Masking Implementation

🔷 Padding Mask

Ignore padding tokens in attention

📐 Shape: **(batch_size, 1, 1, seq_len)**

🎯 Marks which positions are actual tokens vs padding

1	1	1	0	0

🔺 Look-Ahead Mask (Causal)

Prevent attending to future tokens

📐 **Upper triangular matrix** of -inf values

🔒 Ensures causality during decoding

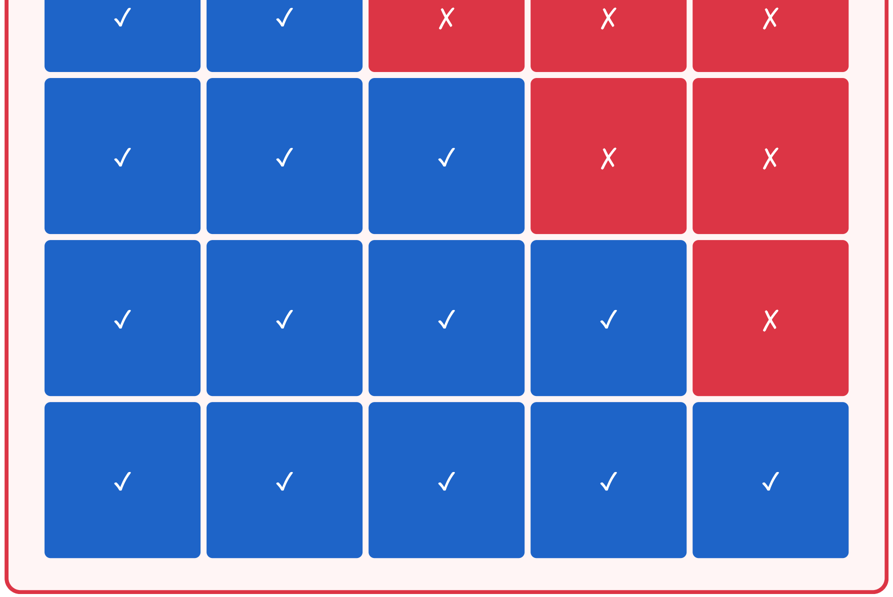
✓	✗	✗	✗	✗

⏱️ Application Timing

Applied **before softmax** in attention calculation

🔢 Mask Values

Use **-inf** for masked positions (becomes 0 after softmax)



✓	✓	✗	✗	✗
✓	✓	✓	✗	✗
✓	✓	✓	✓	✗
✓	✓	✓	✓	✓

🔗 **Combined Masks**

Multiple masks combined using **element-wise OR operation**

⚠

Proper masking is crucial for correct model behavior