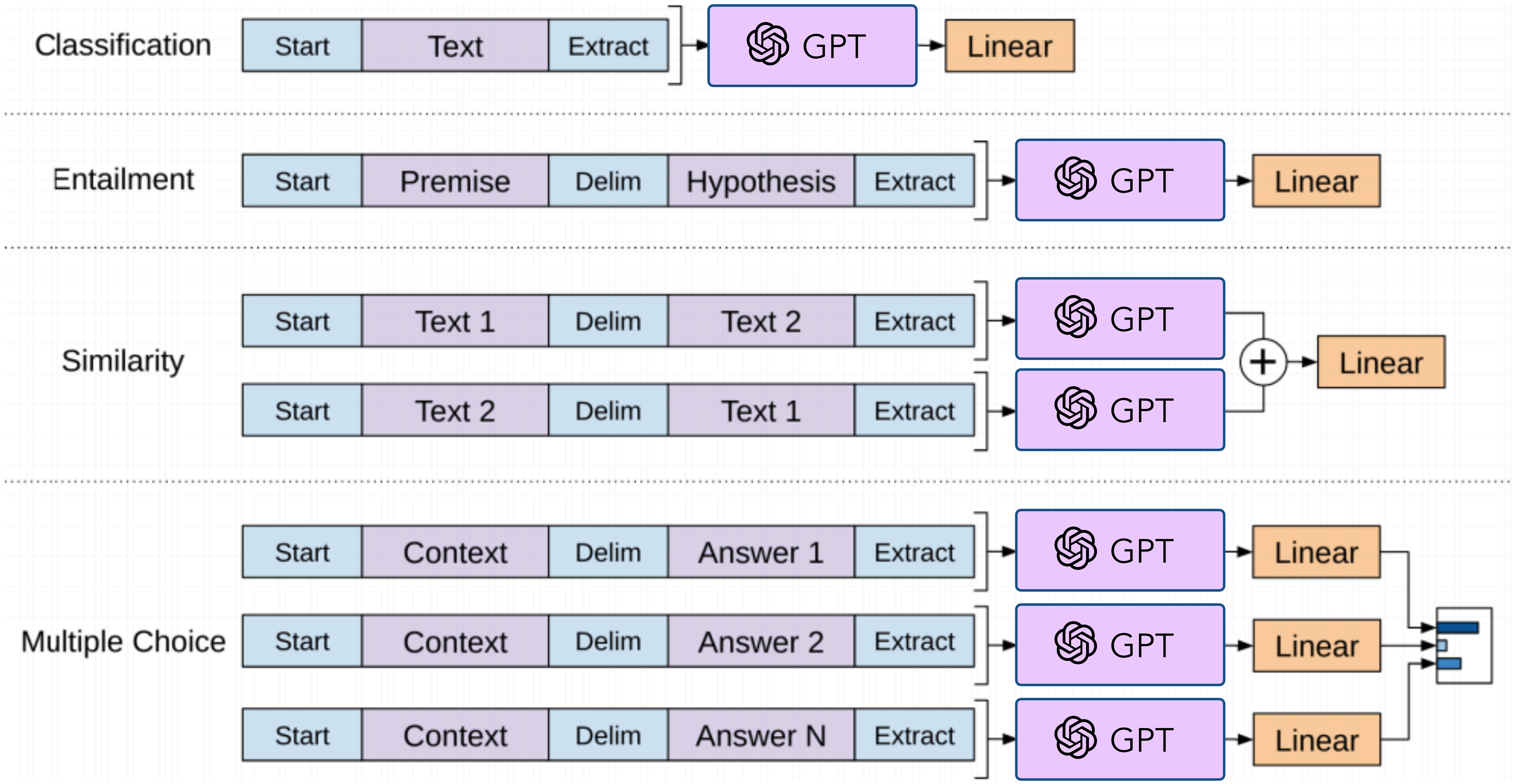
Fine-tuning with LMs



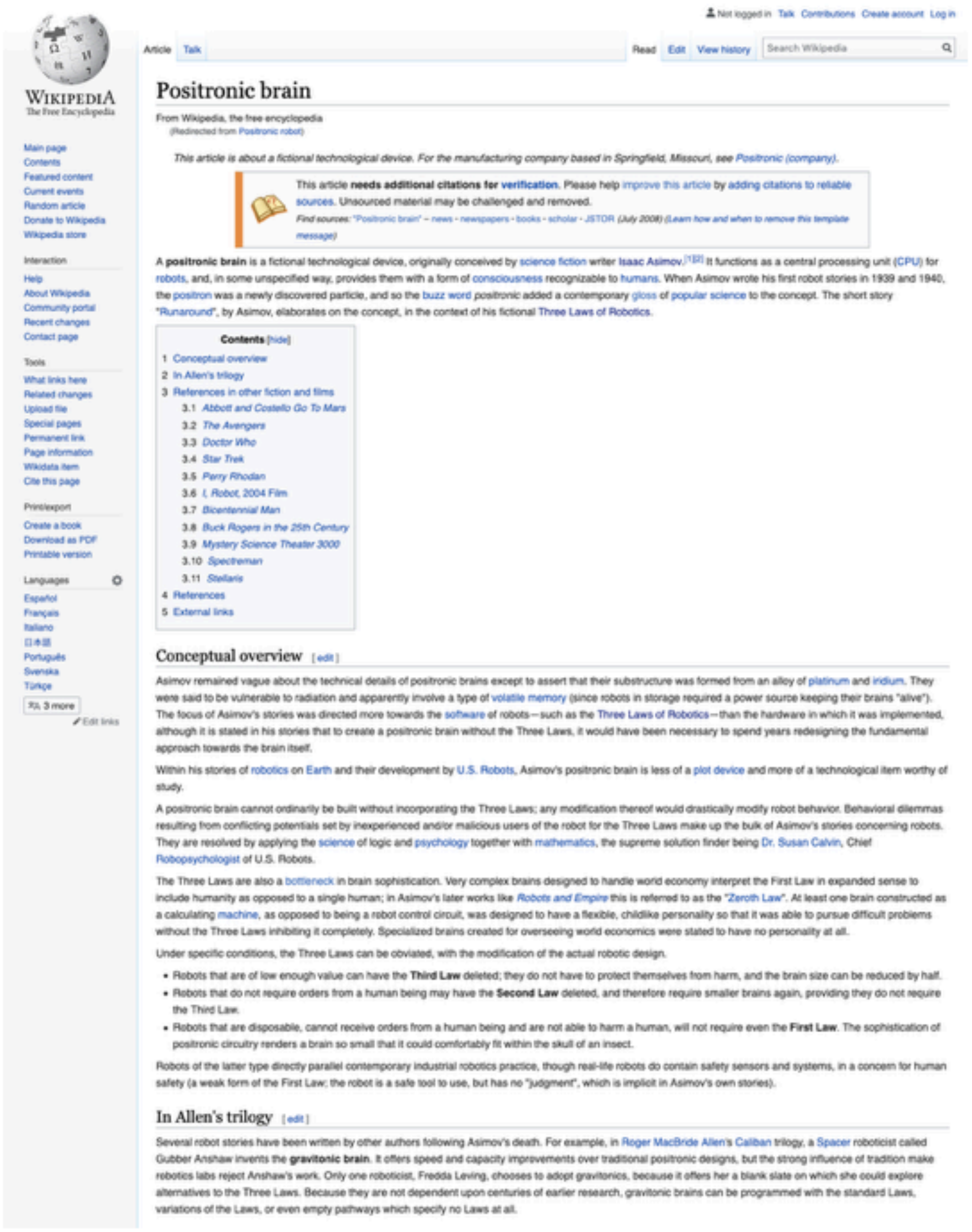
Fine-tuning with LMs



Fine-tuning with LMs



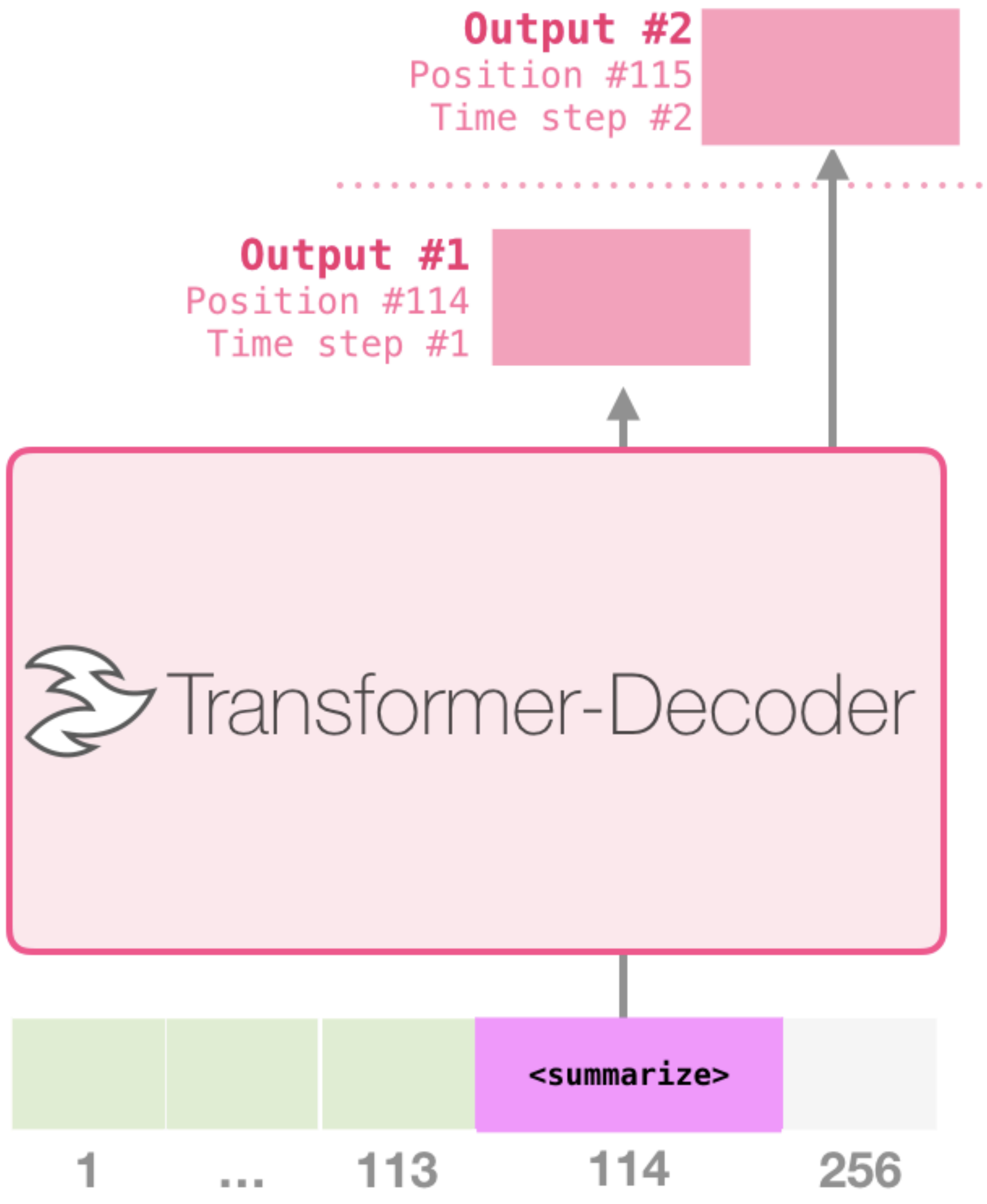
Summarization with LMs



Summarization with LMs

Training Dataset

Article #1 tokens	<summarize>	Article #1 Summary	
Article #2 tokens	<summarize>	Article #2 Summary	padding
Article #3 tokens	<summarize>	Article #3 Summary	



GLUE Benchmark

[Wang+19] GLUE: A Multi-Task Benchmark and Analysis Platform For Natural Language Understanding

https://gluebenchmark.com

Dataset	Description	Data example	Metric
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = **Ungrammatical**	Matthews
SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (Very Positive)**	Accuracy
MRPC	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = **A Paraphrase**	Accuracy / F1
STS-B	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = **4.6 (Very Similar)**	Pearson / Spearman
QQP	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = **Not Similar**	Accuracy / F1
MNLI-mm	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = **Contradiction**	Accuracy
QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = **Answerable**	Accuracy
RTE	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = **Entailed**	Accuracy
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = **Incorrect Referent**	Accuracy

[Figure by Chris McCormick and Nick Ryan]

GLUE Benchmark Results

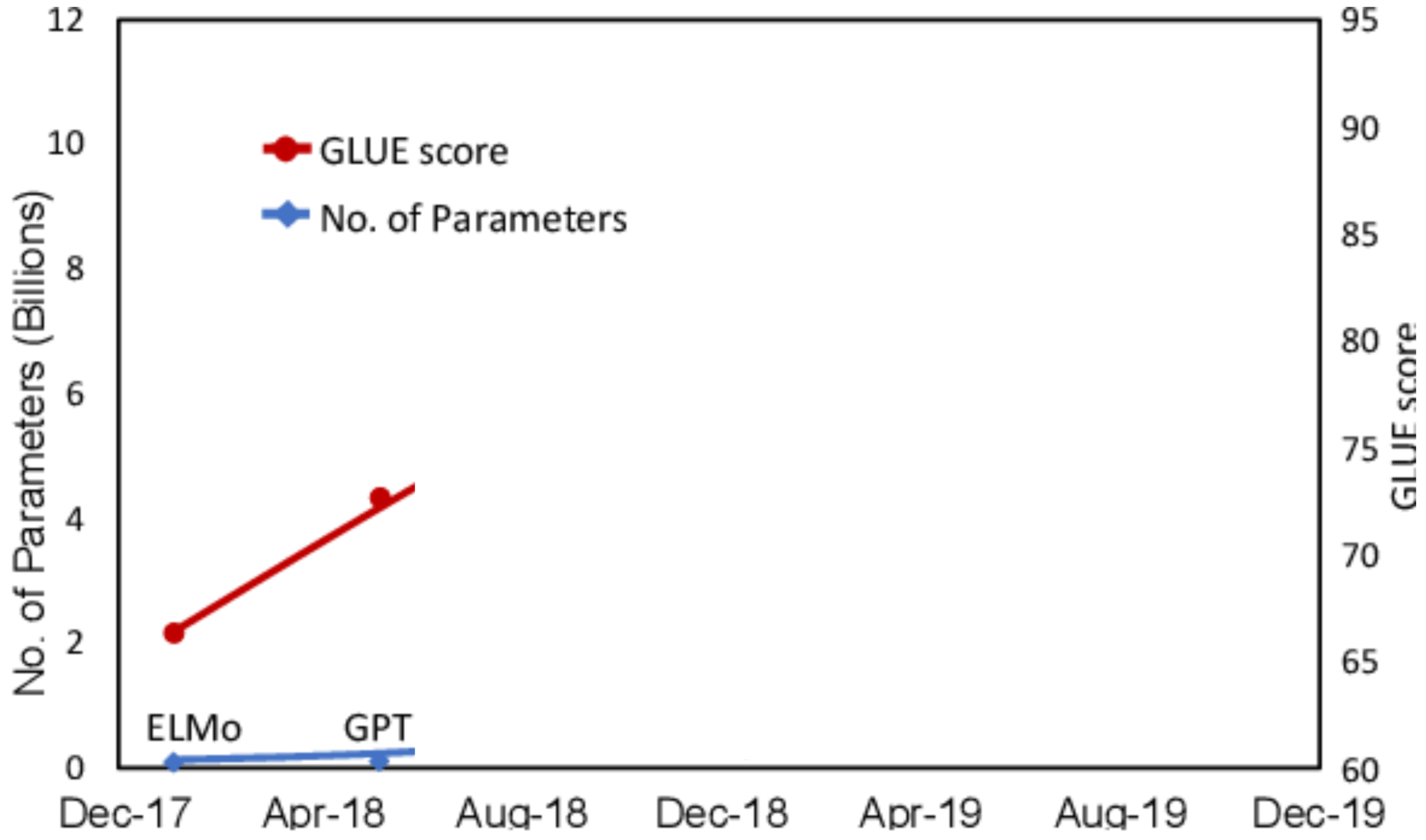
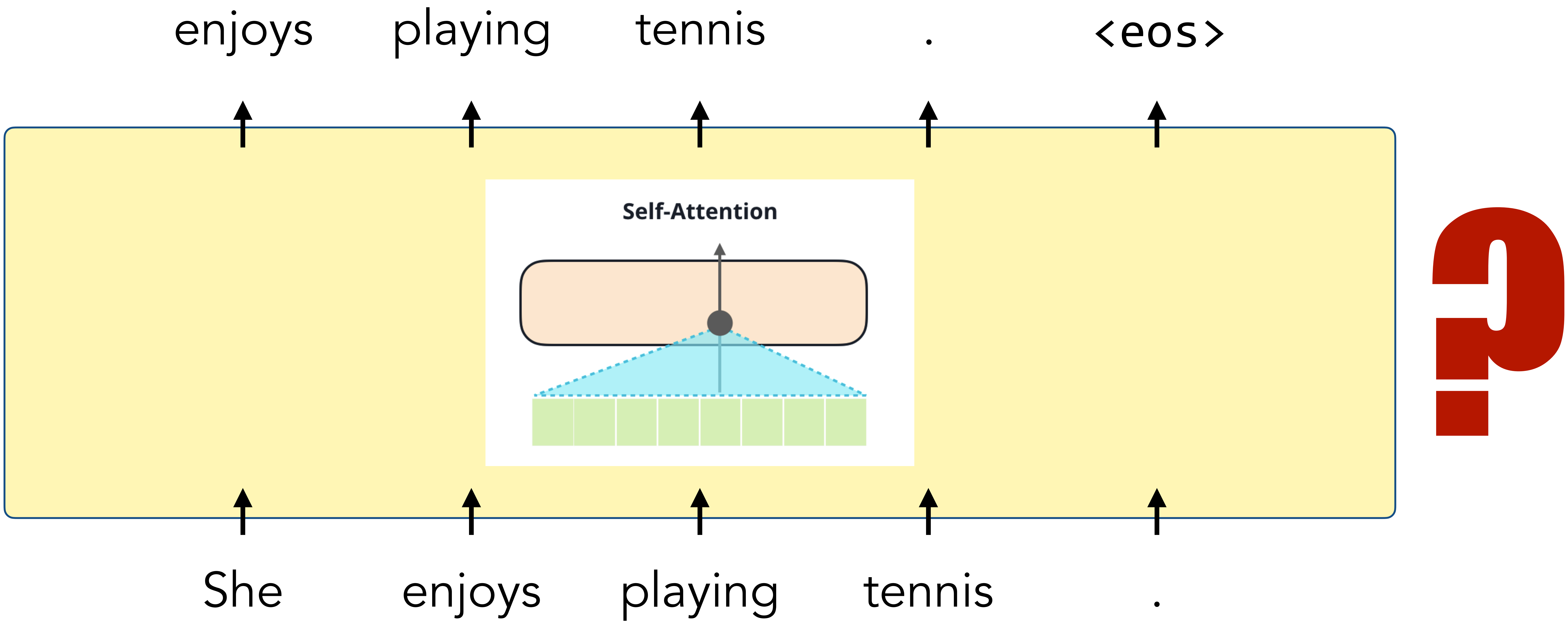


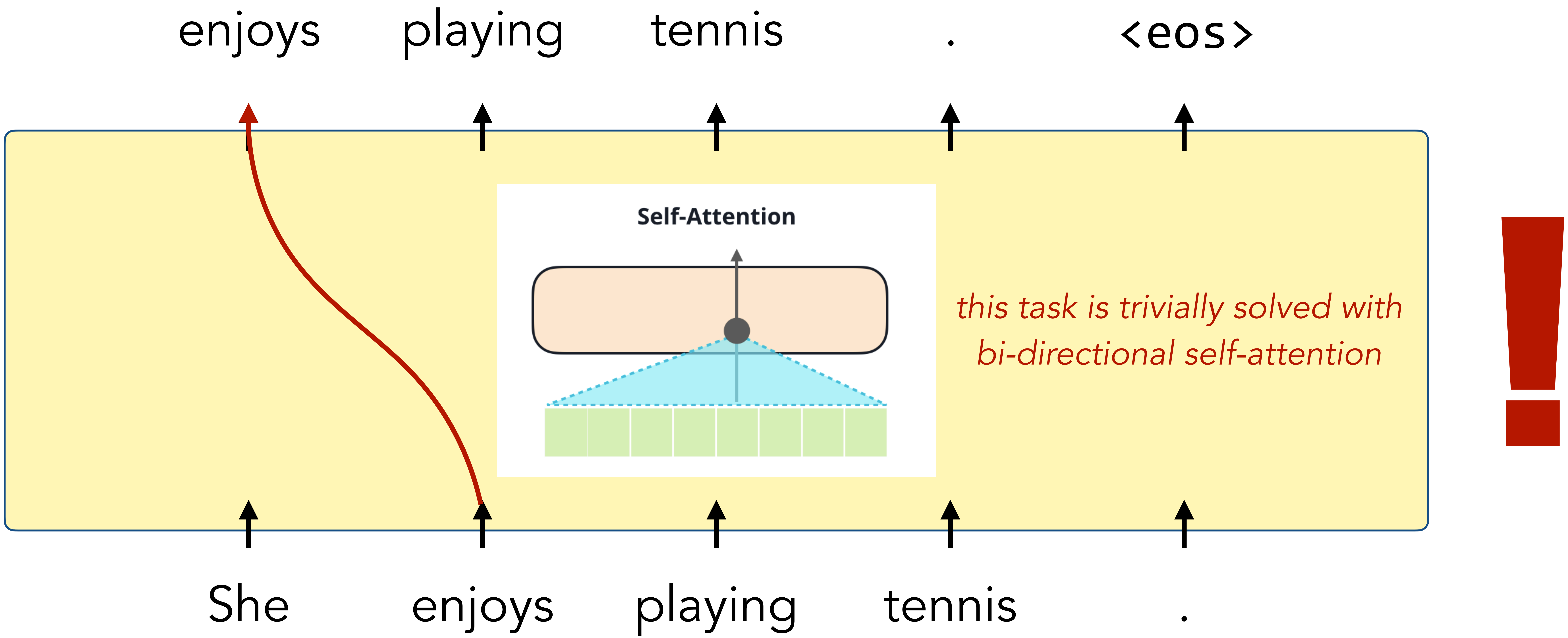
Fig. 1: Language Model Size & GLUE Performance

[Figure from Ahmet & Abdullah, 2020]

Bi-directional Pre-Training

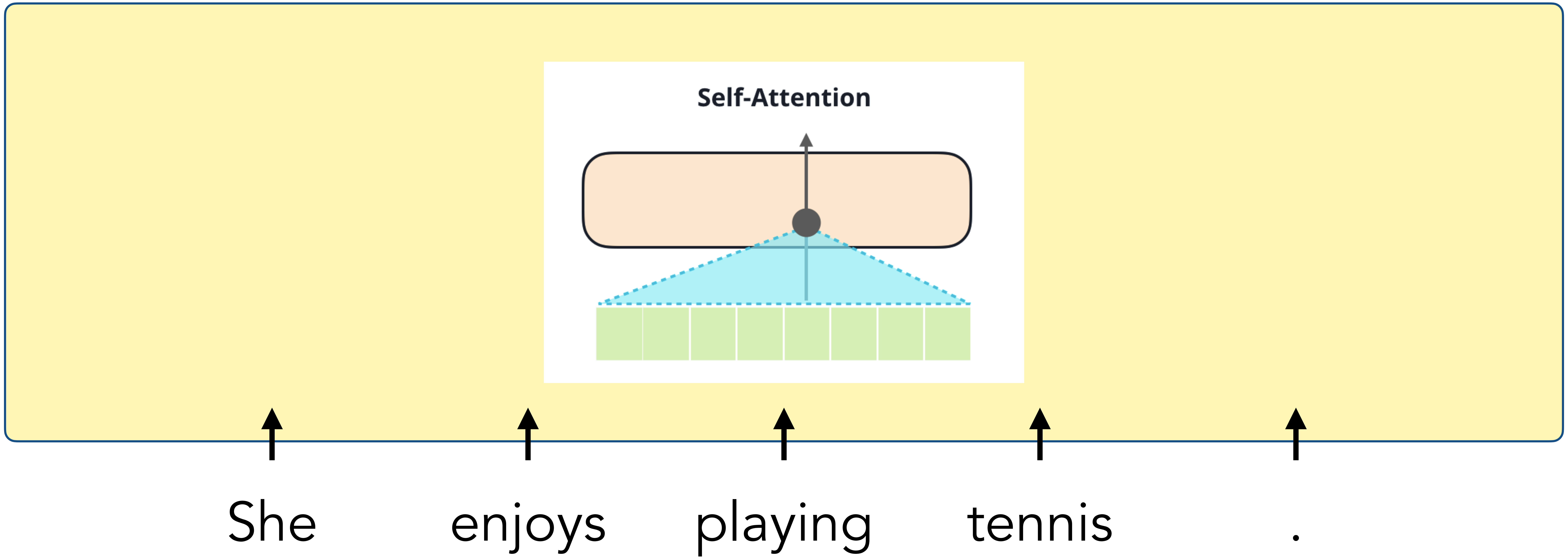


Bi-directional Pre-Training



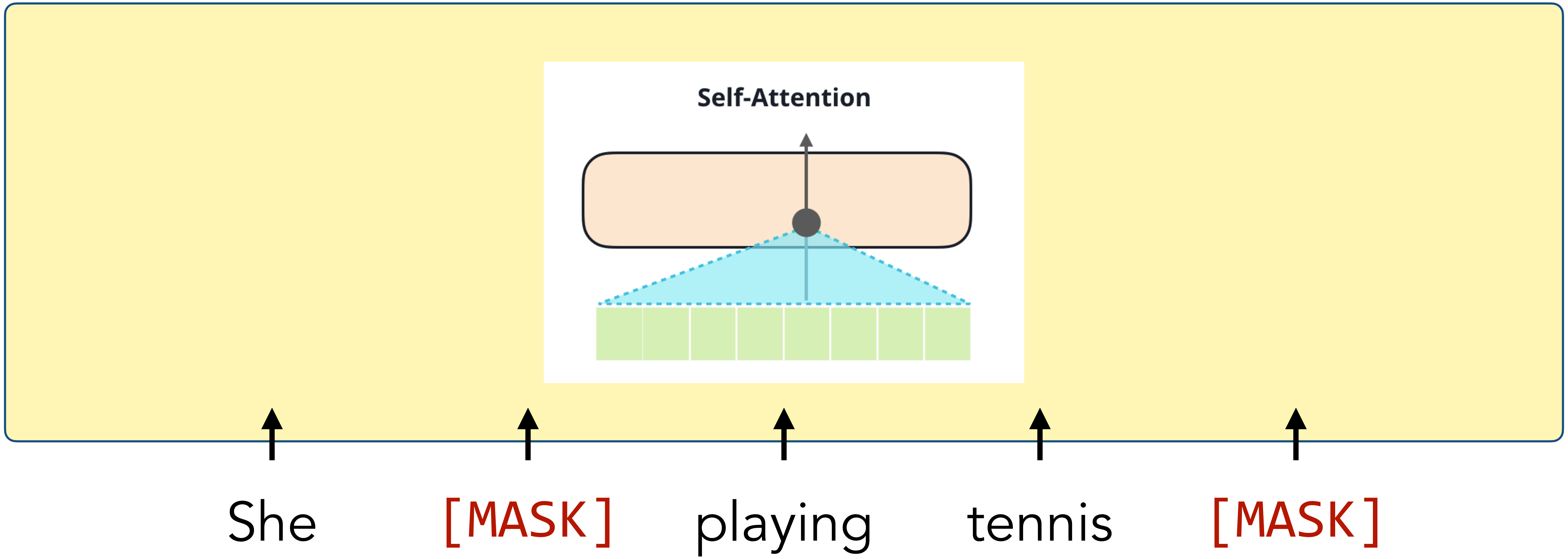
Masked Language Model

Mask out 15% of tokens, then predict the missing tokens



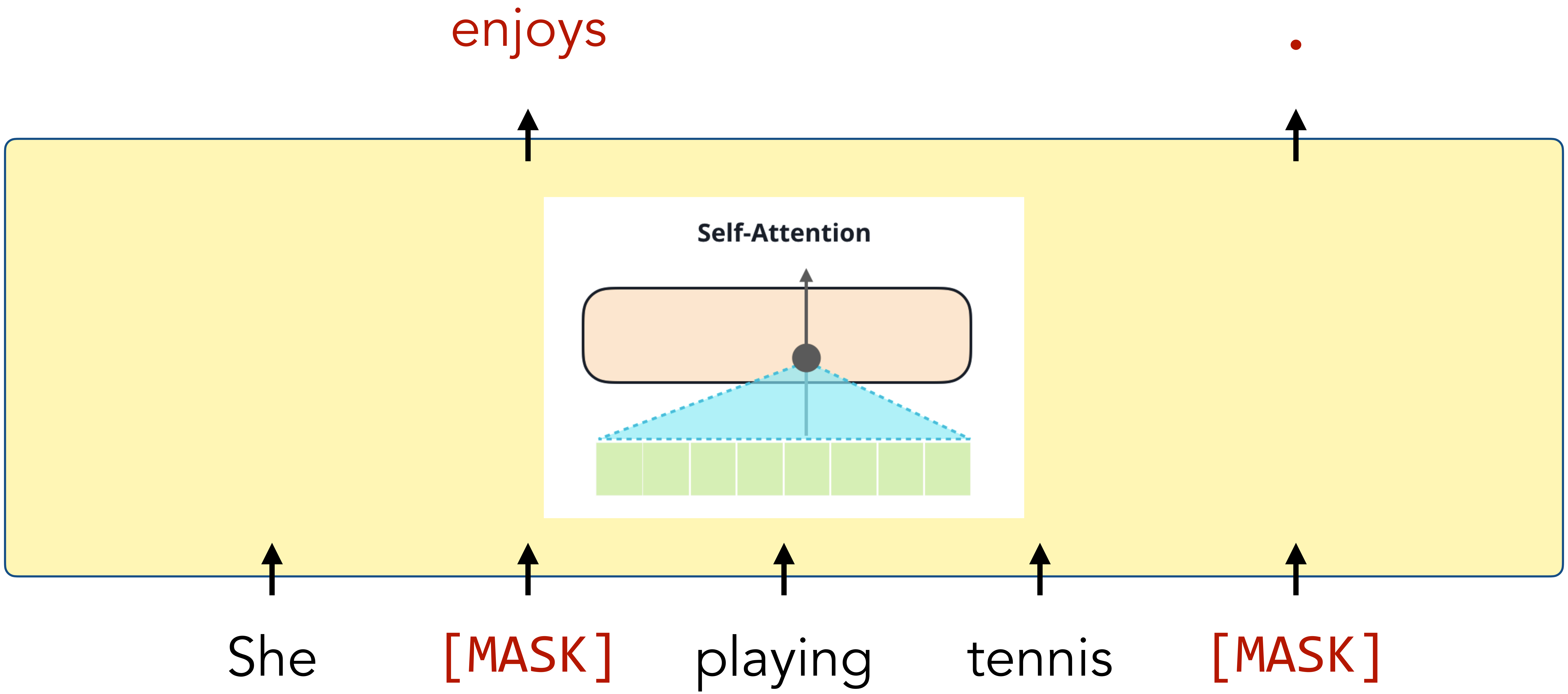
Masked Language Model

Mask out 15% of tokens, then predict the missing tokens



Masked Language Model

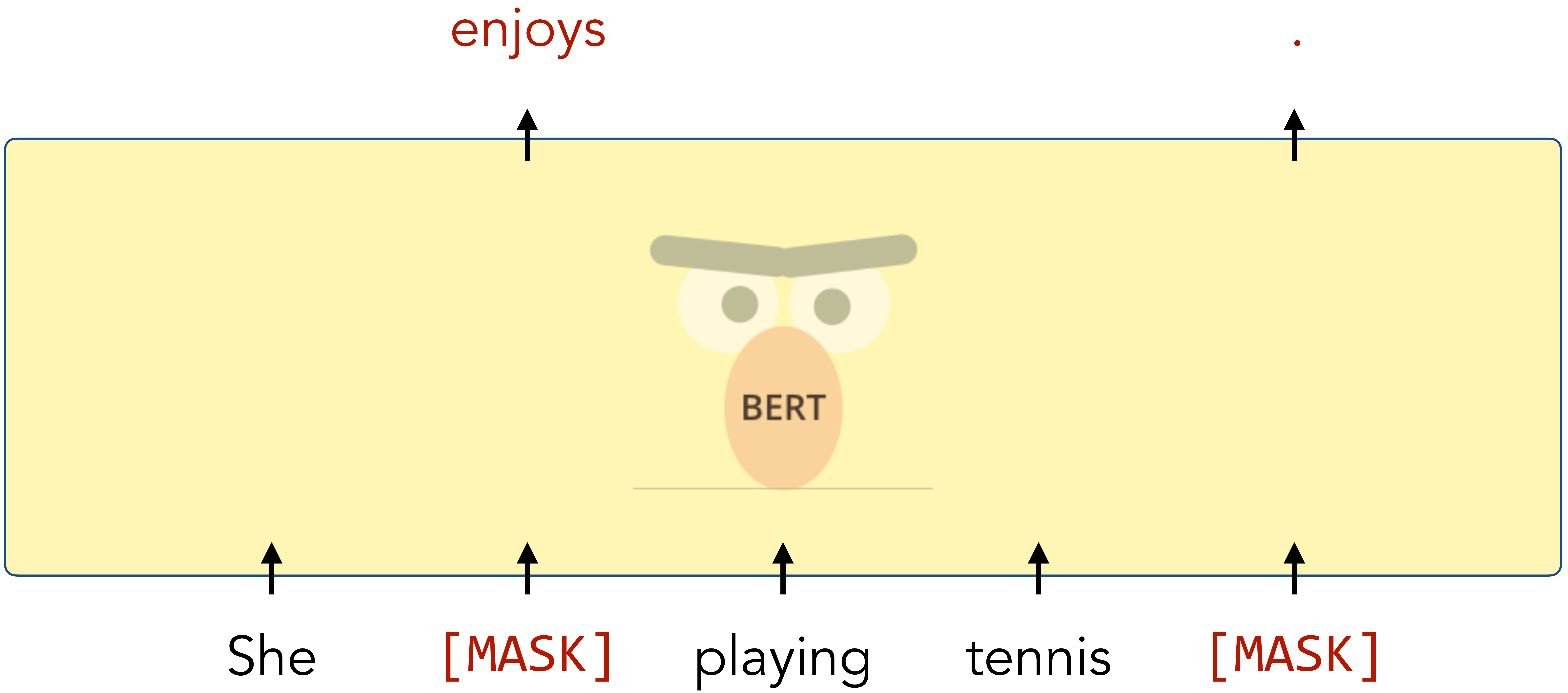
Mask out 15% of tokens, then predict the missing tokens



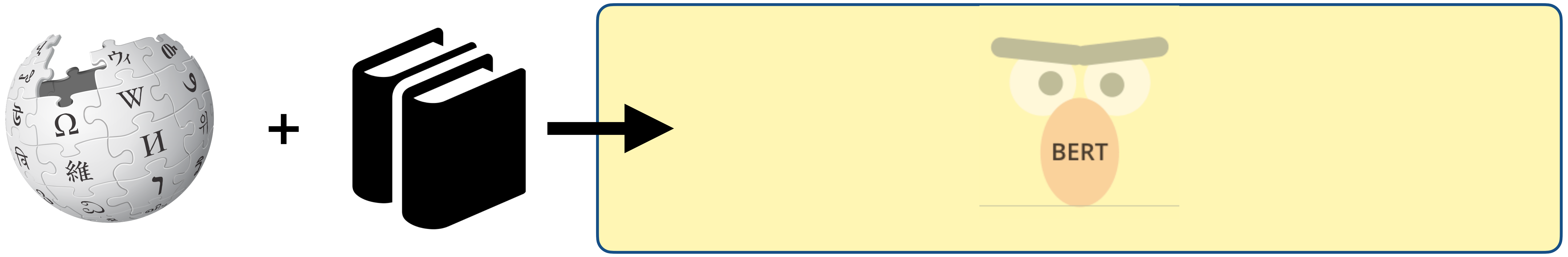
Masked Language Model

Representative Model: BERT

(BERT = Bidirectional Encoder Representations from Transformers)

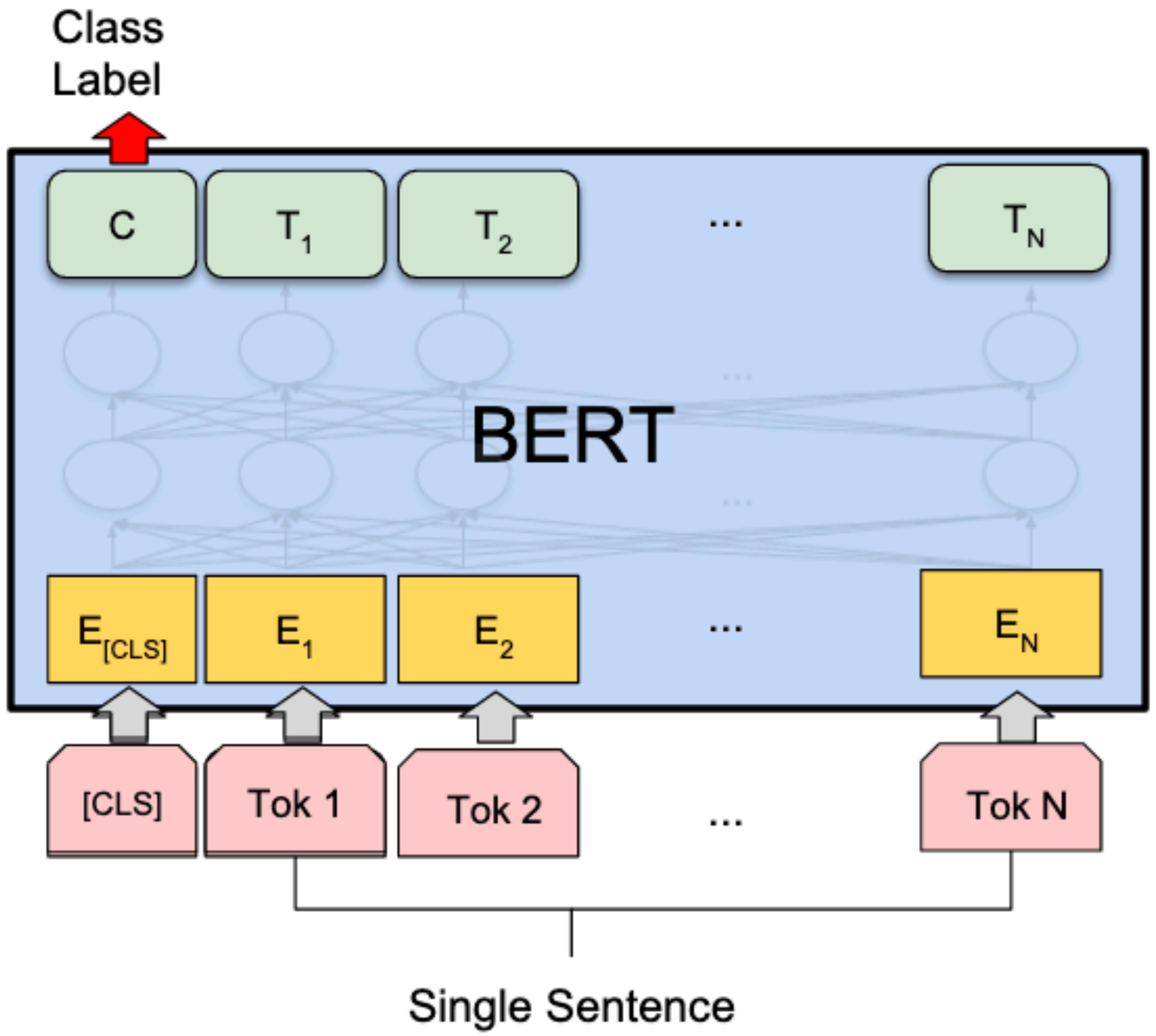


Pre-Training with Masked LMs



pre-training

Fine-Tuning with Masked LMs



Fine-Tuning with Masked LMs

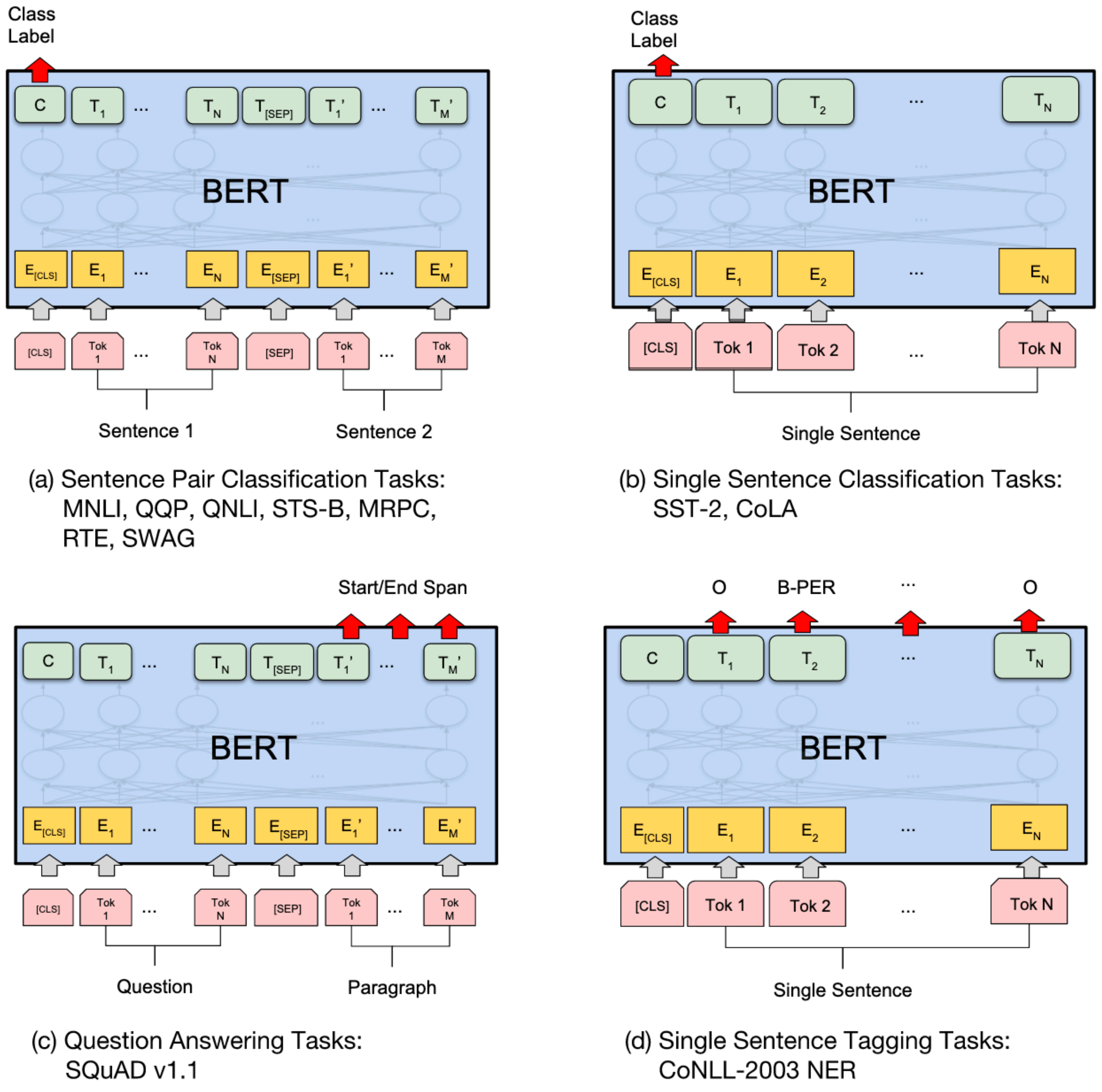
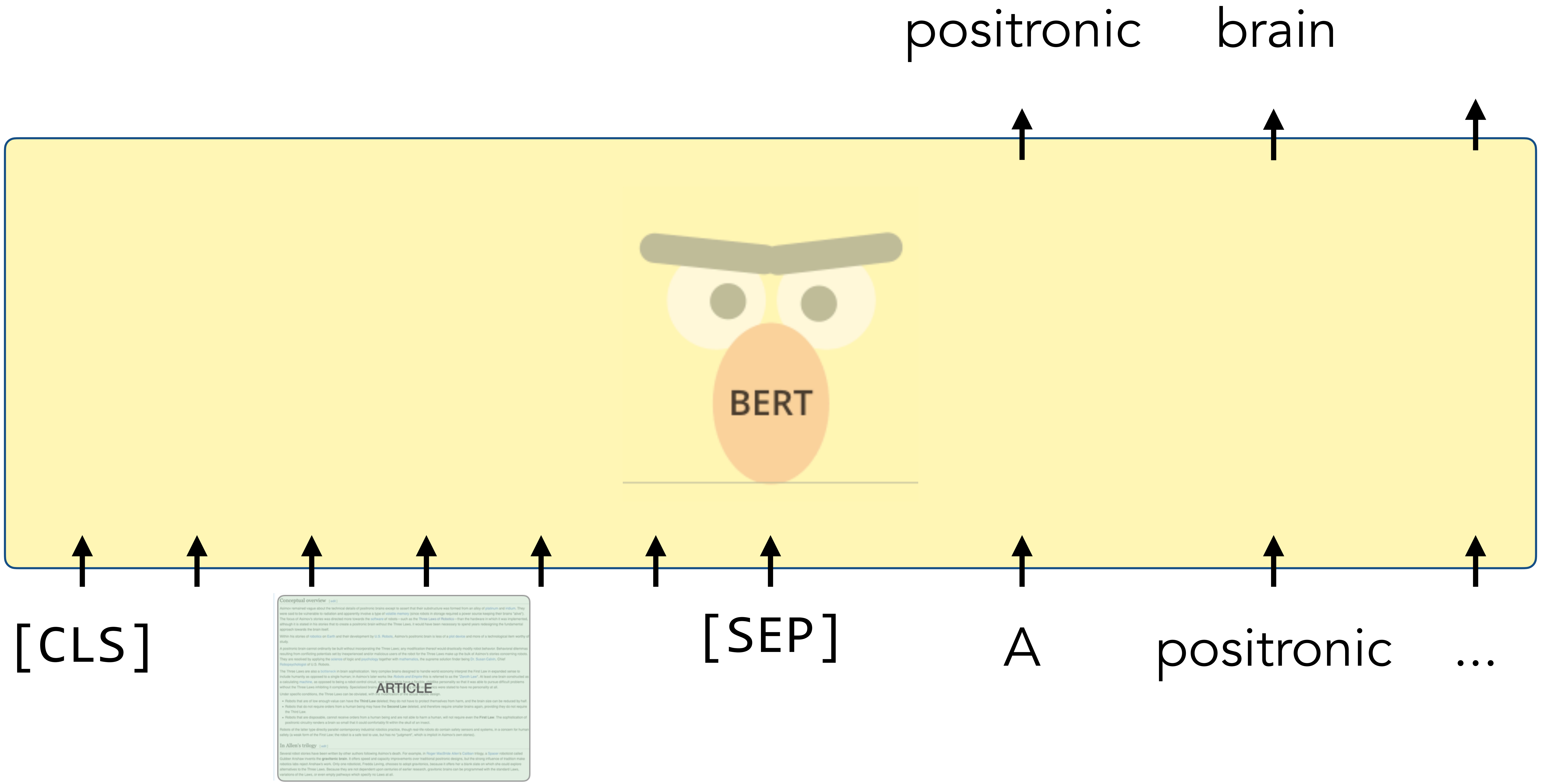


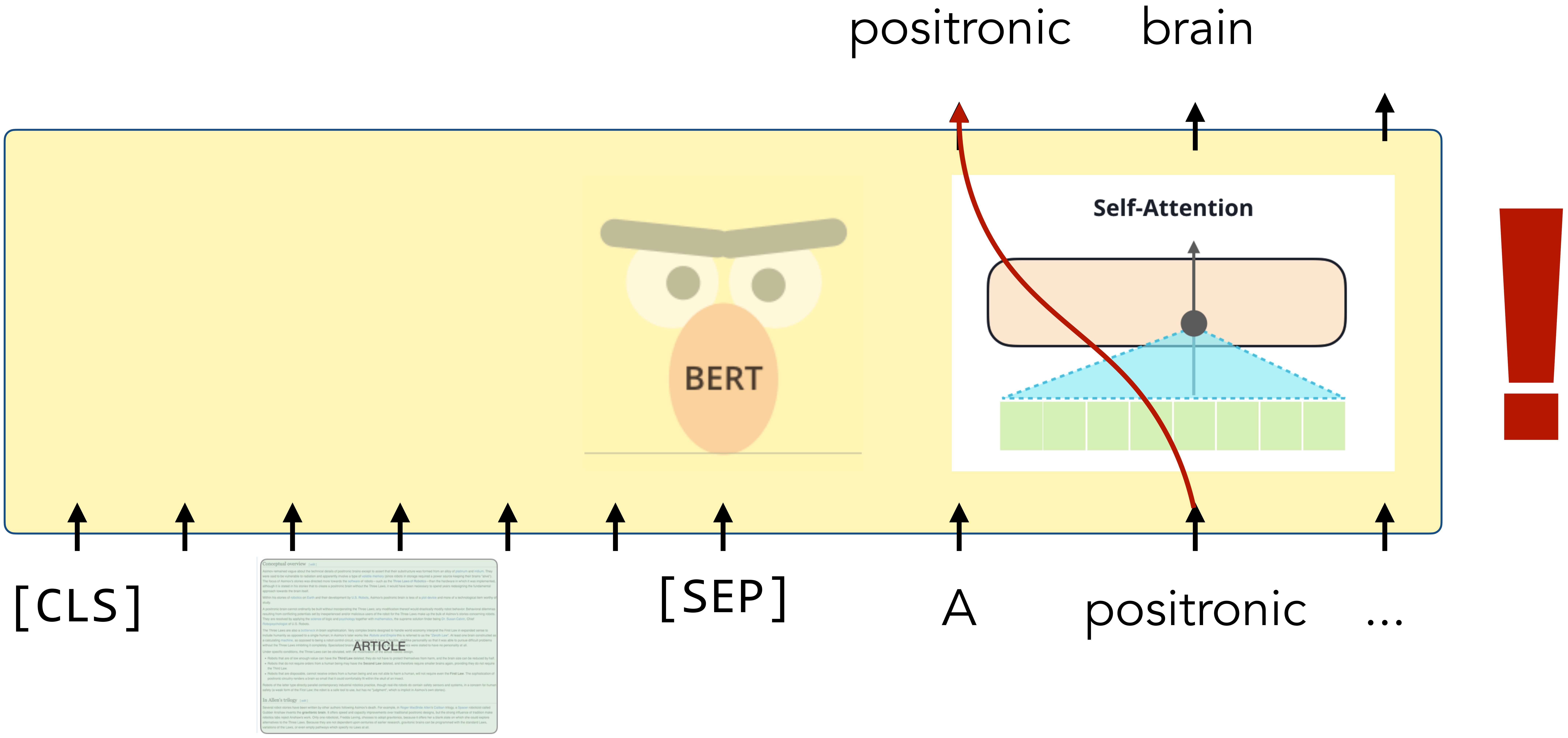
Figure 4: Illustrations of Fine-tuning BERT on Different Tasks.

Summarization with Masked LMs?



Summarization with Masked LMs?

Bi-directional Masked LMs are not ideal for sequence-to-sequence tasks



GLUE Benchmark Results

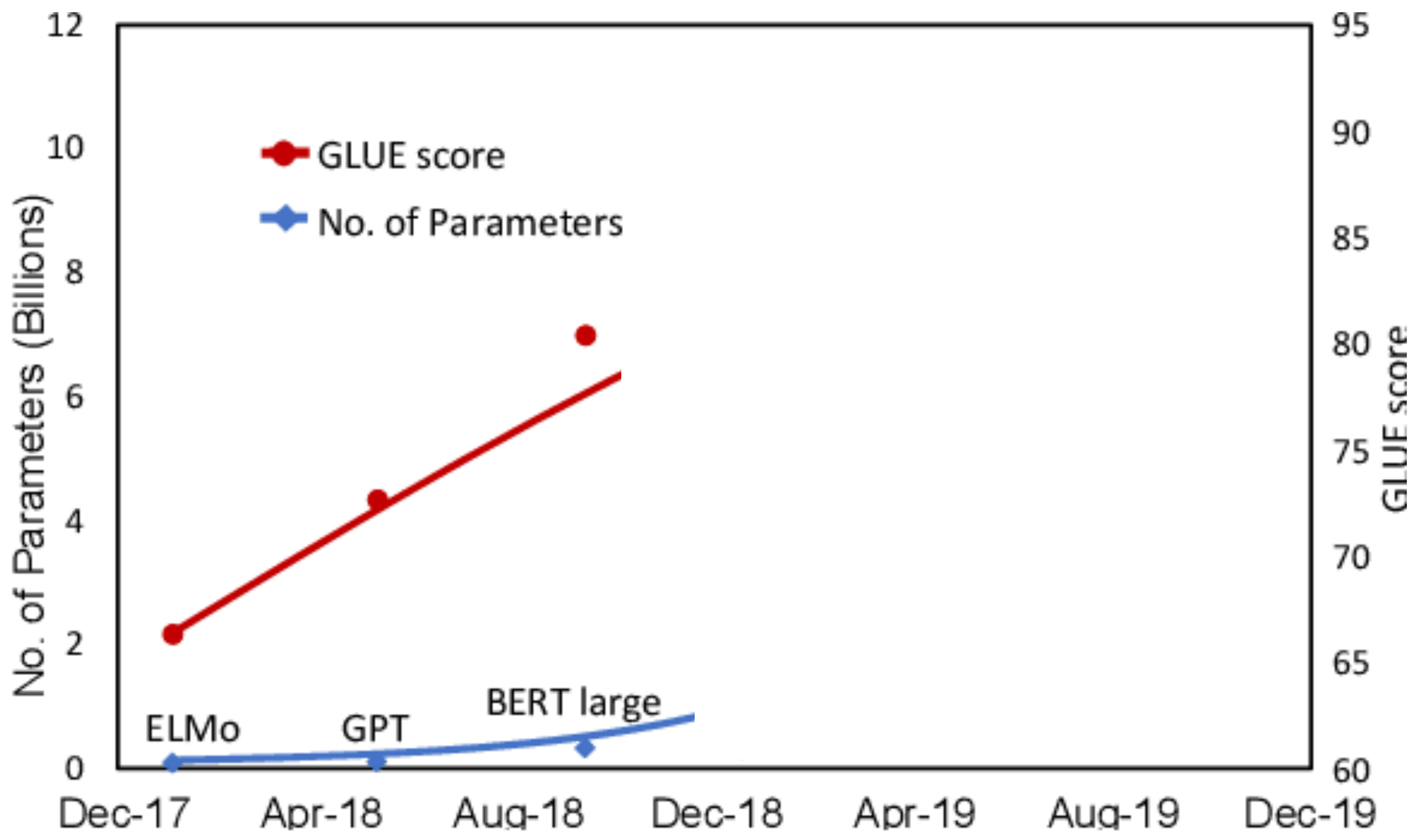
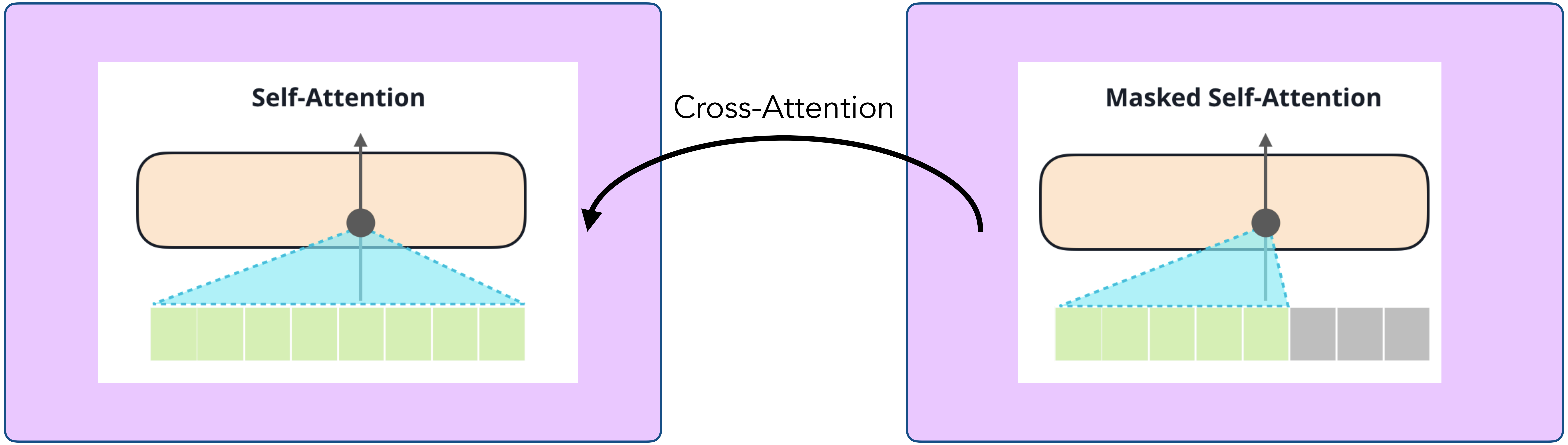


Fig. 1: Language Model Size & GLUE Performance

[Figure from Ahmet & Abdullah, 2020]

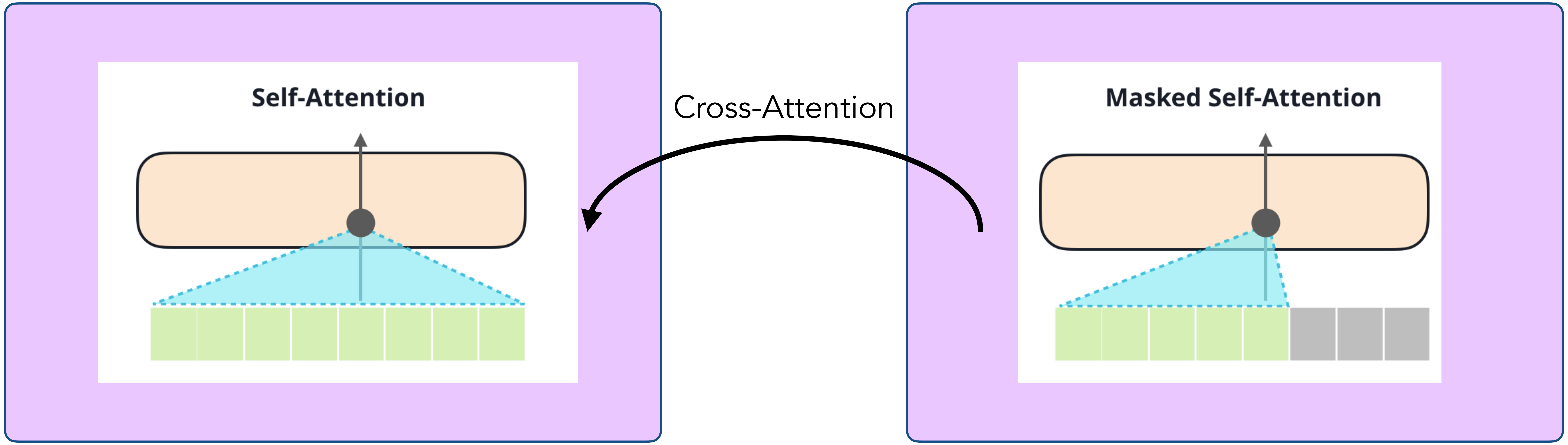
Encoder-Decoder Pre-Training



Encoder-Decoder Pre-Training

Representative Model: T5

(T5 = Text-To-Text Transfer Transformer)



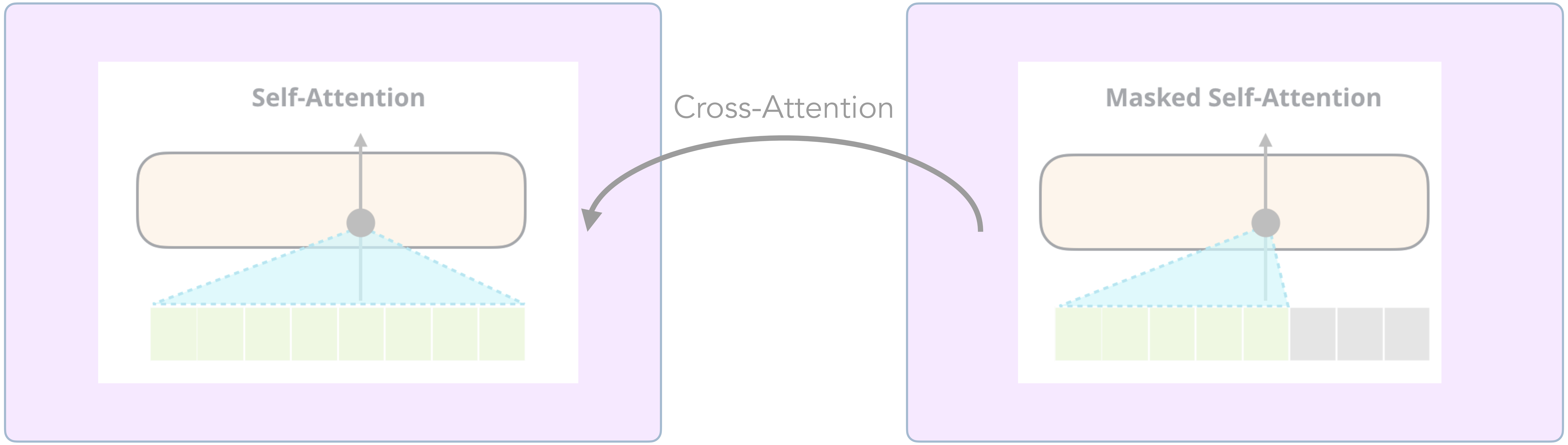
Encoder-Decoder Pre-Training

Original text

Thank you ~~for inviting~~ me to your party ~~last~~ week.

Inputs

Thank you <X> me to your party <Y> week.



Encoder-Decoder Pre-Training

Original text

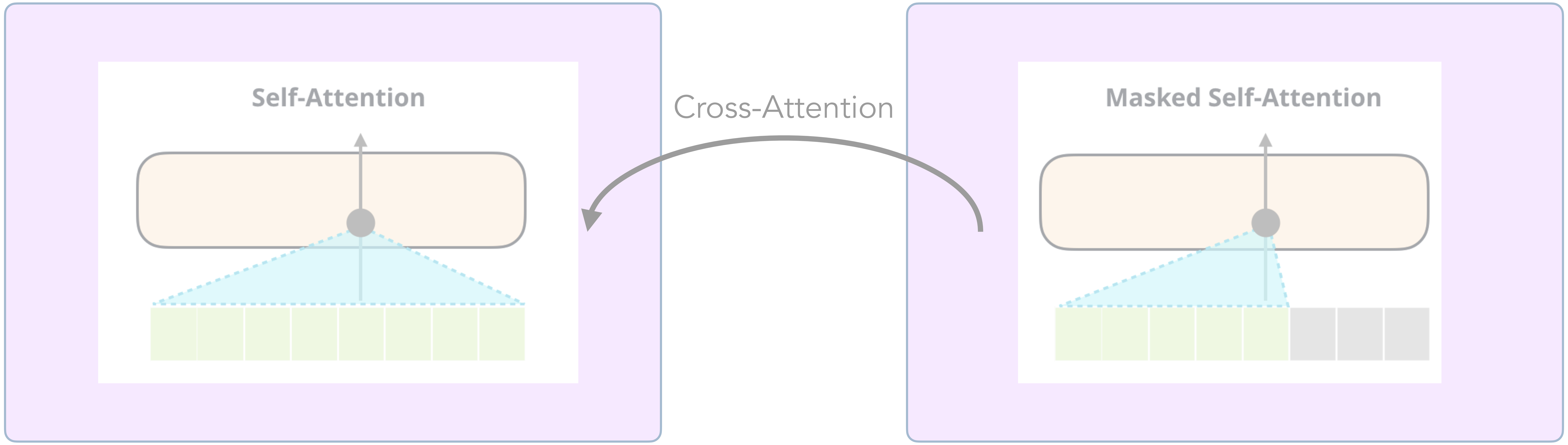
Thank you ~~for inviting~~ me to your party ~~last~~ week.

Inputs

Thank you <X> me to your party <Y> week.

Targets

<X> for inviting <Y> last <Z>



Encoder-Decoder Pre-Training

Original text

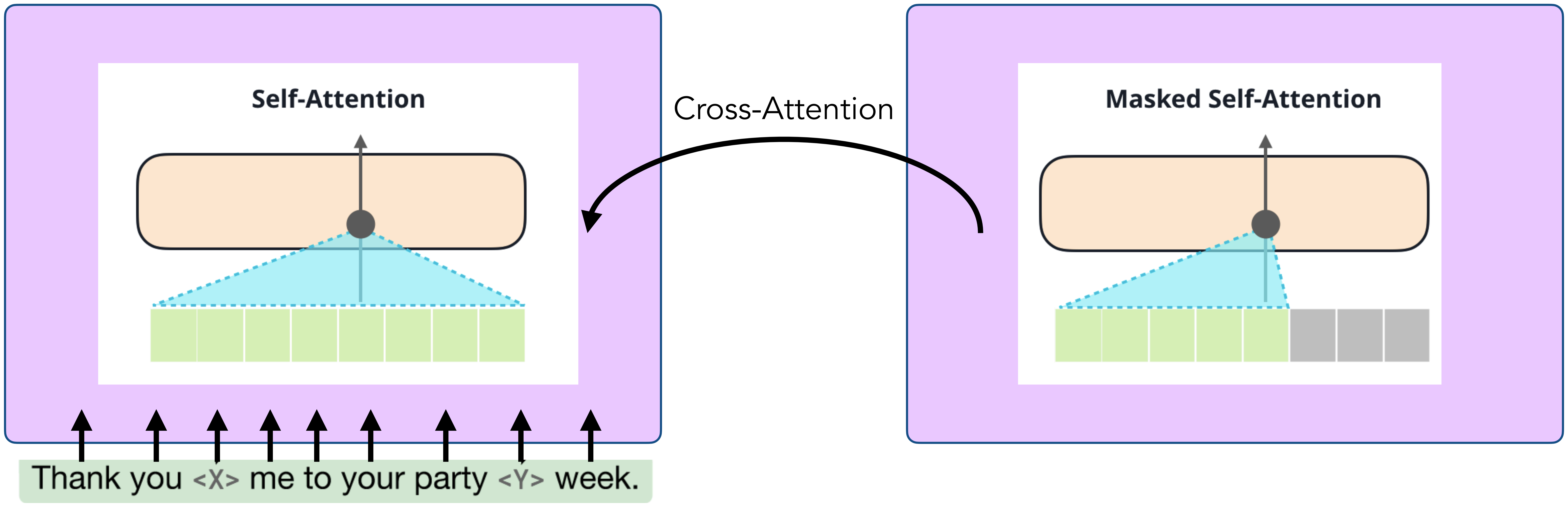
Thank you ~~for inviting~~ me to your party ~~last~~ week.

Inputs

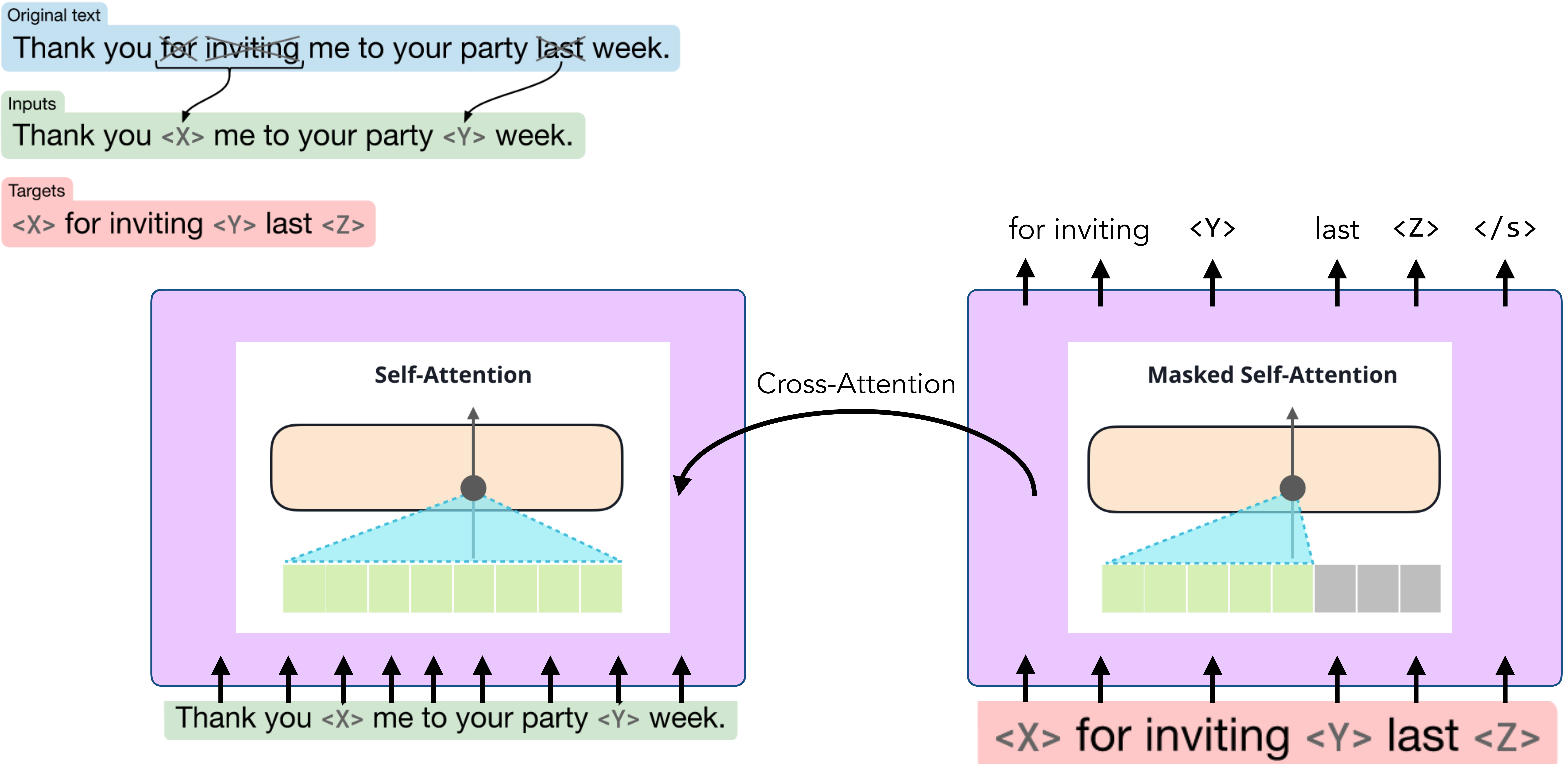
Thank you <X> me to your party <Y> week.

Targets

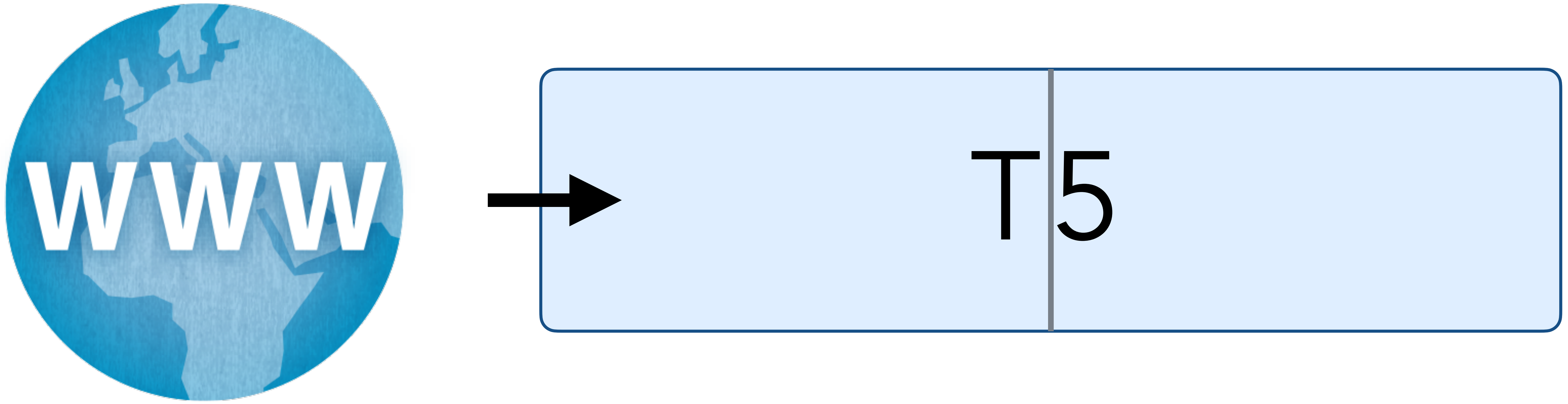
<X> for inviting <Y> last <Z>



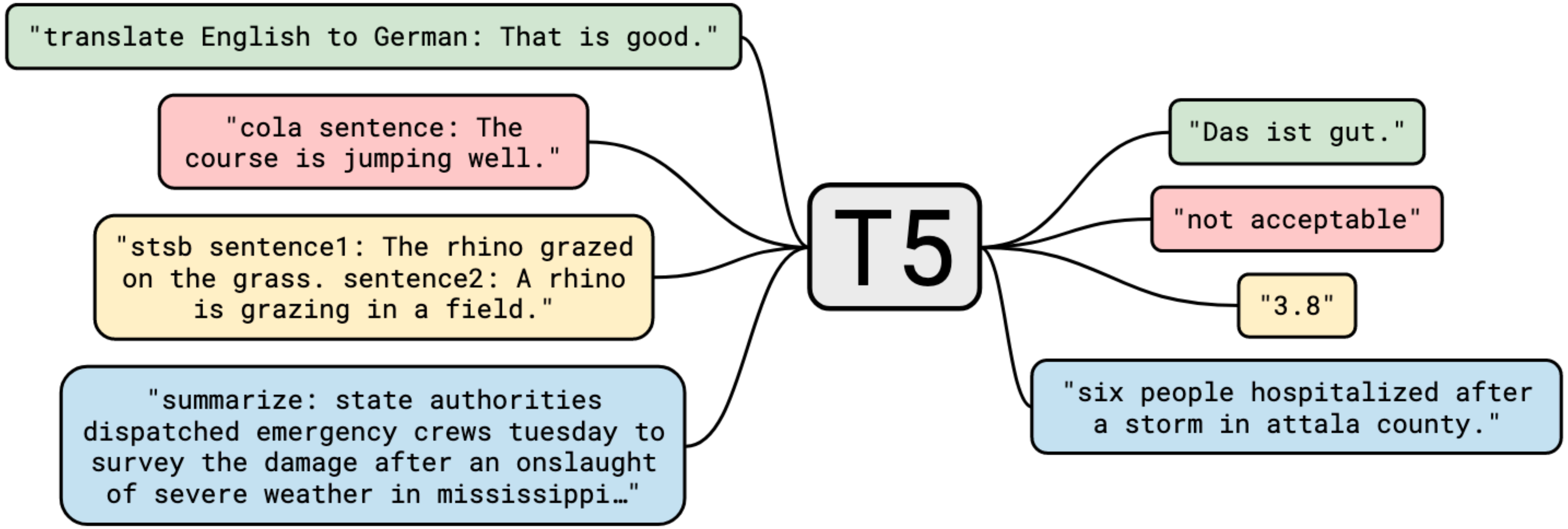
Encoder-Decoder Pre-Training



Encoder-Decoder Pre-Training



Encoder-Decoder Fine-tuning



GLUE Benchmark Results

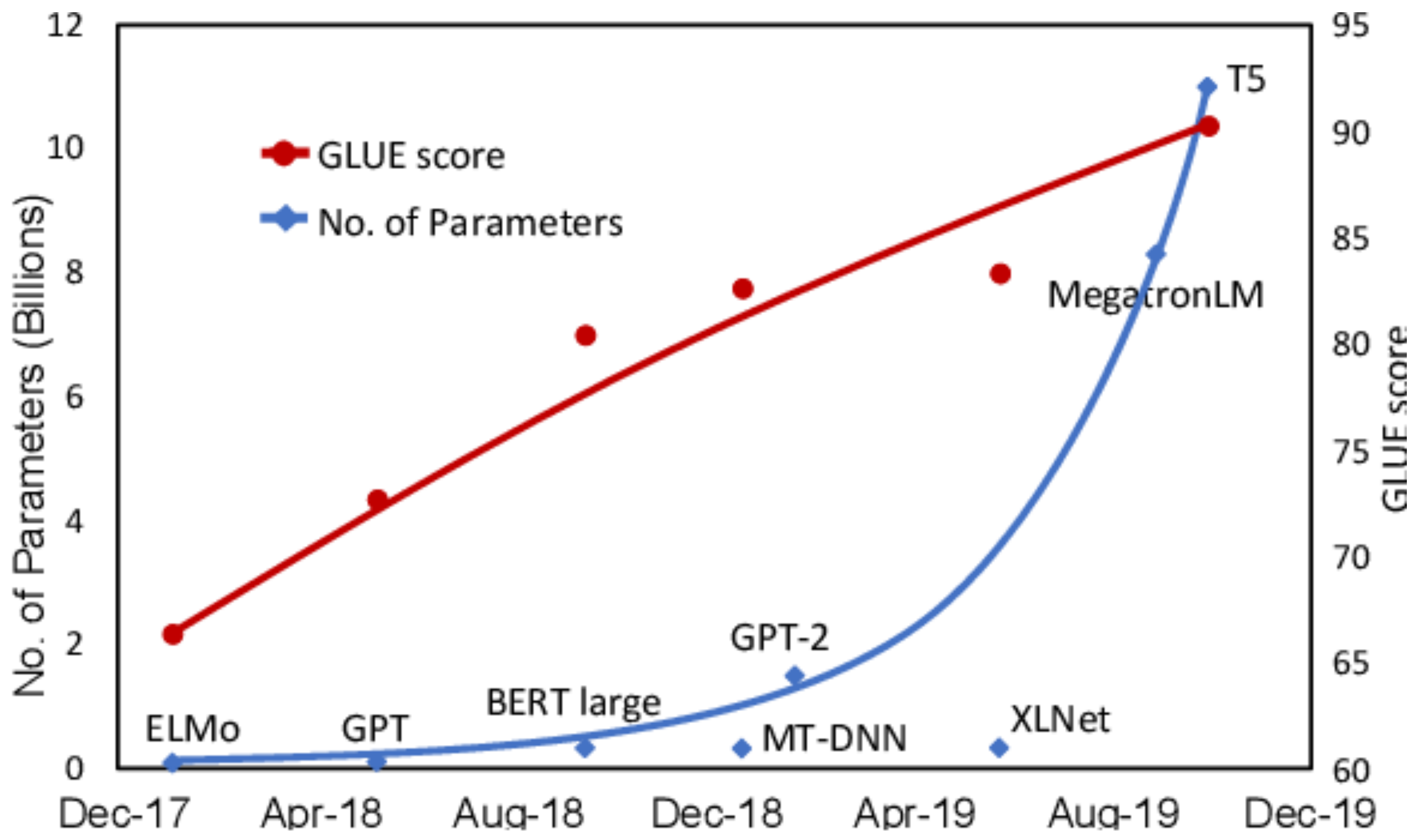


Fig. 1: Language Model Size & GLUE Performance

[Figure from Ahmet & Abdullah, 2020]

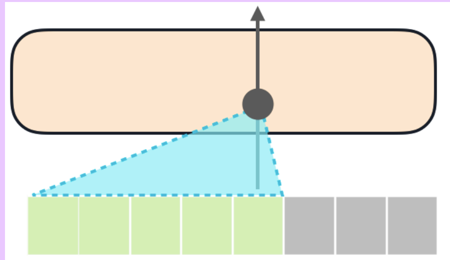
GLUE Benchmark Results

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP
1	T5 Team - Google	T5		89.7	70.8	97.1	91.9/89.2	92.5/92.1	74.6/90.4
2	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5
+ 3	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		89.0	69.2	97.1	93.6/91.5	92.7/92.3	74.4/90.7
4	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.8	68.0	96.8	93.1/90.8	92.4/92.2	74.8/90.3
5	Facebook AI	RoBERTa		88.5	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2
6	XLNet Team	XLNet-Large (ensemble)		88.4	67.8	96.8	93.0/90.7	91.6/91.1	74.2/90.3
+ 7	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9
8	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4
9	Stanford Hazy Research	Snorkel MeTaL		83.2	63.8	96.2	91.5/88.5	90.1/89.7	73.1/89.9
10	XLM Systems	XLM (English only)		83.1	62.9	95.6	90.7/87.1	88.8/88.2	73.2/89.8

[Figure by Chris McCormick and Nick Ryan]

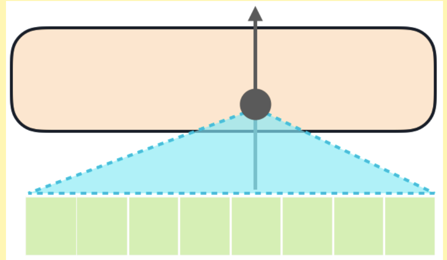
Types of Transformer Models

Decoder only



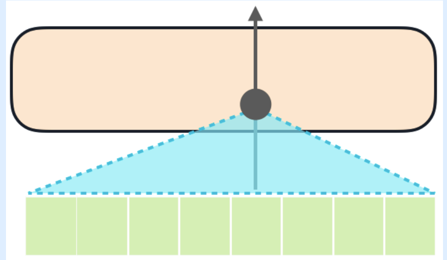
e.g. GPT

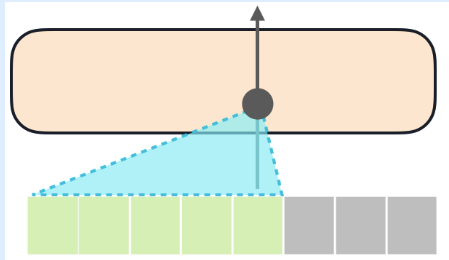
Encoder only



e.g.

Encoder-Decoder





e.g. T5

Types of Task-Specific Adaptation

Fine-tuning: modify existing model parameters

Adapter modules: freeze existing parameters; insert and train new layers

(not covered in these slides)

Prompting: re-formulate a task in natural language (e.g. fill-in-the-blank)

Example from GPT-3 LM

SYSTEM PROMPT (HUMAN-WRITTEN)	*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)	The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow. Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez. Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns. While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic." Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America. While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

[https://openai.com/blog/better-language-models]

Examples of Prompting

Prompt

I am a highly intelligent question answering bot. If you ask me a question that is rooted in truth, I will give you the answer. If you ask me a question that is nonsense, trickery, or has no clear answer, I will respond with "Unknown".

Q: What is human life expectancy in the United States?
A: Human life expectancy in the United States is 78 years.

Q: Who was president of the United States in 1955?
A: Dwight D. Eisenhower was president of the United States in 1955.

Q: Which party did he belong to?
A: He belonged to the Republican Party.

Q: What is the square root of banana?
A: Unknown

Q: How does a telescope work?
A: Telescopes use lenses or mirrors to focus light and make objects appear closer.

Q: Where were the 1992 Olympics held?
A: The 1992 Olympics were held in Barcelona, Spain.

Q: How many squigs are in a bonk?
A: Unknown

Q: Where is the Valley of Kings?
A:

Sample response

The Valley of Kings is located in Luxor, Egypt.

Prompt

The following is a list of companies and the categories they fall into:

Apple, Facebook, Fedex

Apple
Category:

Sample response

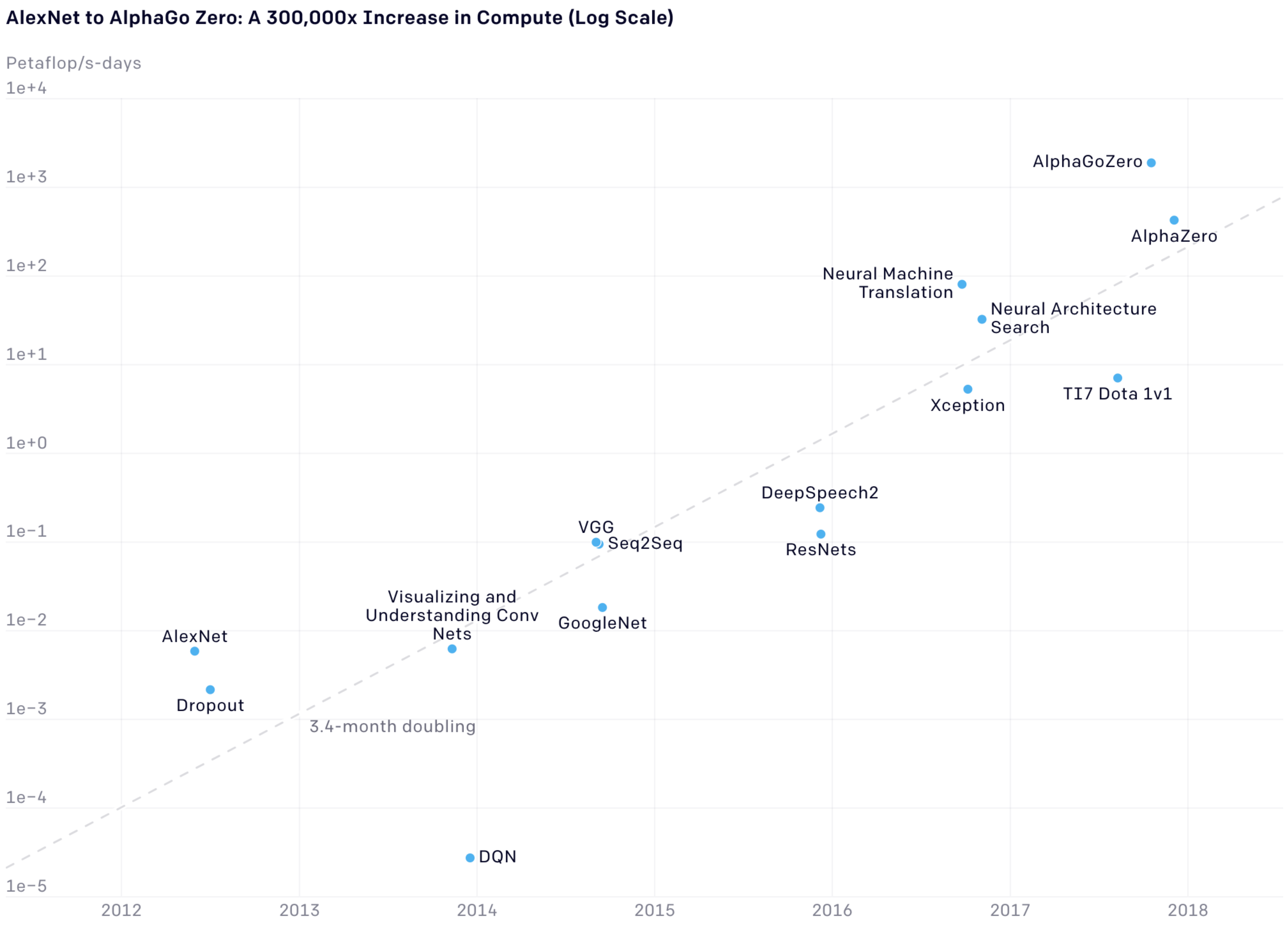
Technology

Facebook
Category: Social Media

Fedex
Category: Delivery

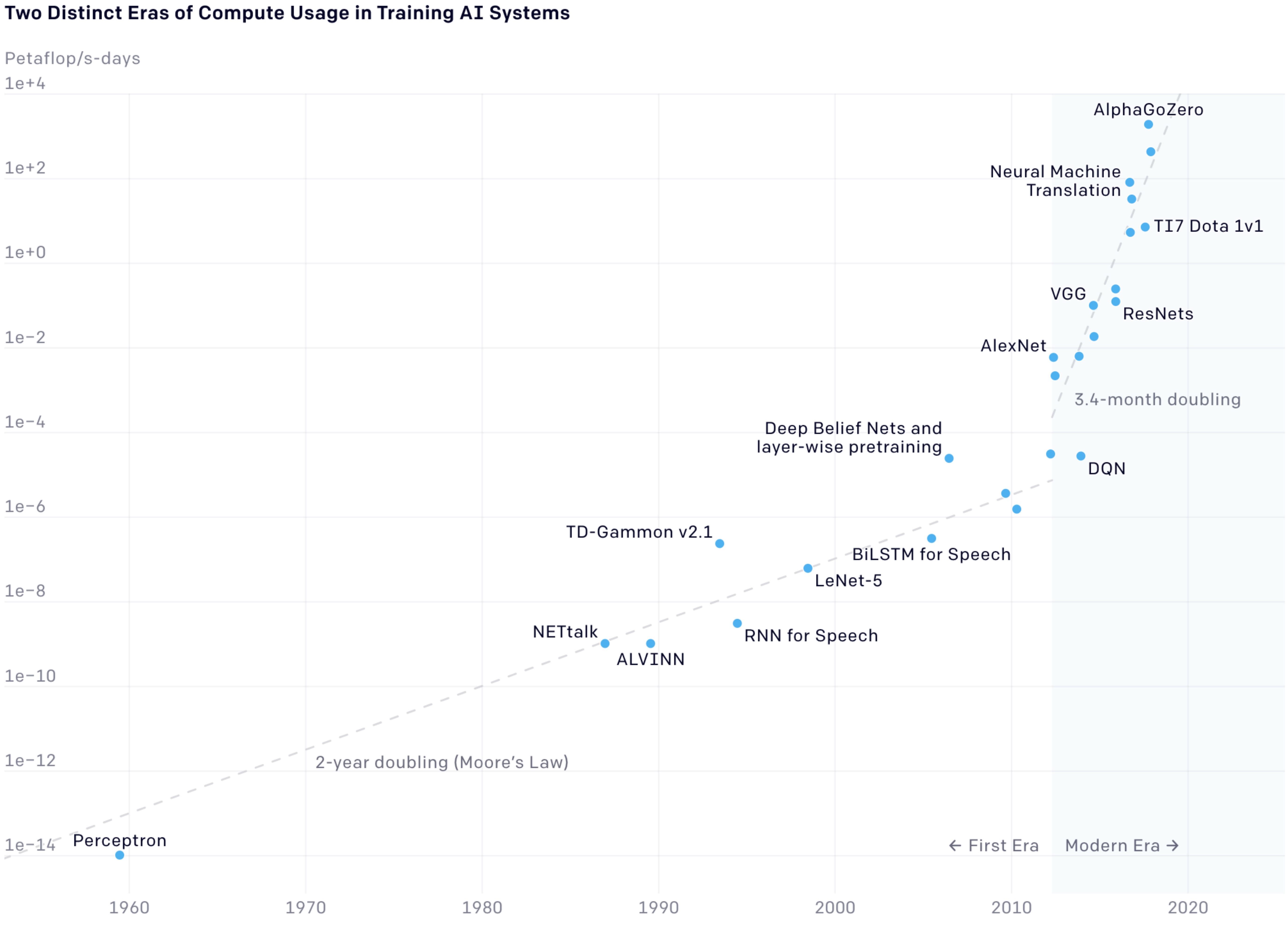
[https://beta.openai.com/examples]

The Era of Rapid Scaling



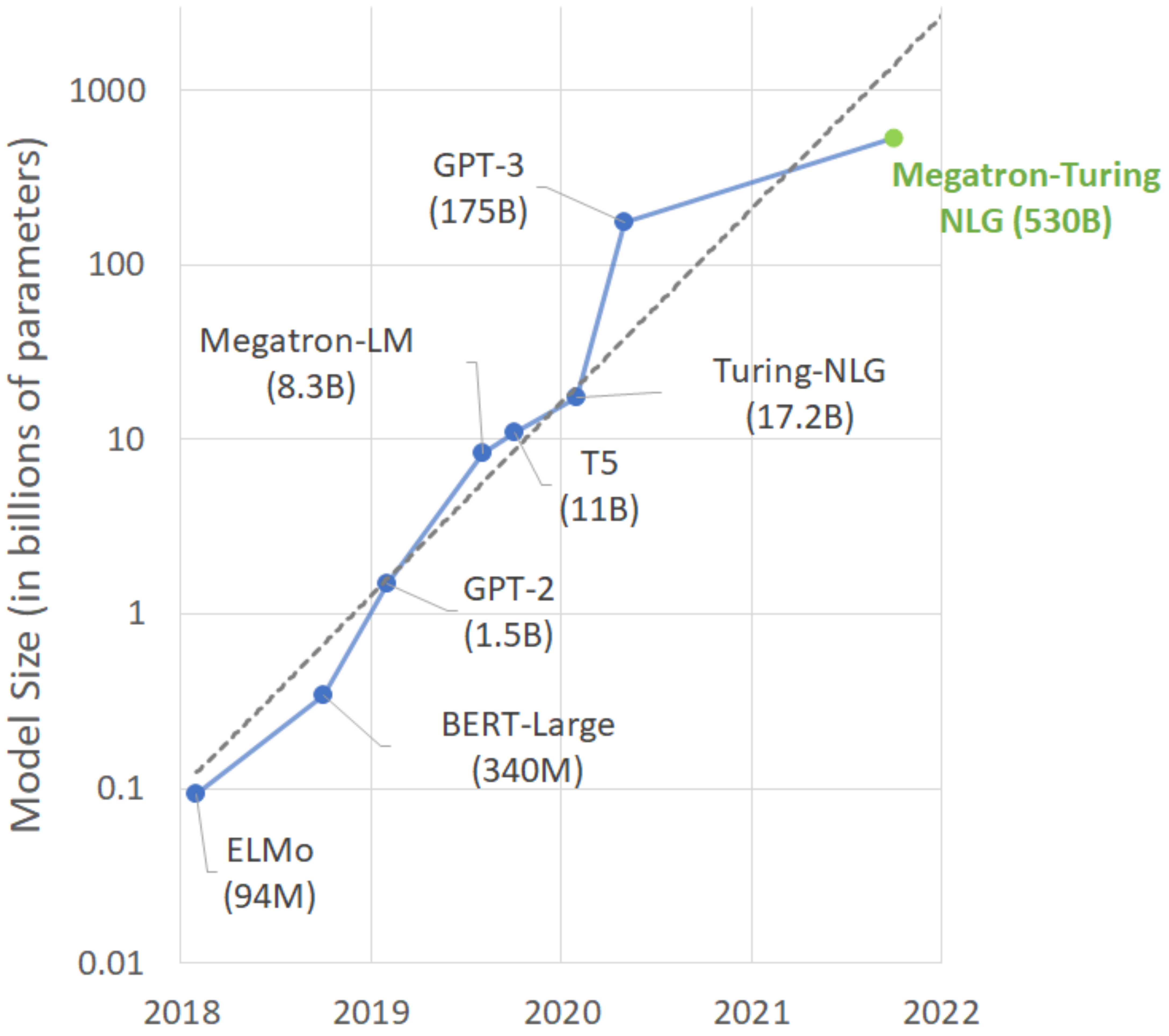
[Amodei, Hernandez, et al. / OpenAI]

The Era of Rapid Scaling



[Amodei, Hernandez, et al. / OpenAI]

The Era of Rapid Scaling in NLP



[NVIDIA]

ELMo is an RNN model; all others are Transformer models