MultiDecode

June 7, 2025

Ted Kyi and Roger Stager

MultiDecode implementation details

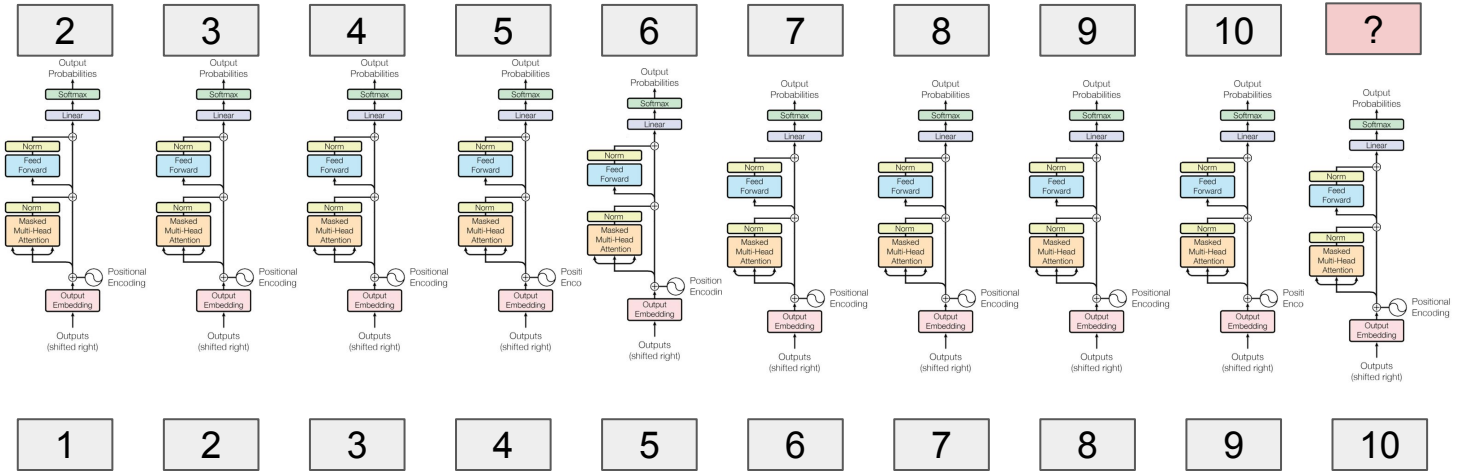
Agenda

Overview of concepts

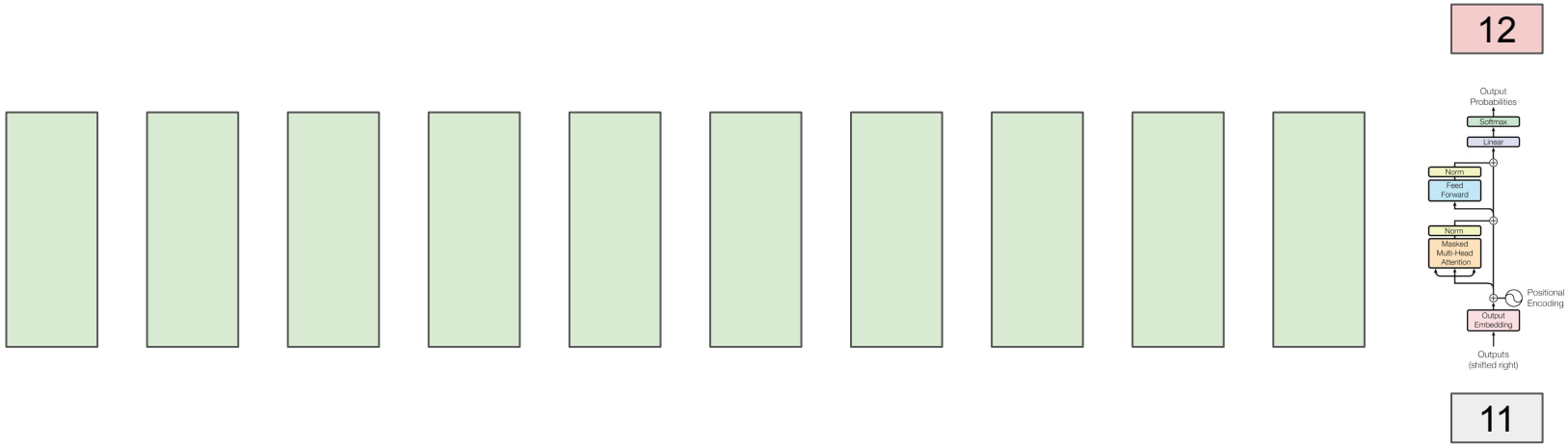
POC implementation walkthrough

Performance

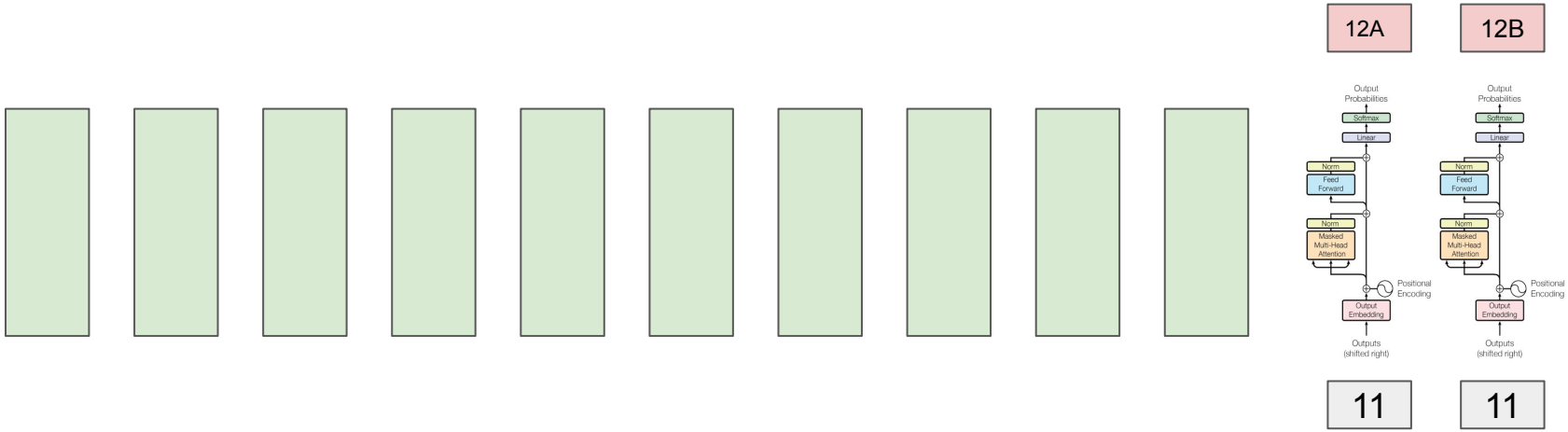
Training vs Inference (generation)



Inference (generation)

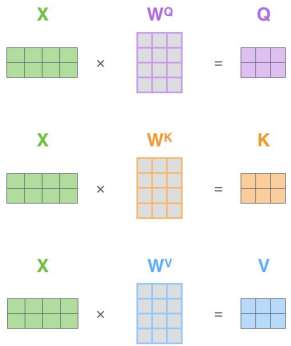


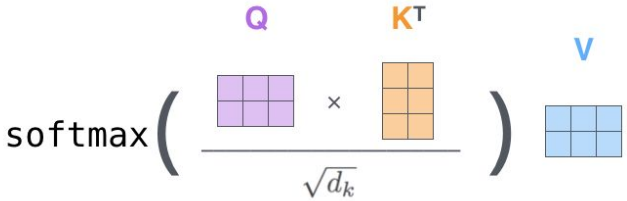
Multi Token Decoding



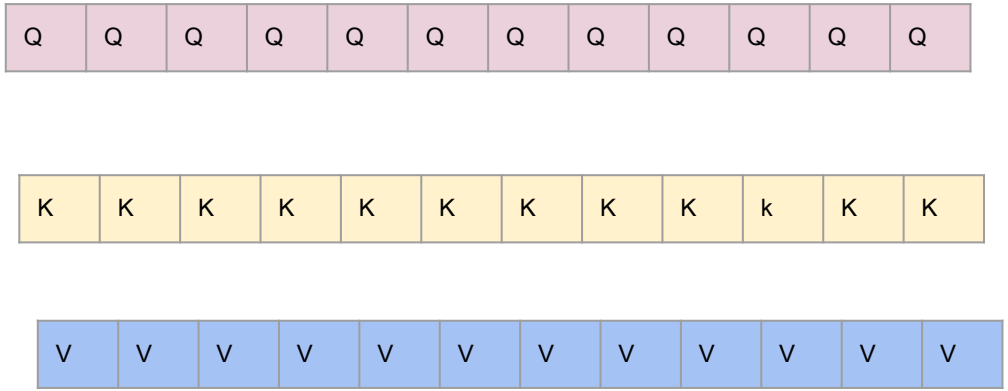
Attention

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

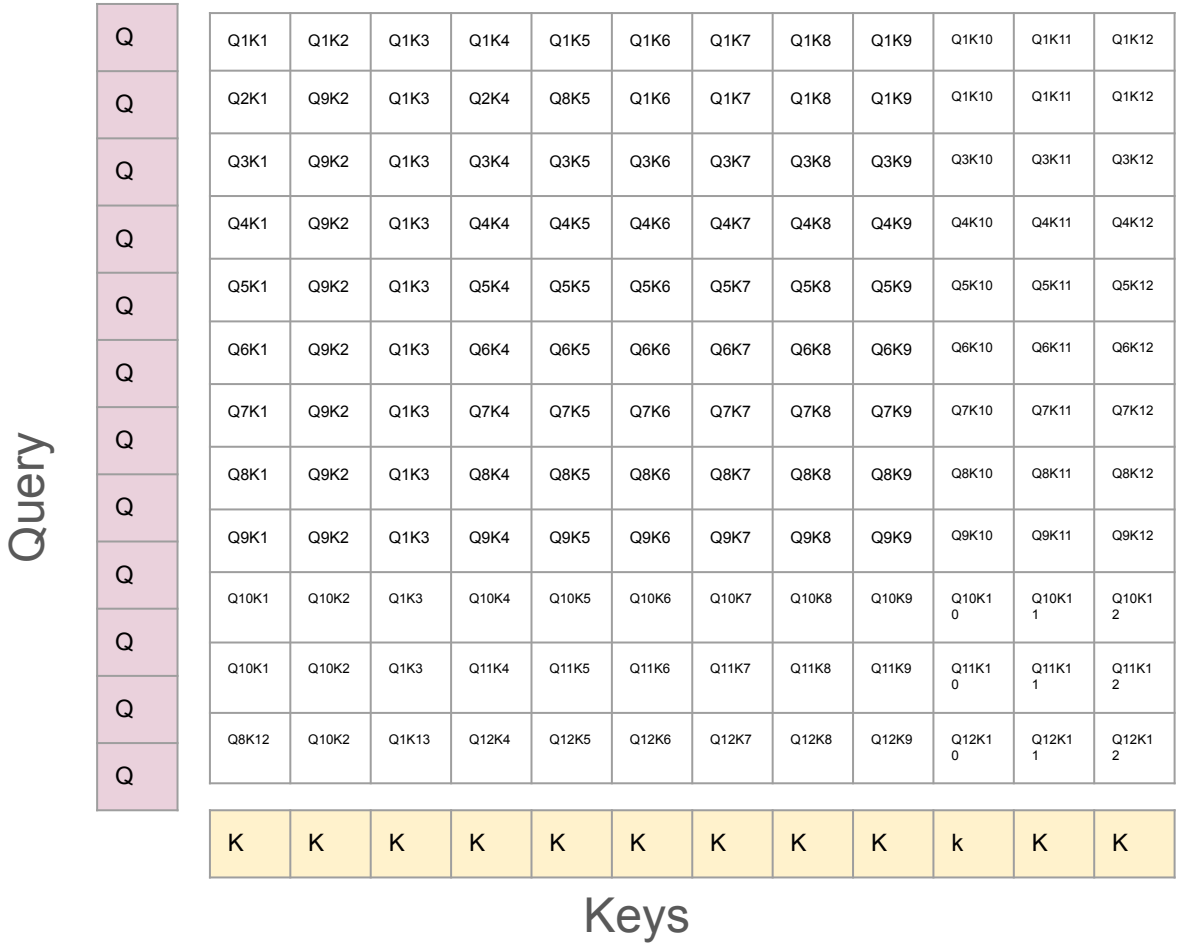




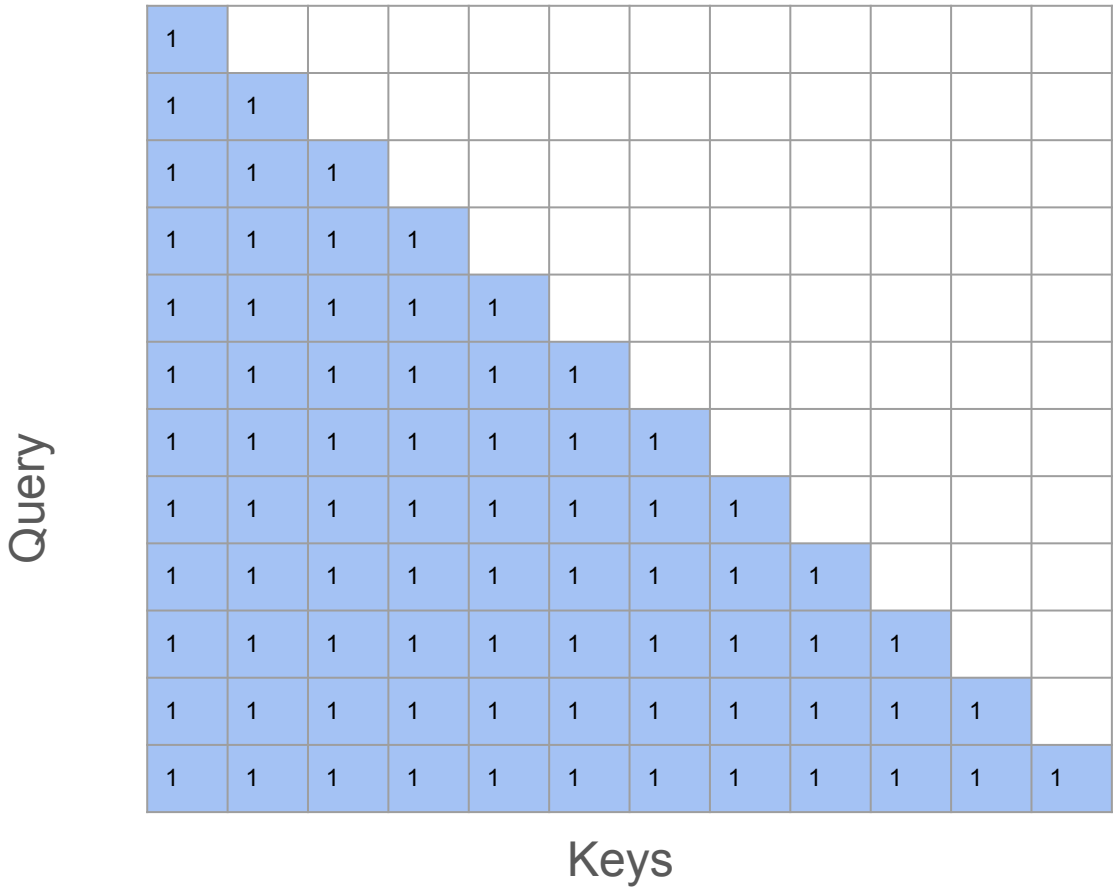
Calculate Q KV for full seqence



Attention: Outer product between Query and Key

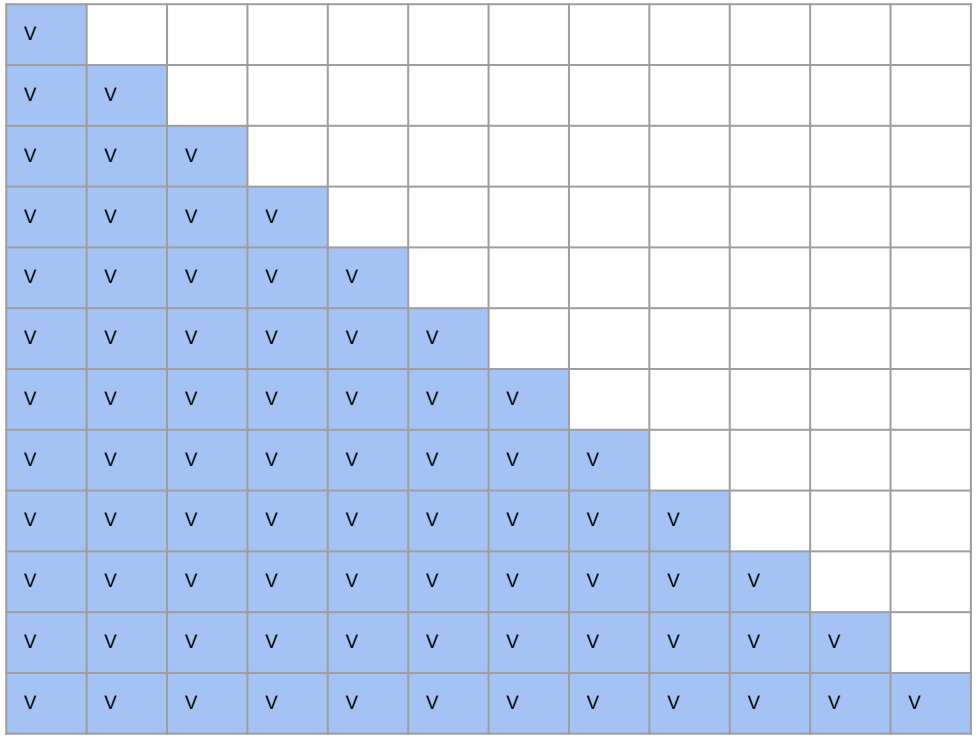


Causal attention mask

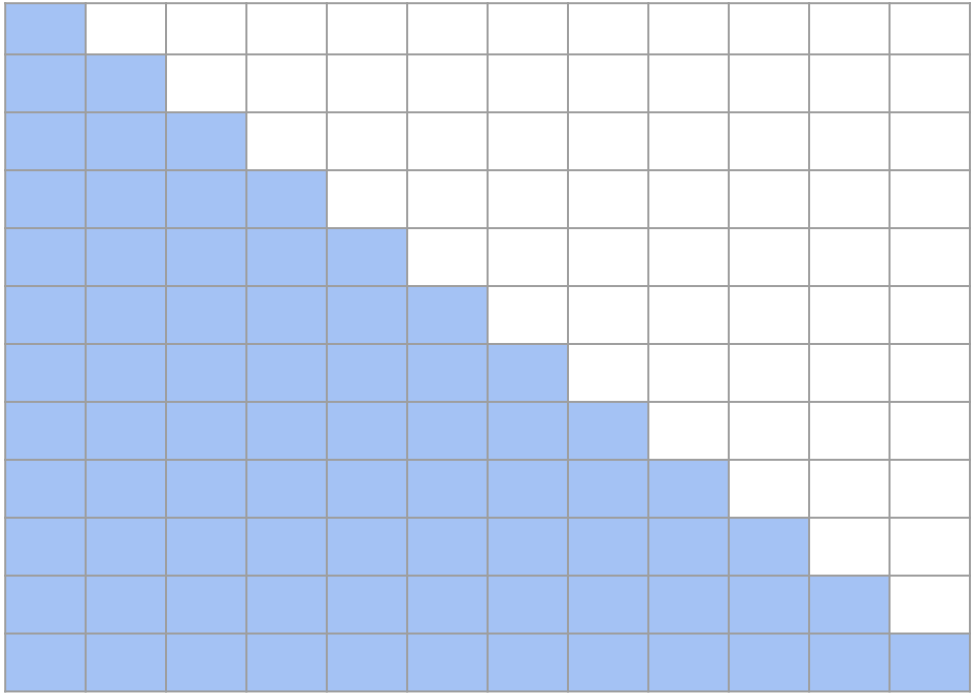


After Causal mask

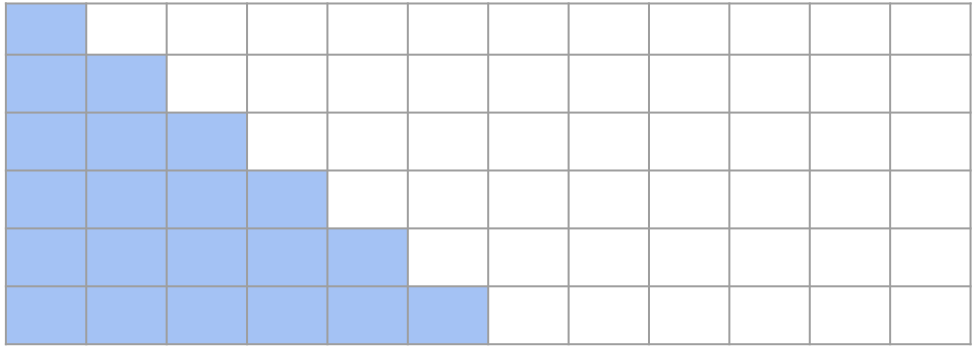
Q1K1											
Q2K1	Q9K2										
Q3K1	Q9K2	Q1K3									
Q4K1	Q9K2	Q1K3	Q4K4								
Q5K1	Q9K2	Q1K3	Q5K4	Q5K5							
Q6K1	Q9K2	Q1K3	Q6K4	Q6K5	Q6K6						
Q7K1	Q9K2	Q1K3	Q7K4	Q7K5	Q7K6	Q7K7					
Q8K1	Q9K2	Q1K3	Q8K4	Q8K5	Q8K6	Q8K7	Q8K8				
Q9K1	Q9K2	Q1K3	Q9K4	Q9K5	Q9K6	Q9K7	Q9K8	Q9K9			
Q10K1	Q10K2	Q1K3	Q10K4	Q10K5	Q10K6	Q10K7	Q10K8	Q10K9	Q10K10		
Q10K1	Q10K2	Q1K3	Q11K4	Q11K5	Q11K6	Q11K7	Q11K8	Q11K9	Q11K10	Q11K11	
Q8K12	Q10K2	Q1K13	Q12K4	Q12K5	Q12K6	Q12K7	Q12K8	Q12K9	Q12K10	Q12K11	Q12K12

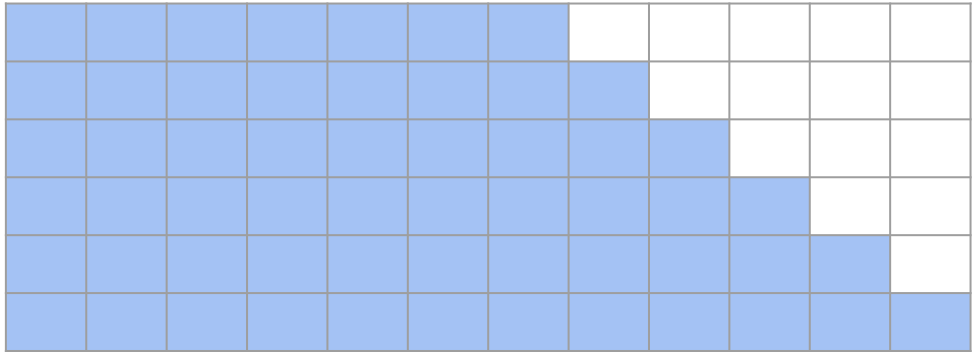

V
V V
V V V
V V V V
V V V V V
V V V V V V
V V V V V V V
V V V V V V V V
V V V V V V V V V
V V V V V V V V V V
V V V V V V V V V V V

This is the O(N^2)

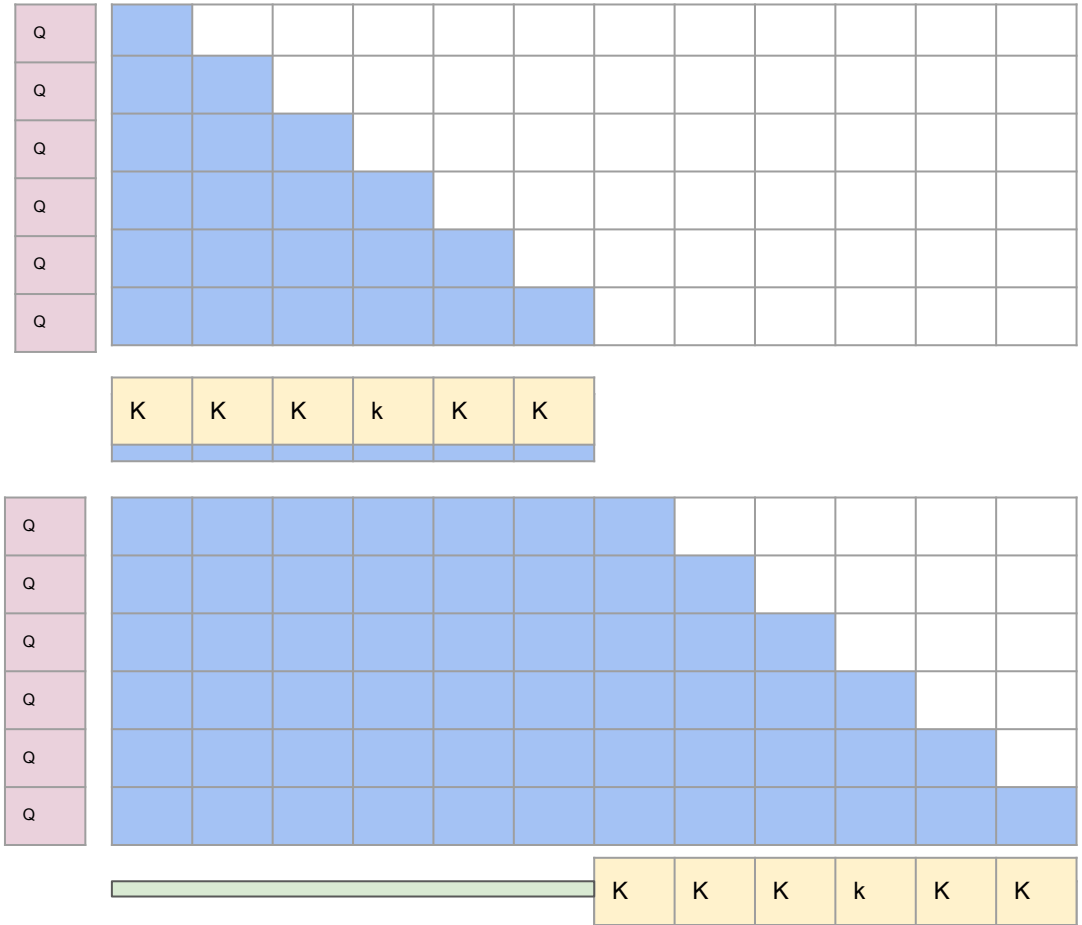


Chunk prefill



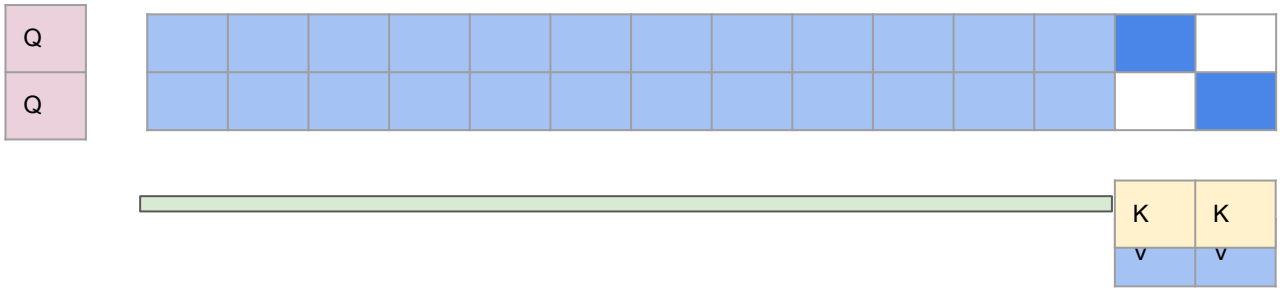


Chunk prefill

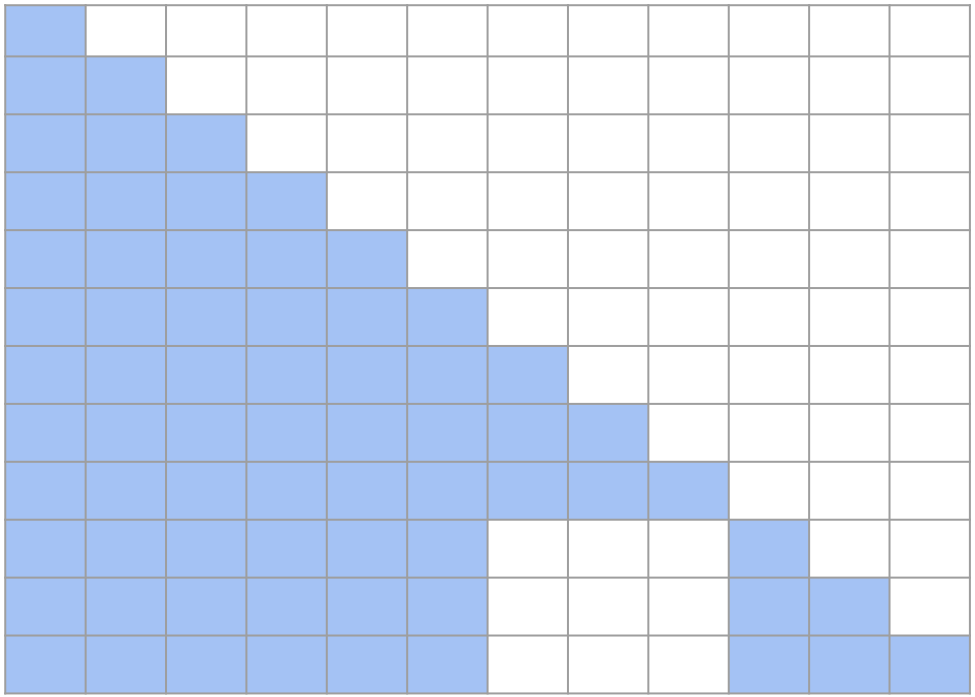


Classic generation

MultiDecode generation

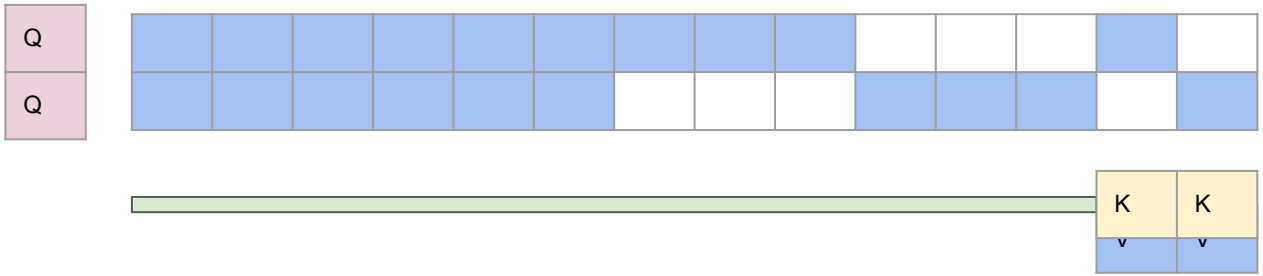


Prefill with multi prompt Attention mask

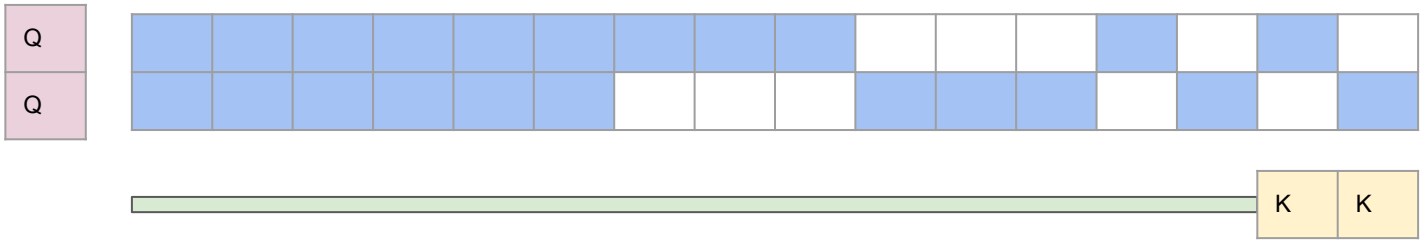


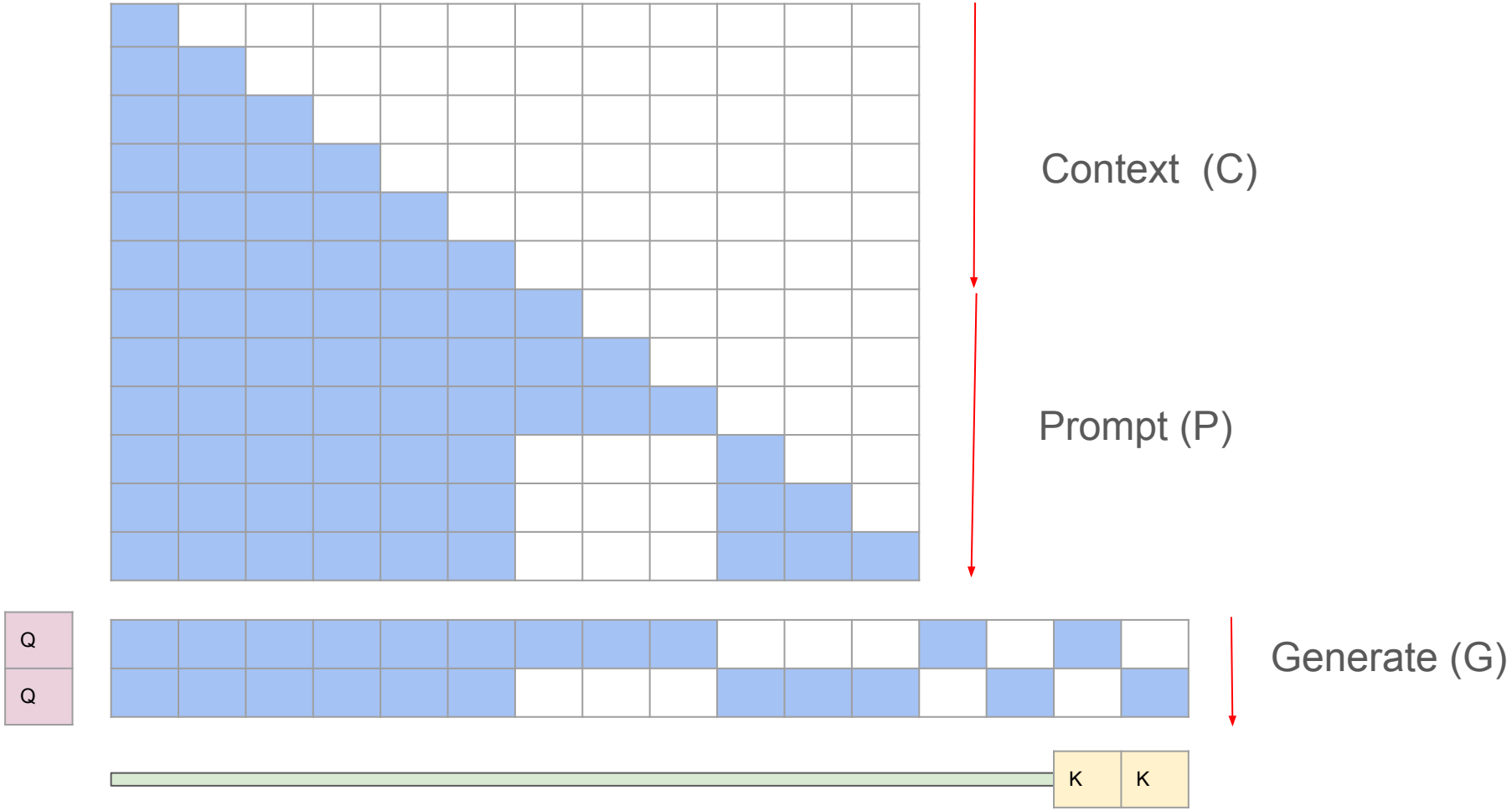
c1	c2	c3	c4	c5	c6	PA1	PA2	PA3	PB1	PB2	PB3

MultiDecode with MultiPrompt Generation

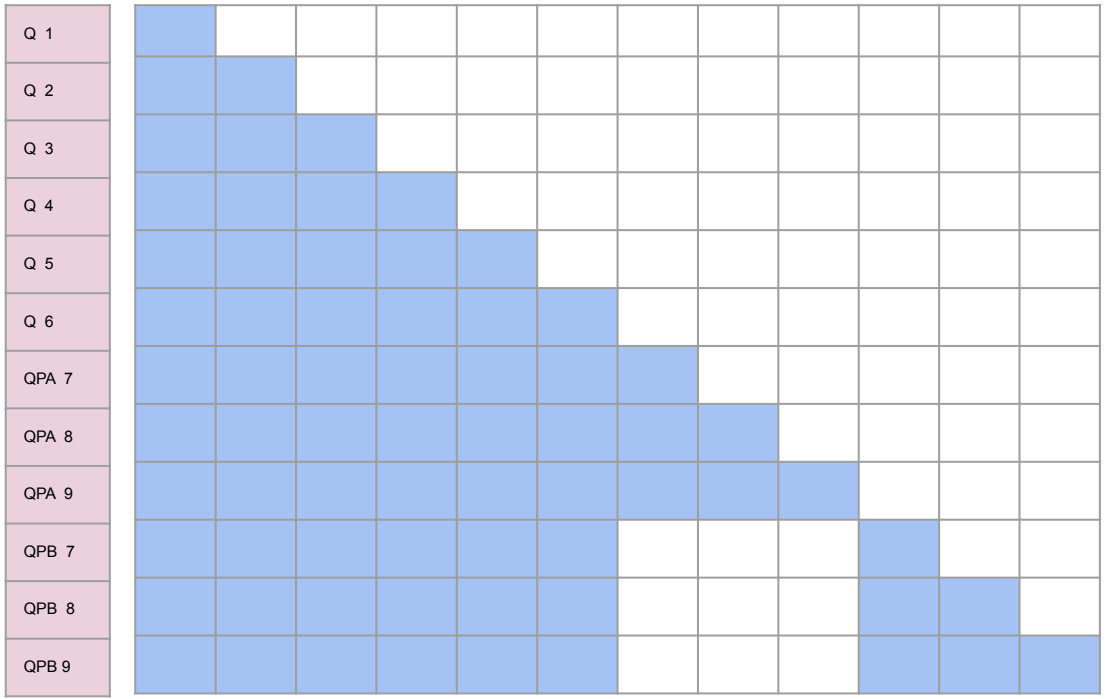


MultiDecode with MultiPrompt Generation



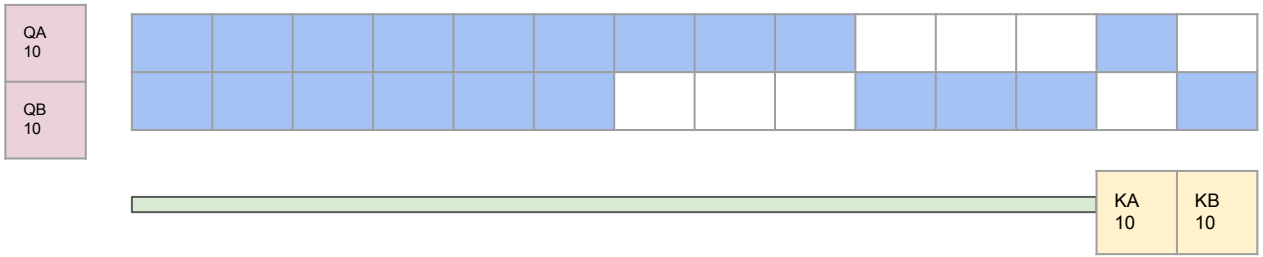

Context (C)
Prompt (P)
Q
Q
Generate (G)
K
K

Prefill with multi prompt Position embed



C 1	C 2	C 3	C 4	C 5	C 6	PA 7	PA 8	PA 9	PB 7	PB 8	PB 9

MultiDecode with MultiPrompt Position embed

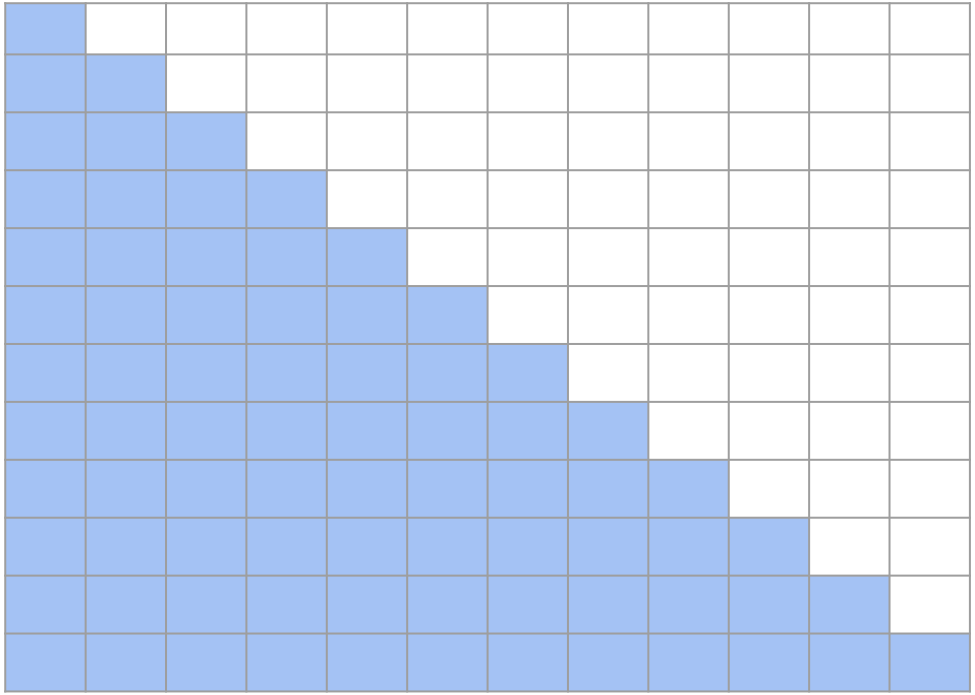


Naive multi-branch (beam)

1) Transformer beam search
    a) Expand the number of rows in the batch. 1 additional row per branch
2) Run the branches sequentially

On to the code

Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
K	K	K	K	K	K	K	K	K	k	K	K
V	V	V	V	V	V	V	V	V	V	V	V



Q
Q
Q
Q
Q
Q
Q
Q
Q
Q
Q
Q