

Attention Weight Properties: True/False

1. The rows of an attention weight matrix always sum to 1.

TRUE ✓

Explanation: Softmax normalization is applied along each row (axis=-1), ensuring all attention weights for a given query sum to 1. This makes each row a probability distribution over all keys.

2. If Q and K are identical, the attention weights will always be uniform.

FALSE X

Explanation: When $Q = K$, we get $Q \cdot Q^T$, which produces different scores for different positions. Each query will have highest attention to itself (diagonal elements) and varying attention to other positions based on similarity. Uniform weights only occur if all queries are orthogonal and equidistant.

3. Multi-head attention simply repeats the same attention mechanism multiple times.

FALSE X

Explanation: Each head uses **different learned projection matrices** (W_Q^h, W_K^h, W_V^h) to project into different subspaces. This allows different heads to learn different types of relationships (e.g., syntactic vs. semantic), not just repeat the same computation.