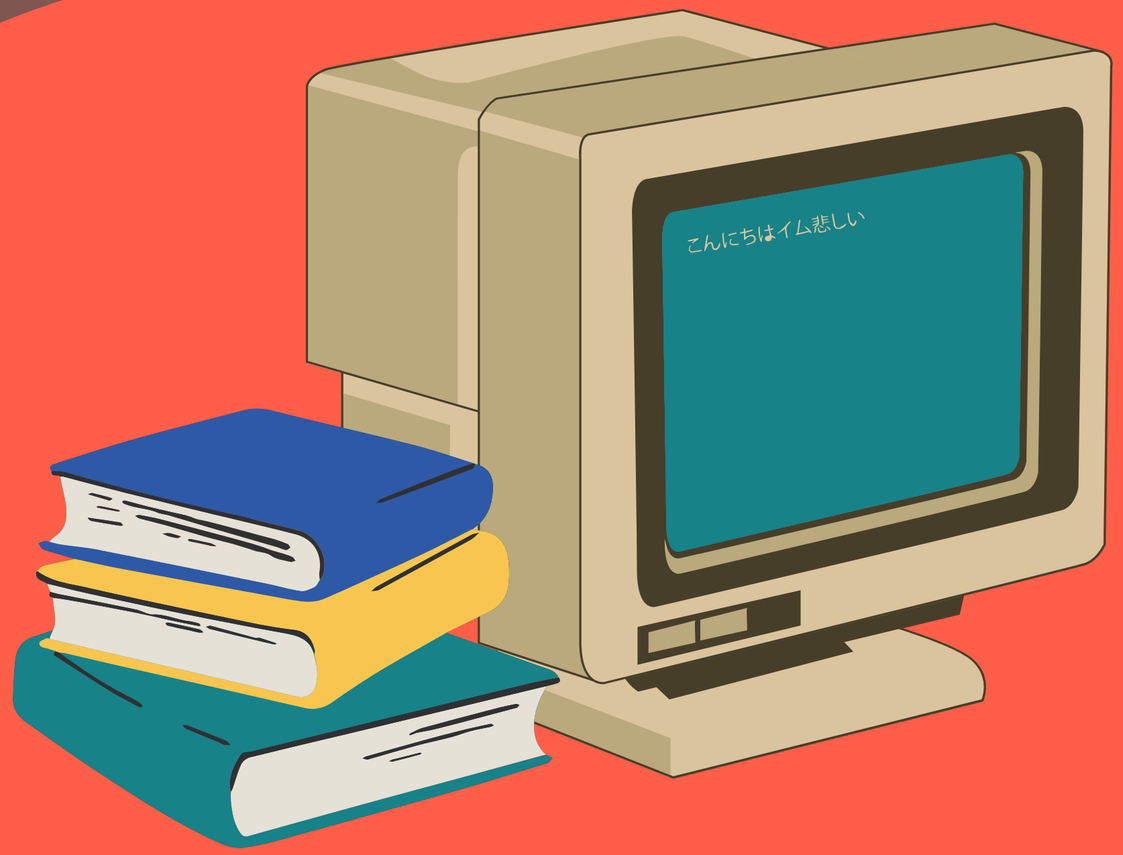
Natural Language Processing

Second School on Data Science and Machine Learning, São Paulo, Brazil

Felipe Serras

IME-USP/C4AI



Agenda

1. **What** is Natural Language Processing and Computational Linguistics?

2. **Why** we need Natural Language Processing and Computational Linguistics

3. **How** we perform NLP and Computational Linguistics?

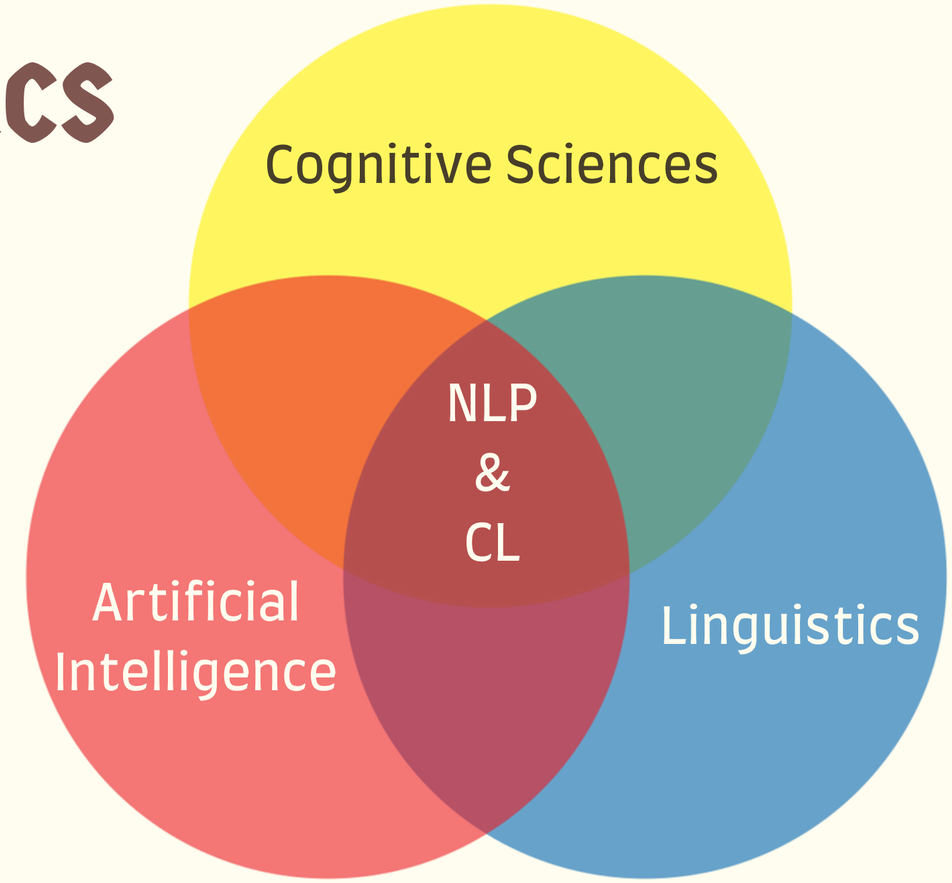
4. **Which** models are state-of-the-art in NLP and Computational Linguistics?

1.What

is Natural Language Processing and Computational Linguistics?

What is Natural Language Processing and Computational Linguistics

- It is a field of scientific and technological research;
- How can computational models be used to process natural language data and better understand the functioning of natural language?
- It is a multidisciplinary field;
- It originated from the attempt to create computer programs to translate texts from Russian to English.



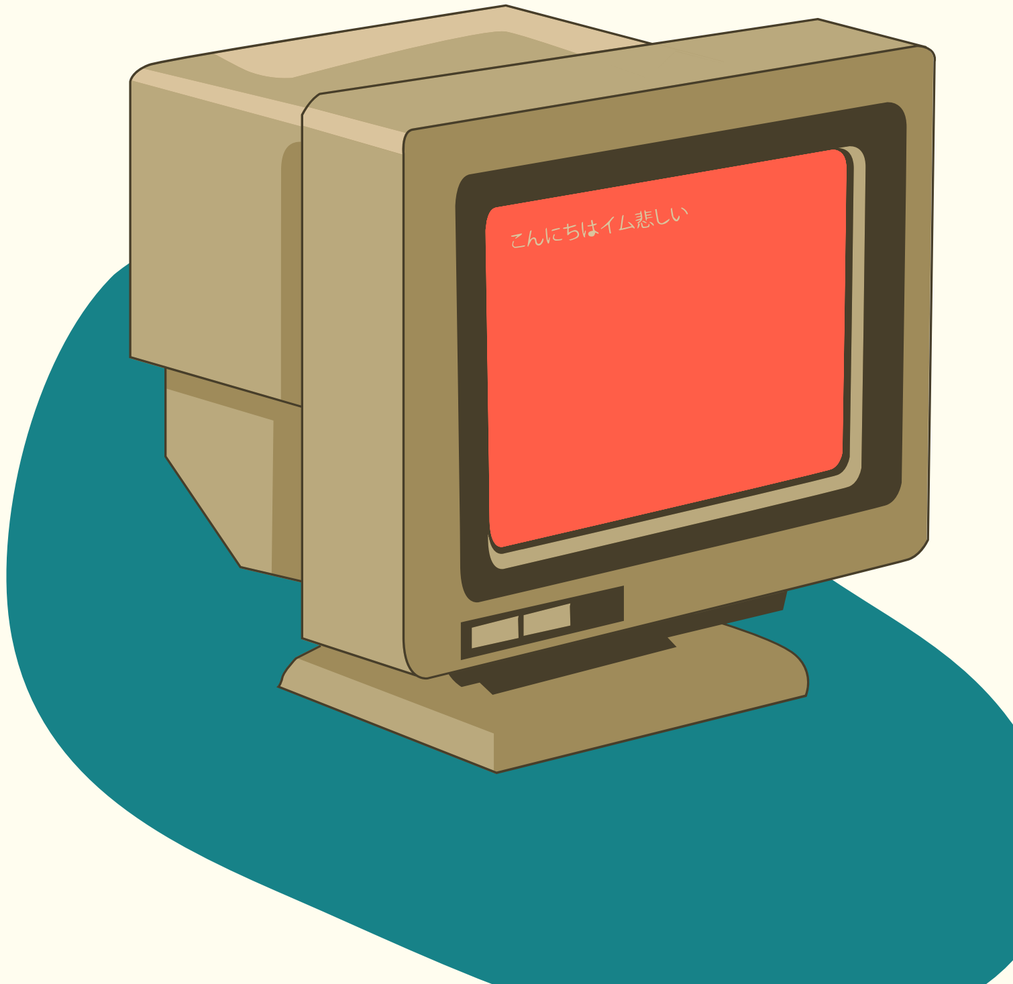
What is the difference between Natural Language Processing and Computational Linguistics

Computational Linguistics is focused on the investigation of human languages and how they function using computational resources.

NLP is focused on the development of computational resources for the accomplishment of tasks using data in human language

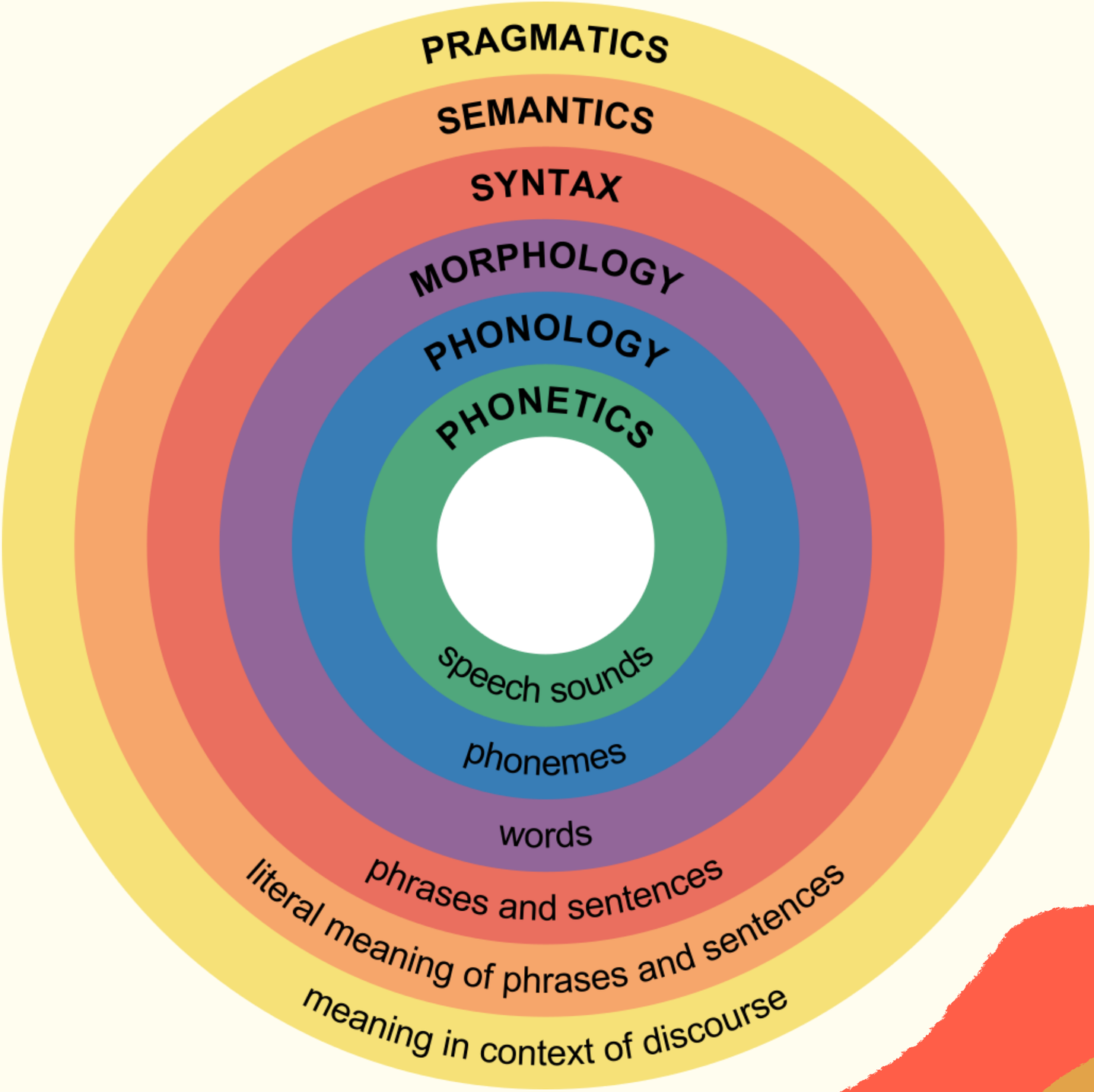
Tasks in Natural Language Processing

- **Translation**
- **Classification**: Sentiment Analysis, Spam Detection, Topic Classification
- **Regression**: Autograding
- **Clustering**: Topic Modeling, Authorship Attribution, Similarity-based Recommendations
- **Tagging**: Named Entity Recognition (NER), Part-of-Speech Tagging
- **Generation**: Conversational Agents, Code Generation



2.Why

do we need Natural Language Processing and Computational Linguistics?


PRAGMATICS
SEMANTICS
SYNTAX
MORPHOLOGY
PHONOLOGY
PHONETICS
speech sounds
phonemes
words
phrases and sentences
literal meaning of phrases and sentences
meaning in context of discourse

Phonetics and Phonology

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2015)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ Bilabial	ɓ Bilabial	ʼ Examples:
ǀ Dental	ɗ Dental/alveolar	pʼ Bilabial
ǃ (Post)alveolar	ʄ Palatal	tʼ Dental/alveolar
ǂ Palatoalveolar	ɠ Velar	kʼ Velar
ǁ Alveolar lateral	ʛ Uvular	sʼ Alveolar fricative

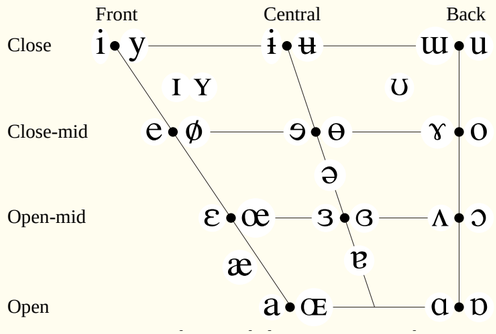
OTHER SYMBOLS

ʍ Voiceless labial-velar fricative
w Voiced labial-velar approximant
ɥ Voiced labial-palatal approximant
ʜ Voiceless epiglottal fricative
ʢ Voiced epiglottal fricative
ʡ Epiglottal plosive

ɕ ʑ Alveolo-palatal fricatives
ɺ Voiced alveolar lateral flap
ɧ Simultaneous ʃ and x

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. t͡s k͡p

VOWELS



OTHER SYMBOLS

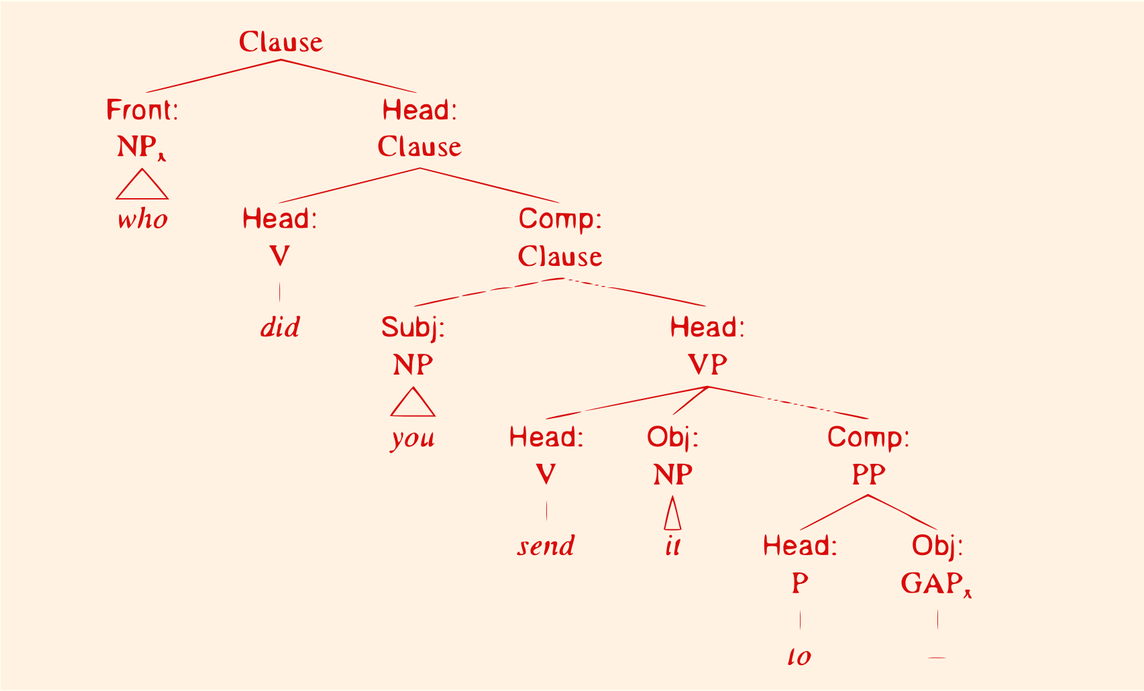
ʍ	Voiceless labial-velar fricative	ɕ ʑ	Alveolo-palatal fricatives
w	Voiced labial-velar approximant	ɺ	Voiced alveolar lateral flap
ɥ	Voiced labial-palatal approximant	ɧ	Simultaneous ʃ and x
ʜ	Voiceless epiglottal fricative		
ʢ	Voiced epiglottal fricative		
ʡ	Epiglottal plosive		

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. t͡s k͡p

DIACRITICS Some diacritics may be placed above a symbol with a descender, e.g. ŋ̊

̥	Voiceless	n̥	d̥	̤	Breathy voiced	b̤	a̤	̪	Dental	t̪	d̪
̬	Voiced	s̬	t̬	̰	Creaky voiced	b̰	a̰	̺	Apical	t̺	d̺
ʰ	Aspirated	tʰ	dʰ	̼	Linguolabial	t̼	d̼	̻	Laminal	t̻	d̻
̹	More rounded	ɔ̹		ʷ	Labialized	tʷ	dʷ	̃	Nasalized	ẽ	
̜	Less rounded	ɔ̜		ʲ	Palatalized	tʲ	dʲ	ⁿ	Nasal release	dⁿ	
̟	Advanced	u̟		ˠ	Velarized	tˠ	dˠ	ˡ	Lateral release	dˡ	
̠	Retracted	e̠		ˤ	Pharyngealized	tˤ	dˤ	̚	No audible release	d̚	
̈	Centralized	ë		̴	Velarized or pharyngealized	ɫ					
̽	Mid-centralized	e̽		̝	Raised	e̝	(ɪ̝ = voiced alveolar fricative)				
̩	Syllabic	n̩		̞	Lowered	e̞	(β̞ = voiced bilabial approximant)				
̯	Non-syllabic	e̯		̘	Advanced Tongue Root	e̘					
˞	Rhoticity	ə˞	a˞	̙	Retracted Tongue Root	e̙					

Syntax

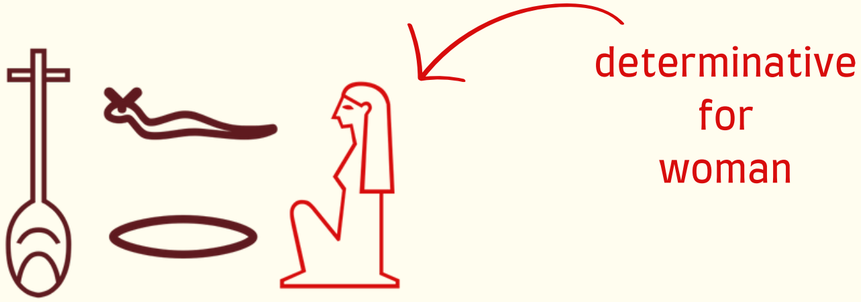


NOUN	ADJECTIVE	DETERMINER
VERB	ADVERB	CONJUNCTION
NUMERAL	PRONOUN	INTERJECTION
ADPOSITION	PARTICLE	...

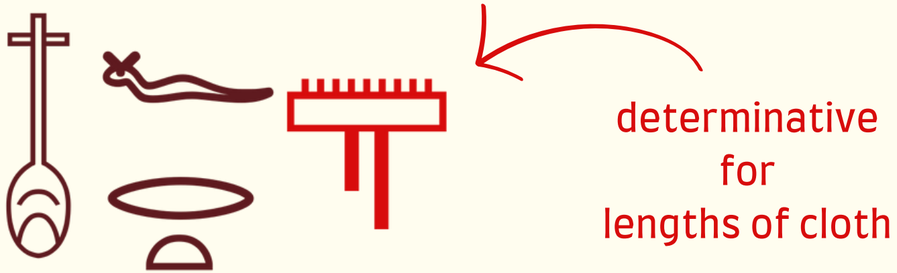
Morphology

- Derivation
  - break → breakable → unbreakable (English)
- Inflection
  - Portuguese:
    - Eu falo, tu falas, nós falamos
  - Spanish
    - Yo hablo, tu hablas, nosotros hablamos.

- Egyptian:



young woman of marriageable age

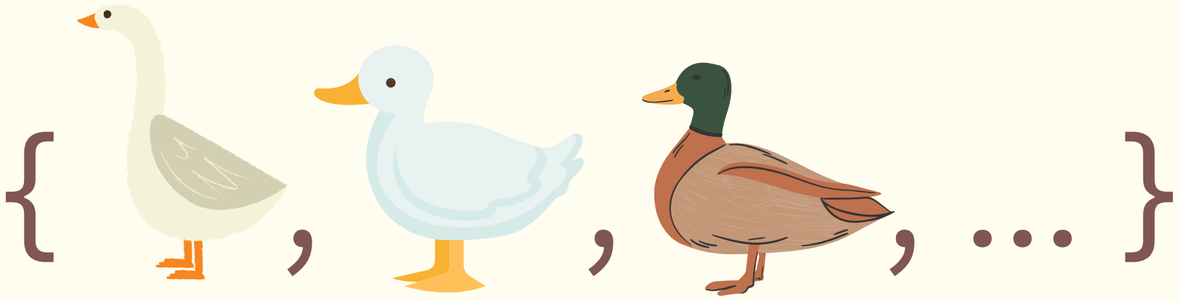


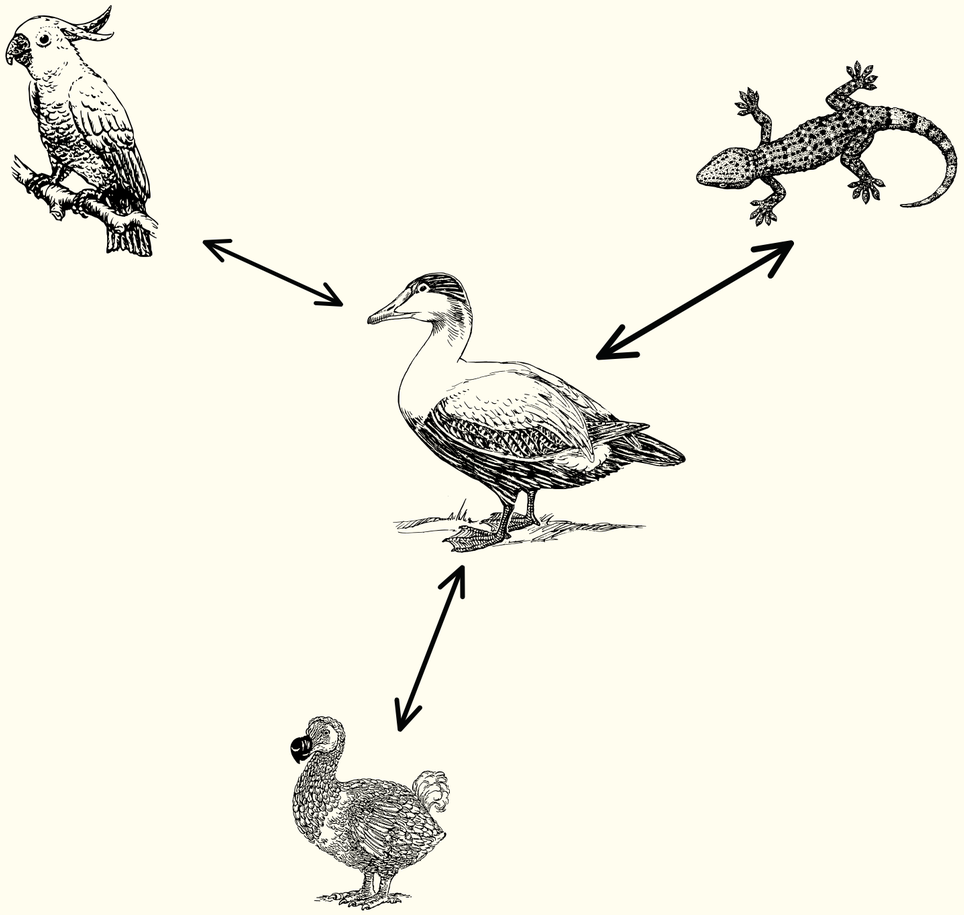
clothing

Semantics

- What is the meaning of “meaning”?
- What is the meaning of “duck”?

Is it the set of all possible ducks?



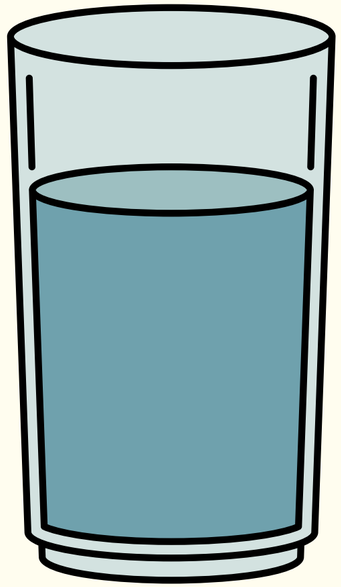


Is it the distance to a prototypical duck?

Pragmatics

- How meaning changes with context?
  - Irony, Implicature
- What is the relationship between meaning and context?
  - Distributional Hypothesis

Can you pass me the water?



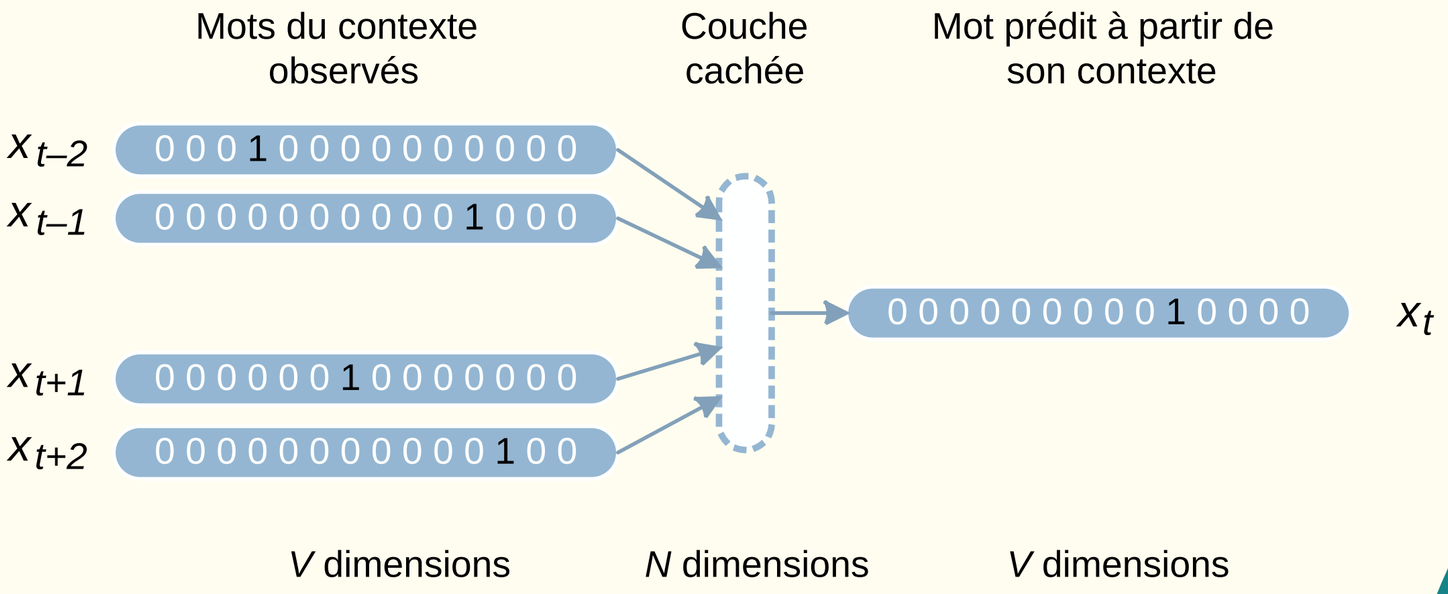
I am fully capable of passing the glass of water. Thanks for asking!

3.How

do we perform Natural Language Processing and Computational Linguistics?

The Problem of Representation

- How to represent data in human language in a format that allows us to perform mathematical operations?
- Vector Semantics → Embeddings

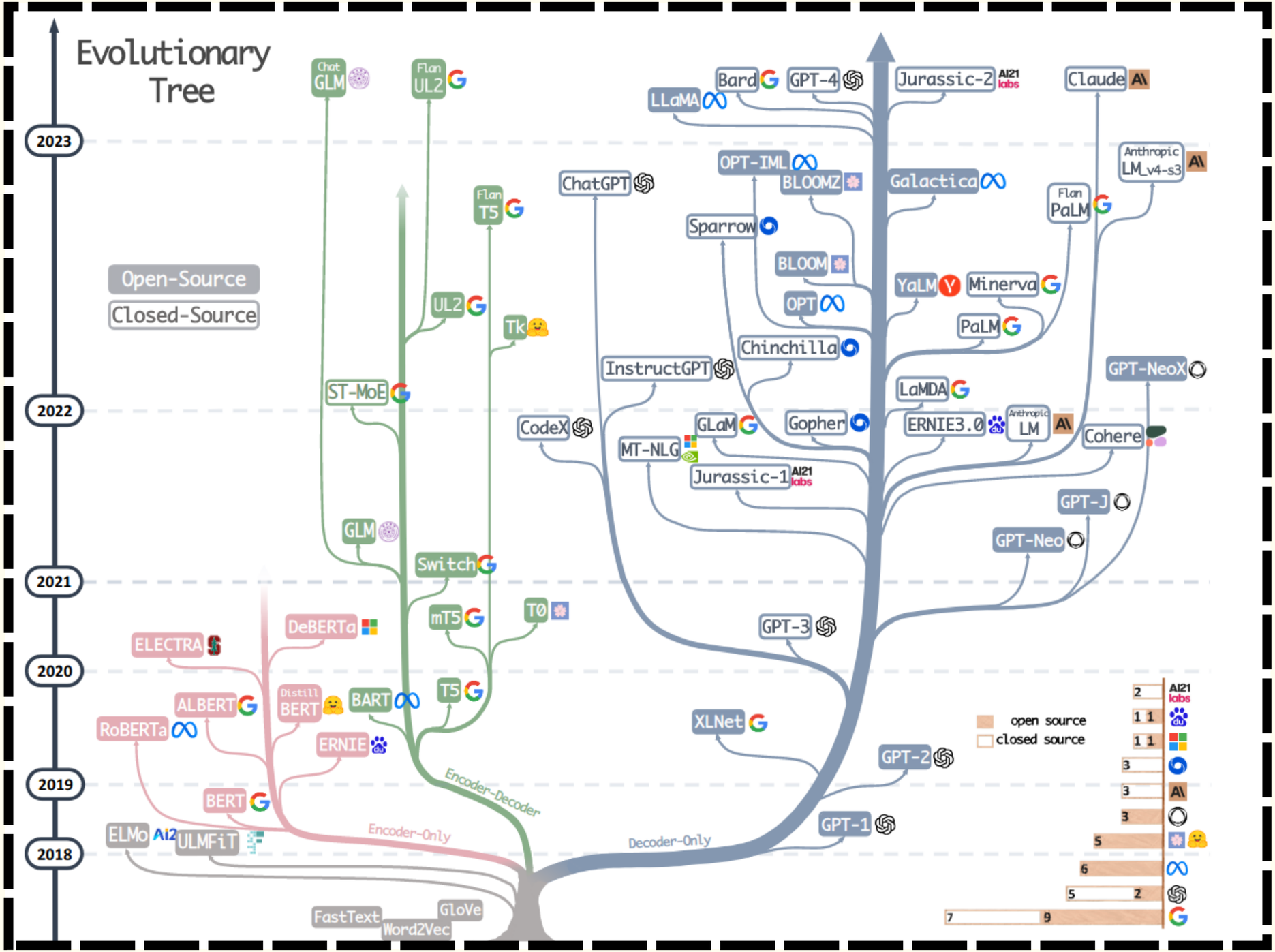


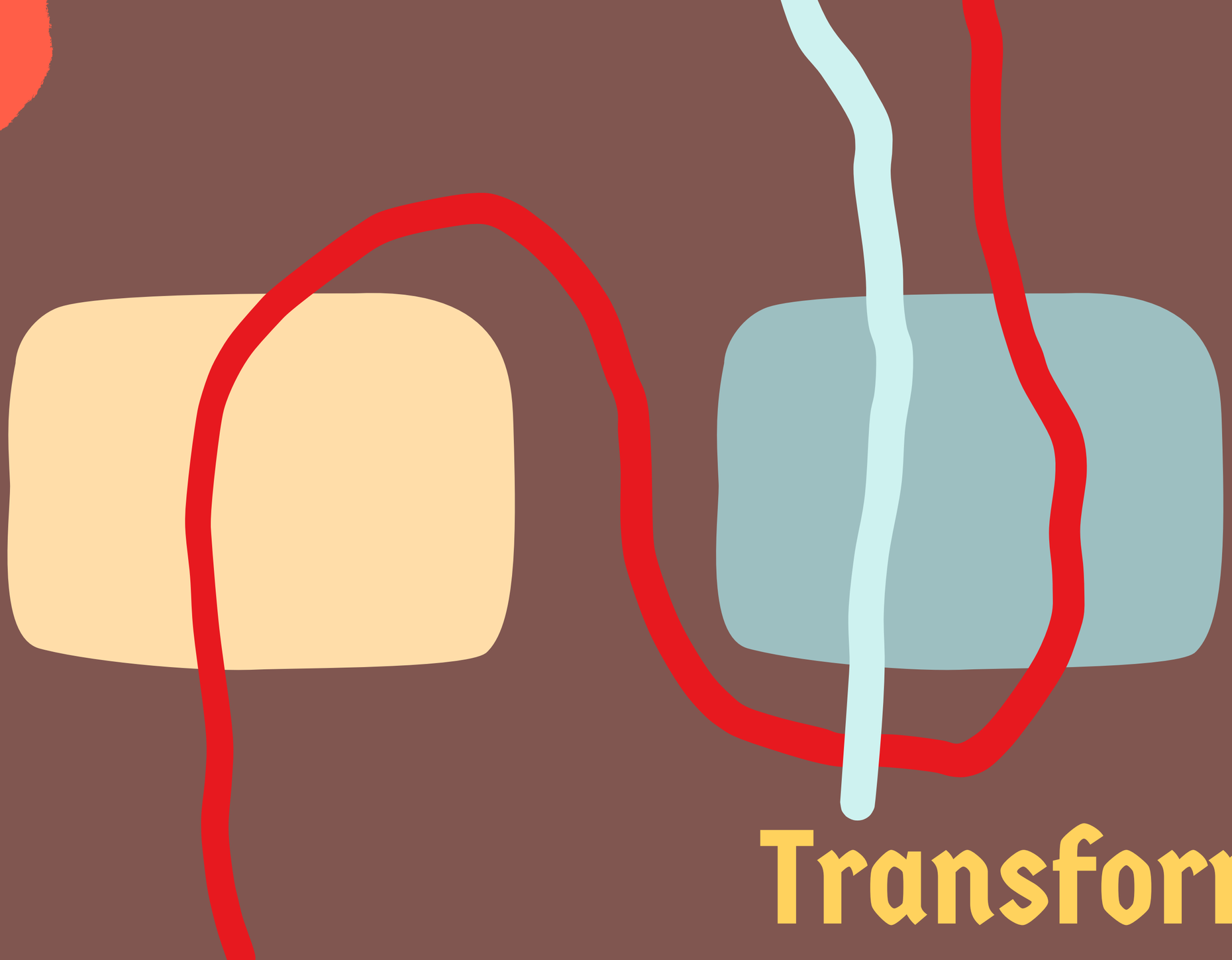
Several Paradigms

- Knowledge Representation
  - Ontologies, Logic Programming, Automated Theorem Proving
- Probabilistic Models
  - Bayesian Networks, Markov Chains, ...
- Neural Networks
  - FNN, CNN, RNN, Attention
- Neuro-symbolic

4.Which

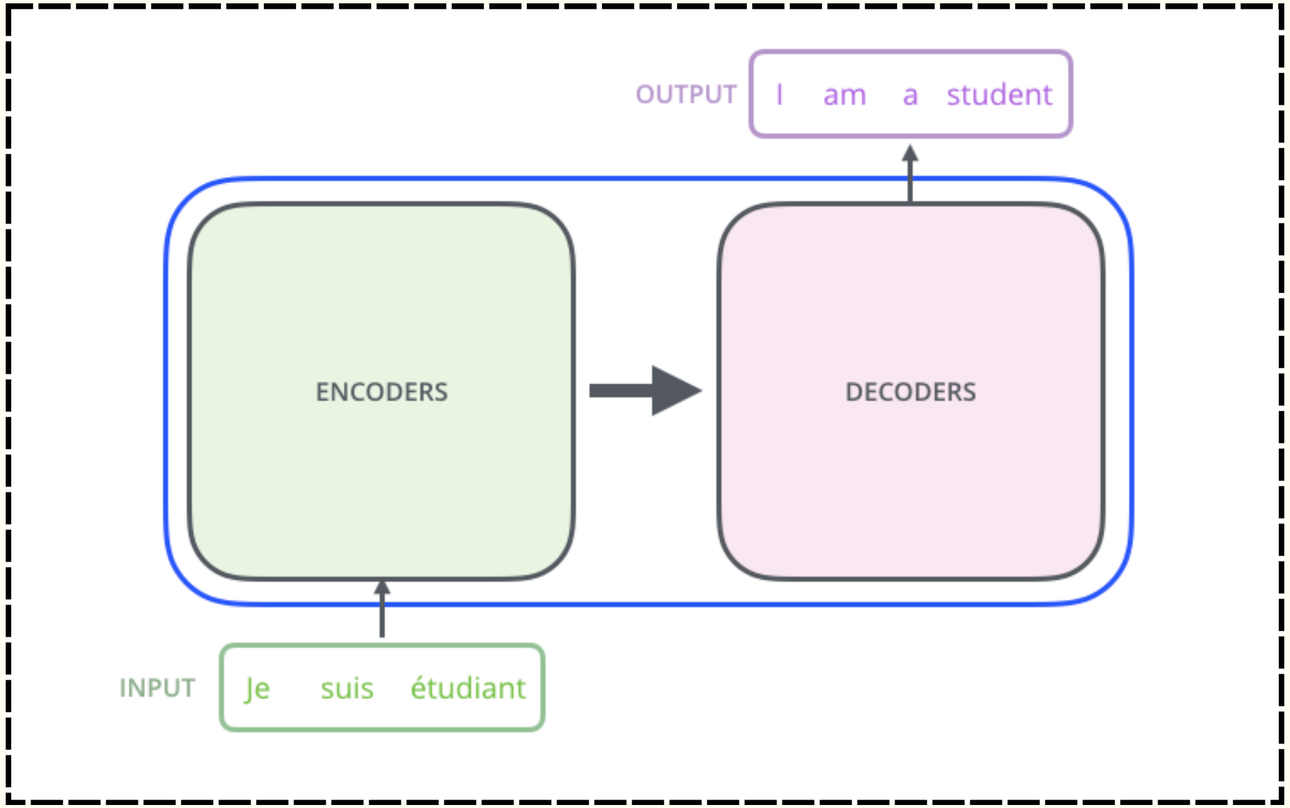
models are state-of-the-art in Natural Language Processing and Computational Linguistics?

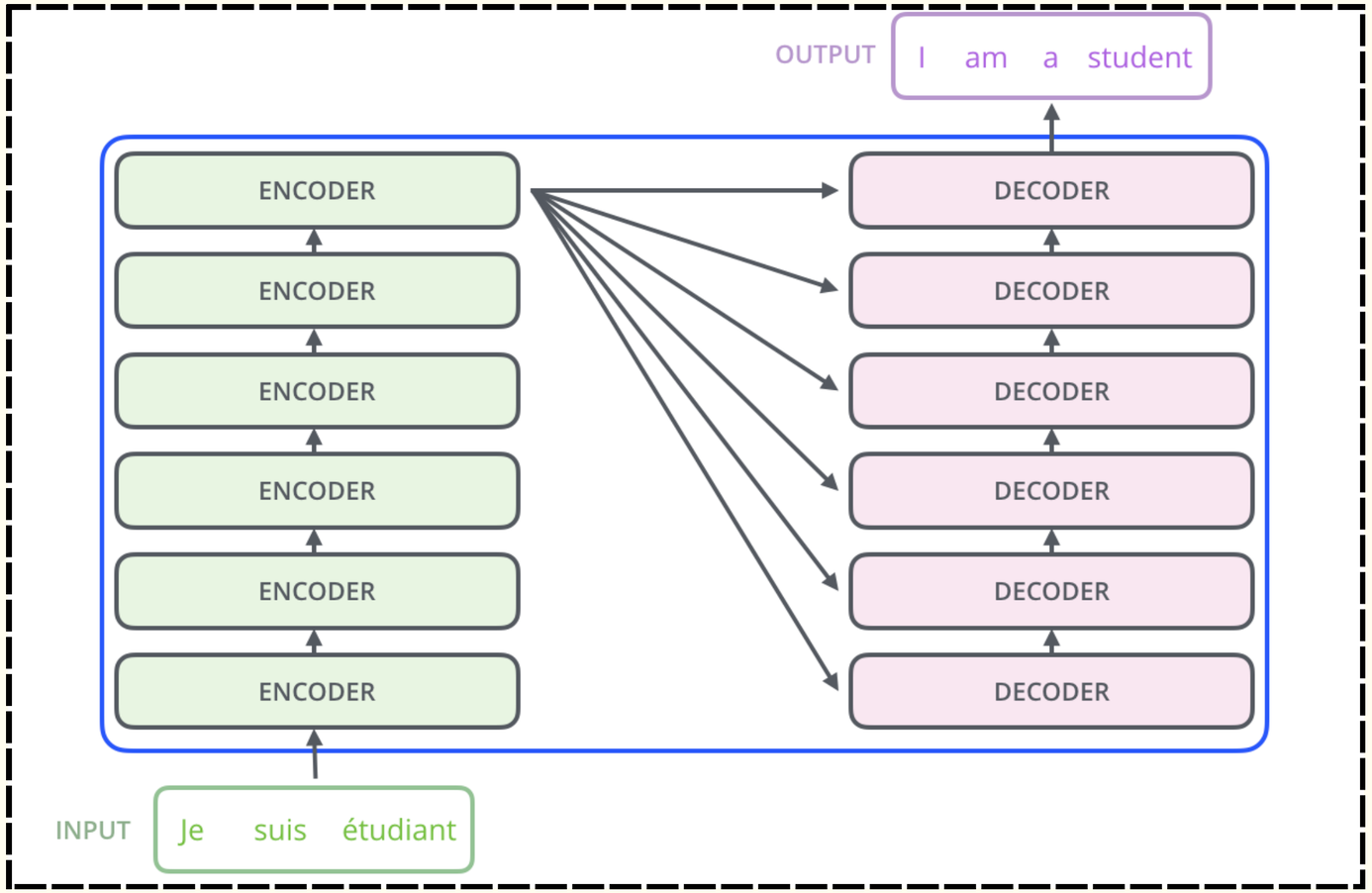

Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Chat GLM
Flan UL2
Flan T5
UL2
Tk
ST-MoE
GLM
Switch
mT5
T0
T5
BART
Distill BERT
ELECTRA
DeBERTa
ALBERT
RoBERTa
ERNIE
BERT
ELMo
ULMFiT
FastText
Word2Vec
GloVe
Encoder-Decoder
Encoder-Only
Decoder-Only
LLaMA
Bard
GPT-4
Jurassic-2
Claude
OPT-IML
BLOOMZ
Galactica
ChatGPT
Sparrow
Flan PaLM
Anthropic LM_v4-s3
BLOOM
YaLM
Minerva
OPT
PaLM
Chinchilla
InstructGPT
GPT-NeoX
LaMDA
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GPT-Neo
GPT-3
XLNet
GPT-2
GPT-1
open source
closed source
2
1 1
1 1
3
3
3
5
6
5 2
7 9

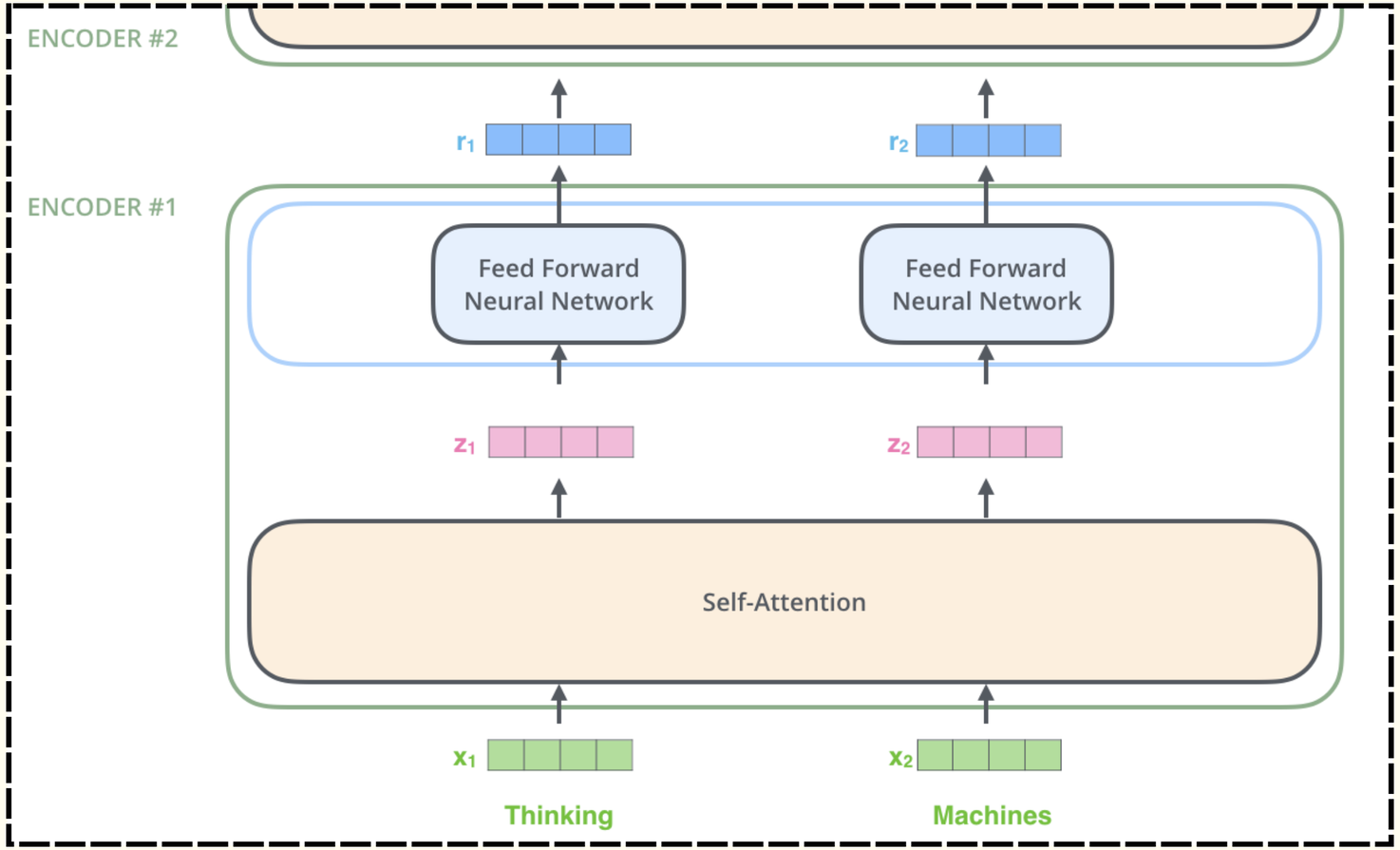


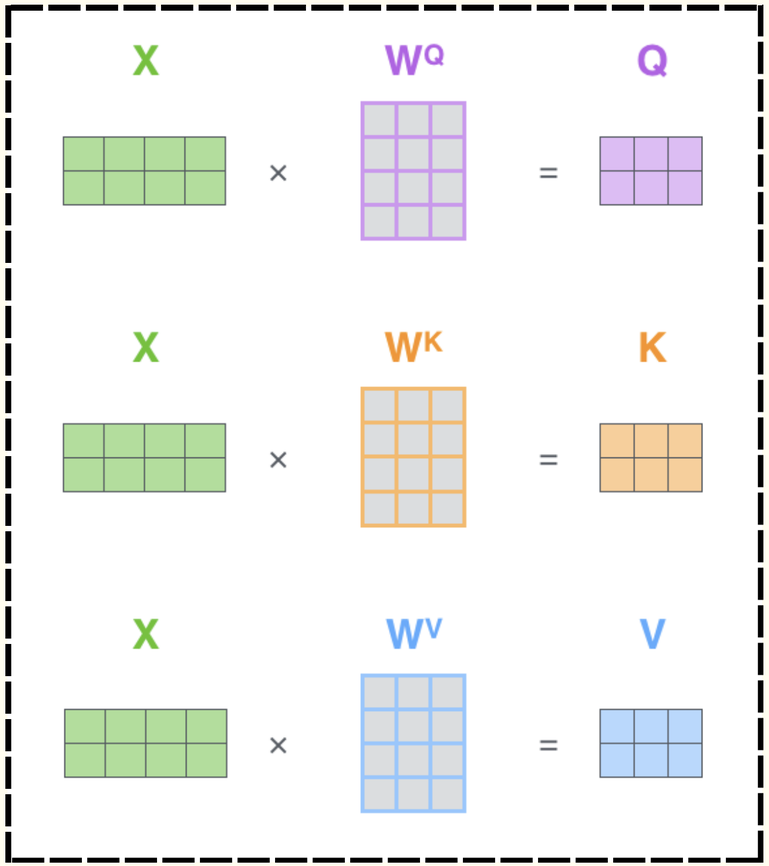
Transformer

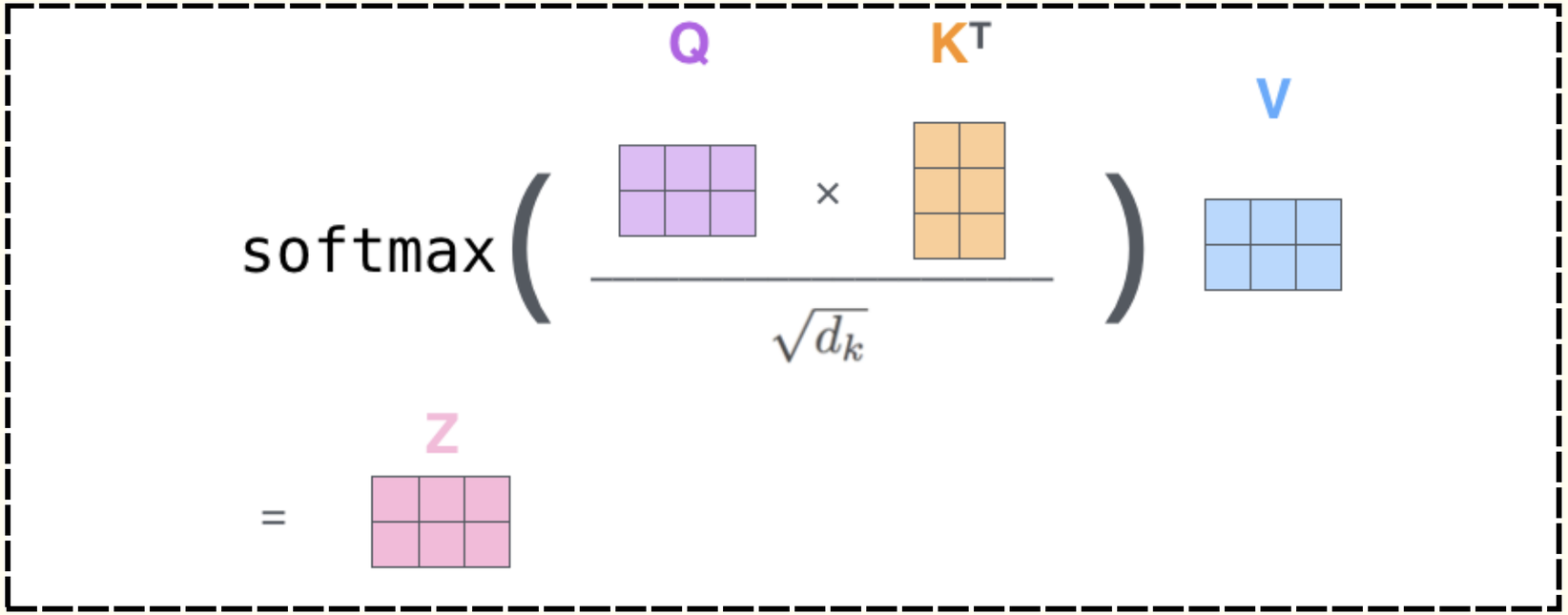

INPUT
Je suis étudiant
THE TRANSFORMER
OUTPUT
I am a student

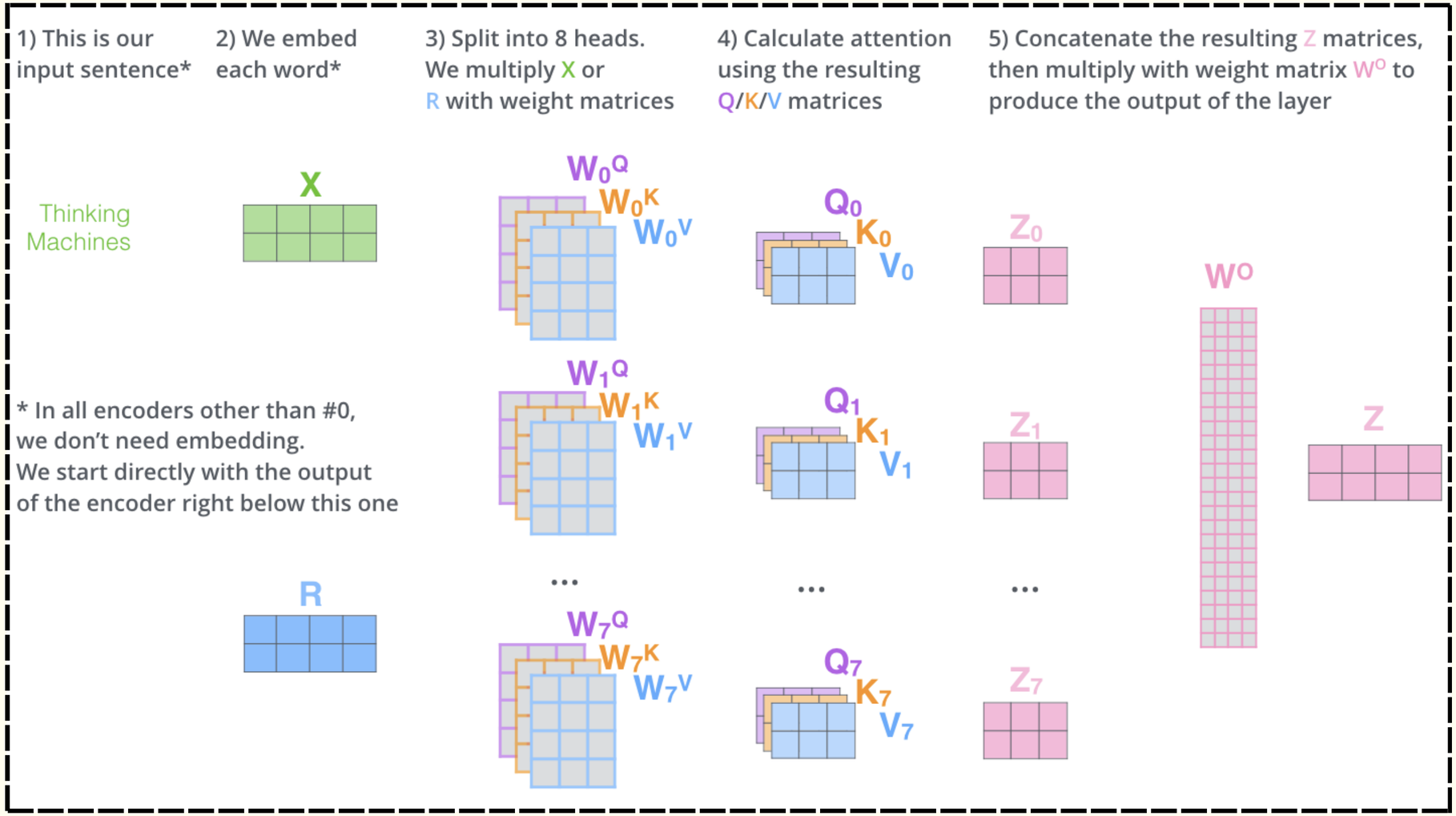

OUTPUT
I am a student
ENCODERS
DECODERS
INPUT
Je suis étudiant

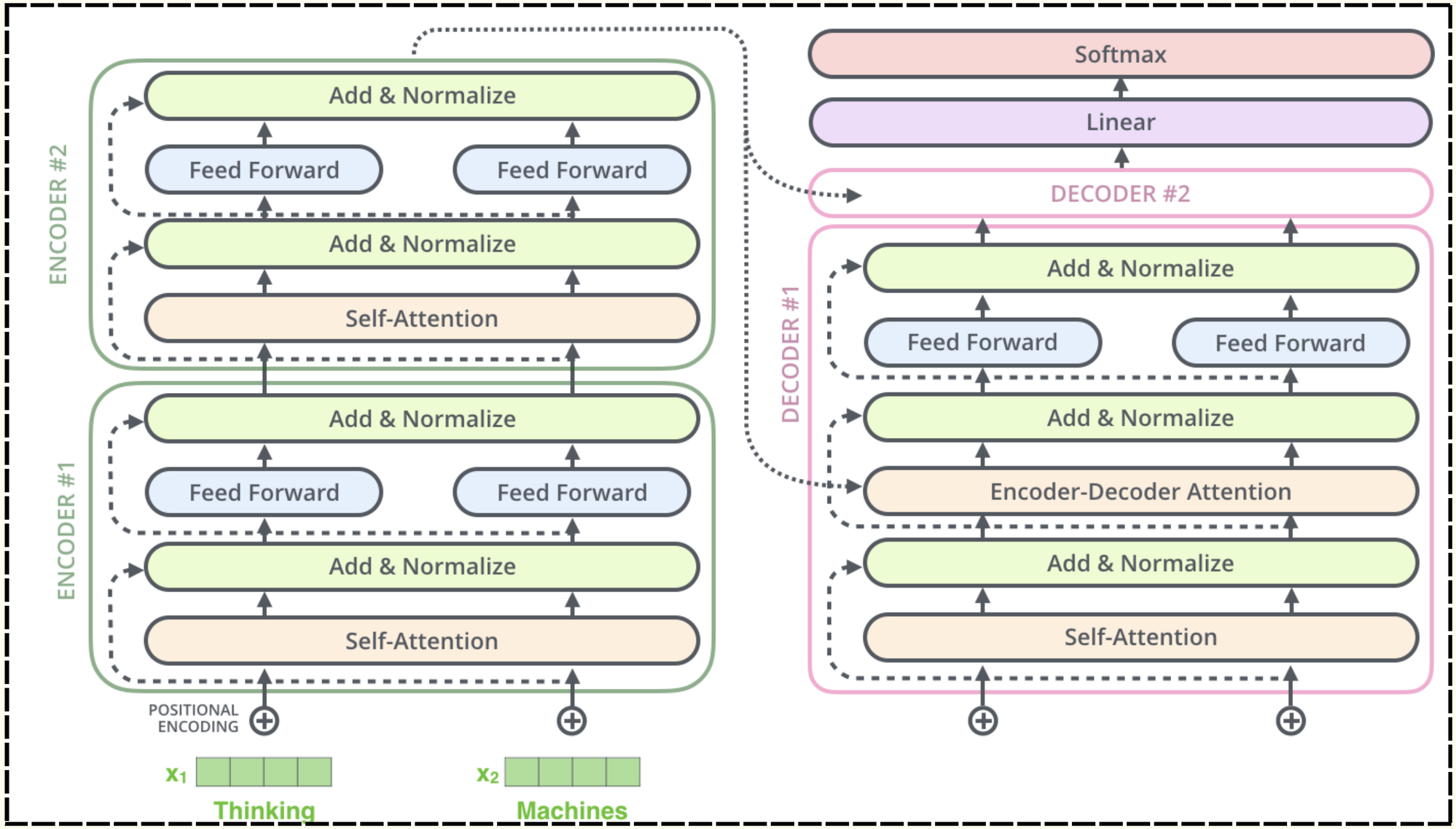

OUTPUT
I am a student
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
DECODER
DECODER
DECODER
DECODER
DECODER
DECODER
INPUT
Je suis étudiant

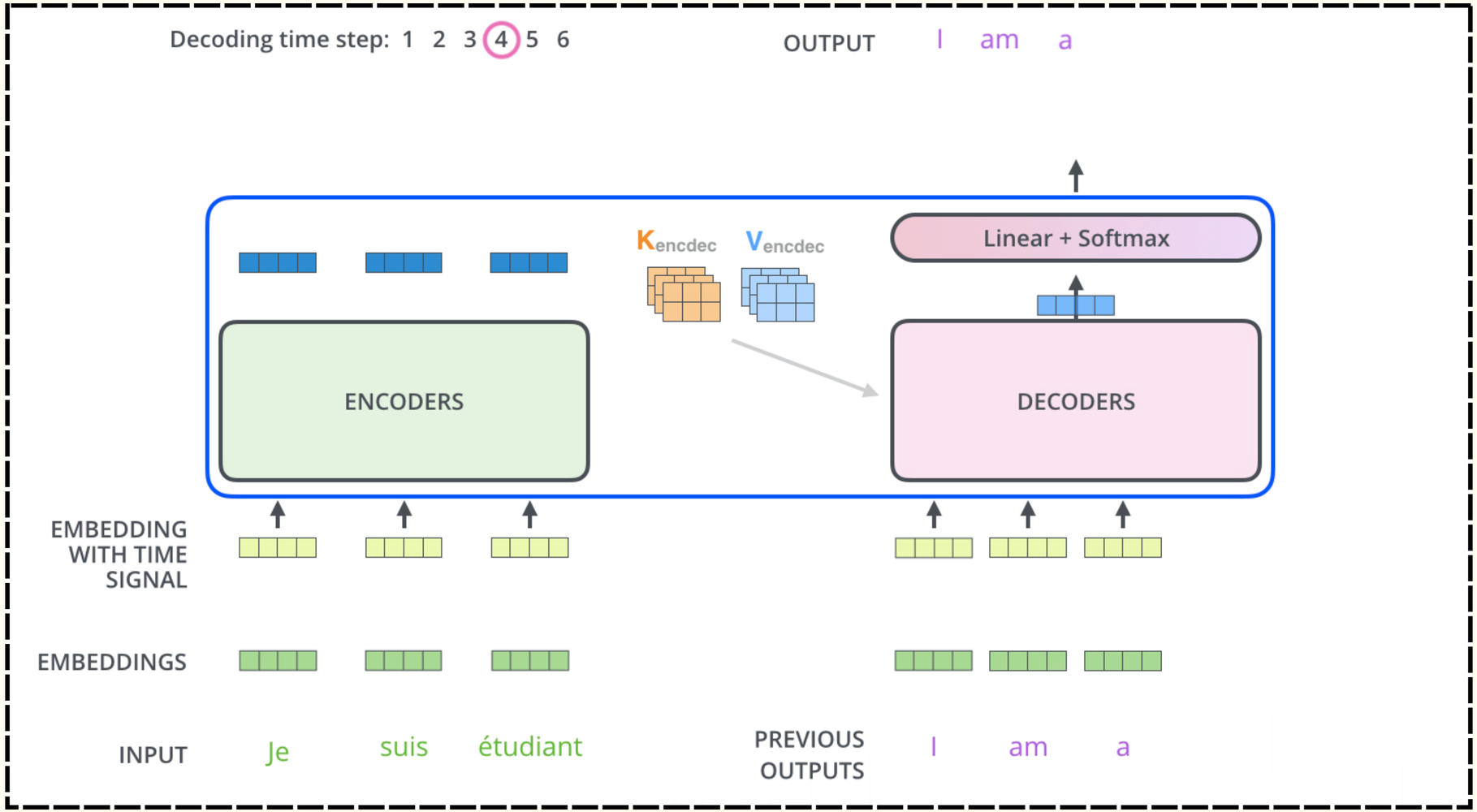

ENCODER #2
ENCODER #1
r1
r2
Feed Forward Neural Network
Feed Forward Neural Network
z1
z2
Self-Attention
x1
x2
Thinking
Machines

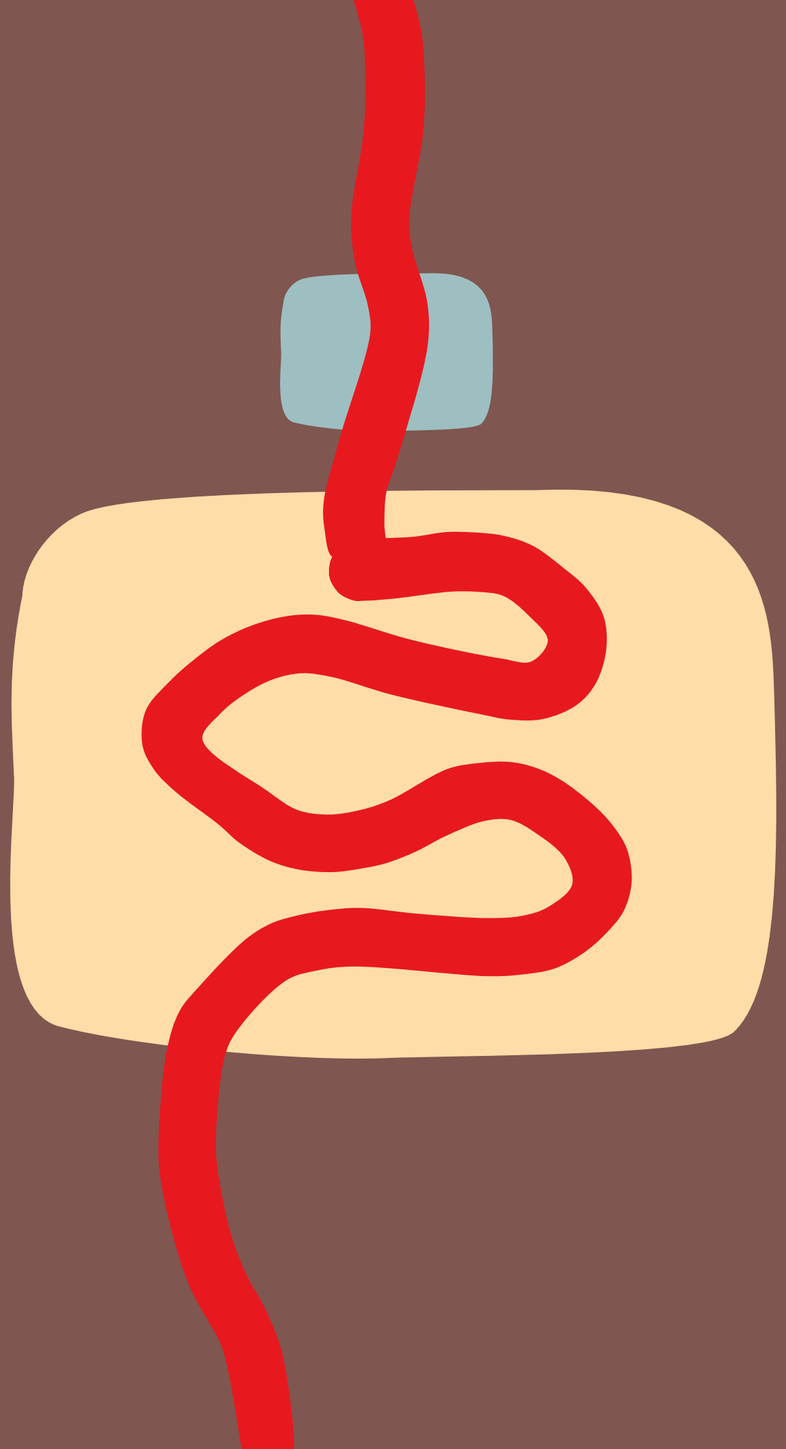

X
W^Q
Q
×
=
X
W^K
K
×
=
X
W^V
V
×
=


Q
K^T
V
softmax
×
$\sqrt{d_k}$
Z
=

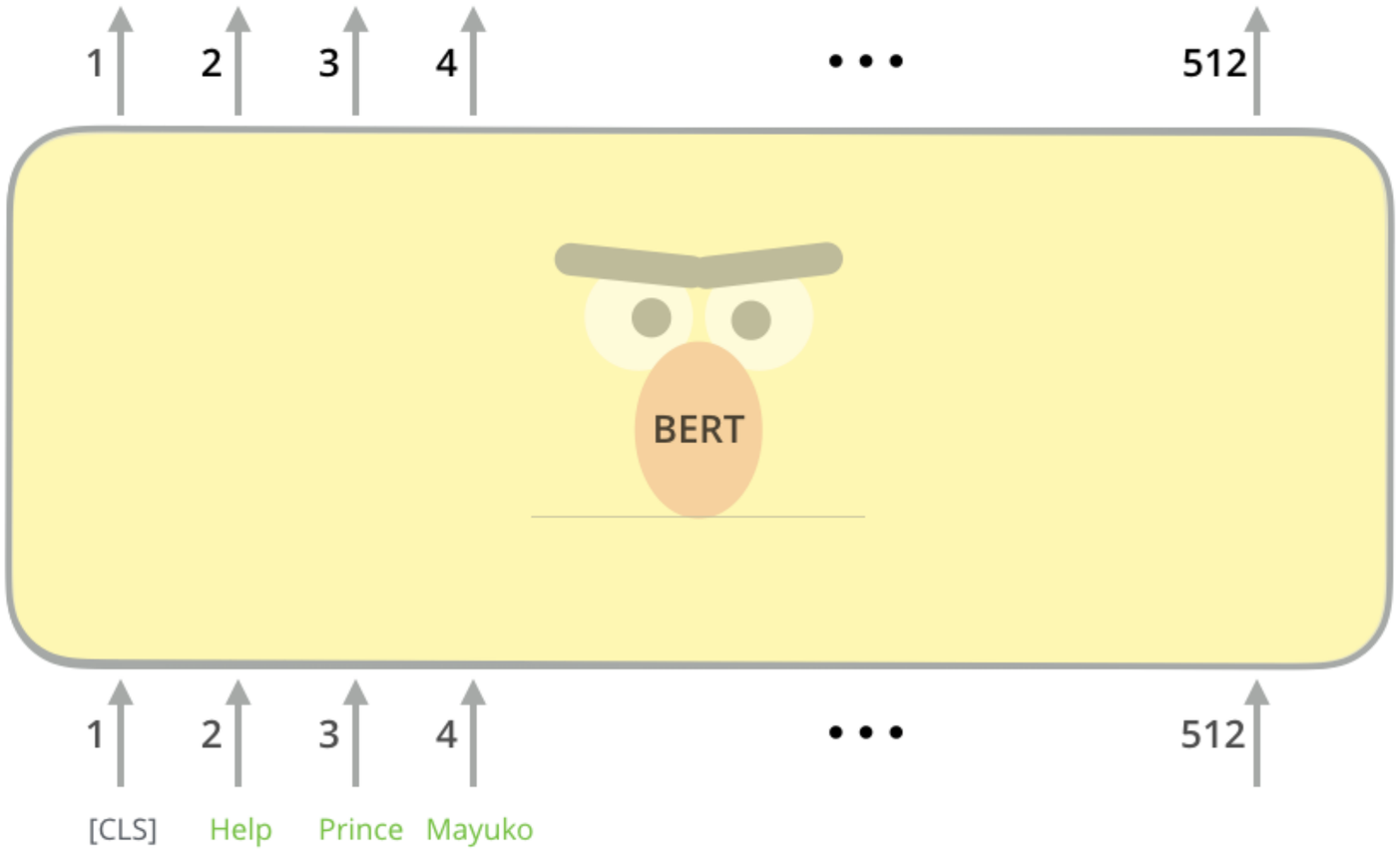

1) This is our input sentence*
2) We embed each word*
3) Split into 8 heads. We multiply X or R with weight matrices
4) Calculate attention using the resulting Q/K/V matrices
5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer
Thinking
Machines
X
R
W_0^Q
W_0^K
W_0^V
W_1^Q
W_1^K
W_1^V
...
W_7^Q
W_7^K
W_7^V
Q_0
K_0
V_0
Q_1
K_1
V_1
...
Q_7
K_7
V_7
Z_0
Z_1
...
Z_7
W^O
Z
* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one

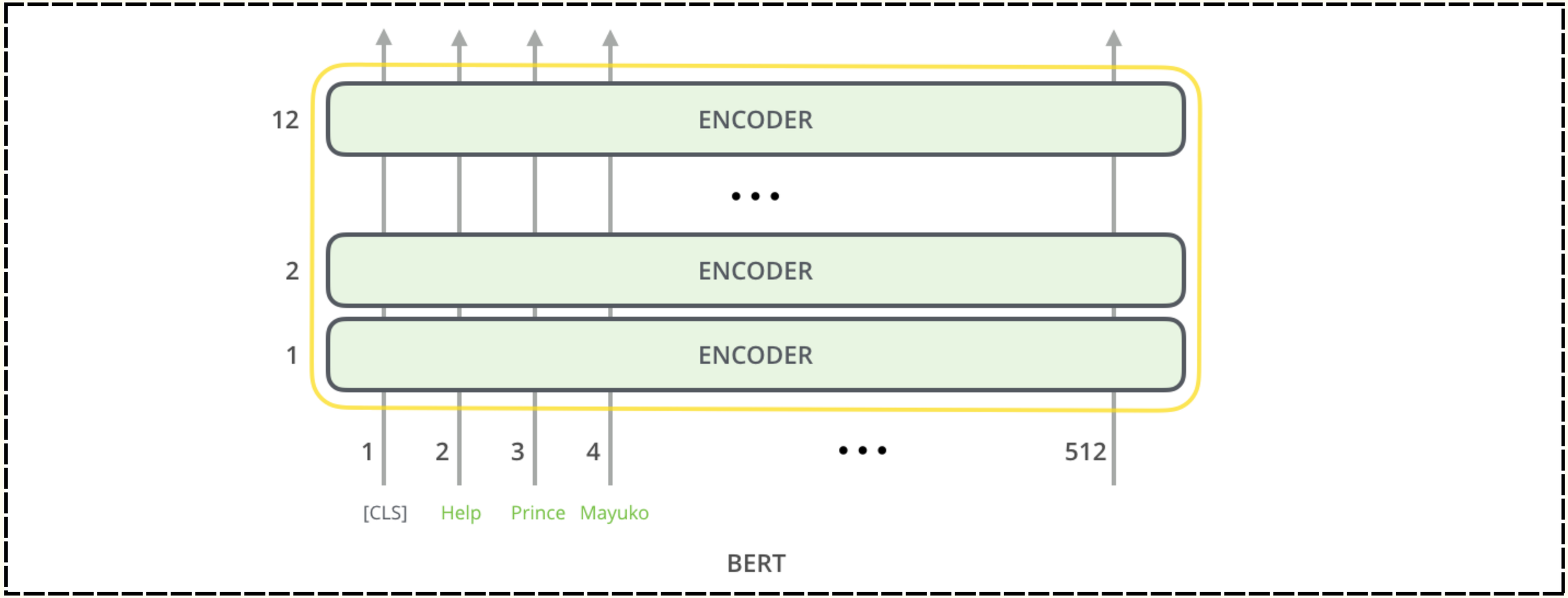

Softmax
Linear
DECODER #2
DECODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Encoder-Decoder Attention
Add & Normalize
Self-Attention
ENCODER #2
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
ENCODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
POSITIONAL ENCODING
x1
x2
Thinking
Machines

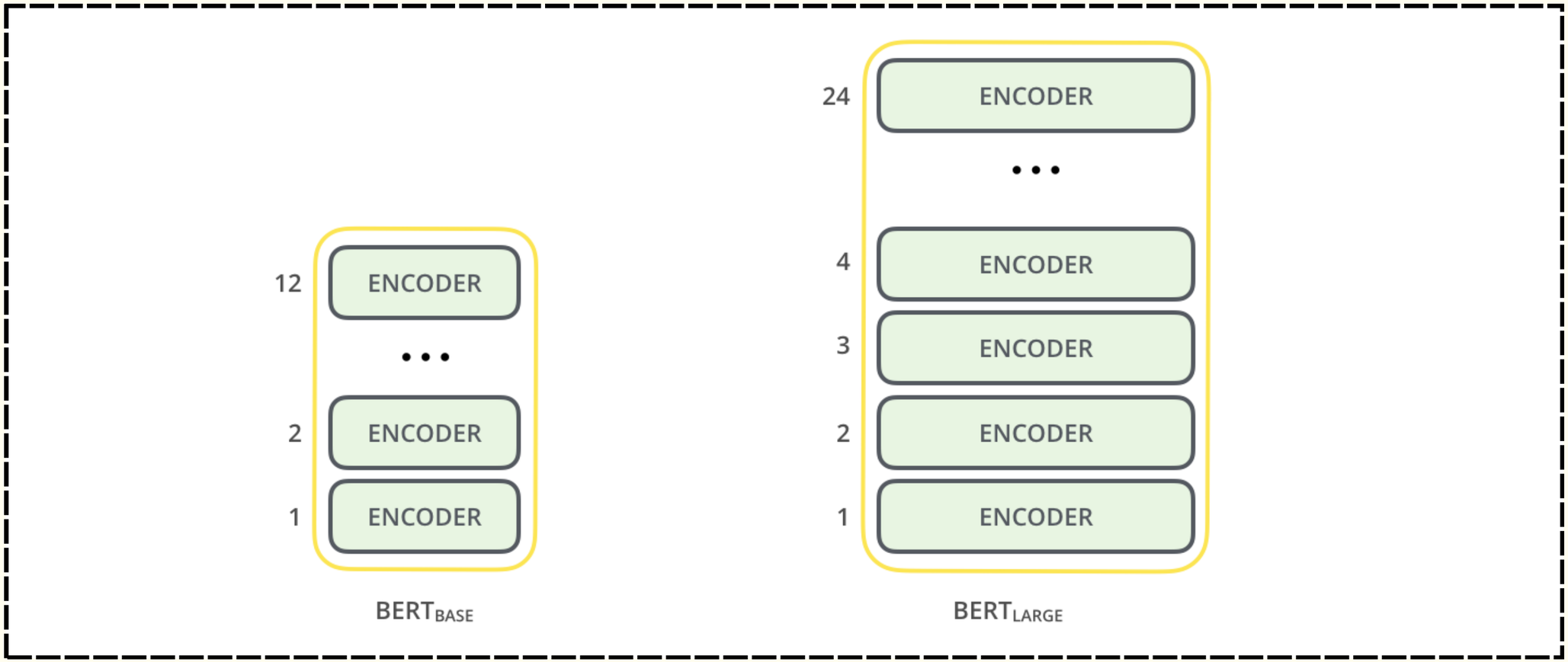

Decoding time step: 1 2 3 4 5 6
OUTPUT
I am a
Linear + Softmax
Kencdec
Vencdec
ENCODERS
DECODERS
EMBEDDING WITH TIME SIGNAL
EMBEDDINGS
INPUT
Je suis étudiant
PREVIOUS OUTPUTS
I am a

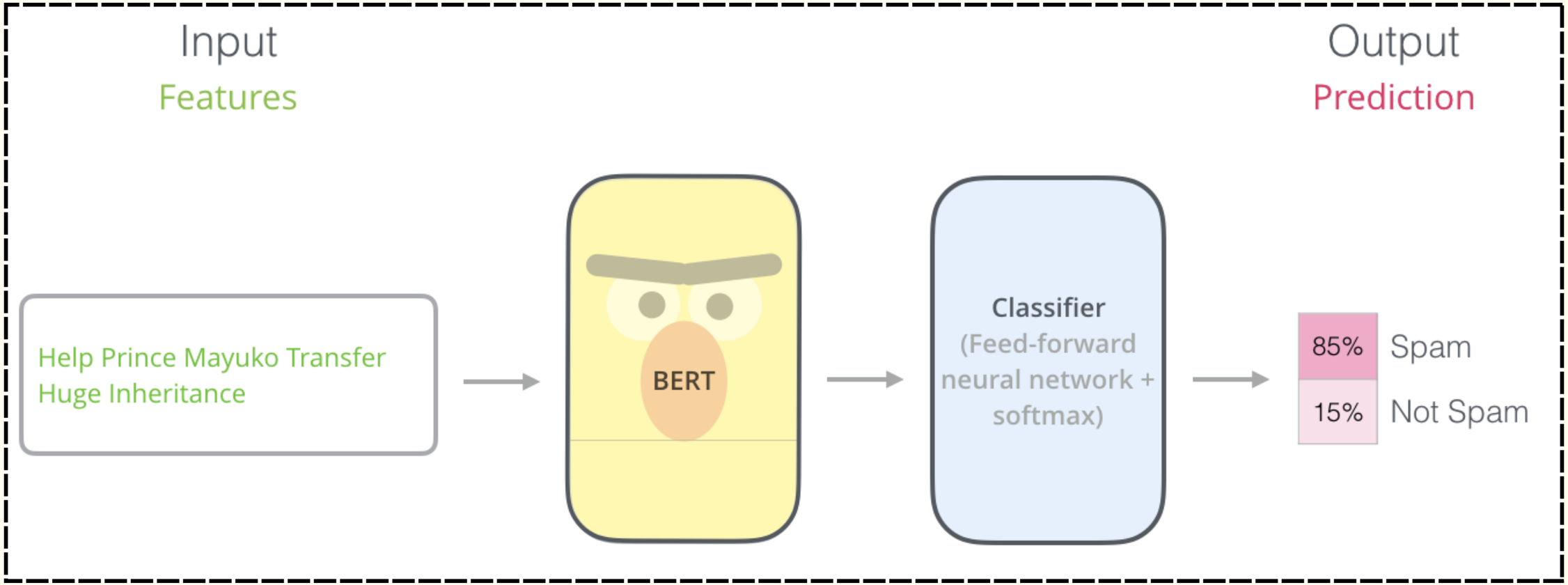


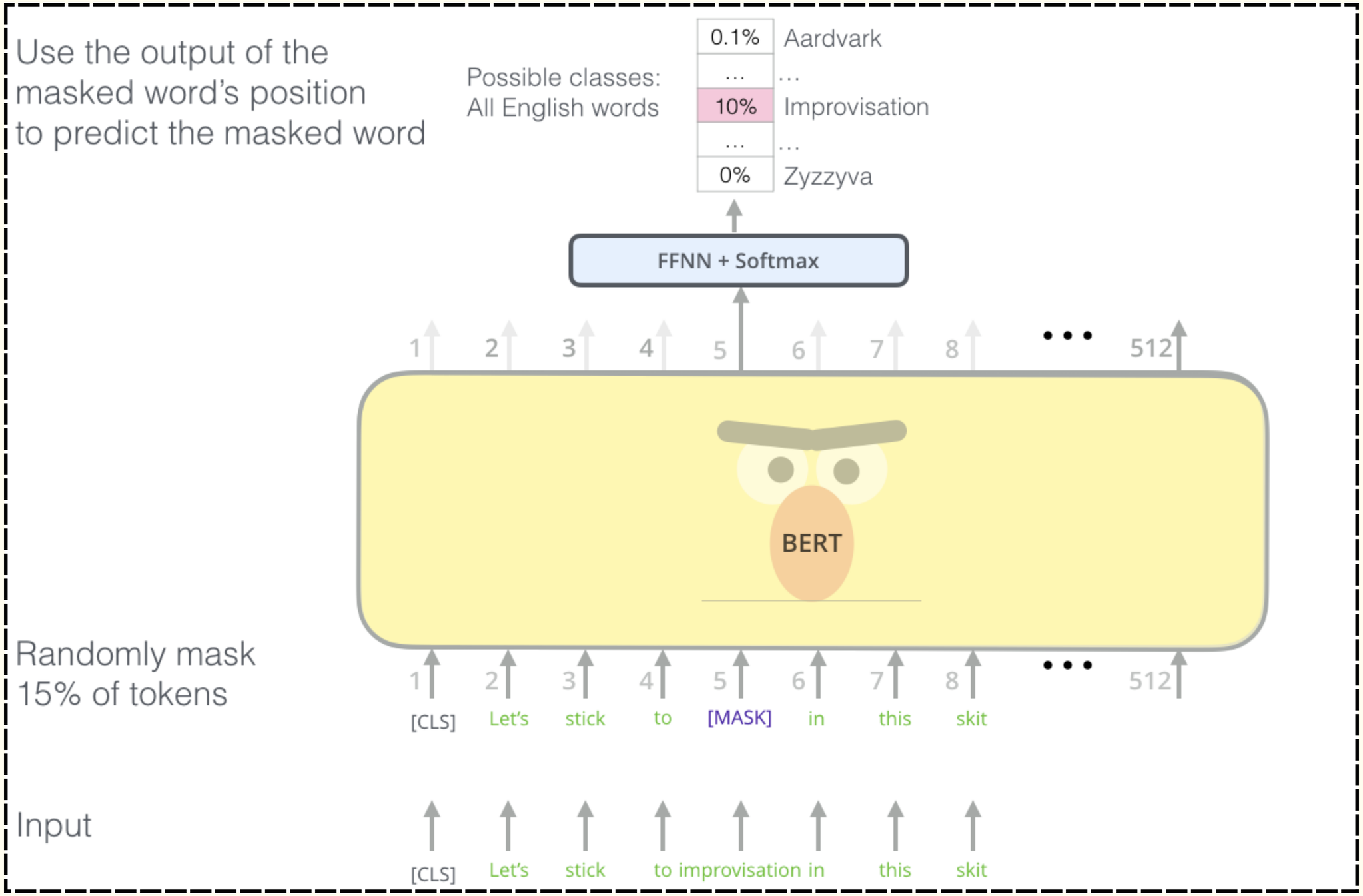
BERT


1
2
3
4
• • •
512
BERT
1
2
3
4
• • •
512
[CLS]
Help
Prince
Mayuko


12
ENCODER
2
ENCODER
1
ENCODER
1
2
3
4
512
[CLS]
Help
Prince
Mayuko
BERT


12
ENCODER
2
ENCODER
1
ENCODER
BERTBASE
24
ENCODER
4
ENCODER
3
ENCODER
2
ENCODER
1
ENCODER
BERTLARGE


Input
Features
Output
Prediction
Help Prince Mayuko Transfer
Huge Inheritance
BERT
Classifier
(Feed-forward
neural network +
softmax)
85%
Spam
15%
Not Spam


Use the output of the masked word's position to predict the masked word
Possible classes: All English words
0.1% Aardvark
... ...
10% Improvisation
... ...
0% Zyzzyva
FFNN + Softmax
1 2 3 4 5 6 7 8 ... 512
BERT
Randomly mask 15% of tokens
1 2 3 4 5 6 7 8 ... 512
[CLS] Let's stick to [MASK] in this skit
Input
[CLS] Let's stick to improvisation in this skit

The end
El fin
O fim
終

Any Questions?

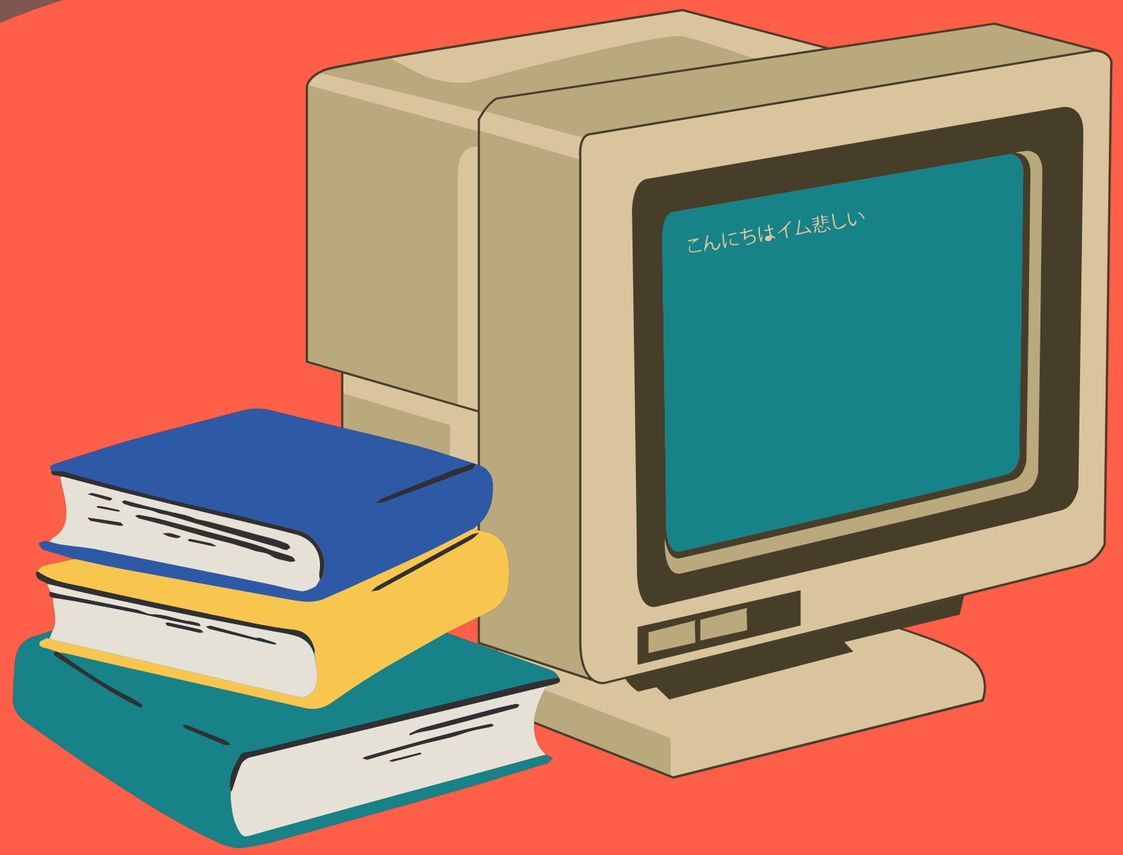


Image Sources

- Blank Venn Diagram Red, Blue and Yellow, by Amousey, under Public Domain
- Major Levels of Linguistic Structure, by James J. Thomas and Kristin A. Cook, and McSush, in Public Domain
- Word Embeddings CBOW, by Jeran Renz, under Creative Commons Attribution-Share Alike 4.0 International
- Transformers Evolutionary Tree, by Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, Xia Hu; extracted from https://arxiv.org/abs/2304.13712
- Transformer Illustrated by Jay Alamar, under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License
  - Alammar, J (2018). The Illustrated Transformer [Blog post]. Retrieved from https://jalammar.github.io/illustrated-transformer/
- Illustrated BERT, by Jay Alamar, under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License
  - Alammar, J (2018). The Illustrated BERT, ELMo and co. [Blog post]. Retrieved from http://jalammar.github.io/illustrated-bert/