(only in training as in inference there's no future tokens too look at) Types of masking LLM (ie setting which tokens do and dont interact commonly by setting their attention to 0) (see Pg 58)

2) no mask (bidirectional): This is the simplest case Every token can attend to every other token, no constraints on direction or position This is what Bert style encoders use, used in cross attention like translation for ex as encoder needs full context to understand everything

What it means in practice: The model can use future, past and current tokens to build representations, not a generative setup, its a fully contextual encoder

2) Casual mask (autoregressive): A casual mask is a triangular mask that prevents any token from seeing tokens to its right, it forces model too condition only On the past (ie model only relies on or uses only past information "previous words" to make prediction, we do this because we want model to predict the next token without cheating by looking at future tokens this makes training match how the model will generate text in real life (left to right)

What it means in practice: nessesorry for generation, forces time direction, forces predict next token, every token j can only attend to tokens 0 to j

3) Padding mask: this mask hides the padding tokens that exist only because batchs need to be equal length if you do not mask padding the model wastes attention on meaning less zeros. purpose: Prevent garbage attention:

Ex sentence 1: I love cats
sentence 2: Dogs bark
} batch 1 → tokenized ⇒ I love cats
Dogs bark <Pad> ⇒ Padding mask S1: 111
S2: 110
↓
can be 4,5 etc len where 3 ↗

→ because ML uses tensor so batch must be rectangular

4) MLM mask (for masked language modeling): Bert style pretraning masks out a subset of tokens so the model must reconstruct them, the mask is not a directional constraint, it is a visibility constraint, you hide some token positions in input and force model to recover them. model still bidirectional except for masked slots (tokens replaced with <mask>) model sees right/left context but must guess missing token
- used in pretraning to help understand language in bidirectional way not just knowing next word predictions (left context). Ex: cat sat on mat ⇒ cat <mask> on mat ⇒ by learning masked words it helps model learn words from later text (left + right context)

5) combined mask (used in real models): combining two or more masks for ex: transformers use casual mask + padding mask aka "attention mask"

6) span masking ⟶ instead of masking random random tokens you mask contiguous blocks (spans) this lets models learn better long-range understanding (variant of MLM masking bidirectional and not for generation)

↗ helps model learn to predict phrases not just single words

Ex I Love eating pizza on Friday ⇒ I 120 love <mask> <mask> on Friday

7) prefix/ prompt masking

Scanned with CamScanner