How Transformers Work

1. What is a Transformer?

A Transformer is a type of artificial intelligence model that helps computers understand and create language — just like humans do.

It is used in tools like ChatGPT, Google Translate, and voice assistants.

Before Transformers, models used to process words one by one (like reading a sentence slowly).

But Transformers can look at the entire sentence at once, which helps them understand context much better.

For example:

If you say — “The cat sat on the mat because it was tired.”

The Transformer can understand that “it” refers to “the cat.”

That’s why Transformers are smart at language tasks — they know which words relate to each other in meaning.

2. The Four Simple Steps

Transformers work in four easy steps. Let’s go step by step:

Step 1: Input

The first step is to take the text we want the model to understand.

For example, the input sentence could be:

“I love eating ice cream.”

Before the Transformer can use this sentence, it must break it into smaller pieces called tokens.

Tokens are like word parts or words — e.g., ["I", "love", "eating", "ice", "cream"].

This process is called tokenization.

It helps the computer handle long texts by splitting them into small understandable chunks.

Step 2: Embedding

Computers don't understand words directly — they understand numbers.

So, every token is converted into a vector, which is a list of numbers that represents the meaning of the word.

Example:

"I" → [0.2, 0.1, 0.7, 0.5]

"love" → [0.9, 0.8, 0.1, 0.3]

"ice" → [0.6, 0.2, 0.4, 0.9]

These numbers are called embeddings.

They help the Transformer understand how similar or different words are.

For instance, "love" and "like" might have similar embeddings because they mean almost the same thing.

Step 3: Self-Attention

This is the heart of the Transformer.

The self-attention mechanism helps the model understand which words are important in a sentence and how they relate to each other.

For example, in the sentence:

"The dog chased the ball because it was rolling."

The word "it" could mean "the ball."

Self-attention allows the model to look at every word and decide — which other words should I pay attention to?

It gives a weight (importance value) to each word.

So, the model learns that "it" refers to "the ball", not "the dog."

In short:

Self-attention helps the model focus on the right words when understanding or generating text.

Step 4: Output

Finally, after understanding the meaning and context, the model produces an output.

This could be:

The next word in a sentence (for text generation).

A translation in another language.

An answer to a question.

Example:

If the input is — “The capital of France is”,

The model predicts — “Paris.”

3. Why Transformers Are So Powerful

Transformers are used in almost every modern AI system because they are:

Fast — They look at all words at once, not one by one.

Smart — They understand context better than older models.

Flexible — They work for text, images, audio, and even video.

For example:

BERT and GPT are both built using Transformer architecture.

They can be trained on large data (like books, articles, websites) to learn language patterns.