

PAT 464/564 (Winter 2026)

Generative AI for Music & Audio Creation

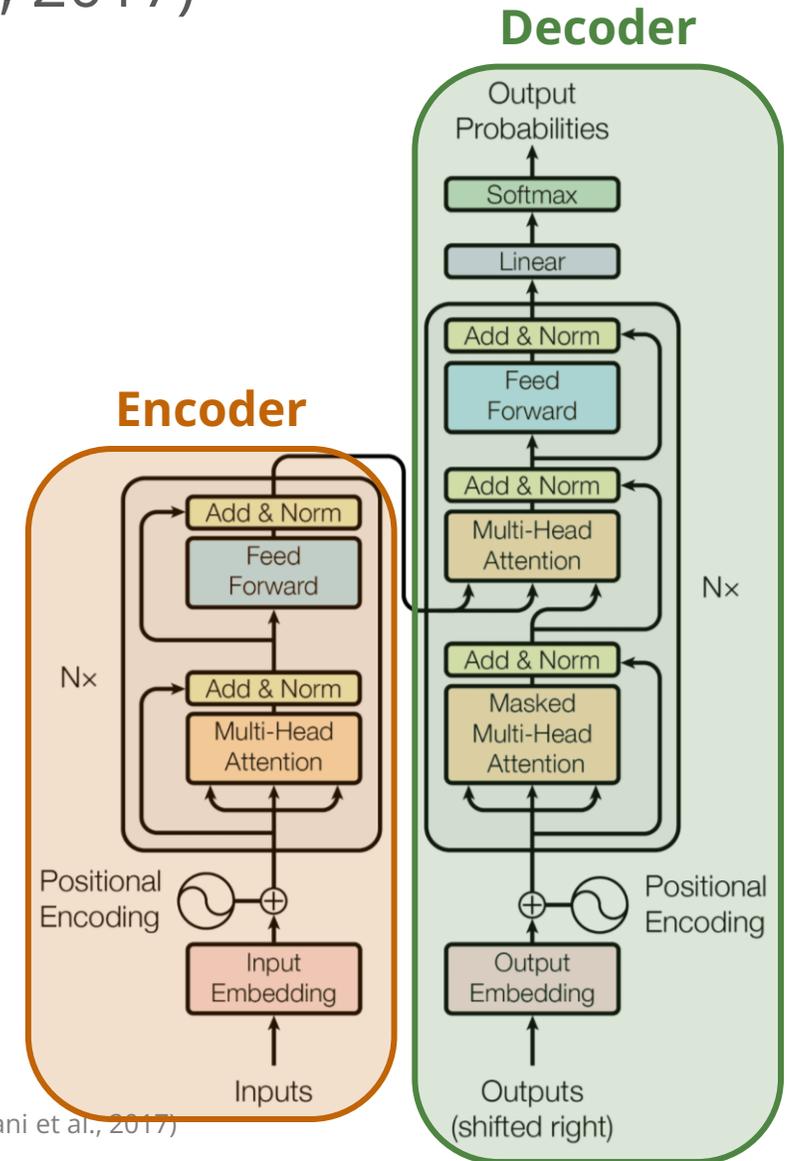
Lecture 11: Transformers II

Instructor: Hao-Wen Dong

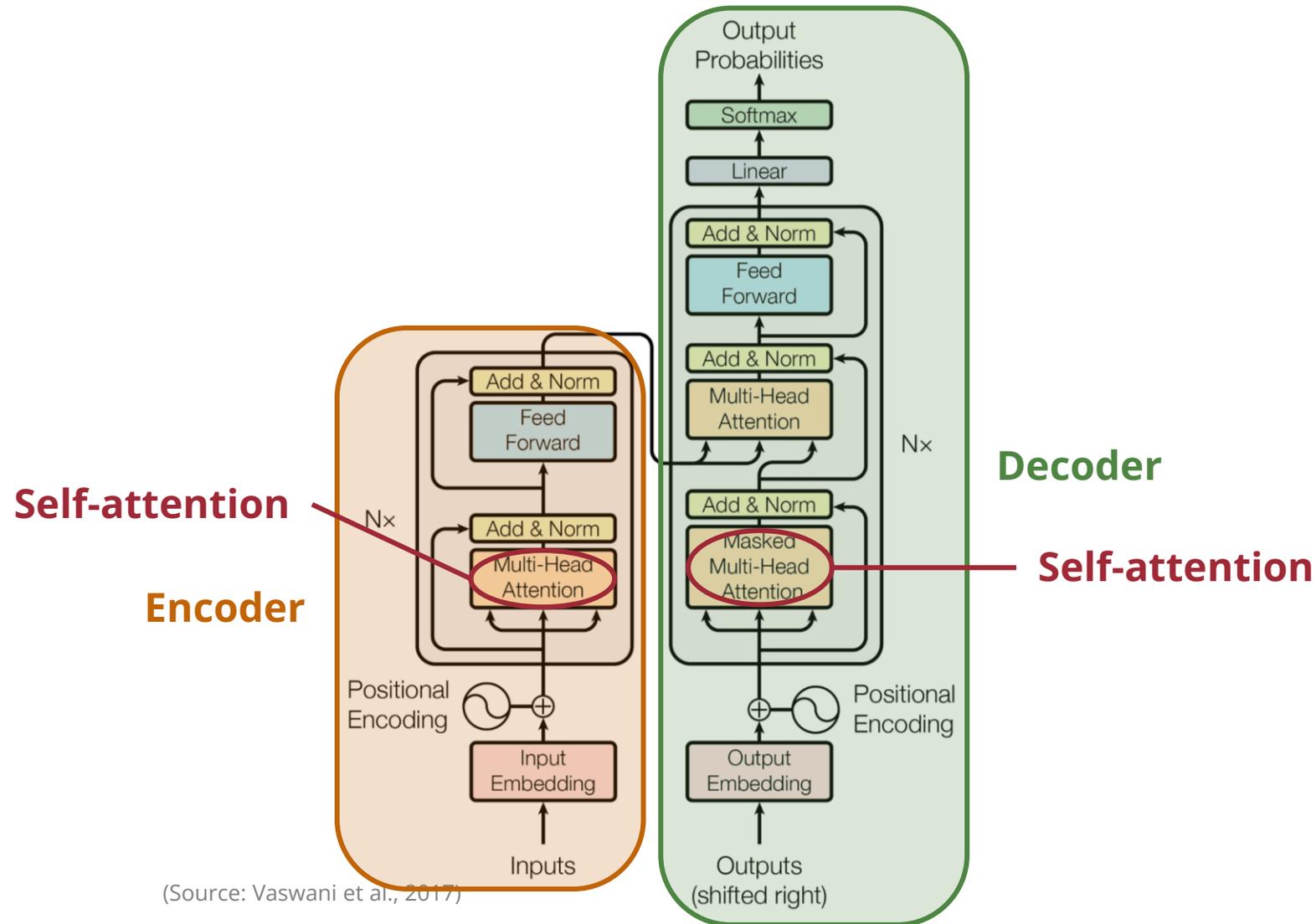
More on Transformers

The Original Transformer (Vaswani et al., 2017)

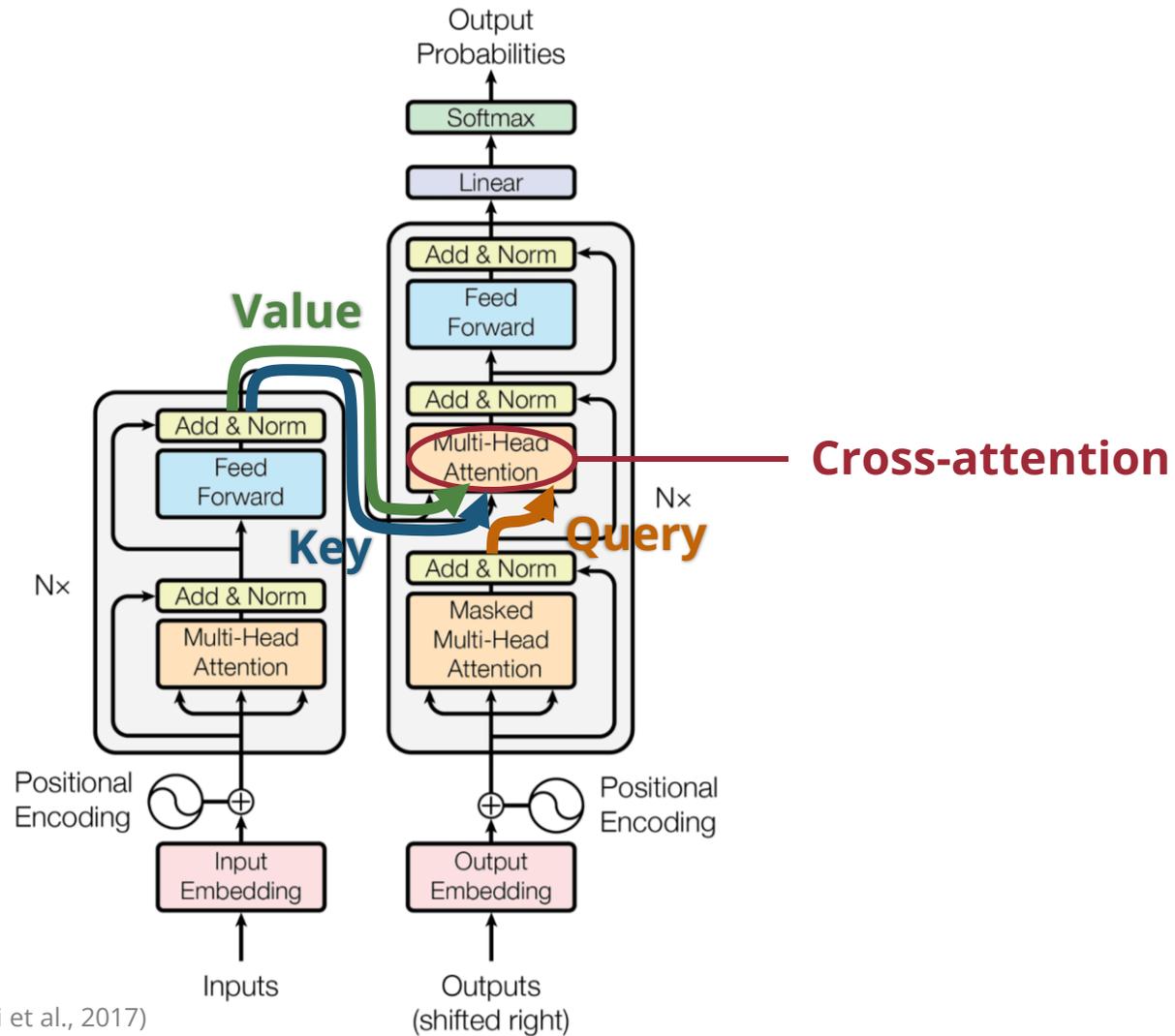
- The original transformer model was proposed for **machine translation**



The Original Transformer (Vaswani et al., 2017)



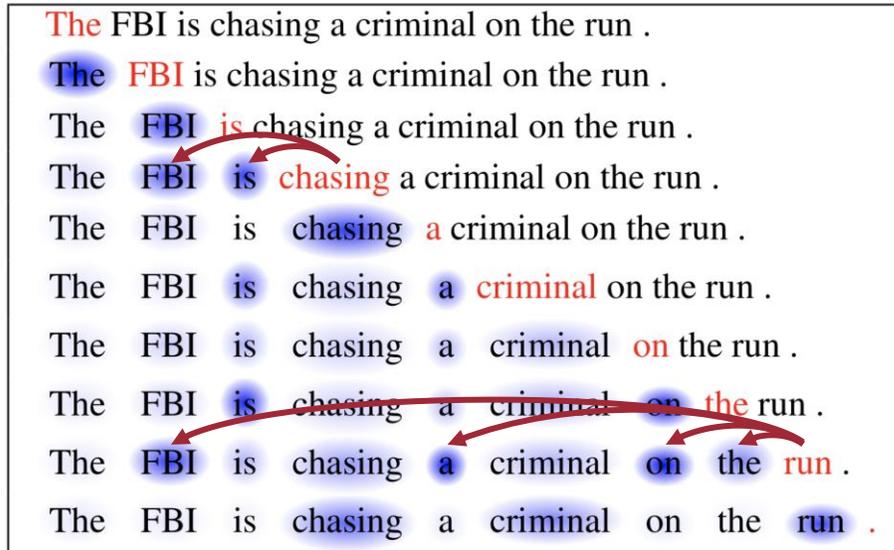
Cross-attention (Vaswani et al., 2017)



(Source: Vaswani et al., 2017)

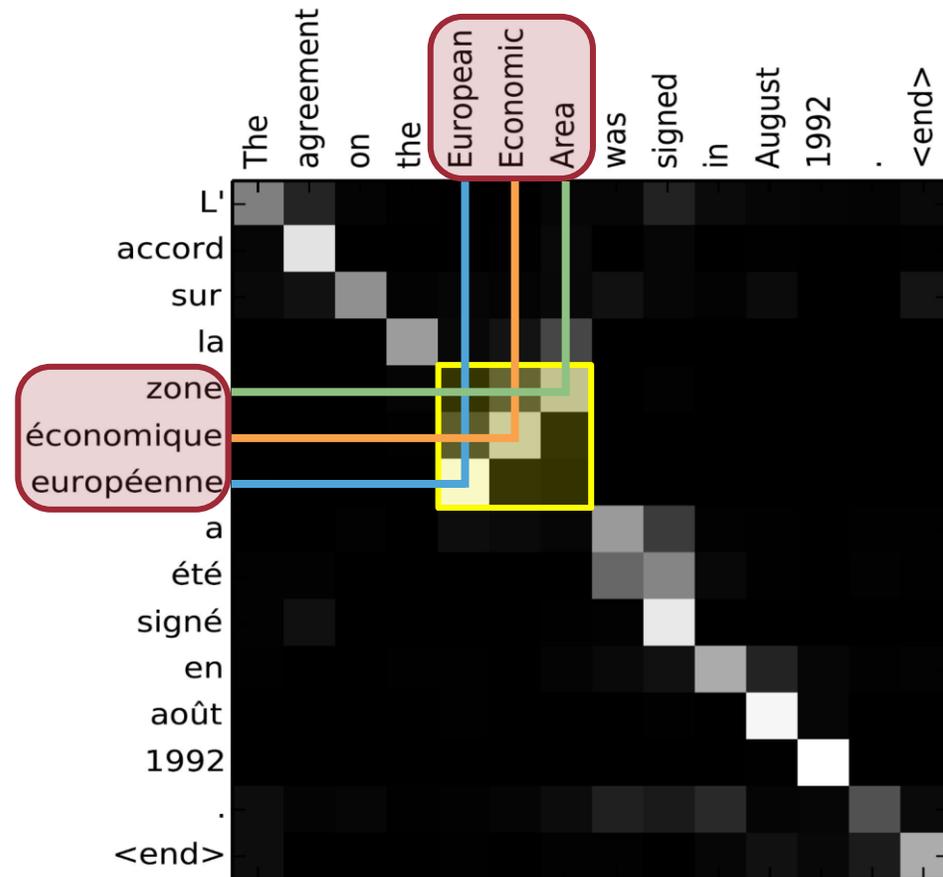
Self-Attention vs. Cross-Attention

Self-attention



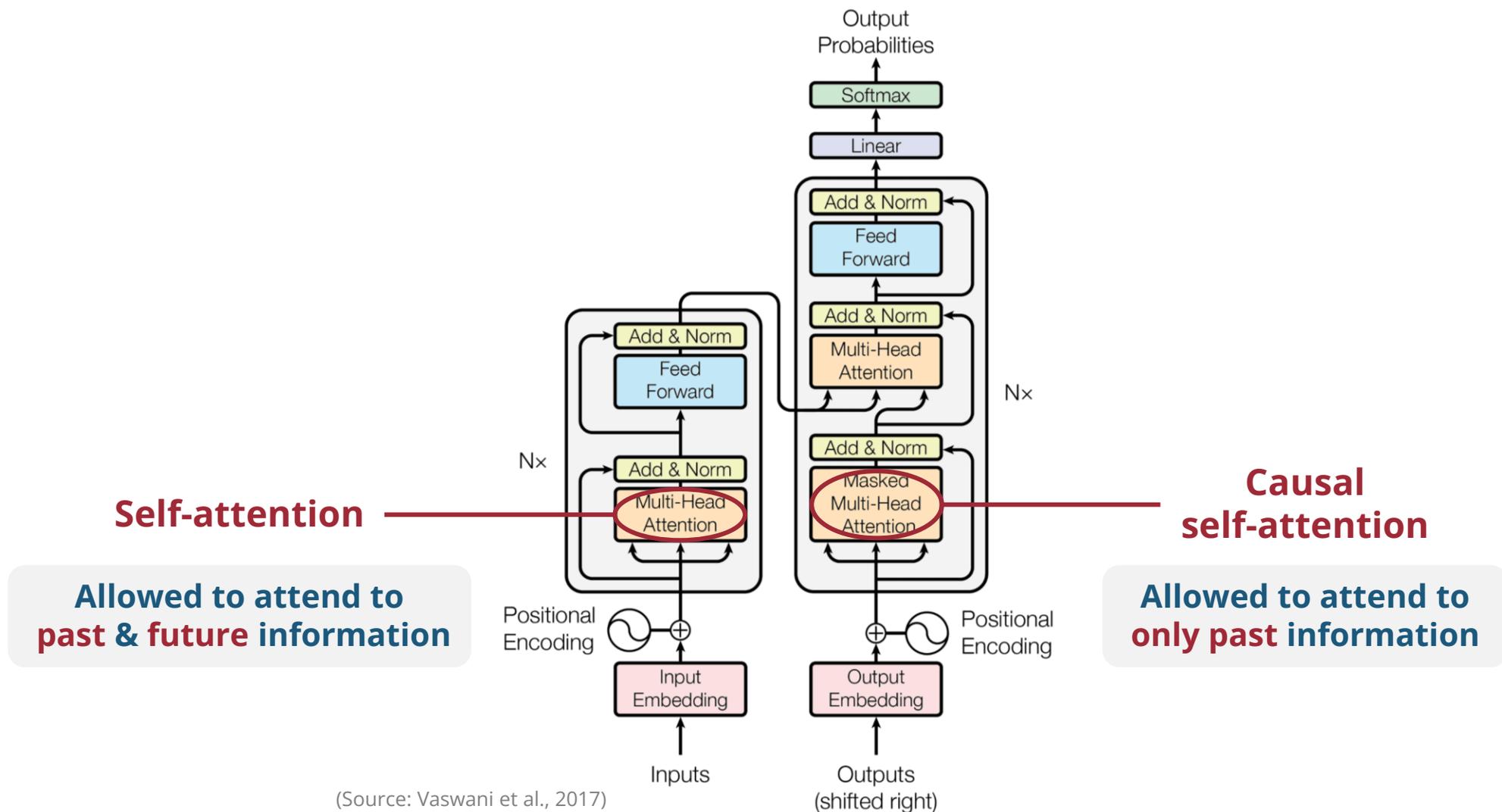
(Source: Cheng et al., 2016)

Cross-attention



(Source: Bahdanau et al., 2015)

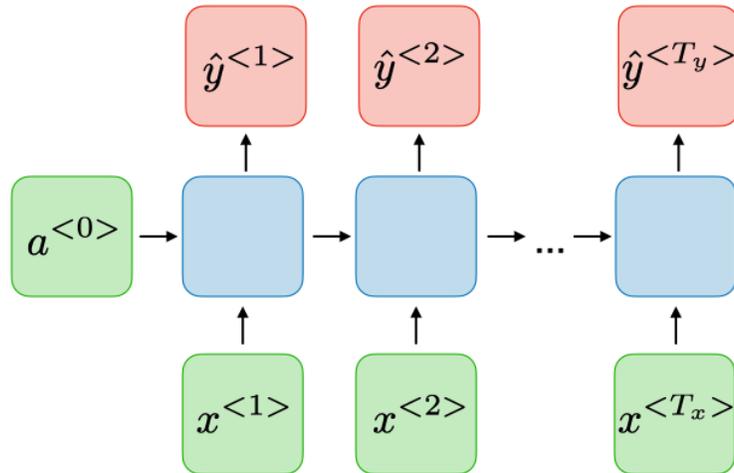
Transformer Encoder vs. Decoder (Vaswani et al., 2017)



(Source: Vaswani et al., 2017)

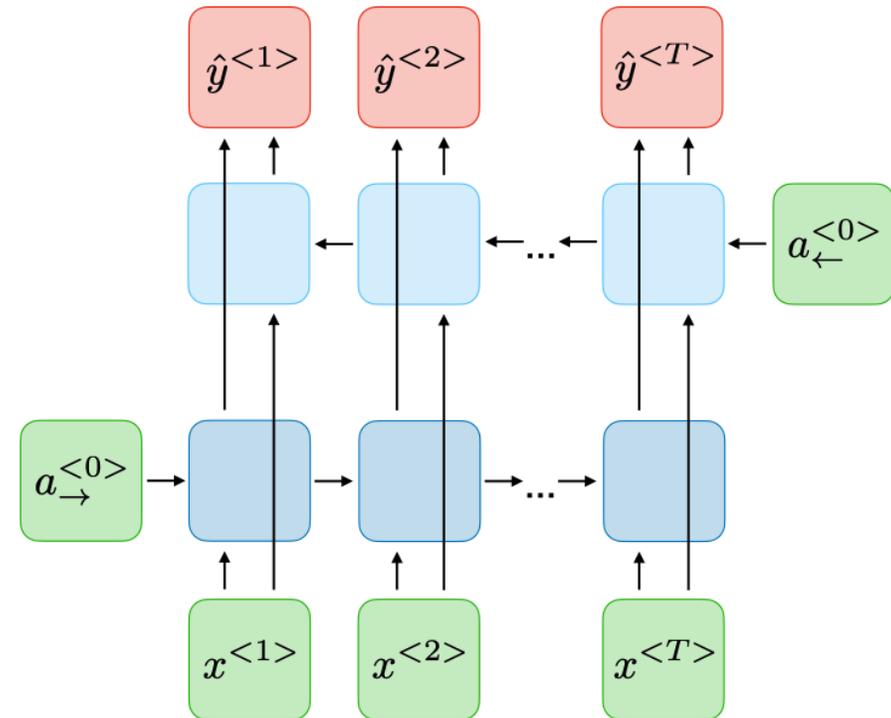
LSTMs vs. BiLSTMs

LSTMs



Access to **only past** information

BiLSTMs

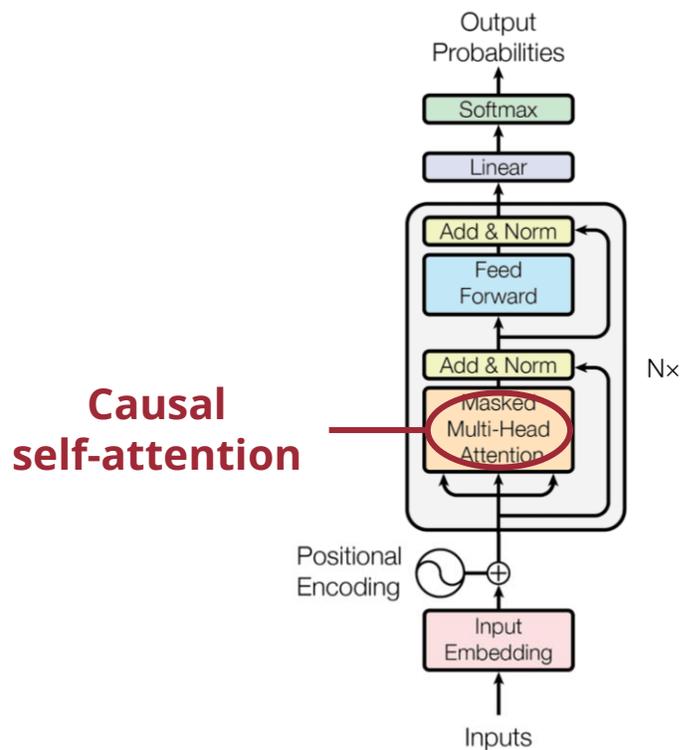


Access to **past & future** information

(Source: Amidi & Amidi, 2019)

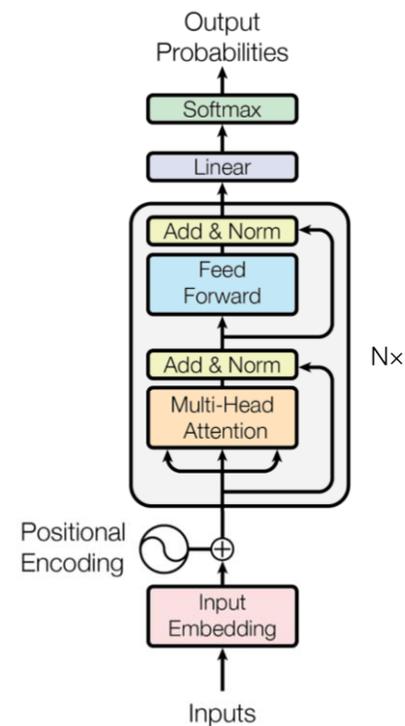
Decoder-only vs. Encoder-only Transformer

Decoder-only transformer



Access to **only past** information

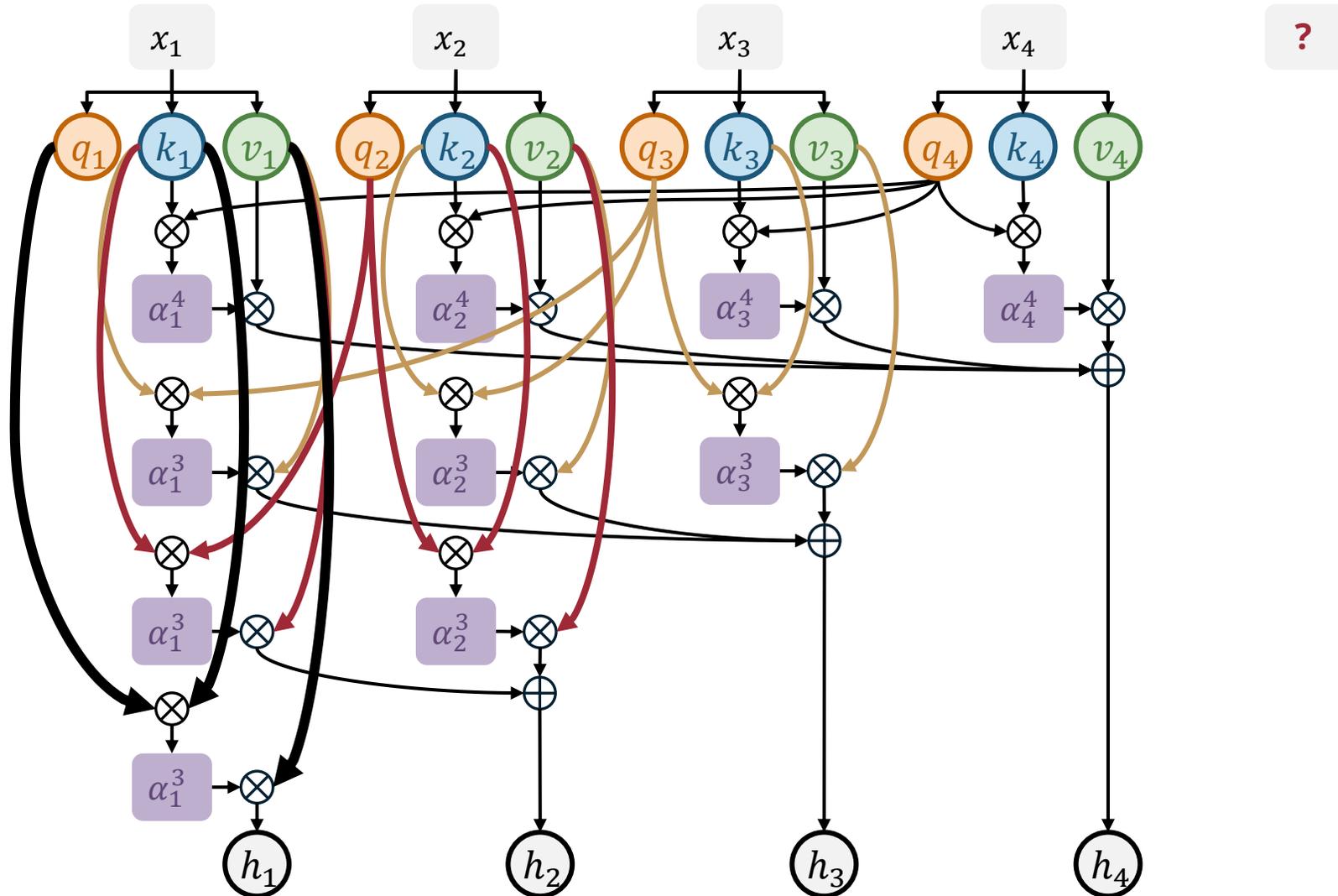
Encoder-only transformer



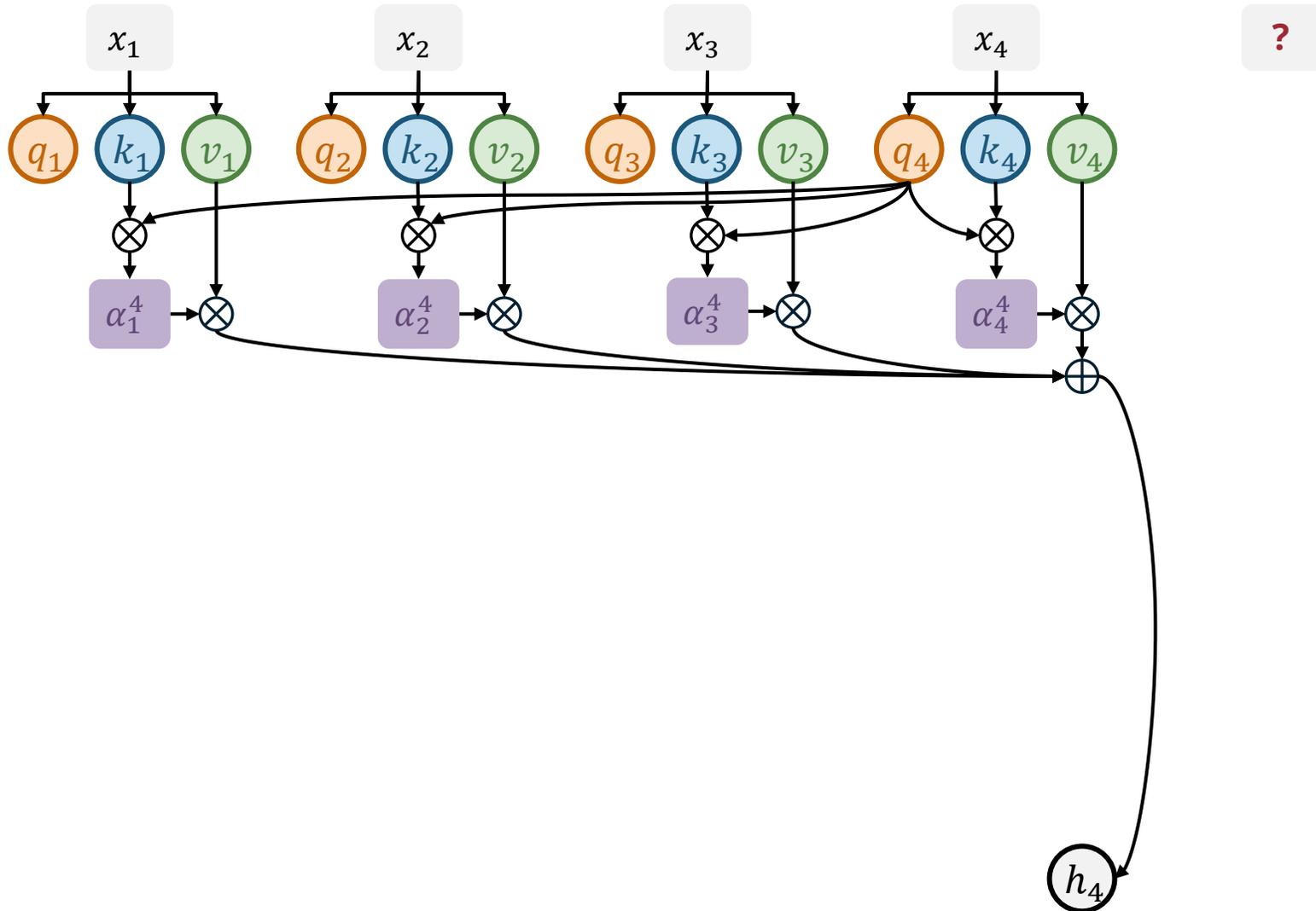
Access to **past & future** information

(Source: Vaswani et al., 2017)

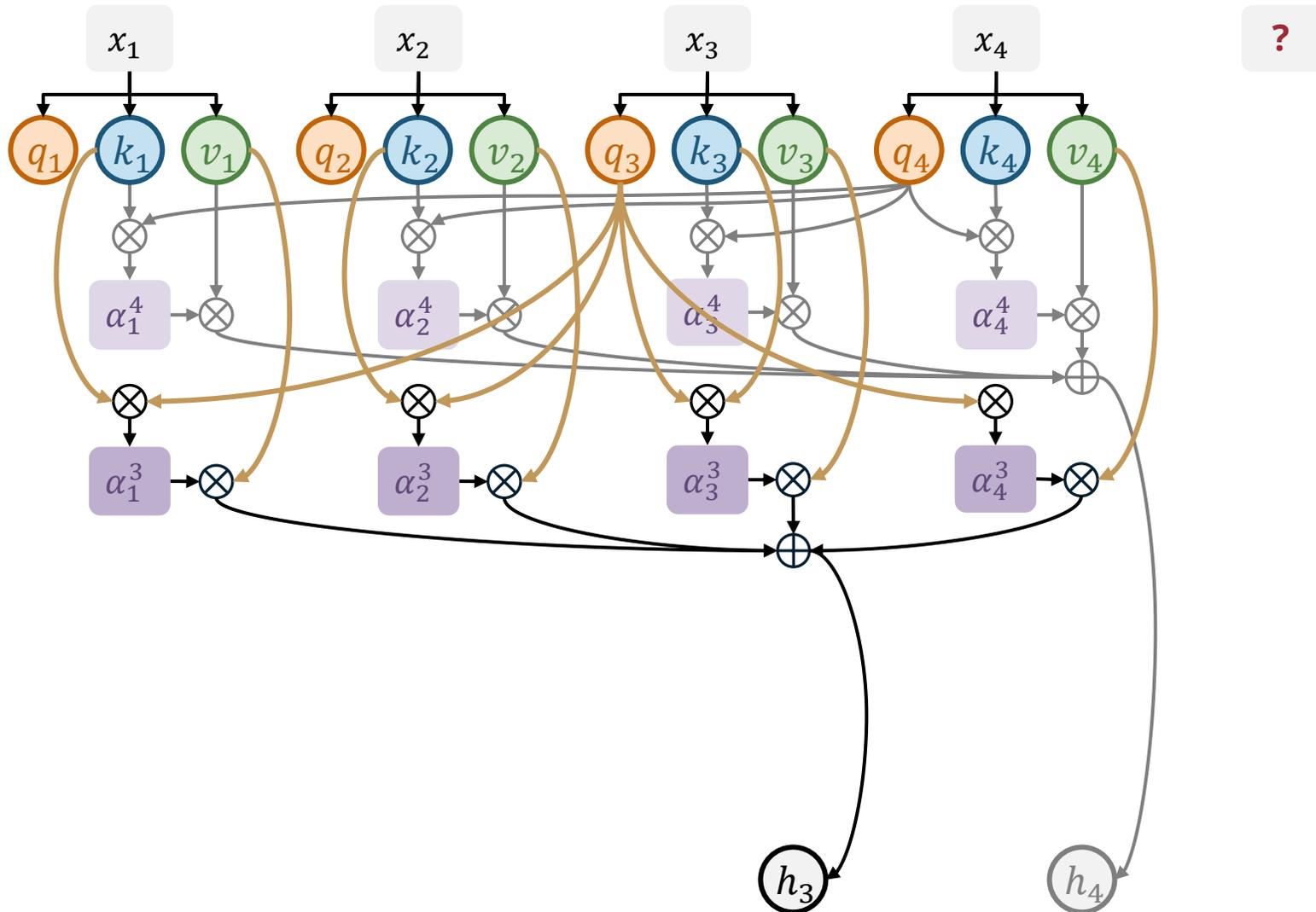
Decoder-only Transformers (Vaswani et al., 2017)



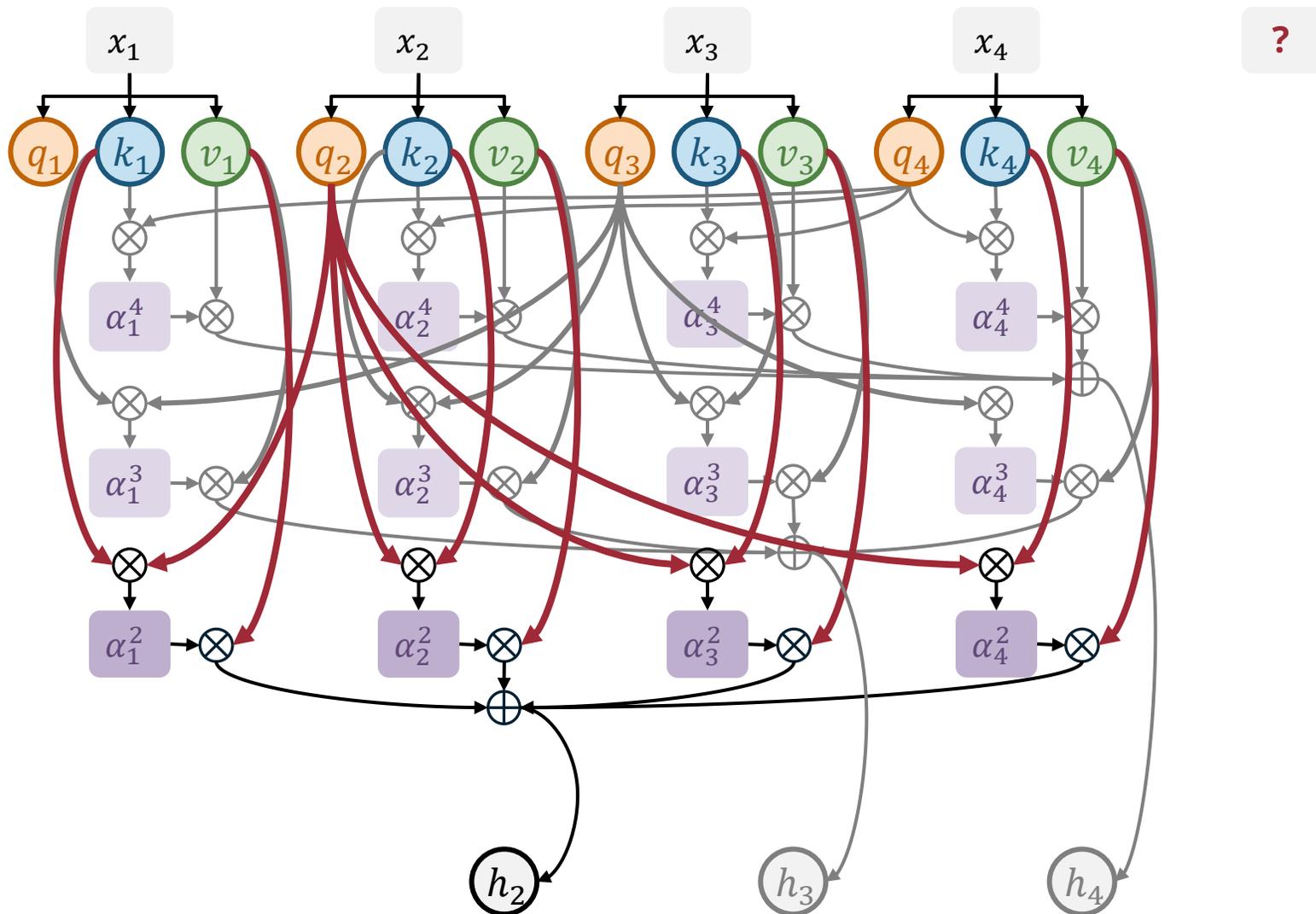
Encoder-only Transformer (Vaswani et al., 2017)



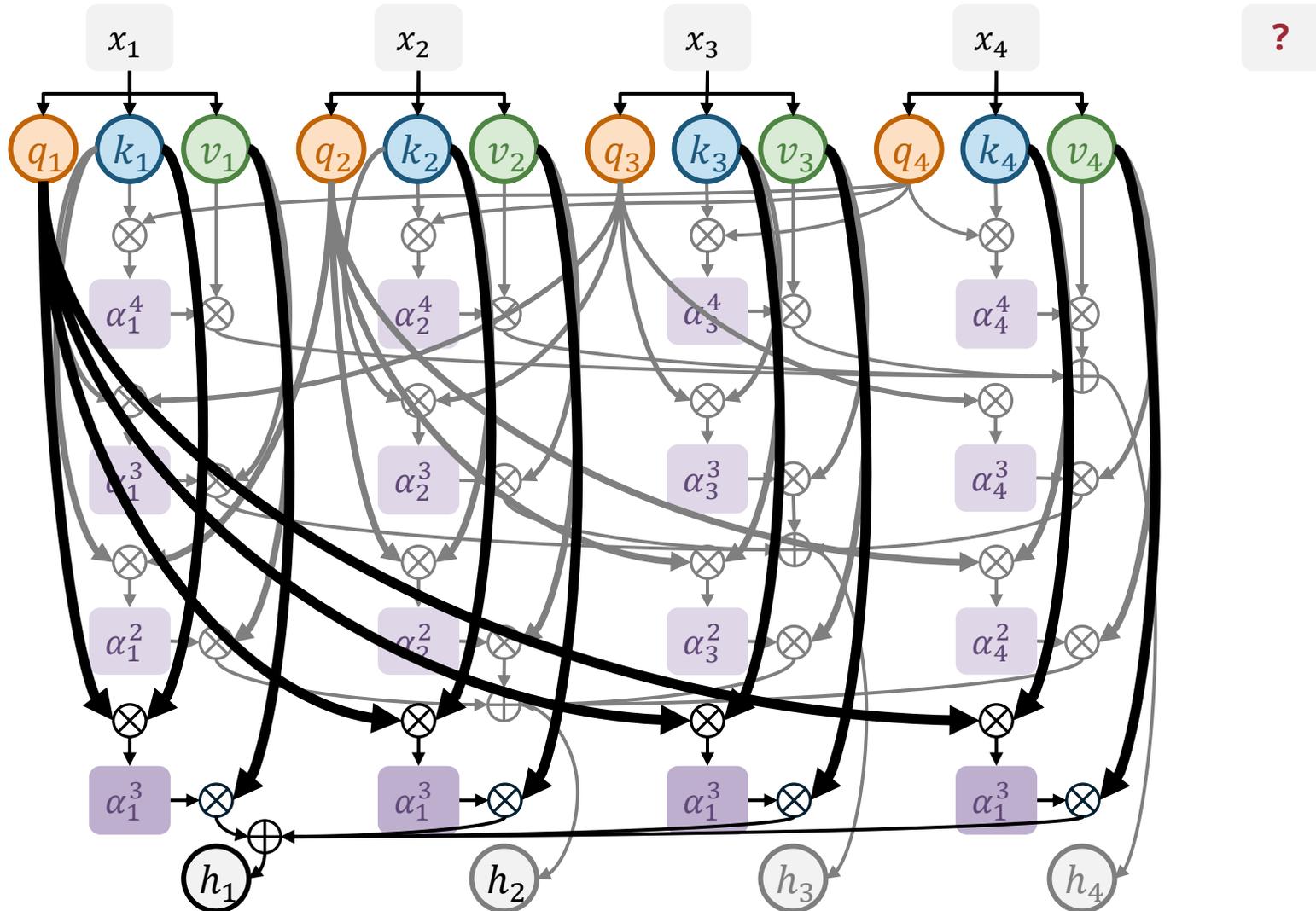
Encoder-only Transformer (Vaswani et al., 2017)



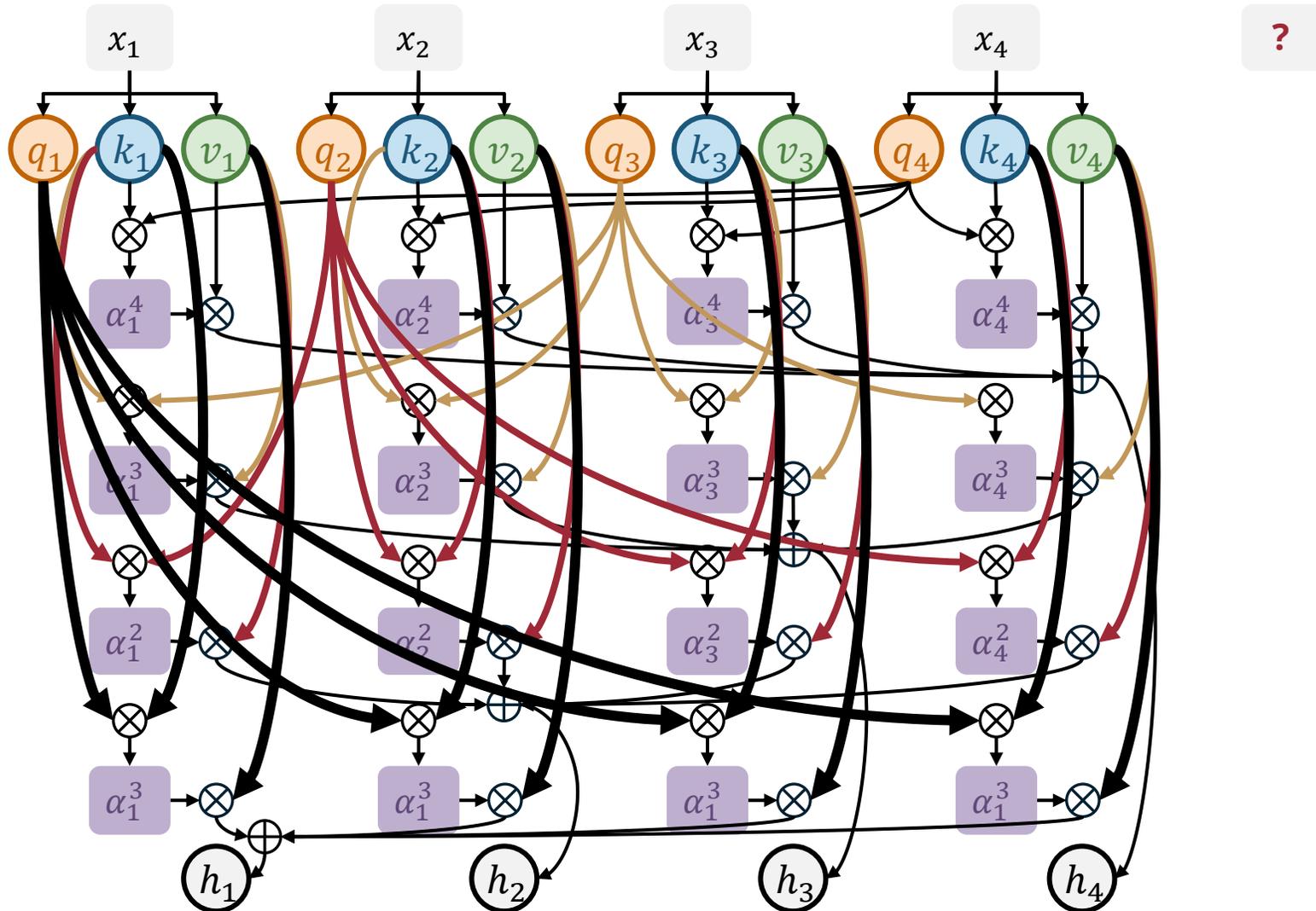
Encoder-only Transformer (Vaswani et al., 2017)



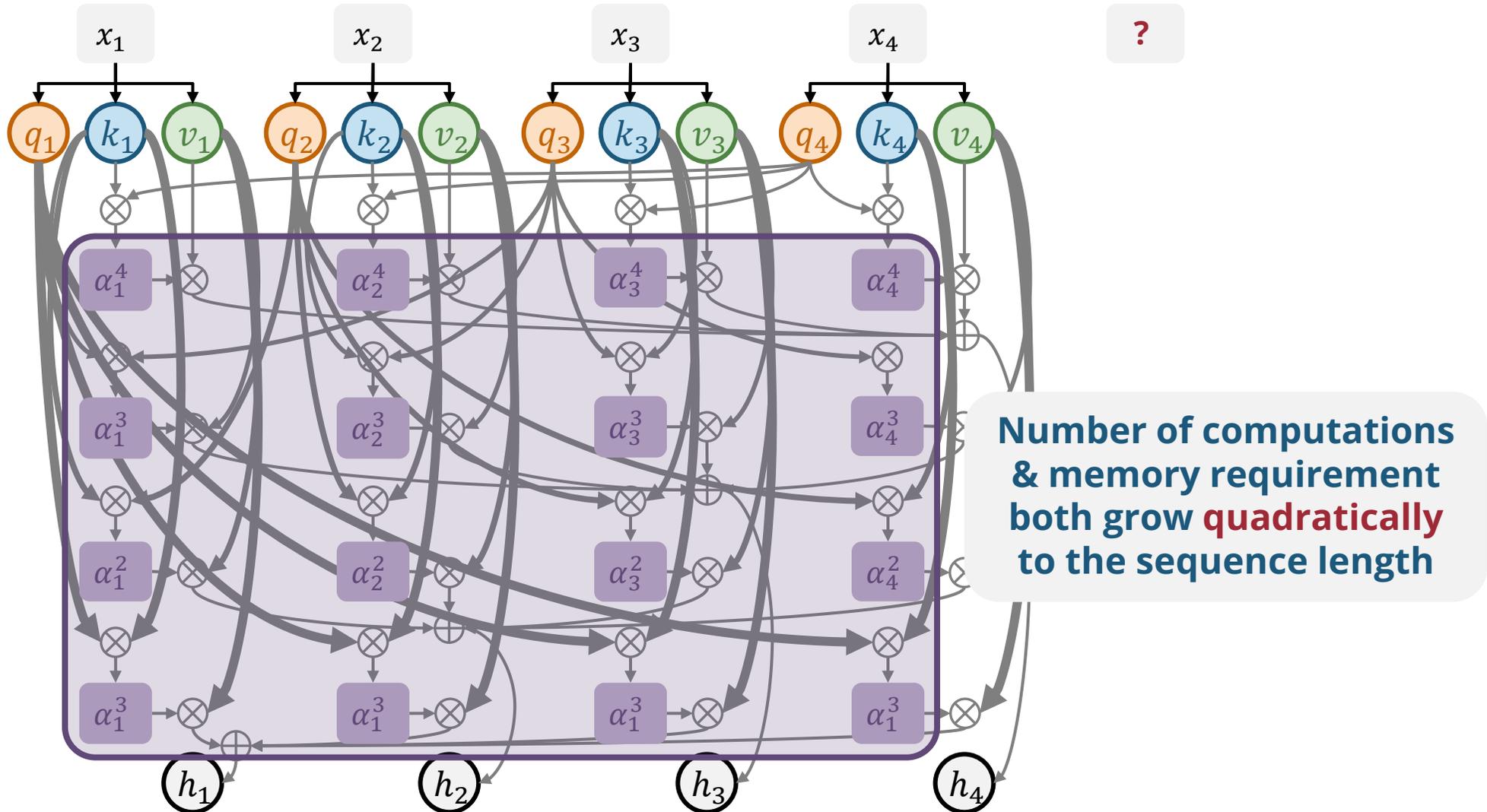
Encoder-only Transformer (Vaswani et al., 2017)



Encoder-only Transformer (Vaswani et al., 2017)

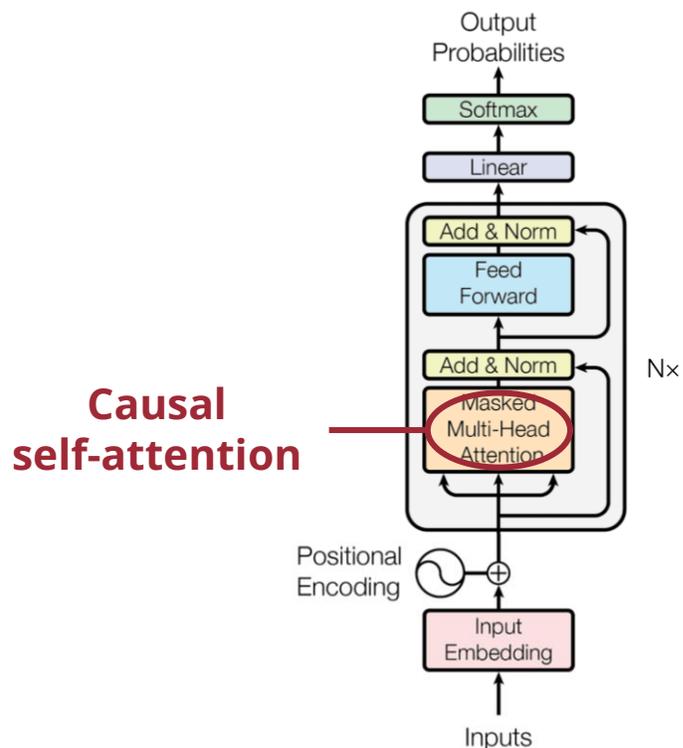


Encoder-only Transformer (Vaswani et al., 2017)



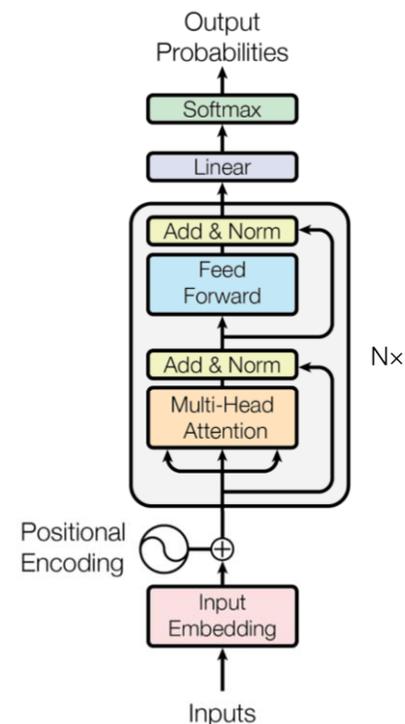
Decoder-only vs. Encoder-only Transformer

Decoder-only transformer



Access to **only past** information

Encoder-only transformer

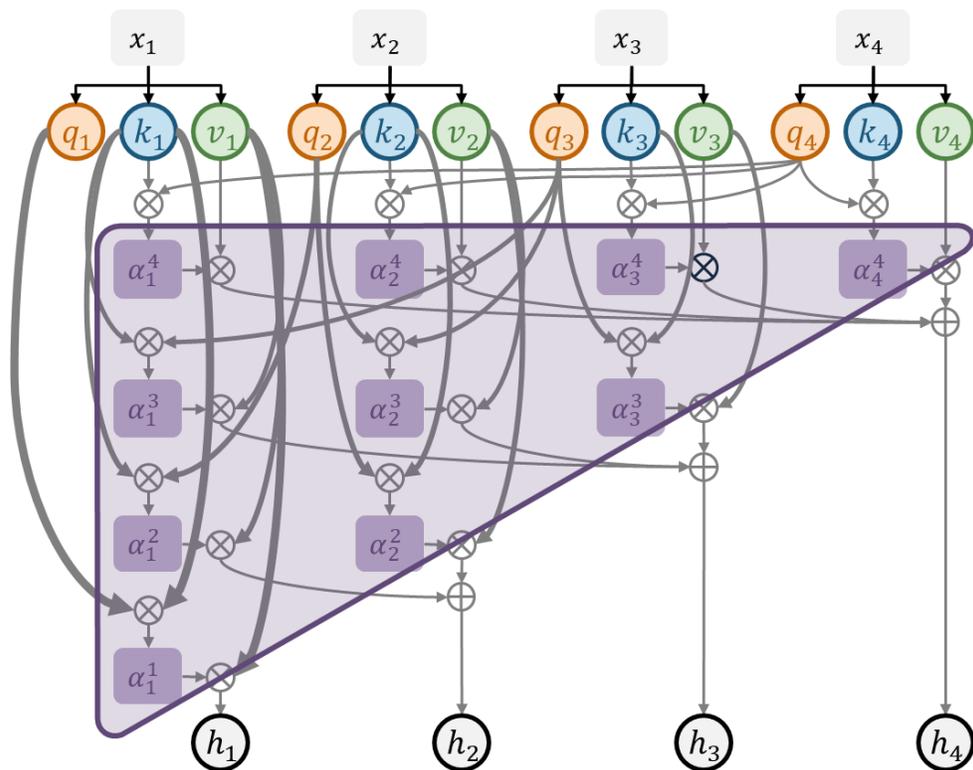


Access to **past & future** information

(Source: Vaswani et al., 2017)

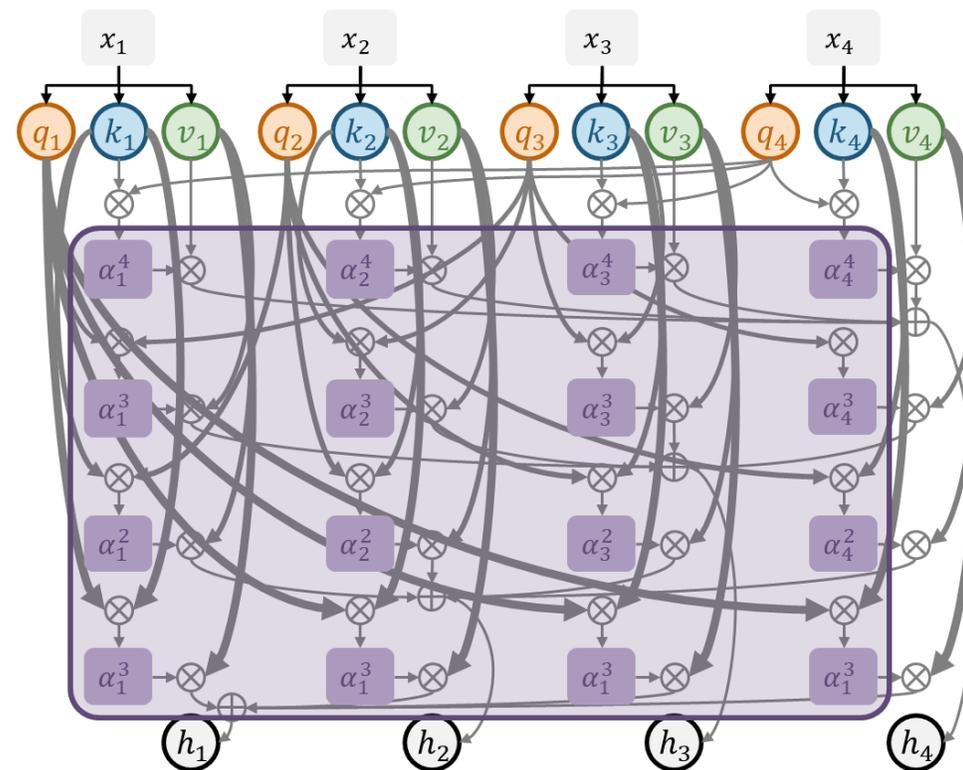
Decoder-only vs. Encoder-only Transformer

Decoder-only transformer



Access to **only past** information

Encoder-only transformer



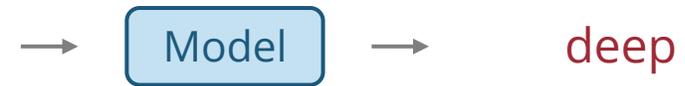
Access to **past & future** information

Decoding Strategies

Deep Autoregressive Models

- **Intuition:** Decompose the generation of a sequence into generating one item after another

A transformer is a



A transformer is a deep



A transformer is a deep learning



A transformer is a deep learning model



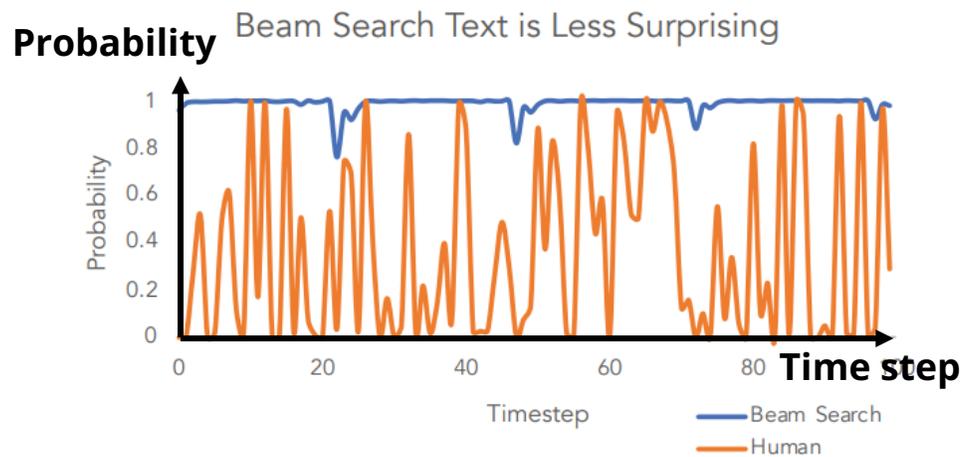
A transformer is a deep learning model introduced



A transformer is a deep learning model introduced in



🤔 Do We Really Want the Most Probable Sequence?



(Source: Holtzman et al., 2020)

Beam Search

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

Human

...which grant increased life span and three years warranty. The Antec HCG series consists of five models with capacities spanning from 400W to 900W. Here we should note that we have already tested the HCG-620 in a previous review and were quite satisfied With its performance. In today's review we will rigorously test the Antec HCG-520, which as its model number implies, has 520W capacity and contrary to Antec's strong beliefs in multi-rail PSUs is equipped...

Deep Autoregressive Models

- **Intuition:** Decompose the generation of a sequence into generating one item after another

$$P(x_i \mid x_1, x_2, \dots, x_{i-1})$$

Next word Previous words

$P(\text{electrical} \mid \text{A transformer is a})$ ↑

$P(\text{character} \mid \text{A transformer is a})$ ↑

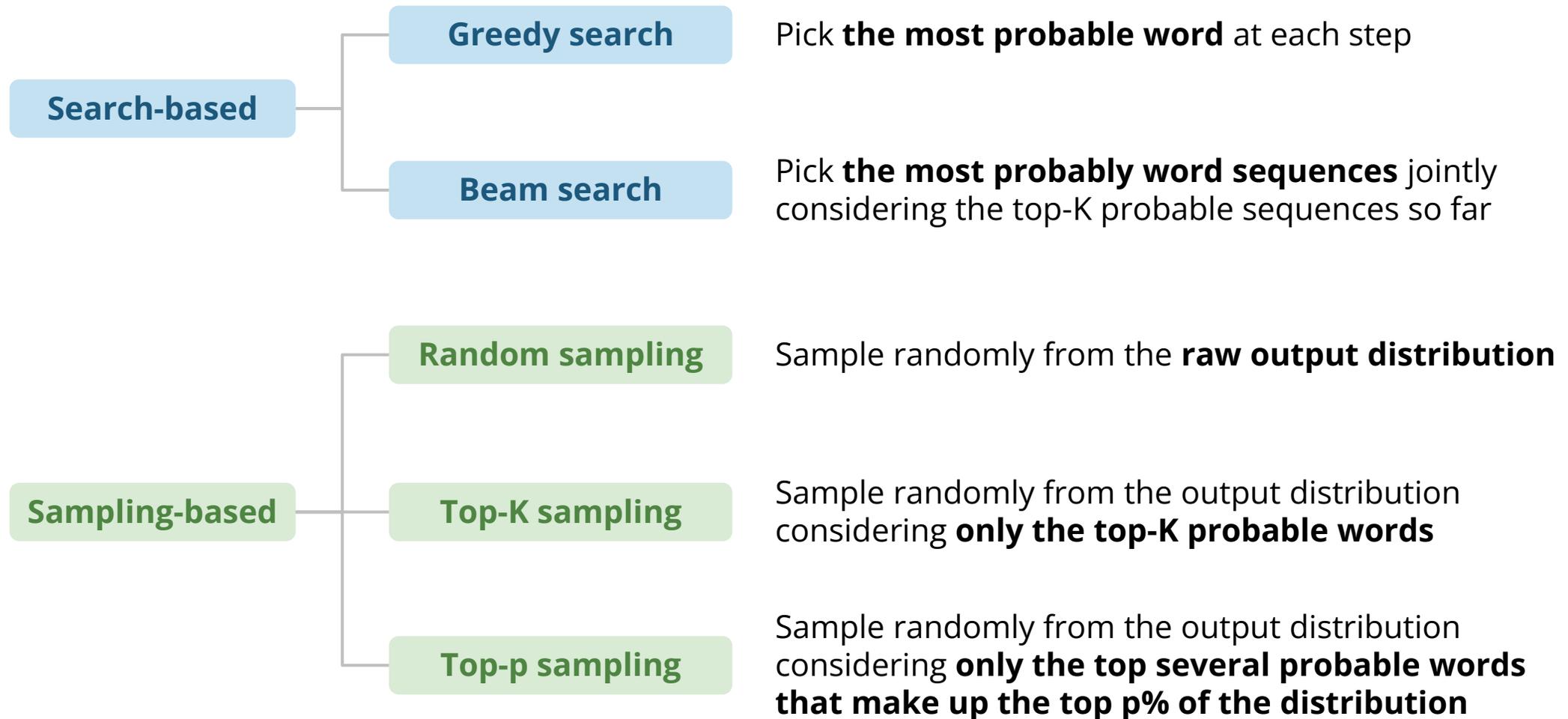
$P(\text{gene} \mid \text{A transformer is a})$ ↑

$P(\text{model} \mid \text{A transformer is a})$ ↑

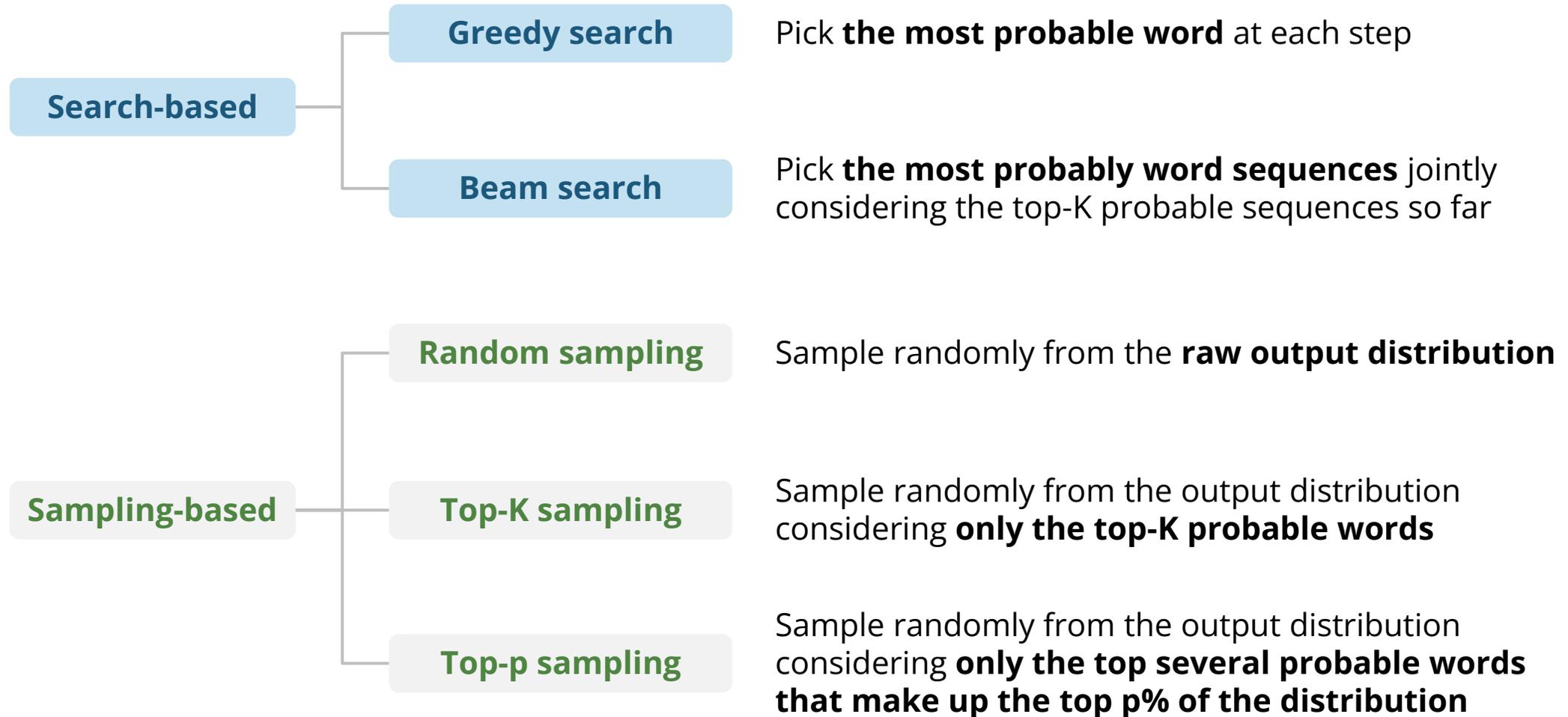
$P(\text{food} \mid \text{A transformer is a})$ ↓

$P(\text{musical} \mid \text{A transformer is a})$ ↓

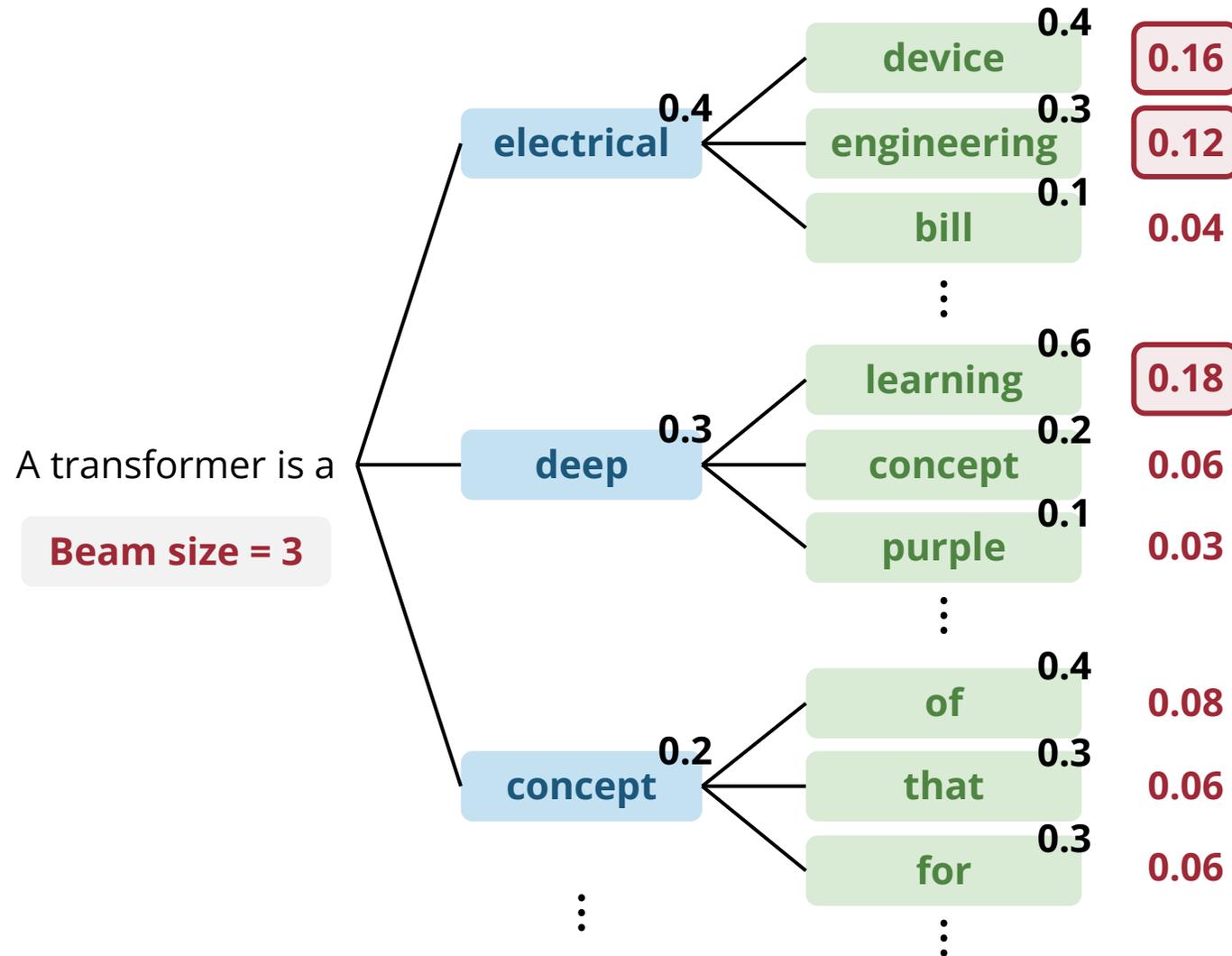
Decoding Strategies



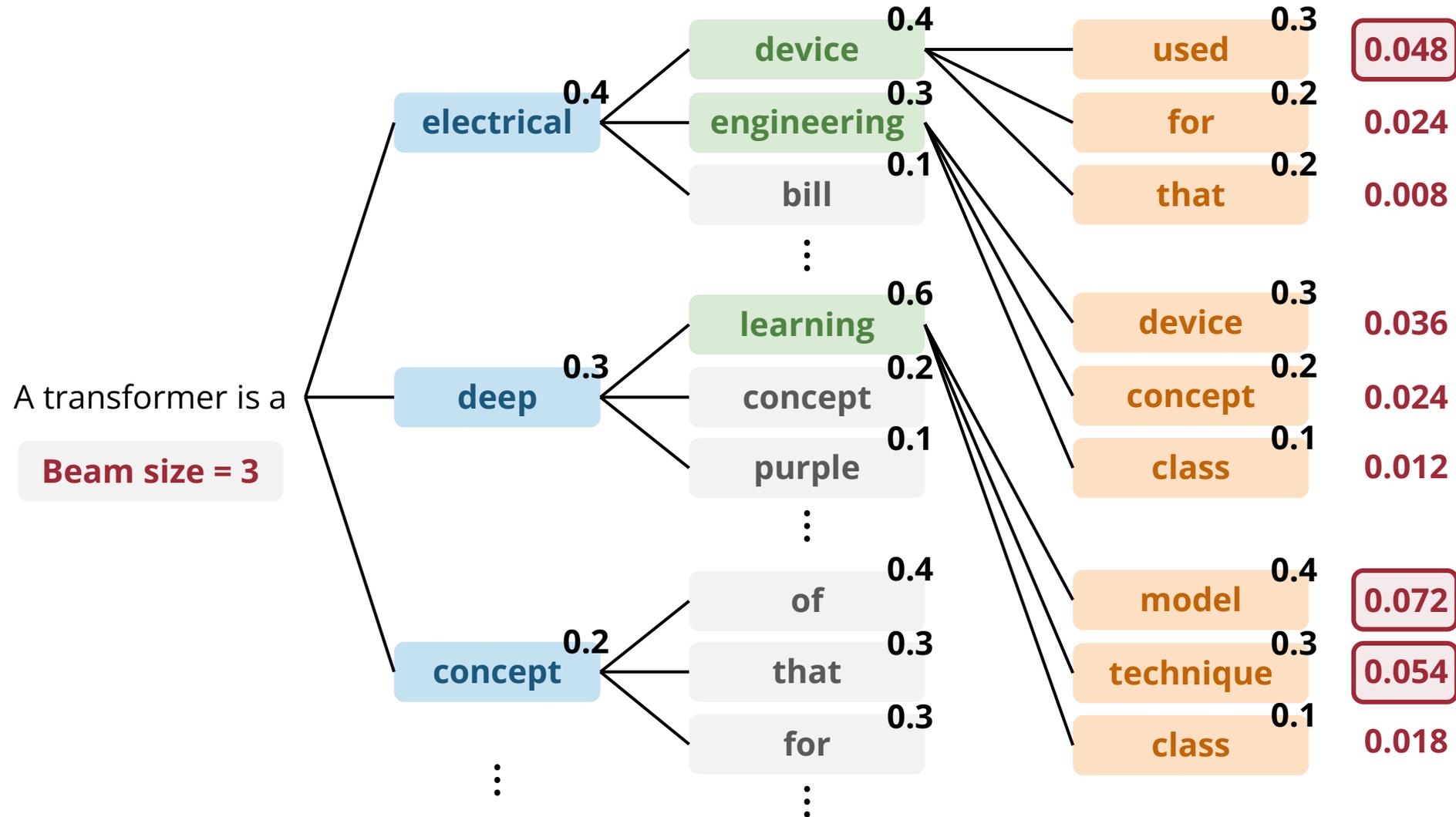
Decoding Strategies



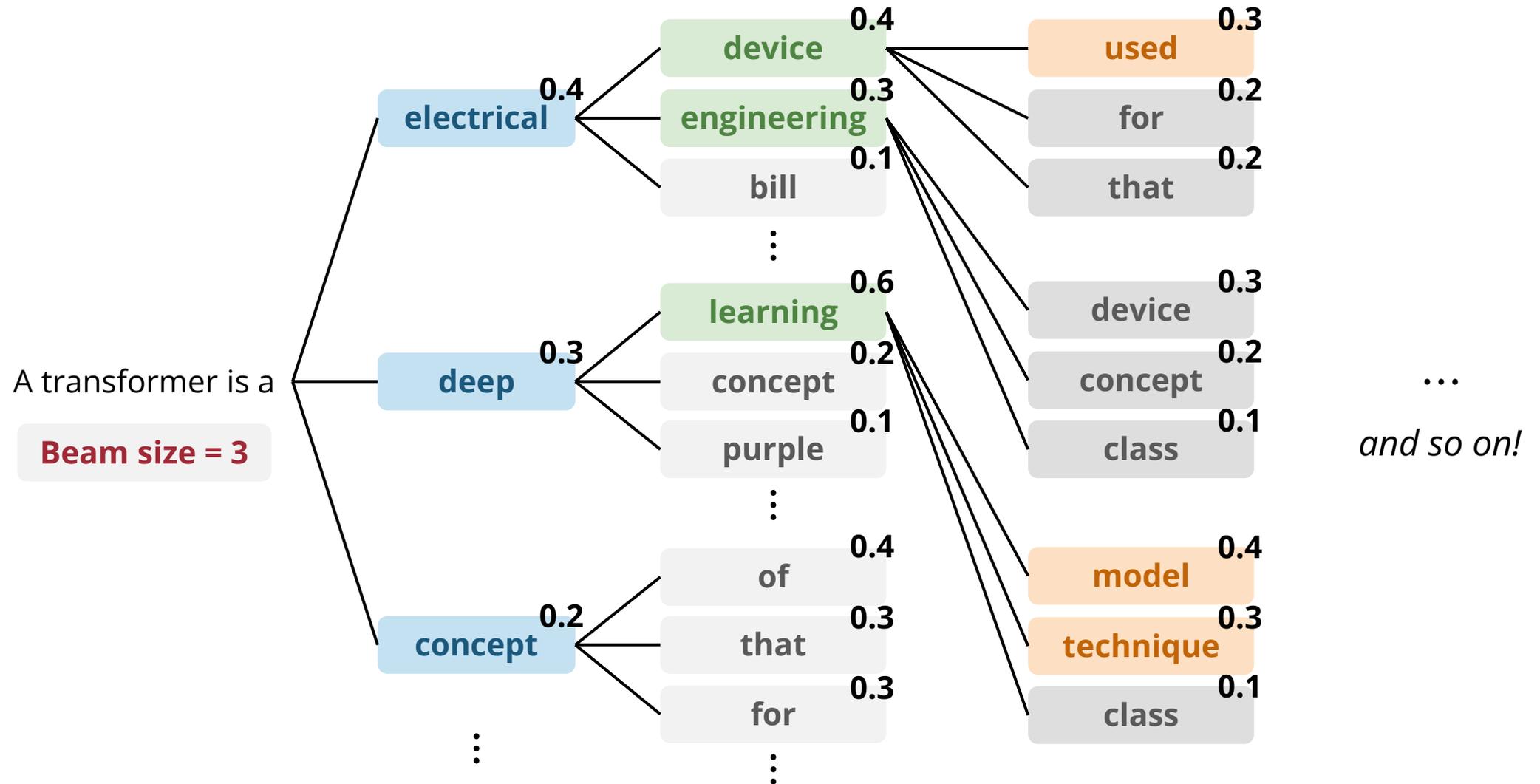
Beam Search



Beam Search

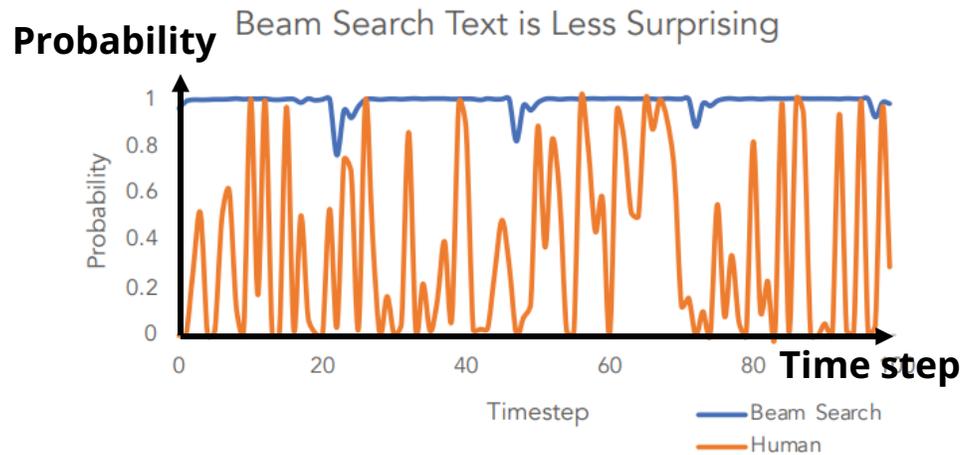


Beam Search





Do We Really Want the Most Probable Sequence?



(Source: Holtzman et al., 2020)

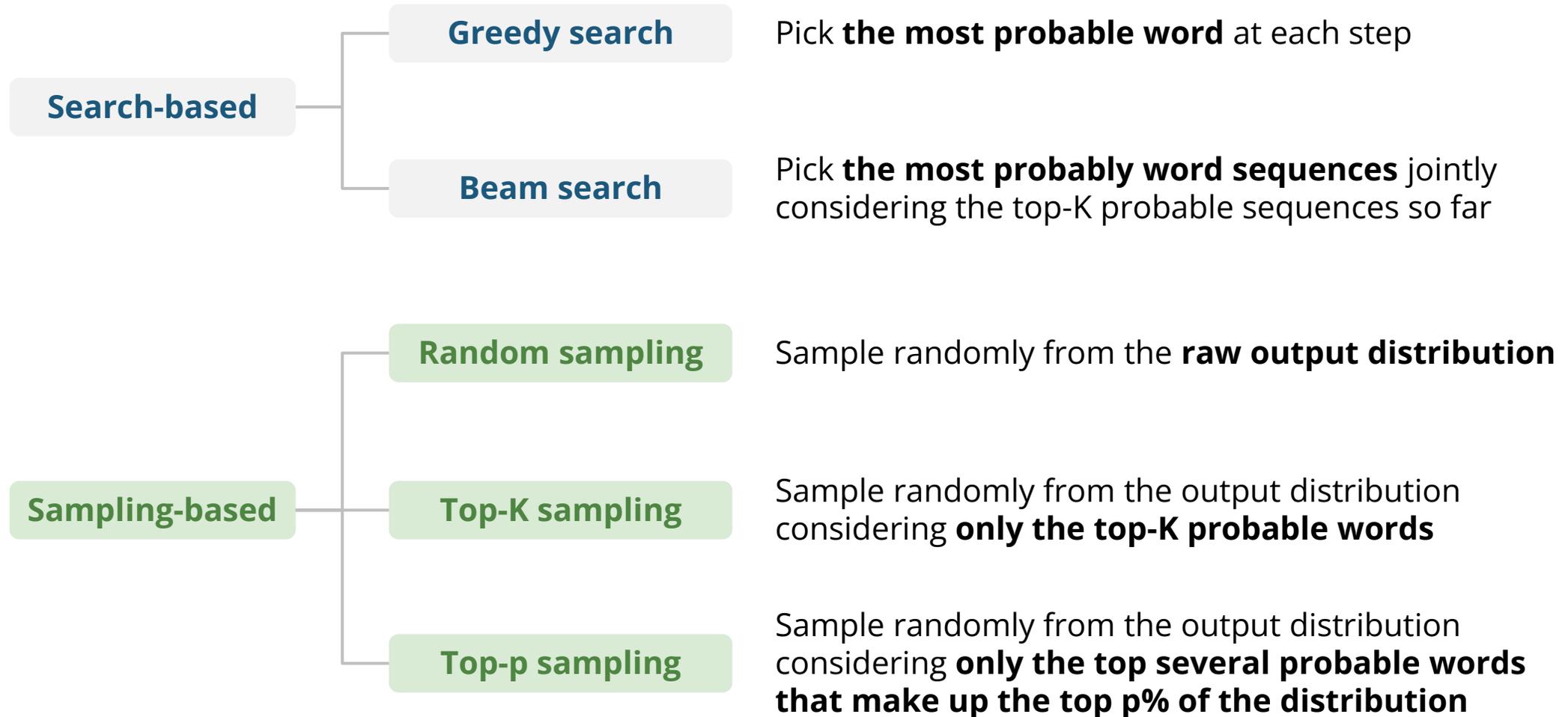
Beam Search

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

Human

...which grant increased life span and three years warranty. The Antec HCG series consists of five models with capacities spanning from 400W to 900W. Here we should note that we have already tested the HCG-620 in a previous review and were quite satisfied With its performance. In today's review we will rigorously test the Antec HCG-520, which as its model number implies, has 520W capacity and contrary to Antec's strong beliefs in multi-rail PSUs is equipped...

Decoding Strategies



Temperature

Softmax

$$\hat{y}_i = \frac{e^{\tilde{y}_i}}{\sum_{j=1}^n e^{\tilde{y}_j}}$$

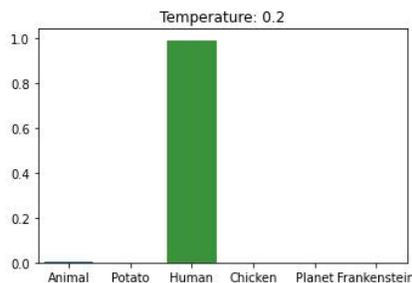


$$\hat{y}_i = \frac{e^{\tilde{y}_i/\tau}}{\sum_{j=1}^n e^{\tilde{y}_j/\tau}}$$

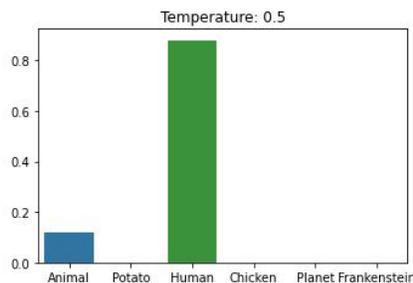
Temperature

Original

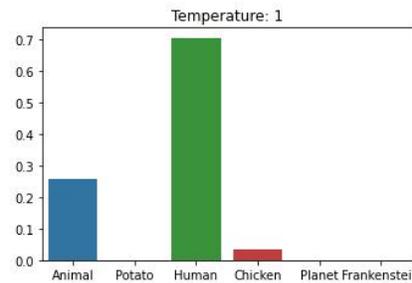
$\tau = 0.2$



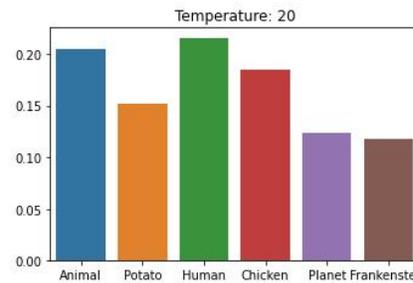
$\tau = 0.5$



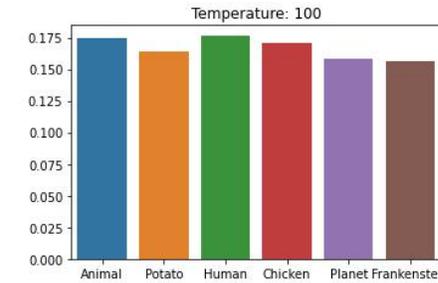
$\tau = 1$



$\tau = 20$



$\tau = 100$



(Source: Mehta, 2023)

Low temperature



High temperature

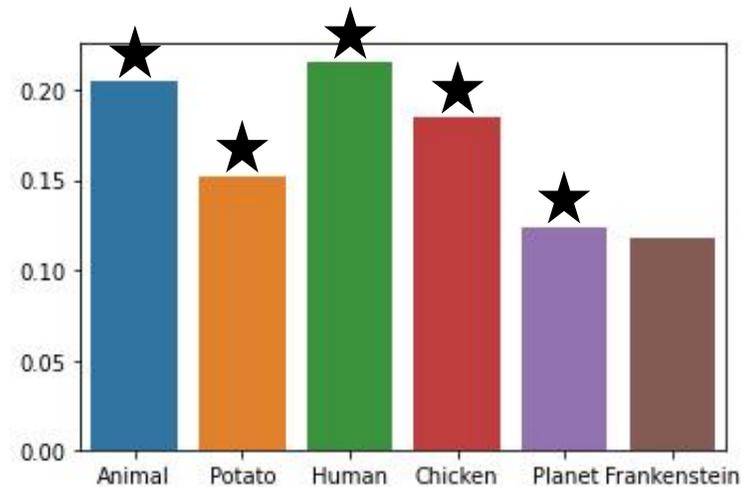
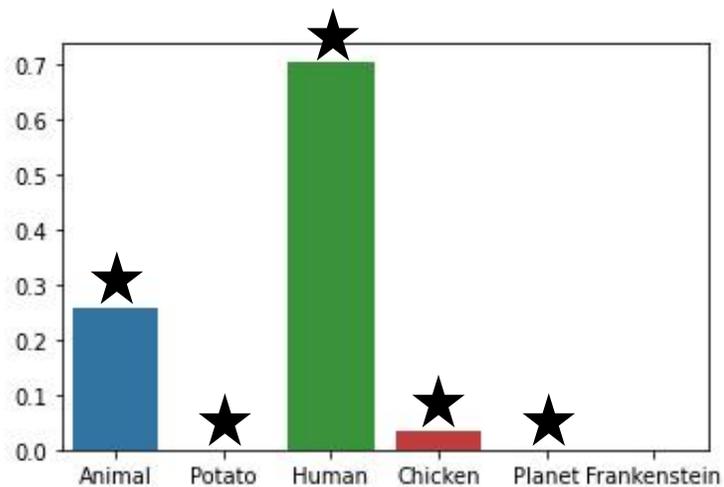


Temperature adjusts the "contrast" of the distribution!

Top-K vs Top-p Sampling

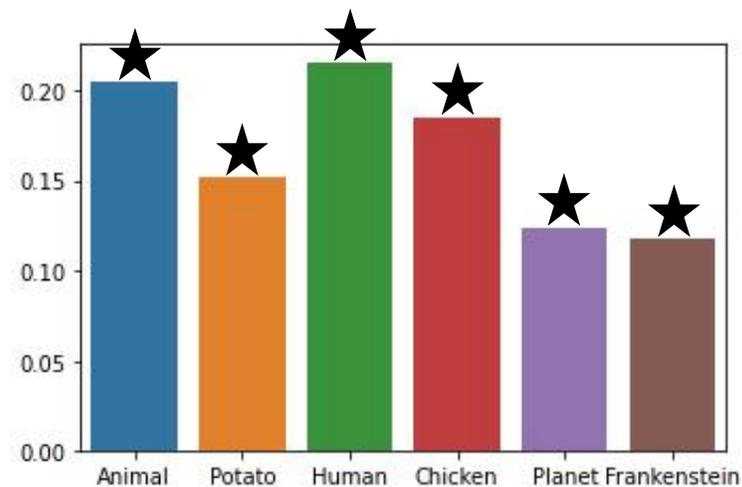
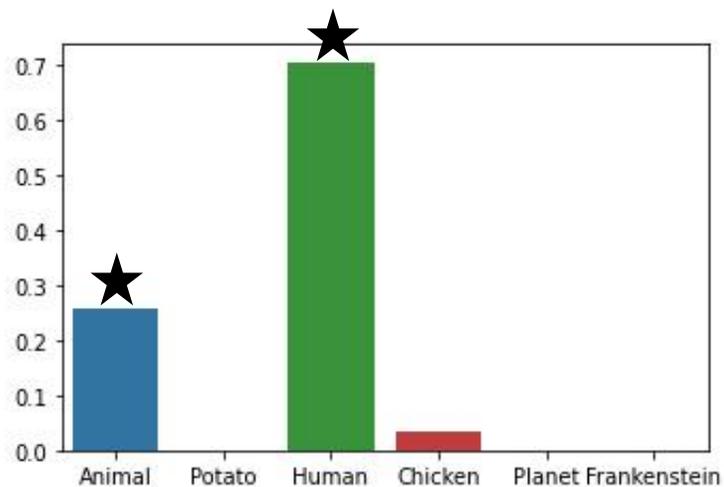
Top-K sampling

$K = 5$



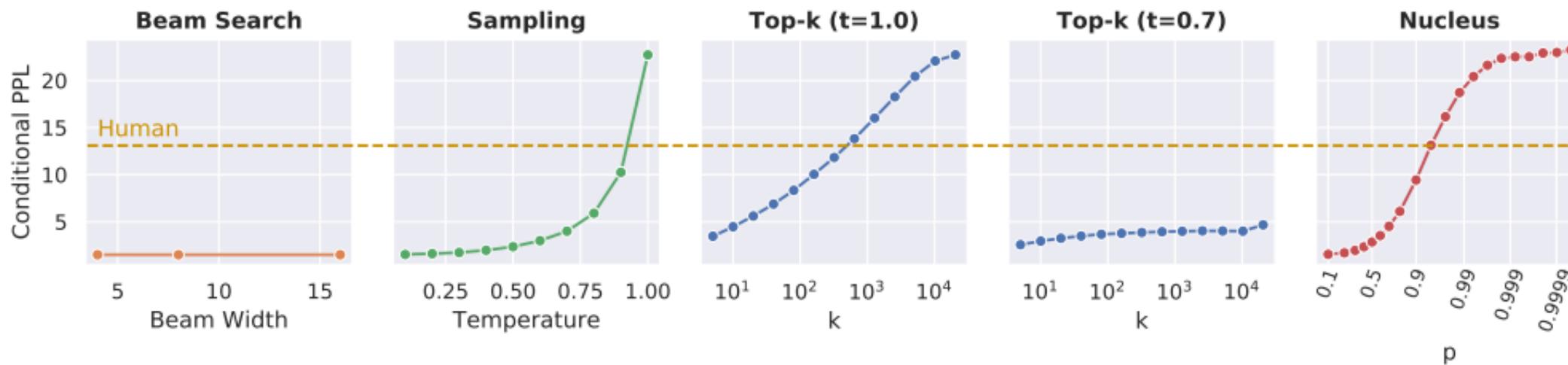
Top-p sampling

$p = 0.95$



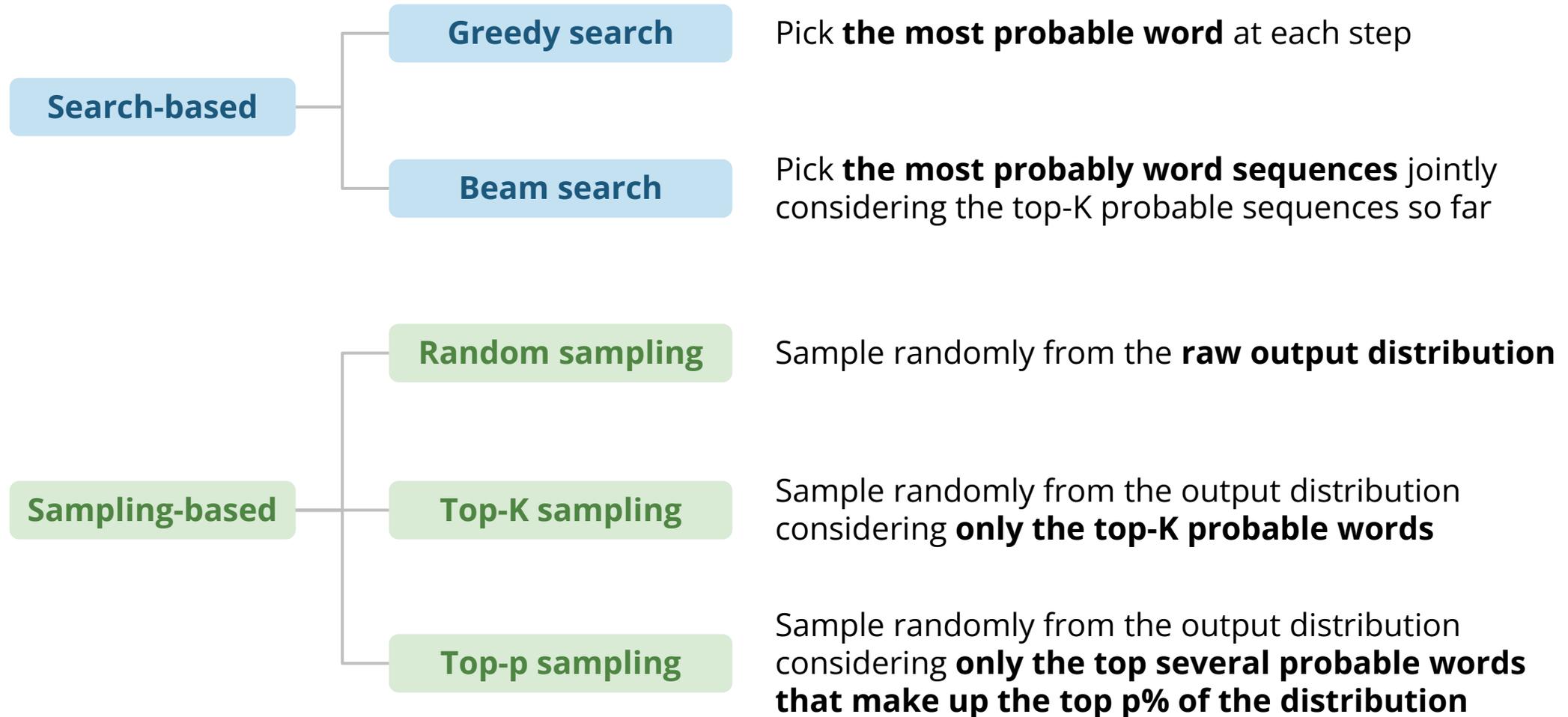
(Source: Mehta, 2023)

Balance between Coherence & Interestingness



(Source: Holtzman et al., 2020)

Decoding Strategies



Tokenization

Breaking Words into Subwords

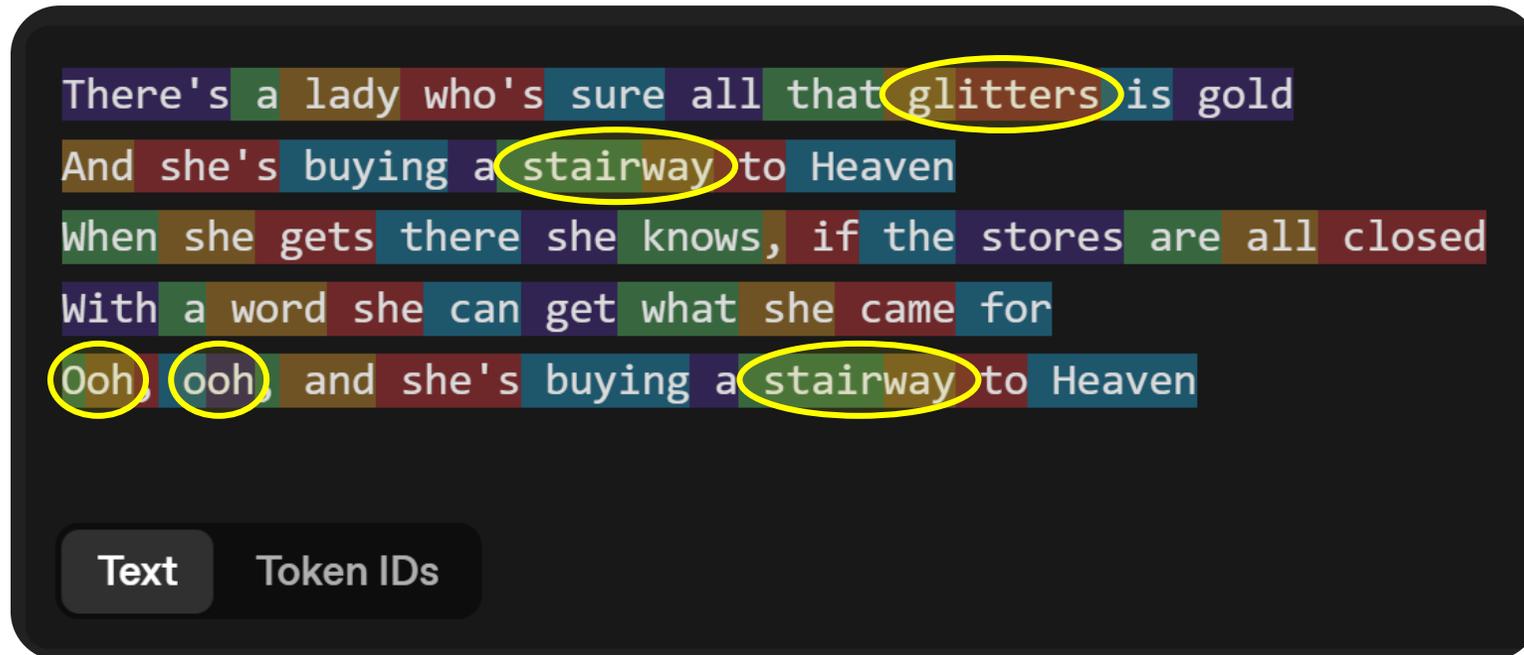
transformer → trans form er

beautiful → beaut iful

whatsoever → what so ever

midwestern → mid west ern

OpenAI's Tokenizer



The screenshot shows the OpenAI tokenizer interface with the following lyrics: "There's a lady who's sure all that glitters is gold", "And she's buying a stairway to Heaven", "When she gets there she knows, if the stores are all closed", "With a word she can get what she came for", and "Ooh, ooh, and she's buying a stairway to Heaven". The words "glitters", "stairway", "Ooh", and "ooh" are highlighted with yellow circles. At the bottom, there are two buttons: "Text" and "Token IDs".

platform.openai.com/tokenizer

Byte-pair Encoding (BPE) (Gage, 1994; Sennrich et al., 2016)

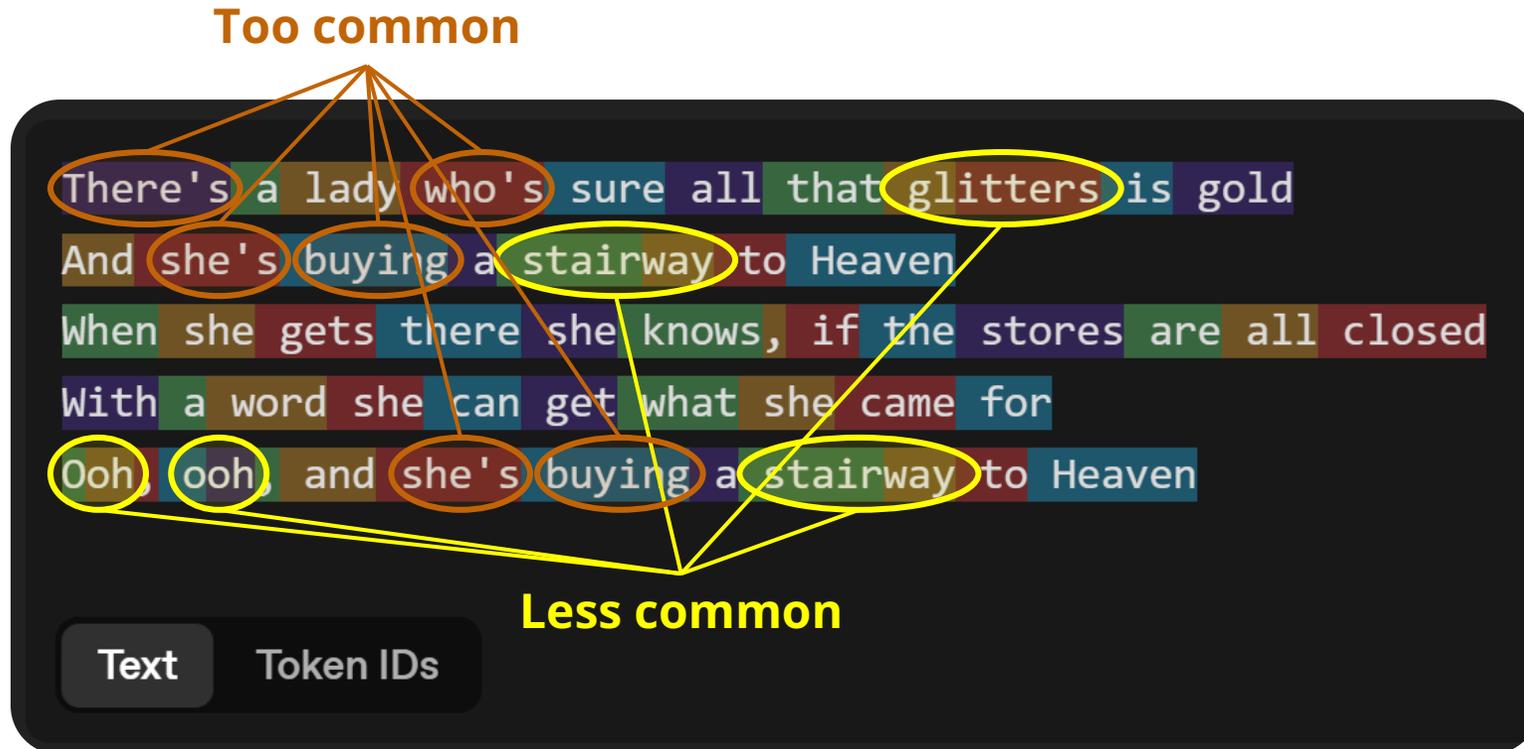
Corpus	Iteration	Learned vocabulary
	1	n e w s r t b l o
new	2	n e w s r t b l o 4x ne
news	3	n e w s r t b l o ne 4x new
newer	4	n e w s r t b l o ne new 2x er
newest	5	n e w s r t b l o ne new er 2x es
best	6	n e w s r t b l o ne new er es 2x est
low	7	n e w s r t b l o ne new er es est 2x lo
lower	8	n e w s r t b l o ne new er es est lo 2x low

Merge inputs in the ordering how the merging rules were learned

Philip Gage, "A New Algorithm for Data Compression," *The C Users Journal*, 12(2):23–38, 1994.

Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural Machine Translation of Rare Words with Subword Units," *ACL*, 2016.

OpenAI's Tokenizer



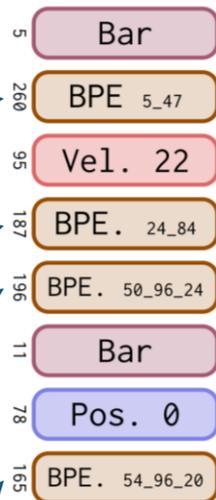
platform.openai.com/tokenizer

Byte-pair Encoding for Music (Fradet et al., 2023)

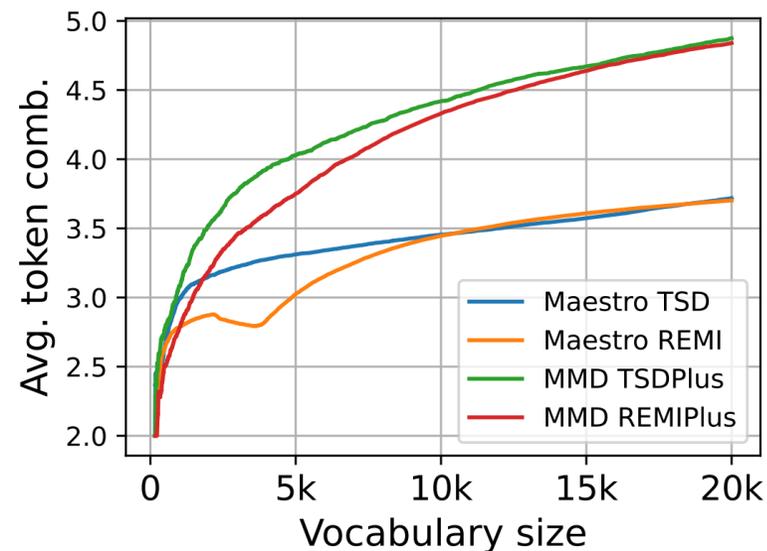
Without BPE



With BPE

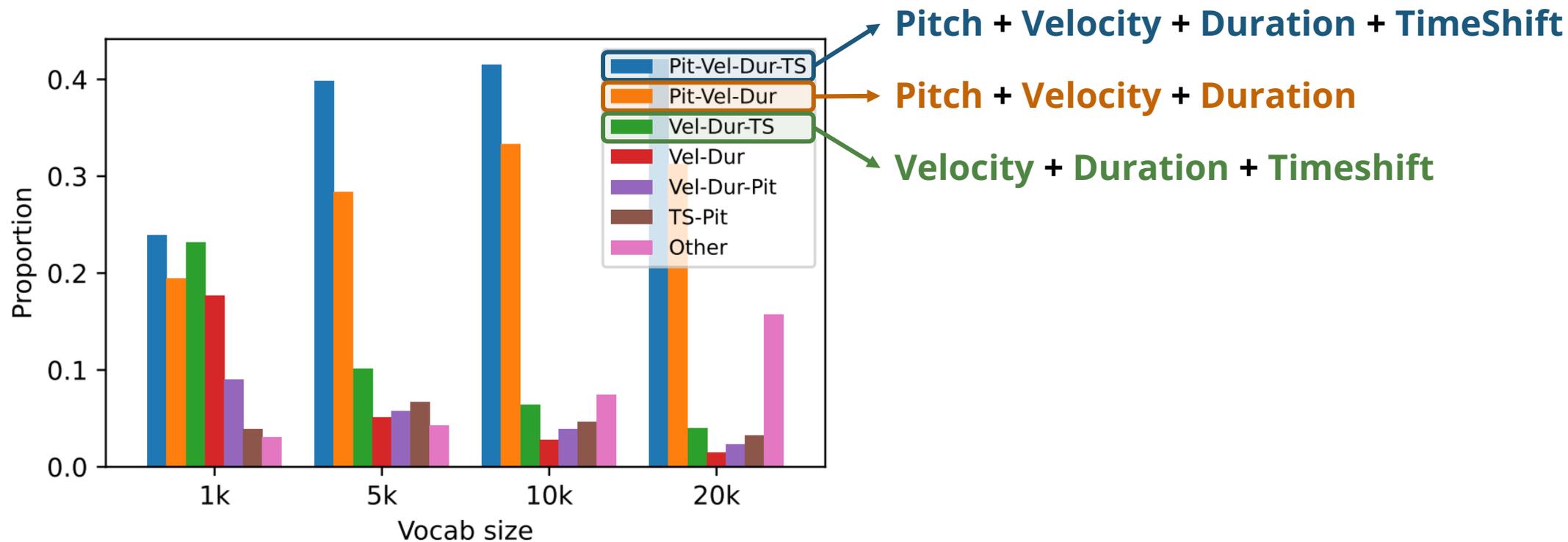


Learned BPE size



(Source: Fradet et al., 2020)

Byte-pair Encoding for Music (Fradet et al., 2023)



(a) Maestro and *TSD*

(Source: Fradet et al., 2020)

Four Paradigms of Music Generation



Symbolic music generation

Text-based

```
Program_change_0,  
Note_on_60, Time_shift_2, Note_off_60,  
Note_on_60, Time_shift_2, Note_off_60,  
Note_on_76, Time_shift_2, Note_off_67,  
Note_on_67, Time_shift_2, Note_off_67,  
...
```

MIDI

So far!

Image-based



Piano roll



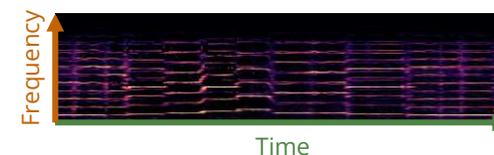
Audio-domain music generation

Time series-based



Waveform

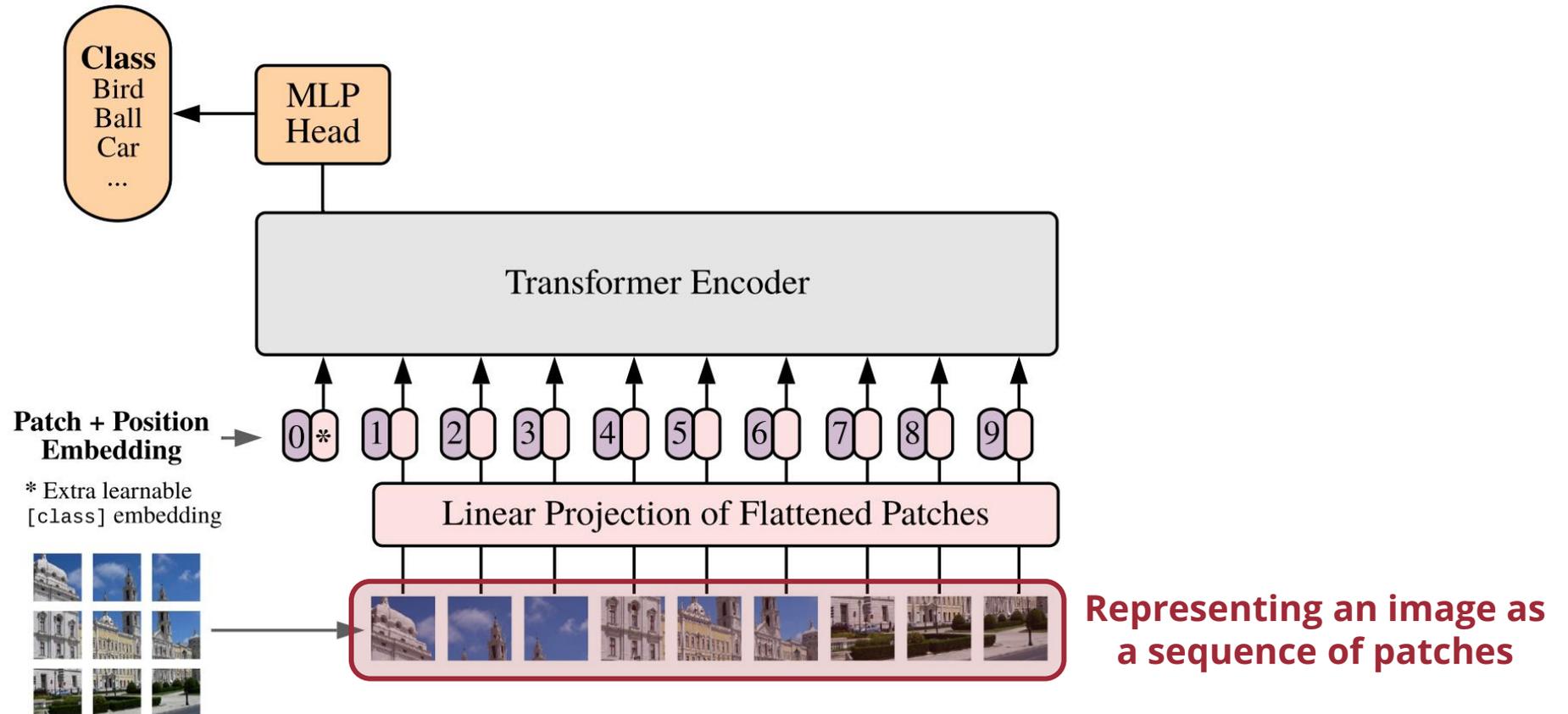
Image-based



Spectrogram

Transformers **beyond** Languages

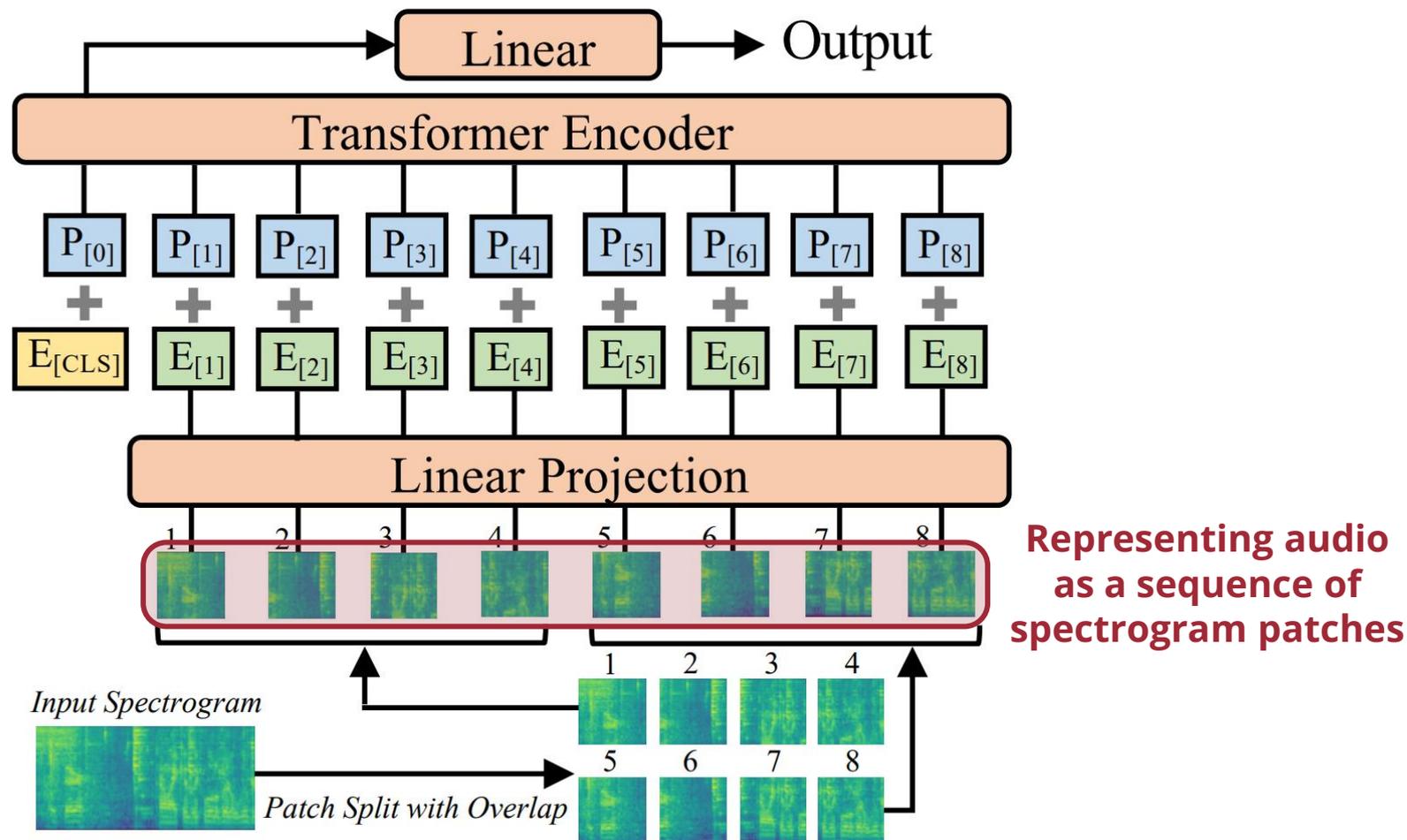
Vision Transformer (ViT) (Dosovitskiy et al., 2021)



(Source: Dosovitskiy et al., 2021)

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021.

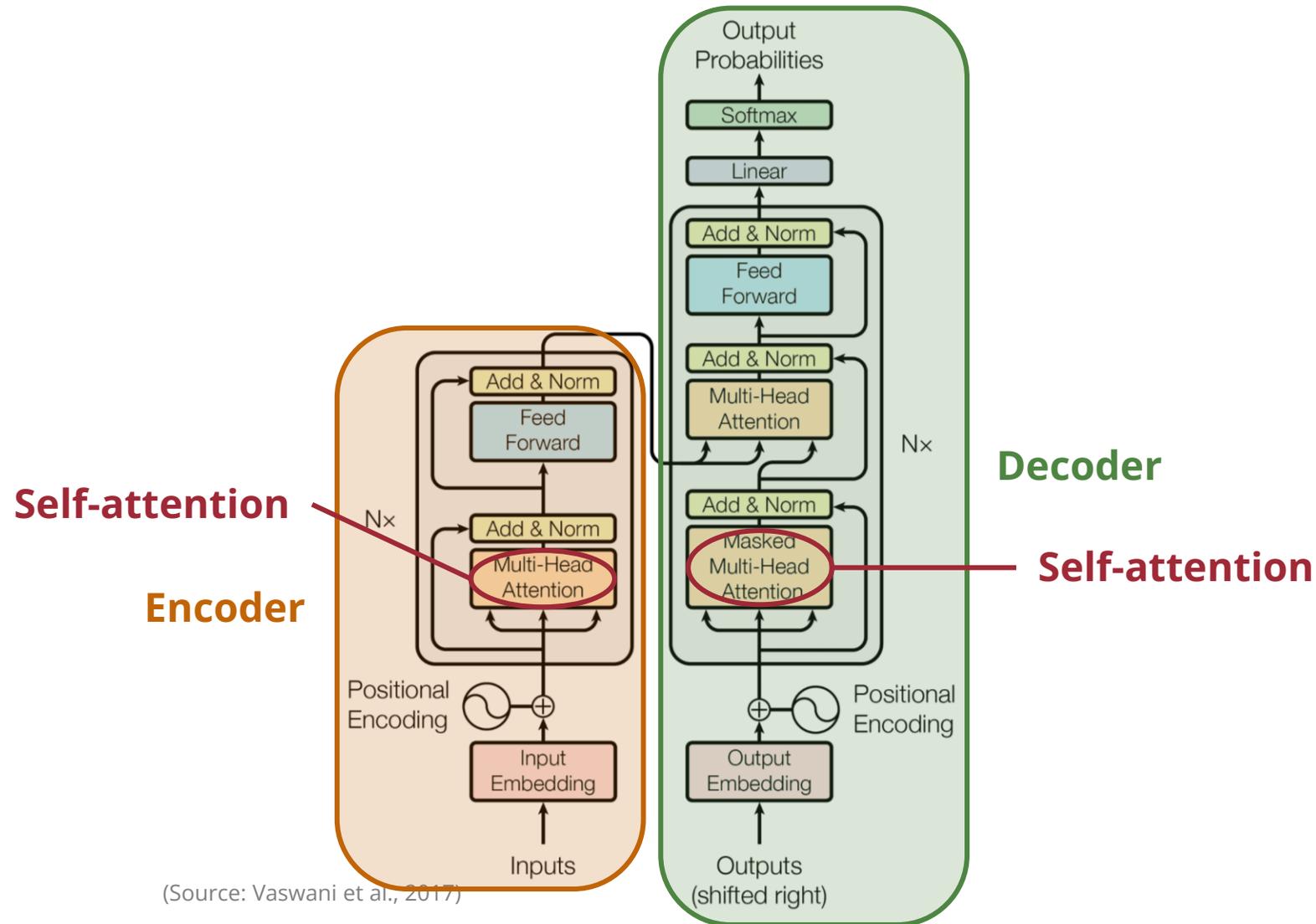
Audio Spectrogram Transformer (AST) (Gong et al., 2021)



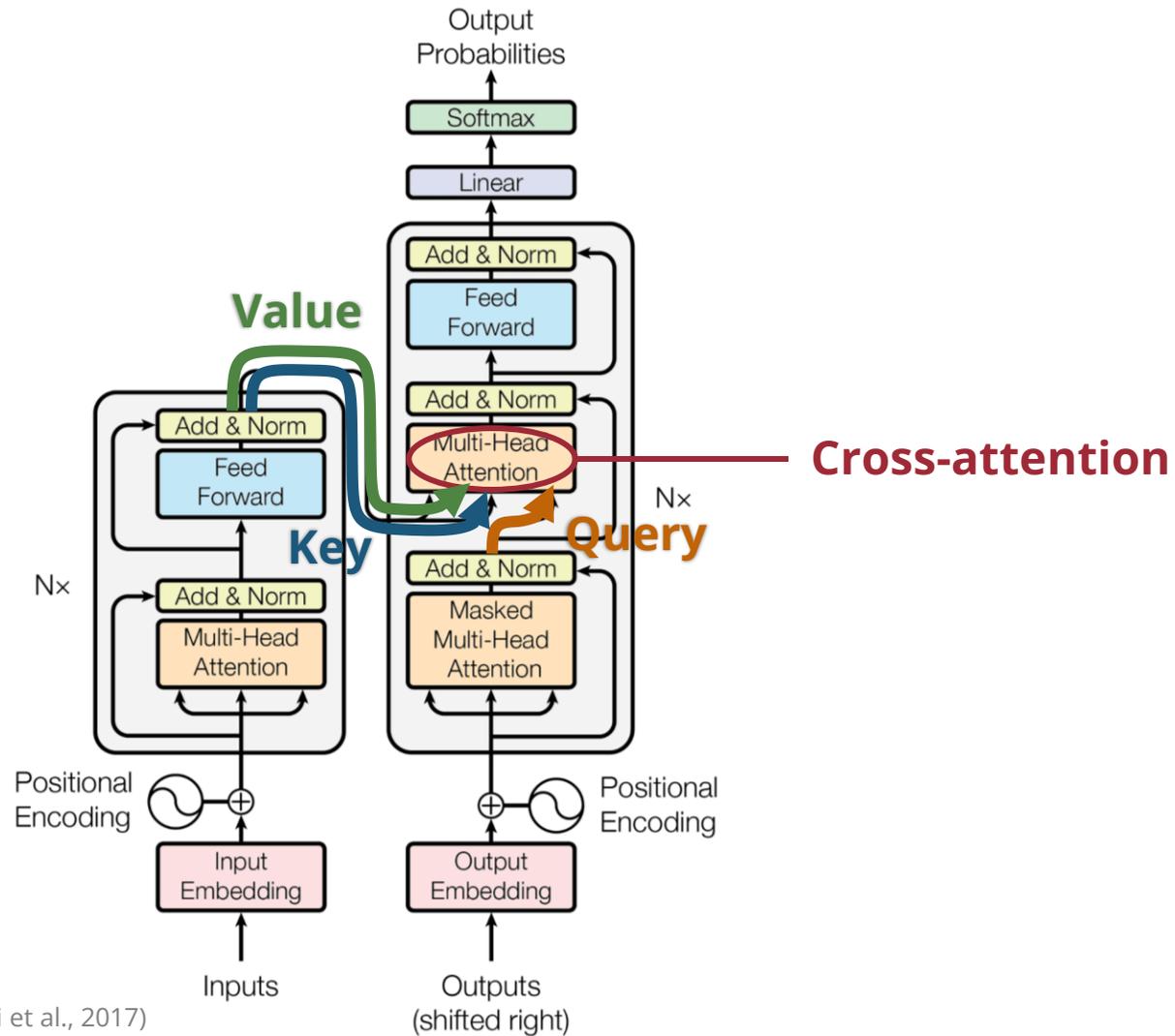
(Source: Gong et al., 2021)

Recap

The Original Transformer (Vaswani et al., 2017)



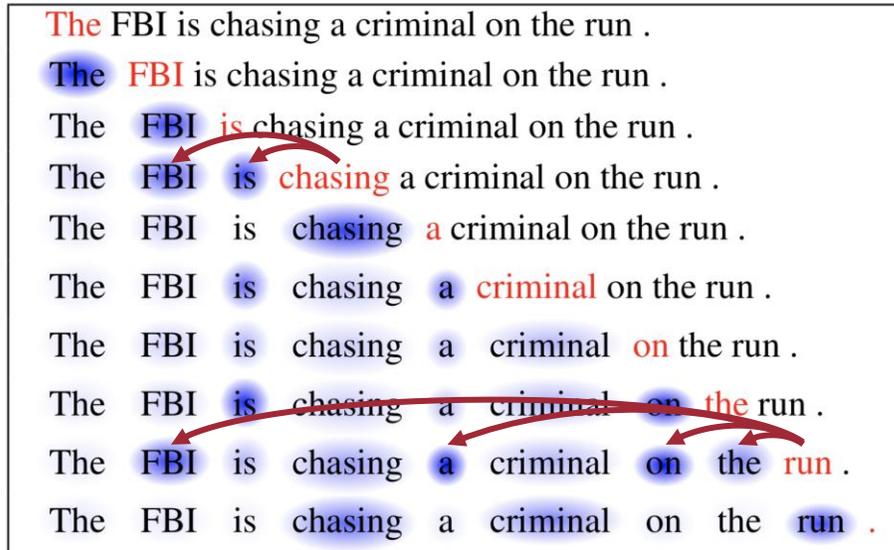
Cross-attention (Vaswani et al., 2017)



(Source: Vaswani et al., 2017)

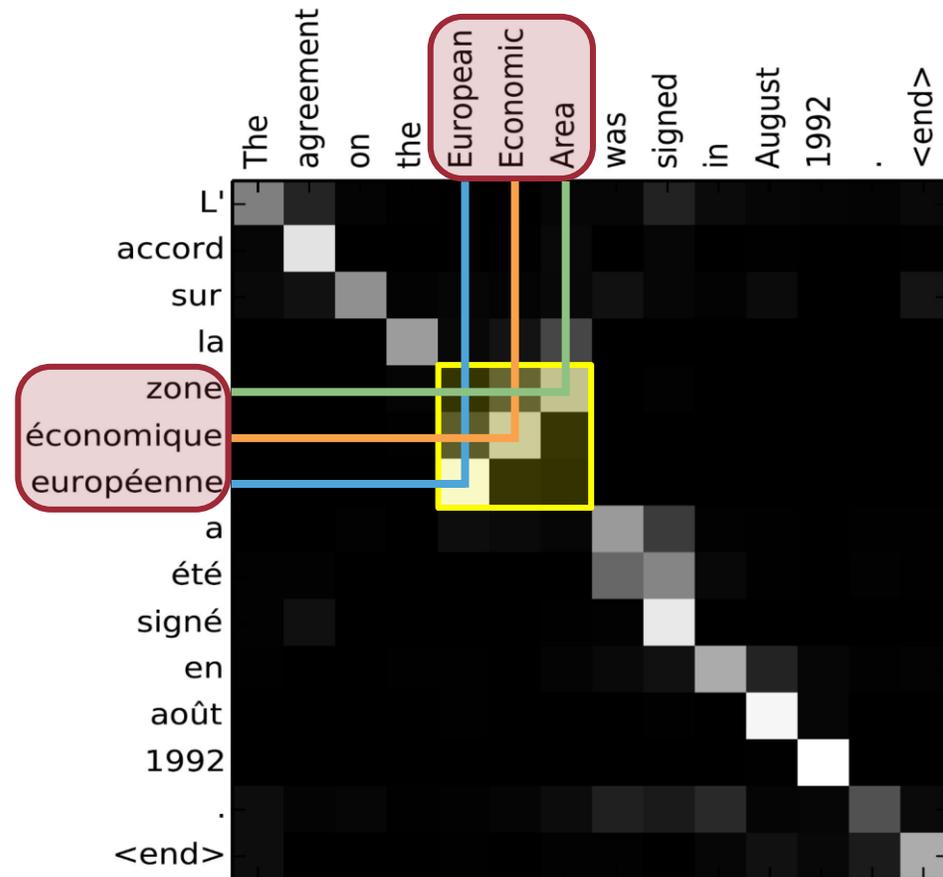
Self-Attention vs. Cross-Attention

Self-attention



(Source: Cheng et al., 2016)

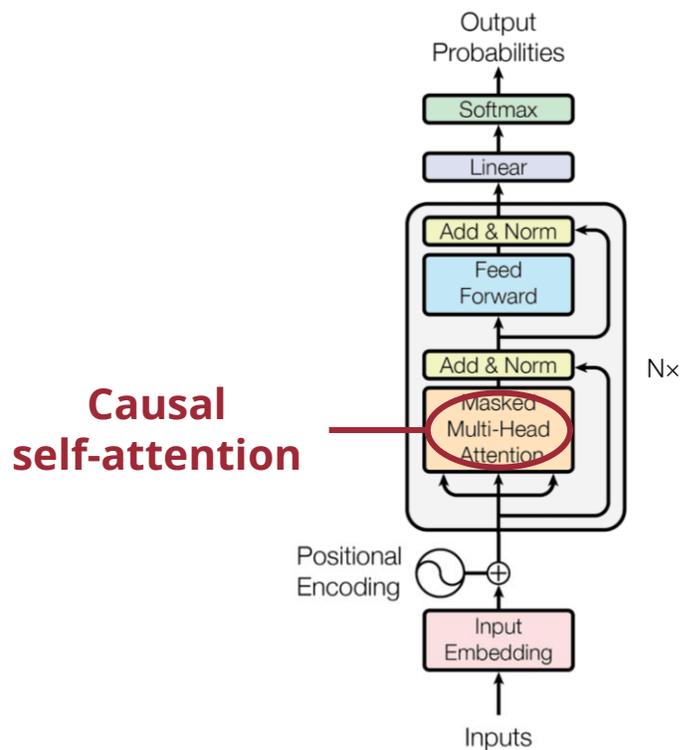
Cross-Attention



(Source: Bahdanau et al., 2015)

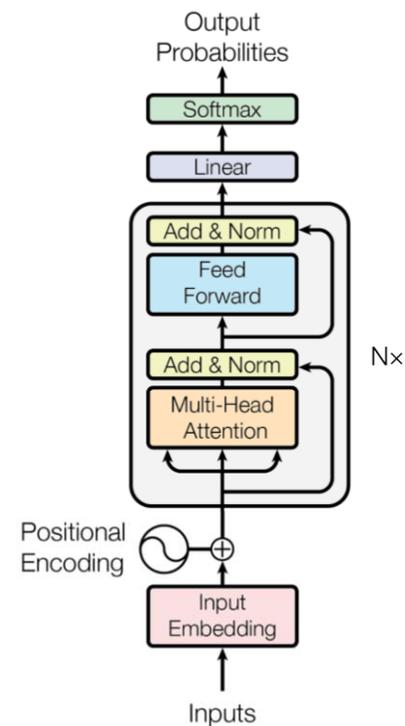
Decoder-only vs. Encoder-only Transformer

Decoder-only transformer



Access to **only past** information

Encoder-only transformer

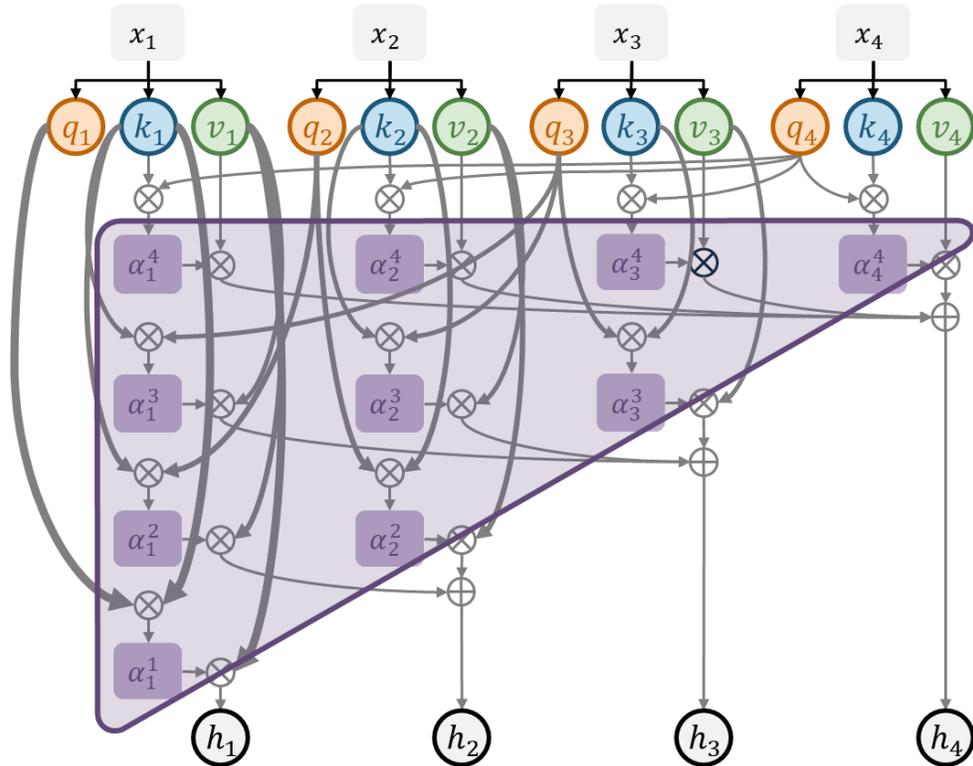


Access to **past & future** information

(Source: Vaswani et al., 2017)

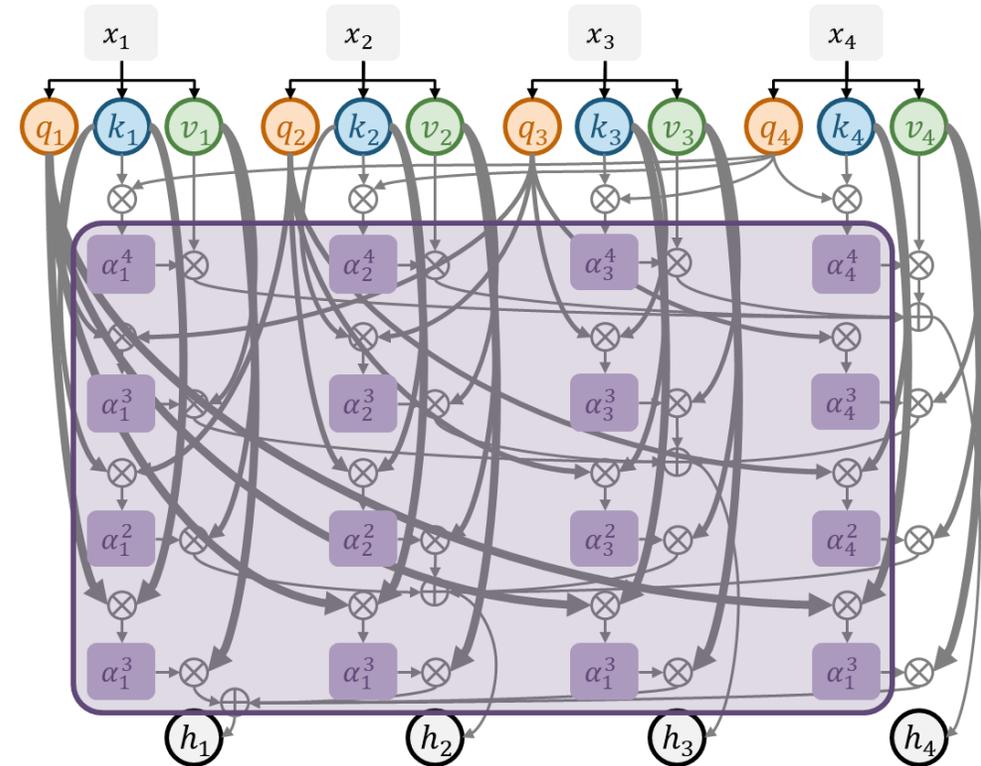
Decoder-only vs. Encoder-only Transformer

Decoder-only transformer



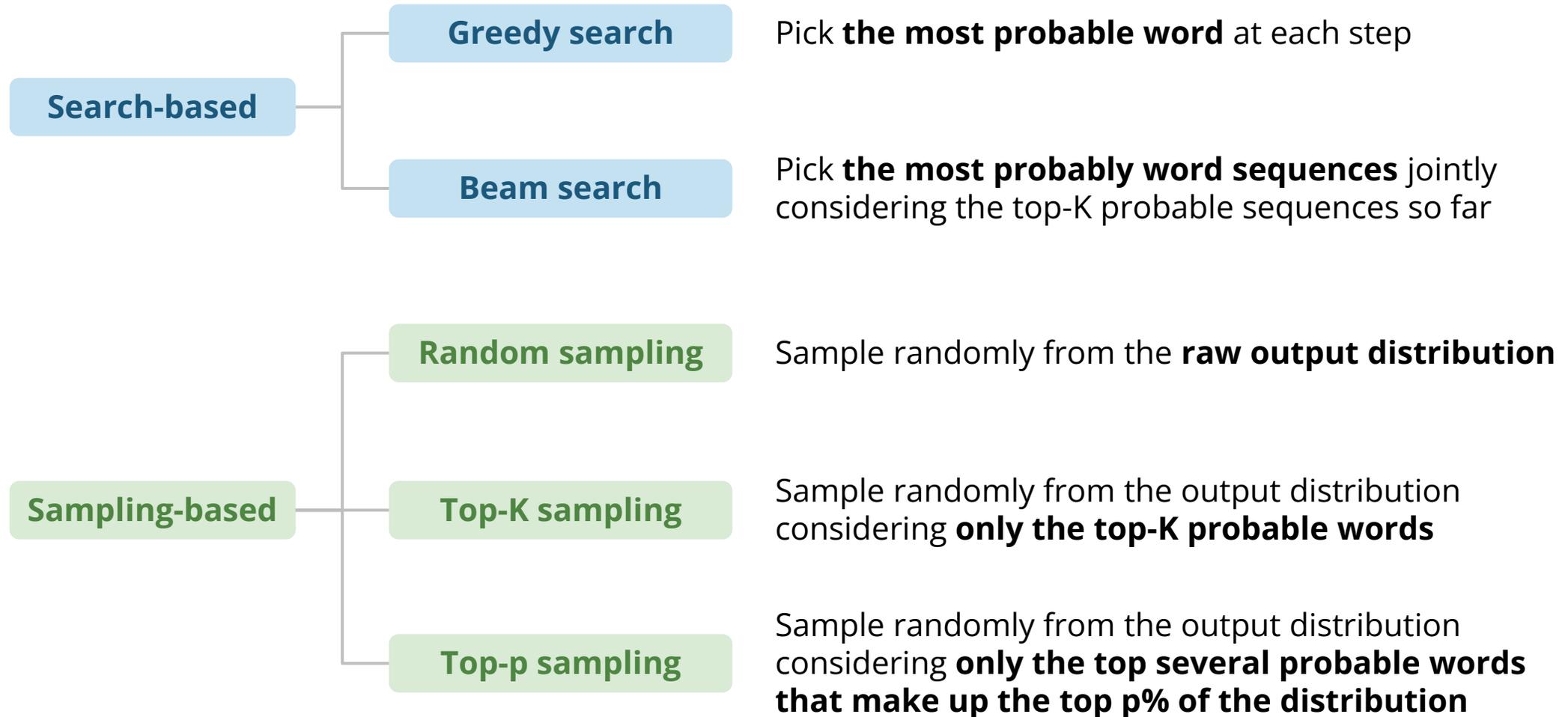
Access to **only past** information

Encoder-only transformer

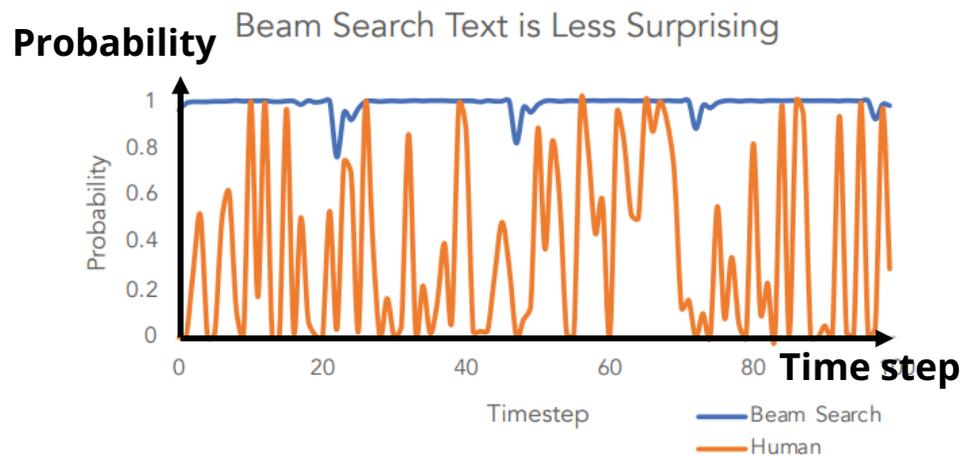


Access to **past & future** information

Decoding Strategies



Do We Really Want the Most Probable Sequence?



(Source: Holtzman et al., 2020)

Beam Search

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

Human

...which grant increased life span and three years warranty. The Antec HCG series consists of five models with capacities spanning from 400W to 900W. Here we should note that we have already tested the HCG-620 in a previous review and were quite satisfied With its performance. In today's review we will rigorously test the Antec HCG-520, which as its model number implies, has 520W capacity and contrary to Antec's strong beliefs in multi-rail PSUs is equipped...

Temperature

Softmax

$$\hat{y}_i = \frac{e^{\tilde{y}_i}}{\sum_{j=1}^n e^{\tilde{y}_j}}$$

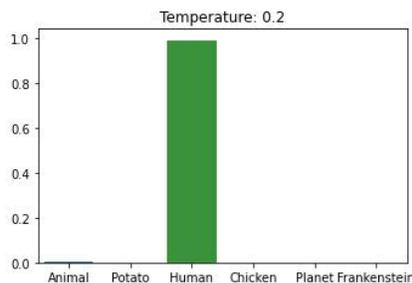


$$\hat{y}_i = \frac{e^{\tilde{y}_i/\tau}}{\sum_{j=1}^n e^{\tilde{y}_j/\tau}}$$

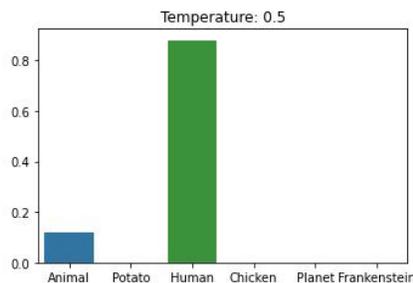
Temperature

Original

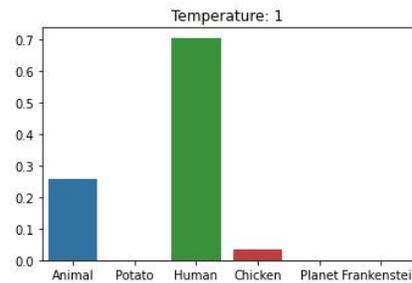
$\tau = 0.2$



$\tau = 0.5$

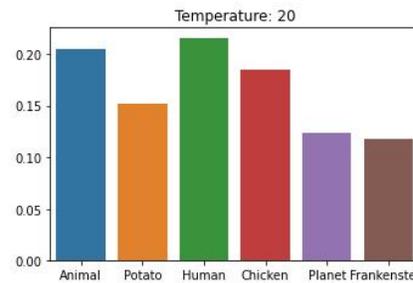


$\tau = 1$

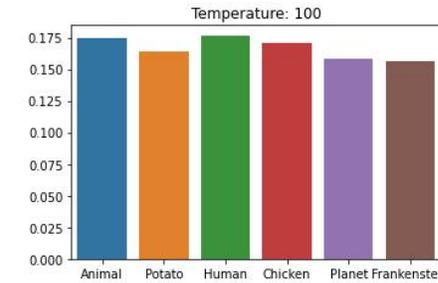


(Source: Mehta, 2023)

$\tau = 20$



$\tau = 100$



Low temperature



High temperature

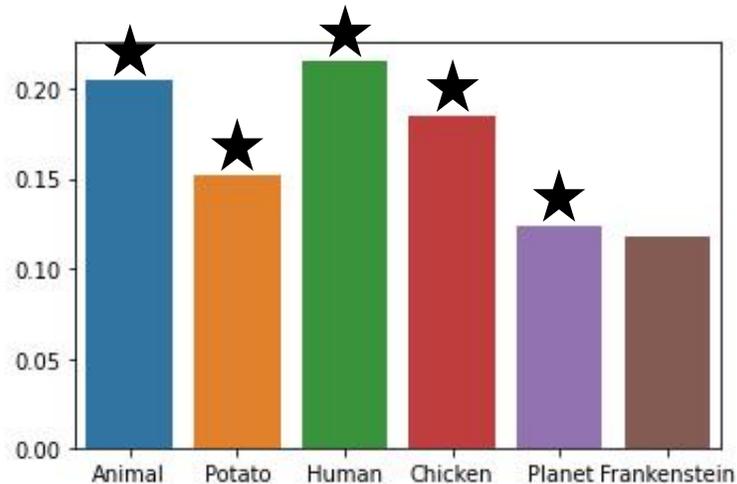
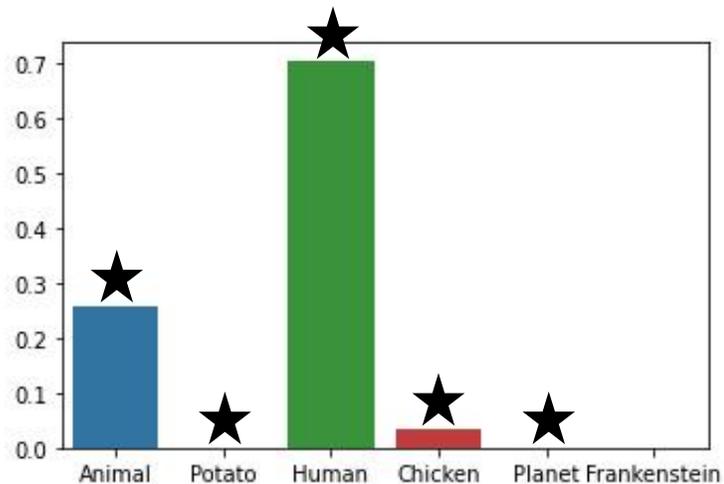


Temperature adjusts the "contrast" of the distribution!

Top-K vs Top-p Sampling

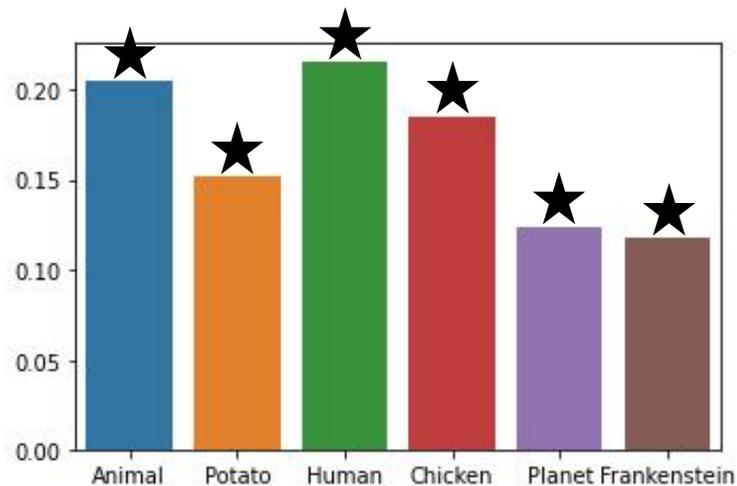
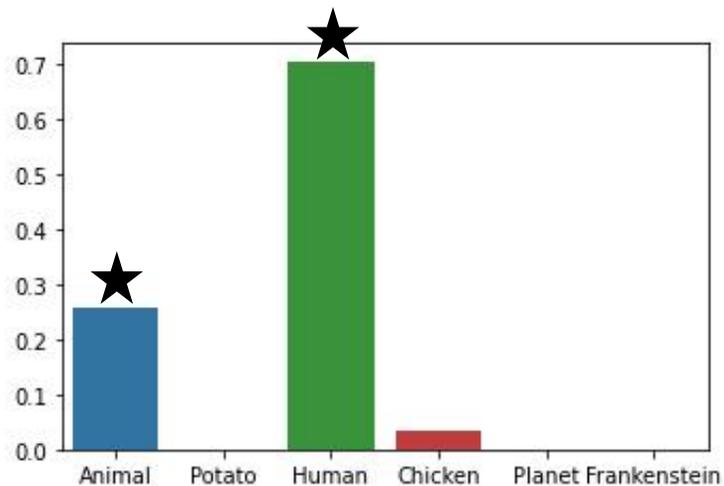
Top-K sampling

$K = 5$



Top-p sampling

$p = 0.95$



Temperature can be applied at the same time

Breaking Words into Subwords

transformer → trans form er

beautiful → beaut iful

whatsoever → what so ever

midwestern → mid west ern

Byte-pair Encoding (BPE) (Gage, 1994; Sennrich et al., 2016)

Corpus	Iteration	Learned vocabulary
	1	n e w s r t b l o
new	2	n e w s r t b l o 4x ne
news	3	n e w s r t b l o ne 4x new
newer	4	n e w s r t b l o ne new 2x er
newest	5	n e w s r t b l o ne new er 2x es
best	6	n e w s r t b l o ne new er es 2x est
low	7	n e w s r t b l o ne new er es est 2x lo
lower	8	n e w s r t b l o ne new er es est lo 2x low

Merge inputs in the ordering how the merging rules were learned

Philip Gage, "A New Algorithm for Data Compression," *The C Users Journal*, 12(2):23-38, 1994.

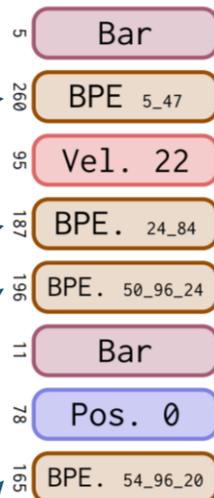
Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural Machine Translation of Rare Words with Subword Units," *ACL*, 2016.

Byte-pair Encoding for Music (Fradet et al., 2023)

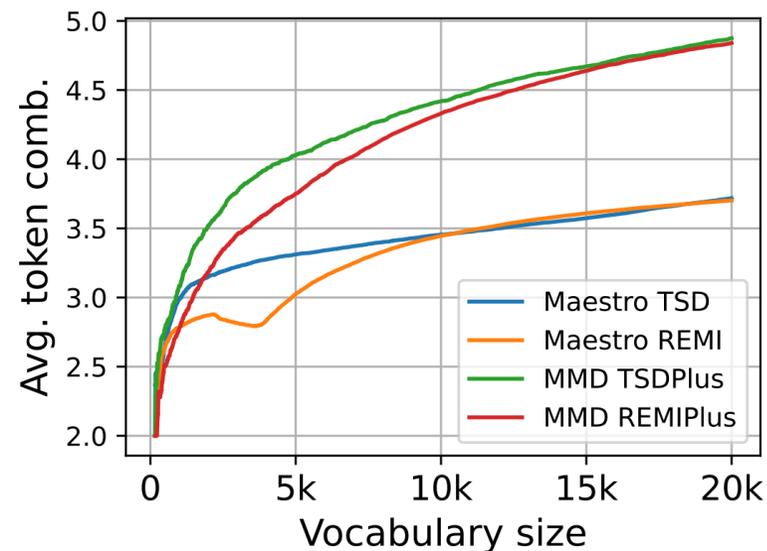
Without BPE



With BPE



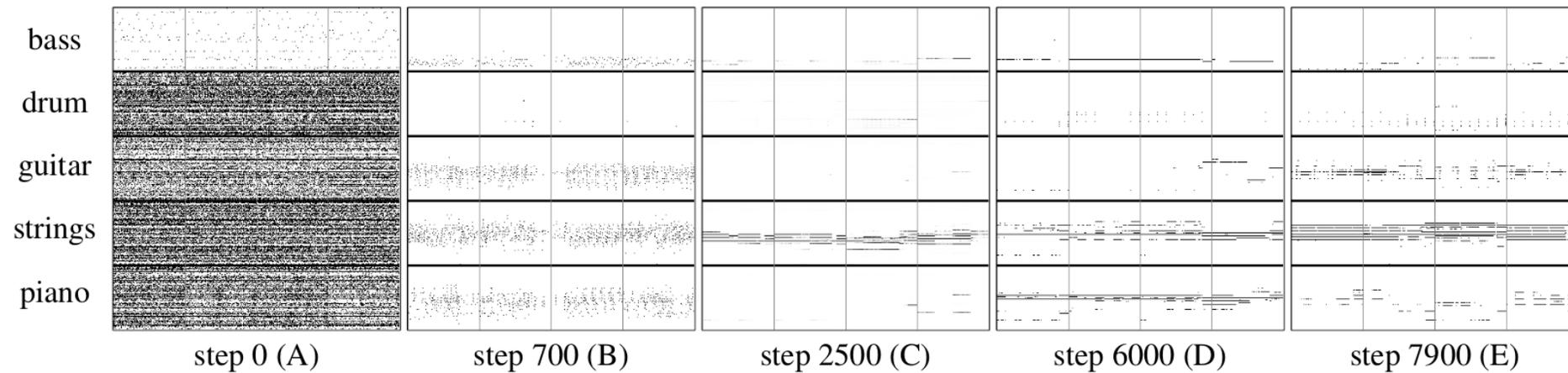
Learned BPE size



(Source: Fradet et al., 2020)

Next Lecture

VAEs & GANs



(Source: Dong et al., 2018)