Multi-Head Attention

What is Multi-Head Attention?

Multi-Head Attention allows a Transformer to **look at the same sentence in multiple ways at the same time**.

Instead of using a single attention mechanism, the model uses **multiple attention heads**, where each head focuses on **different relationships** in the sentence.

Why Multi-Head Attention Is Needed

A sentence contains **many types of information simultaneously**:

- Meaning (semantics)
- Grammar (syntax)
- Time or position
- References (pronouns)
- Long-range dependencies

A **single attention head** can capture only limited patterns. **Multiple heads together provide richer understanding**.

Simple Intuition

Sentence:

```
The bank approved the loan yesterday.
```

Different attention heads may learn:

- Head 1 → bank ↔ loan (meaning)

- Head 2 → approved ↔ bank (action)
- Head 3 → yesterday ↔ approved (time)
- Head 4 → sentence structure

All heads work **in parallel**, and their knowledge is combined.

How Multi-Head Attention Works (Conceptual Steps)

1. Input sentence is converted into embeddings
2. Embeddings are projected into **Queries (Q), Keys (K), and Values (V)**
3. These projections are **split into multiple heads**
4. Each head performs **self-attention independently**
5. Outputs of all heads are **concatenated**
6. A final linear layer produces the output

Mathematical Formulation

1. Scaled Dot-Product Attention (Single Head)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Meaning of Terms

- Q → Query matrix
- K → Key matrix
- V → Value matrix
- d_k → dimension of keys (used for scaling)

2. Multi-Head Attention

For each head i:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Where:

- W_i^Q, W_i^K, W_i^V are learned weight matrices for head i

3. Combine All Heads

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O$$

Where:

- h = number of attention heads
- W^O = output projection matrix

Why Scaling by $\sqrt{d_k}$?

- Prevents dot-product values from becoming too large
- Keeps gradients stable
- Improves training efficiency

Why Multi-Head Attention Is Better Than Single Head

Single Attention	**Multi-Head Attention**
One focus	Multiple focuses
Limited patterns	Rich patterns
Weaker understanding	Stronger understanding

Why LLMs Use Multi-Head Attention

Multi-Head Attention enables:

- Context-aware representations
- Long-range dependency modeling

- Parallel computation
- Scalable training of large models

This is why **Transformers replaced traditional NLP models**.

Key Takeaways

- Multi-Head Attention performs multiple self-attentions in parallel
- Each head learns **different relationships**
- Outputs are combined for **richer representations**
- It is a core component of **Transformers and LLMs**

One-Line Exam / Interview Answer

> *Multi-Head Attention allows Transformers to attend to different parts of a sentence in multiple ways simultaneously, improving contextual understanding and representation power.*