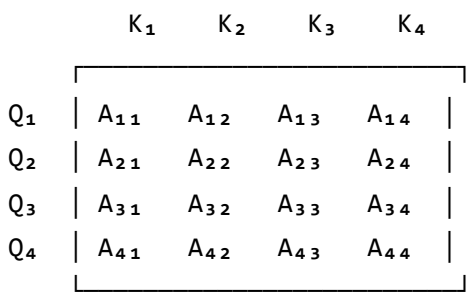
Attention Matrix Properties

Setup: 4×4 Attention Matrix

Consider a sentence with 4 tokens: **[Token$_1$, Token$_2$, Token$_3$, Token$_4$]**

The attention matrix **A** has shape (4, 4):



Question 1: What does element A_{23} represent?

Answer: A_{23} represents the **attention weight** that **Token$_2$** (query) pays to **Token$_3$** (key).

Interpretation:

- **Row index (2)**: The token that is "paying attention" (the query position)
- **Column index (3)**: The token that is "being attended to" (the key position)
- **A_{23}**: How much Token$_2$ focuses on Token$_3$ when computing its representation

Reading the Matrix:

- **Rows**: Each row shows where one token attends to ALL other tokens
- **Columns**: Each column shows which tokens attend to a specific token
- **A_{23}**: Element at row 2, column 3

Question 2: If $A_{23} = 0.8$, what does it mean?

Answer: Token_2 is **strongly attending** to Token_3.

Specific Meaning:

1. **High attention weight**: 0.8 out of maximum 1.0 is very high
2. **Token_2's representation** will be heavily influenced by Token_3's value vector
3. **80% of the "focus"** from Token_2 goes to Token_3

Context (Remember: Rows Sum to 1)

Since softmax ensures row sums equal 1:

$$A_{21} + A_{22} + A_{23} + A_{24} = 1.0$$

If $A_{23} = 0.8$, then:

$$A_{21} + A_{22} + A_{24} = 0.2$$

This means:

- Token_2 pays **80% attention** to Token_3
- Token_2 pays only **20% combined** to Token_1, Token_2 (itself), and Token_4
- Token_3 is **dominant** in determining Token_2's output representation

Example

Sentence: "The cat chased mice"

- Token_1 = "The"
- Token_2 = "cat"
- Token_3 = "chased"
- Token_4 = "mice"

If $A_{23} = 0.8$:

- "cat" strongly attends to "chased"
- When computing the representation for "cat", the model heavily incorporates information from "chased"

- This captures the subject-verb relationship
- The output for "cat" will strongly reflect what action it's performing

Summary

Element	Query (Row)	Key (Column)	Meaning
A_{23}	$Token_2$	$Token_3$	How much $Token_2$ attends to $Token_3$
$A_{23} = 0.8$	"cat"	"chased"	80% of "cat"'s representation comes from "chased"

Key insight: High attention weight (0.8) means strong token interaction - the query token's output will be dominated by the key token's value.