Masked Attention

What Is Masked Attention?

Masked Attention is a special type of self-attention used in **decoder models** (such as GPT), where a token is allowed to attend only to **past tokens** and not to future ones.

In one line:

> *Masked attention prevents the model from seeing future words while predicting the next word.*

Why Masked Attention Is Needed

During text generation, a model predicts **one word at a time**.

Example Sentence

The bank approved the loan

When predicting the word **"approved"**, the model must **not** see:

the loan

Otherwise, the model would **cheat during training**.

Masked attention enforces the following rule:

- Look left (past tokens)
- Look right (future tokens) ×

Where Masked Attention Is Used

Masked attention is used in:

- Decoder blocks

- GPT, GPT-2, GPT-3, GPT-4
- Text generation
- Autoregressive language models

It is **not used in encoder-only models** such as BERT.

How Masked Attention Works (Intuition)

The process works as follows:

1. Create a mask matrix
2. Mask out future positions
3. Apply attention only to allowed tokens

Simple Mask Visualization

For a sentence with five words:

```
Tokens:  The   bank  approved  the   loan
Index:    0     1       2       3     4
```

Mask matrix (1 = allowed, 0 = blocked):

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Each token can attend only to **itself and previous tokens**.

Mathematical Idea (Simple)

Normal Attention

$$\text{softmax}(QK^T)$$

Masked Attention

$$\mathrm{softmax}(QK^T + \mathrm{mask})$$

The mask adds $-\infty$ to future positions, and the softmax function converts those values into **zero attention**.