

Encoder vs Decoder (Transformer Study Notes)

Big Picture (One Line)

Encoder understands the input. Decoder generates the output.

What Is an Encoder?

An **Encoder** reads the **entire input sentence** and converts it into a **context-rich representation**.

What the Encoder Does

- Takes the full sentence at once
- Uses **self-attention**
- Understands **meaning, context, and relationships**
- Does **not generate new text**

Example

Input:

The bank approved the loan

Encoder output:

- A set of vectors that fully represent the **meaning of the sentence**

What Is a Decoder?

A **Decoder** generates text **token by token**, using:

- Previously generated tokens
- Encoder output (if available)

What the Decoder Does

- Predicts the **next word**
- Uses **masked self-attention**
- Can generate long text sequences

Example

Input:

Translate to French: The bank approved the loan

Decoder output:

La banque a approuv le pr t

Key Difference (Core Idea)

Encoder	Decoder
Understands input	Generates output
Reads full sentence	Generates word by word
No masking	Uses masking
Context builder	Text generator

Attention Used in Encoder vs Decoder

Encoder Attention

- **Self-Attention**
- Each word attends to **all other words**

Example:

bank loan approved

Decoder Attention (Two Types)

1. Masked Self-Attention

- Can only attend to **past tokens**
- Cannot see future words

2. Cross-Attention

- Attends to **encoder output**
- Connects input meaning to output generation

Why Masking Is Needed in the Decoder

Without masking, the decoder would **cheat** by seeing future words.

Example during training:

I love AI

When predicting “**love**”, the decoder must **not** see “**AI**”.

Masking ensures:

The next word is predicted using only previous words.

Architecture Comparison

Encoder-Only Models

Used for **understanding tasks**

Examples:

- BERT
- RoBERTa

Use cases:

- Text classification
- Semantic search
- Text embeddings

Decoder-Only Models

Used for **generation tasks**

Examples:

- GPT
- LLaMA

Use cases:

- Chatbots
- Story generation
- Code generation

Encoder-Decoder Models

Used for **input-to-output transformation**

Examples:

- T5
- BART

Use cases:

- Translation
- Summarization
- Question answering

Simple Flow Diagrams (Text)

Encoder-Only

Input sentence

Encoder

Understanding / embeddings

Decoder-Only

Previous tokens

Decoder

Next token

Encoder-Decoder

Input sentence Encoder

Decoder Output sentence

Real-World Examples

Task	Model Type
Sentence embeddings	Encoder
Semantic search	Encoder
ChatGPT	Decoder
Translation	Encoder-Decoder
Summarization	Encoder-Decoder

Key Takeaways

- Encoder understands language
- Decoder generates language
- Masking is used only in the decoder
- GPT is **decoder-only**
- BERT is **encoder-only**

One-Line Interview Answers

Encoder:

The encoder transforms input text into contextual representations.

Decoder:

The decoder generates output text token by token using masked attention.

One-Line Memory Trick

Encoder reads, Decoder writes.