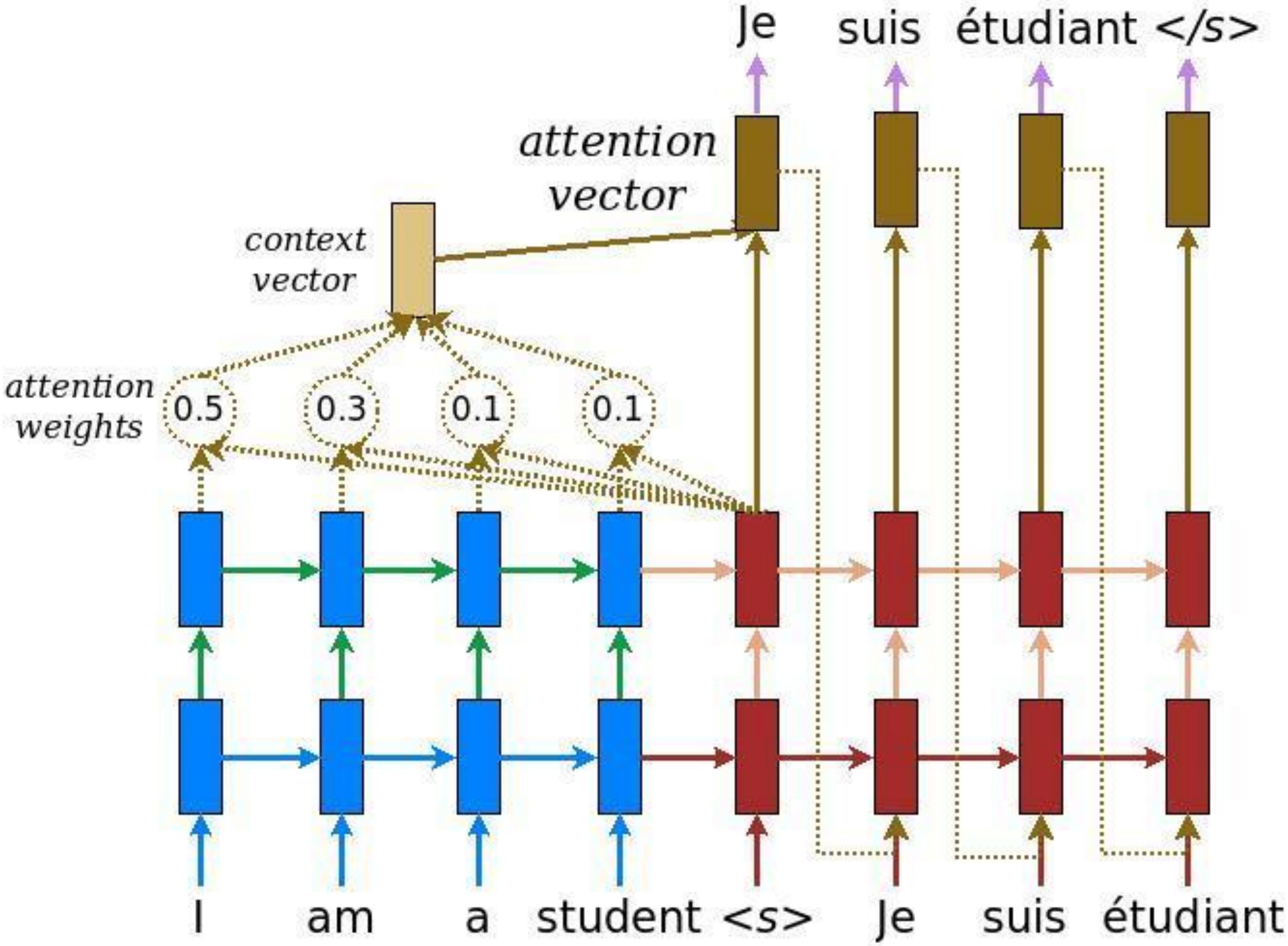
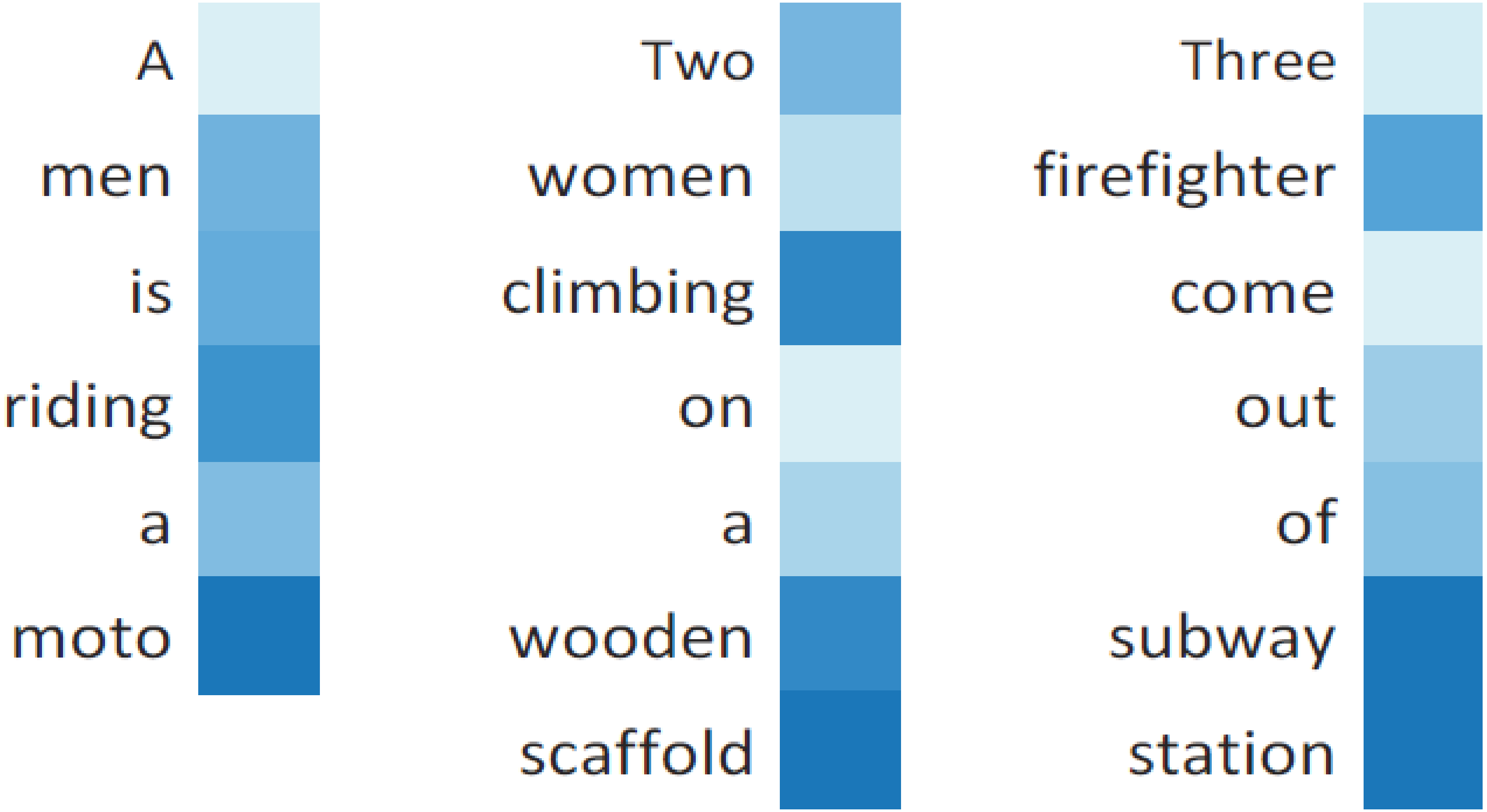
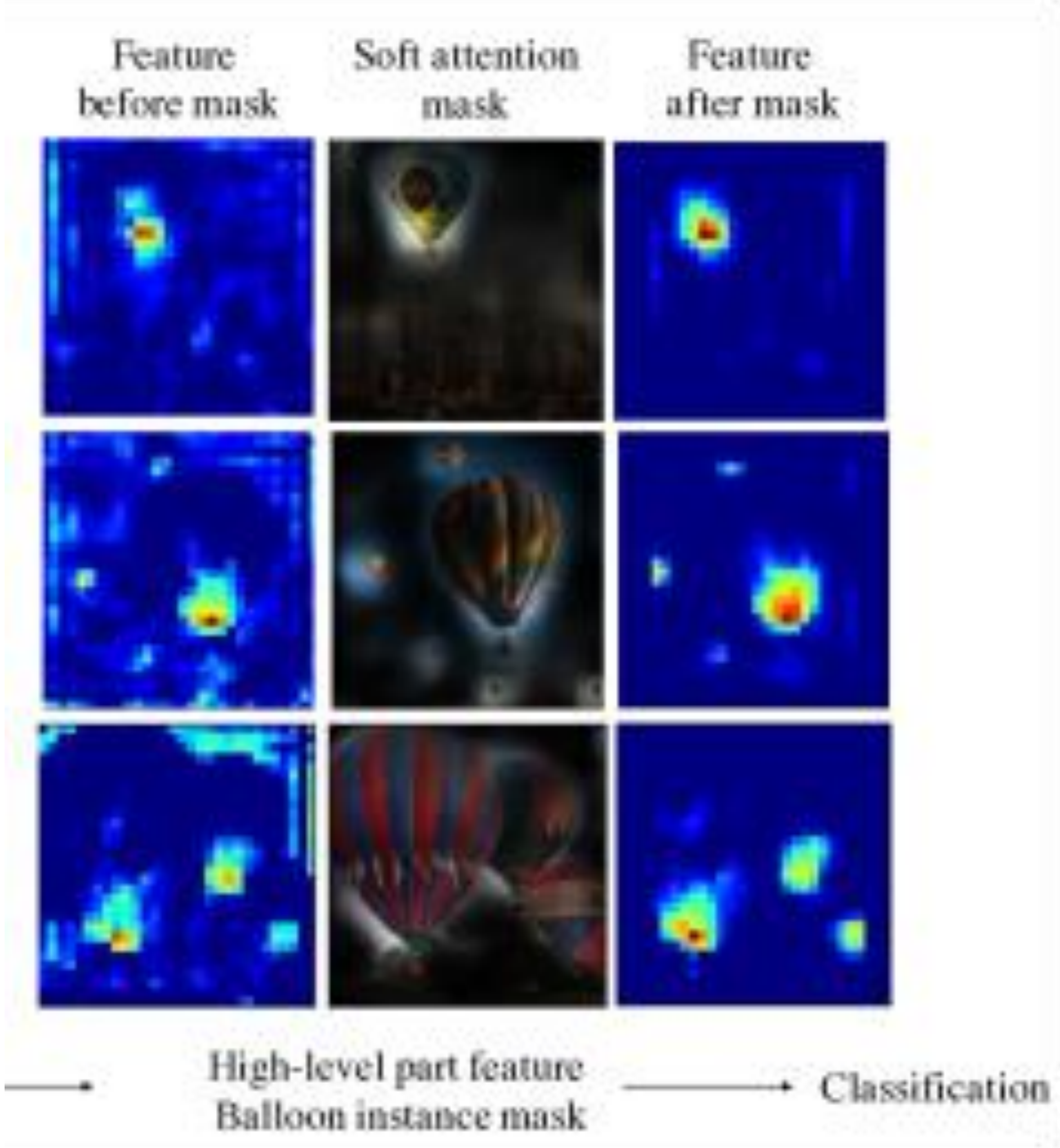
Attention


Je
suis
étudiant
</s>
attention vector
context vector
attention weights
0.5
0.3
0.1
0.1
I
am
a
student
<s>
Je
suis
étudiant

Attention


A
men
is
riding
a
moto
Two
women
climbing
on
a
wooden
scaffold
Three
firefighter
come
out
of
subway
station

Visual attention



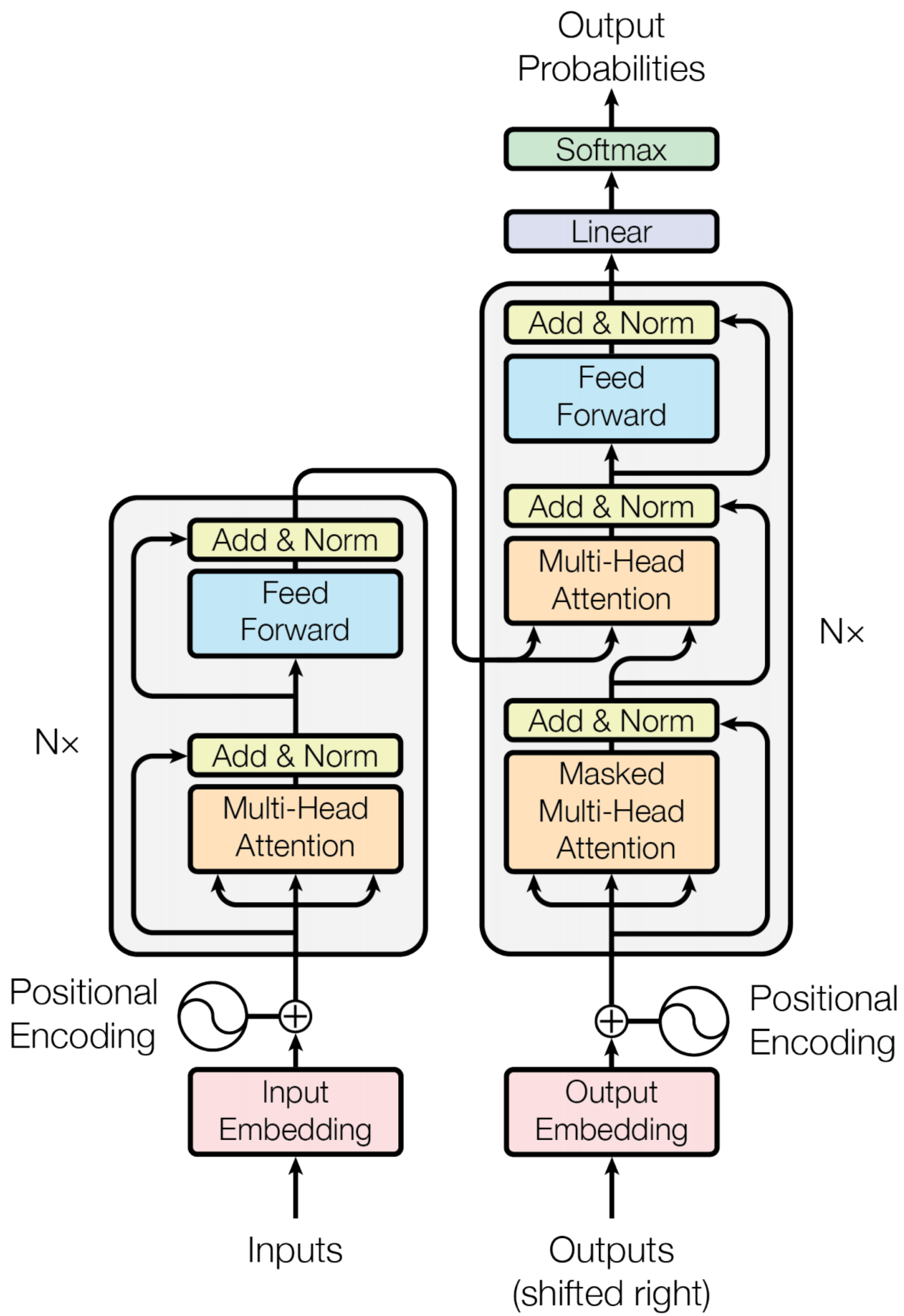
SQuAD1

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Sep 13, 2018	nlnet (single model) *Microsoft Research Asia*	**74.238**	**77.022**
2 Sep 17, 2018	Unet (ensemble) *Fudan University & Liulishuo Lab*	71.553	75.011
2 Aug 15, 2018	Reinforced Mnemonic Reader + Answer Verifier (single model) *NUDT* https://arxiv.org/abs/1808.05759	71.699	74.238
2 Aug 28, 2018	SLQA+ (single model) *Alibaba DAMO NLP* http://www.aclweb.org/anthology/P18-1158	71.451	74.422
3 Sep 14, 2018	SAN (ensemble model) *Microsoft Business Applications Research Group* https://arxiv.org/abs/1712.03556	71.282	73.658

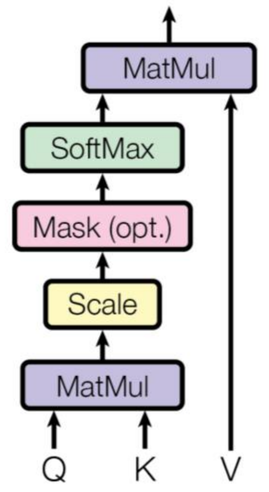
SQuAD2

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	**88.592**	**90.859**
2 Jul 19, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk*	88.050	90.645
3 Jul 23, 2019	XLNet + SG-Net Verifier (single model) *Shanghai Jiao Tong University & CloudWalk*	87.046	89.899
3 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	87.147	89.474
3 Jul 20, 2019	RoBERTa (single model) *Facebook AI*	86.820	89.795
4 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
5 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147

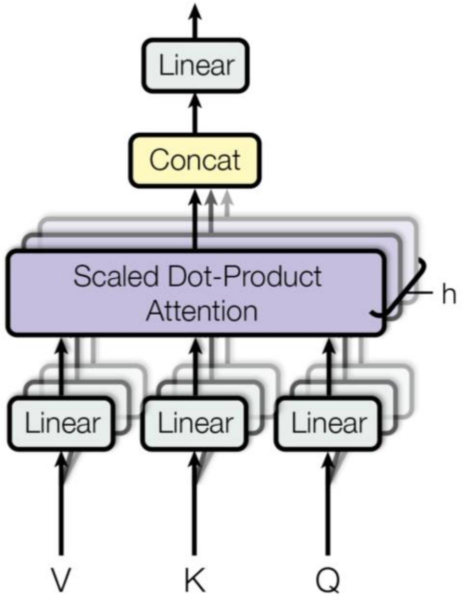
Multi-Head Attention Transformer



Scaled Dot-Product Attention



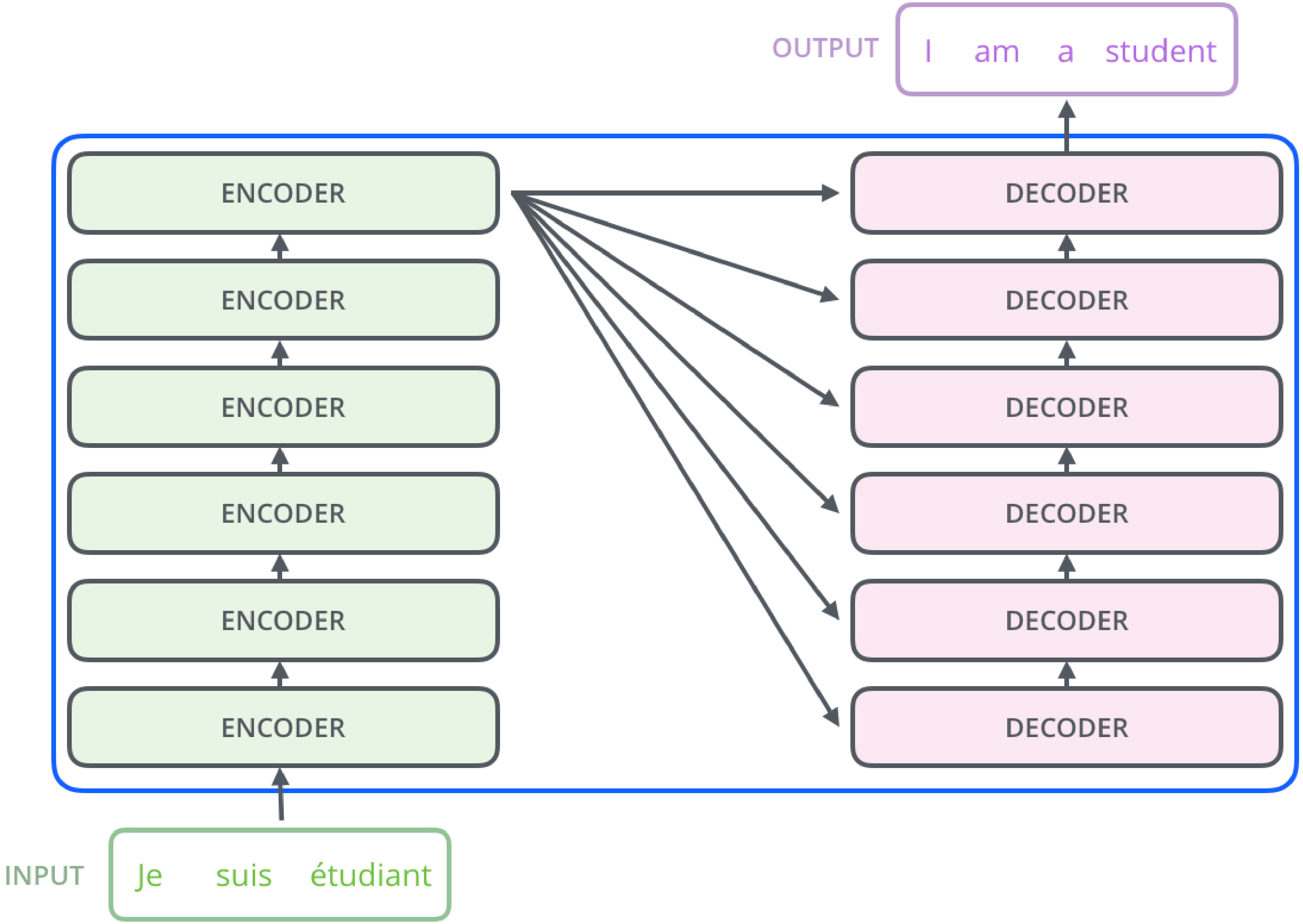
Multi-Head Attention



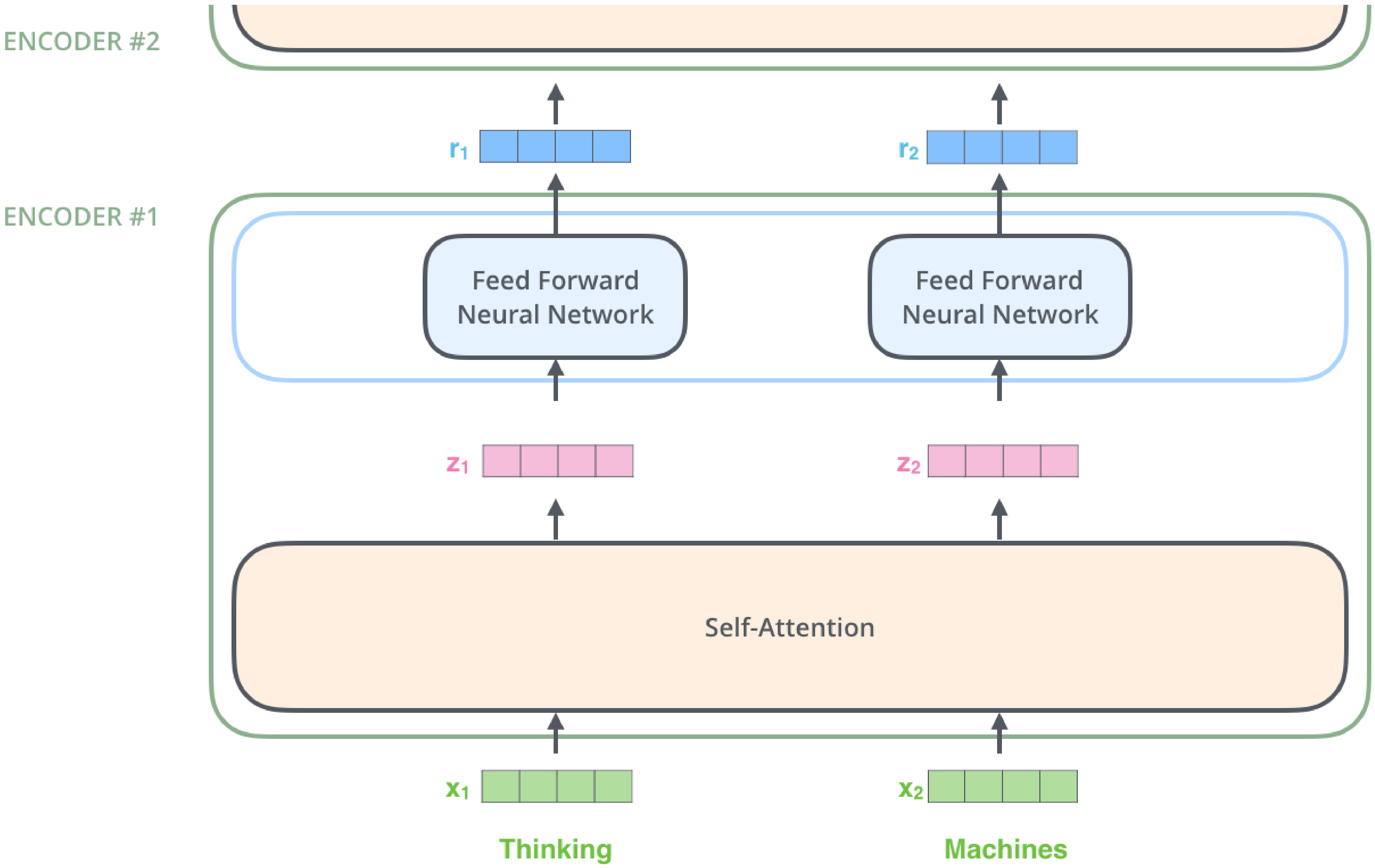
The Illustrated Transformer

http://jalammar.github.io/illustrated-transformer/

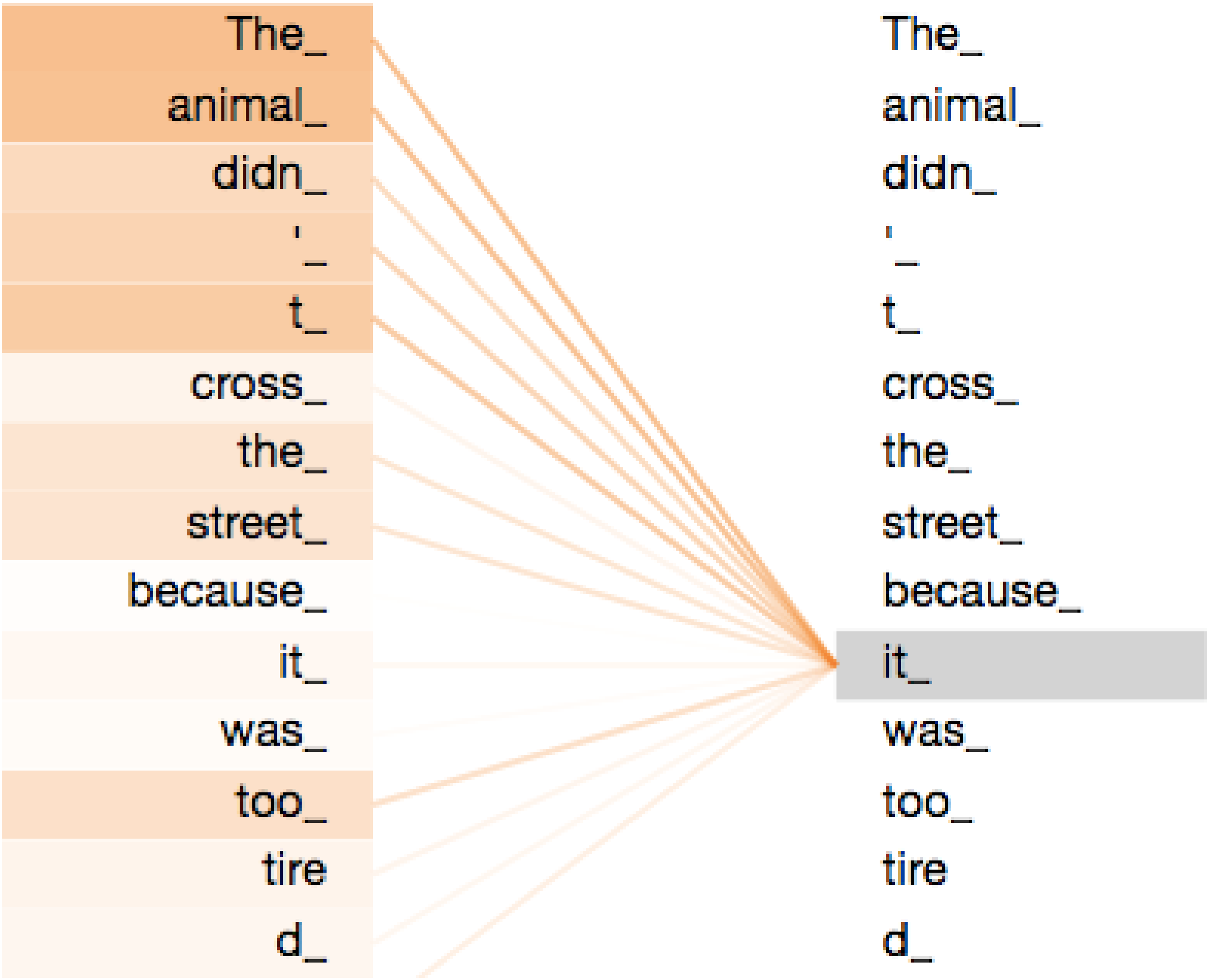
Transformer



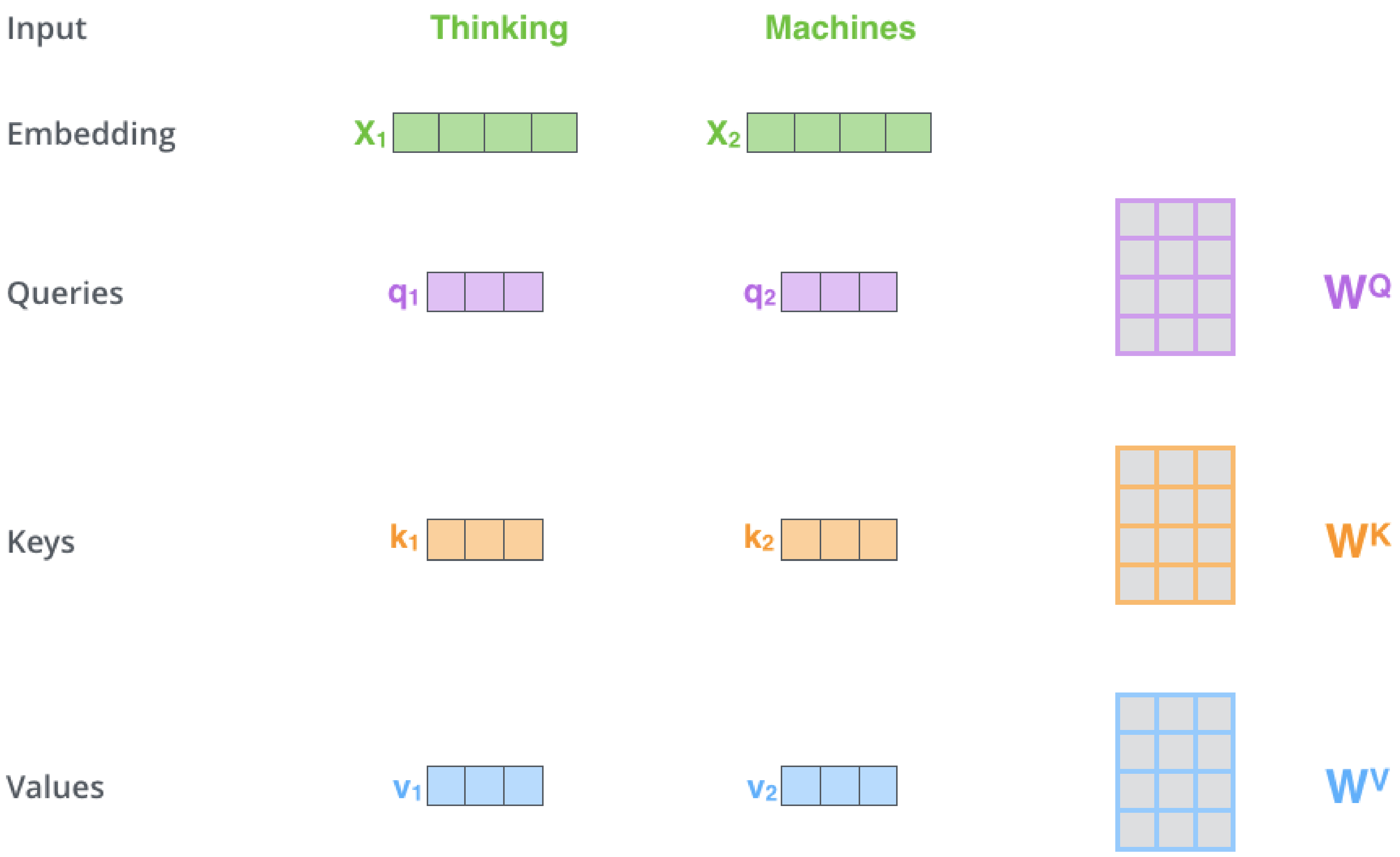
Transformer



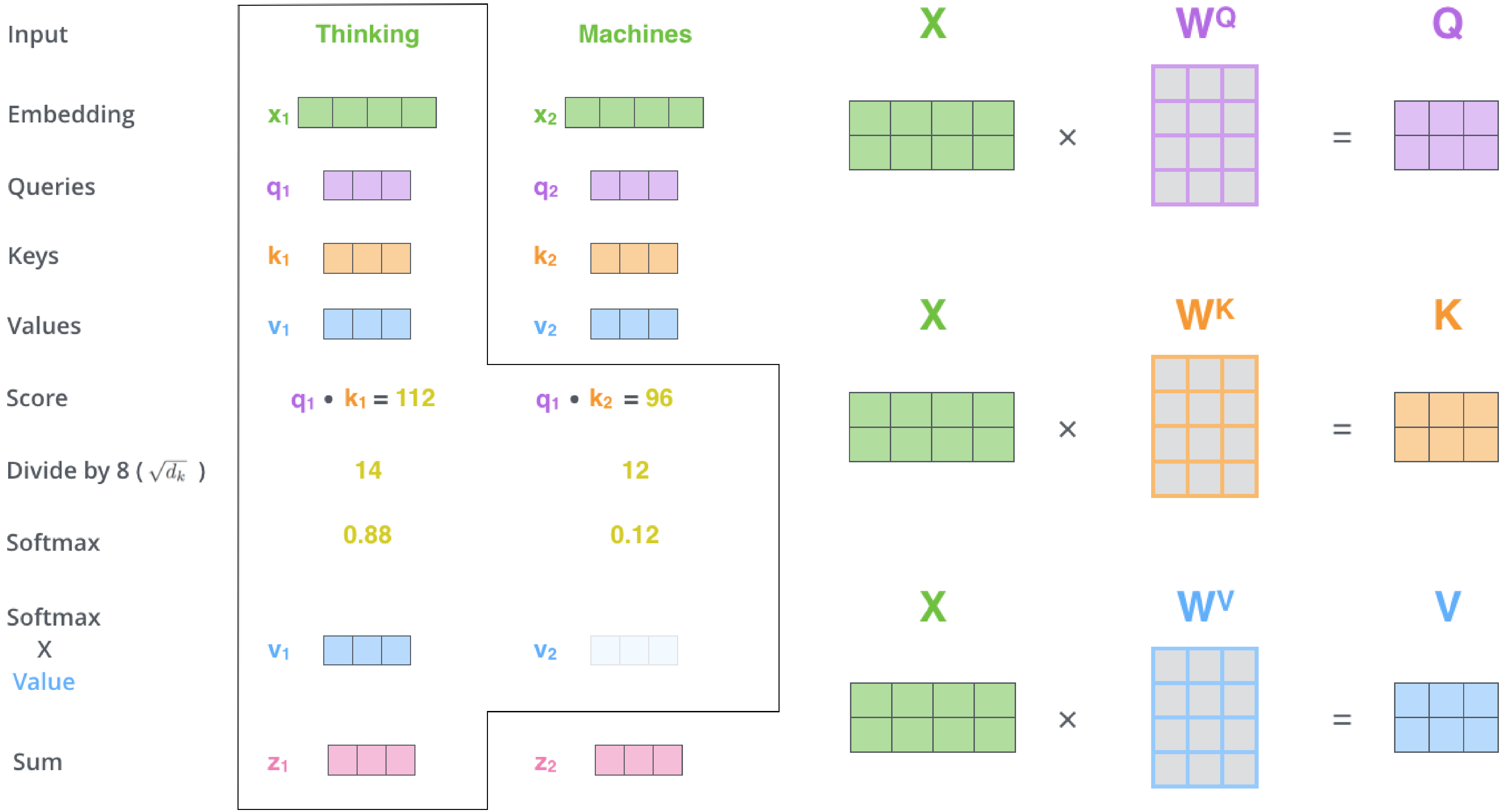
Self-attention



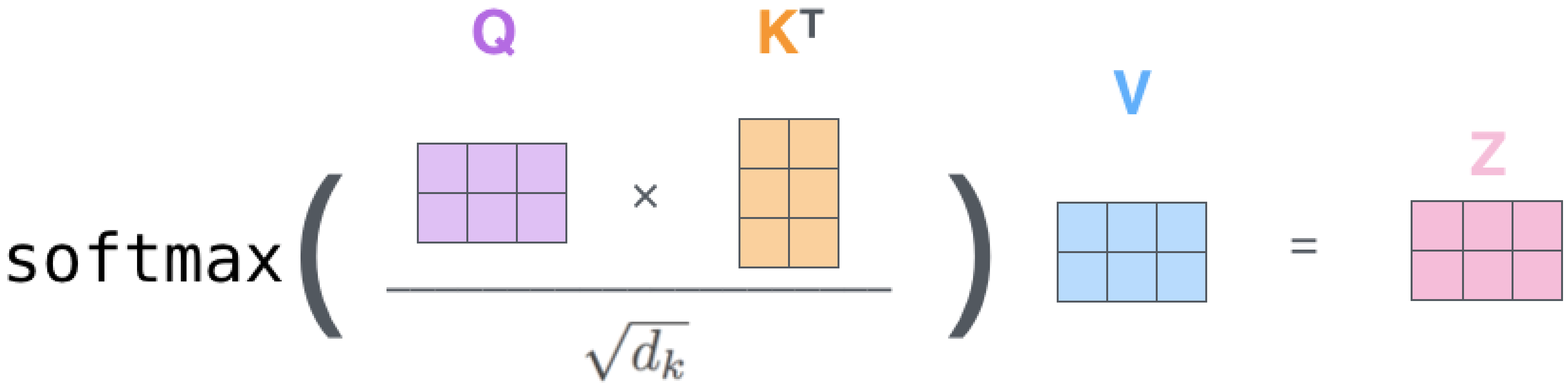
Queries, Keys and Values



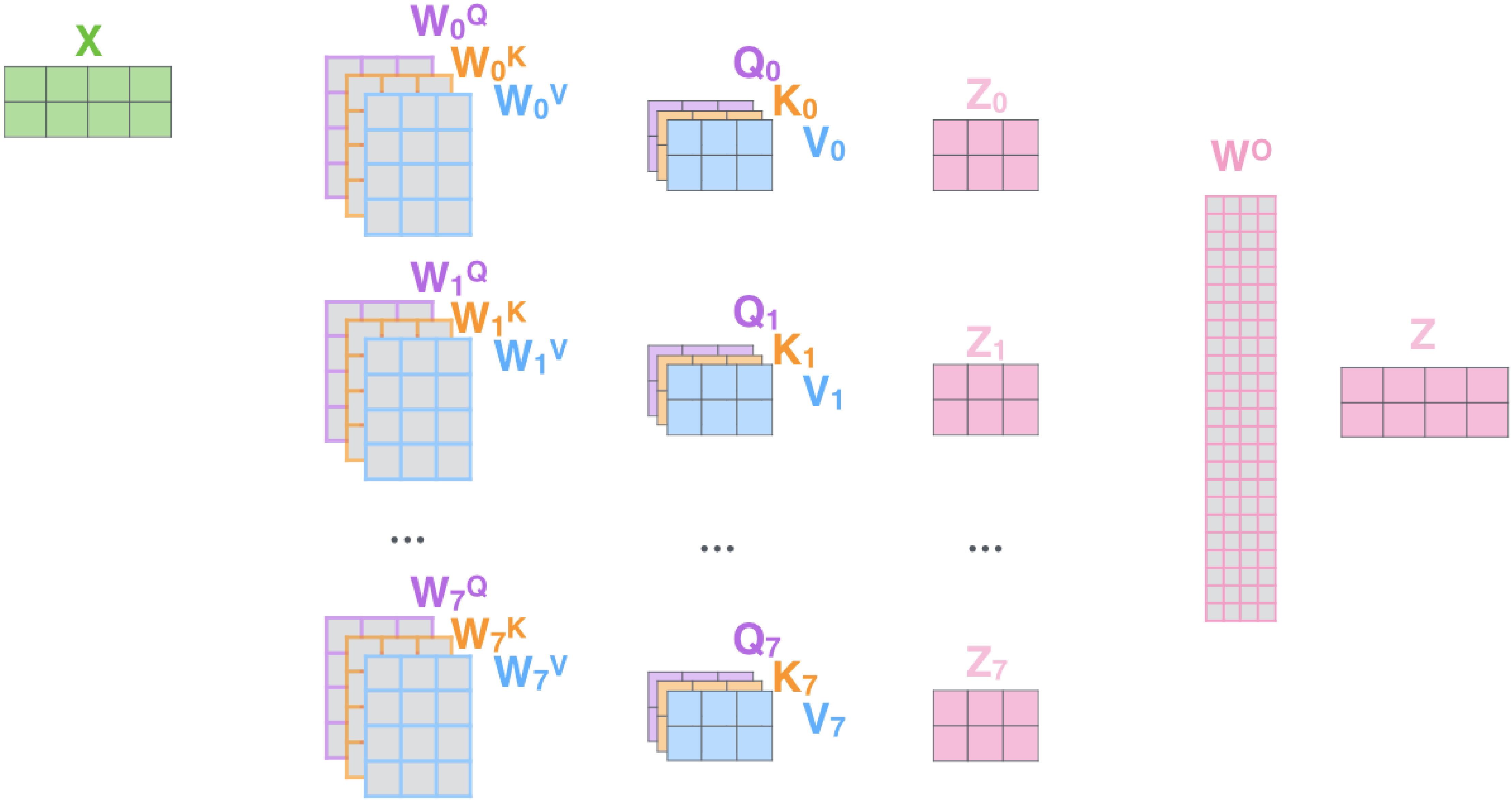
Self attention



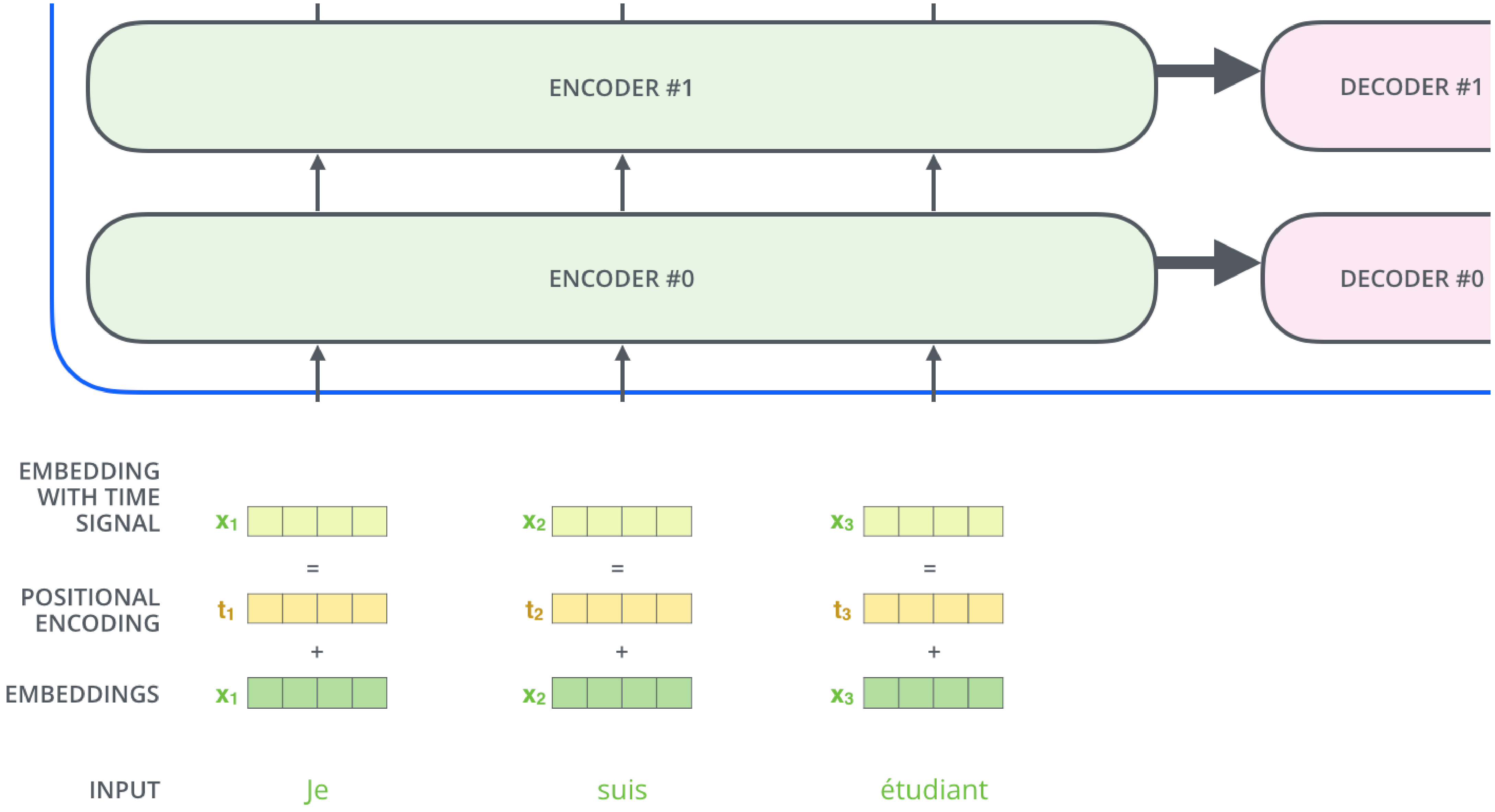
Transformer formula



Multi-headed attention



Positional encoding

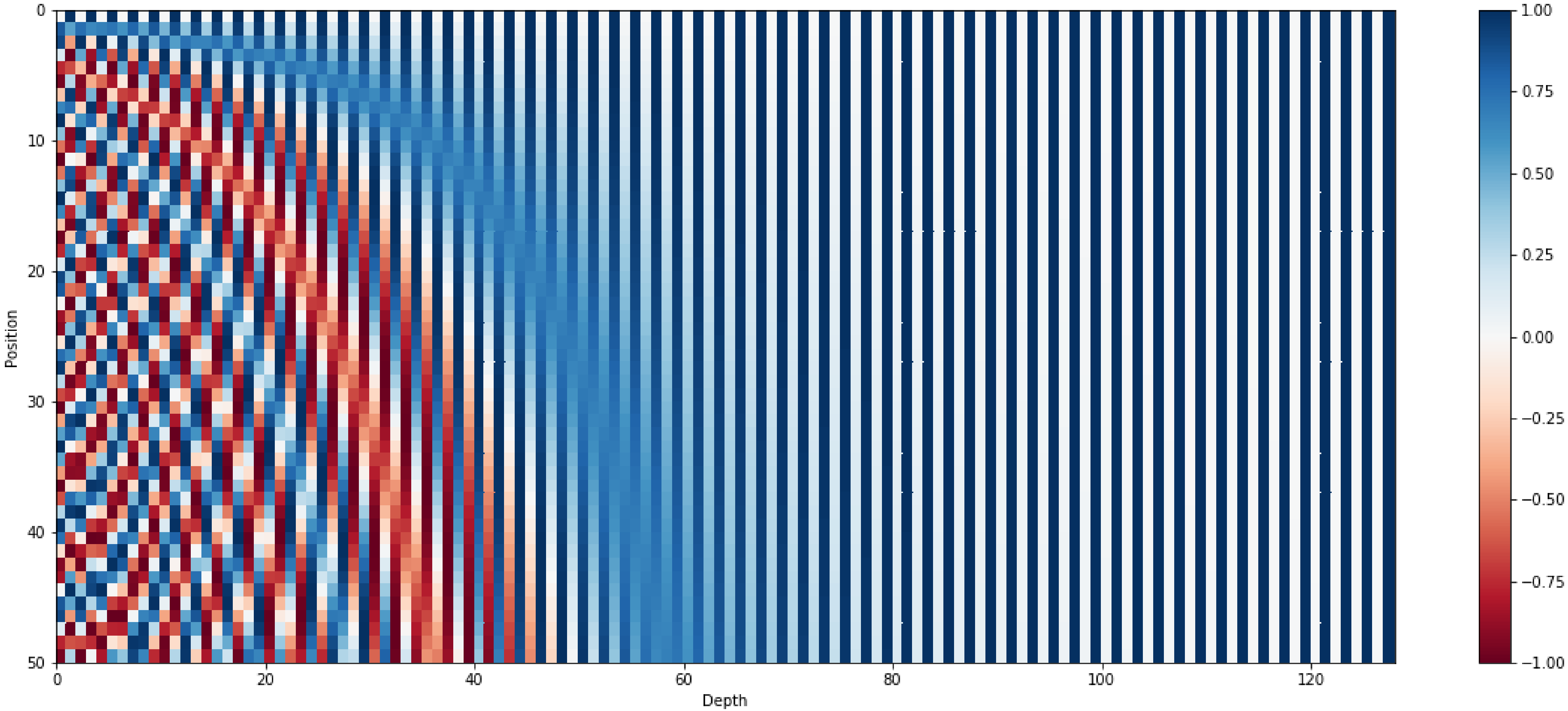


Positional encoding

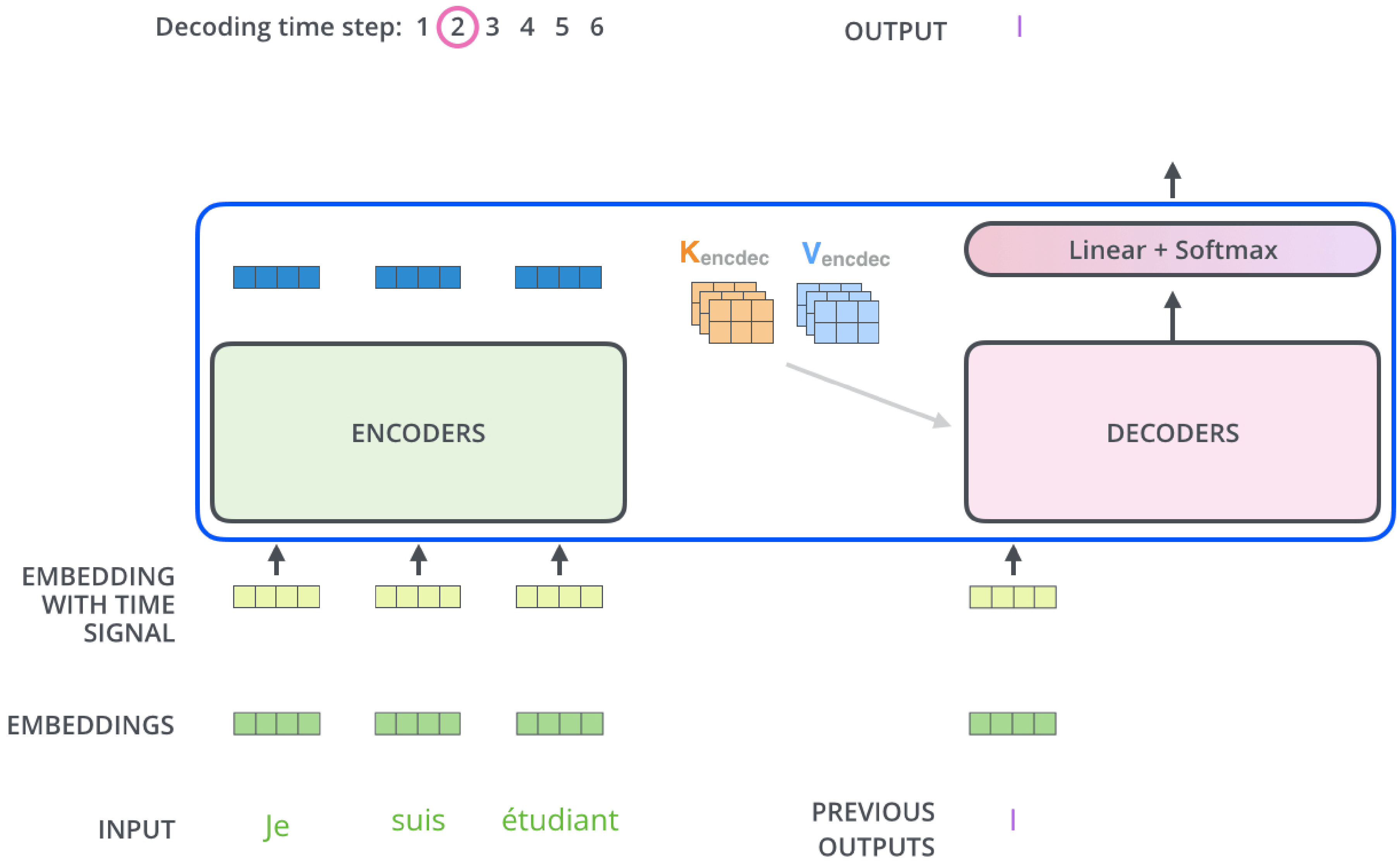
- It should output a unique encoding for each time-step (word's position in a sentence)
- Distance between any two time-steps should be consistent across sentences with different lengths.
- Our model should generalize to longer sentences without any efforts. Its values should be bounded.
- It must be deterministic.

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

Positional encoding



Encoder - decoder

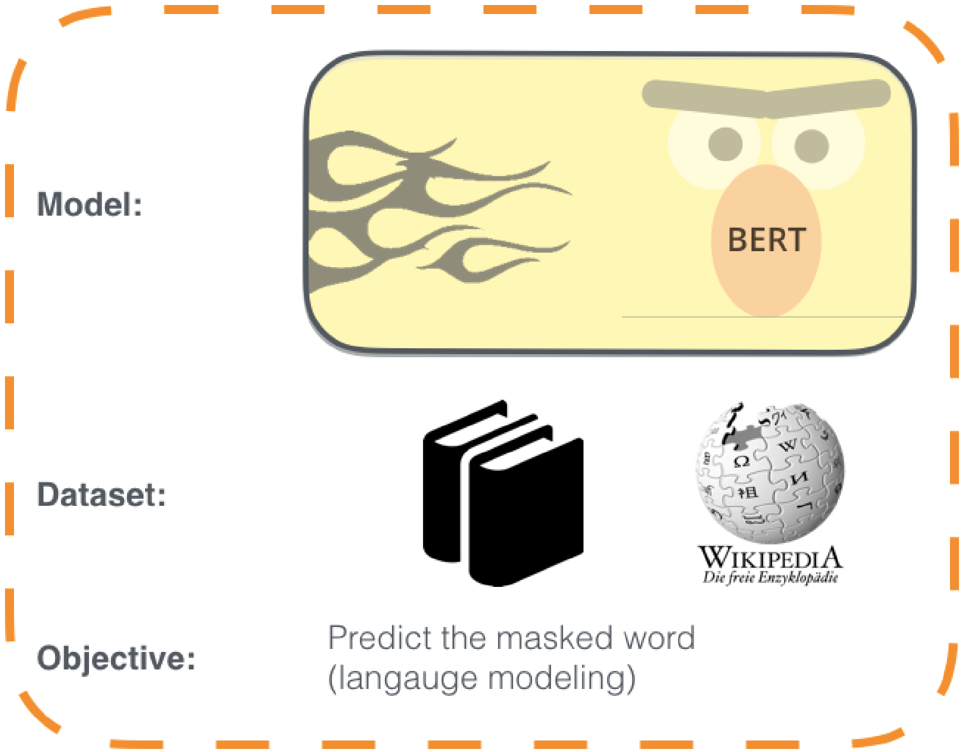


Bidirectional Encoder Representations from Transformers (**BERT**)

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

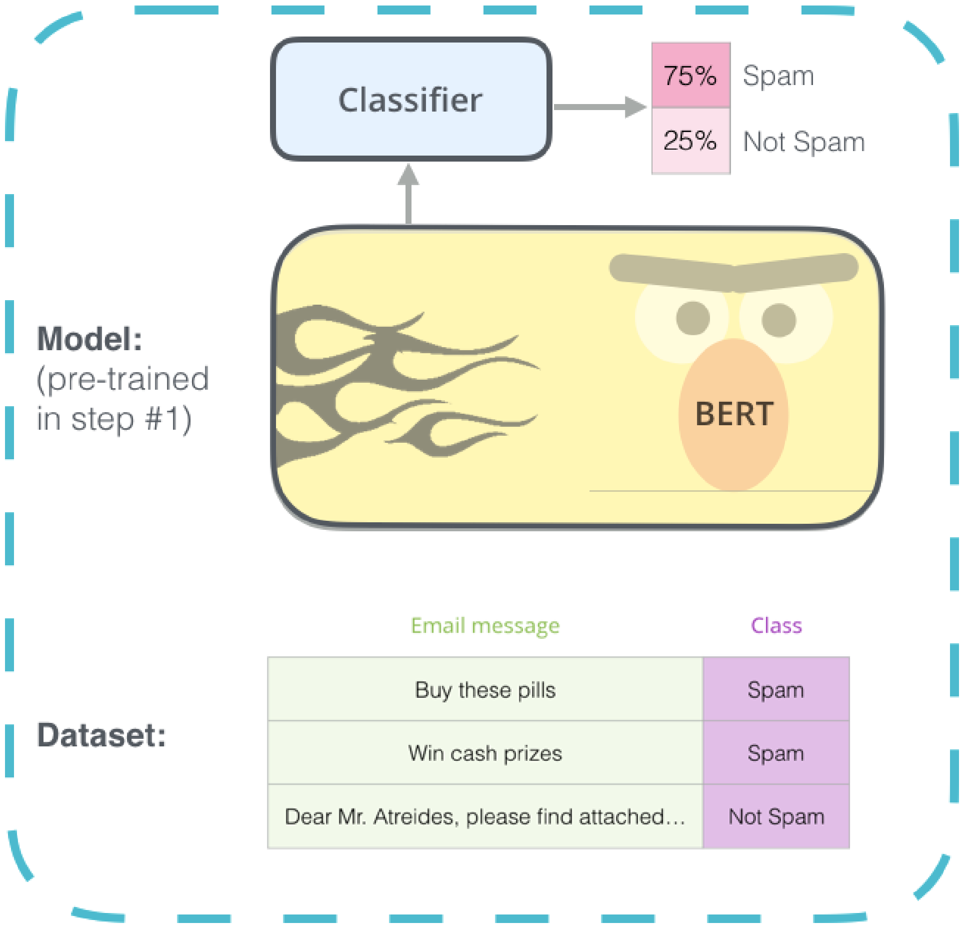
The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step

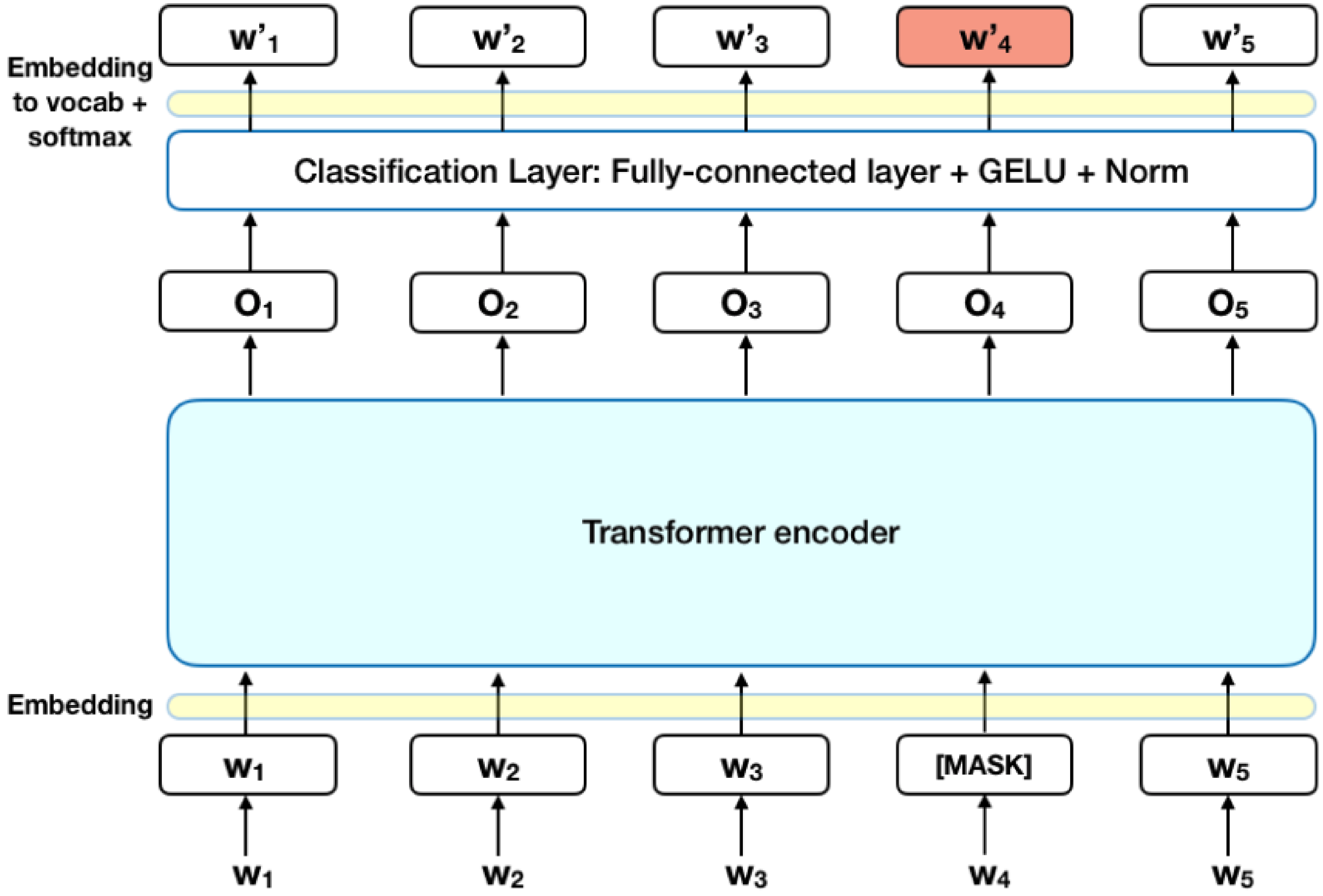


2 - Supervised training on a specific task with a labeled dataset.

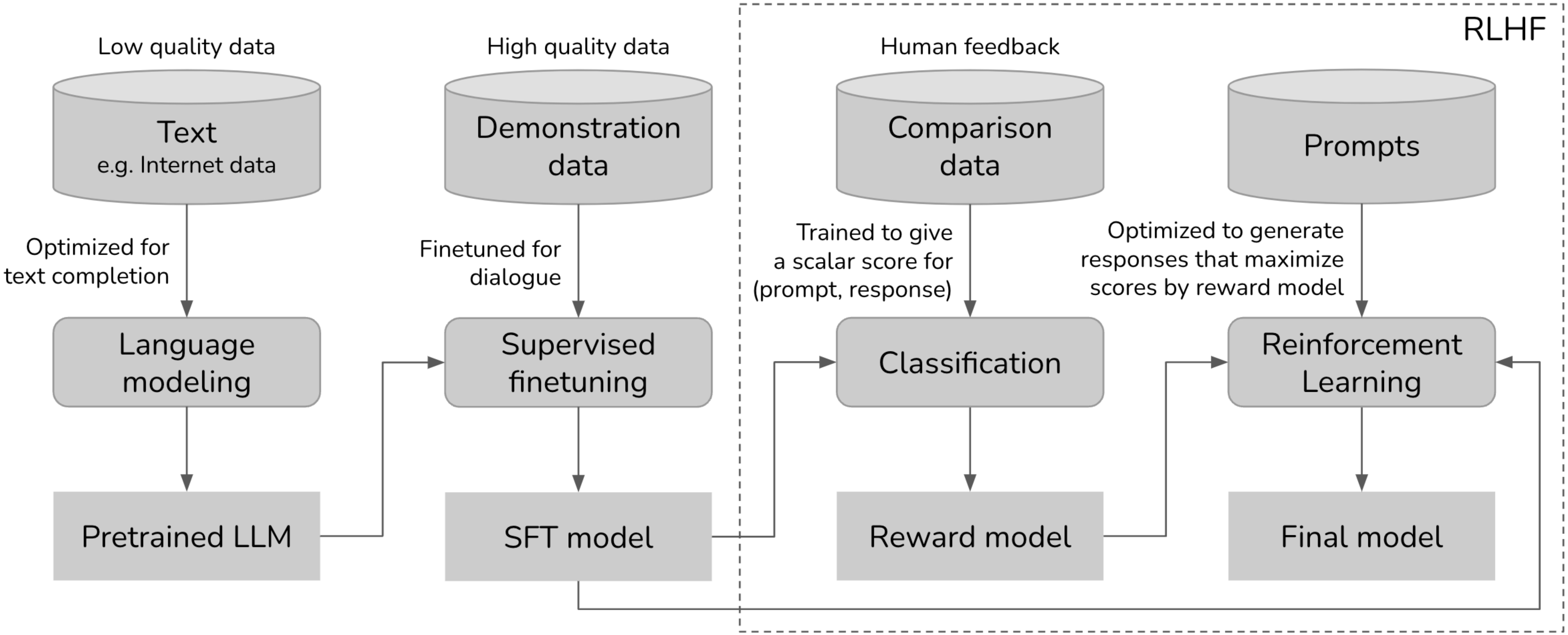
Supervised Learning Step



Bidirectional Encoder Representations from Transformers (**BERT**)



ChatGPT



Scale May '23	**>1 trillion** tokens	**10K - 100K** (prompt, response)	**100K - 1M** comparisons (prompt, winning_response, losing_response)	**10K - 100K** prompts
Examples **Bolded**: open sourced	GPT-x, Gopher, **Falcon**, LLaMa, **Pythia**, **Bloom**, **StableLM**	**Dolly-v2, Falcon-Instruct**		InstructGPT, ChatGPT, Claude, **StableVicuna**