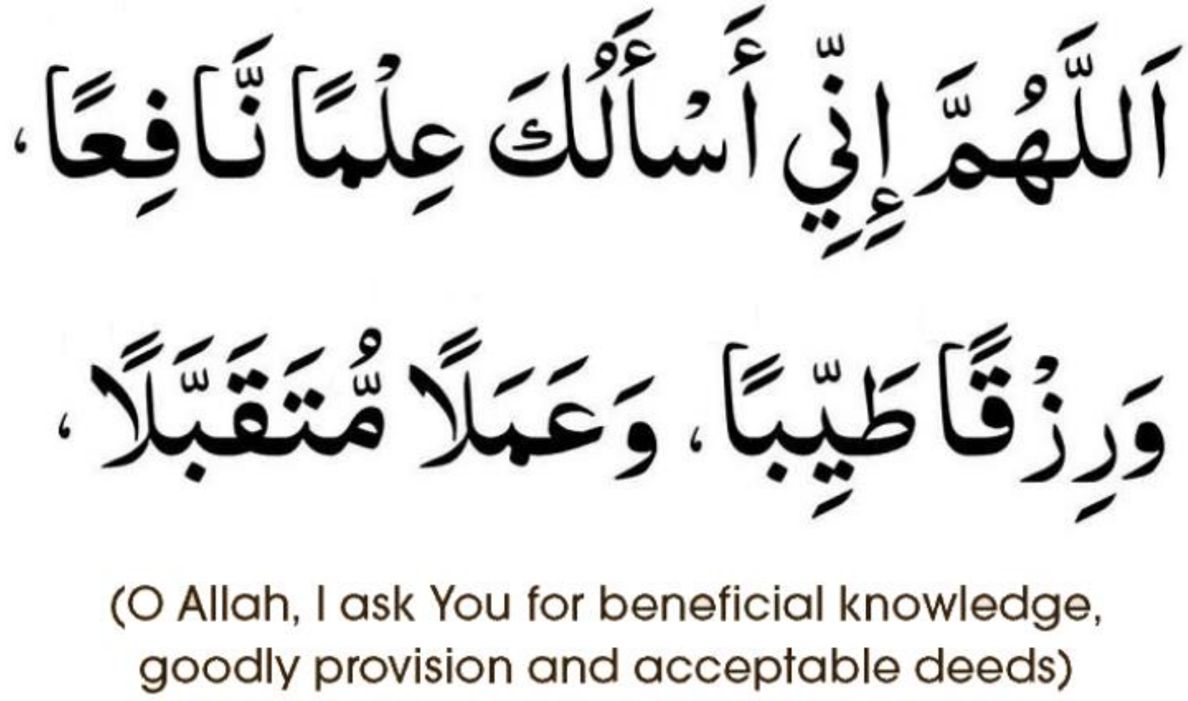
Artificial Intelligence

Unit 11
Transformers and Hugging Face

By:
Syeda Saleha Raza

AL NAFI,
A company with a focus on education,
wellbeing and renewable energy.



اَللّٰهُمَّ إِنِّي أَسْأَلُكَ عِلْمًا نَّافِعًا، وَرِزْقًا طَيِّبًا، وَعَمَلًا مُّتَقَبَّلًا،

(O Allah, I ask You for beneficial knowledge, goodly provision and acceptable deeds)

اے اللہ ، میں آپ سے سوال کرتی ہوں نفع بخش علم کا، طیّب رزق کا، اور اس عمل کا

(Sunan Ibn Majah: 925)

Acknowledgement

- The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)

- Generative AI exists because of the transformer

- Transformers, Explained: Understand the Model Behind GPT-3, BERT, and T5 (daleonai.com)

- What is Hugging Face? The AI Community's Open-Source Oasis | DataCamp

Outline

- Introduction/Intuition
- Attention
- Self-attention
- Architecture
- Applications
- Hands-on

Transformers

Transformers

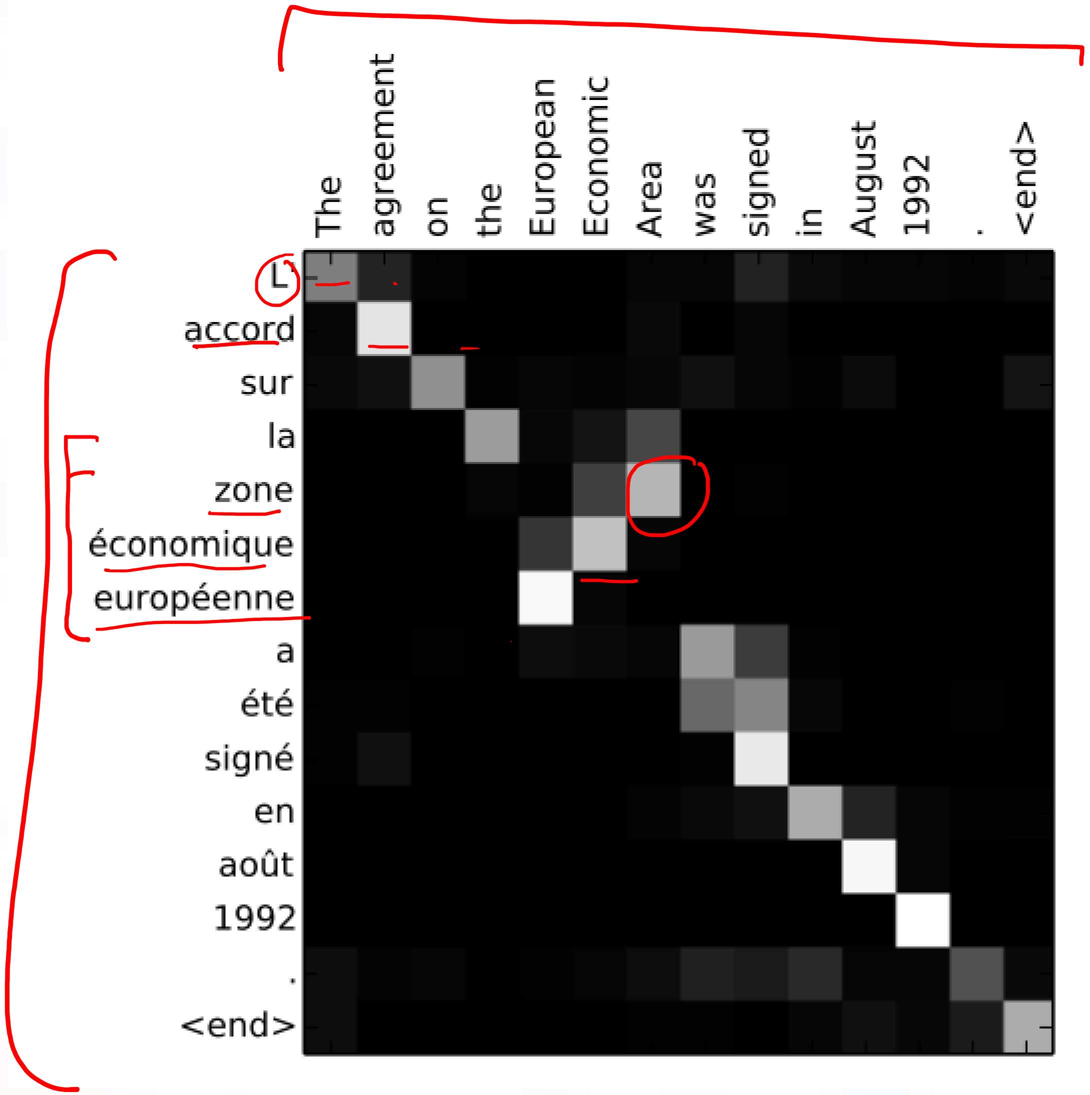
- Transformers have revolutionized the field of natural language processing.
- Transformers process an entire sequence at once — be that a sentence, paragraph or an entire article — analysing all its parts and not just individual words.
- This allows the software to capture context and patterns better, and to translate — or generate — text more accurately.
- This simultaneous processing also makes LLMs much faster to train, in turn improving their efficiency and ability to scale.

Machine Translation

- One bad way to try to translate that sentence would be to go through each word in the English sentence and try to spit out its French equivalent, one word at a time.
- That wouldn't work well for several reasons, but for one, some words in the French translation are flipped: it's
  - "European Economic Area" in English
  - "la zone économique européenne" in French
- Also, French is a language with gendered words. The adjectives "économique" and "européenne" must be in feminine form to match the feminine object "la zone."

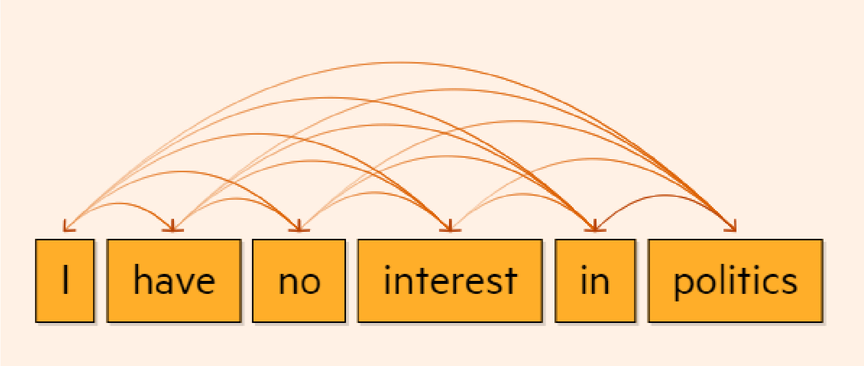
Attention

- Attention is a mechanism that allows a text model to “look at” every single word in the original sentence when making a decision about how to translate words in the output sentence. Here’s a nice visualization from that original attention paper:



Before Transformers….

- Before transformers, the state of the art AI translation methods were recurrent neural networks (RNNs), which scanned each word in a sentence and processed it sequentially.



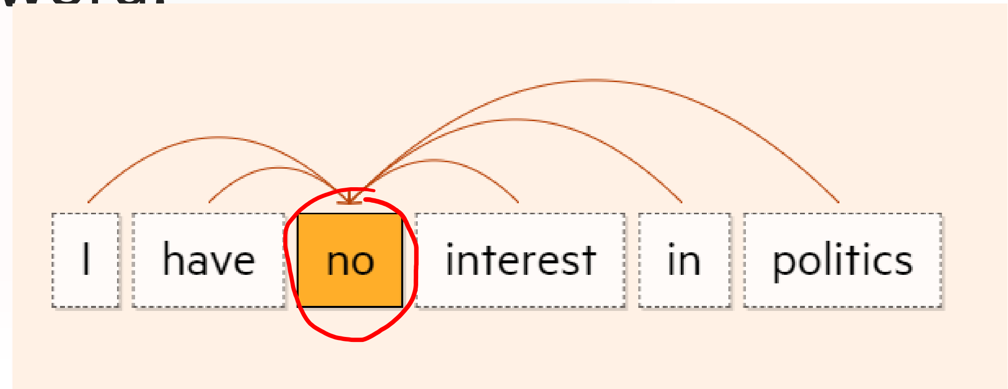
- RNNs become very ineffective when the gap between the relevant information and the point where it is needed become very large.
- Sequential computation inhibits parallelization

Understanding transformers

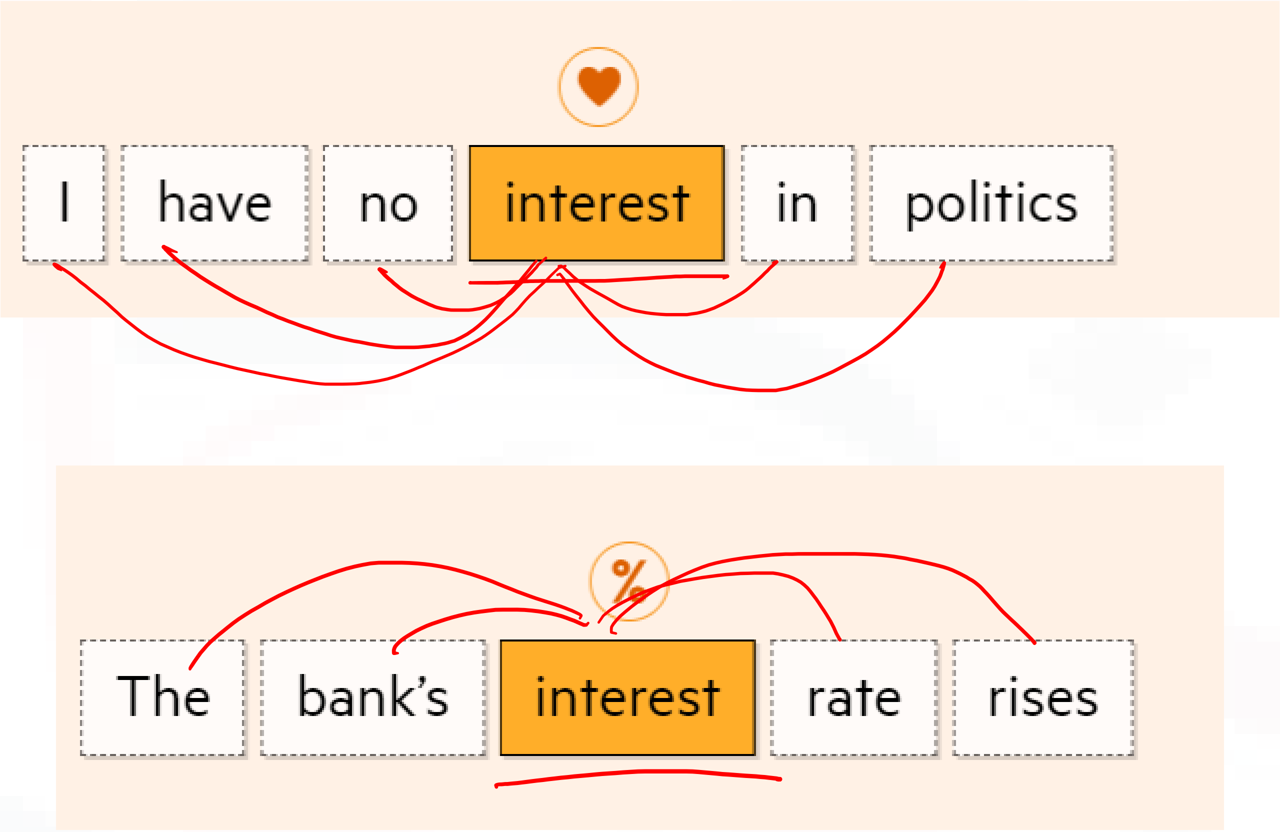
- For example, let’s say that you are trying to predict the last word of the text: **“I grew up in France……… I speak fluent …”.** Recent information suggests that the next word is probably a language, but if we want to narrow down which language, we need context of France, that is further back in the text.

Self-attention

- A key concept of the transformer architecture is self-attention. This is what allows LLMs to understand relationships between words.
- Self-attention determines the relevance of each word in a sequence to every other word.

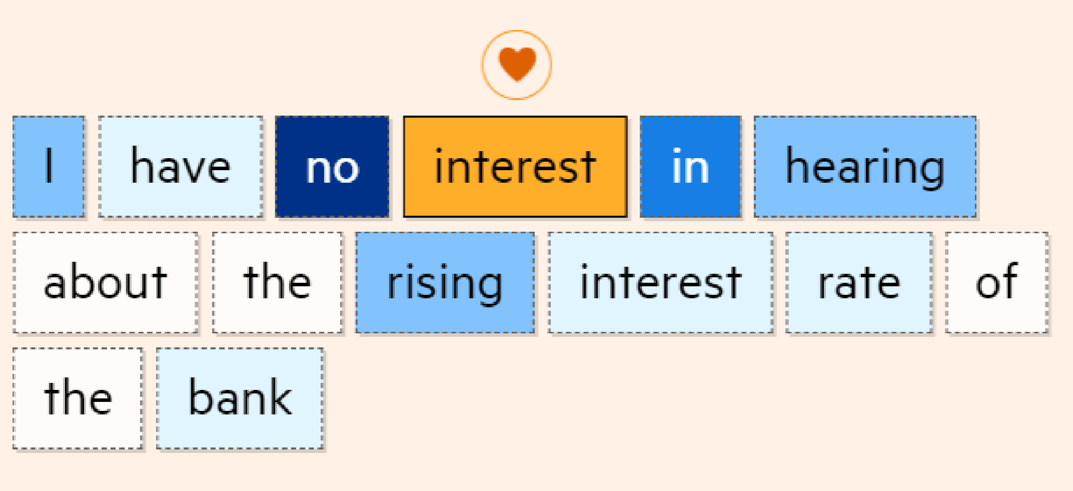


Self-attention

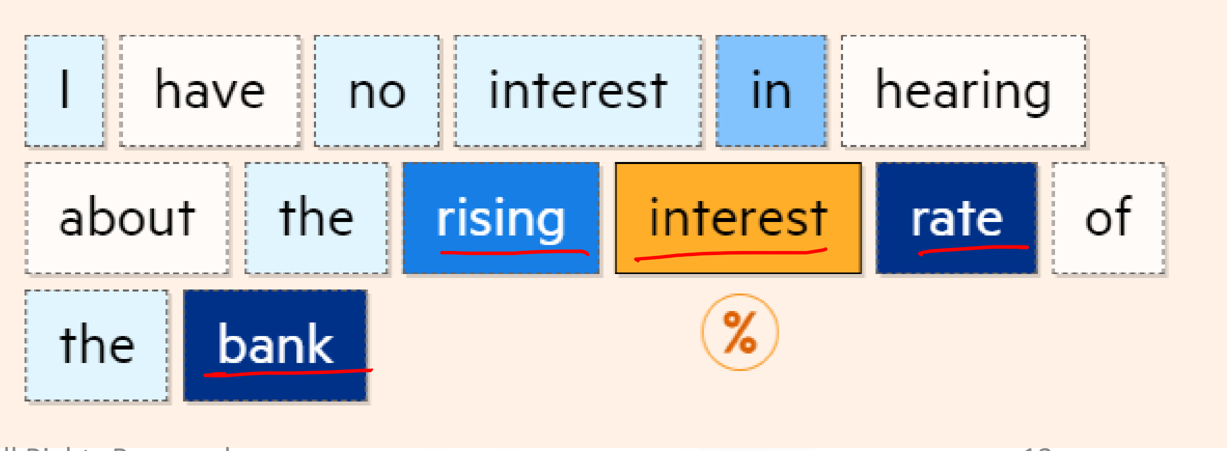


Self-attention

- It answers a fundamental question: “How much should a word attend to another word in the sequence for a specific task?”

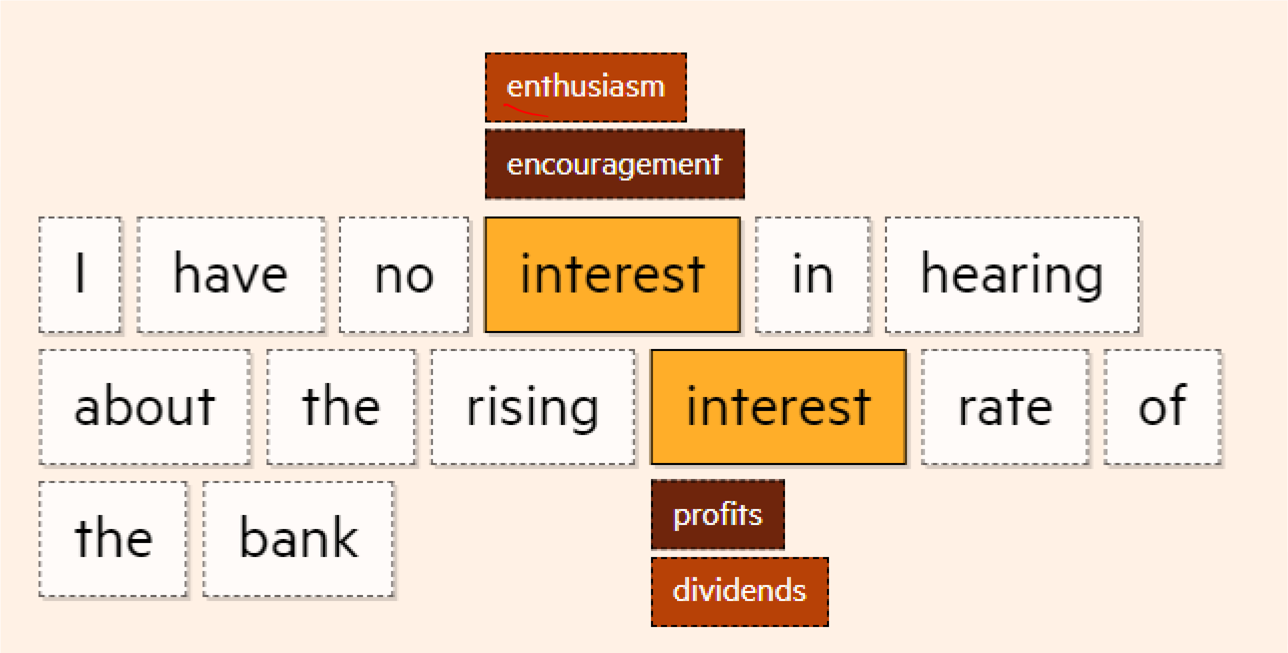


I have no interest in hearing about the rising interest rate of the bank



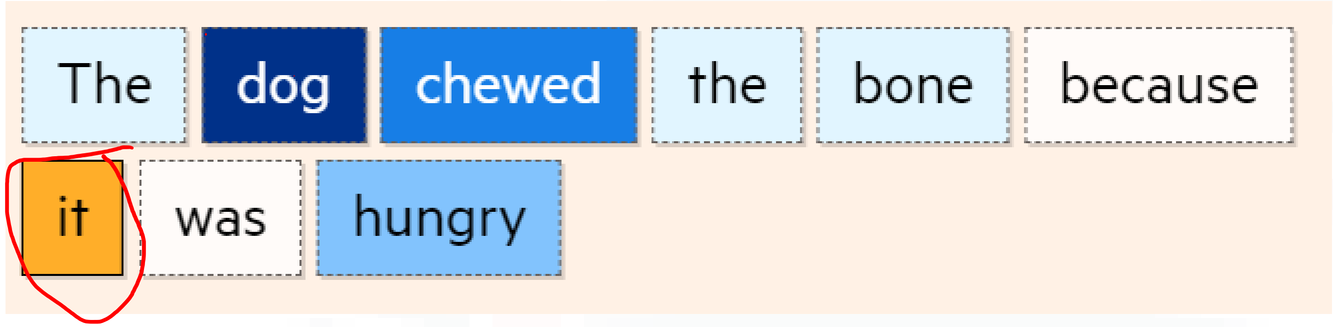
I have no interest in hearing about the rising interest rate of the bank

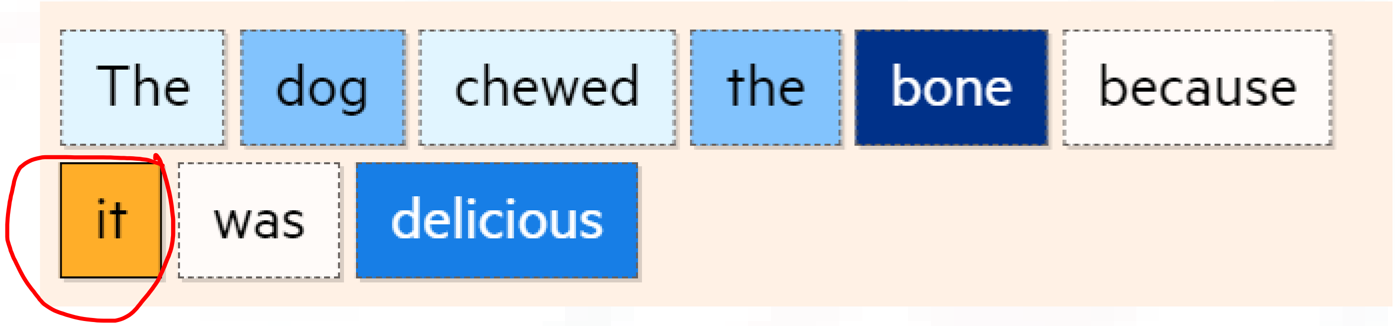
Self-attention



I have no interest in hearing about the rising interest rate of the bank

Self-attention


The dog chewed the bone because it was hungry

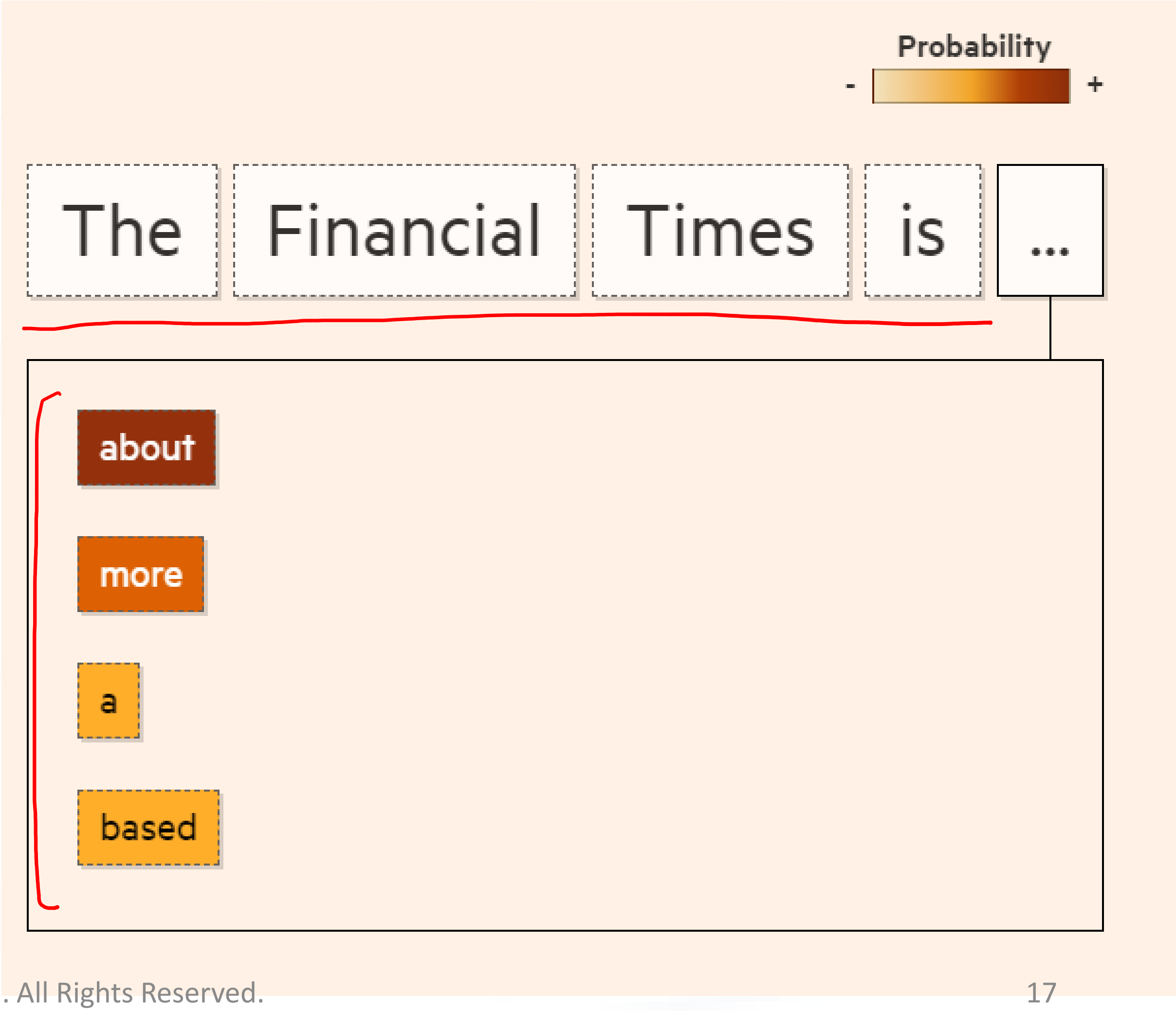

The dog chewed the bone because it was delicious

Self-attention

- The benefits of self-attention for language processing increase the more you scale things up. It allows LLMs to take **context** from beyond sentence boundaries, giving the model a greater understanding of how and when a word is used.

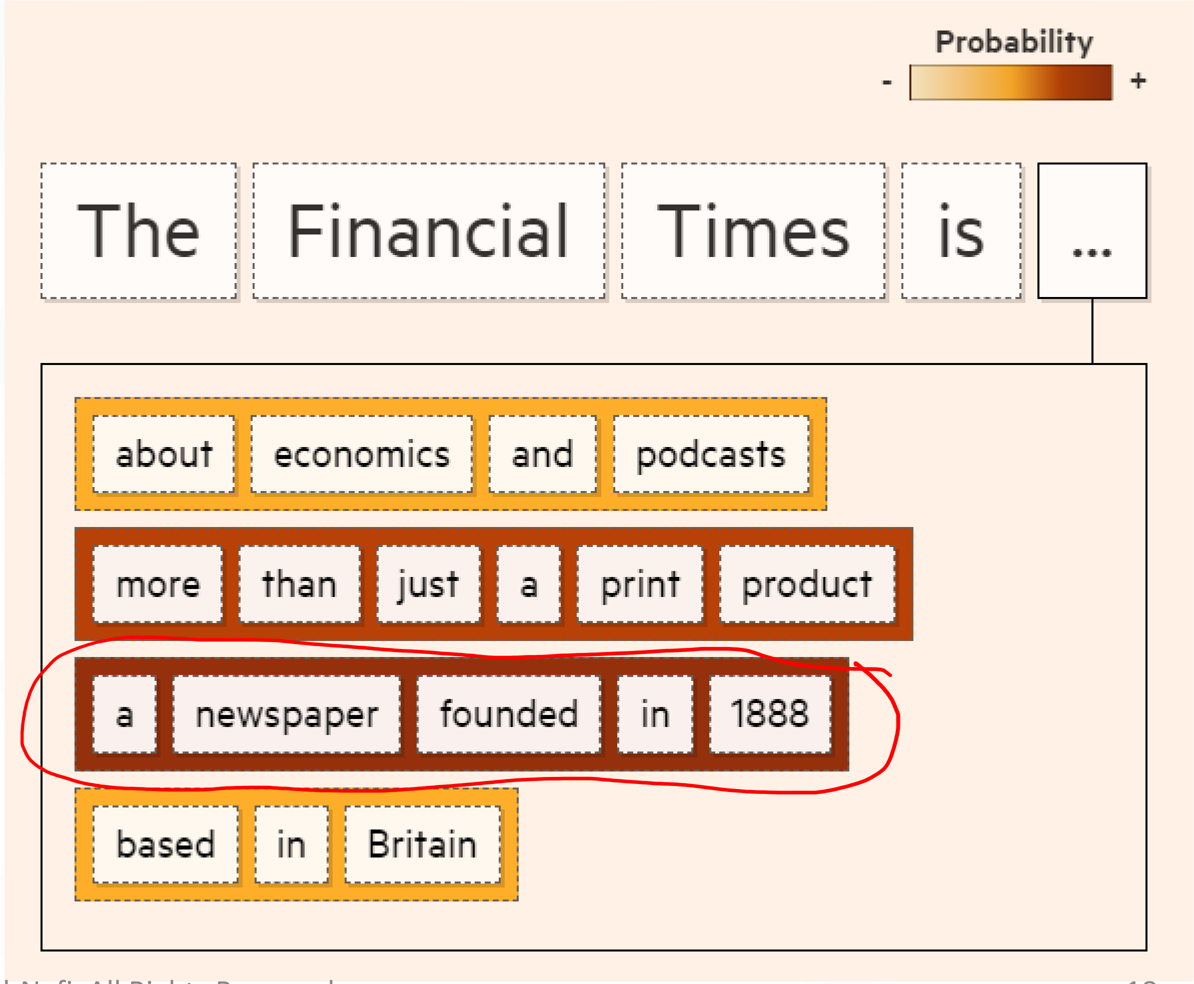
Text generation

- To generate text, the model gives a probability score to each token, which represents the likelihood of it being the next word in the sequence.
- And it continues to do this until it is happy with the text it has produced.



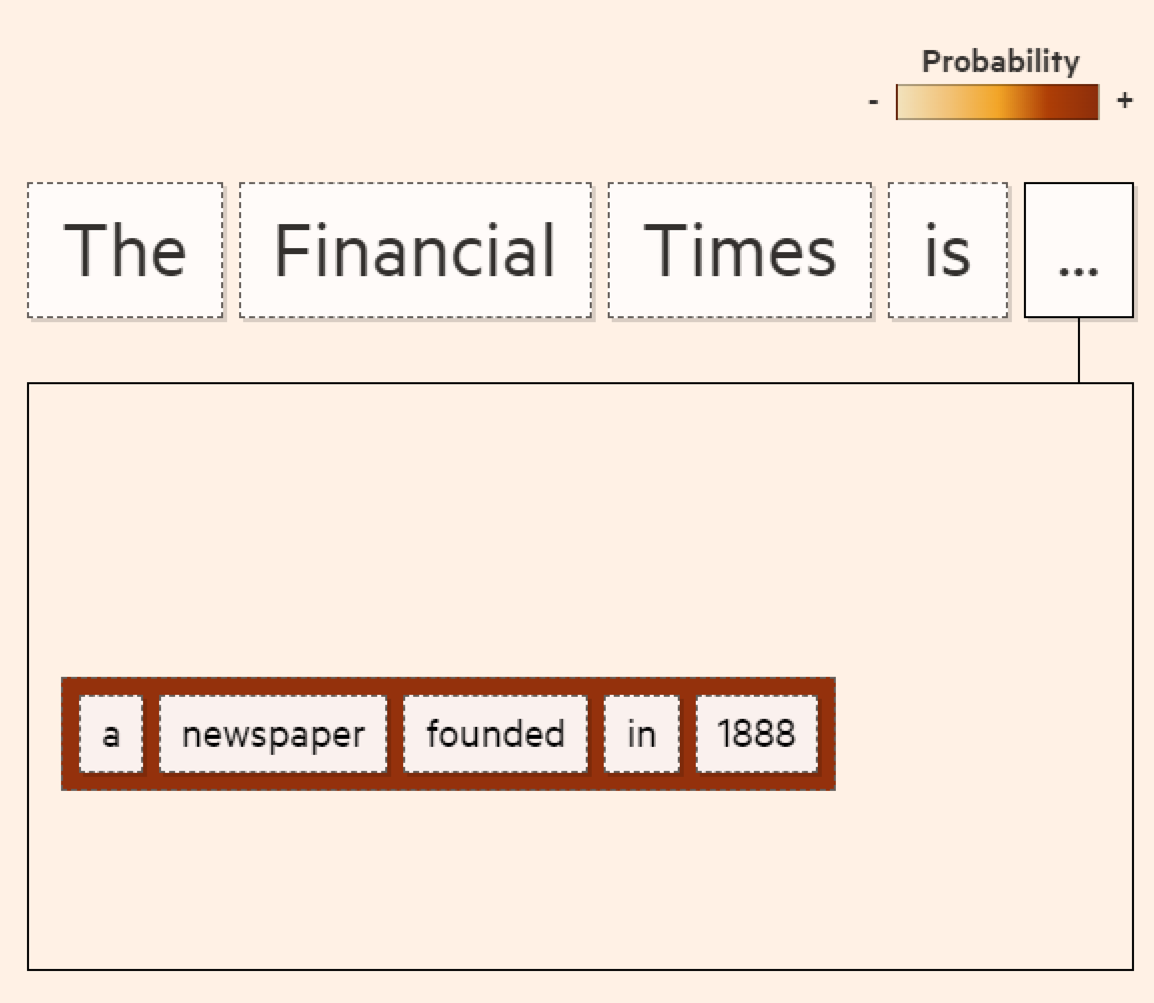
Text generation

- Rather than focusing only on the next word in a sequence, it looks at the probability of a larger set of tokens as a whole.



Text generation

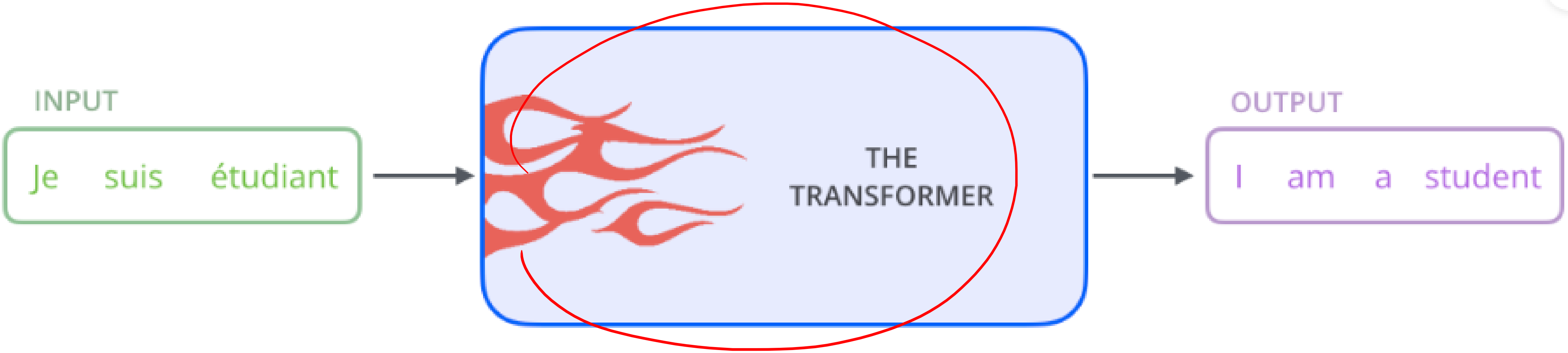
- This produces better results, ultimately leading to more coherent, human-like text.

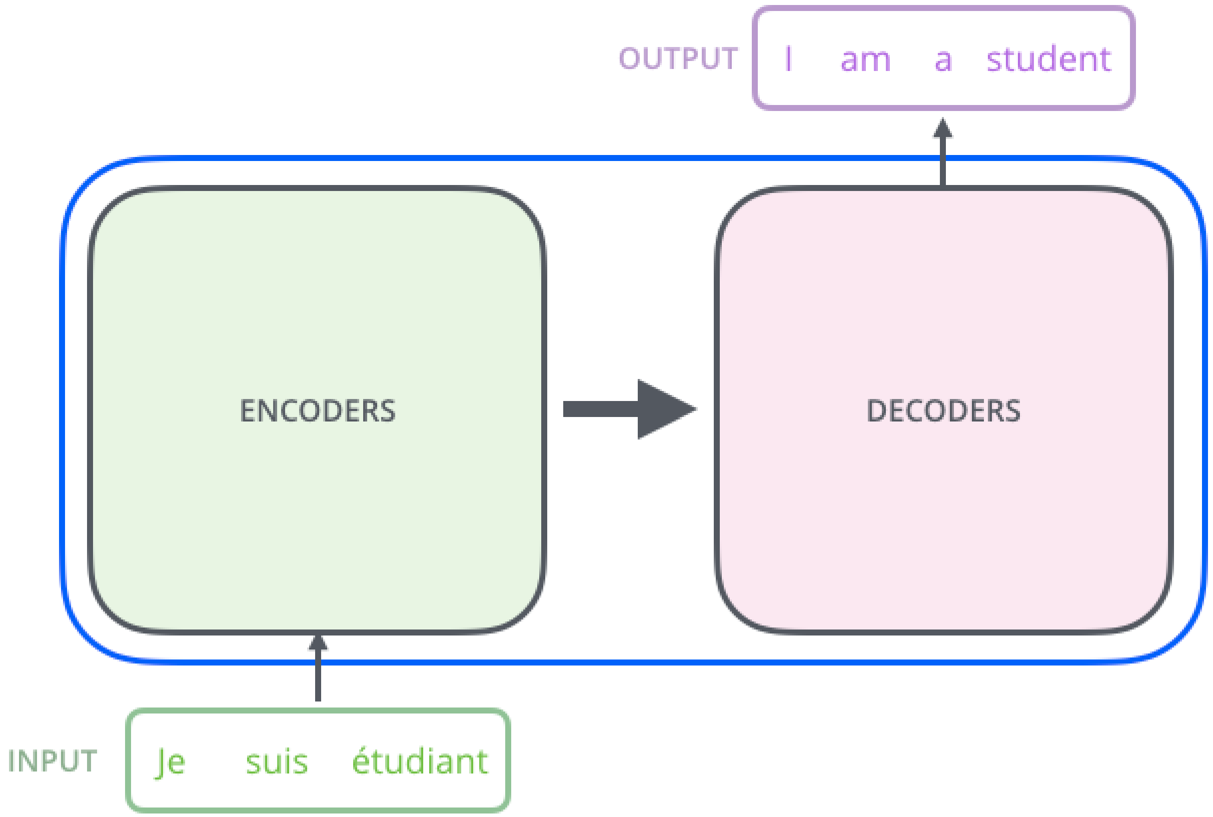


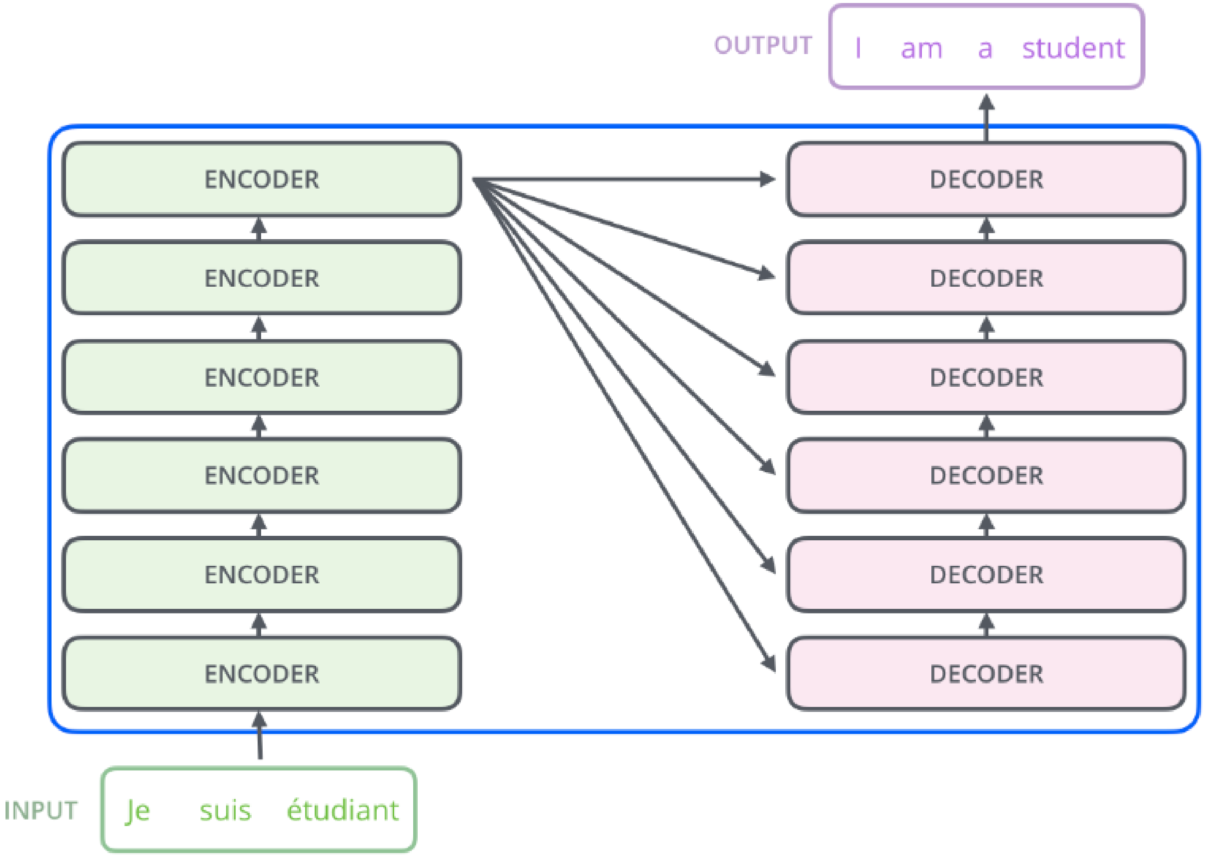
Hallucination

- But things don’t always go to plan. While the text may seem plausible and coherent, it isn’t always factually correct. LLMs are not search engines looking up facts; they are pattern-spotting engines that guess the next best option in a sequence.
- Because of this inherent predictive nature, LLMs can also fabricate information in a process that researchers call “hallucination”. They can generate made-up numbers, names, dates, quotes — even web links or entire articles.

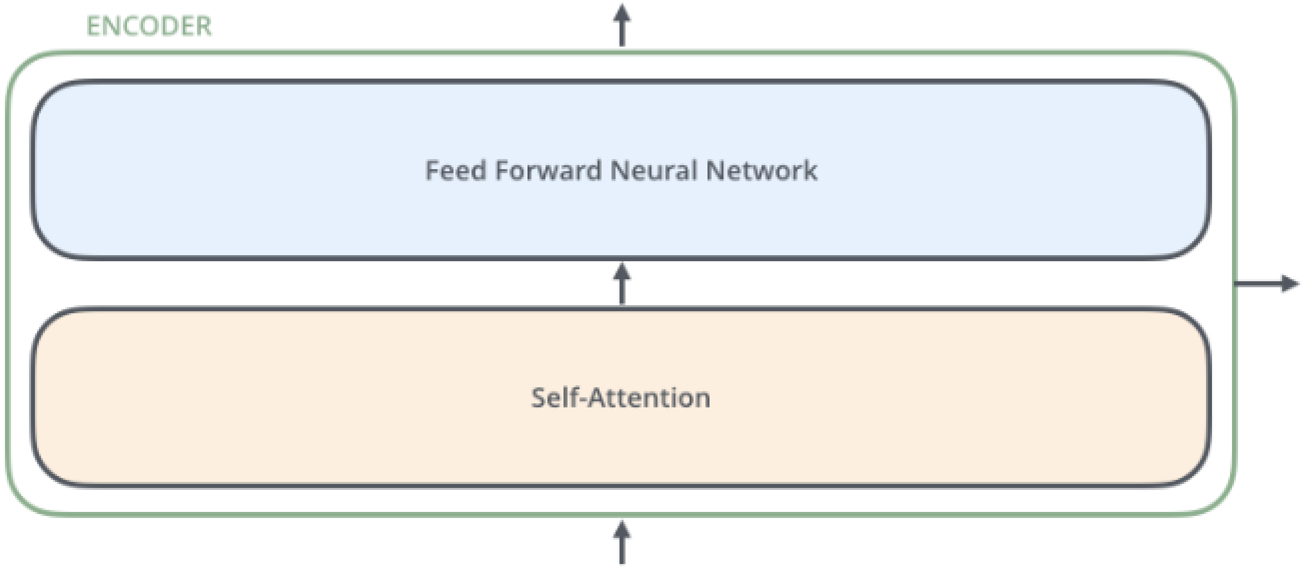
Architecture


INPUT
Je suis étudiant
THE TRANSFORMER
OUTPUT
I am a student

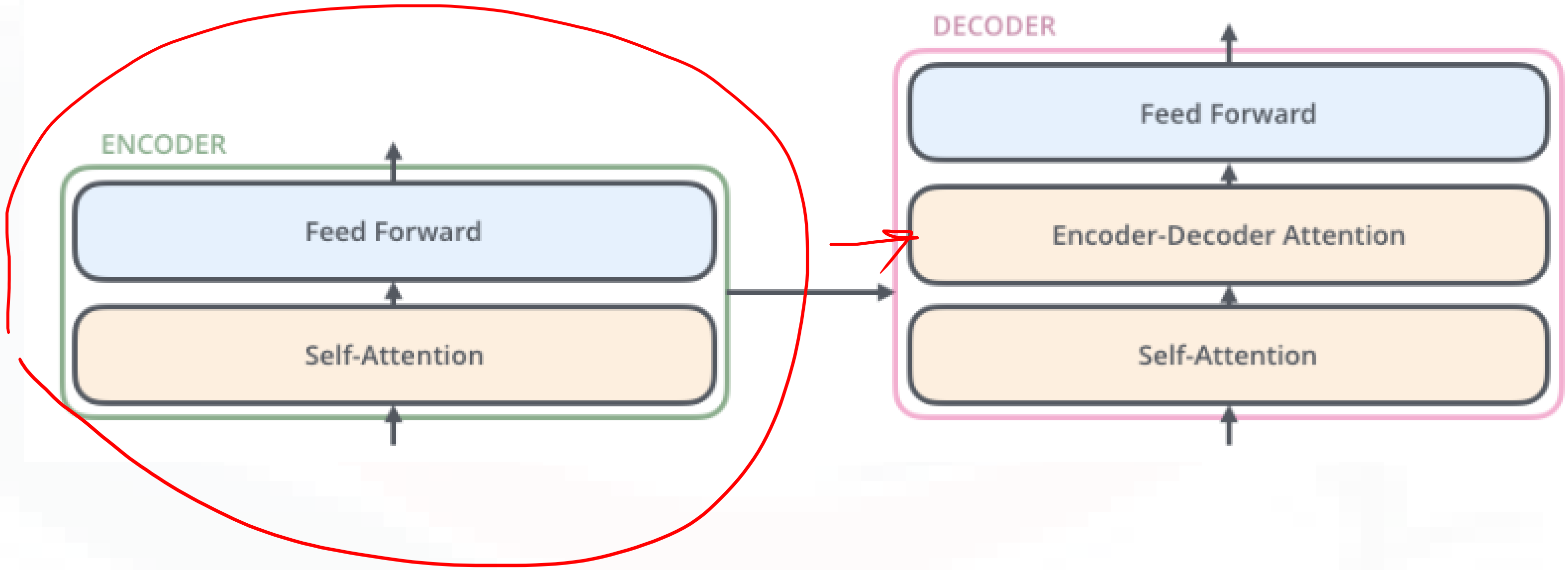

OUTPUT
I am a student
ENCODERS
DECODERS
INPUT
Je suis étudiant

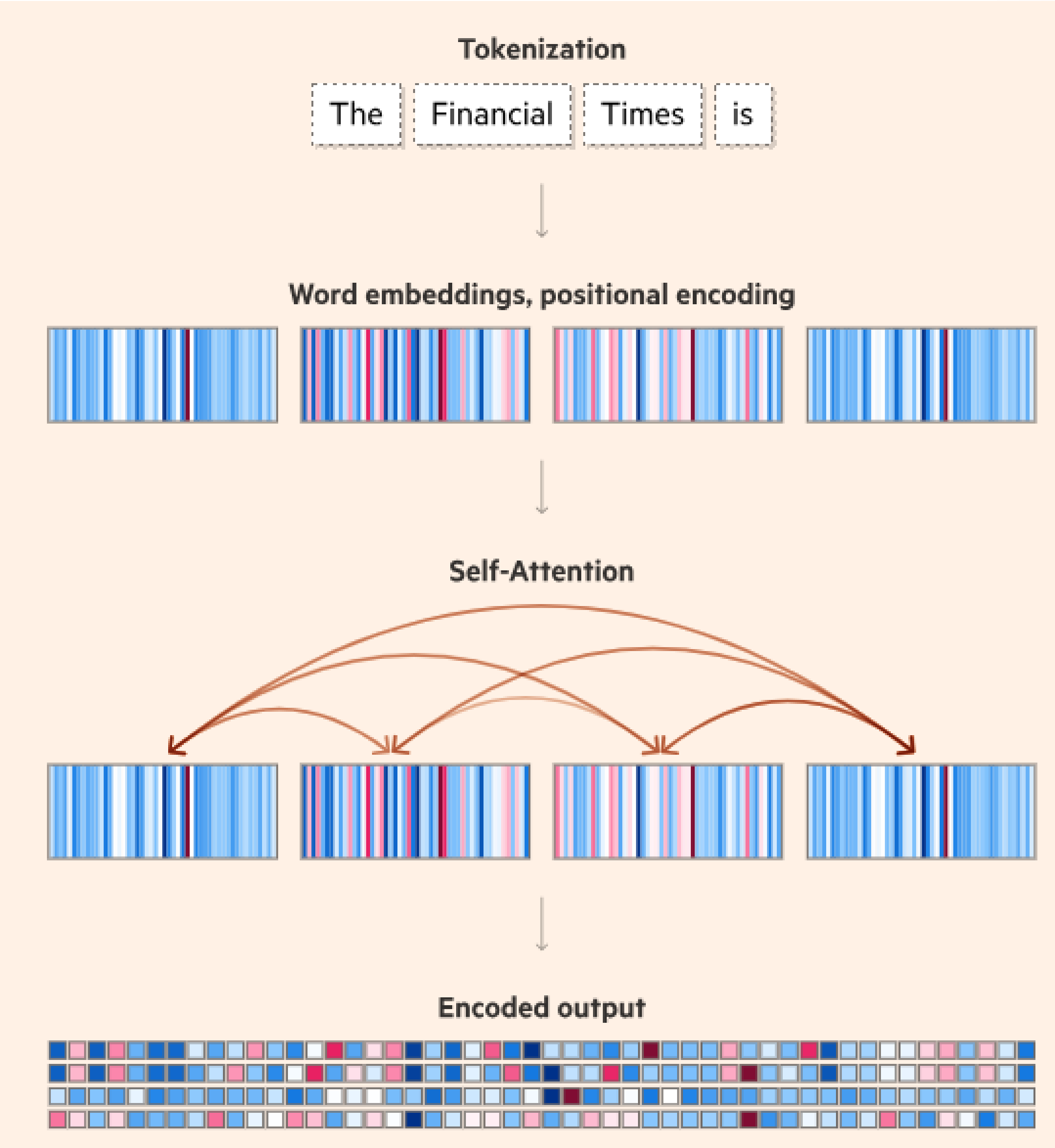

OUTPUT
I am a student
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
DECODER
DECODER
DECODER
DECODER
DECODER
DECODER
INPUT
Je suis étudiant

- The encoders are all identical in structure (yet they do not share weights). Each one is broken down into two sub-layers:



- The decoder has both those layers, but between them is an attention layer that helps the decoder focus on relevant parts of the input sentence




Tokenization
The
Financial
Times
is
Word embeddings, positional encoding
Self-Attention
Encoded output

What can transformers do?

What can transformers do?

- The real power of transformers, however, lies beyond language.
- Transformer models can recognise and predict any repeating motifs or patterns. From **pixels in an image**, using tools such as Dall-E, Midjourney and Stable Diffusion, to **computer code using generators** like GitHub CoPilot.
- It can even predict **notes in music** and **DNA in proteins** to help design drug molecules.

What can transformers do?

- For decades, researchers built specialised models to summarise, translate, search and retrieve. The transformer unified all those actions into a single structure capable of performing a huge variety of tasks.

Applications

- Text summarization
- Question answering
- Classification
- Named entity resolution
- Text similarity
- Offensive message/profanity detection
- Understanding user queries
- a whole lot more

References

- The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)
- Generative AI exists because of the transformer
- Transformers, Explained: Understand the Model Behind GPT-3, BERT, and T5 (daleonai.com)
- What is Hugging Face? The AI Community's Open-Source Oasis | DataCamp

Hands-on

🤗 Hugging Face

Machine Translation

Text Summarization

Text generation

Question/Answering

Sentiment Analysis

جزاك الله

To ask questions, Please use communities link (for respective course) within portal

https://portal.alnafi.com/enrollments