What is a Transformer?

A **Transformer** is a special kind of **artificial intelligence (AI) model** that helps computers **understand language, images, or other types of data**.

It was introduced by Google researchers in 2017 in a paper called **"Attention is All You Need."**

Transformers are the main reason why modern AI systems like **ChatGPT**, **BERT**, and **GPT-5** work so well.

What does a Transformer do?

A Transformer's main job is to **read information (like text)**, **understand the meaning**, and then **generate or predict something new** — like the next word in a sentence.

Example:
If the input is:

> "The sun rises in the …"

The transformer predicts:

> "morning."

How does a Transformer work?

A Transformer has two main parts:

1. **Encoder** – Understands the input.
2. **Decoder** – Generates or predicts the output.

(Some models use only one part — for example, BERT uses only the encoder, GPT uses only the decoder.)

Step-by-step Explanation

1. Input Representation

Words are converted into numbers (called **embeddings**) so that the computer can process them.
For example,
"cat" → [0.2, 0.8, 0.5]

---

2. Positional Encoding

Transformers don't read text in order (left to right).
So, to understand *which word comes first or next*, they add position information to each word.
Example:
"The cat sat" →
cat (position 2), sat (position 3)

---

3. Self-Attention Mechanism

This is the **heart of the Transformer**.

It helps the model decide **which words are important** for understanding each other.

For example:
In the sentence

> "The cat sat on the mat because it was tired."
> The word "it" refers to "cat."
> The attention mechanism helps the model *focus on the word "cat"* when it sees "it."

So the model learns connections between words — not just nearby ones, but even far apart ones.

---

4. Feed-Forward Layers

After attention, each word's understanding is improved through small neural networks called **feed-forward layers**, which learn more complex meanings.

---

5. Stacking Layers

Transformers have many layers (like 12, 24, or even hundreds).
Each layer builds a deeper understanding of the sentence.

The first layer might learn basic grammar.
Higher layers understand relationships and context.

---

6. Decoder (for text generation)

When generating text, the decoder predicts the next word using what it has already seen.
It uses *attention* again to focus on relevant parts of the input.

Example:
If input = “The sun rises in the”,
it predicts the next word “morning”.

Example:

Input: “I love pizza.”

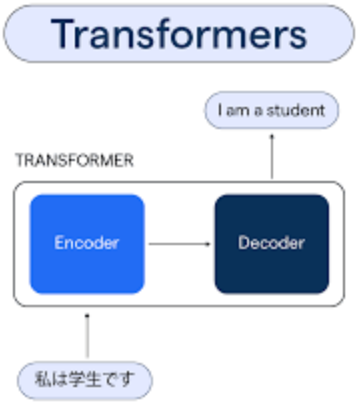
- Attention helps the model understand that “love” connects “I” and “pizza.”
- So it learns the meaning: *the person has positive feelings about pizza.*

Why is it called a “Transformer”?

Because it **transforms** the input (like a sentence) into something useful (like a translation, summary, or answer).

Why Transformers Are Powerful

- They process **all words at once**, not one by one (unlike older models like RNNs).
- They understand **long-distance relationships** in text.
- They can be trained on **huge amounts of data**.
- They work for **text, images, audio, and even video** now.


Transformers
I am a student
TRANSFORMER
Encoder
Decoder
私は学生です