

Efficient LLM Decoding

Large Language Models: Introduction and Recent Advances

ELL881 · AIL821



Yatin Nandwani
Research Scientist, IBM Research

Training Vs Inference in LLMs

Forward Pass through an LLM

Transformer based LLM (θ)

<S>	The	cat	sat	on	a	mat	</S>
0	1	2	3	4	5	6	7



Forward Pass through an LLM

Probability distribution over
all the tokens at each step
(simultaneously)

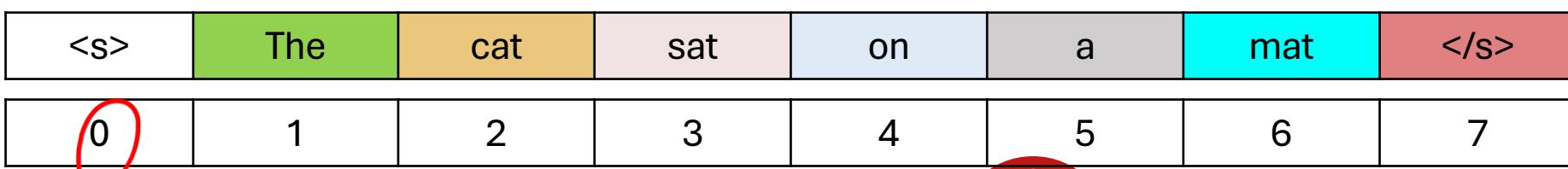
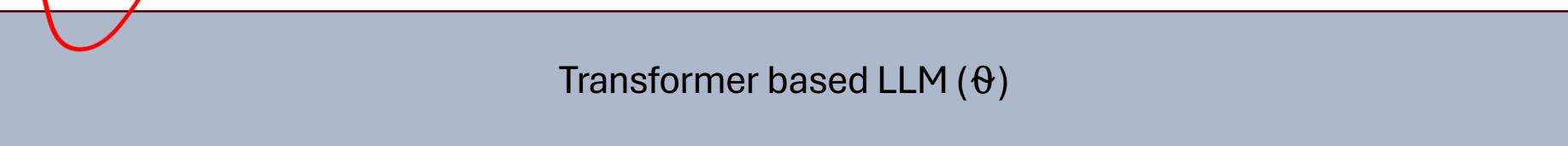
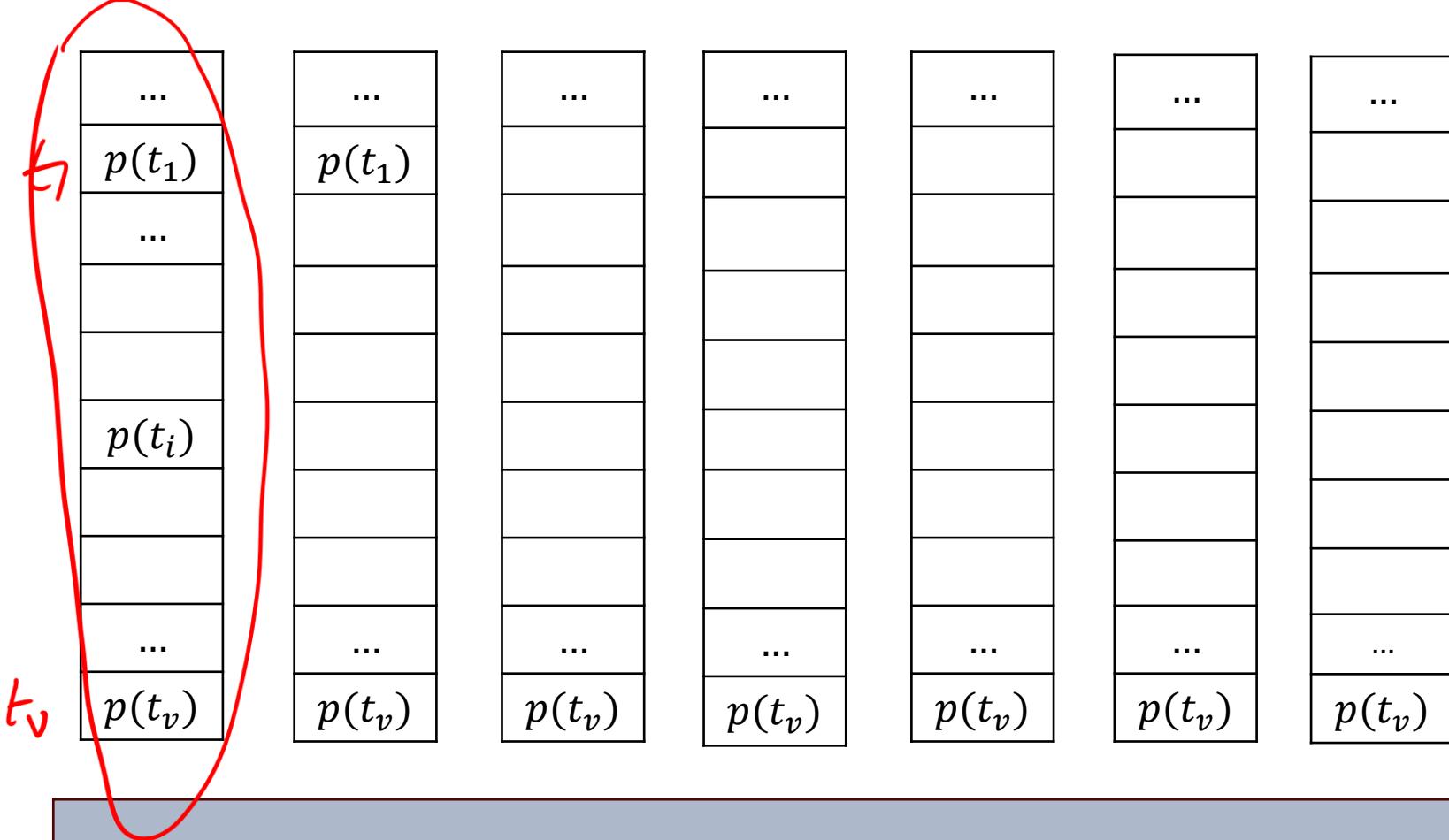
Transformer based LLM (θ)

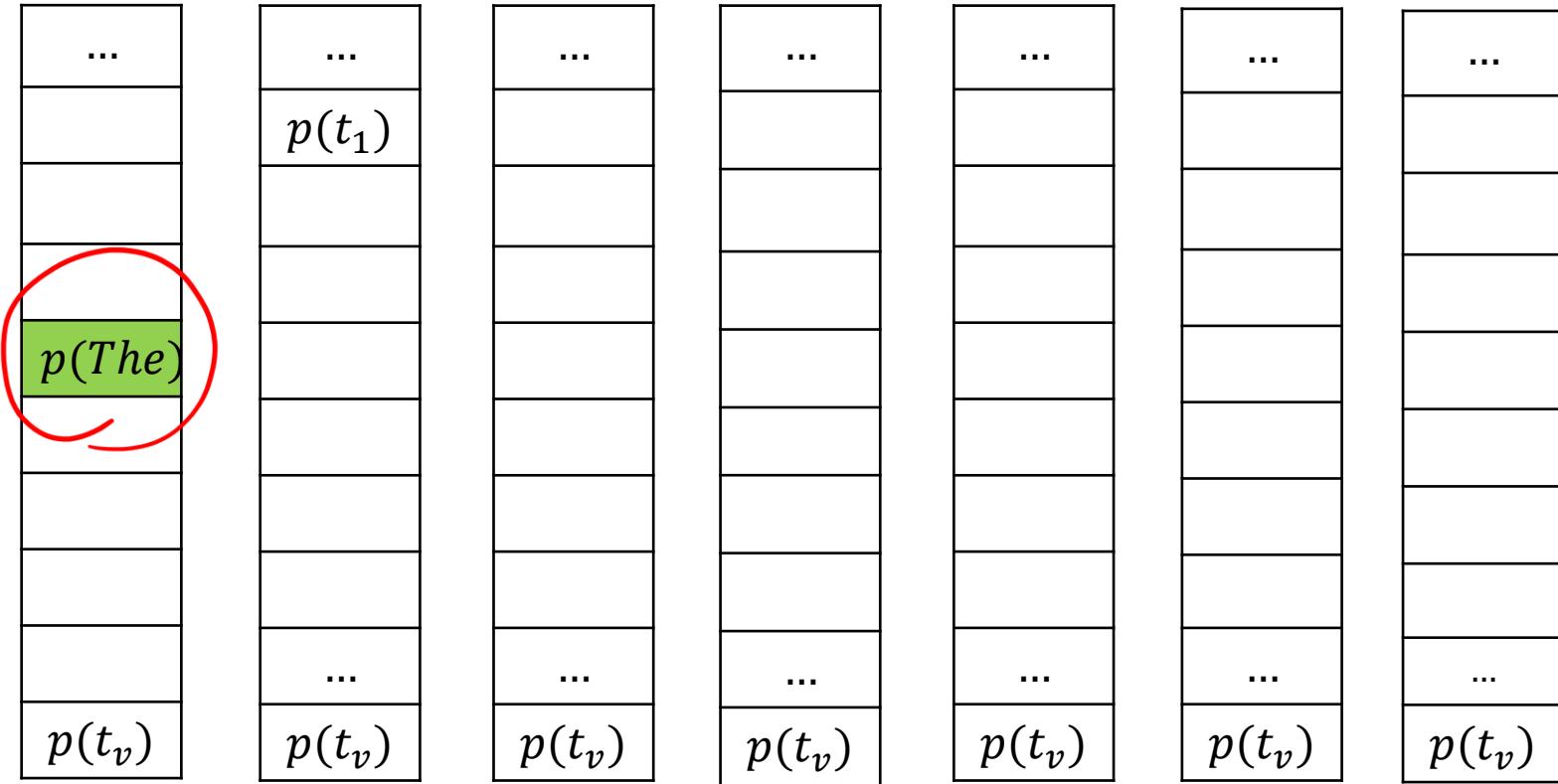
<s>	The	cat	sat	on	a	mat	</s>
0	1	2	3	4	5	6	7



Forward Pass through an LLM

Probability distribution over all the tokens at each step (simultaneously)





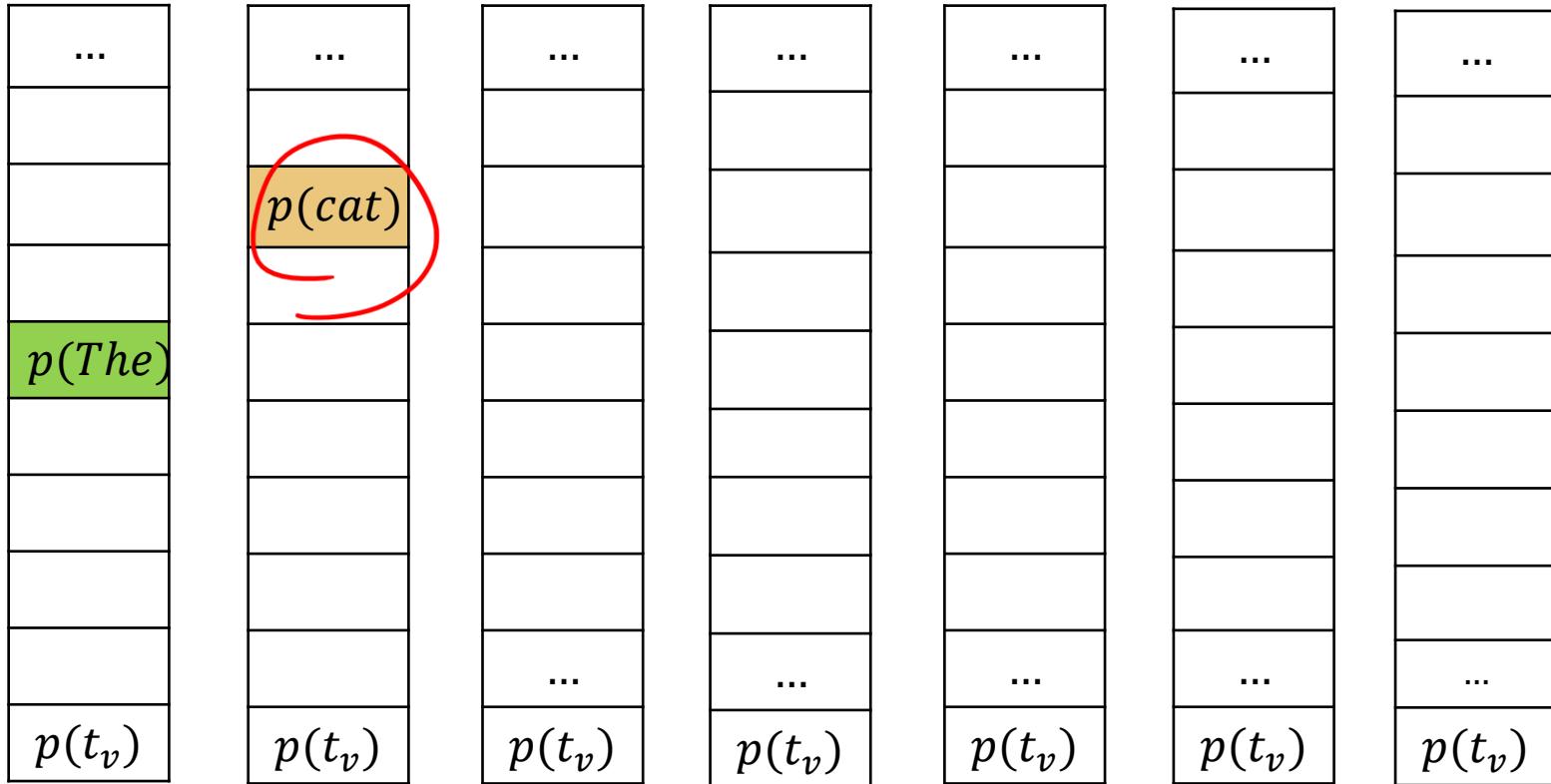
Forward Pass through an LLM

Train to maximize prob. of
The at step 0

Transformer based LLM (θ)

<S>	The	cat	sat	on	a	mat	</S>
0	1	2	3	4	5	6	7





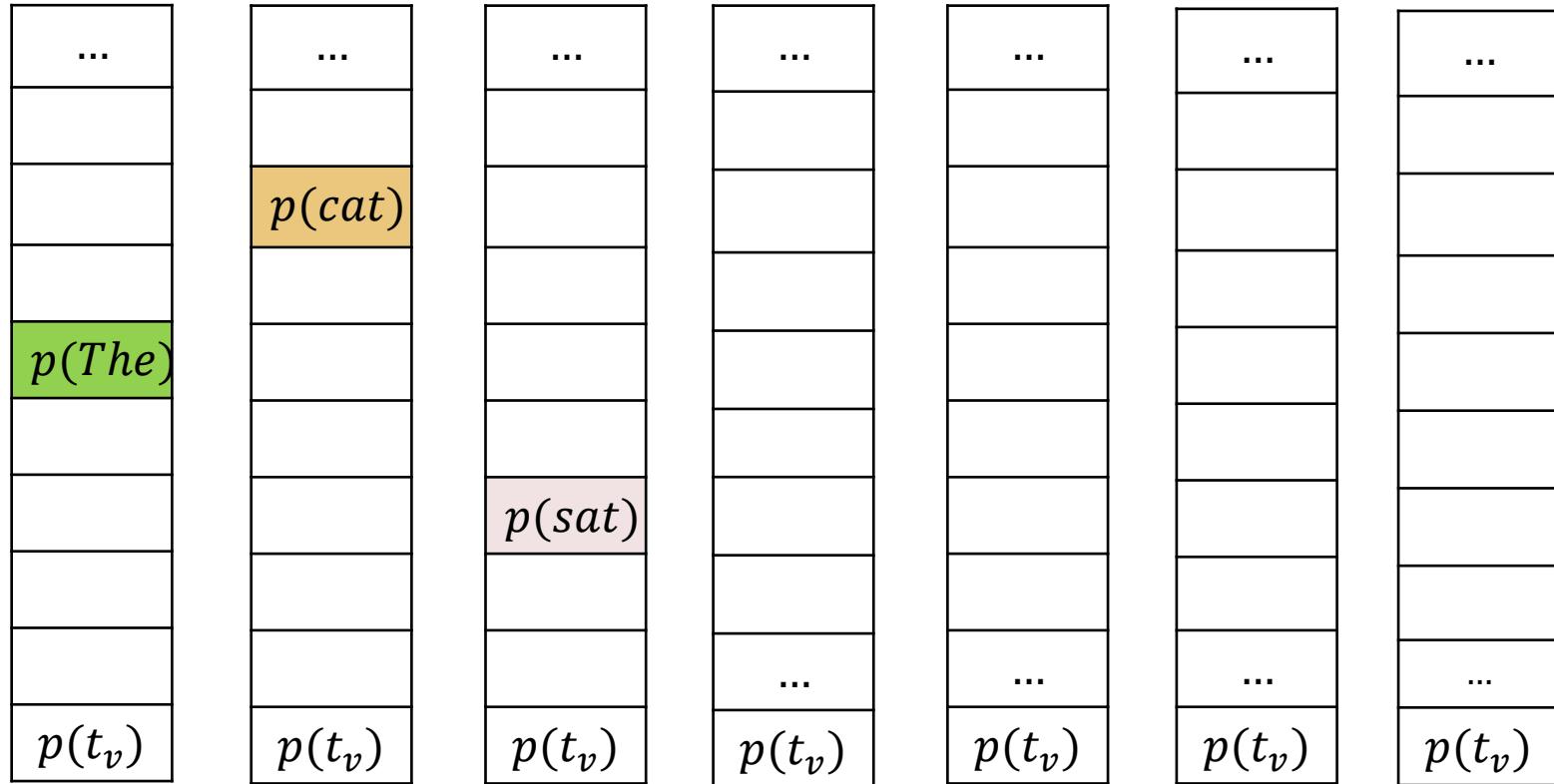
Forward Pass through an LLM

Train to maximize prob. of
cat at step 1

Transformer based LLM (θ)

<s>	The	cat	sat	on	a	mat	</s>
0	1	2	3	4	5	6	7





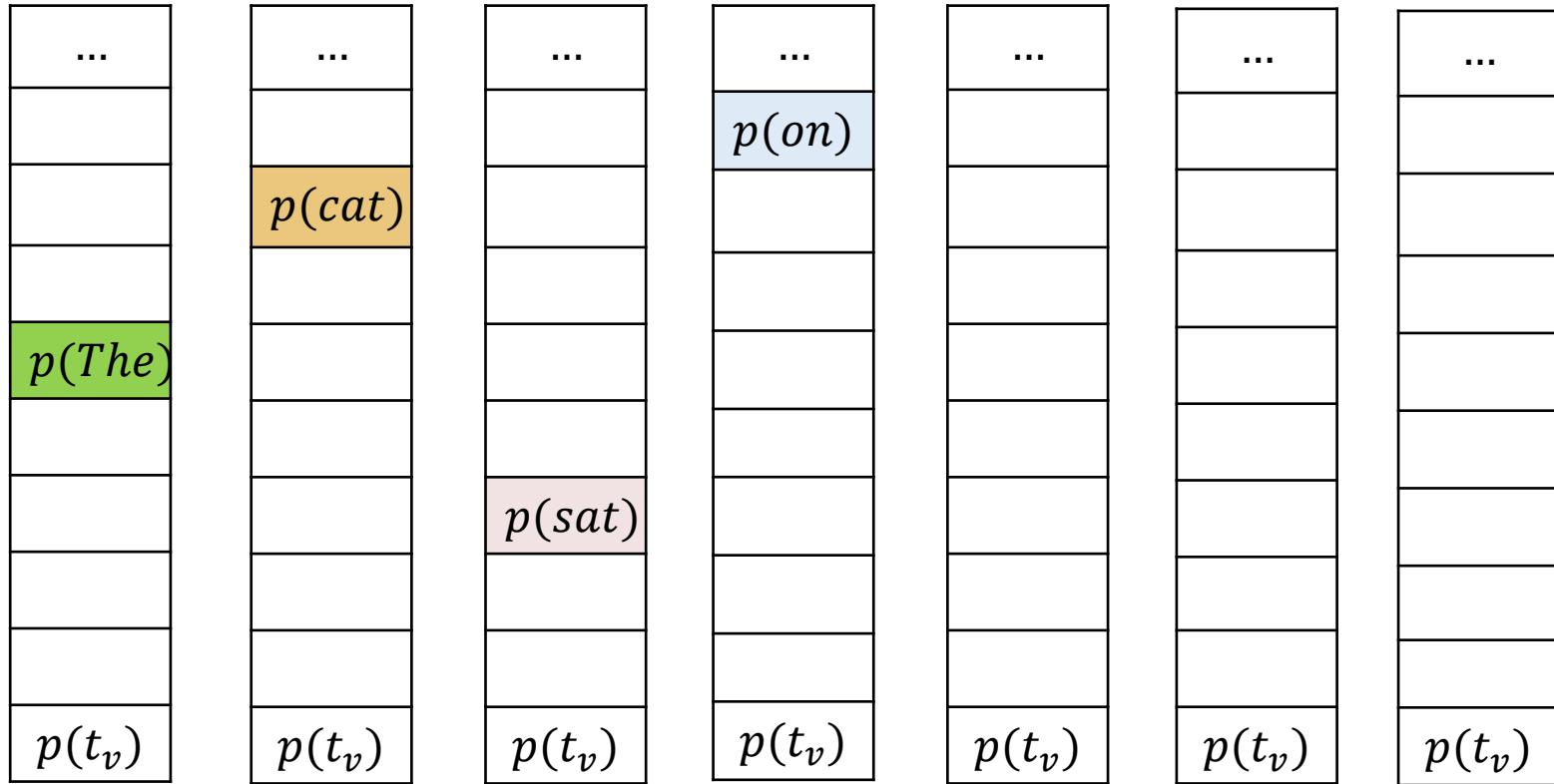
Forward Pass through an LLM

Train to maximize prob. of
sat at step 2

Transformer based LLM (θ)

< s >	The	cat	sat	on	a	mat	< /s >
0	1	2	3	4	5	6	7





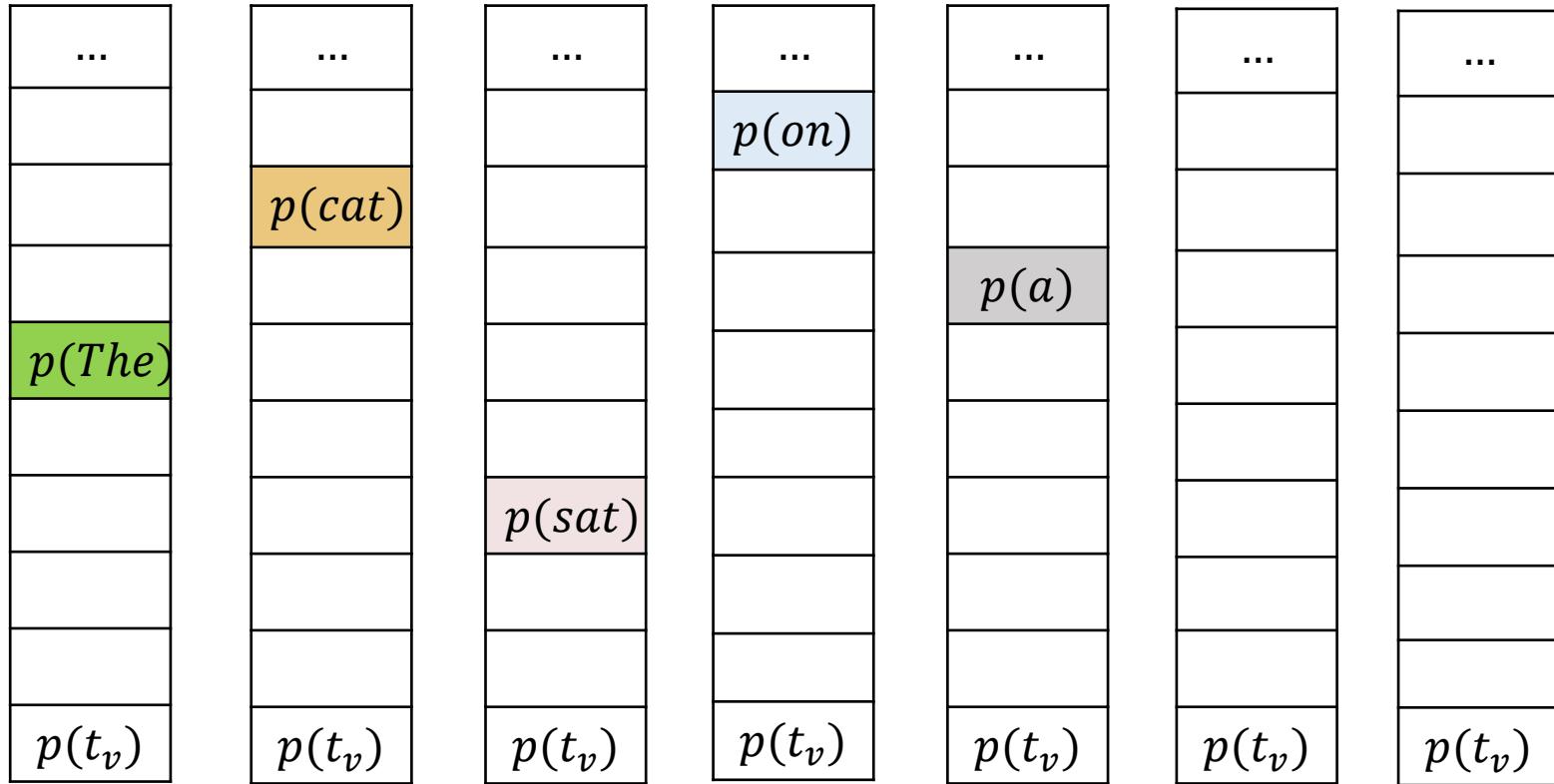
Forward Pass through an LLM

Train to maximize prob. of
on at step 3

Transformer based LLM (θ)

< s >	The	cat	sat	on	a	mat	< /s >
0	1	2	3	4	5	6	7





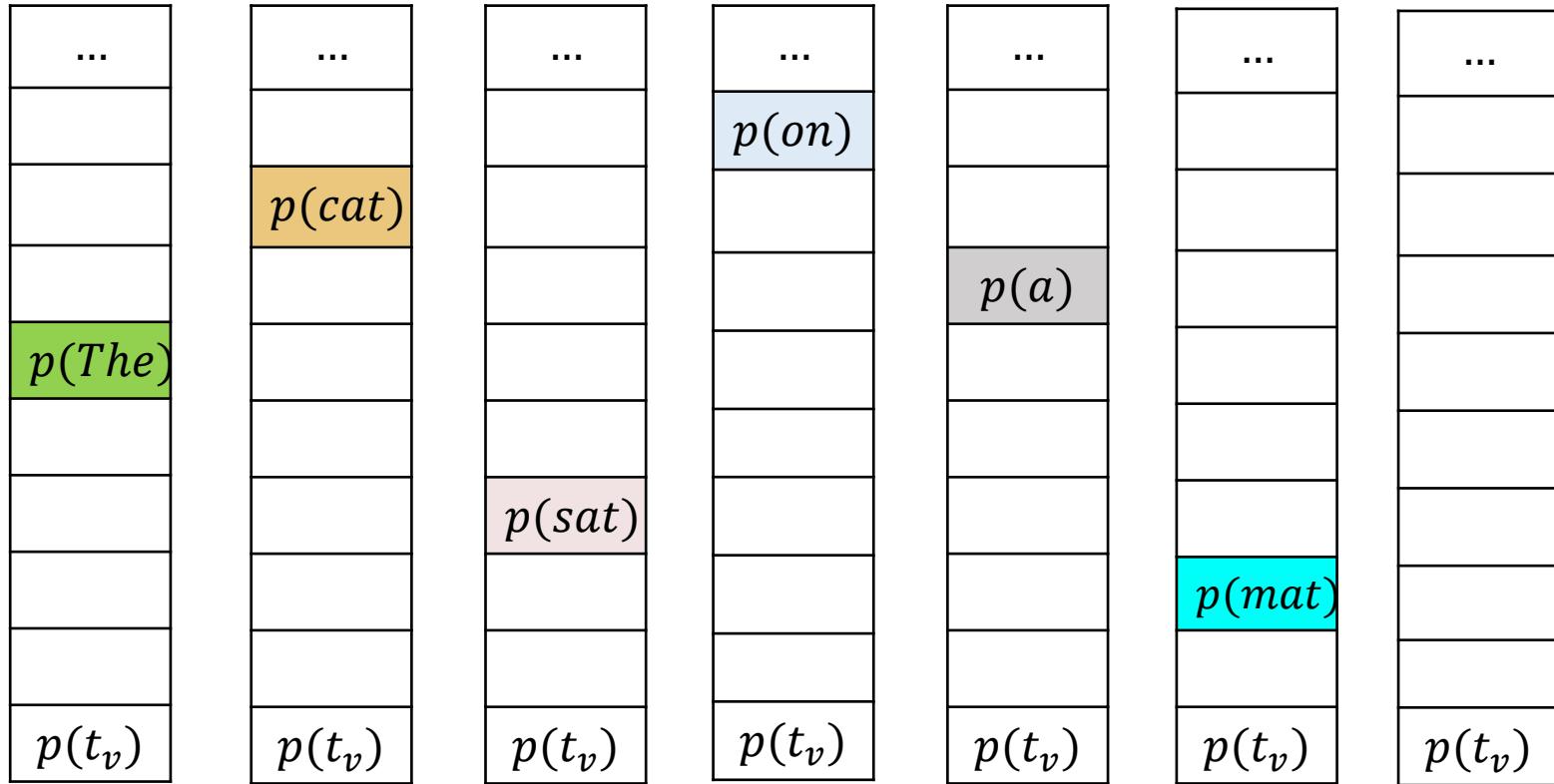
Forward Pass through an LLM

Train to maximize prob. of a at step 4

Transformer based LLM (θ)

<s>	The	cat	sat	on	a	mat	</s>
0	1	2	3	4	5	6	7





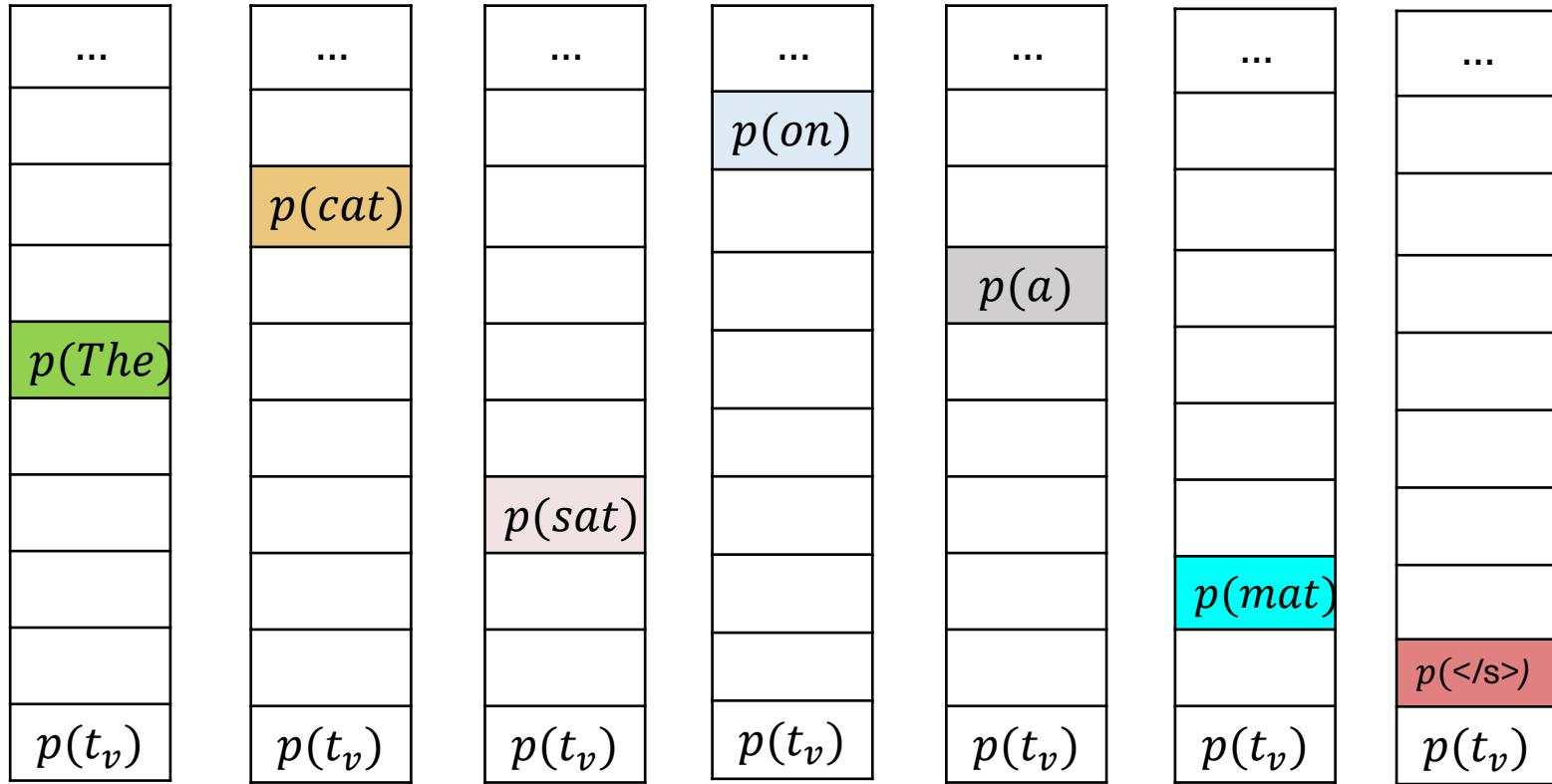
Forward Pass through an LLM

Train to maximize prob. of
mat at step 5

Transformer based LLM (θ)

<s>	The	cat	sat	on	a	mat	</s>
0	1	2	3	4	5	6	7





Forward Pass through an LLM

Train to maximize prob. of
 $</s>$ at step 6

Transformer based LLM (θ)

<s>	The	cat	sat	on	a	mat	</s>
0	1	2	3	4	5	6	7



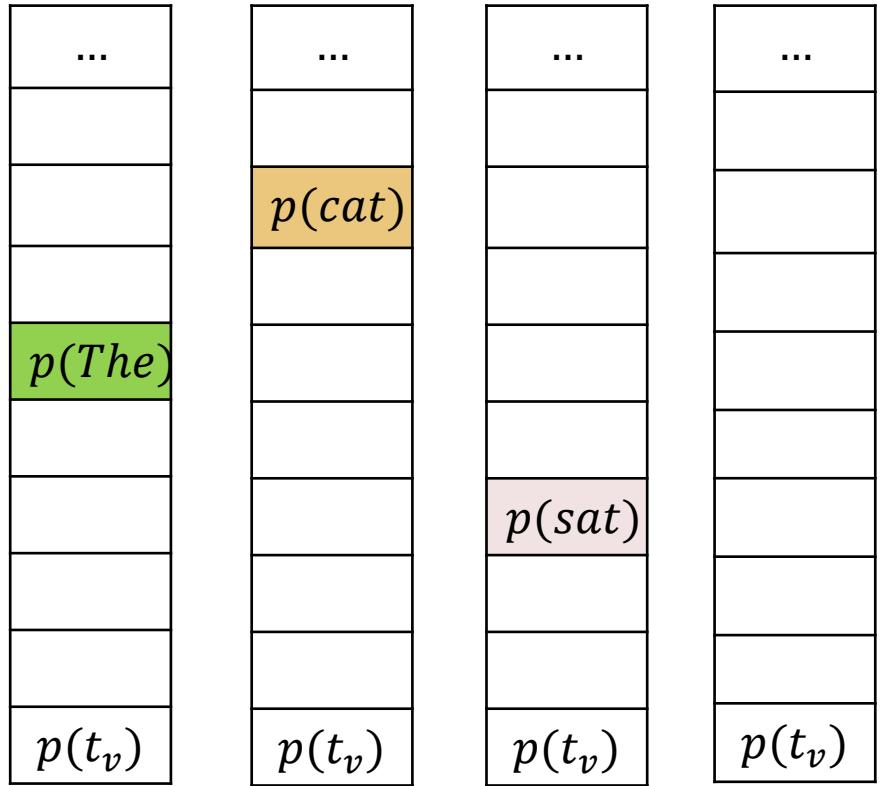
Inference through an LLM

Forward Pass (#1)

Transformer based LLM (θ)

<s>	The	cat	sat				
0	1	2	3	4	5	6	7





Inference through an LLM

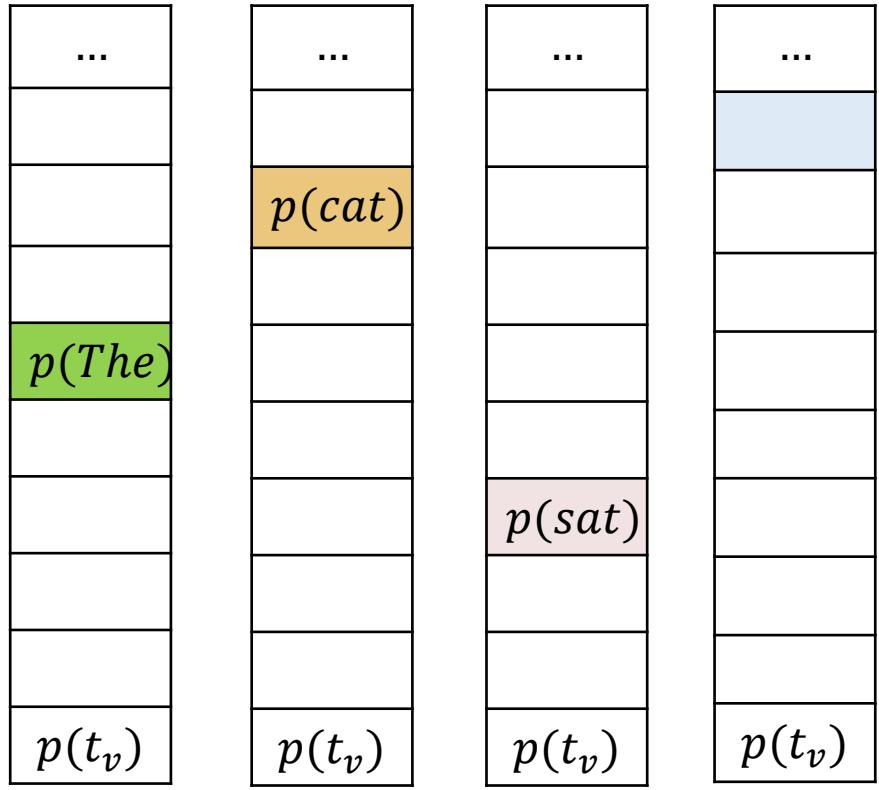
Prob. Dist. at all steps

Transformer based LLM (θ)

Sequence of tokens:

<s>	The	cat	sat				
0	1	2	3	4	5	6	7





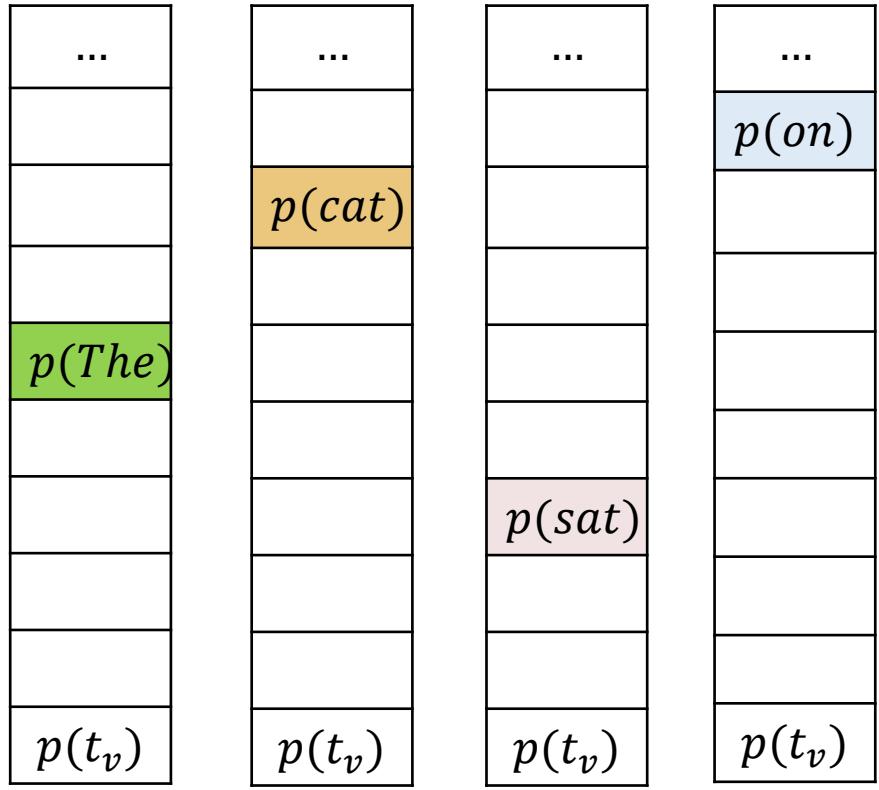
Inference through an LLM

Pick the token having max. probability at step 3

Transformer based LLM (θ)

<s>	The	cat	sat				
0	1	2	3	4	5	6	7





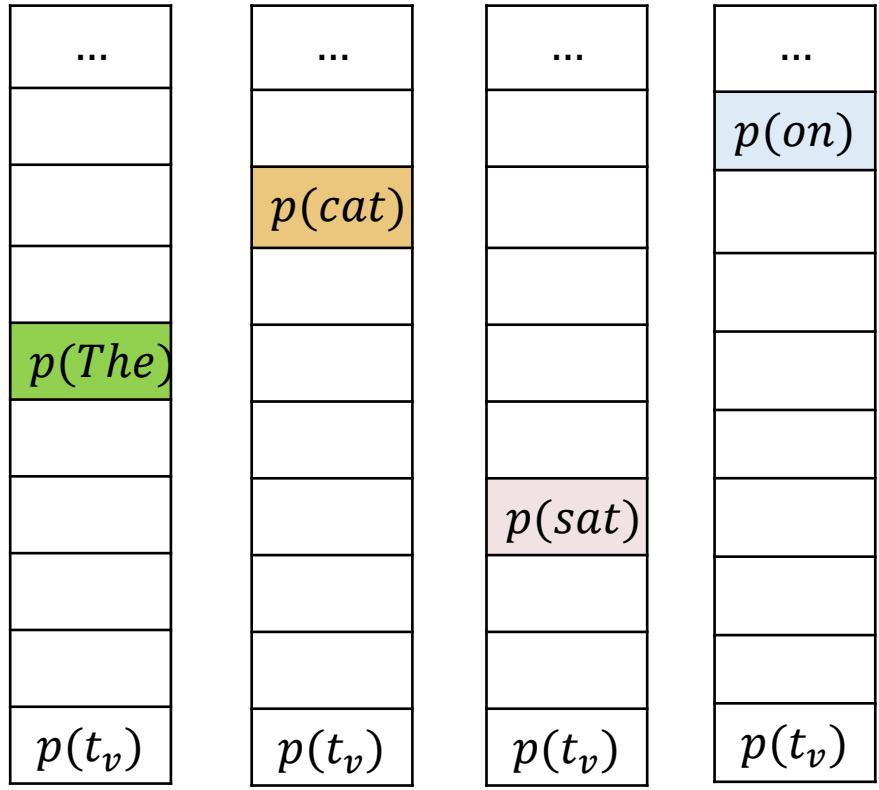
Inference through an LLM

Pick the token having max. probability at step 3

Transformer based LLM (θ)

<s>	The	cat	sat				
0	1	2	3	4	5	6	7





Inference through an LLM

Fill at step 4

Transformer based LLM (θ)

<s>	The	cat	sat	on			
0	1	2	3	4	5	6	7



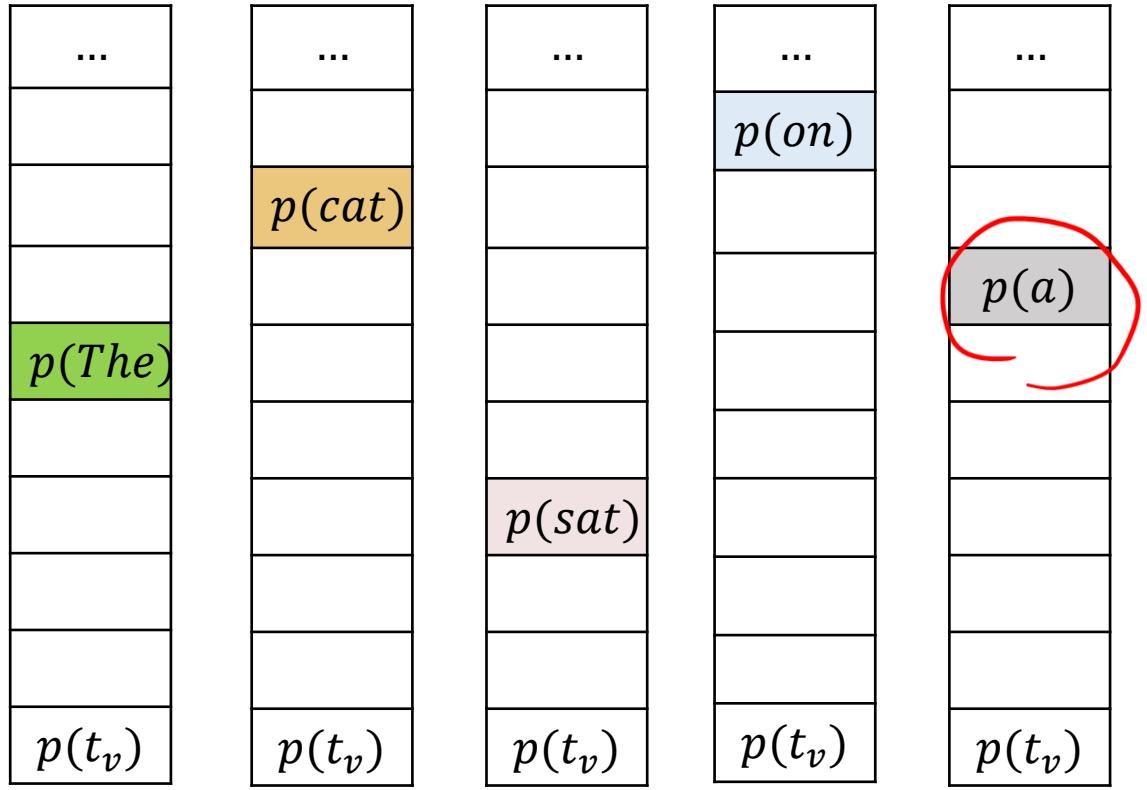
Inference through an LLM

Fwd. Pass (#2)

Transformer based LLM (θ)

<s>	The	cat	sat	on				
0	1	2	3	4	5	6	7	





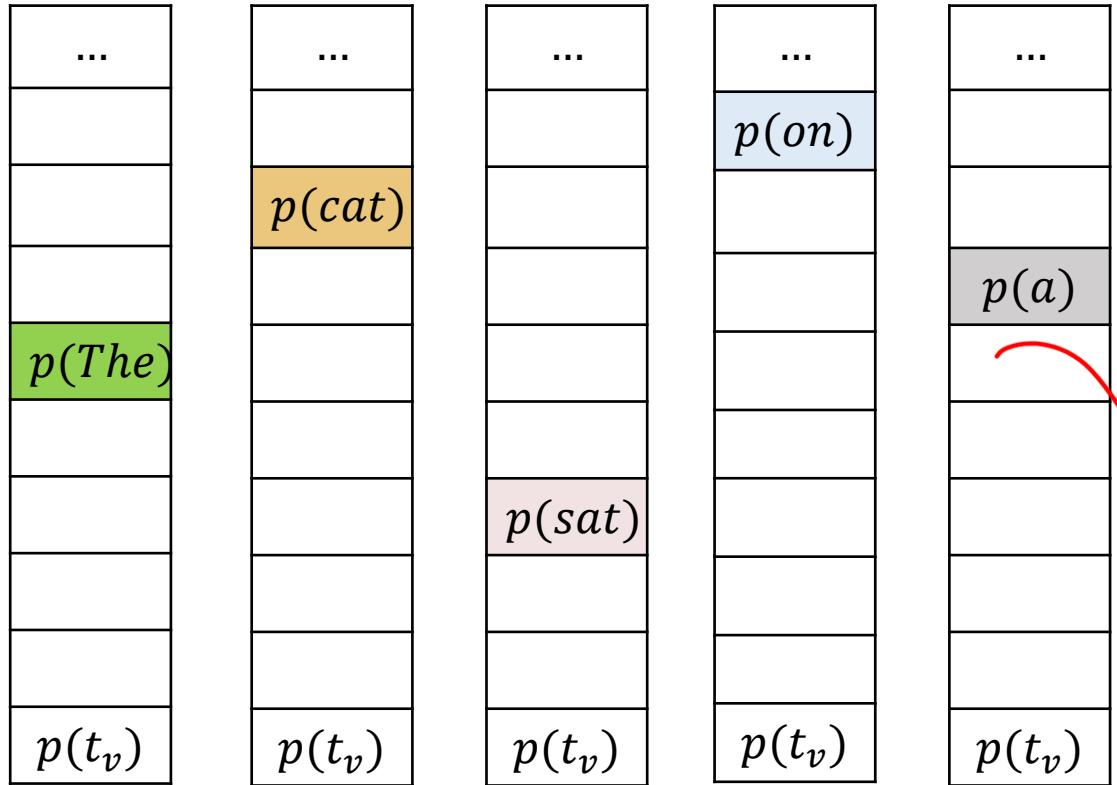
Inference through an LLM

Fwd. pass (#2) to get distribution at step 4

Transformer based LLM (θ)

<s>	The	cat	sat	on			
0	1	2	3	4	5	6	7





Inference through an LLM

Fill at step 5

Transformer based LLM (θ)

<S>	The	cat	sat	on	a		
0	1	2	3	4	5	6	7



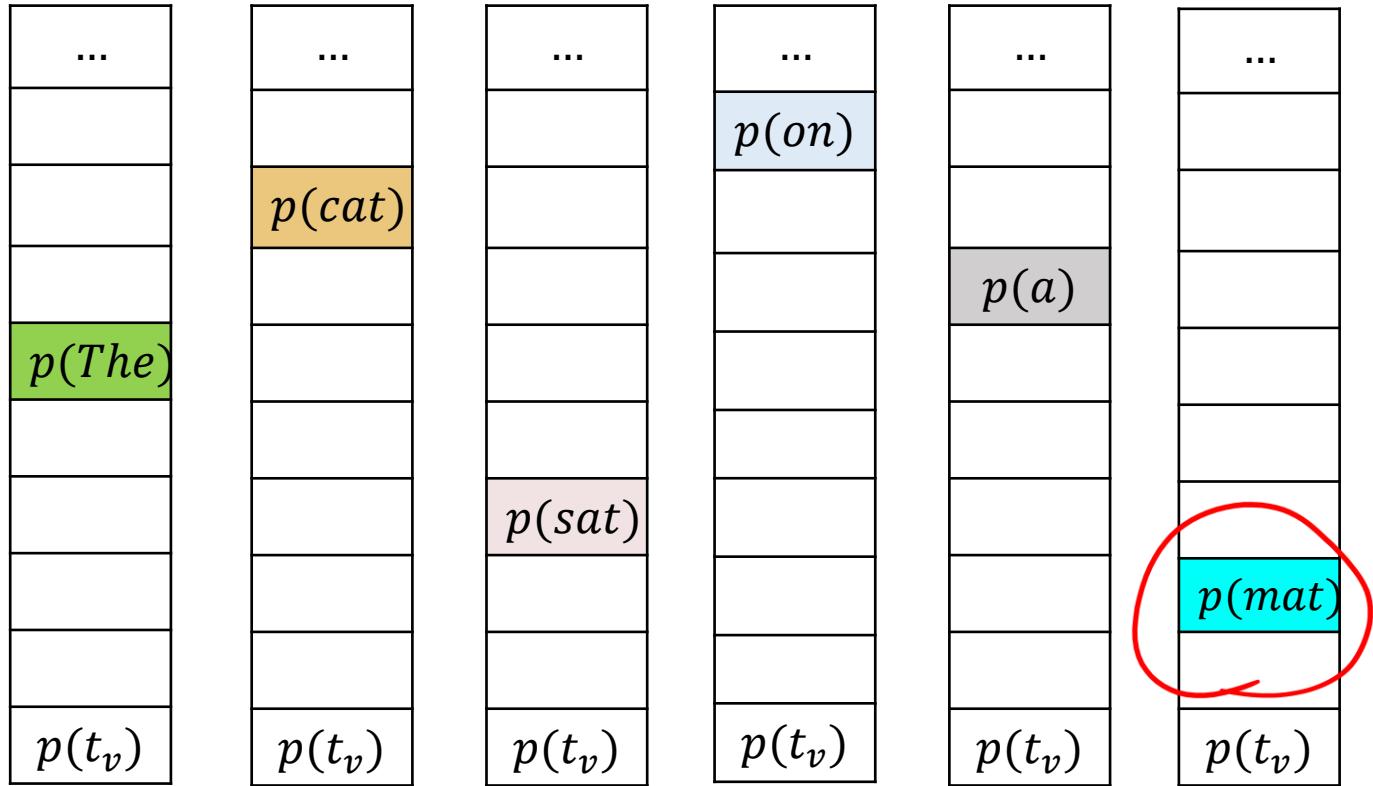
Inference through an LLM

Fwd. pass again (#3)

Transformer based LLM (θ)

<s>	The	cat	sat	on	a		
0	1	2	3	4	5	6	7





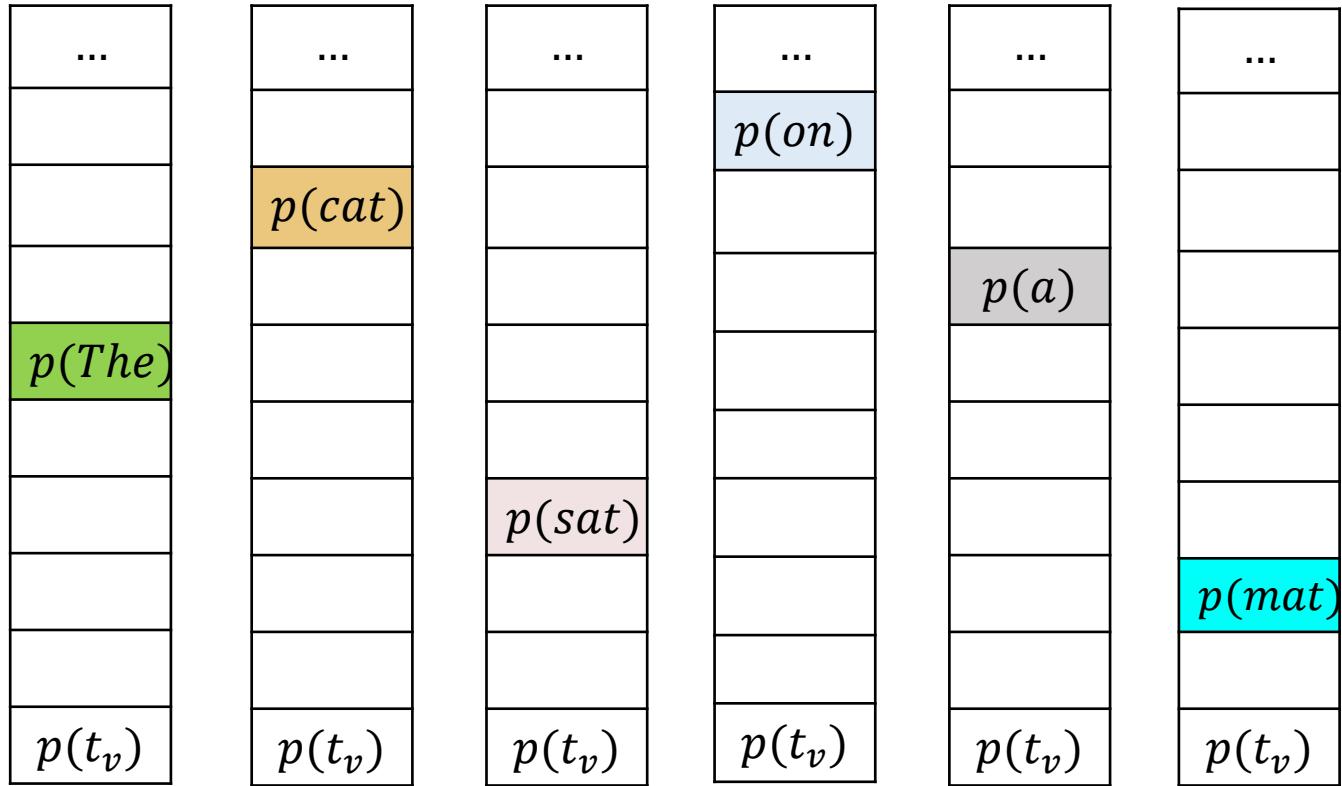
Inference through an LLM

Fwd. pass again (#3)

Transformer based LLM (θ)

<s>	The	cat	sat	on	a		
0	1	2	3	4	5	6	7





Inference through an LLM

Fill at step 6

Transformer based LLM (θ)

<s>	The	cat	sat	on	a	mat	
0	1	2	3	4	5	6	7



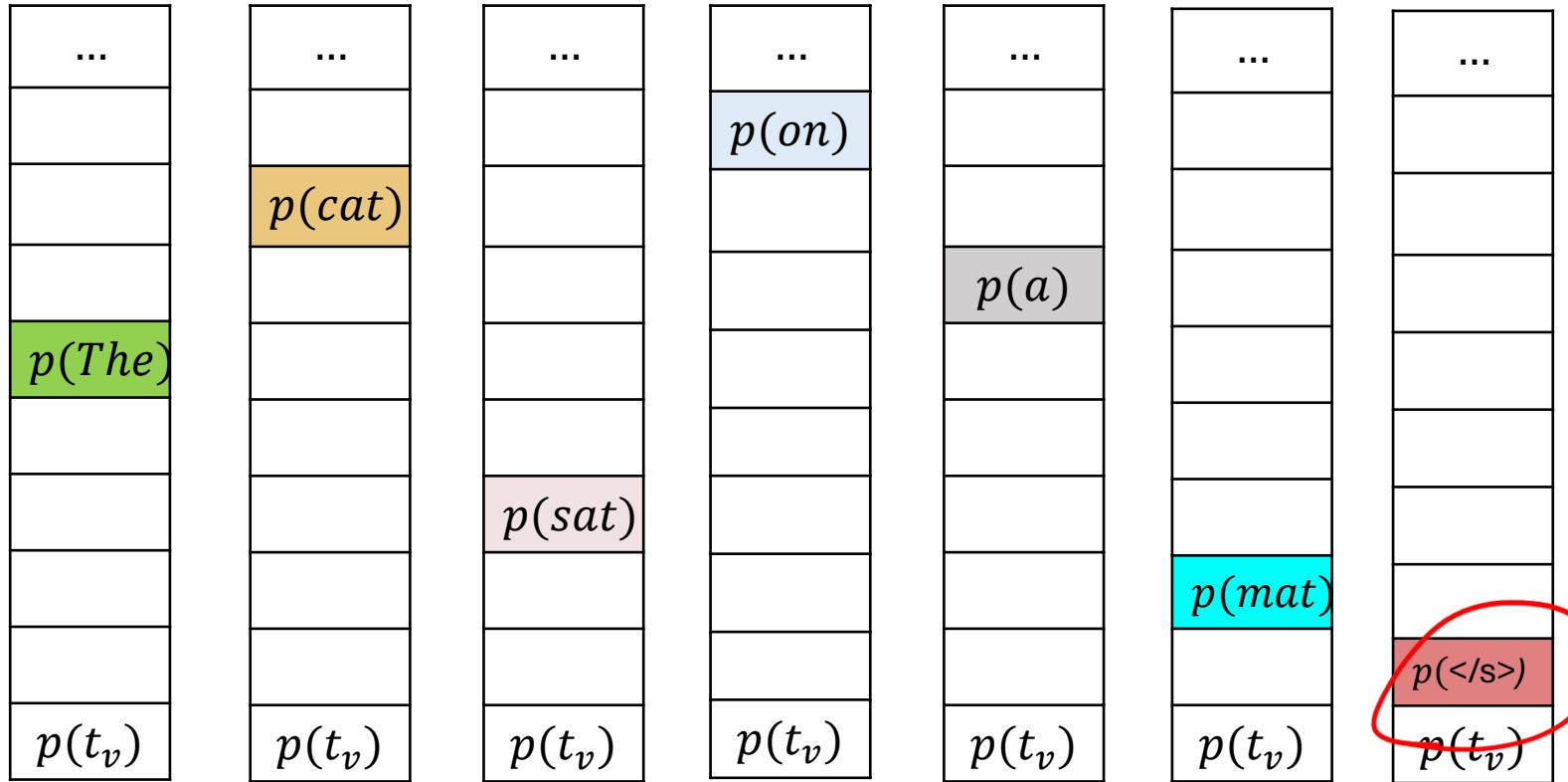
Inference through an LLM

Fwd. pass again (#4)

Transformer based LLM (θ)

<s>	The	cat	sat	on	a	mat	
0	1	2	3	4	5	6	7





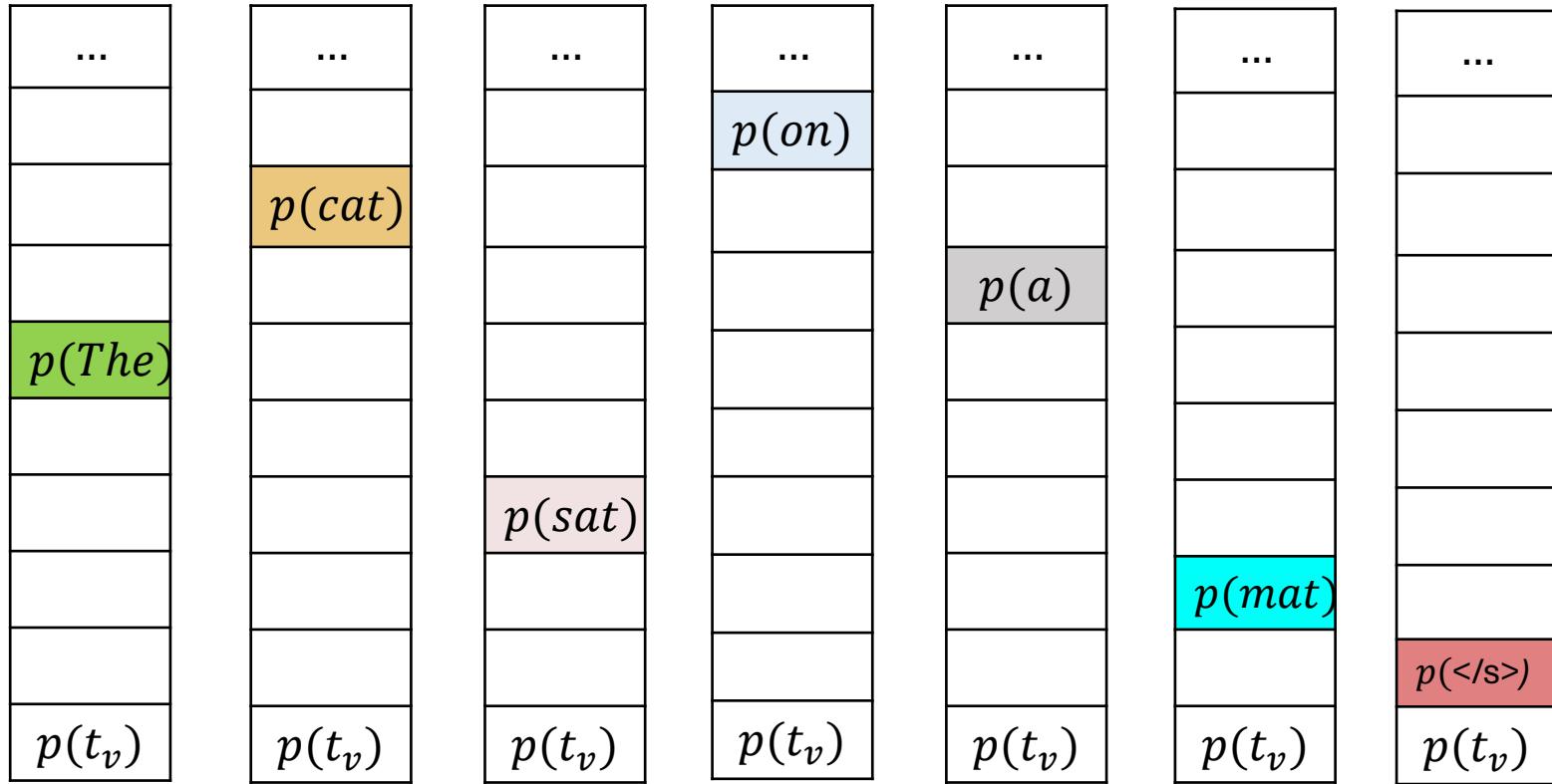
Inference through an LLM

Fwd. pass again (#4)

Transformer based LLM (θ)

<s>	The	cat	sat	on	a	mat	
0	1	2	3	4	5	6	7





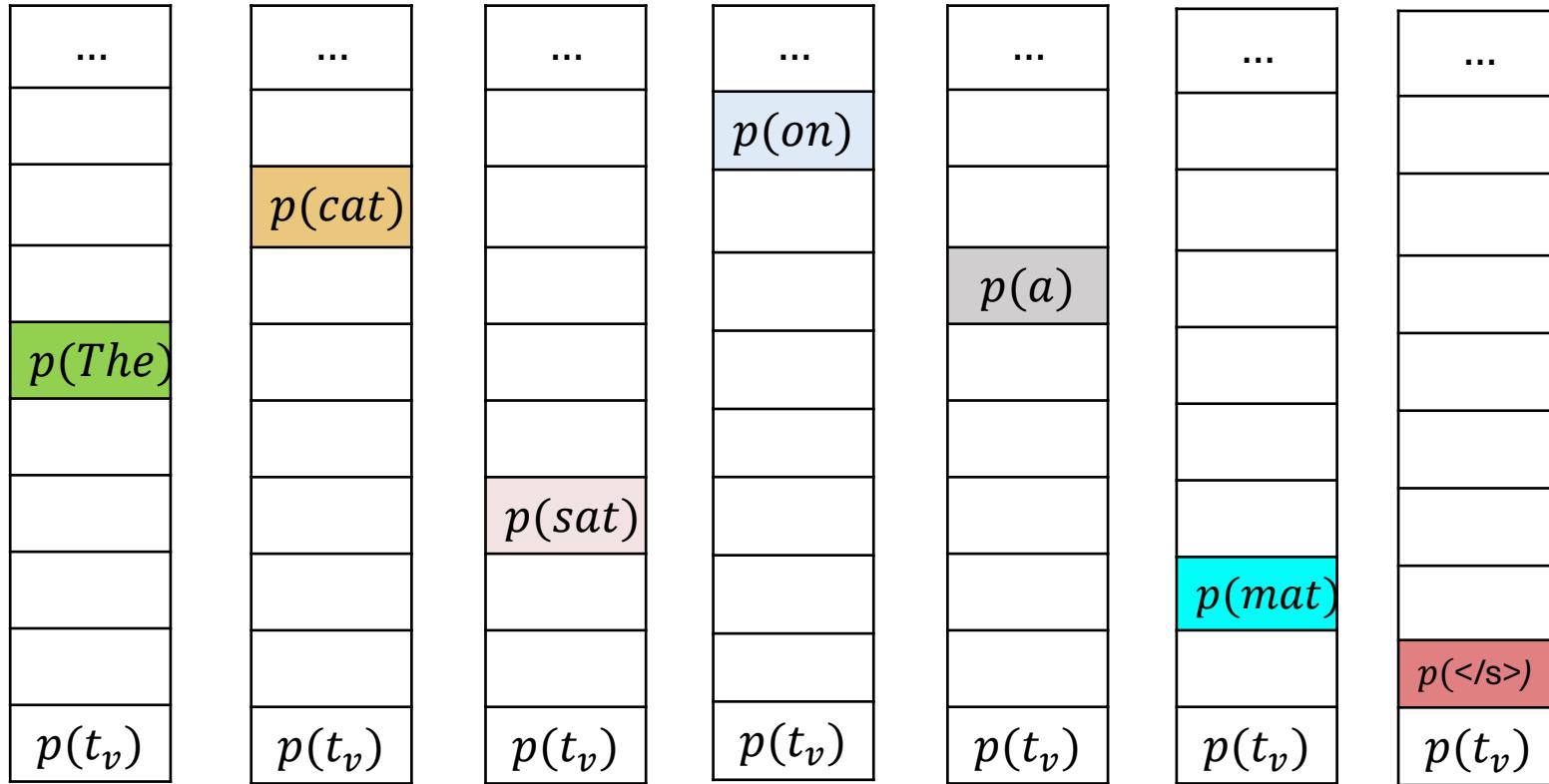
Inference through an LLM

Stop at end of seq. token:
 $</s>$

Transformer based LLM (θ)

<s>	The	cat	sat	on	a	mat	</s>
0	1	2	3	4	5	6	7





Inference through an LLM

Fwd Passes: 4
#Tokens: 4

Transformer based LLM (θ)

<s>	The	cat	sat	on	a	mat	</s>
0	1	2	3	4	5	6	7



Inference through an LLM

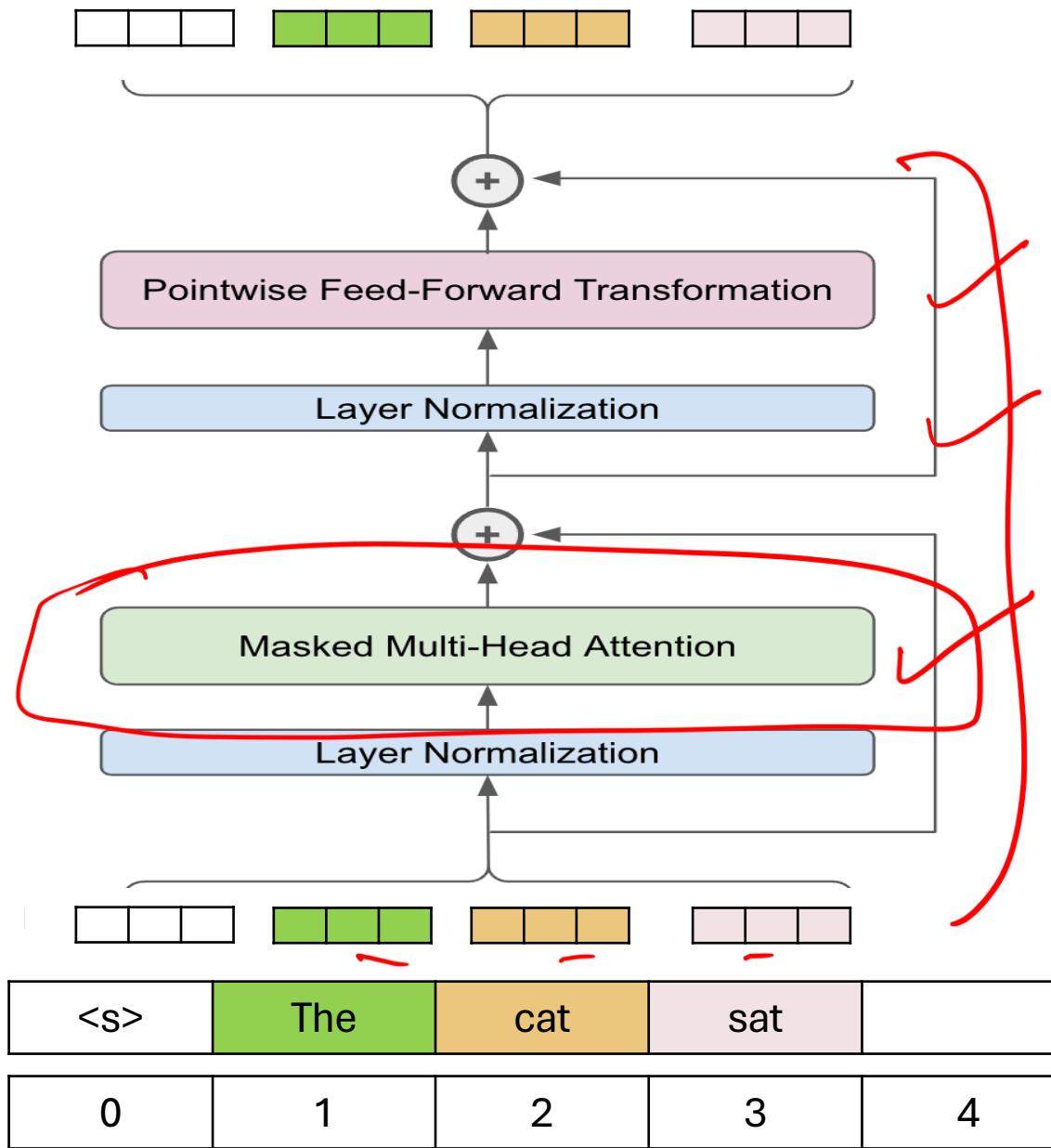
- ❑ 4 forward passes for 4 tokens
- ❑ Not feasible at production scale
- ❑ Let us revisit forward pass through and see if we can optimize
- ❑ We will focus on attention layer as that is the bottleneck

Fwd Passes: 4
#Tokens: 4

Transformer based LLM (θ)

<s>	The	cat	sat	on	a	mat	</s>
0	1	2	3	4	5	6	7

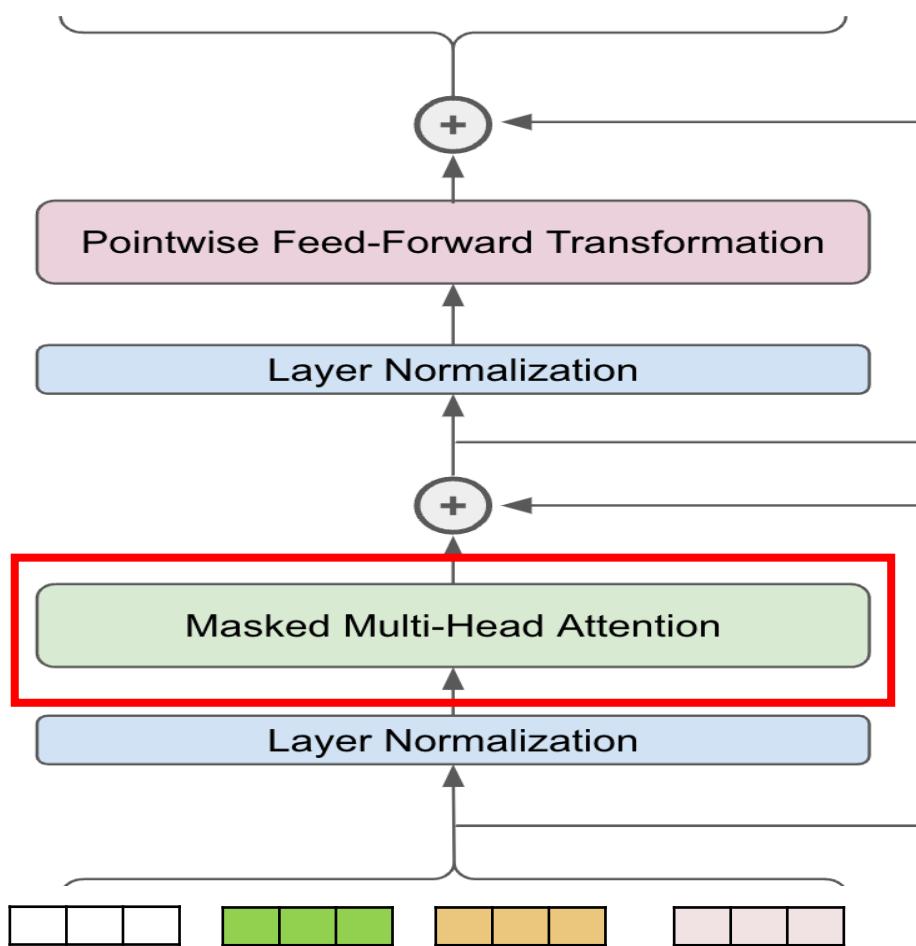




Inference through an LLM

Forward Pass #1





Inference through an LLM

Forward Pass #1

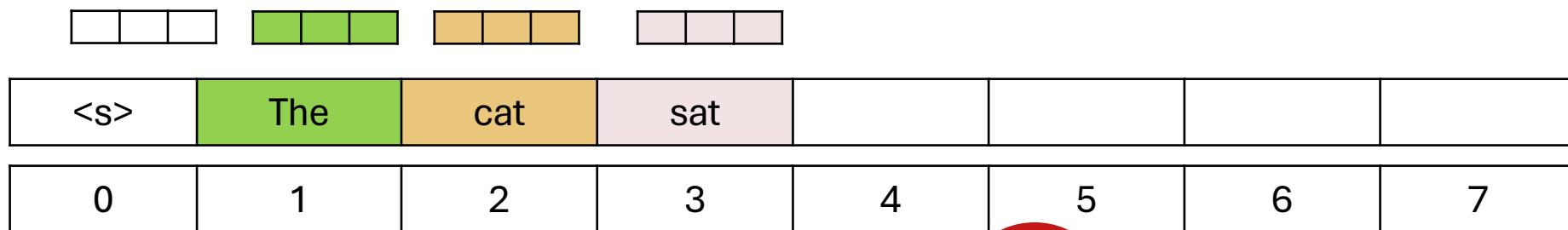
<s>	The	cat	sat				
0	1	2	3	4	5	6	7

Content credits: <https://cameronwolfe.substack.com/p/decoder-only-transformers-the-workhorse>



Inference through an LLM

Forward Pass #1



Content credits: <https://cameronwolfe.substack.com/p/decoder-only-transformers-the-workhorse>





Inference through an LLM

Forward Pass #1

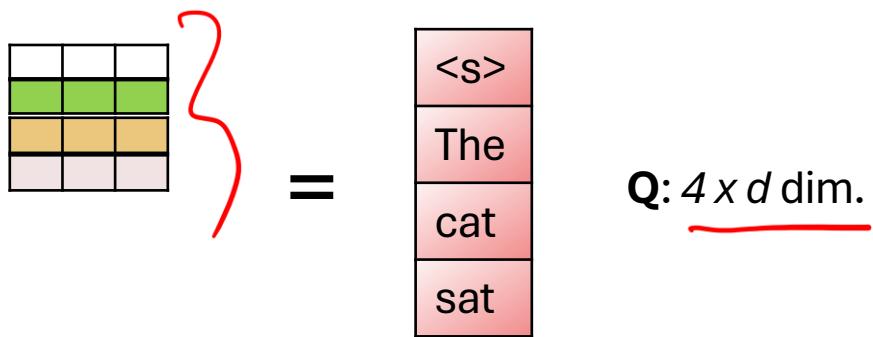


<s>	The	cat	sat				
0	1	2	3	4	5	6	7

Content credits: <https://cameronwolfe.substack.com/p/decoder-only-transformers-the-workhorse>



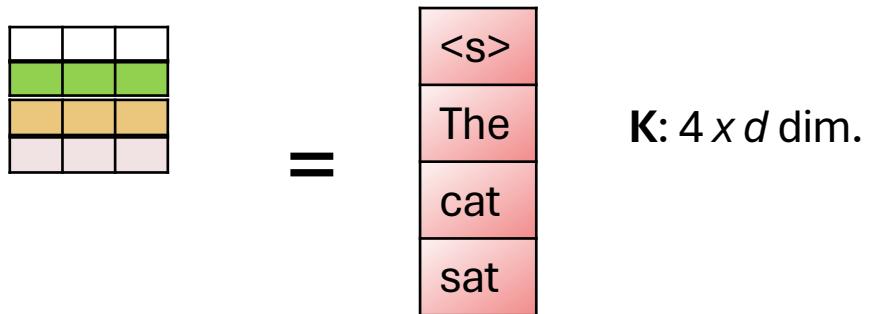
W_Q



$Q: 4 \times d \text{ dim.}$

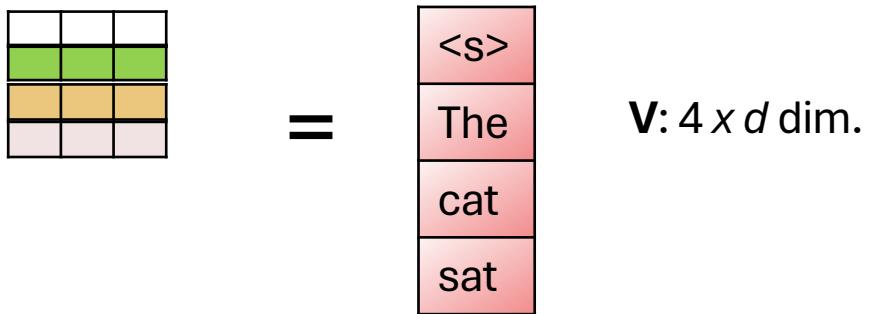
Inference through an LLM

W_K



$K: 4 \times d \text{ dim.}$

W_V



$V: 4 \times d \text{ dim.}$

Forward Pass #1

<s>	The	cat	sat				
0	1	2	3	4	5	6	7

Content credits: <https://cameronwolfe.substack.com/p/decoder-only-transformers-the-workhorse>



Inference through an LLM

$Q: 4 \times d$ dim.

<s>
The
cat
sat

$V: 4 \times d$ dim.

<s>
The
cat
sat

<s>	The	cat	sat
-----	-----	-----	-----

$K^T: d \times 4$ dim.

<s>	The	cat	sat				
0	1	2	3	4	5	6	7

Forward Pass #1



Inference through an LLM

Q: $4 \times d$ dim.

<S>
The
cat
sat

A: 4×4 dim.

$$A = \text{softmax} \left(\frac{QK^T}{\sqrt{d}} \right)$$

<S> The cat sat

V: $4 \times d$ dim.

<S>
The
cat
sat

Forward Pass #1

K^T: $d \times 4$ dim.

<S>	The	cat	sat				
0	1	2	3	4	5	6	7



Inference through an LLM

$Q: 4 \times d$ dim.

<s>
The
cat
sat

$A: 4 \times 4$ dim.

1			
0.2	0.8		
0.1	0.3	0.6	
0.01	0.19	0.3	0.5

$V: 4 \times d$ dim.

<s>
The
cat
sat

<s>	The	cat	sat
-----	-----	-----	-----

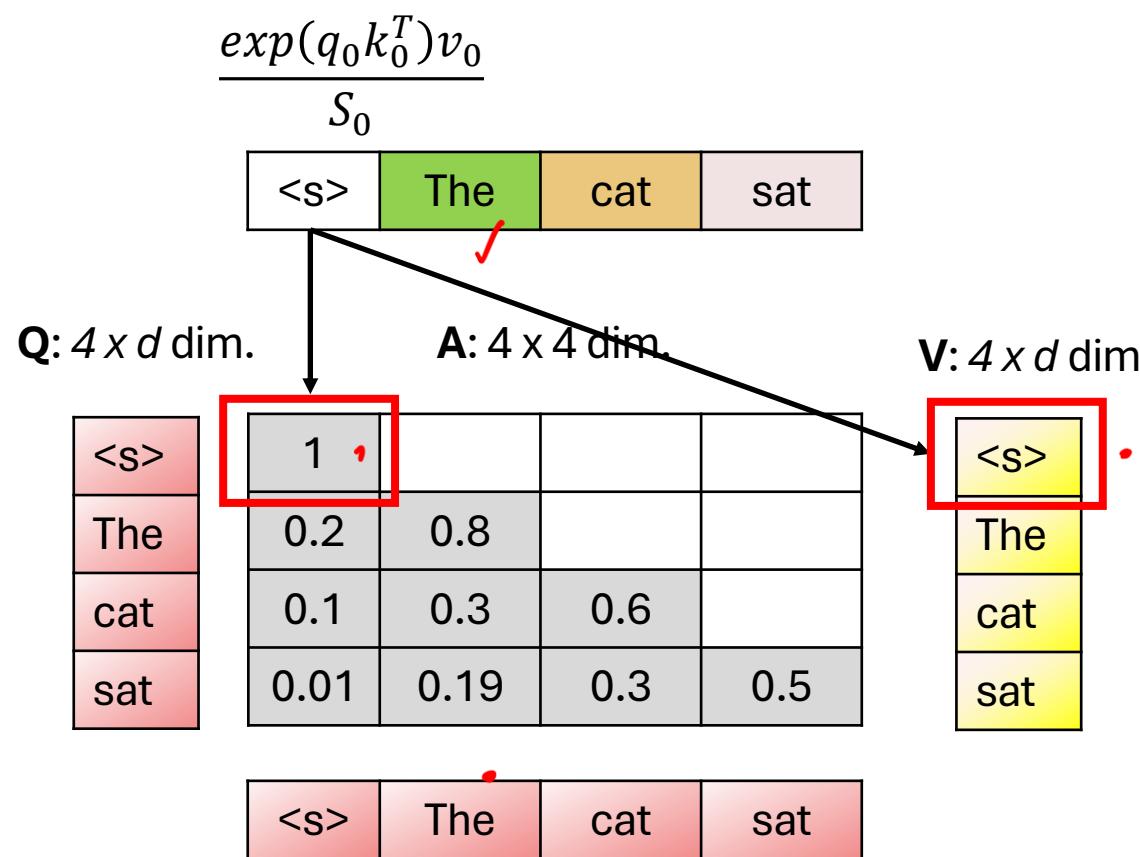
$K^T: d \times 4$ dim.

<s>	The	cat	sat				
0	1	2	3	4	5	6	7

Forward Pass #1



Inference through an LLM

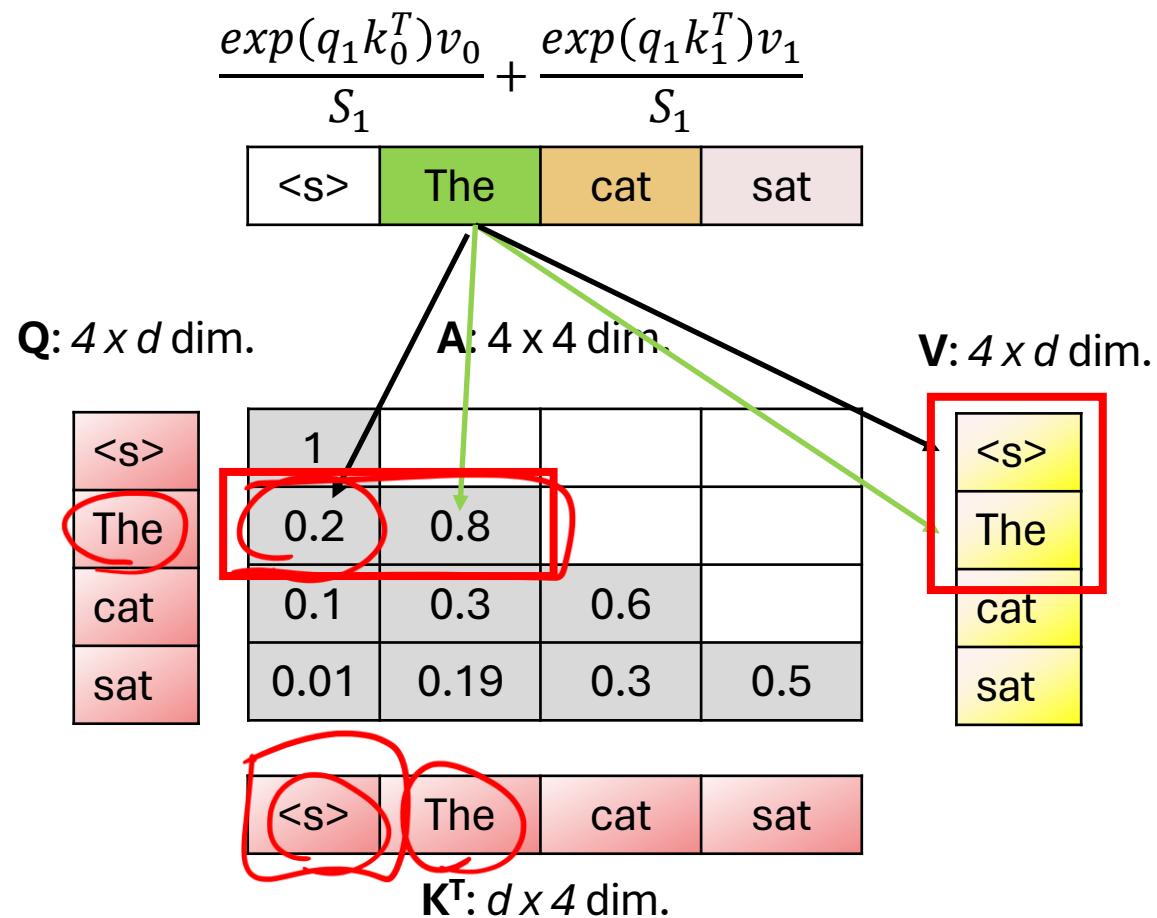


Forward Pass #1

<s>	The	cat	sat				
0	1	2	3	4	5	6	7



Inference through an LLM



Forward Pass #1

<s>	The	cat	sat				
0	1	2	3	4	5	6	7



Inference through an LLM

$$\frac{\exp(q_2 k_0^T) v_0}{S_2} + \frac{\exp(q_2 k_1^T) v_1}{S_2} + \frac{\exp(q_2 k_2^T) v_2}{S_2}$$

<s>	The	cat	sat
-----	-----	-----	-----

Q: $4 \times d$ dim.

<s>
The
cat
sat

1			
0.2	0.8		
0.1	0.3	0.6	
0.01	0.19	0.3	0.5

A: 4×4 dim.

V: $4 \times d$ dim.

<s>
The
cat
sat

Forward Pass #1

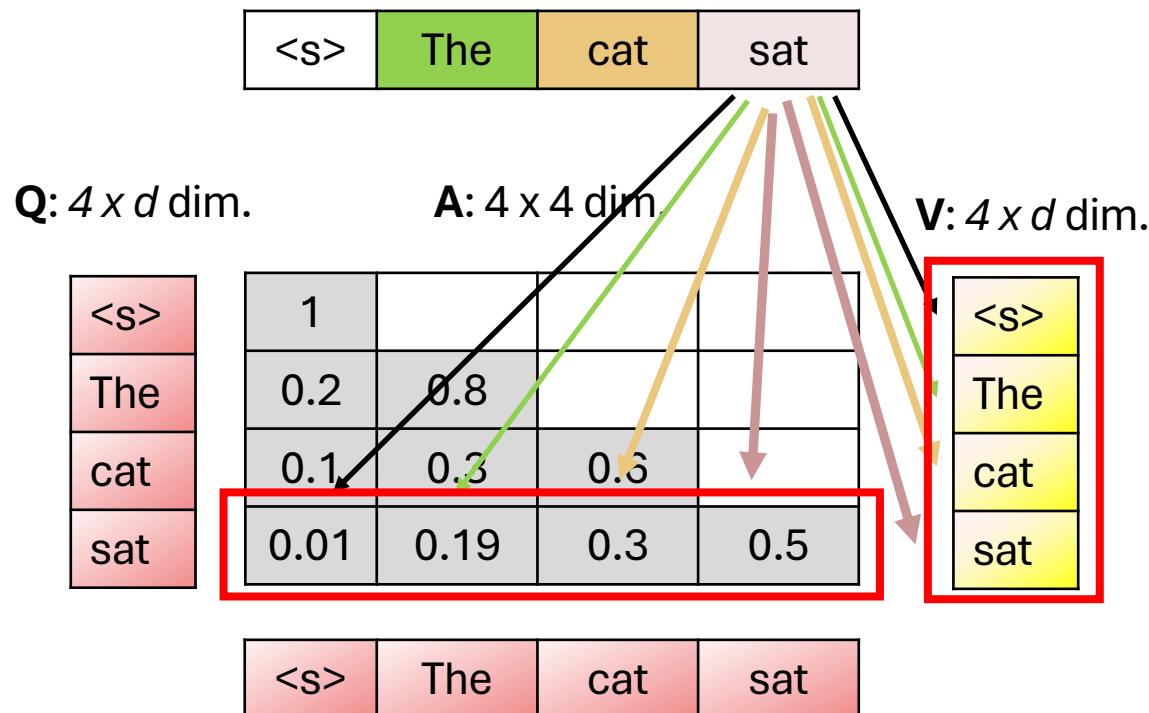
<s>	The	cat	sat
-----	-----	-----	-----

K^T: $d \times 4$ dim.

<s>	The	cat	sat				
0	1	2	3	4	5	6	7



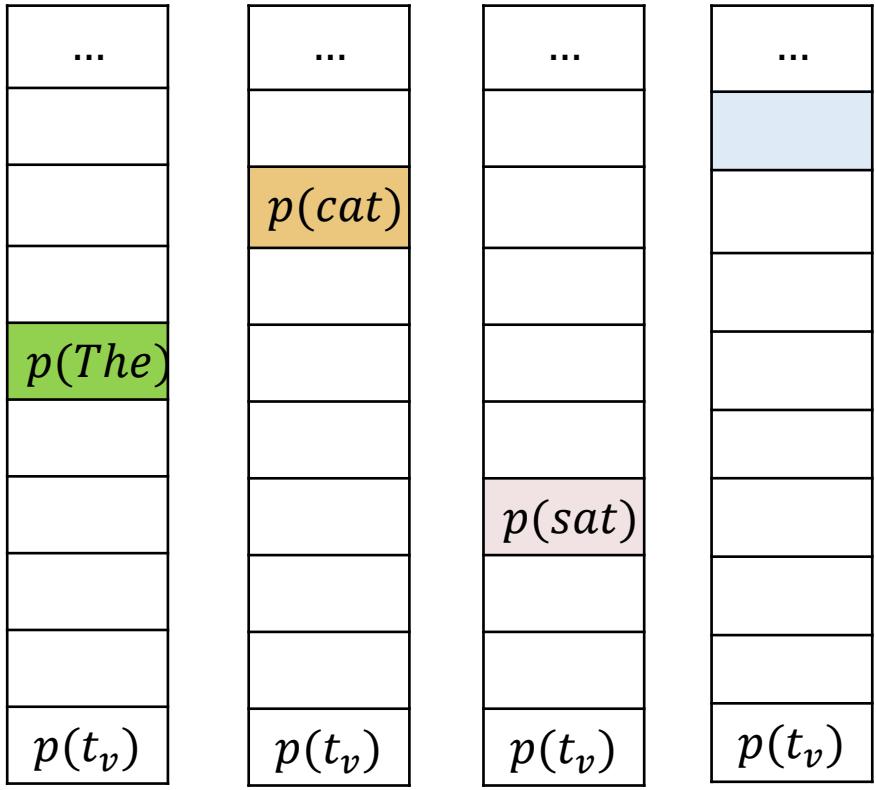
$$\frac{\exp(q_3 k_0^T) v_0}{S_3} + \frac{\exp(q_3 k_1^T) v_1}{S_3} + \frac{\exp(q_3 k_2^T) v_2}{S_3} + \frac{\exp(q_3 k_3^T) v_3}{S_3}$$



Inference through an LLM

Forward Pass #1





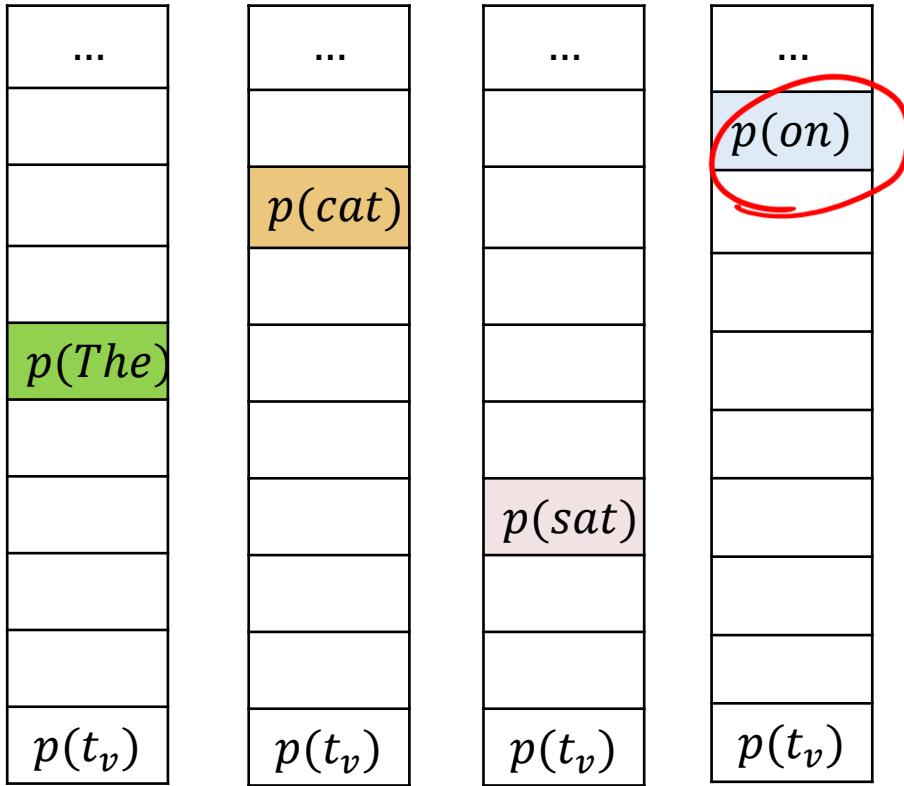
Inference through an LLM

- Emb. of `sat` at the last layer
- Pass through classifier to get distribution over tokens
- Pick the token having max. probability at step 3

Transformer based LLM (θ)

<s>	The	cat	sat				
0	1	2	3	4	5	6	7





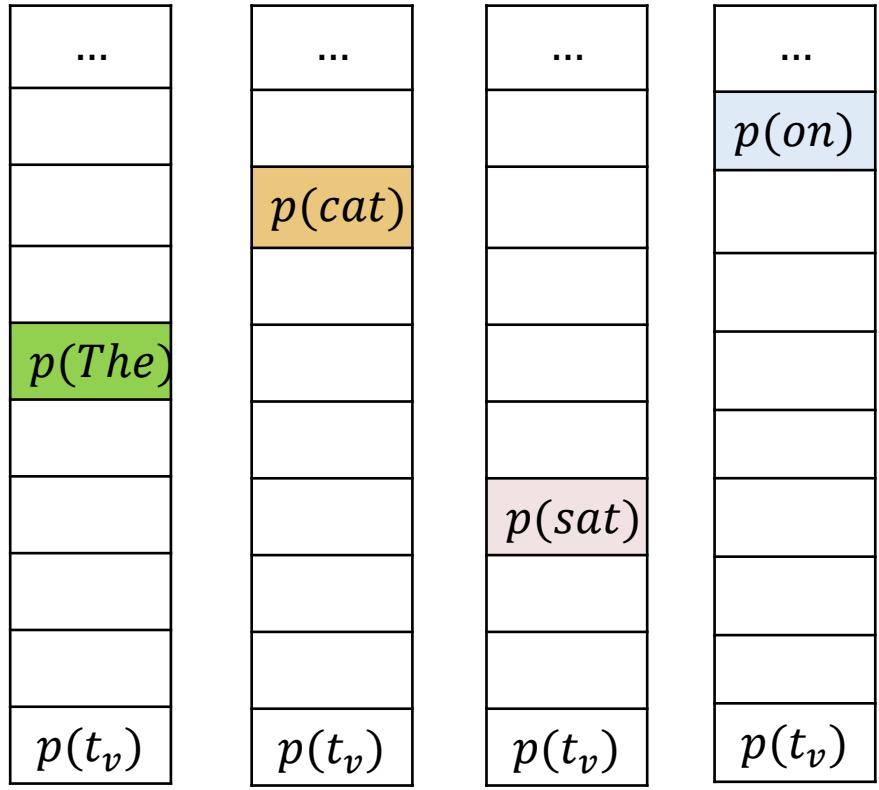
Inference through an LLM

- Emb. of sat at the last layer
- Pass through classifier to get distribution over tokens
- Pick the token having max. probability at step 3

Transformer based LLM (θ)

<s>	The	cat	sat				
0	1	2	3	4	5	6	7





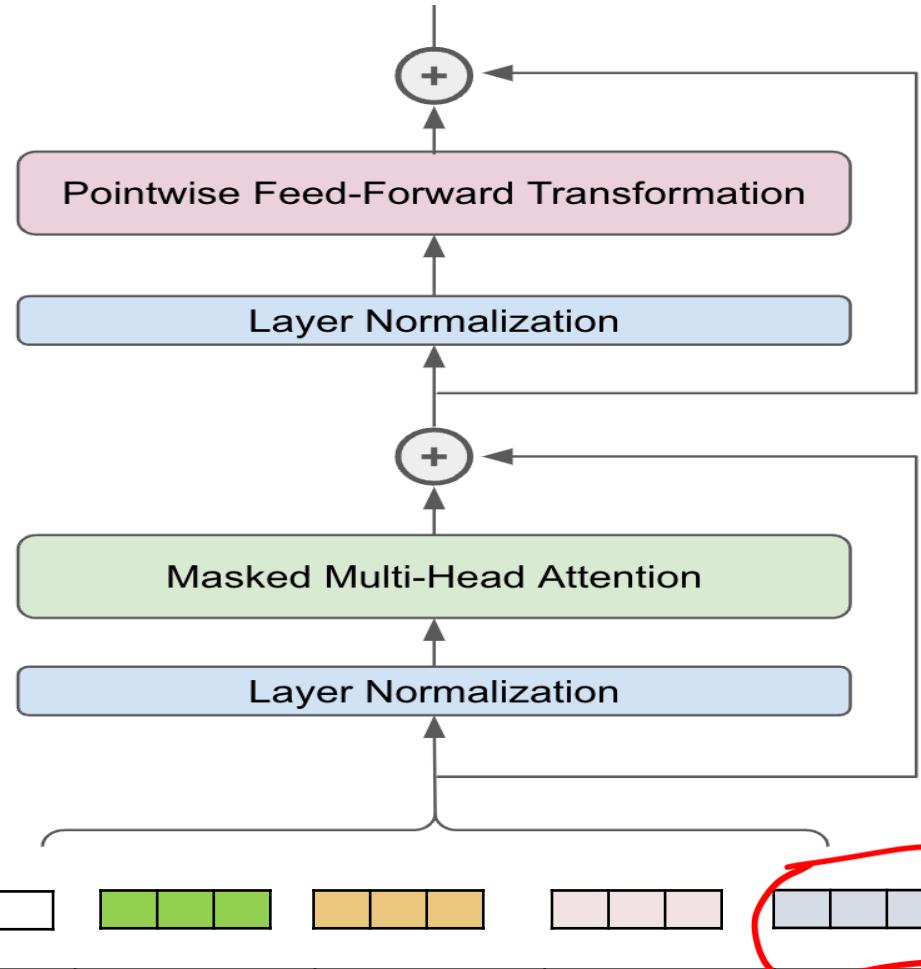
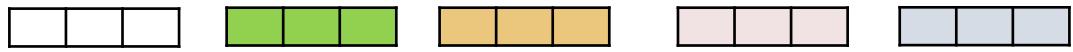
Inference through an LLM

Fill at step 4

Transformer based LLM (θ)

<s>	The	cat	sat	on			
0	1	2	3	4	5	6	7





Inference through an LLM

Forward Pass #2

<s>	The	cat	sat	on			
0	1	2	3	4	5	6	7

Content credits: <https://cameronrwolfe.substack.com/p/decoder-only-transformers-the-workhorse>



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Inference through an LLM

Q: $5 \times d$ dim.

<s>
The
cat
sat
on

A: 5×5 dim.

$$A = \text{softmax} \left(\frac{QK^T}{\sqrt{d}} \right)$$

<s> The cat sat on

V: $5 \times d$ dim.

<s>
The
cat
sat
on

K^T: $d \times 5$ dim.

<s>	The	cat	sat	on			
0	1	2	3	4	5	6	7

- A lot of computation already done in Fwd. pass #1

Forward Pass #2



Inference through an LLM

$Q: 5 \times d$ dim.

<s>
The
cat
sat
on

$A: 5 \times 5$ dim.

1				
0.2	0.8			
0.1	0.3	0.6		
0.01	0.19	0.3	0.5	

$V: 5 \times d$ dim.

<s>
The
cat
sat
on

<s>	The	cat	sat	on
-----	-----	-----	-----	----

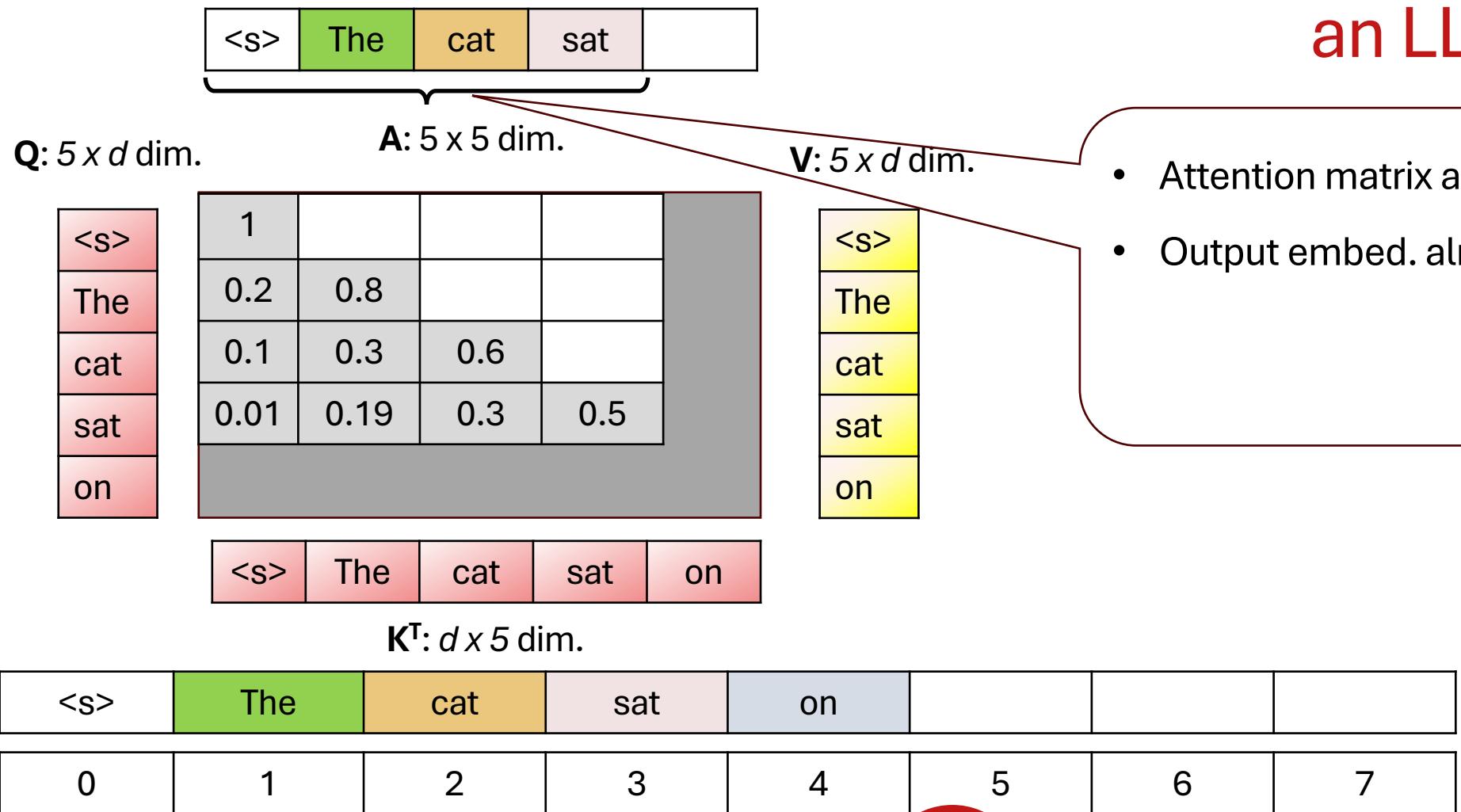
$K^T: d \times 5$ dim.

<s>	The	cat	sat	on			
0	1	2	3	4	5	6	7

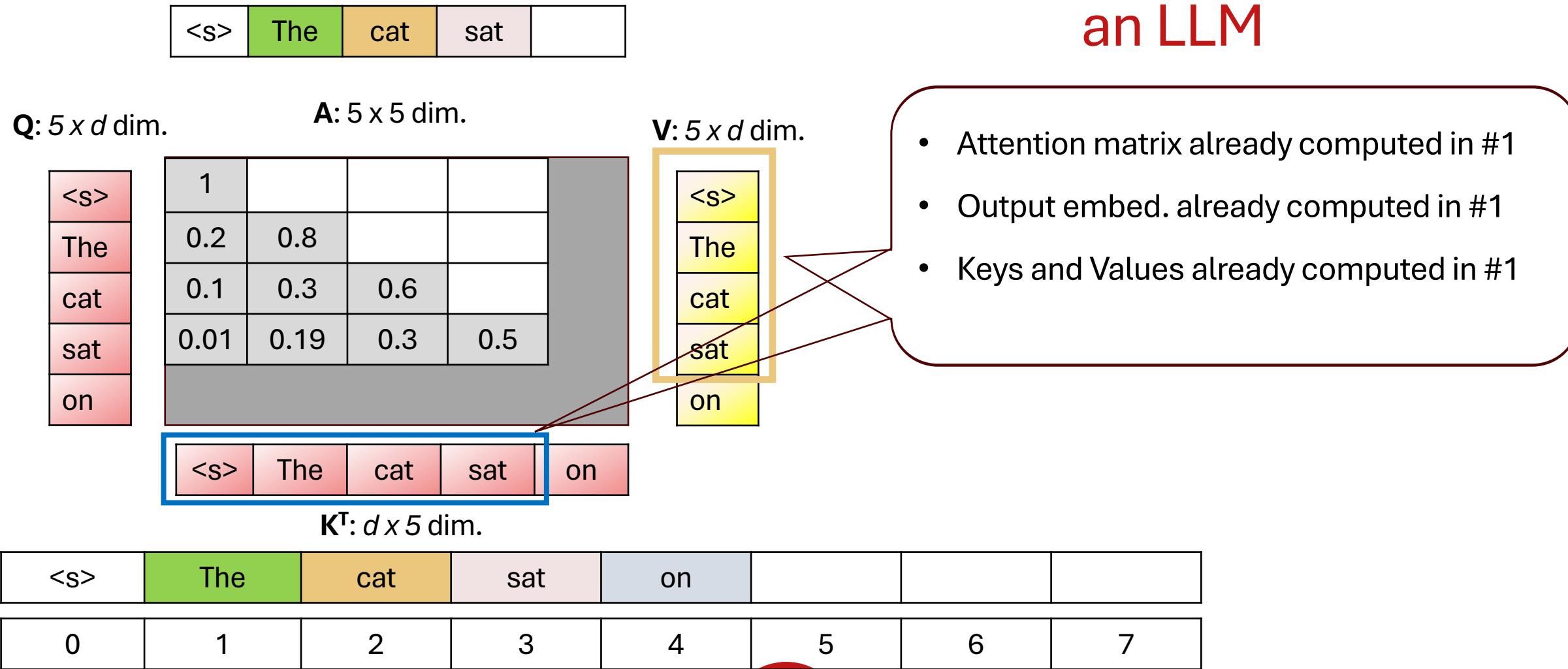
- Attention matrix already computed in #1



Inference through an LLM



Inference through an LLM



Inference through an LLM

<s>	The	cat	sat	
-----	-----	-----	-----	--

Q: $5 \times d$ dim.

<s>
The
cat
sat
on

A: 5×5 dim.

1				
0.2	0.8			
0.1	0.3	0.6		
0.01	0.19	0.3	0.5	

V: $5 \times d$ dim.

<s>
The
cat
sat
on

- Attention matrix already computed in #1
- Output embed. already computed in #1
- Keys and Values already computed in #1
- Queries not required in #2

$K^T: d \times 5$ dim.

<s>	The	cat	sat	on			
0	1	2	3	4	5	6	7



Inference through an LLM

<s>	The	cat	sat	
-----	-----	-----	-----	--

Q: $5 \times d$ dim.

<s>				
The				
cat				
sat				
on				

A: 5×5 dim.

1				
0.2	0.8			
0.1	0.3	0.6		
0.01	0.19	0.3	0.5	

V: $5 \times d$ dim.

<s>				
The				
cat				
sat				
on				

- Cache the already computed matrices

<s>	The	cat	sat	on
-----	-----	-----	-----	----

$K^T: d \times 5$ dim.

<s>	The	cat	sat	on			
0	1	2	3	4	5	6	7



Inference through an LLM

<s>	The	cat	sat	
-----	-----	-----	-----	--

Q: $5 \times d$ dim.

<s>
The
cat
sat
on

A: 5×5 dim.

1				
0.2	0.8			
0.1	0.3	0.6		
0.01	0.19	0.3	0.5	

V: $5 \times d$ dim.

<s>
The
cat
sat
on

K cache

<s>
The
cat
sat
on

$K^T: d \times 5$ dim.

<s>	The	cat	sat	on			
0	1	2	3	4	5	6	7



Inference through an LLM

<s>	The	cat	sat	
-----	-----	-----	-----	--

Q: $5 \times d$ dim.

<s>
The
cat
sat
on

A: 5×5 dim.

1					
0.2	0.8				
0.1	0.3	0.6			
0.01	0.19	0.3	0.5		

V: $5 \times d$ dim.

<s>
The
cat
sat
on

K cache

<s>
The
cat
sat

$K^T: d \times 5$ dim.

<s>	The	cat	sat	on			
0	1	2	3	4	5	6	7



Inference through an LLM

Q: $5 \times d$ dim.

<s>
The
cat
sat
on

A: 5×5 dim.

1					
0.2	0.8				
0.1	0.3	0.6			
0.01	0.19	0.3	0.5		

V: $5 \times d$ dim.

<s>
The
cat
sat
on

K cache

<s>
The
cat
sat
on

V cache

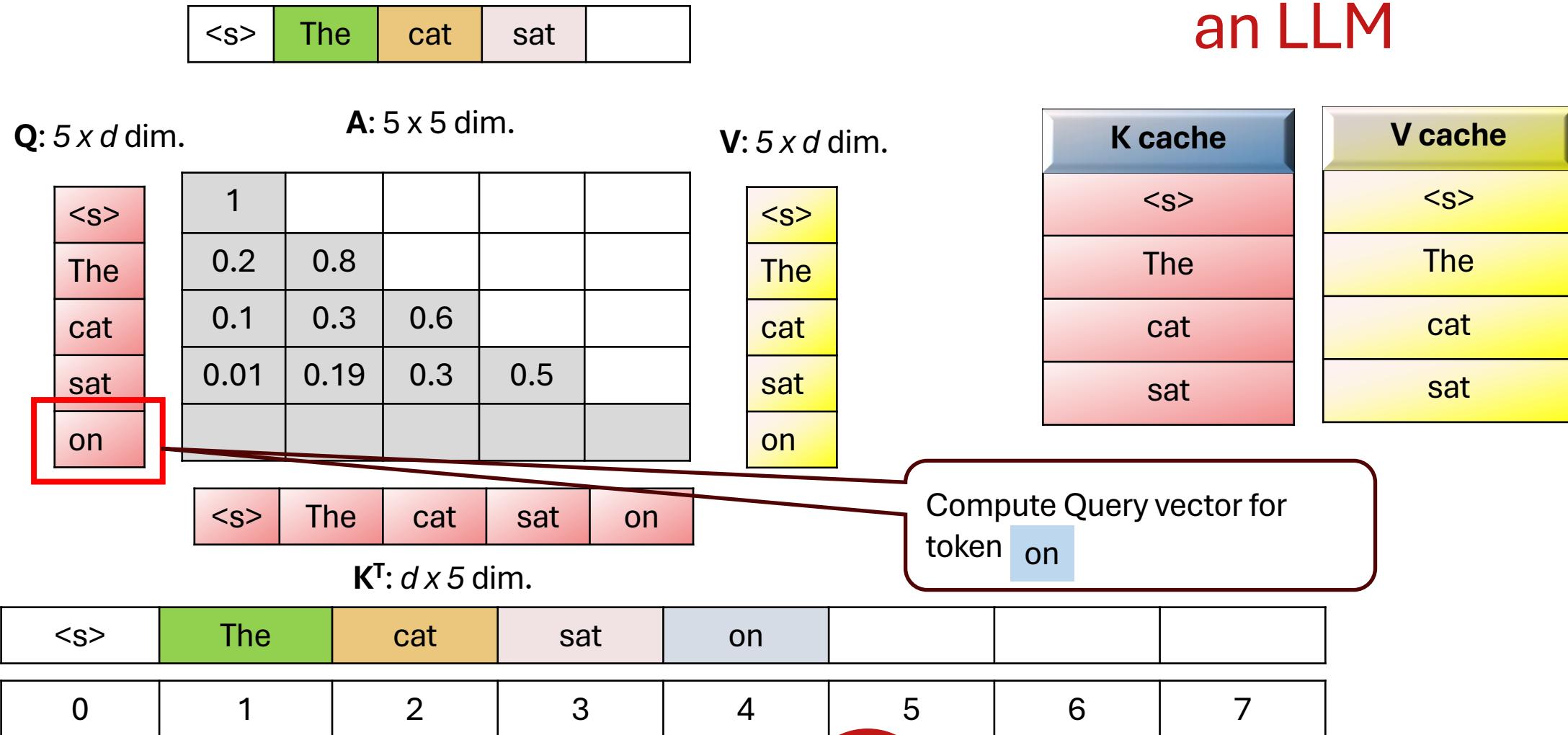
<s>
The
cat
sat
on

$K^T: d \times 5$ dim.

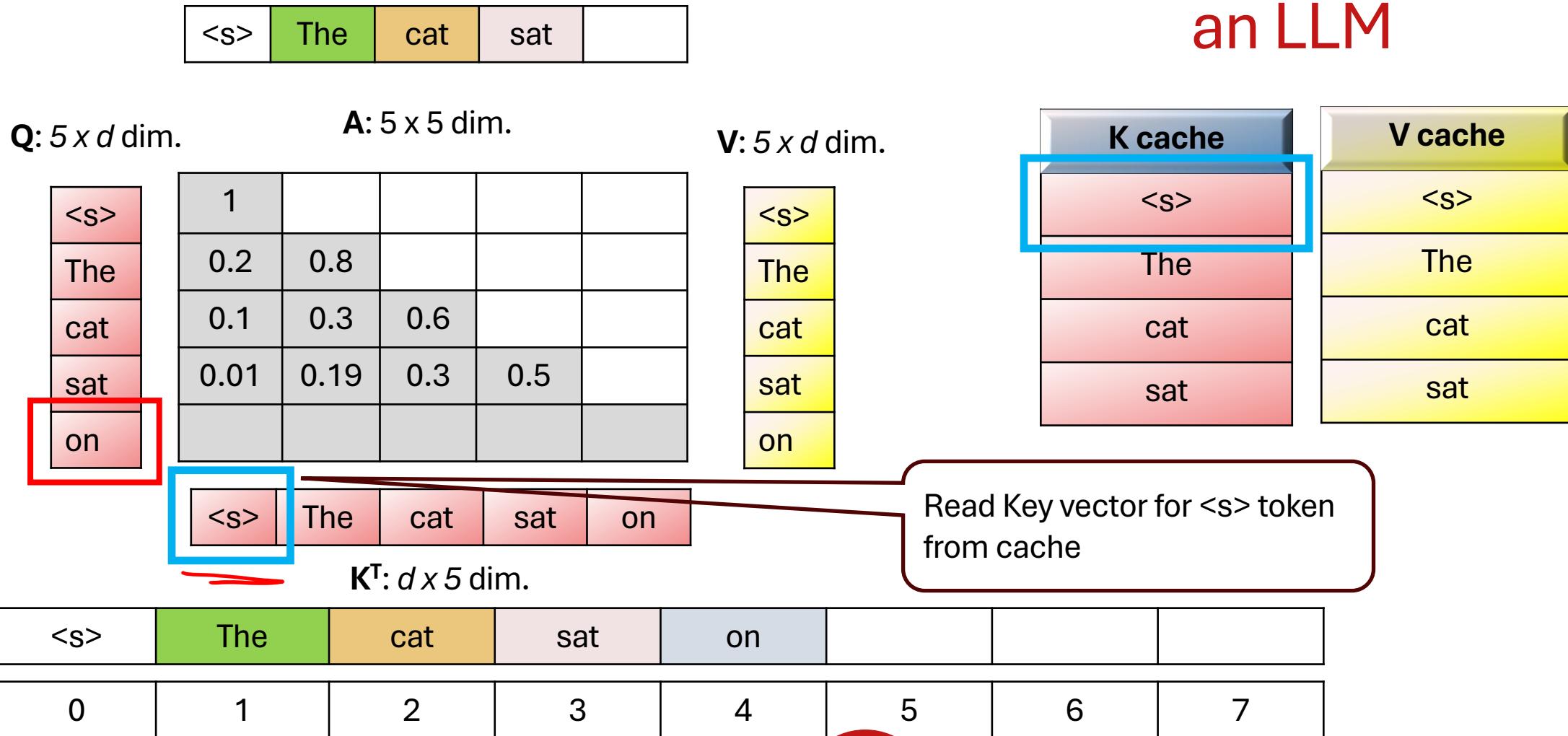
<s>	The	cat	sat	on			
0	1	2	3	4	5	6	7



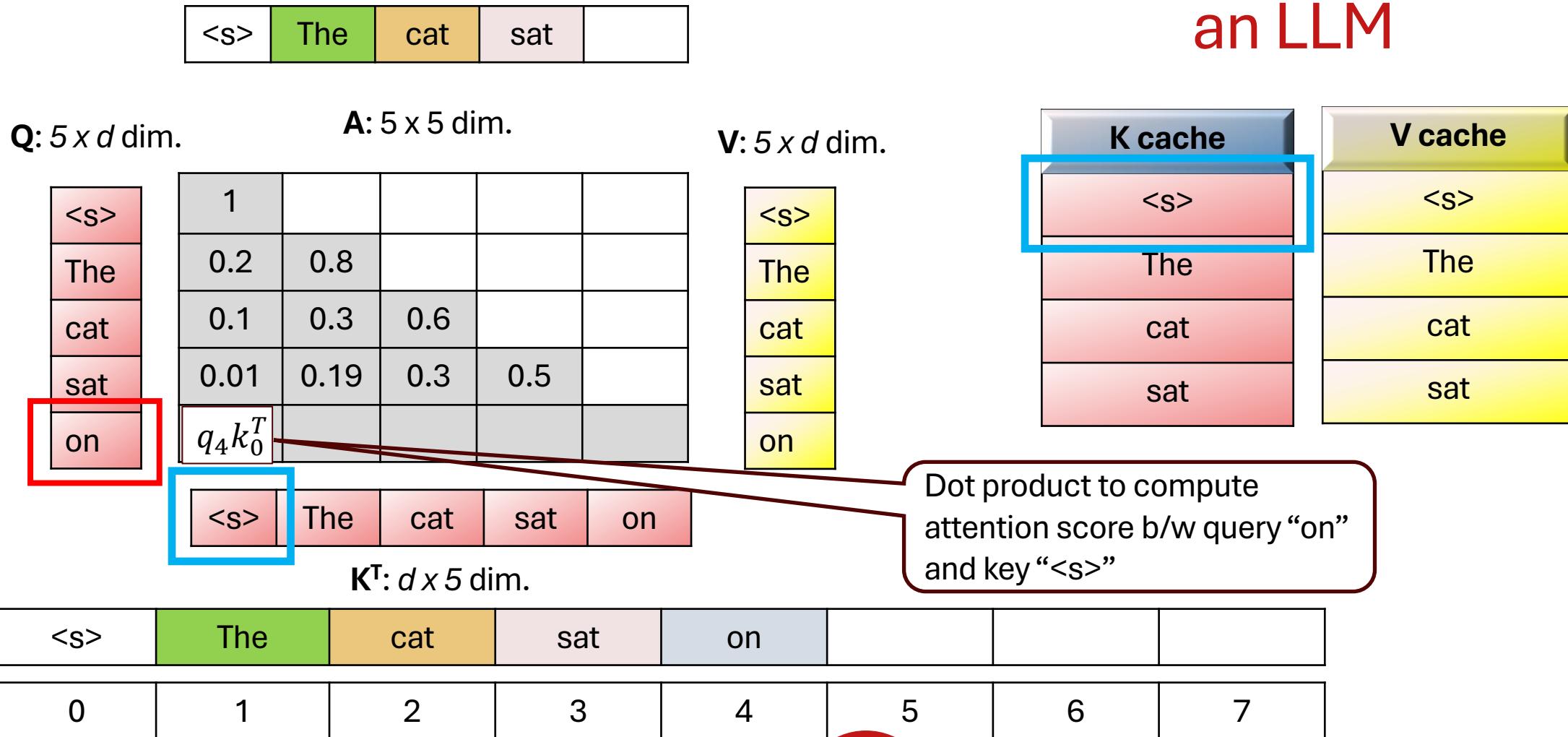
Inference through an LLM



Inference through an LLM



Inference through an LLM



Inference through an LLM

<s>	The	cat	sat	
-----	-----	-----	-----	--

Q: $5 \times d$ dim.

<s>
The
cat
sat
on

1				
0.2	0.8			
0.1	0.3	0.6		
0.01	0.19	0.3	0.5	
$q_4 k_0^T$	$q_4 k_1^T$			

<s>	The	cat	sat	on
-----	-----	-----	-----	----

$K^T: d \times 5$ dim.

<s>	The	cat	sat	on			
0	1	2	3	4	5	6	7

V: $5 \times d$ dim.

<s>
The
cat
sat
on

K cache
<s>
The
cat
sat
on

V cache
<s>
The
cat
sat
on



Inference through an LLM

<s>	The	cat	sat	
-----	-----	-----	-----	--

Q: $5 \times d$ dim.

<s>
The
cat
sat
on

1				
0.2	0.8			
0.1	0.3	0.6		
0.01	0.19	0.3	0.5	
$q_4 k_0^T$	$q_4 k_1^T$	$q_4 k_2^T$		

<s>	The	cat	sat	on
-----	-----	-----	-----	----

$K^T: d \times 5$ dim.

<s>	The	cat	sat	on			
0	1	2	3	4	5	6	7

V: $5 \times d$ dim.

<s>
The
cat
sat
on

K cache
<s>
The
cat
sat
on

V cache
<s>
The
cat
sat
on



Inference through an LLM

<s>	The	cat	sat	
-----	-----	-----	-----	--

Q: $5 \times d$ dim.

<s>
The
cat
sat
on

1				
0.2	0.8			
0.1	0.3	0.6		
0.01	0.19	0.3	0.5	
$q_4 k_0^T$	$q_4 k_1^T$	$q_4 k_2^T$	$q_4 k_3^T$	

<s>	The	cat	sat	on
-----	-----	-----	-----	----

$K^T: d \times 5$ dim.

<s>	The	cat	sat	on			
0	1	2	3	4	5	6	7

V: $5 \times d$ dim.

<s>
The
cat
sat
on

K cache

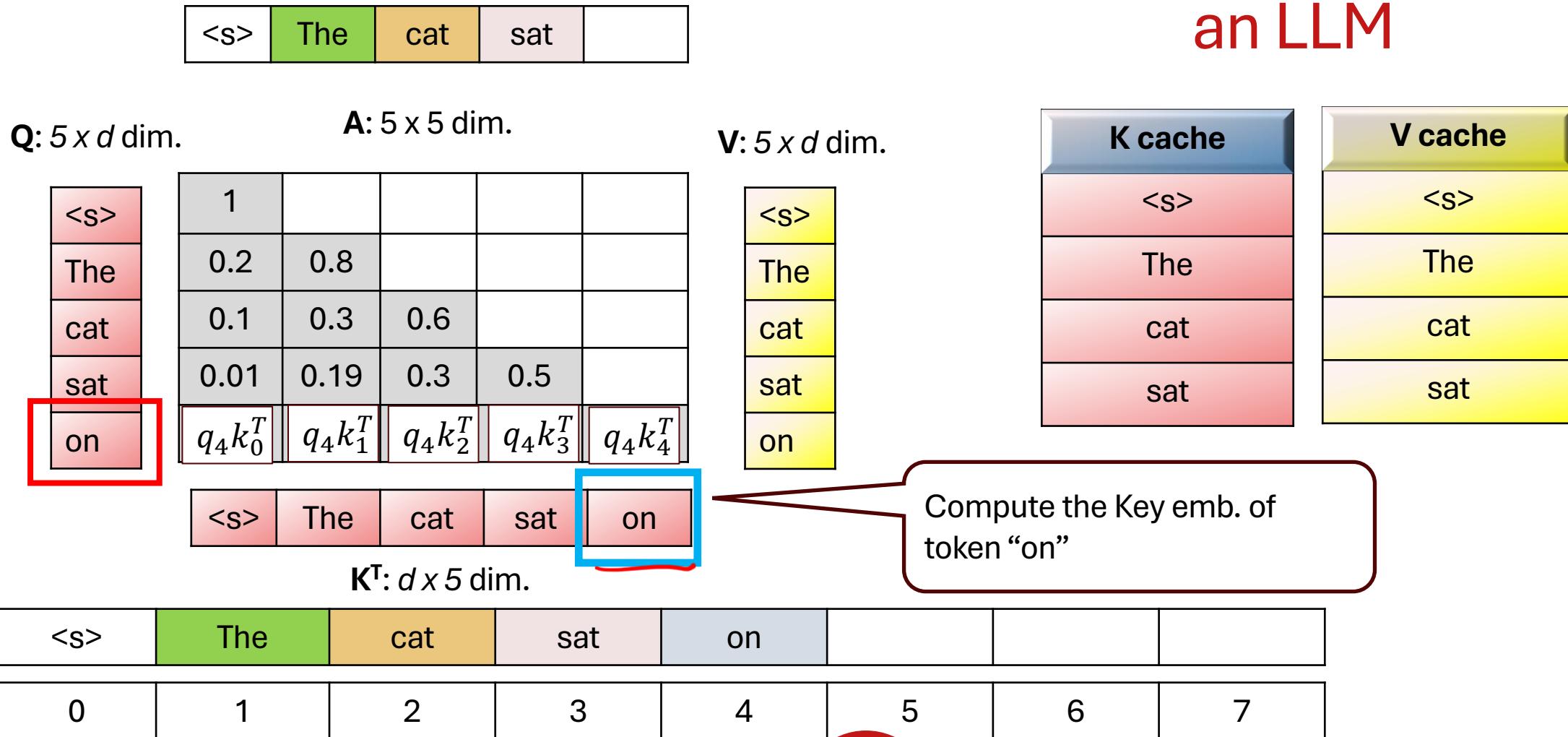
<s>
The
cat
sat
on

V cache

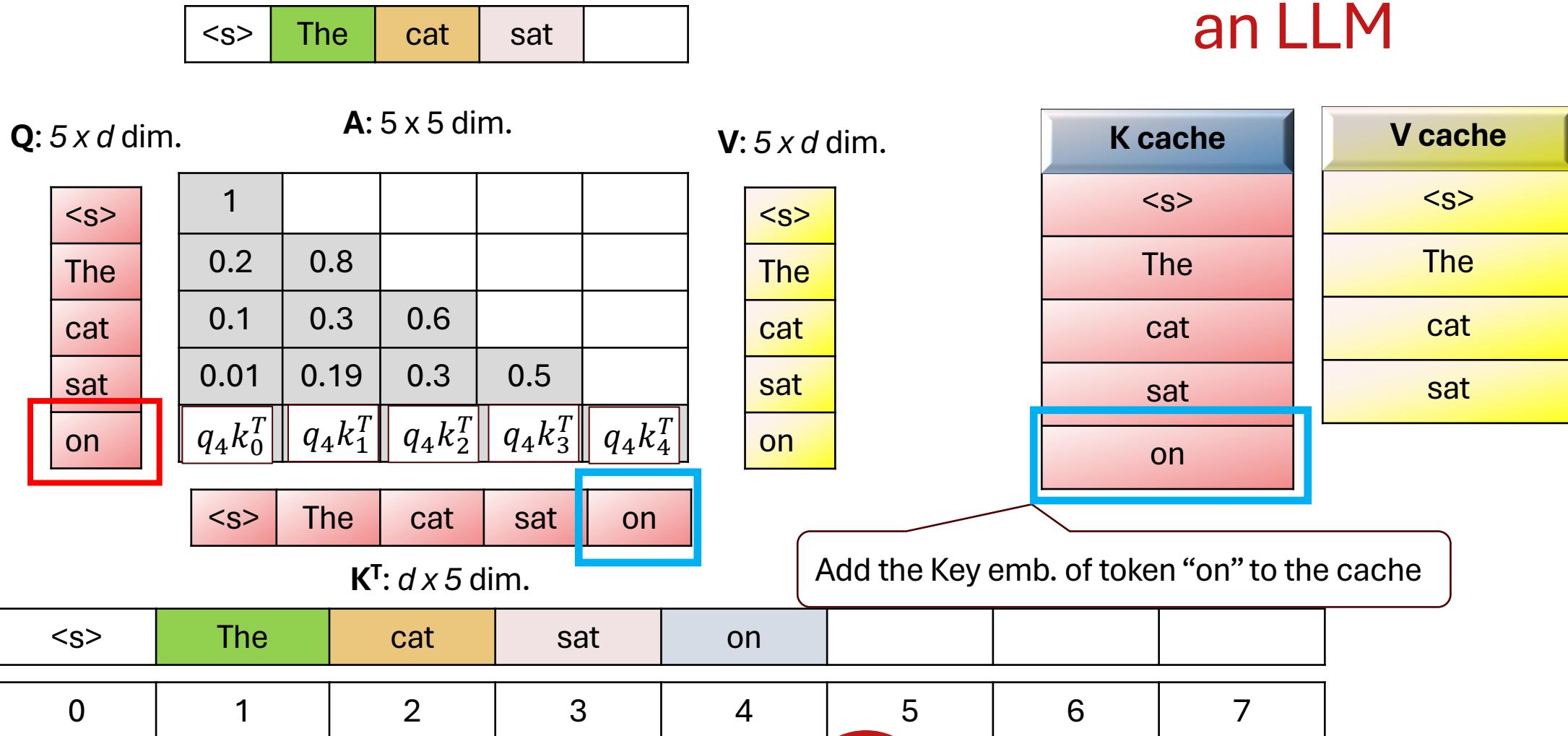
<s>
The
cat
sat
on



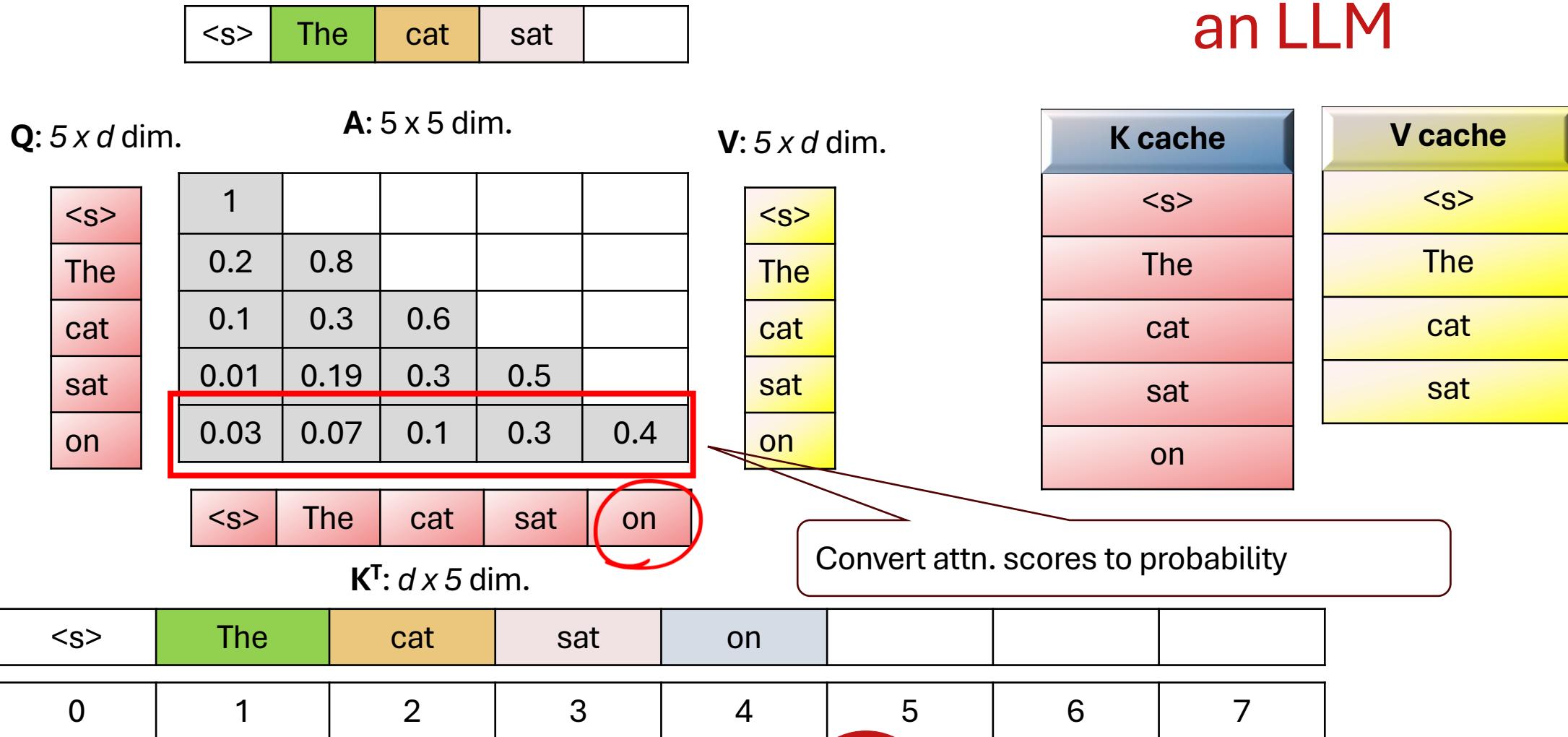
Inference through an LLM



Inference through an LLM



Inference through an LLM



Inference through an LLM

<s>	The	cat	sat	
-----	-----	-----	-----	--

Q: $5 \times d$ dim.

<s>
The
cat
sat
on

1				
0.2	0.8			
0.1	0.3	0.6		
0.01	0.19	0.3	0.5	
0.03	0.07	0.1	0.3	0.4

<s>	The	cat	sat	on
-----	-----	-----	-----	----

$A: 5 \times 5$ dim.

V: $5 \times d$ dim.

<s>
The
cat
sat
on

K cache

<s>
The
cat
sat
on

V cache

<s>
The
cat
sat
on

Load Value vectors from V cache

<s>	The	cat	sat	on				
0	1	2	3	4	5	6	7	



$$\frac{(q_4 k_0^T) v_0}{S_4}$$

<s>	The	cat	sat	
-----	-----	-----	-----	--

Q: $5 \times d$ dim.

<s>	1				
The	0.2	0.8			
cat	0.1	0.3	0.6		
sat	0.01	0.19	0.3	0.5	
on	0.03	0.07	0.1	0.3	0.4

<s>	The	cat	sat	on
-----	-----	-----	-----	----

A: 5×5 dim.

V: $5 \times d$ dim.

<s>
The
cat
sat
on

Inference through an LLM

K cache
<s>
The
cat
sat
on

V cache
<s>
The
cat
sat
on

K^T : $d \times 5$ dim.

<s>	The	cat	sat	on				
0	1	2	3	4	5	6	7	



$$\frac{(q_4 k_0^T) v_0}{S_4} + \frac{(q_4 k_1^T) v_1}{S_4}$$

<s>	The	cat	sat	
-----	-----	-----	-----	--

Q: $5 \times d$ dim.

<s>
The
cat
sat
on

1				
0.2	0.8			
0.1	0.3	0.6		
0.01	0.19	0.3	0.5	
0.03	0.07	0.1	0.3	0.4

<s>	The	cat	sat	on
-----	-----	-----	-----	----

A: 5×5 dim.

V: $5 \times d$ dim.

<s>
The
cat
sat
on

K cache
<s>
The
cat
sat
on

V cache
<s>
The
cat
sat

K^T : $d \times 5$ dim.

<s>	The	cat	sat	on				
0	1	2	3	4	5	6	7	

Inference through an LLM



$$\frac{(q_4 k_0^T) v_0}{S_4} + \frac{(q_4 k_1^T) v_1}{S_4} + \frac{(q_4 k_2^T) v_2}{S_4}$$

<s>	The	cat	sat	
-----	-----	-----	-----	--

Q: $5 \times d$ dim.

<s>
The
cat
sat
on

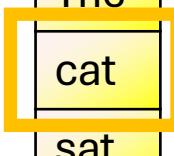
1				
0.2	0.8			
0.1	0.3	0.6		
0.01	0.19	0.3	0.5	
0.03	0.07	0.1	0.3	0.4

<s>	The	cat	sat	on
-----	-----	-----	-----	----

A: 5×5 dim.

V: $5 \times d$ dim.

<s>
The
cat
sat
on



Inference through an LLM

K cache
<s>
The
cat
sat
on

V cache
<s>
The
cat
sat
on

K^T : $d \times 5$ dim.

<s>	The	cat	sat	on				
0	1	2	3	4	5	6	7	



$$\frac{(q_4 k_0^T) v_0}{S_4} + \frac{(q_4 k_1^T) v_1}{S_4} + \frac{(q_4 k_2^T) v_2}{S_4} + \frac{(q_4 k_3^T) v_3}{S_4}$$

<s>	The	cat	sat	
-----	-----	-----	-----	--

Q: $5 \times d$ dim.

<s>
The
cat
sat
on

1				
0.2	0.8			
0.1	0.3	0.6		
0.01	0.19	0.3	0.5	
0.03	0.07	0.1	0.3	0.4

<s>	The	cat	sat	on
-----	-----	-----	-----	----

$K^T: d \times 5$ dim.

<s>	The	cat	sat	on				
0	1	2	3	4	5	6	7	

A: 5×5 dim.

<s>
The
cat
sat
on

V: $5 \times d$ dim.

K cache

<s>
The
cat
sat
on

V cache

<s>
The
cat
sat
on

Inference through an LLM



$$\frac{(q_4 k_0^T) v_0}{S_4} + \frac{(q_4 k_1^T) v_1}{S_4} + \frac{(q_4 k_2^T) v_2}{S_4} + \frac{(q_4 k_3^T) v_3}{S_4}$$

<s>	The	cat	sat	
-----	-----	-----	-----	--

Q: $5 \times d$ dim.

<s>
The
cat
sat
on

1				
0.2	0.8			
0.1	0.3	0.6		
0.01	0.19	0.3	0.5	
0.03	0.07	0.1	0.3	0.4

<s>	The	cat	sat	on
-----	-----	-----	-----	----

$K^T: d \times 5$ dim.

<s>	The	cat	sat	on				
0	1	2	3	4	5	6	7	

V: $5 \times d$ dim.

<s>
The
cat
sat
on

K cache

<s>
The
cat
sat
on

V cache

<s>
The
cat
sat
on

Compute V emb. of on



$$\frac{(q_4 k_0^T) v_0}{S_4} + \frac{(q_4 k_1^T) v_1}{S_4} + \frac{(q_4 k_2^T) v_2}{S_4} + \frac{(q_4 k_3^T) v_3}{S_4}$$

<s>	The	cat	sat	
-----	-----	-----	-----	--

Q: $5 \times d$ dim.

<s>
The
cat
sat
on

1				
0.2	0.8			
0.1	0.3	0.6		
0.01	0.19	0.3	0.5	
0.03	0.07	0.1	0.3	0.4

<s>	The	cat	sat	on
-----	-----	-----	-----	----

$K^T: d \times 5$ dim.

<s>	The	cat	sat	on				
0	1	2	3	4	5	6	7	

V: $5 \times d$ dim.

<s>
The
cat
sat
on

K cache

<s>
The
cat
sat
on

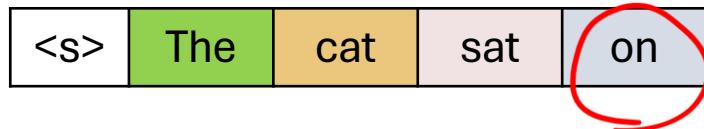
V cache

<s>
The
cat
sat
on

Add V emb. of on to V-cache



$$\frac{(q_4 k_0^T) v_0}{S_4} + \frac{(q_4 k_1^T) v_1}{S_4} + \frac{(q_4 k_2^T) v_2}{S_4} + \frac{(q_4 k_3^T) v_3}{S_4} + \frac{(q_4 k_4^T) v_4}{S_4}$$



Q: $5 \times d$ dim.

< s >
The
cat
sat
on

1				
0.2	0.8			
0.1	0.3	0.6		
0.01	0.19	0.3	0.5	
0.03	0.07	0.1	0.3	0.4

< s >	The	cat	sat	on
-------	-----	-----	-----	----

K^T: $d \times 5$ dim.

V: $5 \times d$ dim.

< s >
The
cat
sat
on

K cache

< s >
The
cat
sat
on

V cache

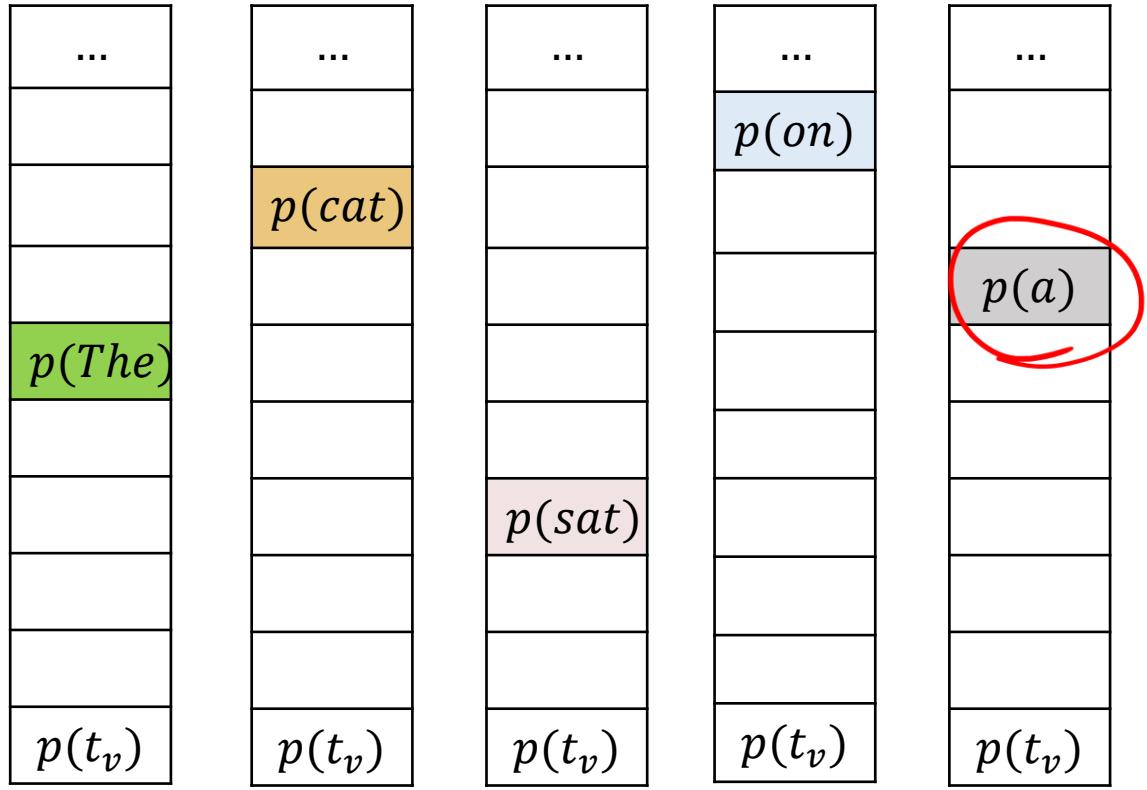
< s >
The
cat
sat
on

We get output emb. of on

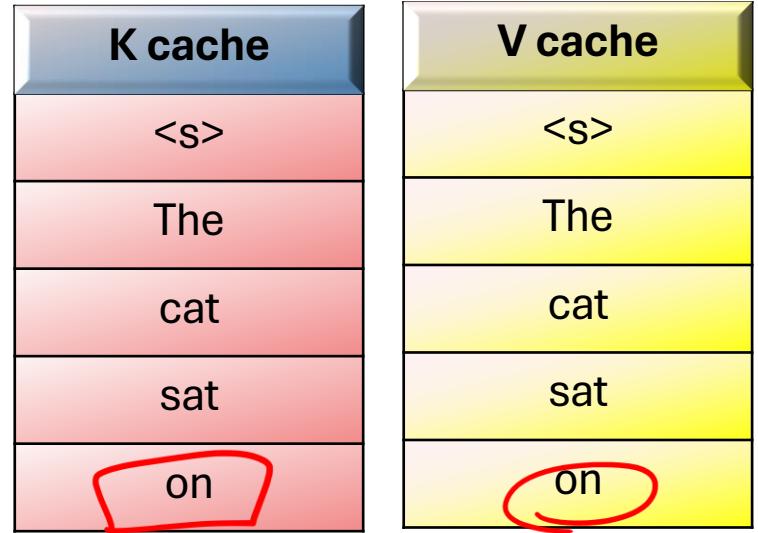
< s >	The	cat	sat	on				
0	1	2	3	4	5	6	7	

Inference through an LLM





Inference through an LLM

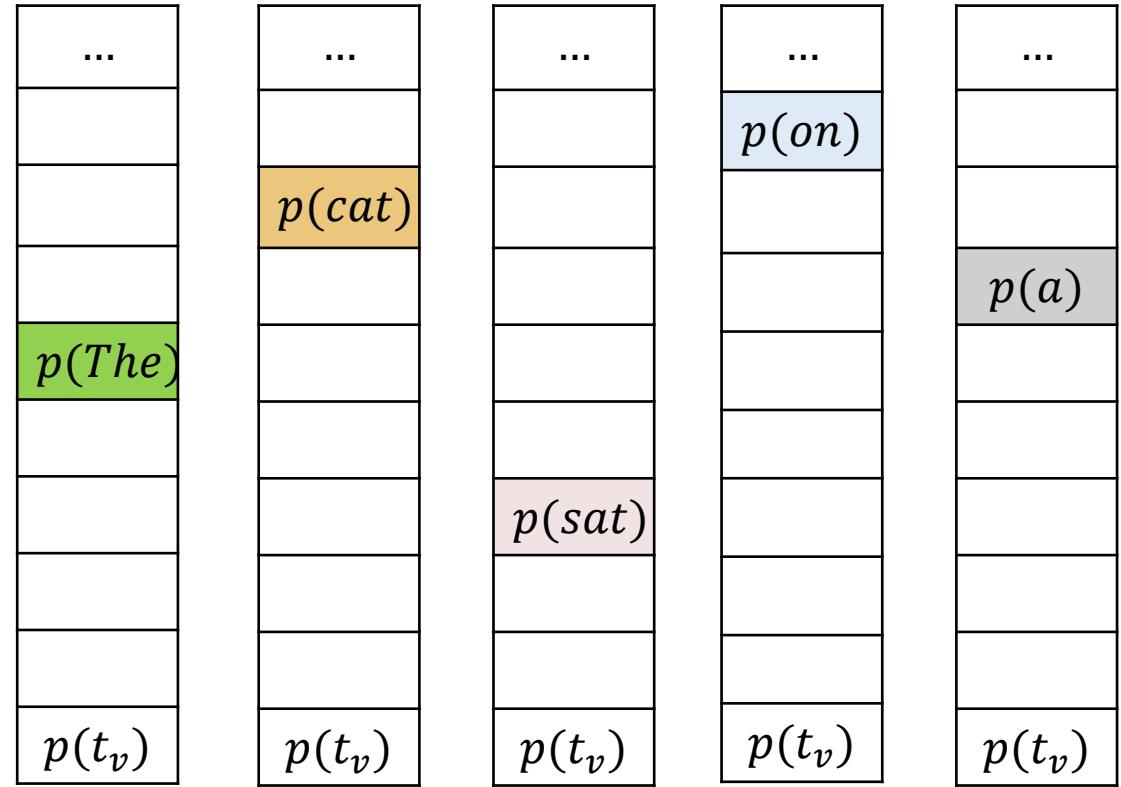


Transformer based LLM (θ)

<s>	The	cat	sat	on			
0	1	2	3	4	5	6	7



Inference through an LLM



Fill at step 5

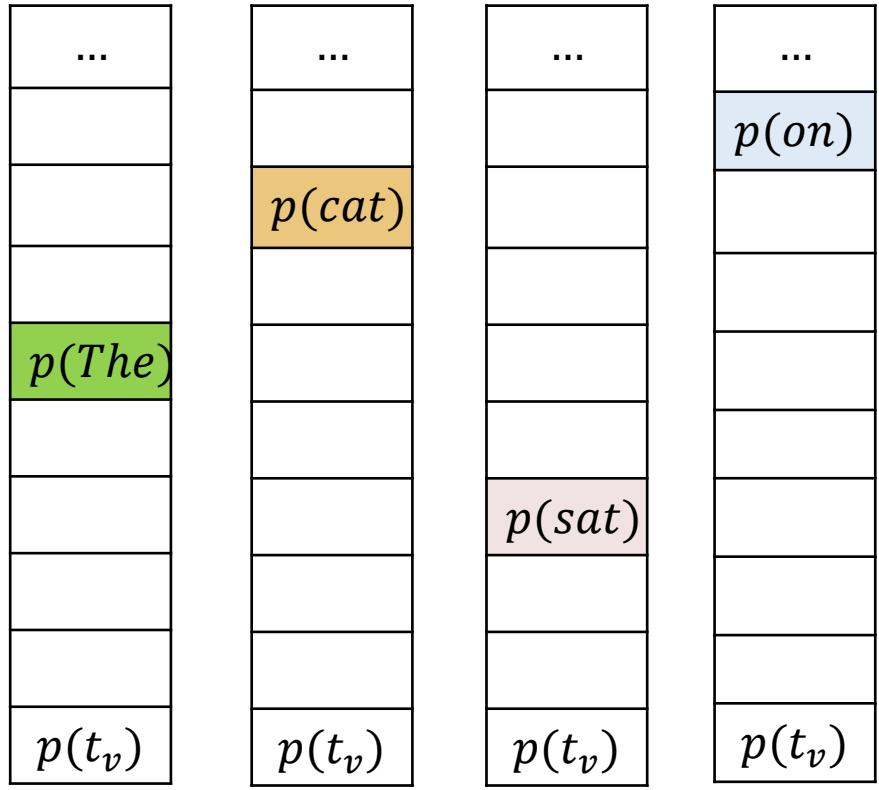
K cache
<s>
The
cat
sat
on

V cache
<s>
The
cat
sat
on

Transformer based LLM (θ)

<s>	The	cat	sat	on	a		
0	1	2	3	4	5	6	7





Inference through an LLM

K cache
<s>
The
cat
sat
on

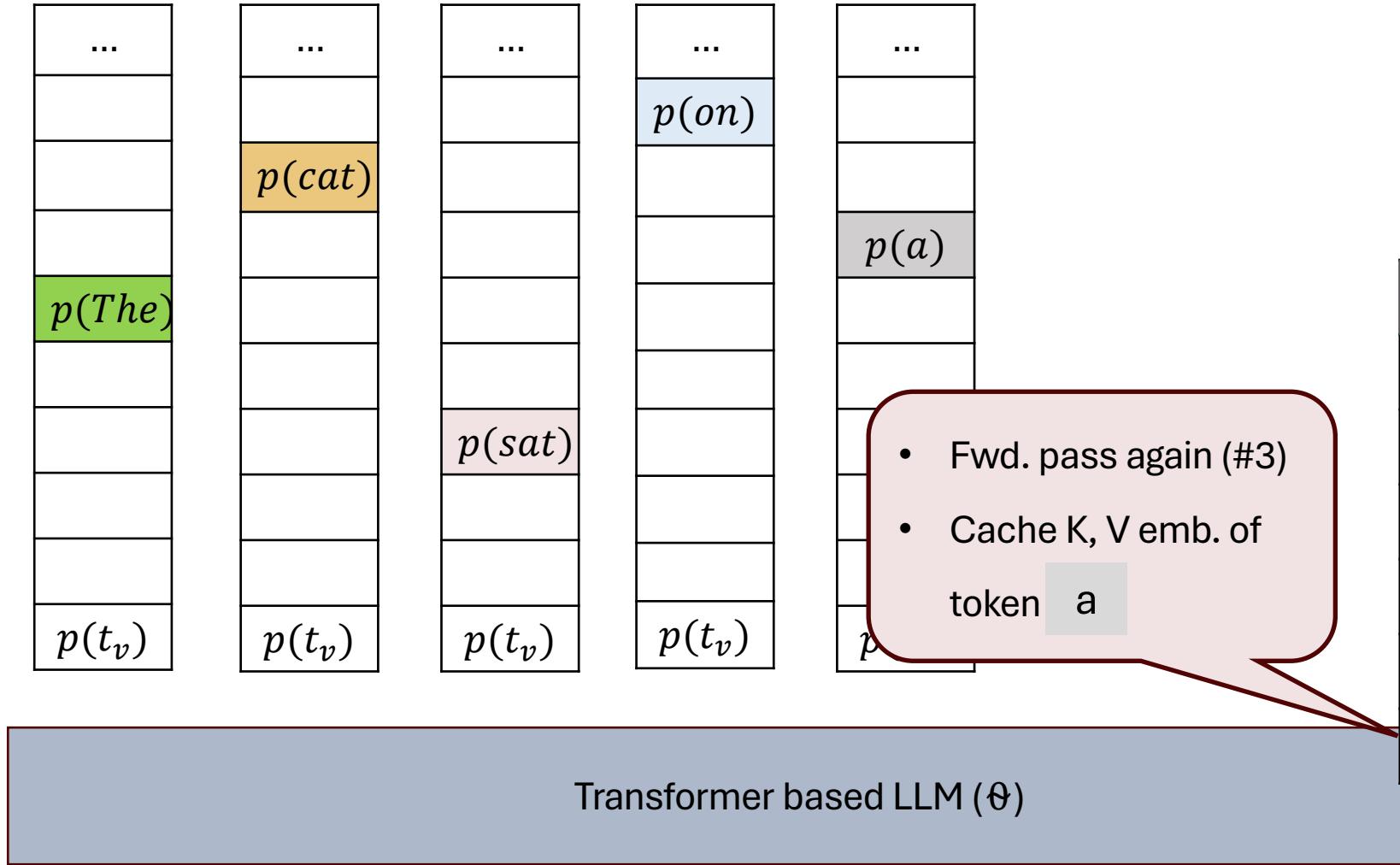
V cache
<s>
The
cat
sat
on

Transformer based LLM (θ)

<s>	The	cat	sat	on	a		
0	1	2	3	4	5	6	7



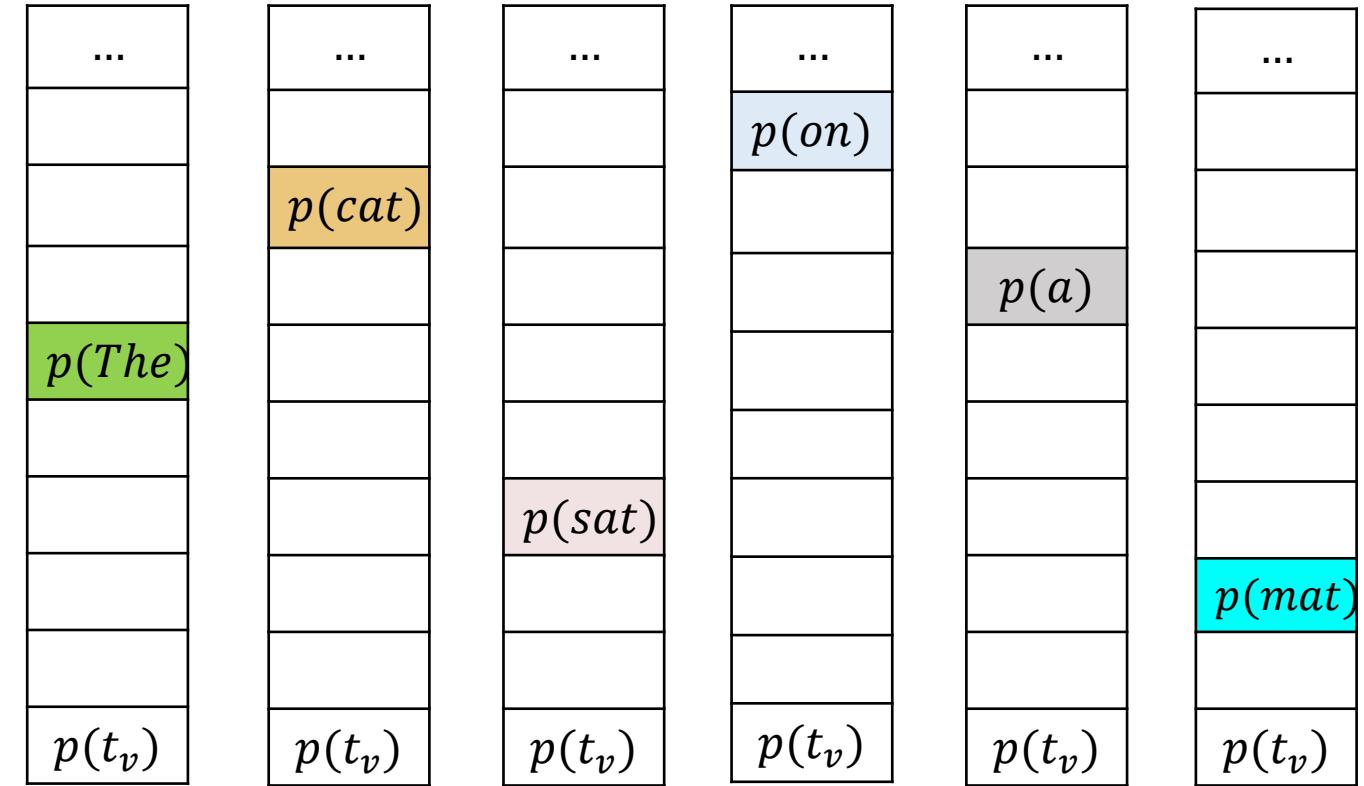
Inference through an LLM



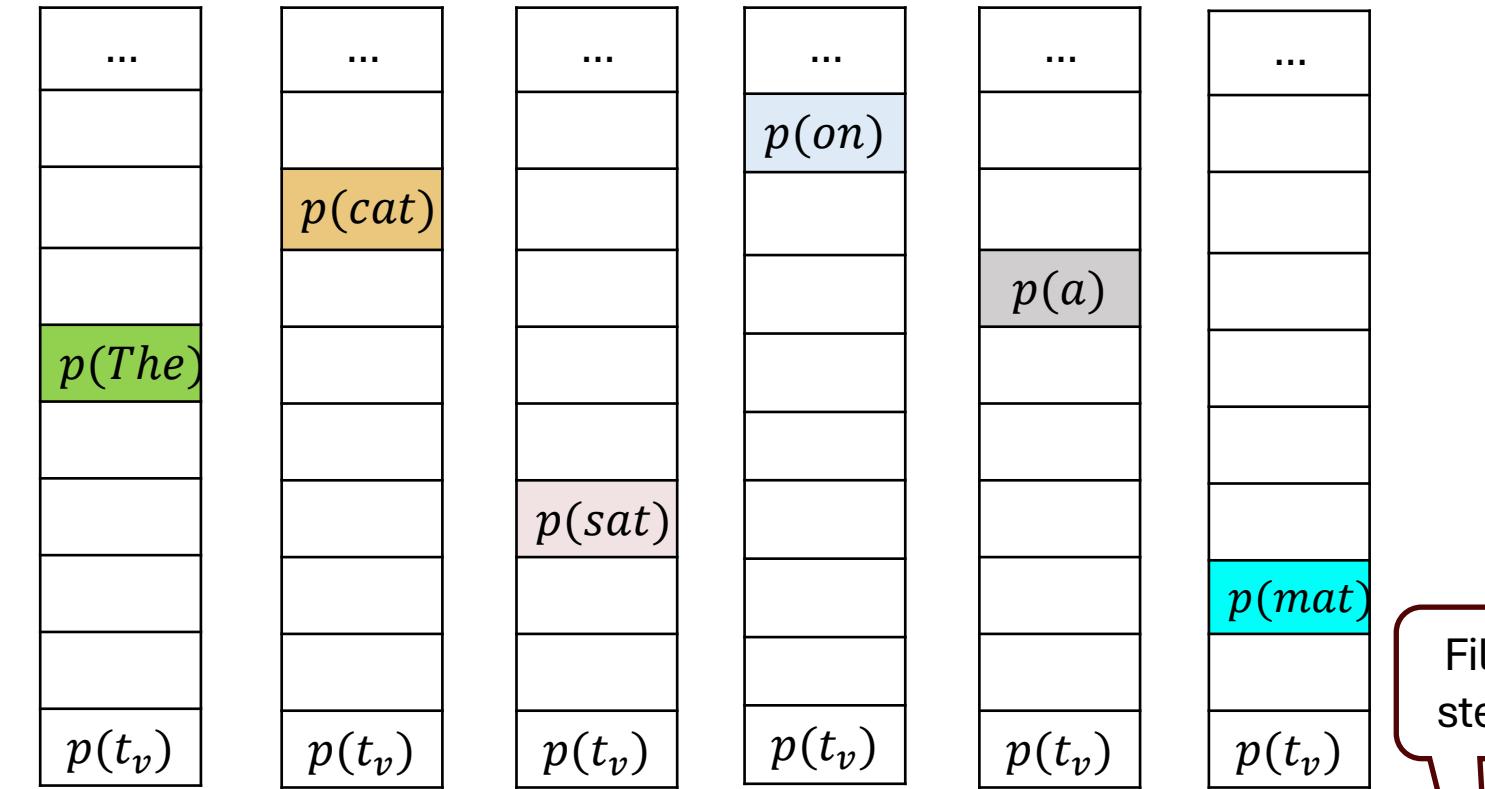
<s>	The	cat	sat	on	a		
0	1	2	3	4	5	6	7



Inference through an LLM



Inference through an LLM



Transformer based LLM (θ)

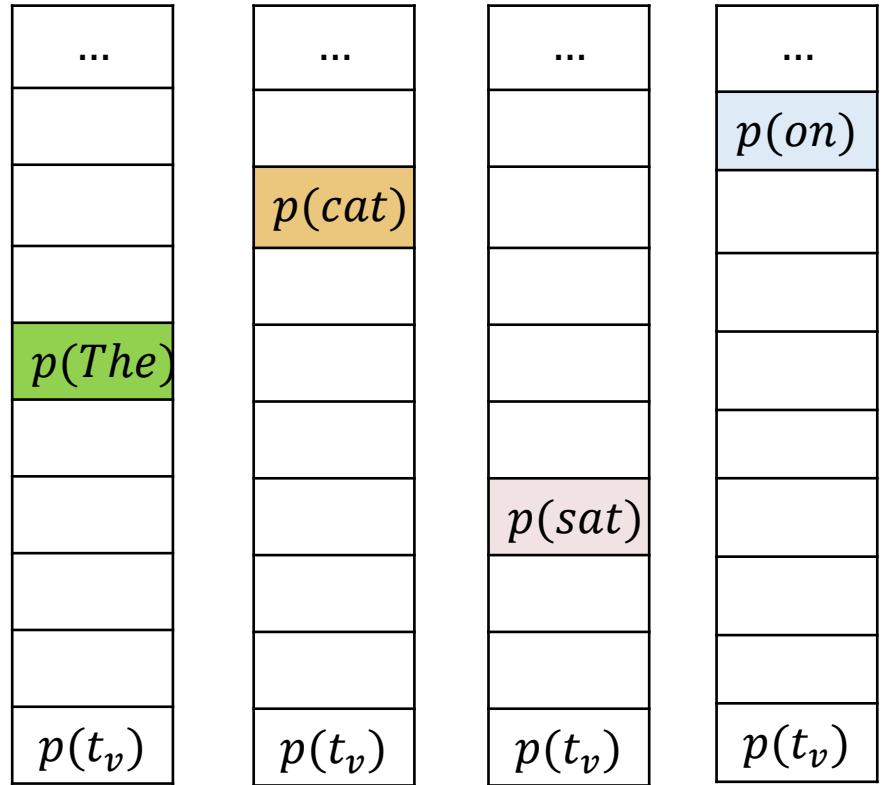
Fill at step 6

K cache
<s>
The
cat
sat
on
a

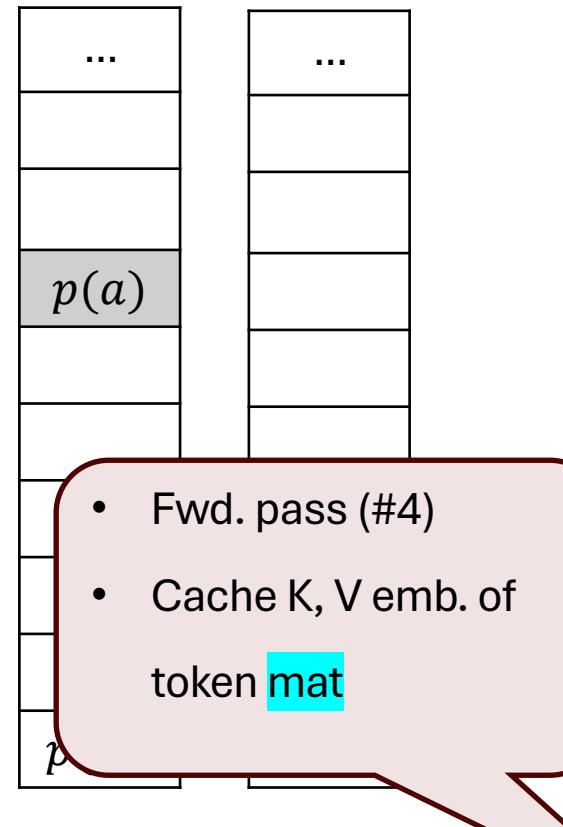
V cache
<s>
The
cat
sat
on
a

<s>	The	cat	sat	on	a	mat	
0	1	2	3	4	5	6	7

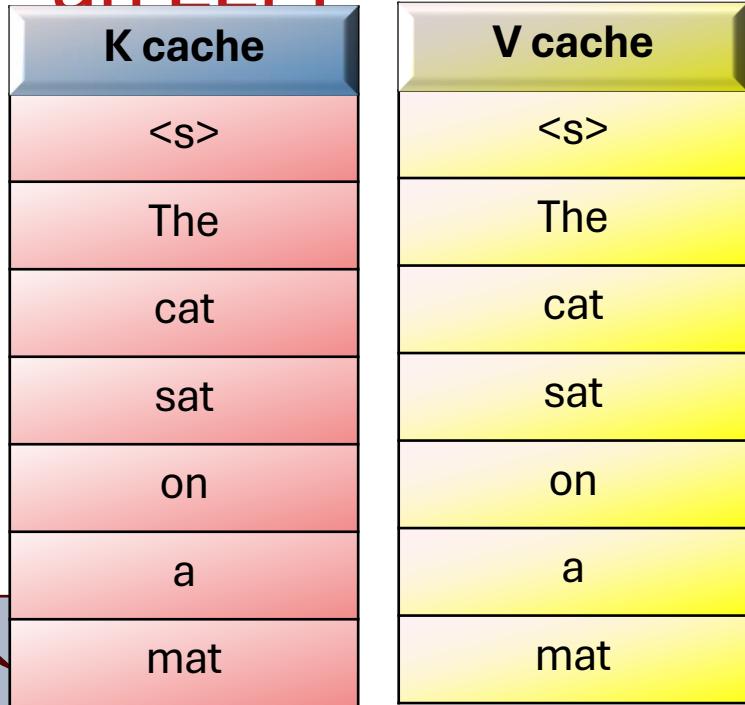




Transformer based LLM (θ)



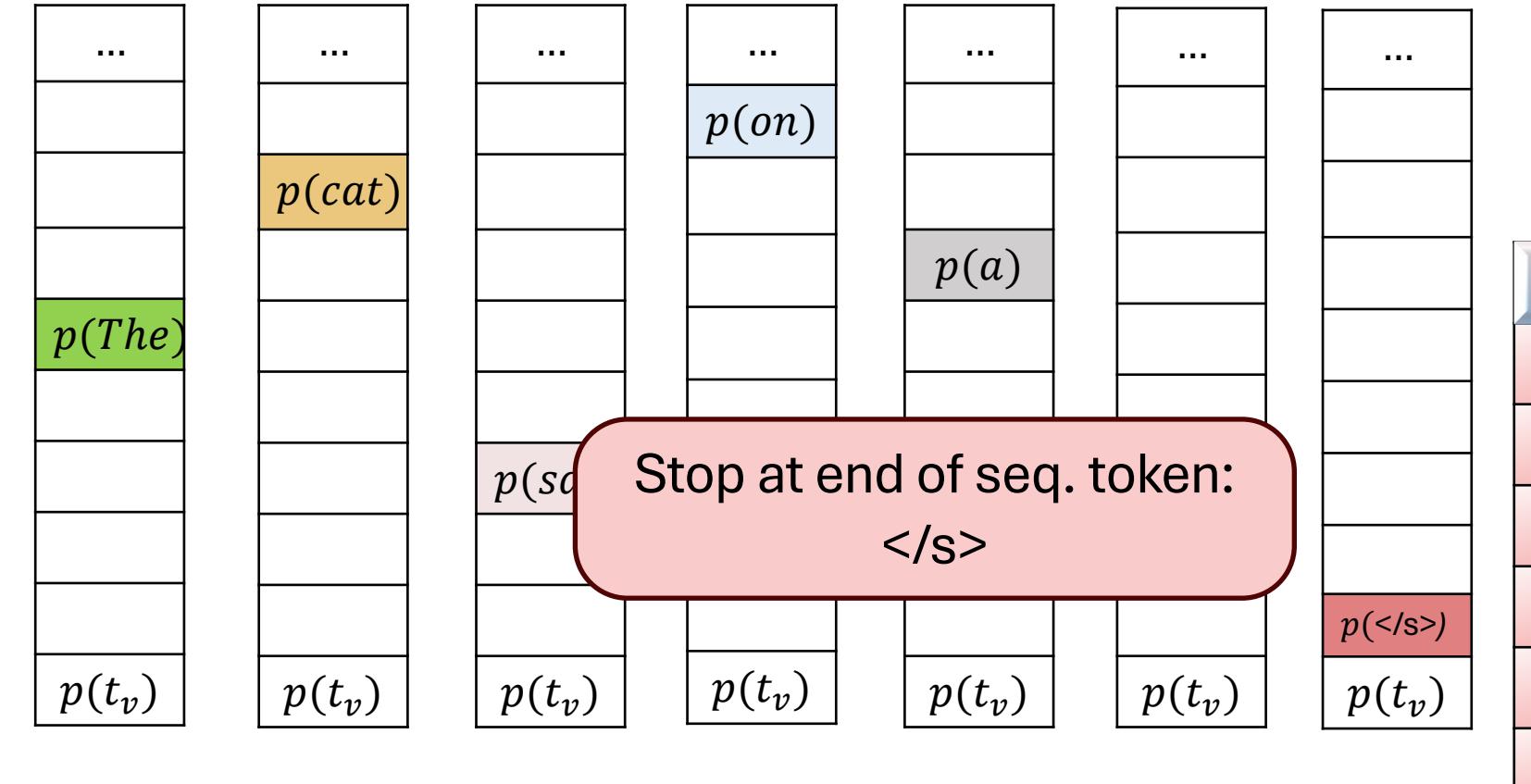
Inference through an LLM



<S>	The	cat	sat	on	a	mat	
0	1	2	3	4	5	6	7

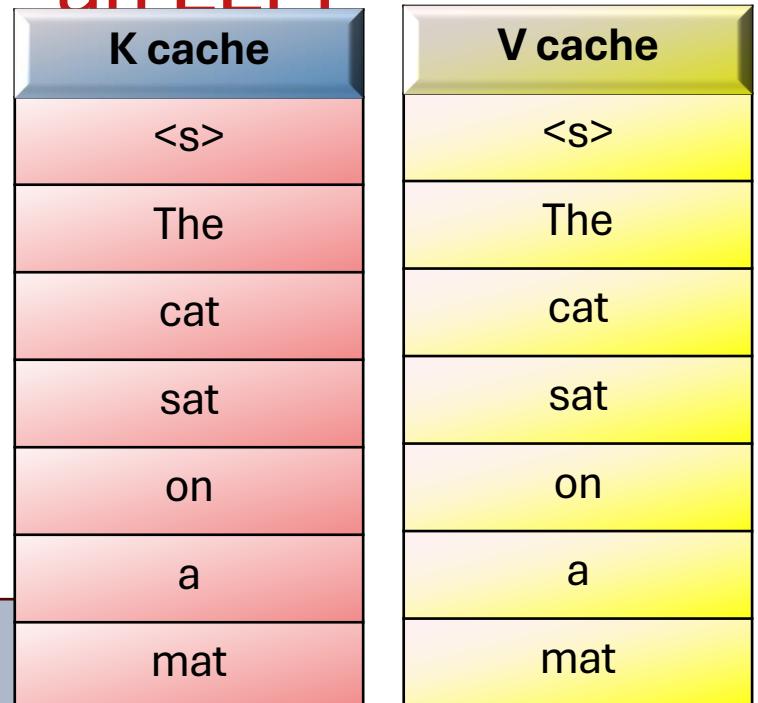


Inference through an LLM

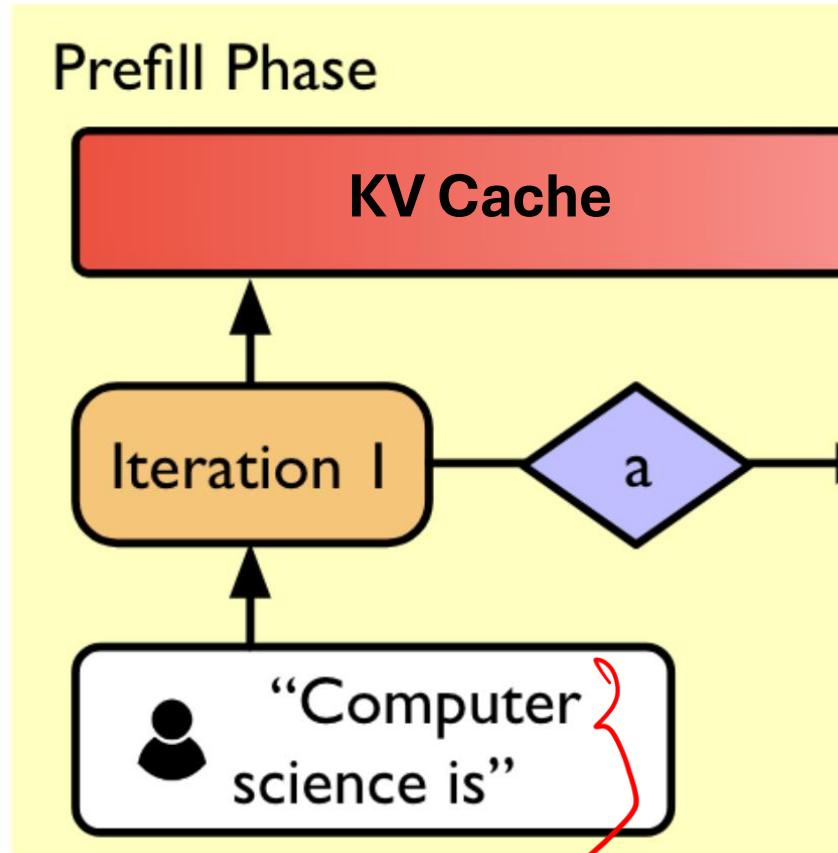


Transformer based LLM (θ)

<s>	The	cat	sat	on	a	mat	</s>
0	1	2	3	4	5	6	7



Two stages of LLM inference



- 1st forward pass (**Pre-fill step**) **Highly parallel**
 - ❖ The entire prompt is embedded and encoded – High latency
 - ❖ Multi-head attention computes the keys and values (KV)
 - ❖ Large matrix multiplication, high usage of the hardware accelerator

Content credits: [Li et al, 2024 LLM Inference Serving: Survey of Recent Advances](#)



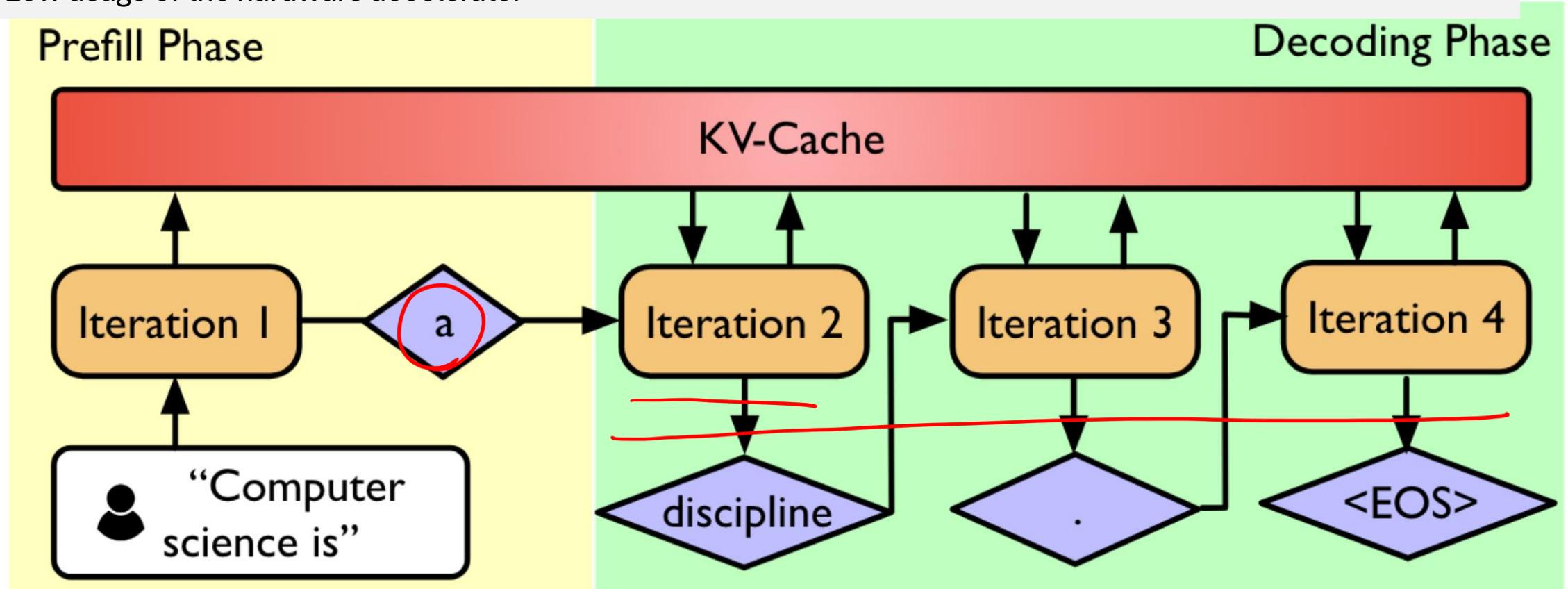
LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Remaining forward passes (Output generation): **sequential**

- The answer is generated **one token** at a time – Low latency per step
- Each generated token is **appended** to the previous input
- The process is repeated until the **stopping criteria** is met (max. length or EOS)
- Low usage of the hardware accelerator



Content credits: [Li et al, 2024 LLM Inference Serving: Survey of Recent Advances and Opportunities](#)



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Inference through an LLM

- 1st forward pass (**Pre-fill step**) **Highly parallel**
 - The entire prompt is embedded and encoded – High latency
 - Multi-head attention computes the keys and values (KV)
 - Large matrix multiplication, high usage of the hardware accelerator
- Remaining forward passes (**Output generation**): **sequential**
 - The answer is generated **one token** at a time – Low latency per step
 - Each generated token is **appended** to the previous input
 - The process is repeated until the **stopping criteria** is met (max. length or EOS)
 - Low usage of the hardware accelerator

Content credits: <https://www.slideshare.net/slideshow/julien-simon-deep-dive-optimizing-llm-inference-69d3/270921961>



Memory Usage of KV cache

$$2 * \underbrace{precision}_{k/v} * \underbrace{N_{layers}}_{-} * \underbrace{d_{model}}_{-} * \underbrace{seqlen}_{-} * \underbrace{batch}_{-}$$

2 : Two matrices for K and V

precision : bytes per parameter (e.g. 4 for fp32)

N_{layers} : layers in the model

d_{model} : dimension of embeddings

seqlen : length of context in tokens

batch : batch size

Content credits: https://www.youtube.com/watch?v=80blUggRJf4&t=1s&ab_channel=EfficientNLP



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Memory Usage of KV cache: Example OPT-13B

$$2 * \text{precision} * N_{\text{layers}} * d_{\text{model}} * \text{seqlen} * \text{batch}$$

- 2 : Two matrices for K and V
precision : bytes per parameter (e.g. 4 for fp32)
N_{layers} : layers in the model
d_{model} : dimension of embeddings
seqlen : length of context in tokens
batch : batch size

2 (KV)
2 bytes (fp16)
40 layers
5120 dim.
2048 tokens
10



Content credits: https://www.youtube.com/watch?v=80bIUggRJf4&t=1s&ab_channel=EfficientNLP



Memory Usage of KV cache: Example OPT-13B

$$2 * \text{precision} * N_{\text{layers}} * d_{\text{model}} * \text{seqlen} * \text{batch}$$

KV Cache: 17 GB

Model Size: $2^*13 = 26$ GB

On a 40GB A100

- 65% (26GB) used by model parameters
- ~30% (12 GB) available for KV cache
- Expected throughput ~ 8 batch size of 2048 tokens

2 (KV)
2 bytes (fp16)
40 layers
5120 dim.
2048 tokens
10

Content credits: https://www.youtube.com/watch?v=80blUggRJf4&t=1s&ab_channel=EfficientNLP



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Memory Management of KV Cache

Prompt A: “*The cat sat*”

Max Tokens: 2048

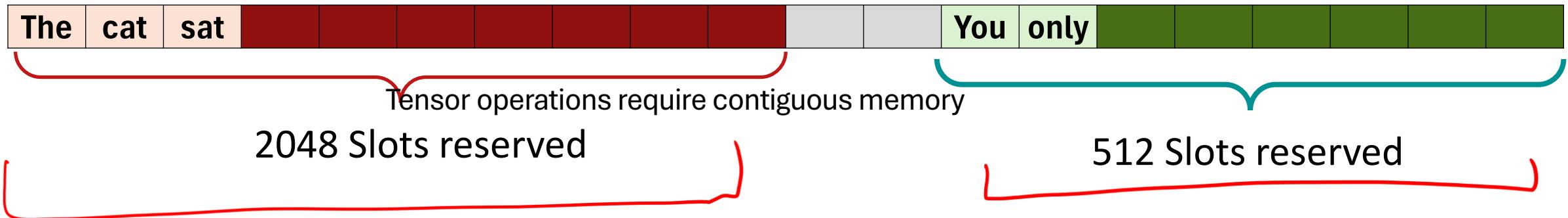
Prompt B : “*You only*”

Max Tokens: 512



Memory Management of KV Cache

Tensor operations require contiguous memory

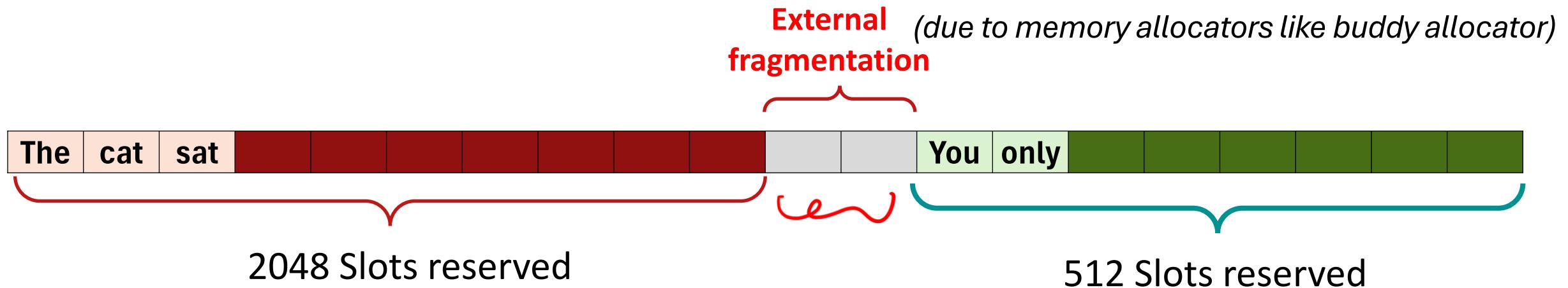


Prompt A : “**The cat sat**”
Max Tokens: **2048**

Prompt B : “**You only**”
Max Tokens: **512**



Memory Management of KV Cache



Prompt A : **"The cat sat"**

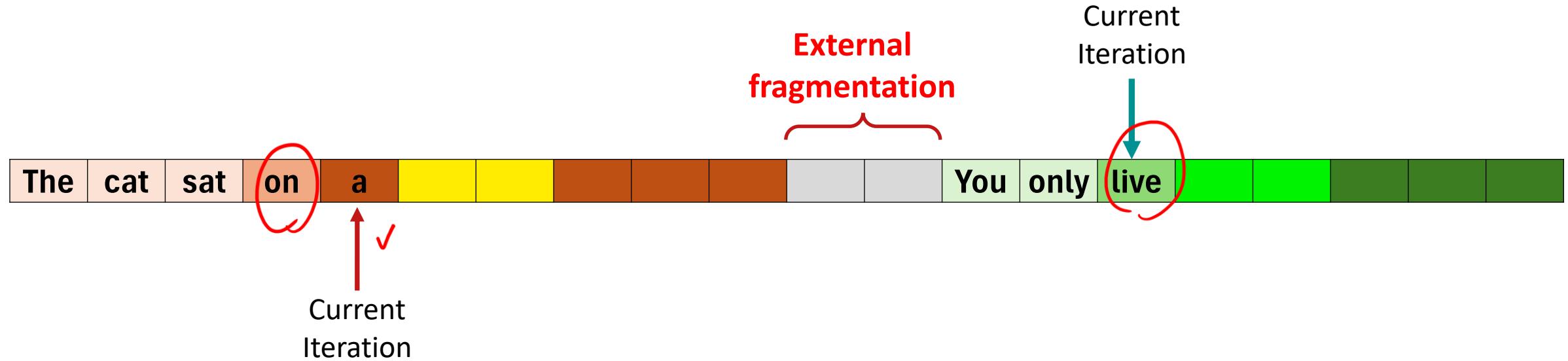
Max Tokens: **2048**

Prompt B : **"You only"**

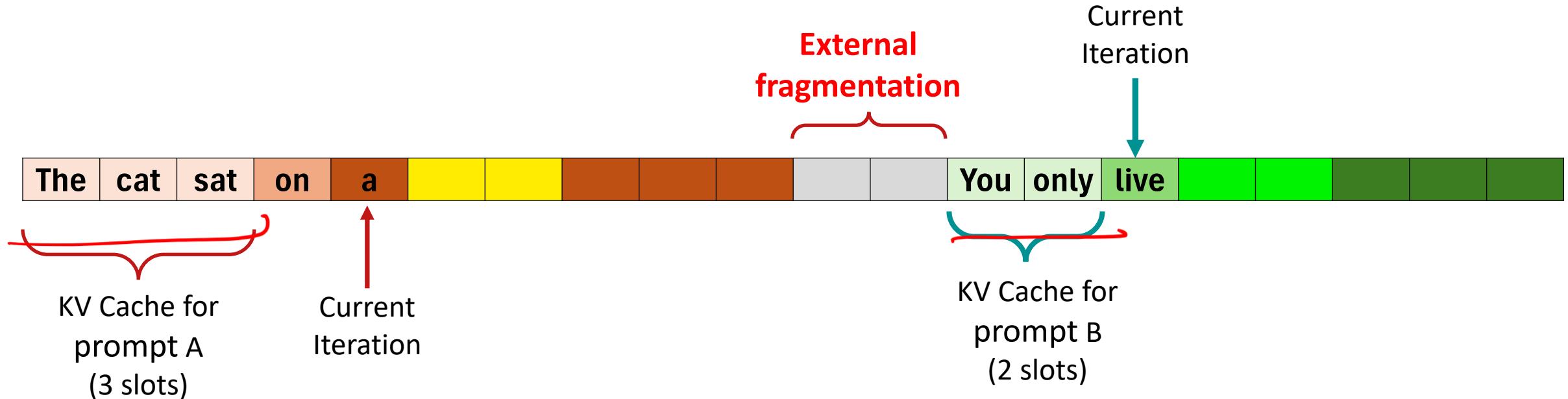
Max Tokens: **512**



Memory Management of KV Cache



Memory Management of KV Cache



Prompt A: “*The cat sat*”

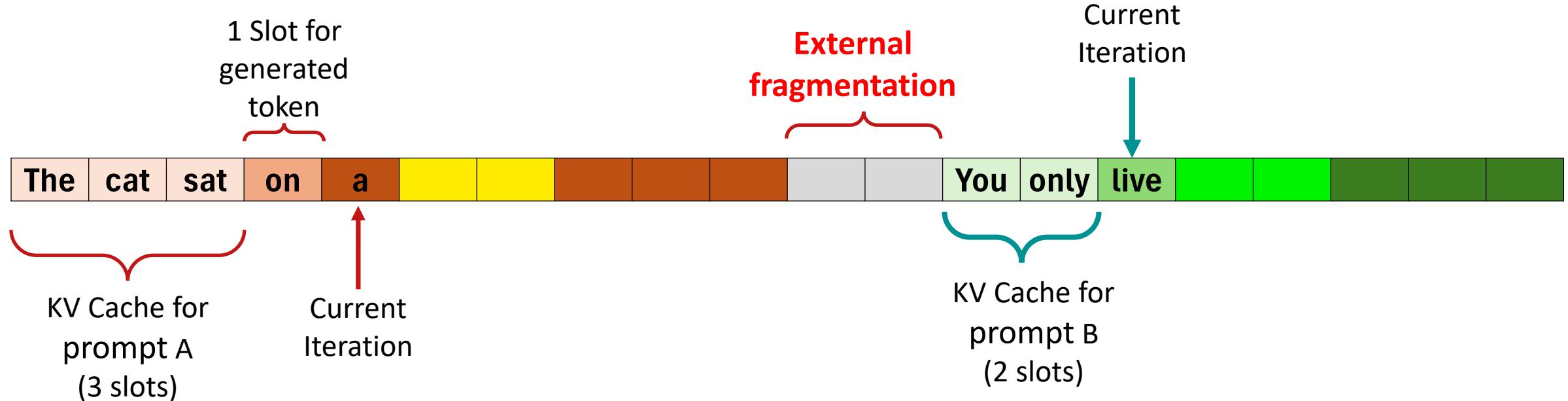
Max Tokens: 2048

Prompt B: “*You only*”

Max Tokens: 512



Memory Management of KV Cache



Prompt A: "*The cat sat*"

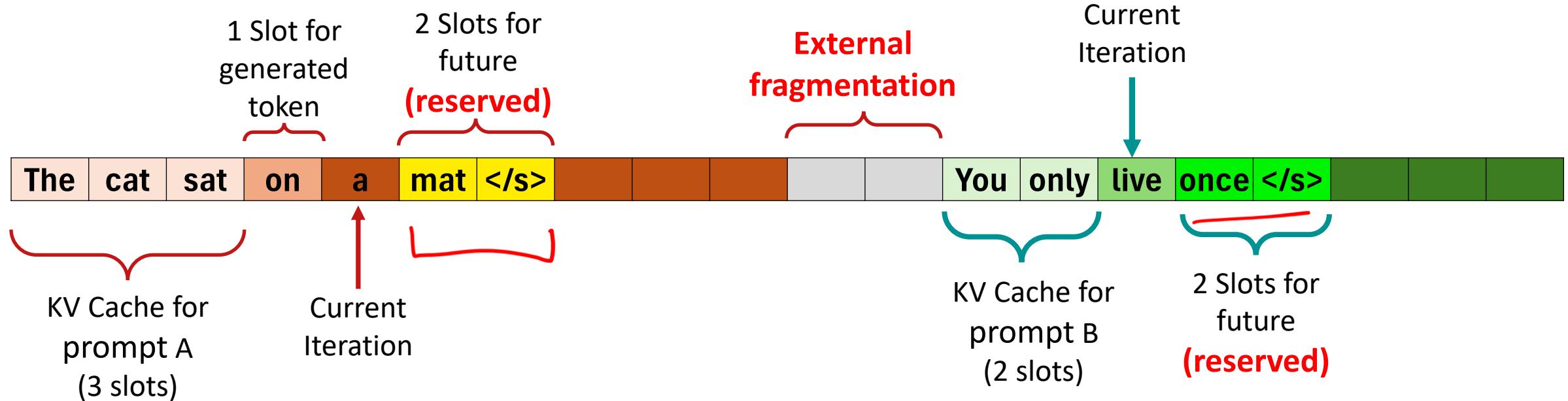
Max Tokens: 2048

Prompt B: "*You only*"

Max Tokens: 512



Memory Management of KV Cache



Prompt A: "*The cat sat*"

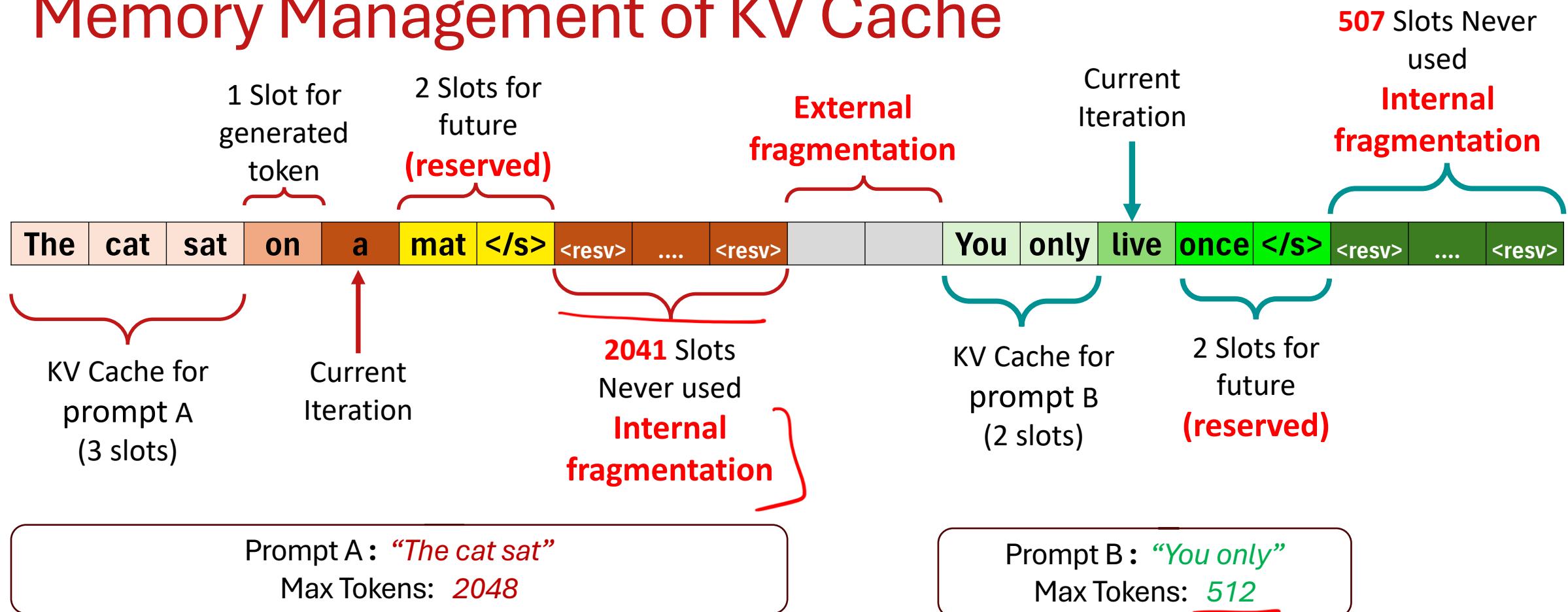
Max Tokens: 2048

Prompt B: "*You only*"

Max Tokens: 512



Memory Management of KV Cache



Memory Management of KV Cache

Chunk Pre-allocation scheme

- KV cache stored in contiguous memory
- Chunks of memory allocated statically, based on max. tokens.
- Actual input or eventual output length ignored while allocating memory



Memory Management of KV Cache

Chunk Pre-allocation scheme

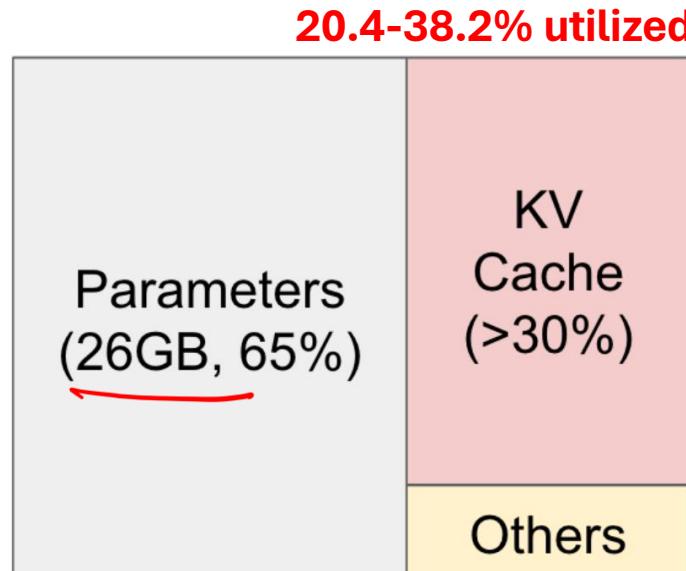
- KV cache stored in contiguous memory
- Chunks of memory allocated statically, based on max. tokens.
- Actual input or eventual output length ignored while allocating memory

Results in 3 types of memory wastes –

- **Reserved slots** for future tokens
- **Internal fragmentation** due to over-provisioning for maximum sequence lengths
- **External fragmentation** from the memory allocator.



Memory Layout for 13B-OPT model on A100 (40GB)



NVIDIA A100 40GB

Content credits: https://www.youtube.com/watch?v=5ZlavKF_98U&t=1646s&ab_channel=Anyscale

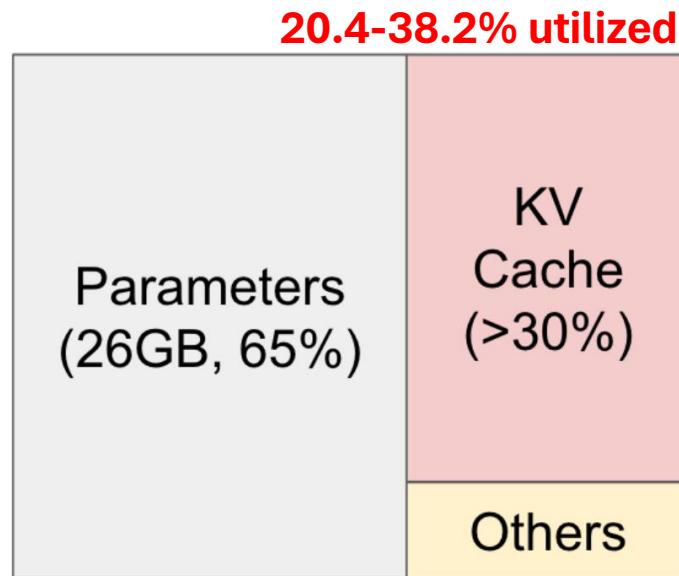


LLMs: Introduction and Recent Advances

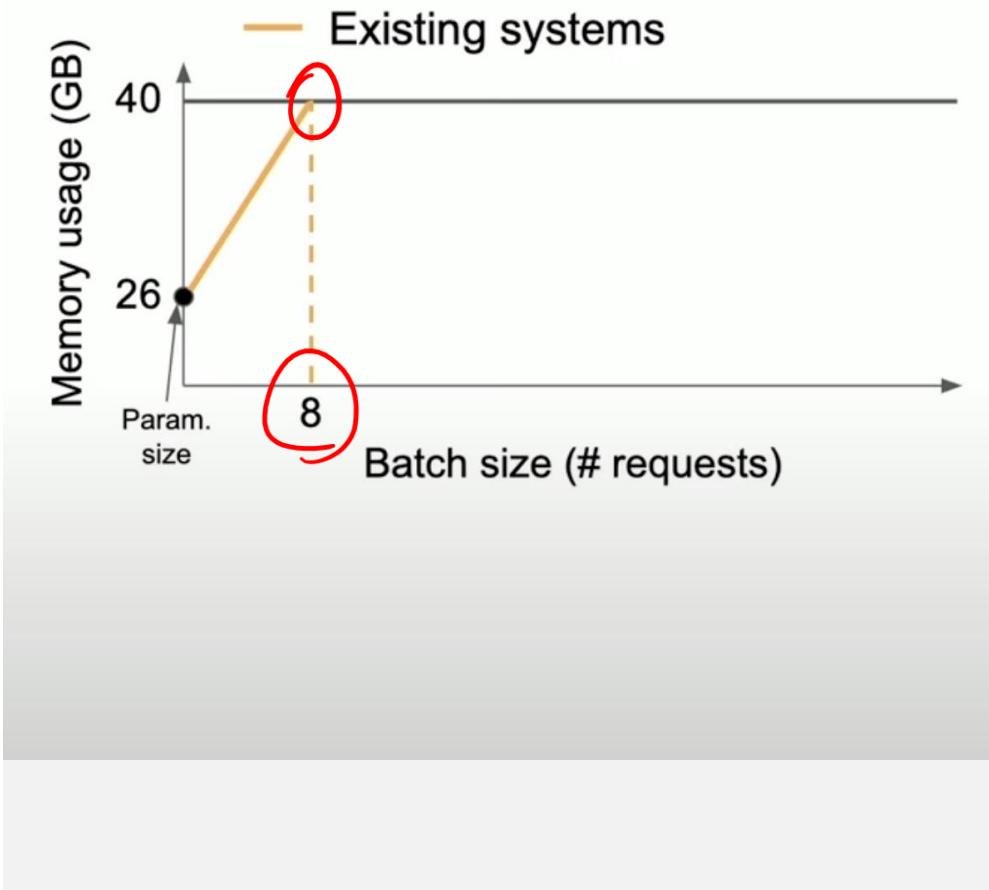


Tanmoy Chakraborty

Memory Layout for 13B-OPT model on A100 (40GB)



NVIDIA A100 40GB



Existing systems

- max batch size - 8

Content credits: https://www.youtube.com/watch?v=5ZlavKF_98U&t=1646s&ab_channel=Anyscale

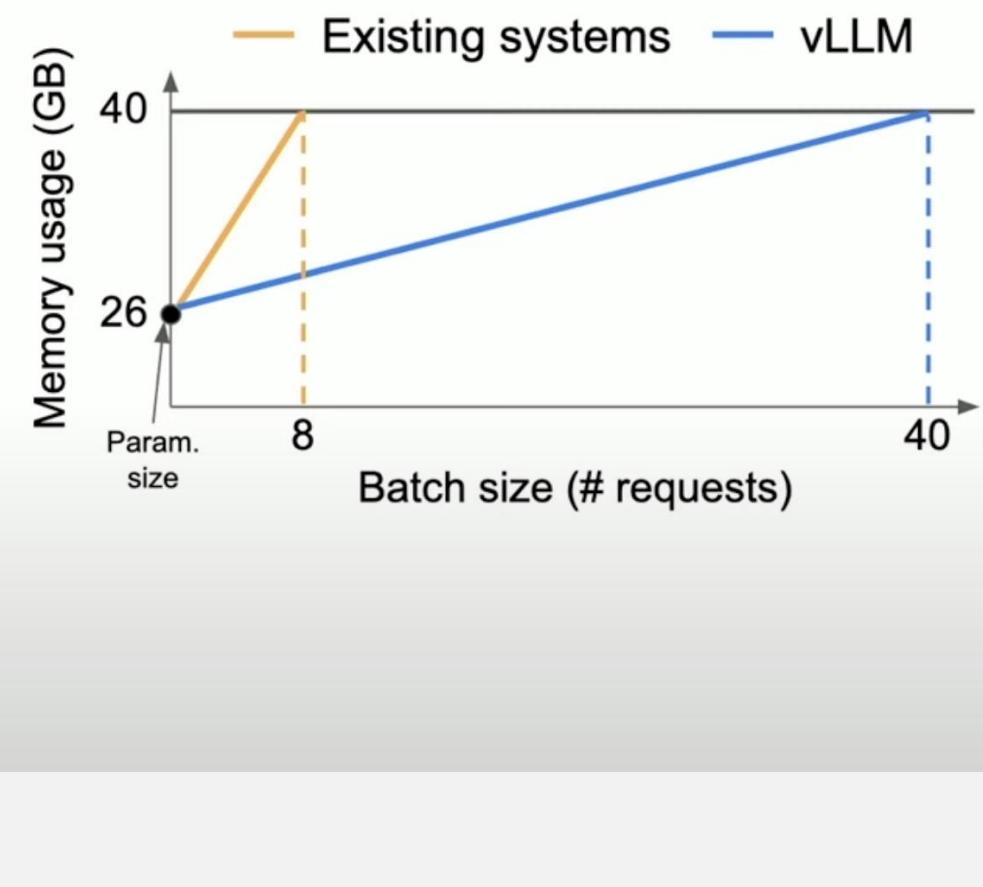
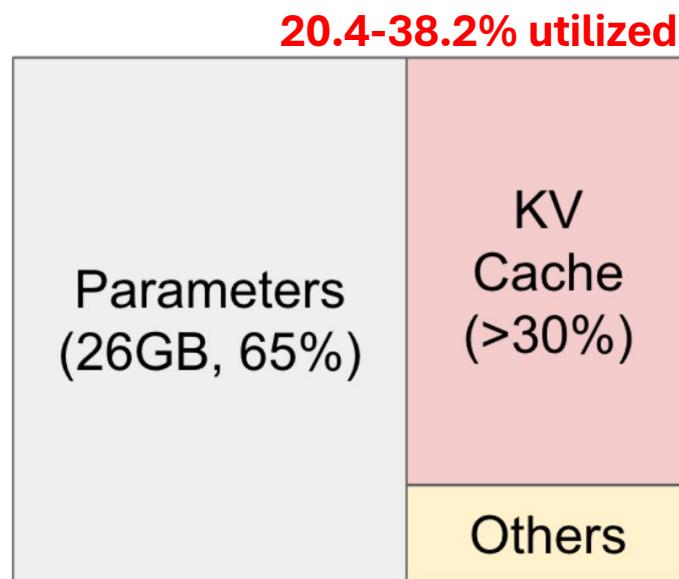


LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Memory Layout for 13B-OPT model on A100 (40GB)



Existing systems

- max batch size - 8

vLLM (paged attention)

- Max batch size ~ 40

Content credits: https://www.youtube.com/watch?v=5ZlavKF_98U&t=1646s&ab_channel=Anyscale

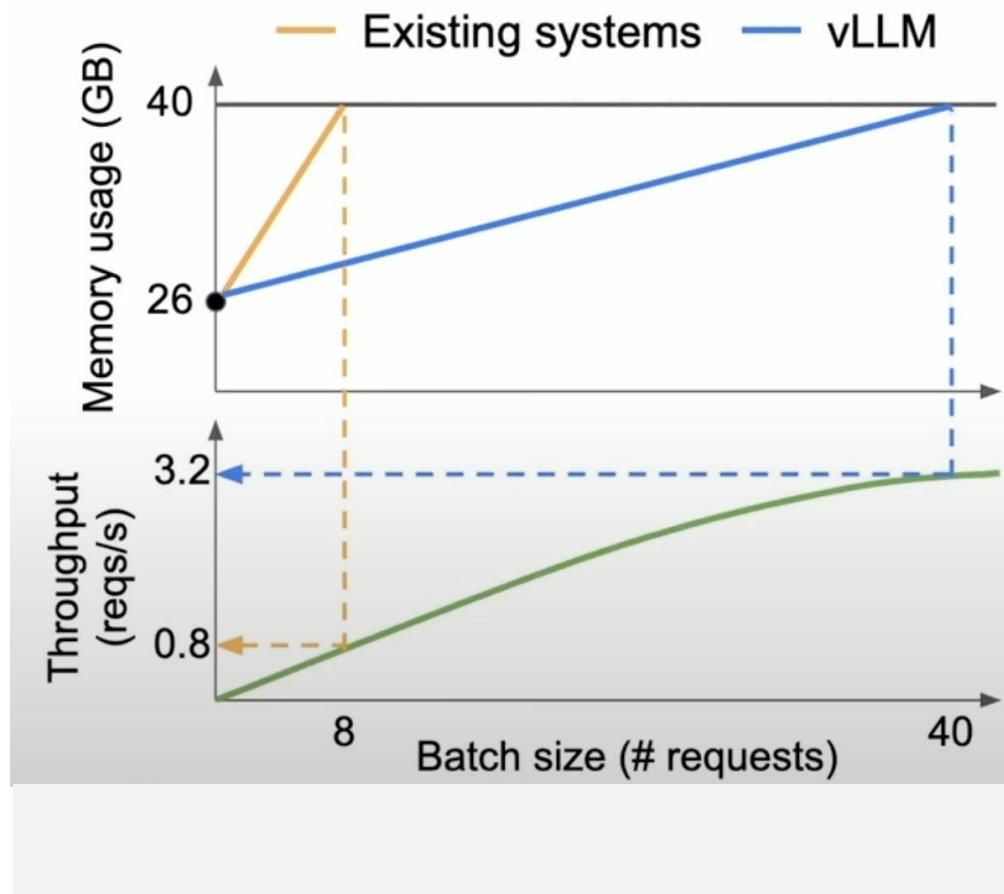
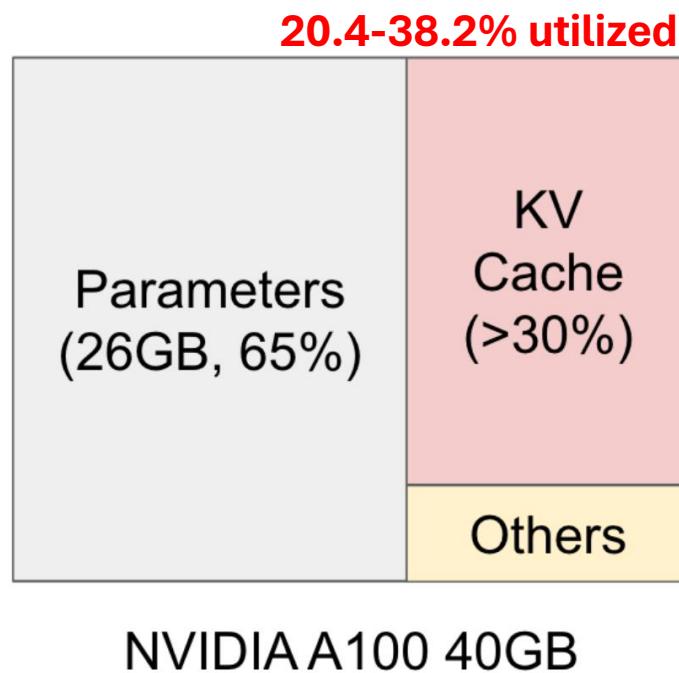


LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Memory Layout for 13B-OPT model on A100 (40GB)



Existing systems

- max batch size - 8
- ~ 0.8 requests / sec

vLLM (paged attention)

- Max batch size ~ 38
- ~ 3.2 requests per sec

Content credits: https://www.youtube.com/watch?v=5ZlavKF_98U&t=1646s&ab_channel=Anyscale



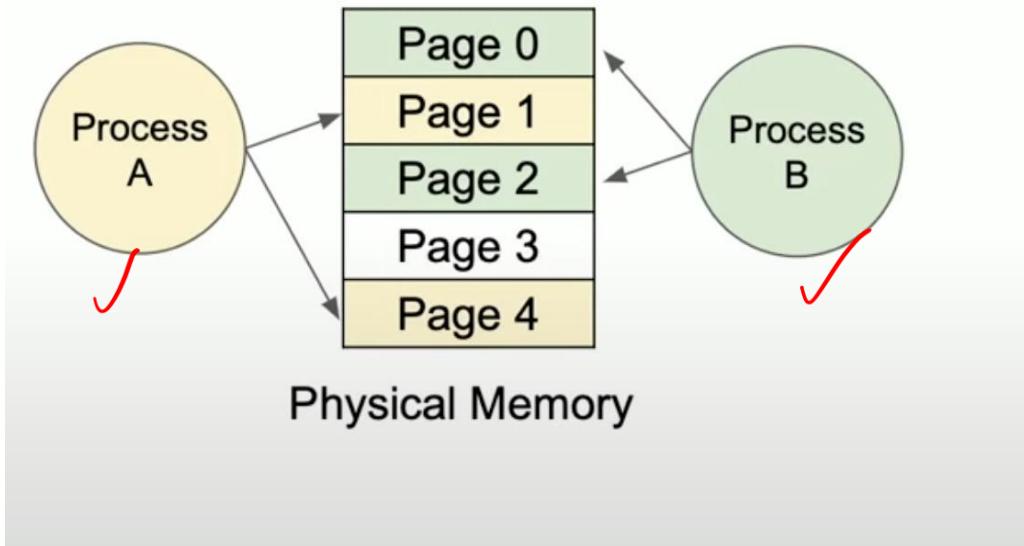
LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

vLLM: Efficient KV cache management

Inspired by **Virtual memory** and paging



Memory management in OS

Content credits: https://www.youtube.com/watch?v=5ZlavKF_98U&t=1646s&ab_channel=Anyscale



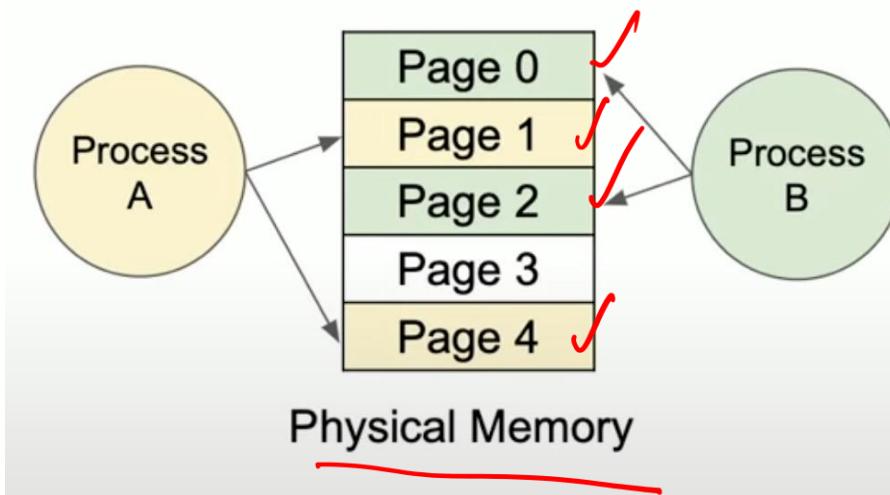
LLMs: Introduction and Recent Advances



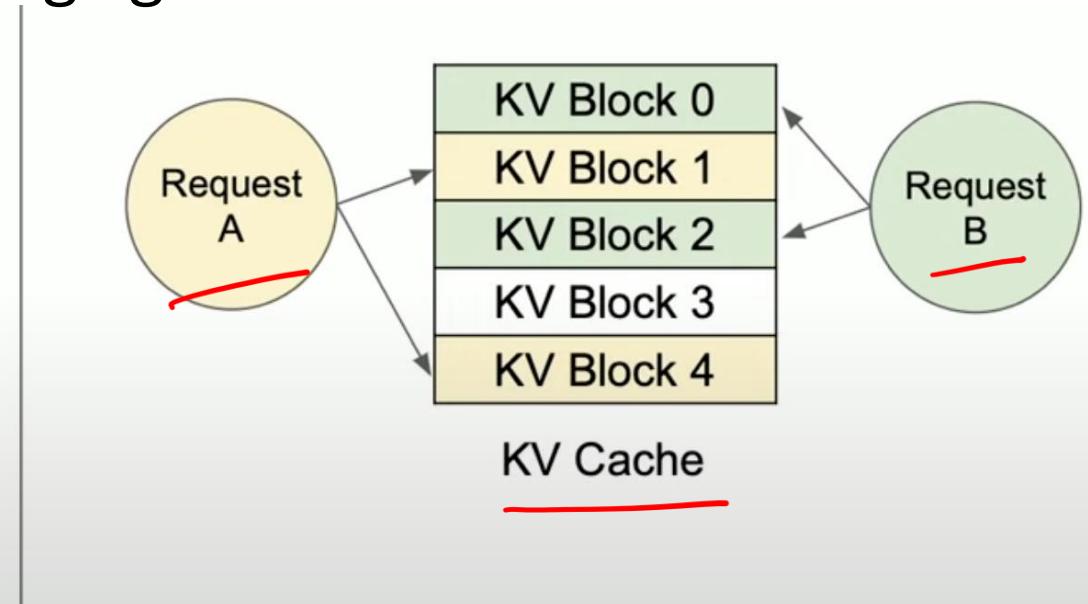
Tanmoy Chakraborty

vLLM: Efficient KV cache management

Inspired by **Virtual memory** and paging



Memory management in OS



Memory management in vLLM

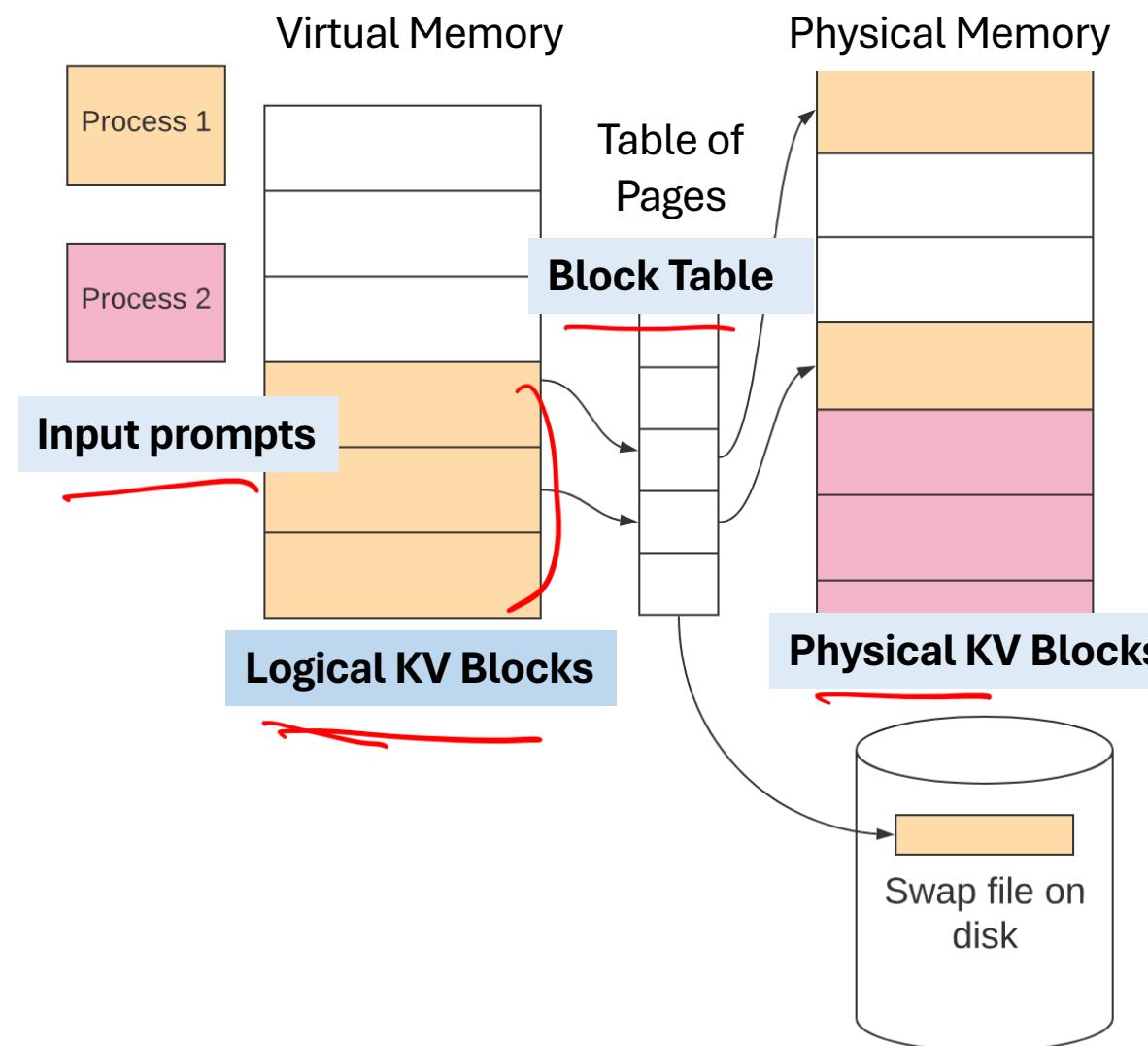
Content credits: https://www.youtube.com/watch?v=5ZlavKF_98U&t=1646s&ab_channel=Anyscale



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty



Efficient KV cache management

Inspired by **Virtual memory** and paging

- ❑ Processes as **incoming requests** (input to the model)
- ❑ Virtual Memory to **Logical KV Blocks**
- ❑ Physical Memory to **Physical KV Blocks**
- ❑ Page table to **Block Table**



KV Blocks

KV Cache

Content credits: https://www.youtube.com/watch?v=5ZlavKF_98U&t=1646s&ab_channel=Anyscale



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

KV Blocks

Block size 4

Block 0			
Block 1			
Block 2			
Block 3			
Block 4			
Block 5			
Block 6			
Block 7			

Content credits: https://www.youtube.com/watch?v=5ZlavKF_98U&t=1646s&ab_channel=Anyscale

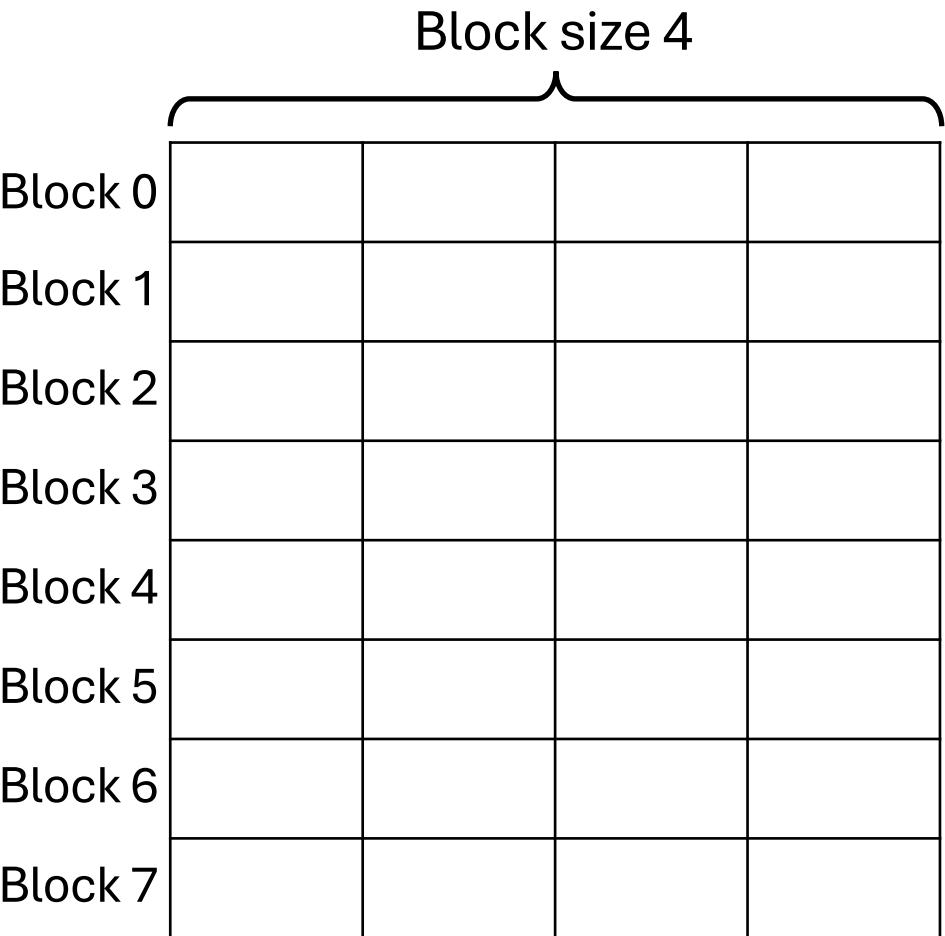


LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

KV Blocks



Physical KV Blocks

Content credits: https://www.youtube.com/watch?v=5ZlavKF_98U&t=1646s&ab_channel=Anyscale

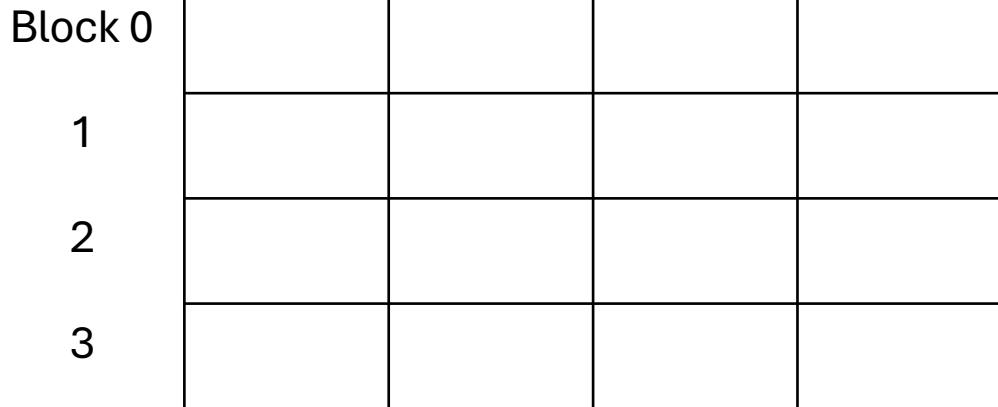


LLMs: Introduction and Recent Advances

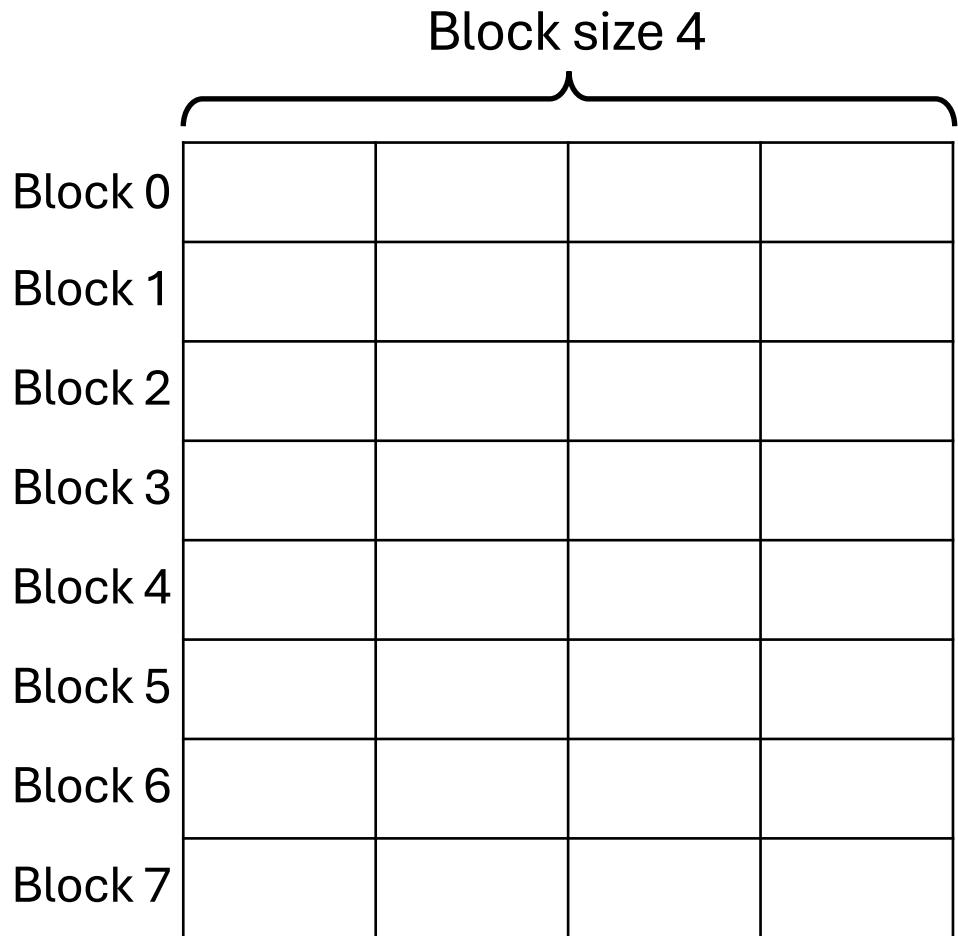


Tanmoy Chakraborty

Physical vs Logical KV Blocks



Logical KV Blocks



Physical KV Blocks

Content credits: https://www.youtube.com/watch?v=5ZlavKF_98U&t=1646s&ab_channel=Anyscale



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Physical vs Logical KV Blocks

	Block 0	Block 1	Block 2	Block 3
1				
2				
3				

Logical KV Blocks

Phys. Block	# Filled

Block Table

Block size 4

Block 0			
Block 1			
Block 2			
Block 3			
Block 4			
Block 5			
Block 6			
Block 7			

Physical KV Blocks

Content credits: https://www.youtube.com/watch?v=5ZlavKF_98U&t=1646s&ab_channel=Anyscale



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Physical vs Logical KV Blocks

Prompt: “Today we are learning about LLMs and”

Block 0	Today	we	are	learning
1	about	LLMs	and	
2				
3				

Logical KV Blocks

Phys. Block	# Filled

Block Table

Block 0				
Block 1				
Block 2				
Block 3				
Block 4				
Block 5				
Block 6				
Block 7				

Physical KV Blocks

Content credits: <https://youtu.be/yVXtLTcdO1Q?si=XO2Dk-VYOShUMH1u>



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Physical vs Logical KV Blocks

Prompt: “Today we are learning about LLMs and”

Block 0	Today	we	are	learning
1	about	LLMs	and	
2				
3				

Logical KV Blocks

Phys. Block	# Filled
7	4

Block Table

Block 0				
1				
2				
3				
4				
5				
6				
7	Today	we	are	learning

Physical KV Blocks

Content credits: <https://youtu.be/yVxtLTcdO1Q?si=XO2Dk-VYOShUMH1u>



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Physical vs Logical KV Blocks

Prompt: “Today we are learning about LLMs and”

Block 0			
1	Today	we	are
2	about	LLMs	and
3			
4			
5			
6			
7			

Logical KV Blocks

Phys. Block	# Filled
7	4
1	3

Block Table

Block size 4			
Block 0	1	2	3
	about	LLMs	and
4			
5			
6			
7	Today	we	are
8	learning		

Physical KV Blocks

Content credits: <https://youtu.be/yVXtLTcdO1Q?si=XO2Dk-VYOShUMH1u>



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Physical vs Logical KV Blocks

Prompt: “Today we are learning about LLMs and”

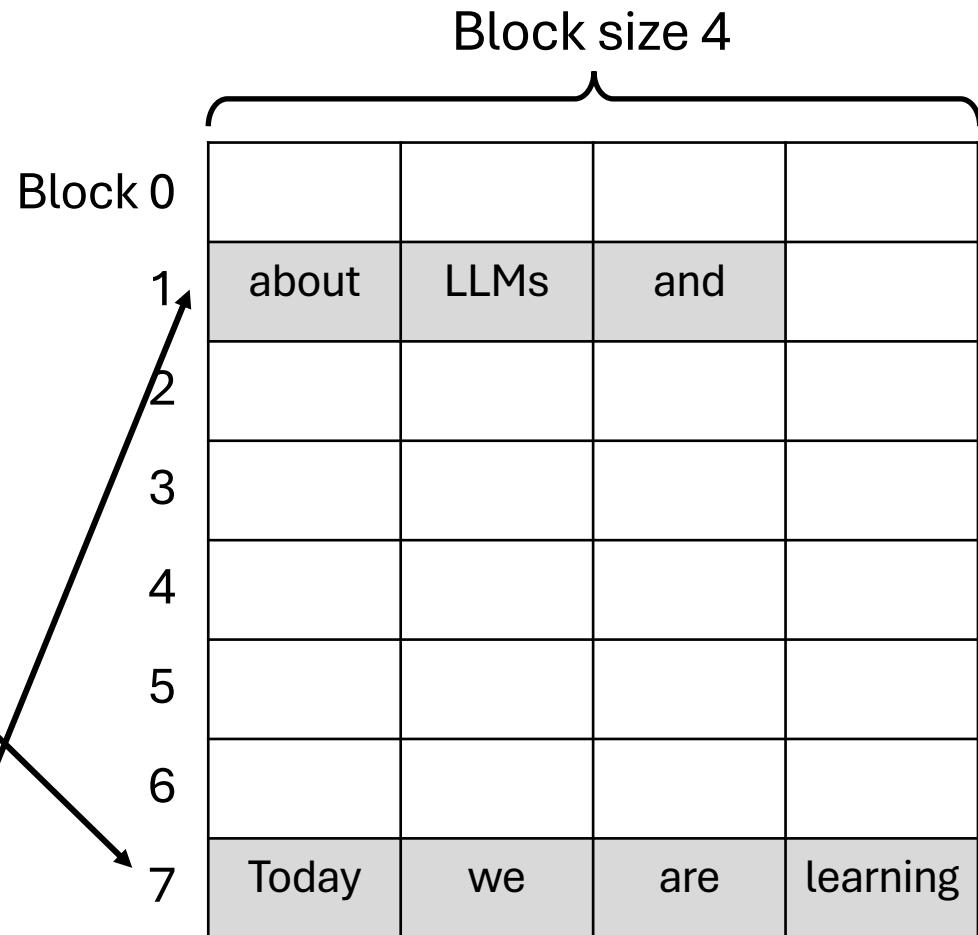
Completion: “*memory*”

Block 0	Today	we	are	learning
1	about	LLMs	and	memory
2				.
3				

Logical KV Blocks

Phys. Block	# Filled
7	4
1	4

Block Table



Physical KV Blocks

Content credits: <https://youtu.be/yVXtLTcdO1Q?si=XO2Dk-VYOShUMH1u>



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Physical vs Logical KV Blocks

Prompt: “Today we are learning about LLMs and”

Completion: “**memory**”

	Block 0	1	2	3
0	Today	we	are	learning
1	about	LLMs	and	memory
2				
3				

Logical KV Blocks

Phys. Block	# Filled
7	4
1	4

Block Table

Block size 4

Block 0	1	2	3	4	5	6	7
about	LLMs	and	memory				
Today	we	are	learning				

Physical KV Blocks

Content credits: <https://youtu.be/yVXtLTcdO1Q?si=XO2Dk-VYOShUMH1u>



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Physical vs Logical KV Blocks

Prompt: “Today we are learning about LLMs and”

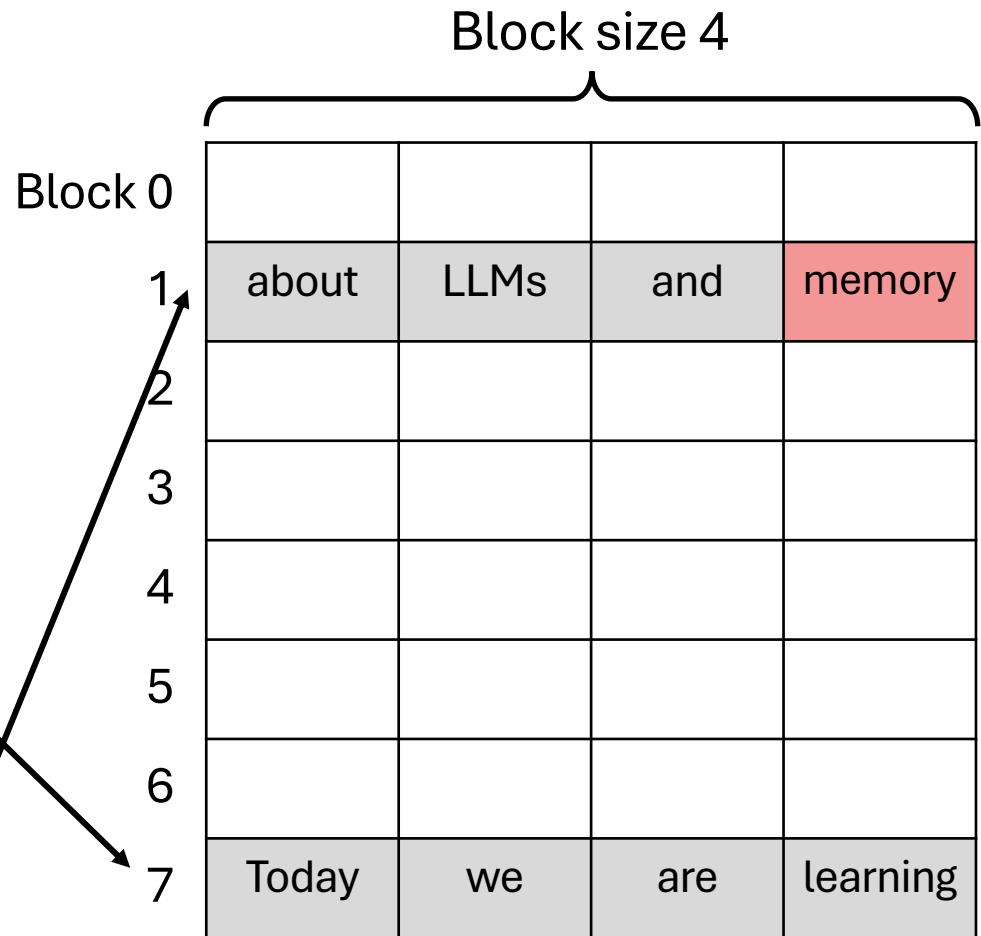
Completion: “**memory on**”

	Block 0	1	2	3
0	Today	we	are	learning
1	about	LLMs	and	memory
2	on			
3				

Logical KV Blocks

Phys. Block	# Filled
7	4
1	4

Block Table



Content credits: <https://youtu.be/yVXtLTcdO1Q?si=XO2Dk-VYOShUMH1u>



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Physical vs Logical KV Blocks

Prompt: “Today we are learning about LLMs and”

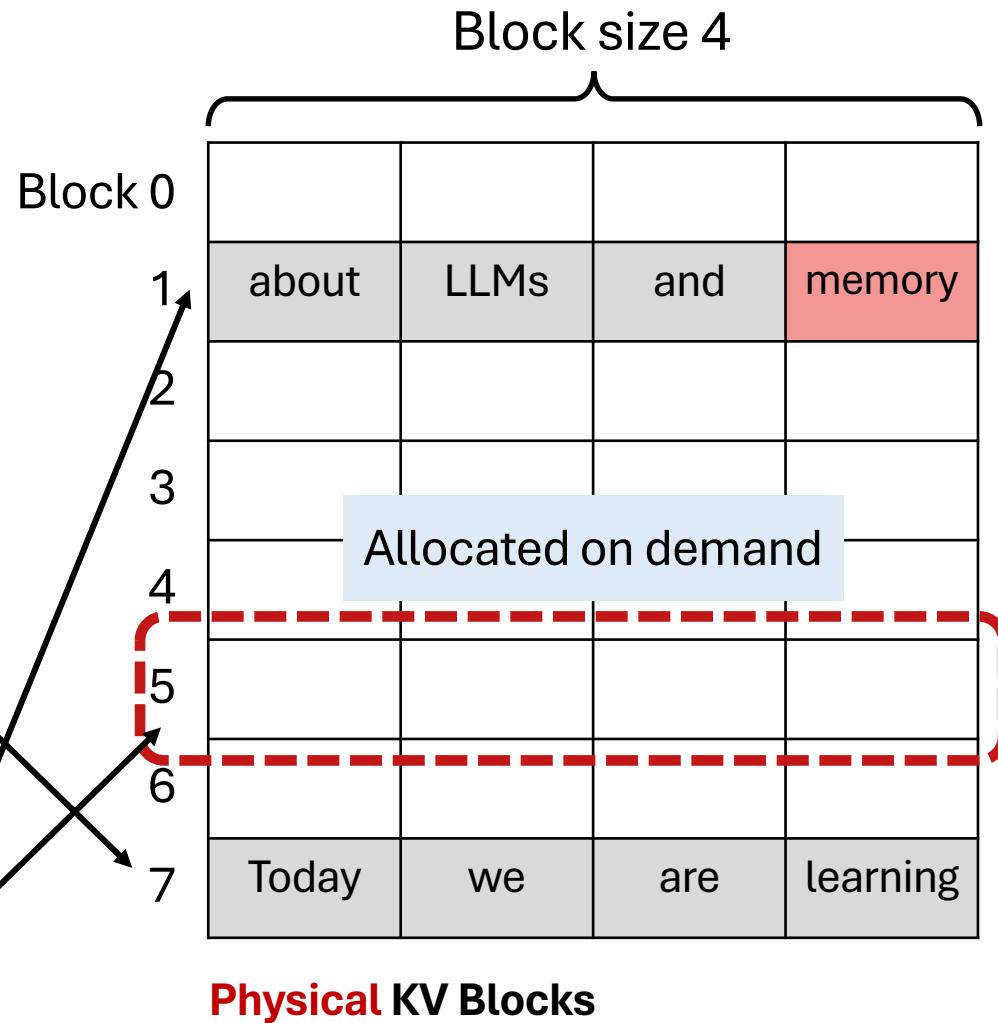
Completion: “**memory on**”

Today	we	are	learning
about	LLMs	and	memory
on			

Logical KV Blocks

Phys. Block	# Filled
7	4
1	4
5	1

Block Table



Content credits: <https://youtu.be/yVXtLTcdO1Q?si=XO2Dk-VYOShUMH1u>



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Physical vs Logical KV Blocks

Prompt: “Today we are learning about LLMs and”

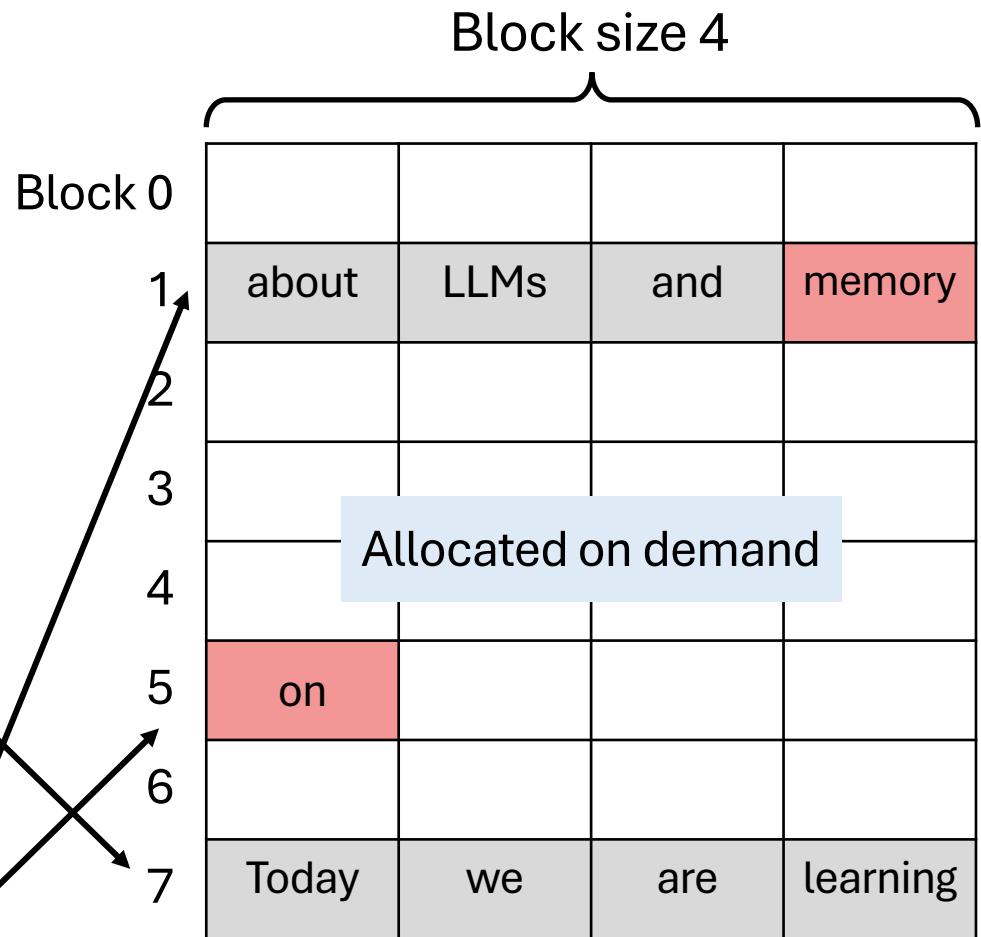
Completion: “**memory on**”

	Block 0	1	2	3
0	Today	we	are	learning
1	about	LLMs	and	memory
2	on			
3				

Logical KV Blocks

Phys. Block	# Filled
7	4
1	4
5	1

Block Table



Physical vs Logical KV Blocks

Prompt: “Today we are learning about LLMs and”

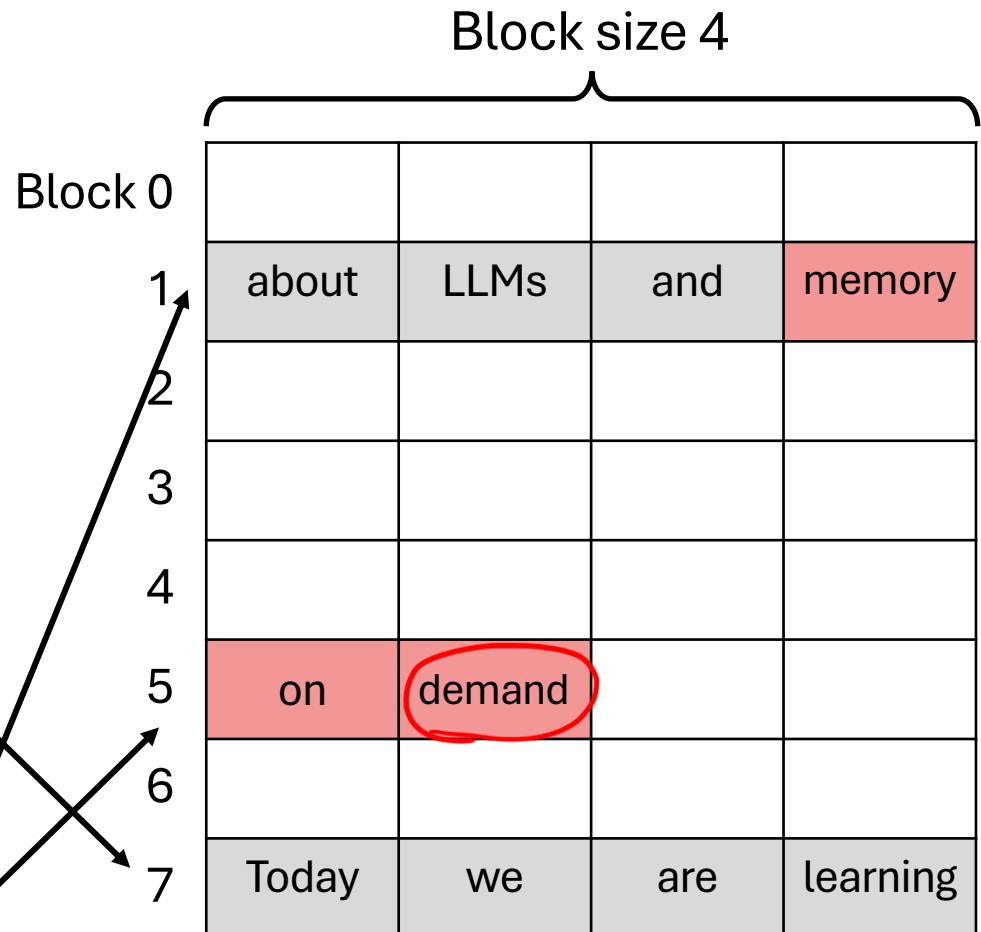
Completion: “*memory on demand*”

Block 0	Today	we	are	learning
1	about	LLMs	and	memory
2	on	demand		
3				

Logical KV Blocks

Phys. Block	# Filled
7	4
1	4
5	2

Block Table



Content credits: <https://youtu.be/yVxtLTcdO1Q?si=XO2Dk-VYOShUMH1u>

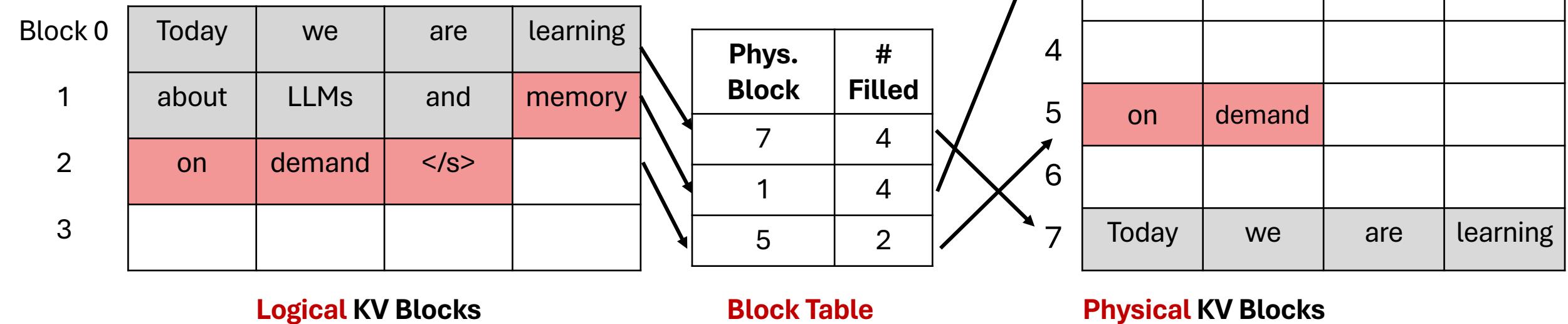


LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

Physical vs Logical KV Blocks



Prompt A: “Today we are learning about LLMs and”

Completion: “**memory on demand </s>**”

Content credits: <https://youtu.be/yVxtLTcdO1Q?si=XO2Dk-VYOSHUMH1u>



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

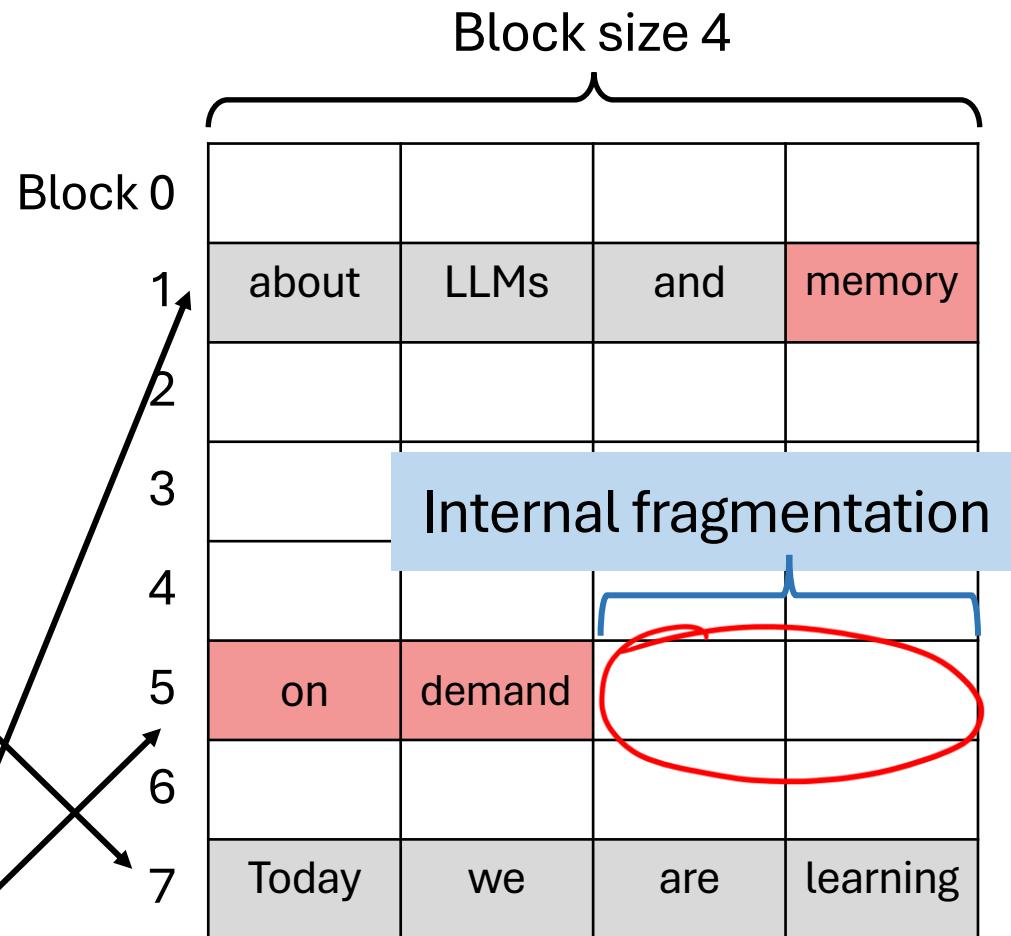
Physical vs Logical KV Blocks

Block 0	Today	we	are	learning
1	about	LLMs	and	memory
2	on	demand	</s>	
3				

Logical KV Blocks

Phys. Block	# Filled
7	4
1	4
5	2

Block Table



Prompt A: "Today we are learning about LLMs and"

Completion: "**memory on demand </s>**"

Content credits: <https://youtu.be/yVxtLTcdO1Q?si=XO2Dk-VYOSHUMH1u>



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

0	Today	we	are	learning
1	about	LLMs	and	
2				
3				

Logical KV Blocks - B

0	Today	we	are	learning
1	about	LLMs	and	memory
2	on	demand	</s>	
3				

Logical KV Blocks - A

Phys. Block	# Filled
7	4
1	4
5	2

Block Table -A

Block size 4

0			
1	about	LLMs	and
2			memory
3			
4			
5	on	demand	
6			
7	Today	we	are learning

Physical KV Blocks

Prompt A: "Today we are learning about LLMs and"
Completion: "memory on demand</s>"

Prompt B: "Today we are learning about LLMs and"
Completion:

Content credits: <https://youtu.be/yVxtLTcdO1Q?si=XO2Dk-VYOSHUMH1u>



LLMs: Introduction and Recent Advances



Tanmoy Chakraborty

0	Today	we	are	learning
1	about	LLMs	and	
2				
3				

Logical KV Blocks - B

Phys. Block	# Filled

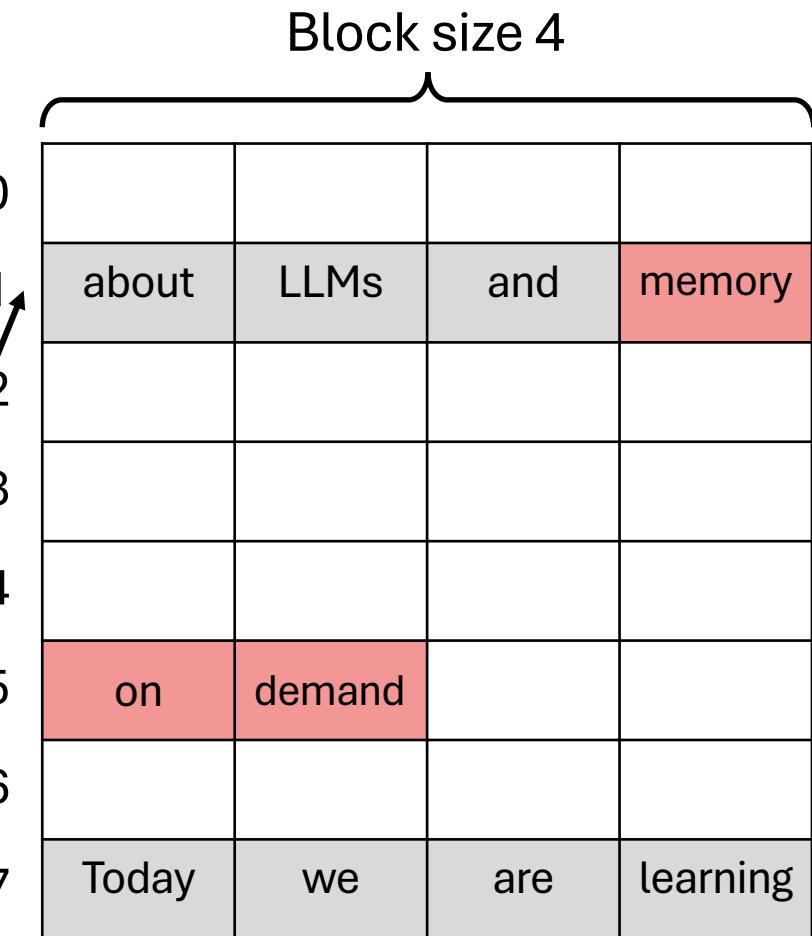
Block Table - B

0	Today	we	are	learning
1	about	LLMs	and	memory
2	on	demand	</s>	
3				

Logical KV Blocks - A

Phys. Block	# Filled
7	4
1	3
5	2

Block Table - A



Physical KV Blocks

Prompt A: “Today we are learning about LLMs and”
Completion: “**memory on demand </s>**”

Prompt B: “**Today we are learning about LLMs and**”
Completion:



0	Today	we	are	learning
1	about	LLMs	and	
2				
3				

Logical KV Blocks - B

Phys. Block	# Filled
3	4
6	3

Block Table - B

0	Today	we	are	learning
1	about	LLMs	and	memory
2	on	demand	</s>	
3				

Logical KV Blocks - A

Phys. Block	# Filled
7	4
1	3
5	2

Block Table - A

Block size 4			
about	LLMs	and	memory
Today	we	are	learning
on	demand		
about	LLMs	and	
Today	we	are	learning

Physical KV Blocks

Prompt A: "Today we are learning about LLMs and"
Completion: "memory on demand </s>"

Prompt B: "Today we are learning about LLMs and"
Completion:



0	Today	we	are	learning
1	about	LLMs	and	memory
2	management	</s>		
3				

Logical KV Blocks - B

Phys. Block	# Filled
3	4
6	4
2	1

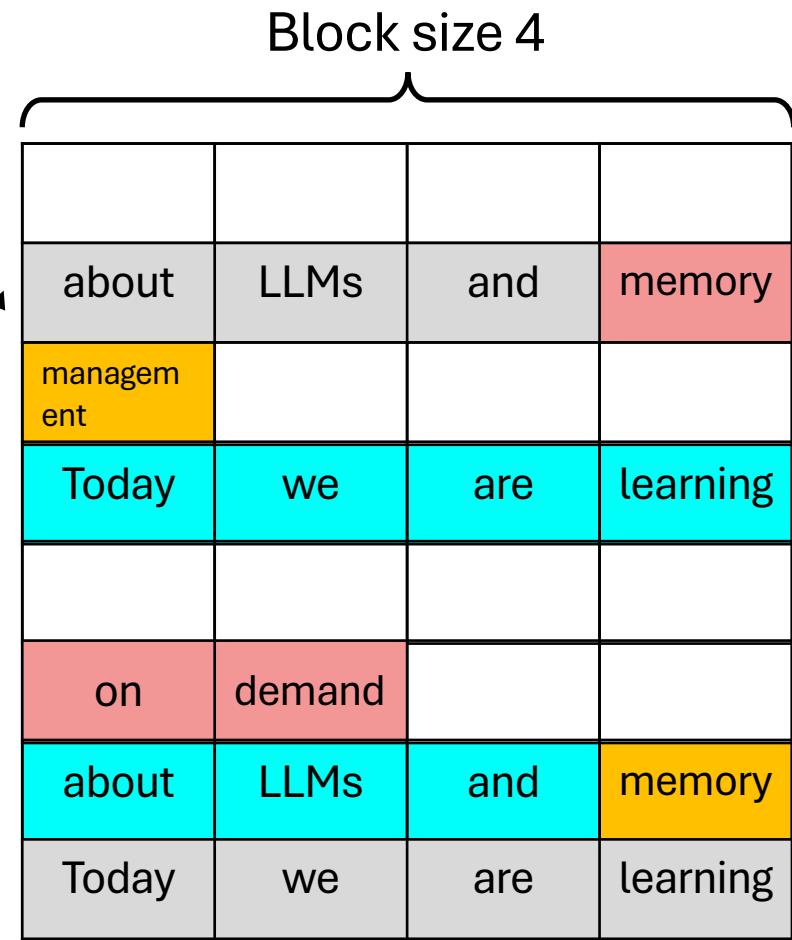
Block Table - B

0	Today	we	are	learning
1	about	LLMs	and	memory
2	on	demand	</s>	
3				

Logical KV Blocks - A

Phys. Block	# Filled
7	4
1	3
5	2

Block Table - A



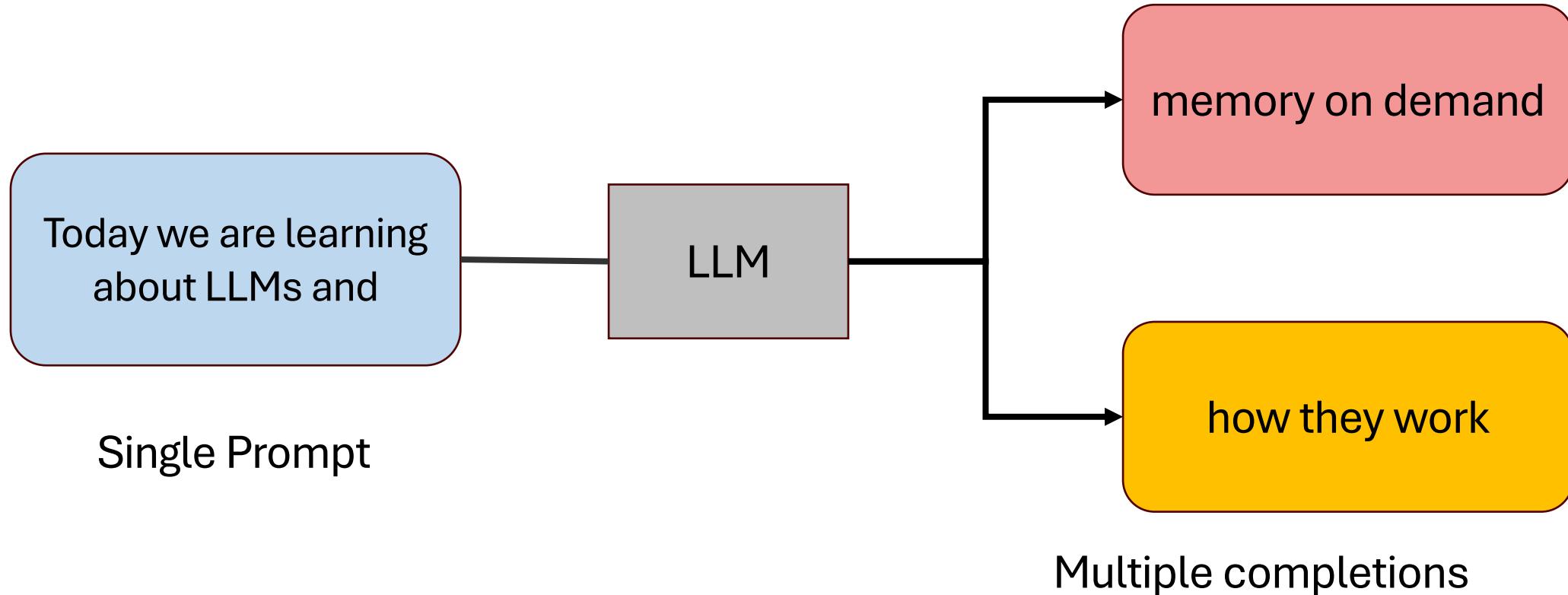
Physical KV Blocks

Prompt A: "Today we are learning about LLMs and"
Completion: "memory on demand </s>"

Prompt B: "Today we are learning about LLMs and"
Completion: "memory management </s>"



Dynamic block mapping enables sharing



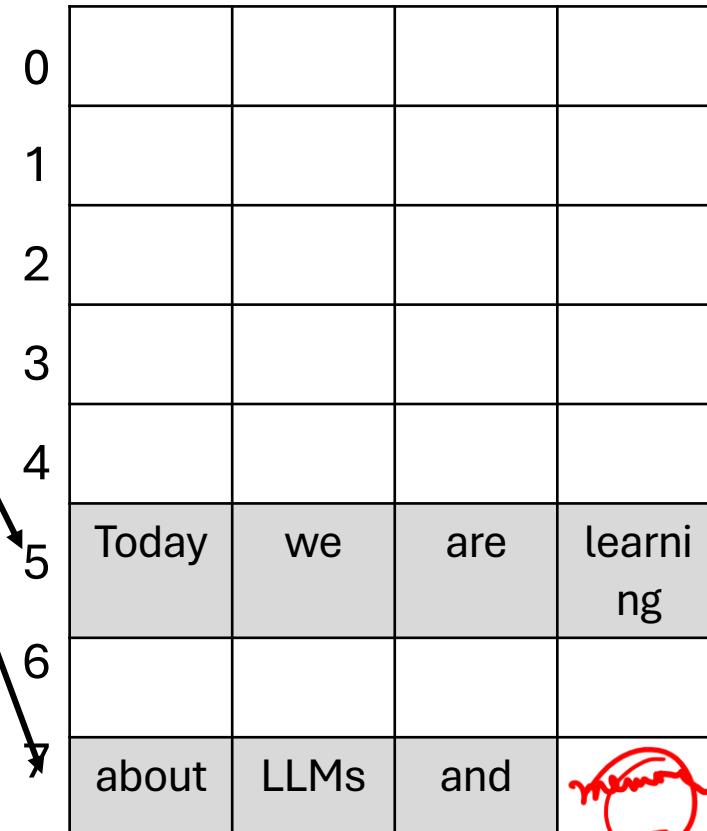
Sharing KV blocks in parallel sampling

B.T. for Kt#

Phys. Block	# Filled
5	4
7	3

Today	we	are	learning
about	LLMs	and	<i>memory</i>

Logical KV Blocks - A



Physical KV Blocks

P.T. for Kt#

Phys. Block	# Filled
7 5	4
5 7	3

Today	we	are	learning
about	LLMs	and	<i>how</i>

Logical KV Blocks - B

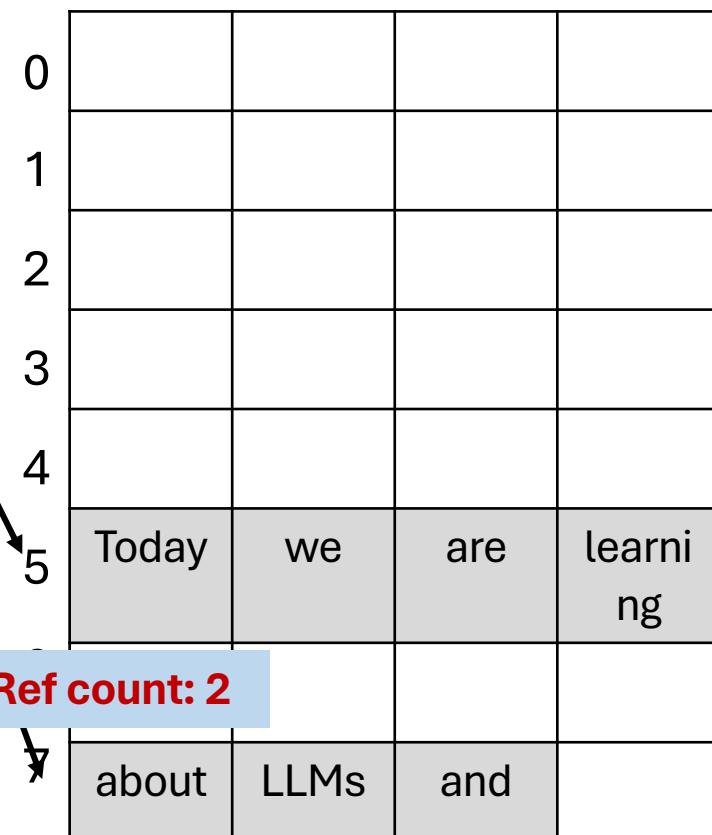


Sharing KV blocks in parallel sampling

Phys. Block	# Filled
5	4
7	3

Today	we	are	learning
about	LLMs	and	

Logical KV Blocks - A



Phys. Block	# Filled
7	4
5	3

Today	we	are	learning
about	LLMs	and	

Logical KV Blocks - B

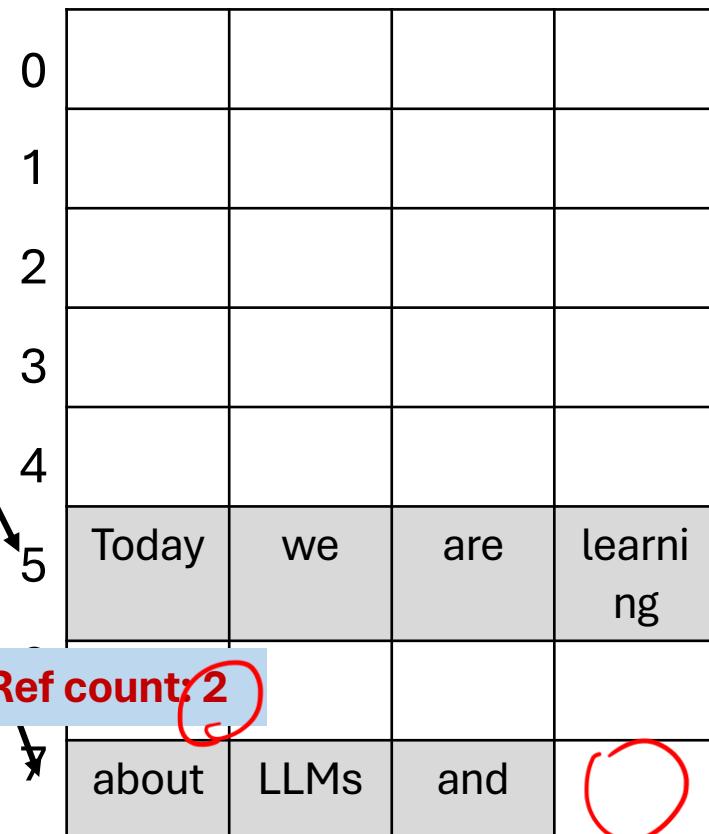


Sharing KV blocks in parallel sampling

Phys. Block	# Filled
5	4
7	3

Today	we	are	learning
about	LLMs	and	memory

Logical KV Blocks - A



Phys. Block	# Filled
7	4
5	3

Today	we	are	learning
about	LLMs	and	how

Logical KV Blocks - B

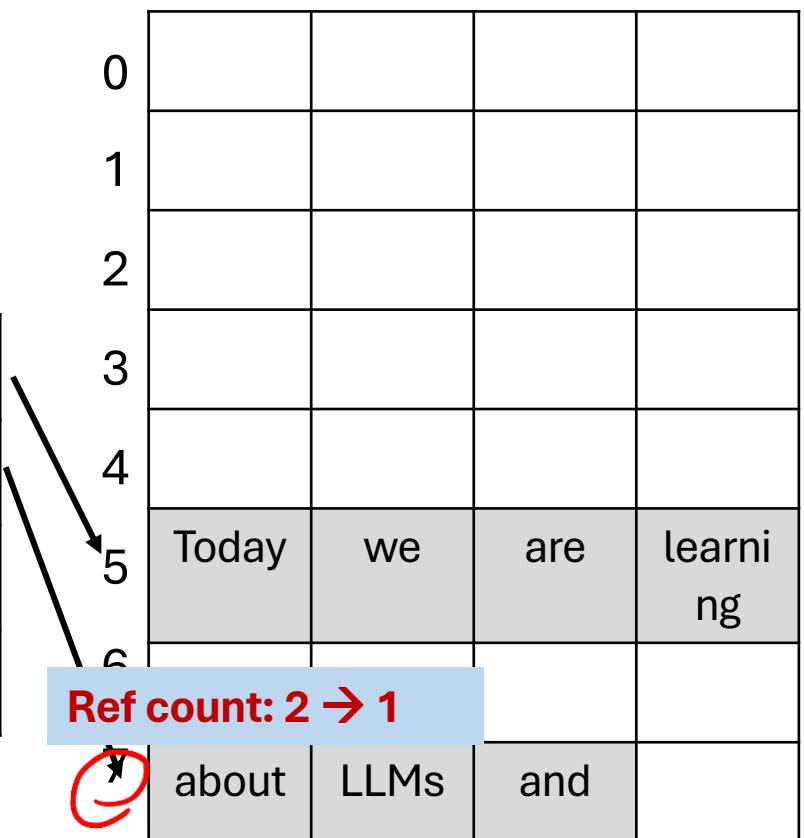


Sharing KV blocks in parallel sampling

Phys. Block	# Filled
5	4
7	3

Today	we	are	learning
about	LLMs	and	memory

Logical KV Blocks - A



Phys. Block	# Filled
7	4
5	3

Today	we	are	learning
about	LLMs	and	how

Logical KV Blocks - B

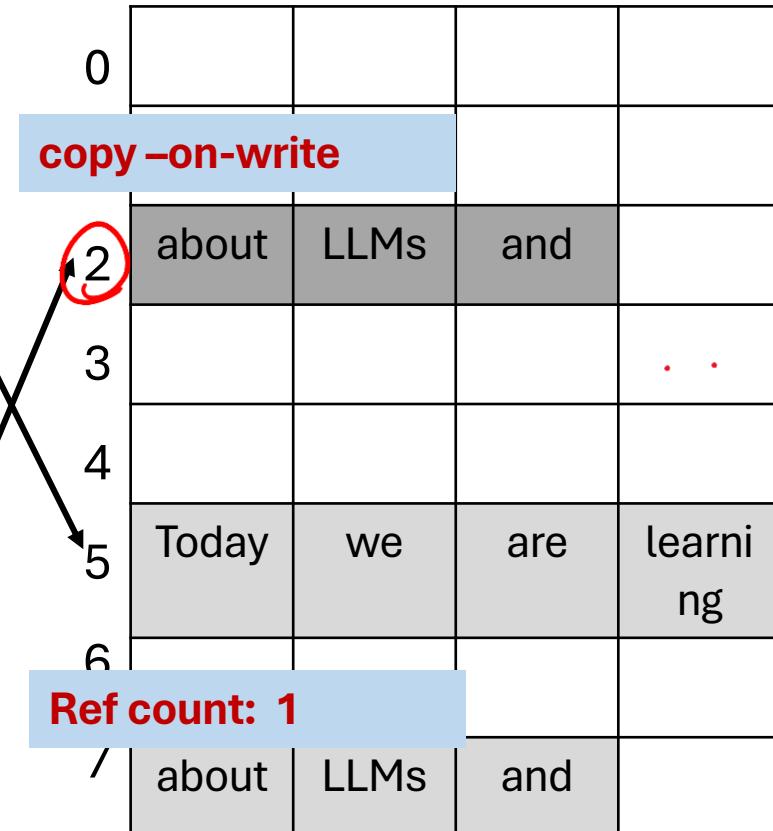


Sharing KV blocks in parallel sampling

Phys. Block	# Filled
5	4
7	3

Today	we	are	learning
about	LLMs	and	memory

Logical KV Blocks - A

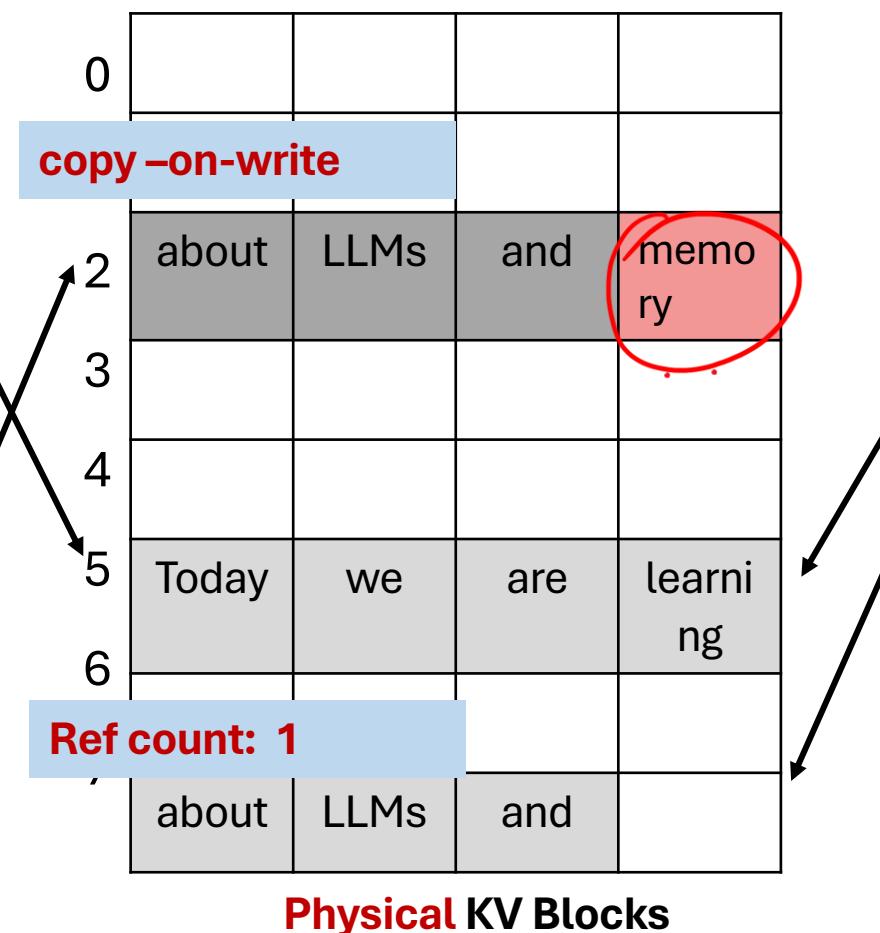


Sharing KV blocks in parallel sampling

Phys. Block	# Filled
5	4
7	3

Today	we	are	learning
about	LLMs	and	memory

Logical KV Blocks - A



Phys. Block	# Filled
7	4
5	3

Today	we	are	learning
about	LLMs	and	how

Logical KV Blocks - B

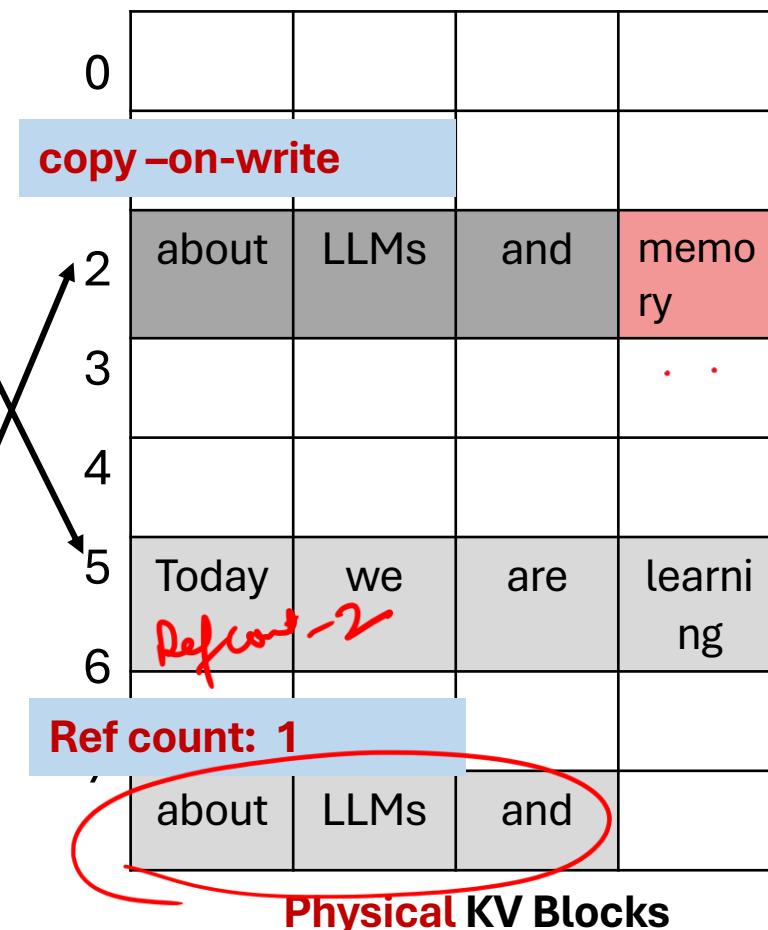


Sharing KV blocks in parallel sampling

Phys. Block	# Filled
5	4
2	4

Today	we	are	learning
about	LLMs	and	memory

Logical KV Blocks - A

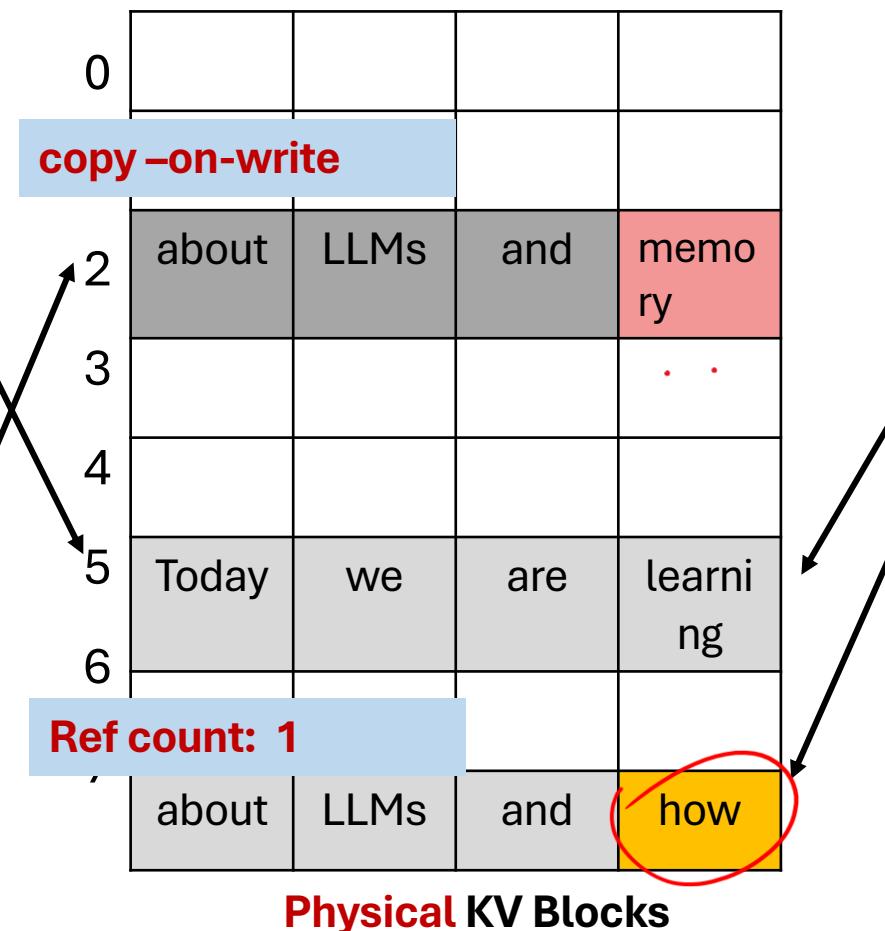


Sharing KV blocks in parallel sampling

Phys. Block	# Filled
5	4
2	4

Today	we	are	learning
about	LLMs	and	memory

Logical KV Blocks - A



Phys. Block	# Filled
7	4
5	3

Today	we	are	learning
about	LLMs	and	how

Logical KV Blocks - B

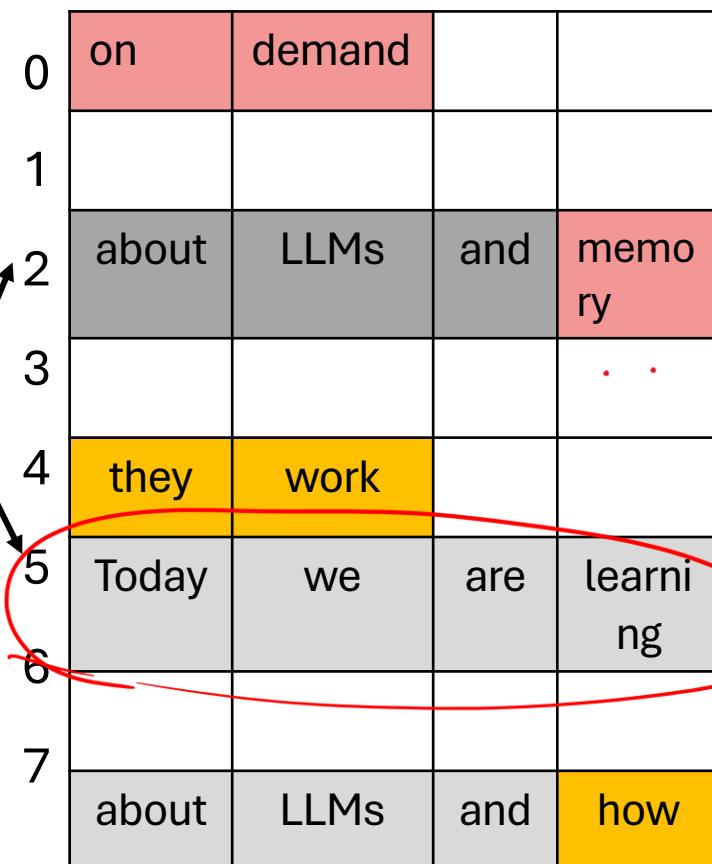


Sharing KV blocks in parallel sampling

Phys. Block	# Filled
5	4
2	4
0	2

Today	we	are	learning
about	LLMs	and	memory
on	demand		

Logical KV Blocks - A



Physical KV Blocks

Phys. Block	# Filled
5	4
2	4
4	2

Today	we	are	learning
about	LLMs	and	how
they	work		

Logical KV Blocks - B

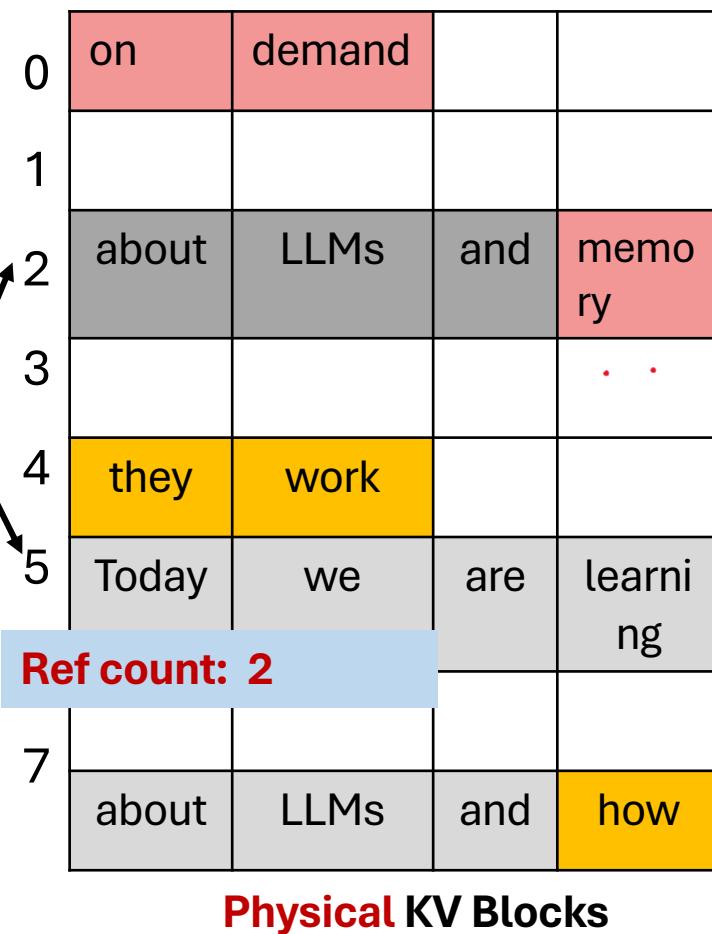


Sharing KV blocks in parallel sampling

Phys. Block	# Filled
5	4
2	4
0	2

Today	we	are	learning
about	LLMs	and	memory
on	demand		

Logical KV Blocks - A

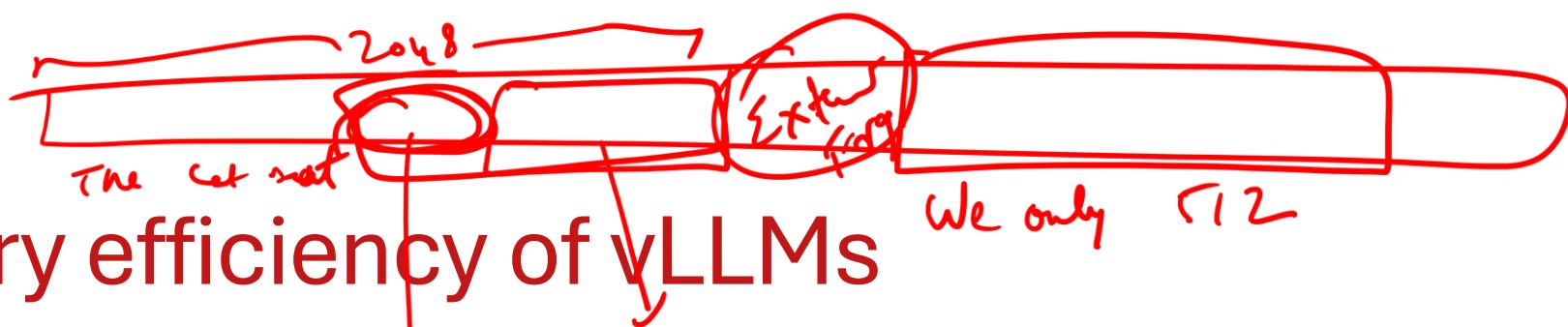


Phys. Block	# Filled
7	4
5	4
4	2

Today	we	are	learning
about	LLMs	and	how
they	work		

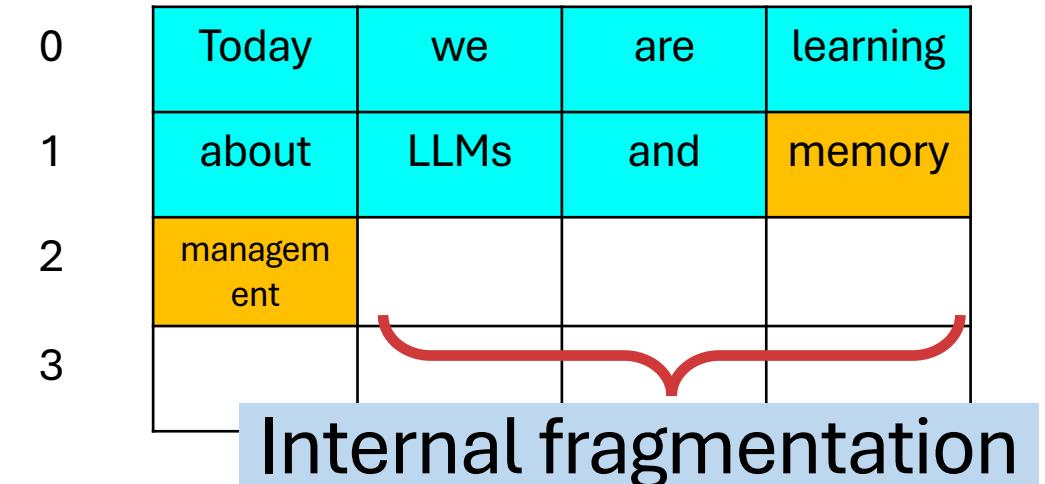
Logical KV Blocks - B





Memory efficiency of vLLMs

- ✓ Minimal internal fragmentation
 - Only happens at the last block of a sequence
 - **# wasted tokens / seq < block size**
 - Sequence: O(100) or O(1000) tokens
 - Block size: 16 or 32 tokens



- ✓ No external fragmentation
- ✓ On average, wasted space < **4%** of KV cache
- ✓ **3-5x** improved memory utilization!

Content credits: https://www.youtube.com/watch?v=5ZlavKF_98U&t=1646s&ab_channel=Anyscale

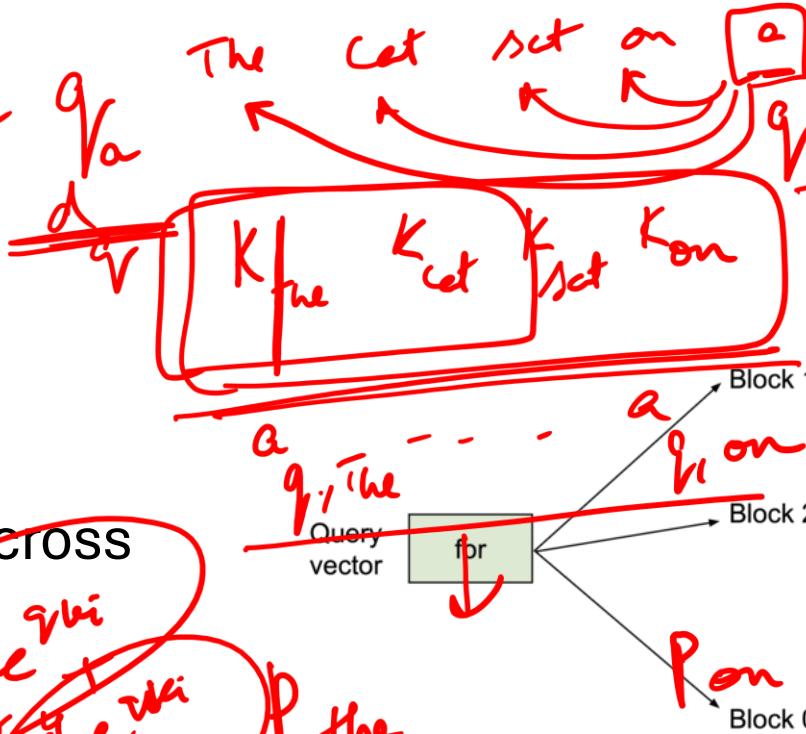


$$\sum_{i=1}^4 e^{q_k i} = \sum_{i=1}^4 e^{q_k r_i}$$

$$W_q a = q_a$$

Paged Attention

- Tensor operations require contiguous memory
- How to compute attention softmax across fragmented memory?
- Paged Attention!



QK^T

Key and value vectors

computer	scientist	and	mathematician
renowned	for		
The	cat		
Alan	Turing	is	a

$$\sum_{i=1}^4 e^{q_k i} = \sum_{i=1}^4 e^{q_k r_i}$$

$\sum_{i=1}^4 e^{q_k i}$

$$\sum_{i \in \text{the cat}} p_i (j_i)$$

$$\alpha = \sum_{i=1}^4 \beta_i R_i$$

$$\text{softmax}([A_1, A_2]) = [\alpha \text{softmax}(A_1), \beta \text{softmax}(A_2)]$$

$$\text{softmax}([A_1, A_2]) \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \alpha \text{softmax}(A_1) * V_1 + \beta \text{softmax}(A_2) * V_2$$



How vLLM & Paged Attention results in efficient inference?

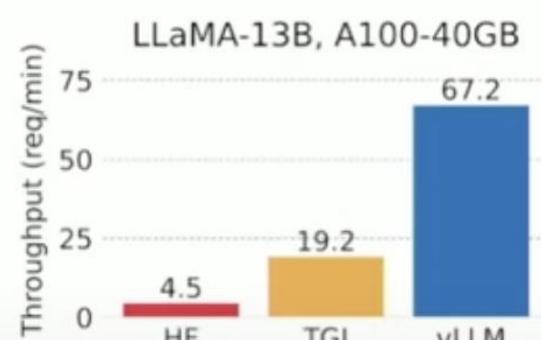
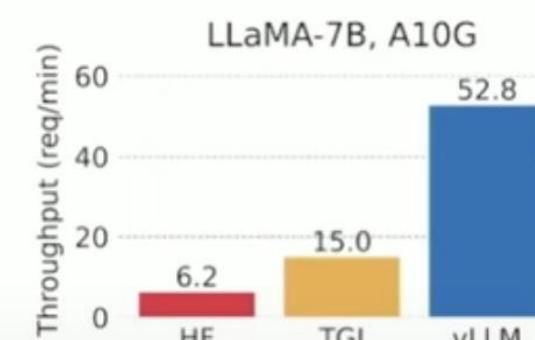
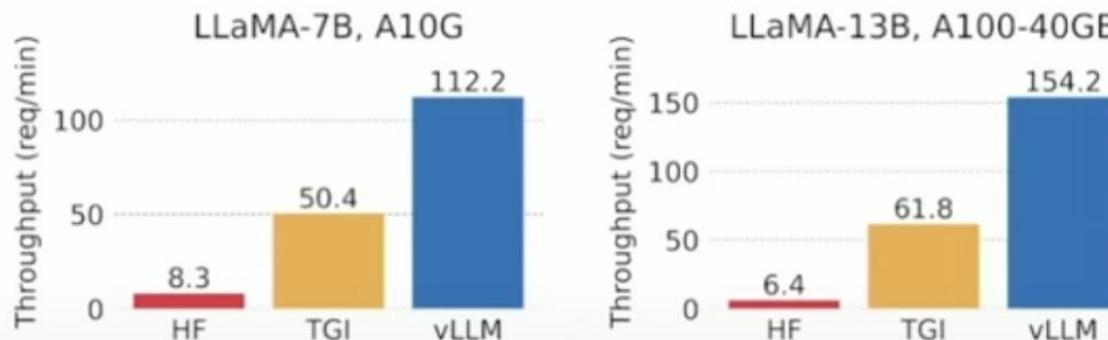
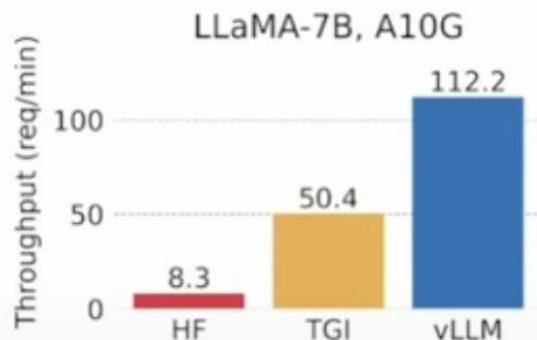
Reduce memory fragmentation with paging

Further reduce memory usage with sharing

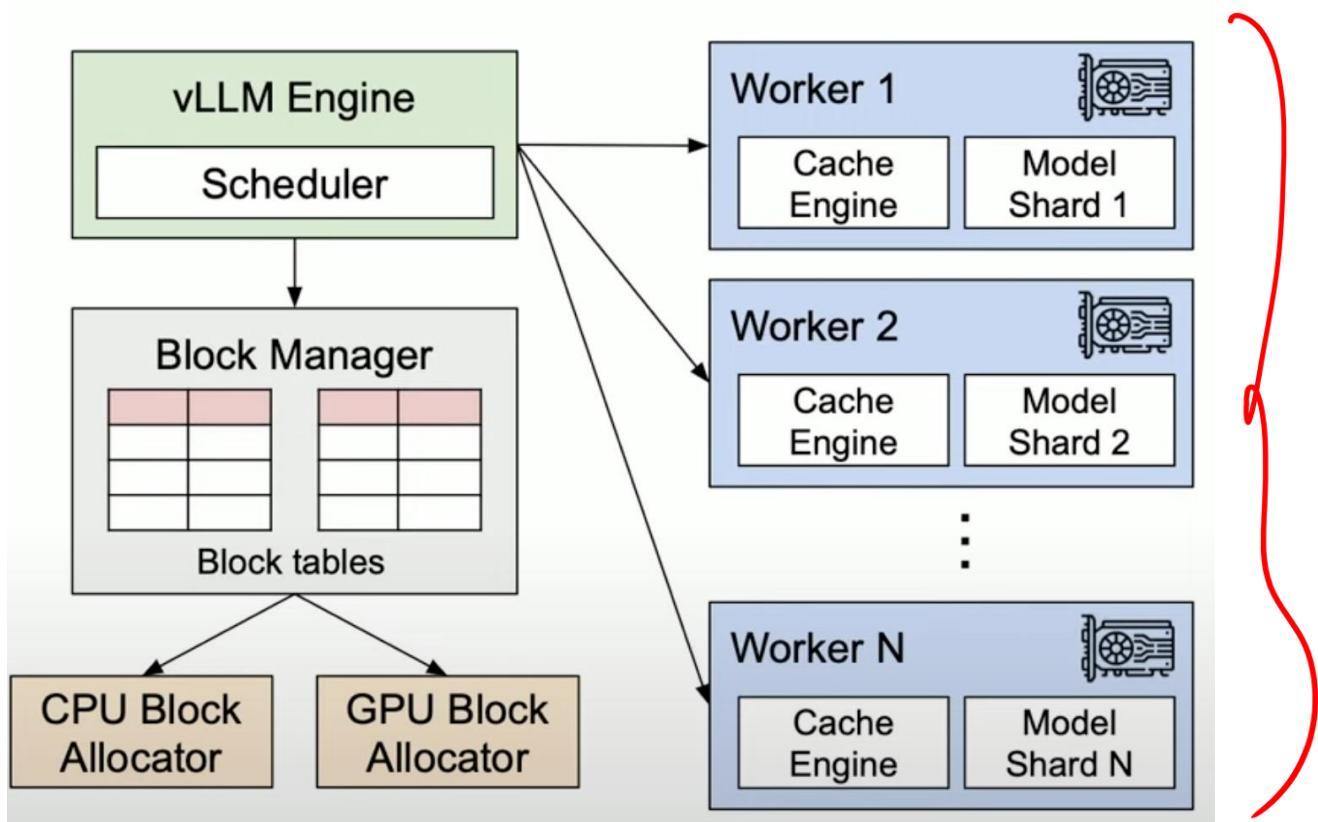


Comparison with HuggingFace and TGI (2023)

- Up to 24x higher throughput than HuggingFace (HF)
- Up to **3.5x** higher throughput than Text Generation Inference (TGI)



System Architecture and Implementation



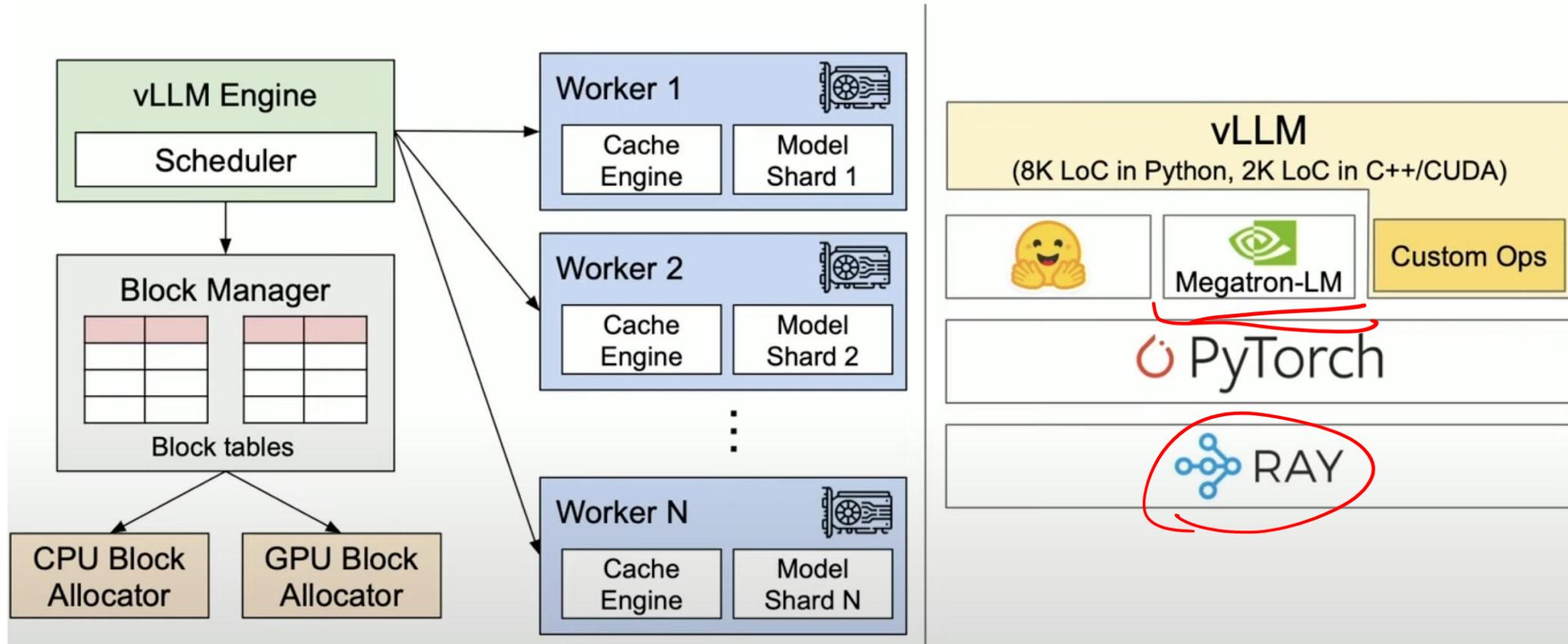
End to end llm serving engine

3 components –

- A frontend
- A distributed model executor
- A scheduler

A centralized engine that managers block table





Till now...

- **KV caching** – avoids re-computation of Keys and Value matrices
- **Paged Attention and vLLM** - efficient memory management
- Can we speed up attention computation?
- Flash Attention?

✓ Redundant computation

✓ GPU KV cache

✓ Efficient memory management

Speedy computation

Speculative decoding

