Spring25 CS598YP

24.2: Block Attention

Yongjoo Park

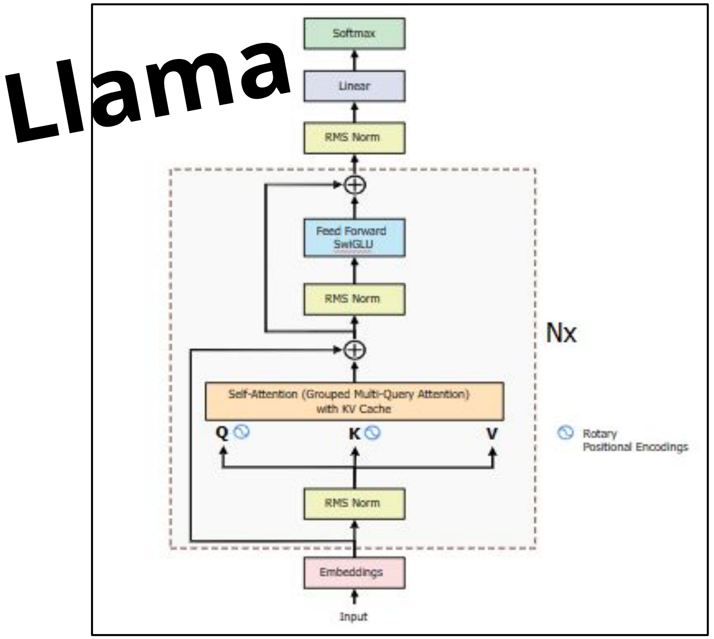
University of Illinois Urbana-Champaign

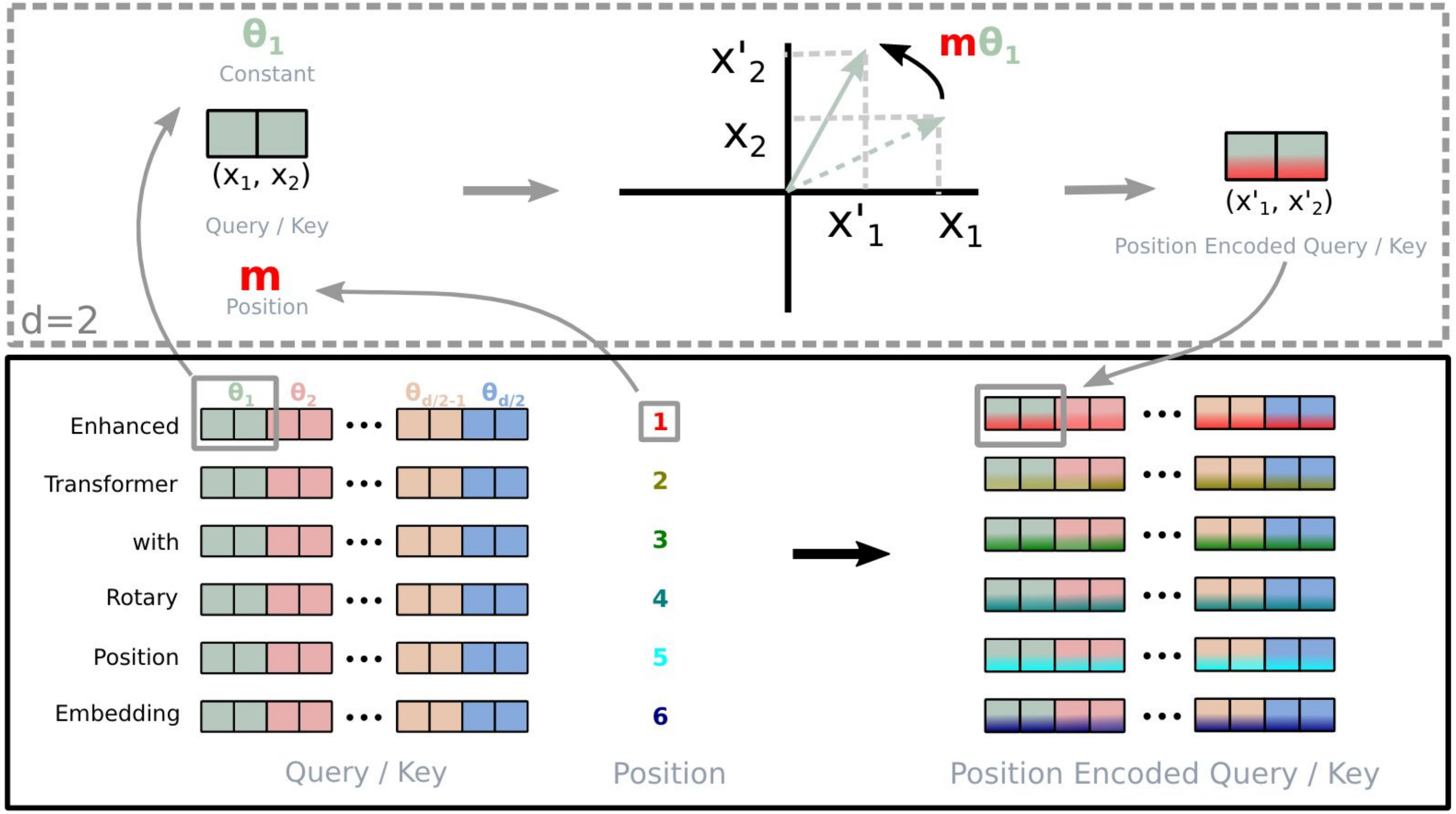
Outline

- Recap: **RoPE** (Rotary Position Embedding)
- **Block Attention:** fine-tuning and position re-coding
- Limitation and opportunities

RoPE

RoPE: Rotary position embedding





We perturb Q and K by **rotating** vectors (its angle *proportional to* the position *m*)

Block Attention (ICLR'25)

Block attention: cross attention only for query

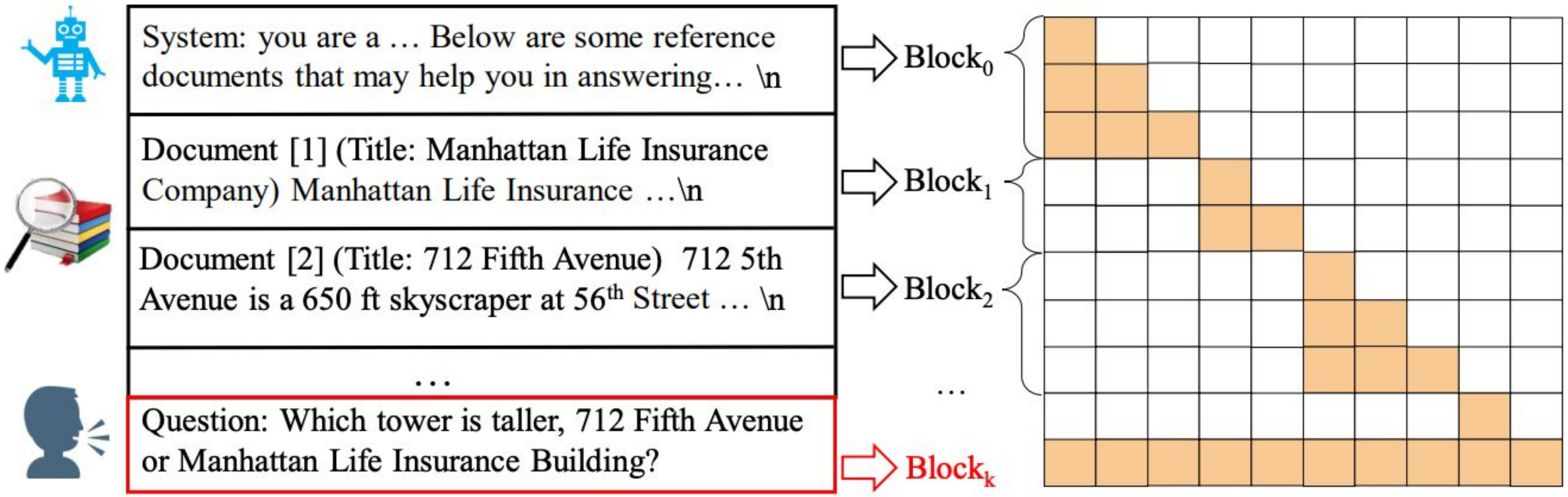
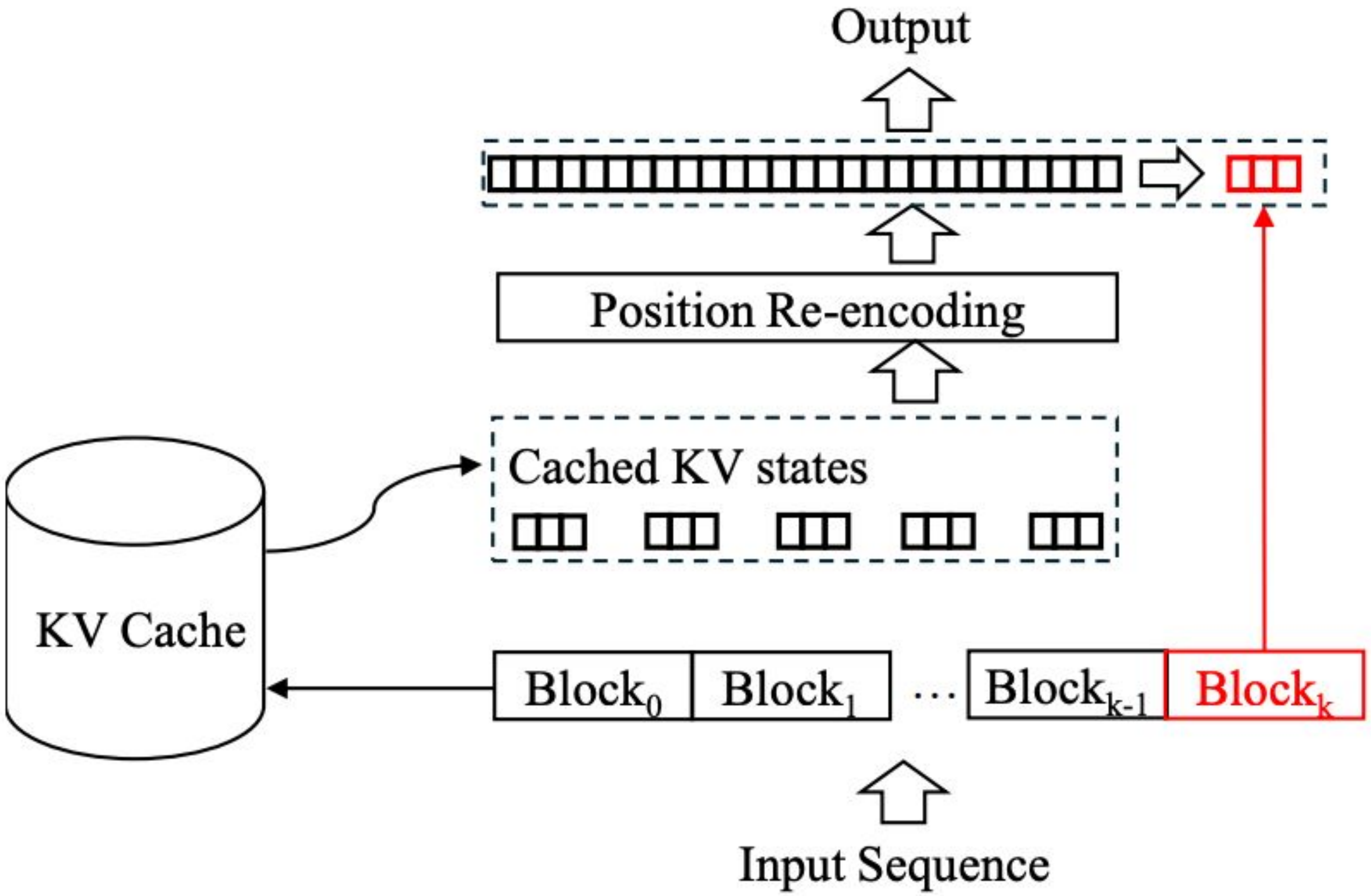


Figure 1: The Block-attention Masks

This allows re-using KV states of the ***Blocks***

Block attention: pipeline



What is the **expected problem**? (based on the lessons from CacheBlend)

Problem: No cross-attention, poor performance

	Models	2wiki	HQA	NQ	TQA
baseline instruct model	*Tulu3-SFT*	62.0	68.4	58.6	75.7
fine-tuned for RAG: **ideal** model	*Tulu3-RAG*	**73.2**	**74.8**	**61.5**	**75.8**
a previous work	*Tulu3-RAG-Superposition*	30.1	32.3	35.9	58.9
another previous work	*Tulu3-RAG-promptCache*	32.4	31.6	44.4	61.8

Block attention proposes ***fine-tuning***, instead of *re-computing* cross attention (as in CacheBlend)

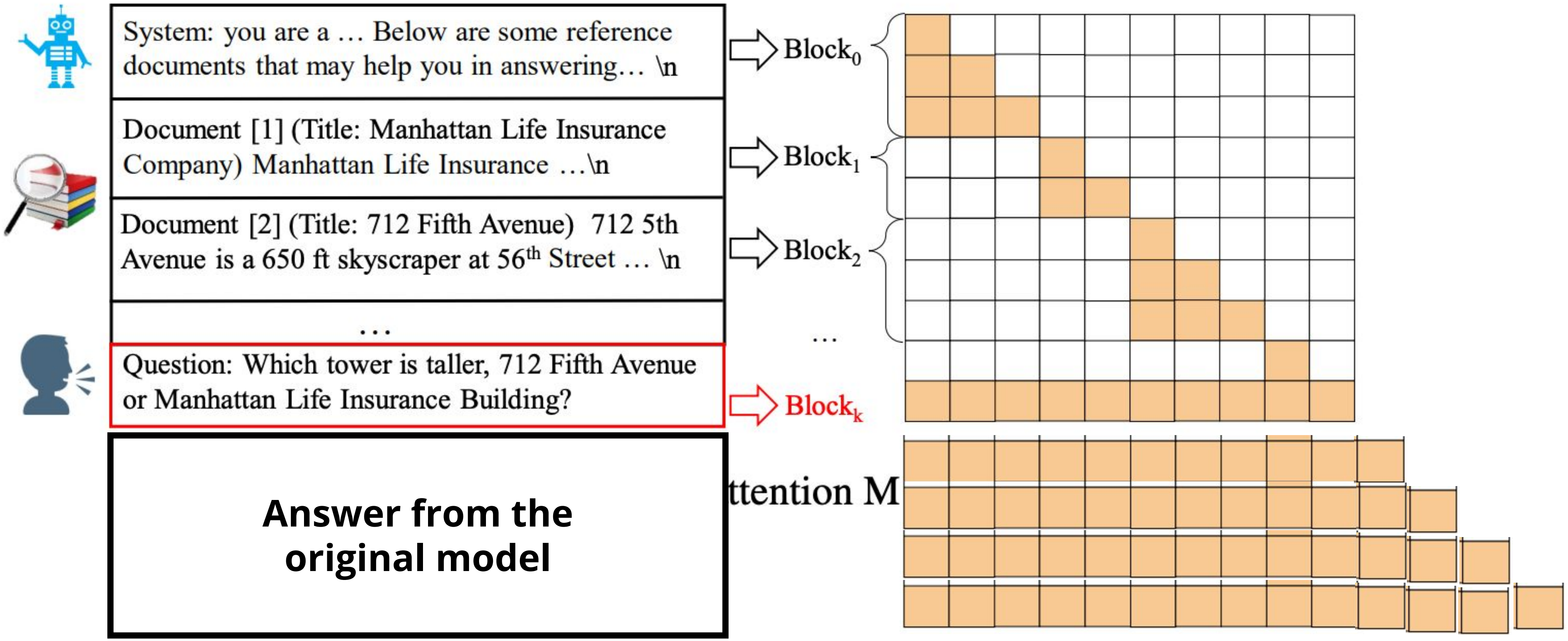
Fine-tuning and **position re-encoding** improves accuracy

	Models	2wiki	HQA	NQ	TQA
baseline instruct model	*Tulu3-SFT*	62.0	68.4	58.6	75.7
fine-tuned for RAG: **ideal** model	*Tulu3-RAG*	**73.2**	**74.8**	**61.5**	**75.8**
a previous work	*Tulu3-RAG-Superposition*	30.1	32.3	35.9	58.9
another previous work	*Tulu3-RAG-promptCache*	32.4	31.6	44.4	61.8
proposed model	*Tulu3-block-ft*	**72.2**	72.3	**60.4**	**75.1**
proposed model + cross-attention	*Tulu3-block-ft-full*	**73.6**	**75.2**	**62.2**	**76.2**
proposed model – **positional re-encoding**	*Tulu3-block-ft-w/o-pos*	68.9	69.9	59.2	74.4
proposed model – fine-tuning	*Tulu3-block-w/o-ft*	42.9	42.1	48.3	66.5

Table 1: Accuracy of different models on four RAG benchmarks.

Block attention proposes ***fine-tuning***, instead of *re-computing* cross attention (as in CacheBlend)

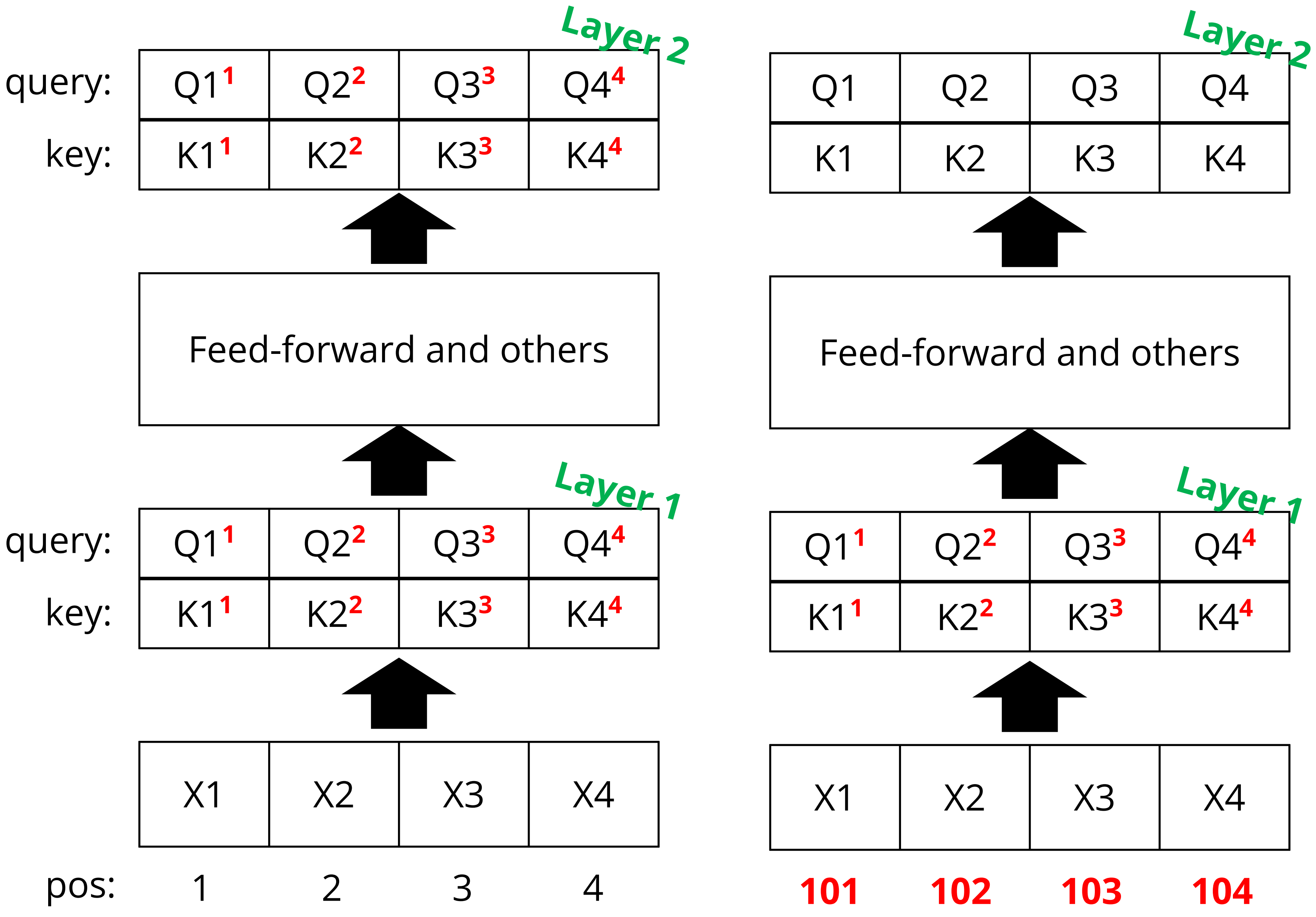
During fine-tuning, mask/erase cross-attention

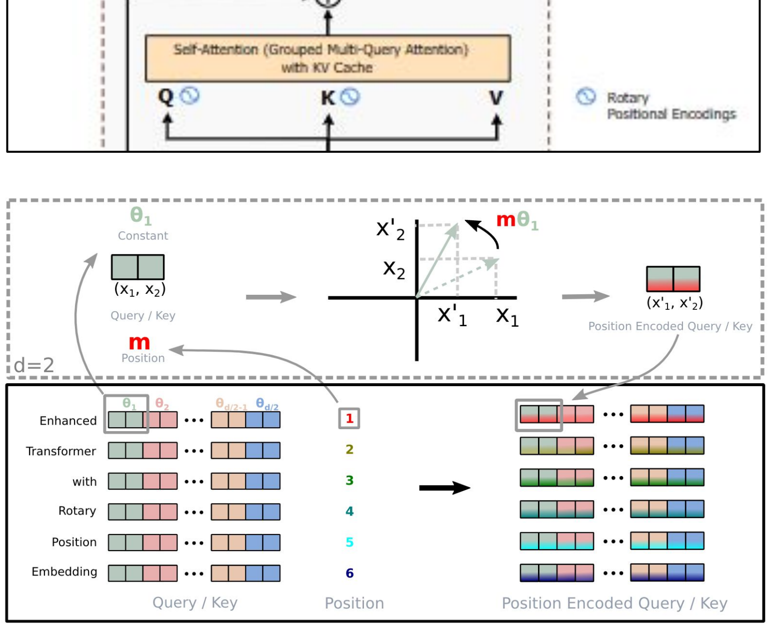


We let model the produce the same answer, *without cross attention*

When will it work? or not work?

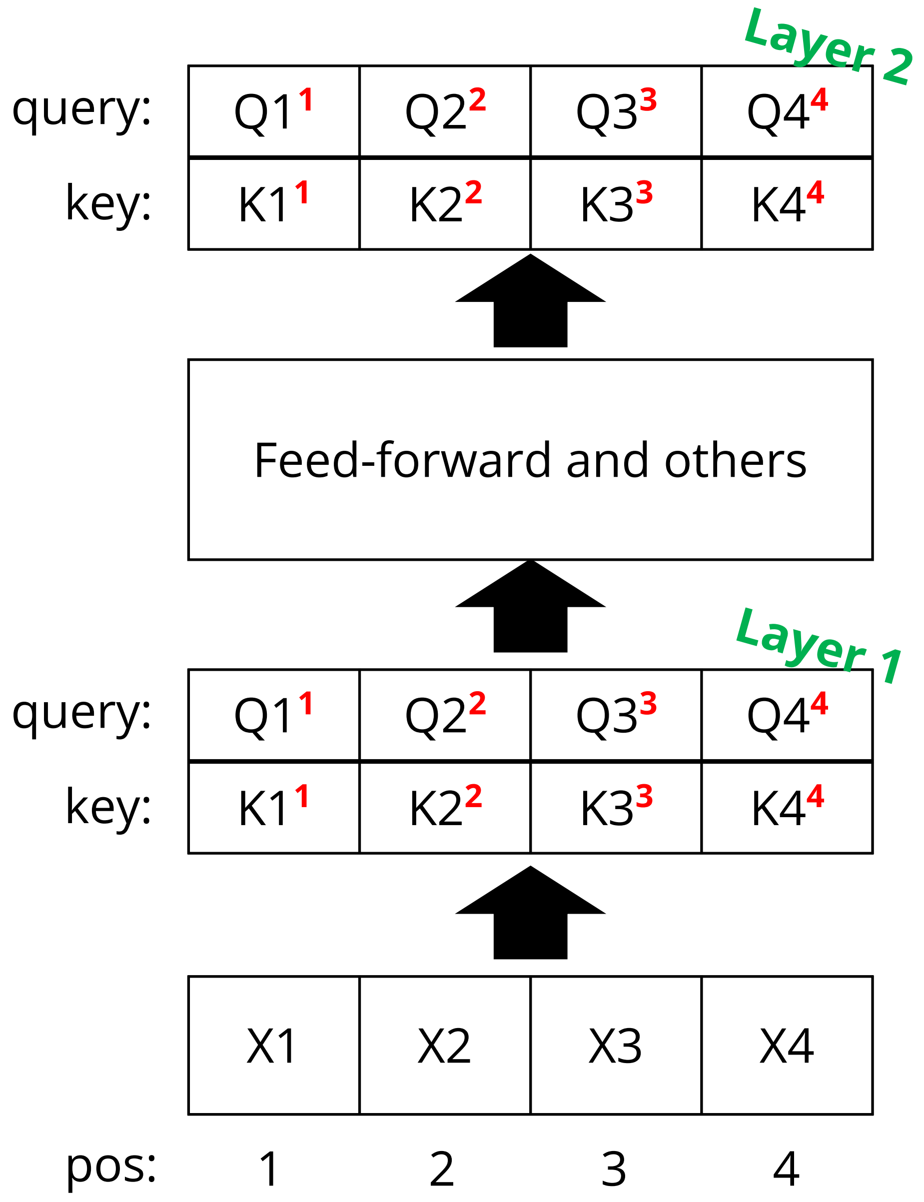
Position re-encoding

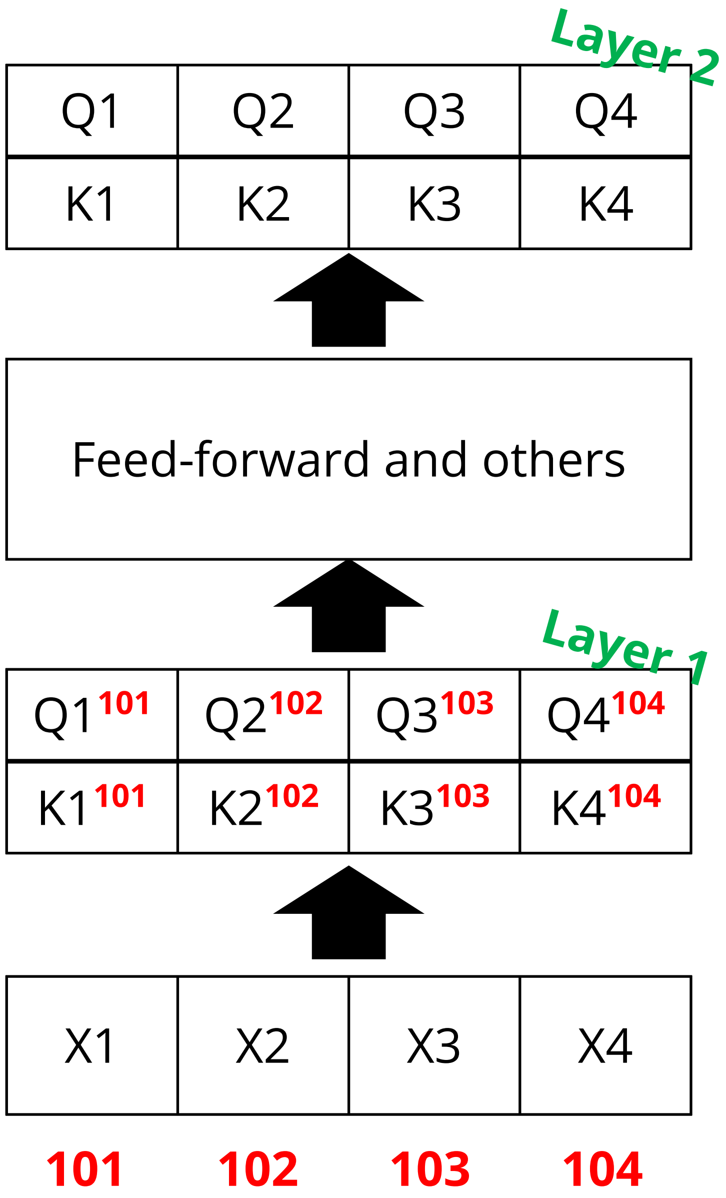


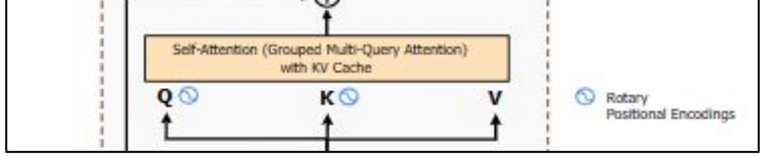


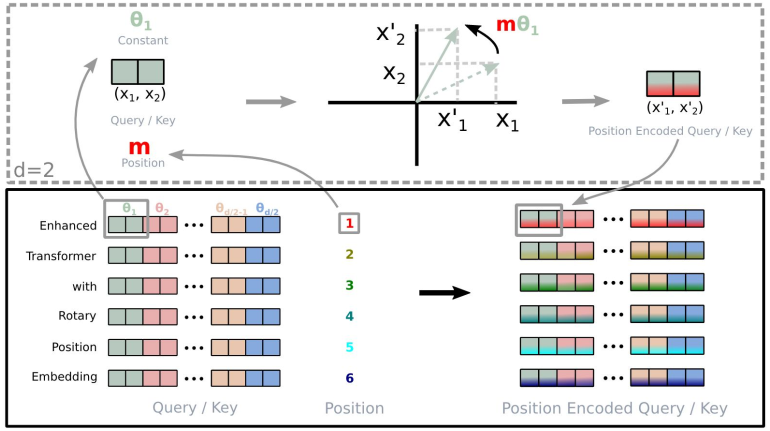
$$sim(K^{\theta_1}, Q^{\theta_2}) = f(K, Q, \boldsymbol{\theta_1} - \boldsymbol{\theta_2})$$

Position re-encoding



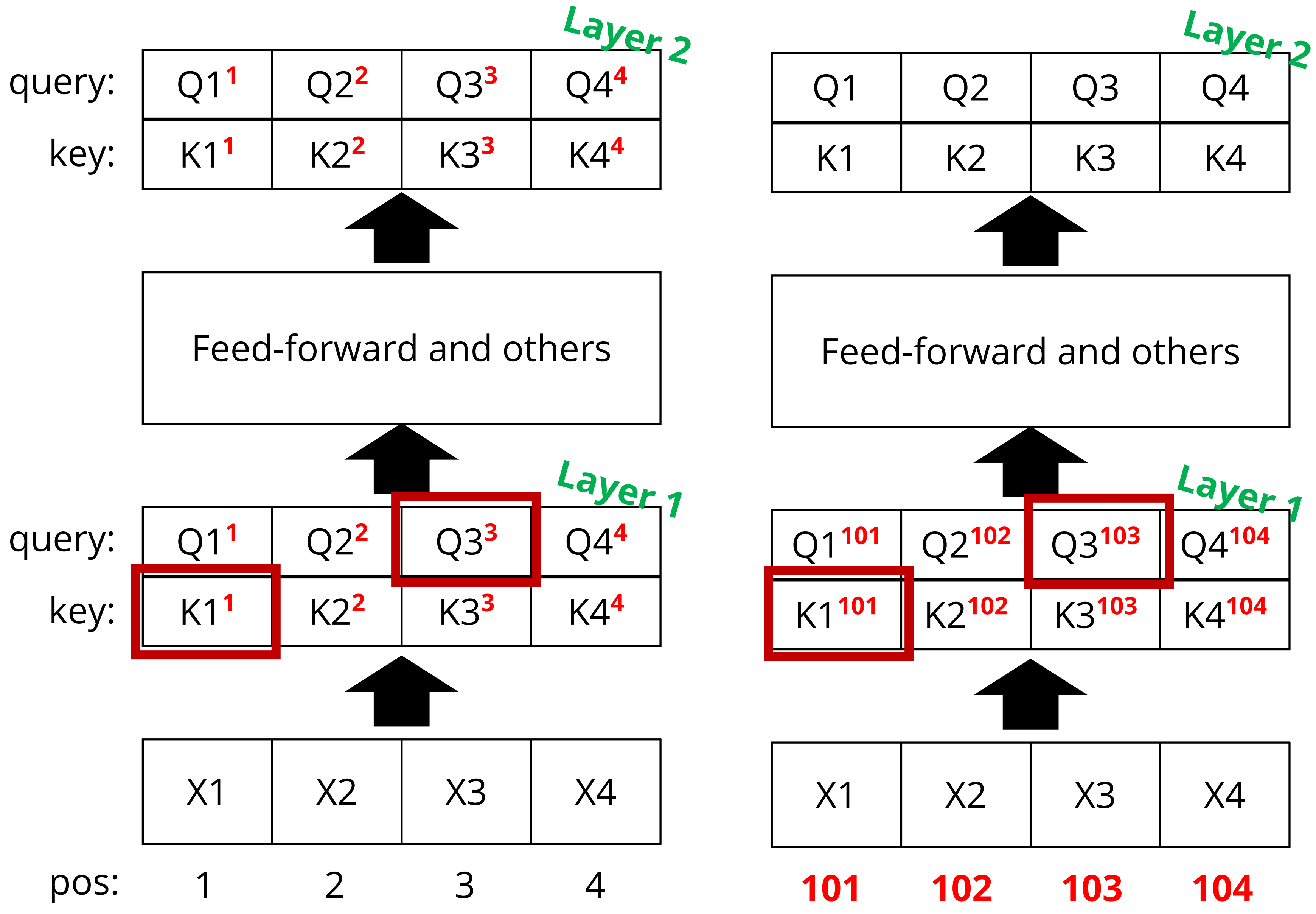


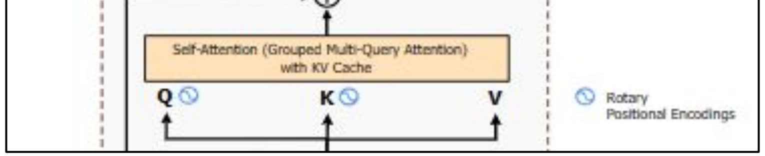


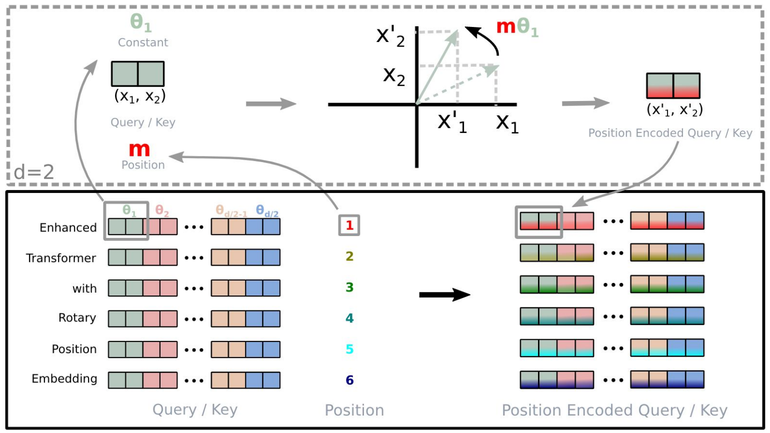


$$sim(K^{\theta_1}, Q^{\theta_2}) = f(K, Q, \boldsymbol{\theta_1} - \boldsymbol{\theta_2})$$

Position re-encoding

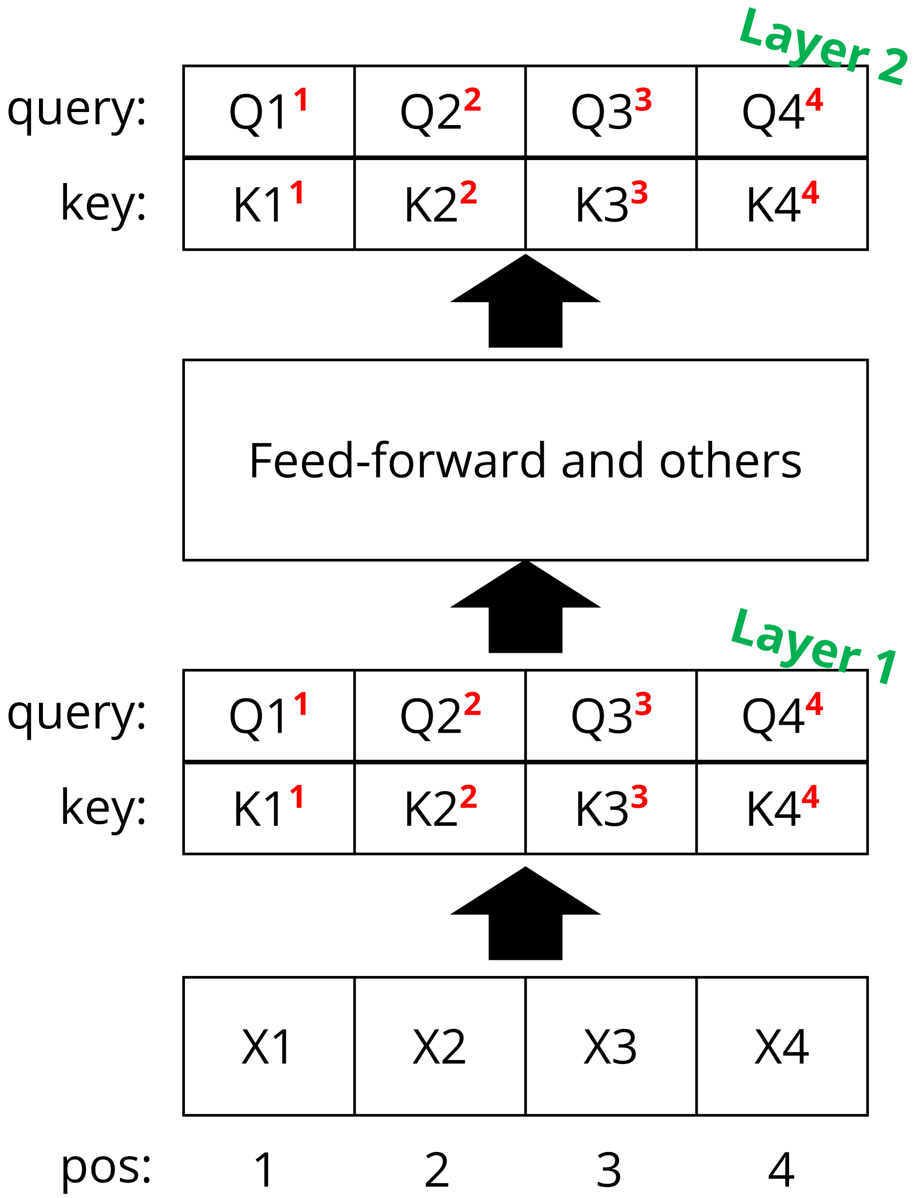


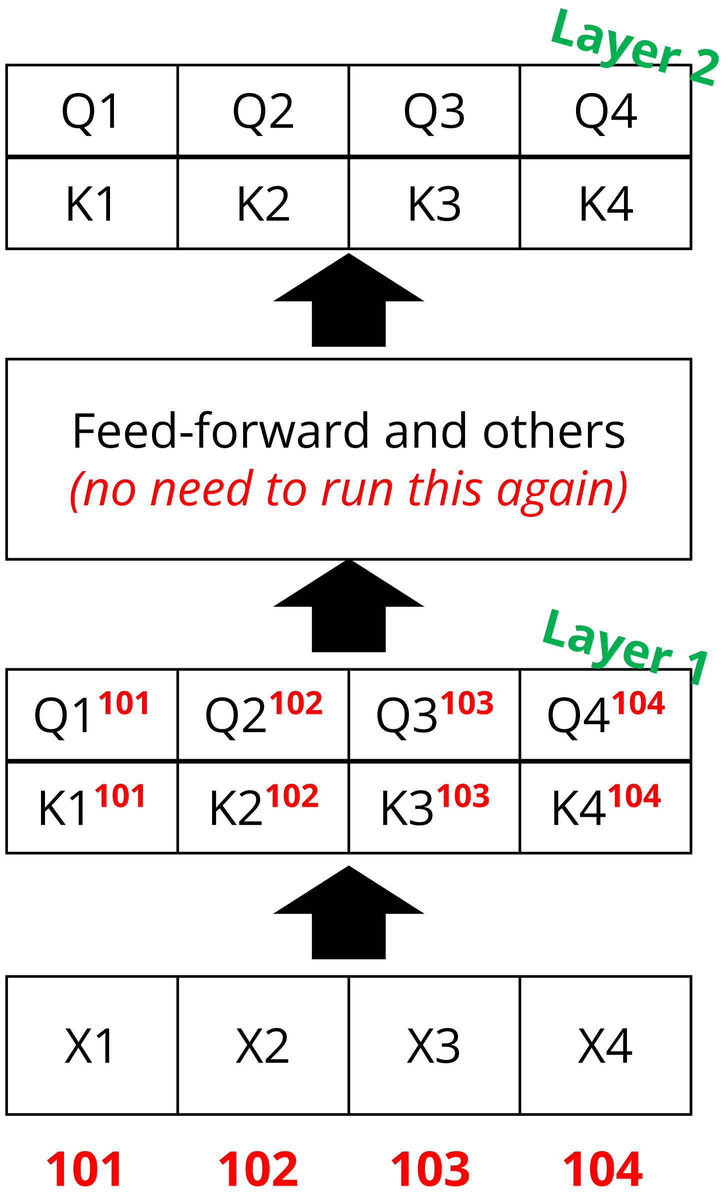


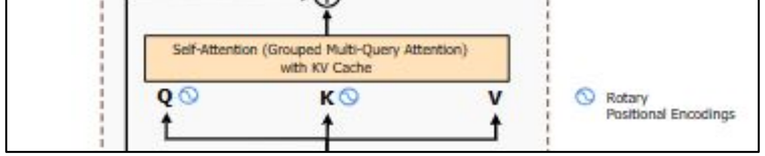


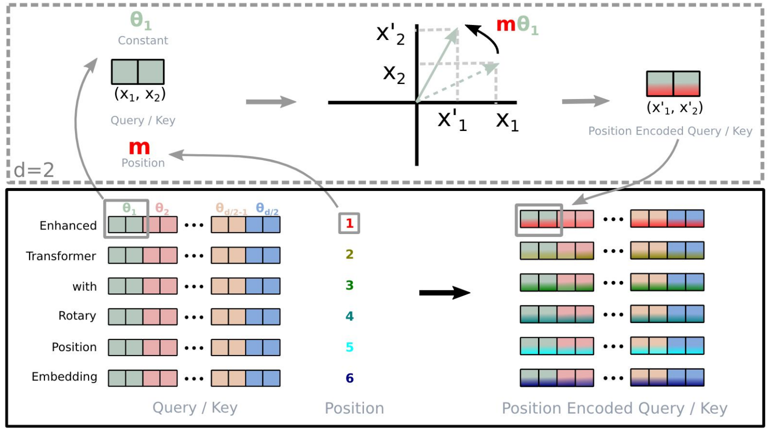
$$sim(K^{\theta_1}, Q^{\theta_2}) = f(K, Q, \boldsymbol{\theta_1} - \boldsymbol{\theta_2})$$

Position re-encoding





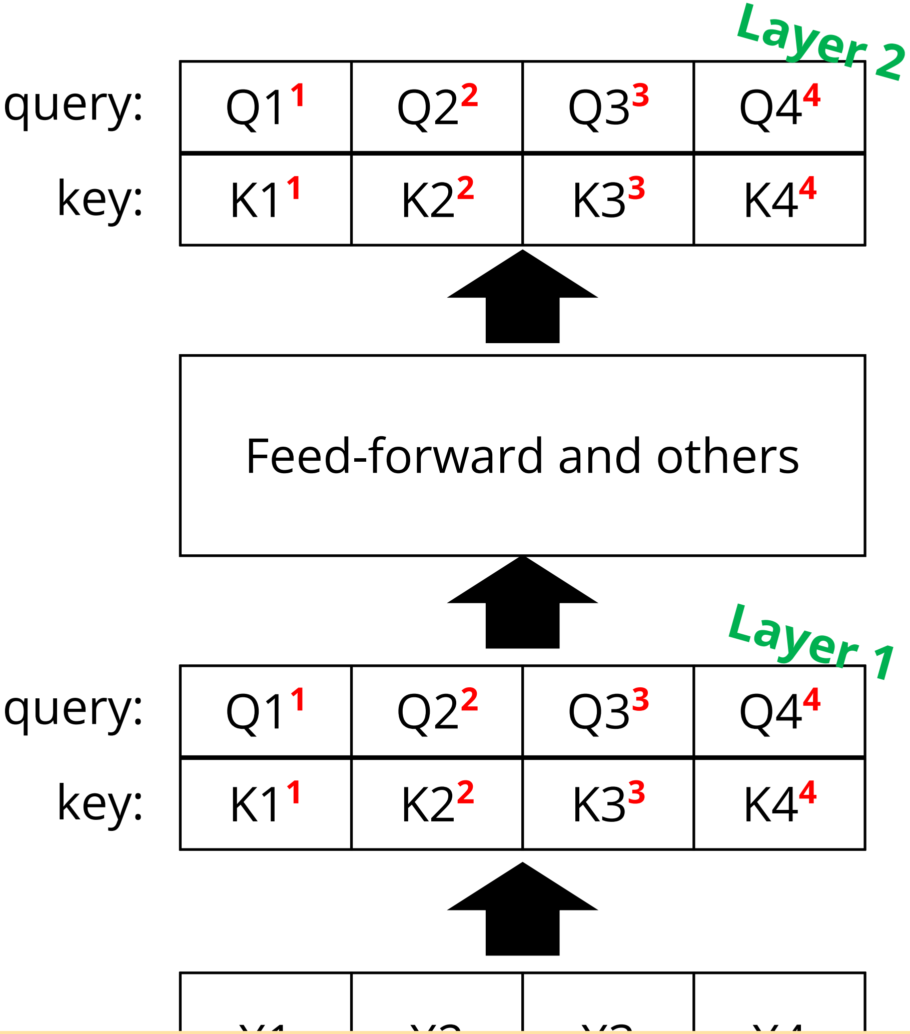


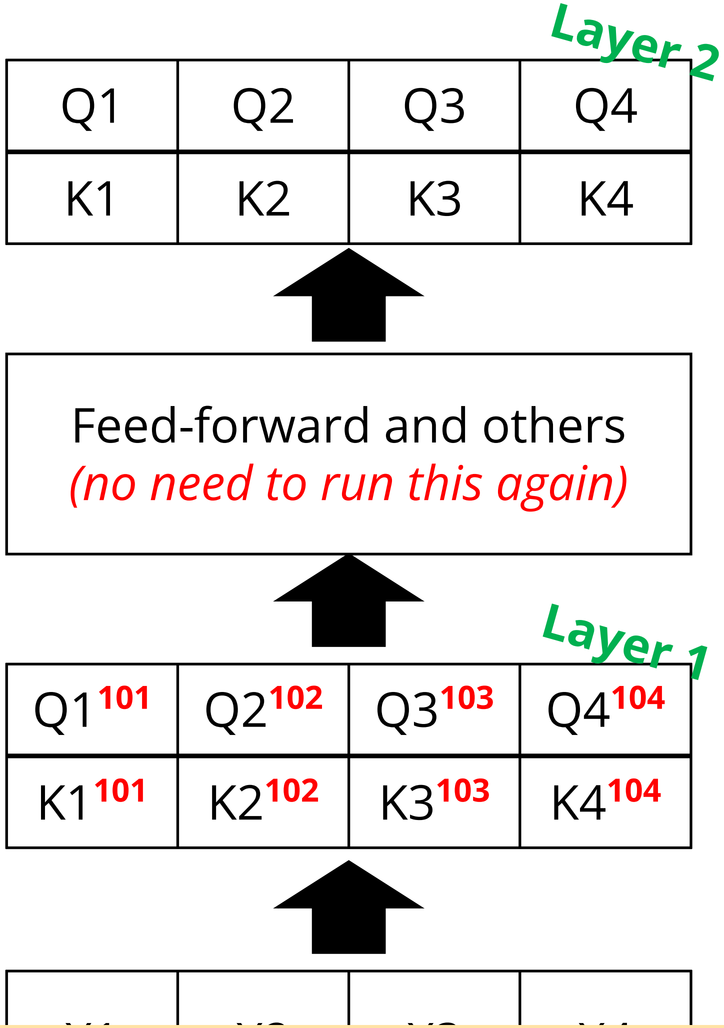


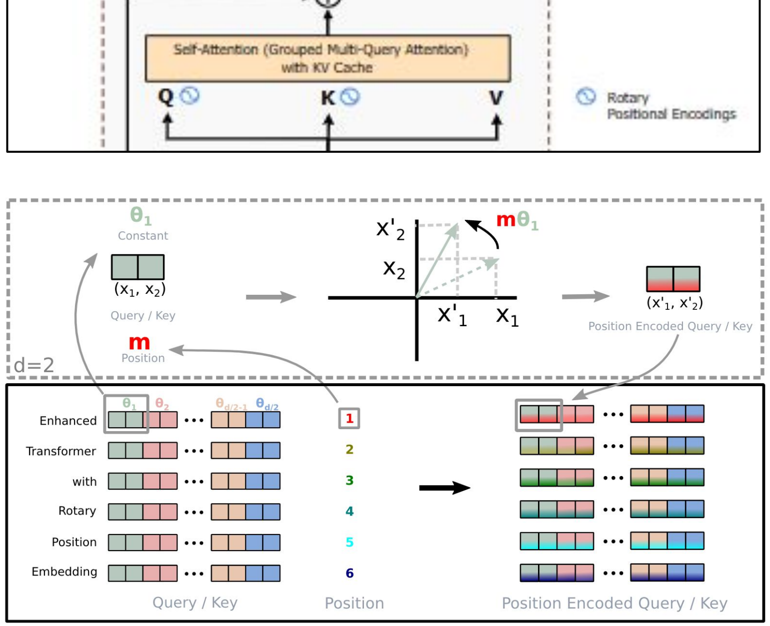
$$sim(K^{\theta_1}, Q^{\theta_2}) = f(K, Q, \boldsymbol{\theta_1} - \boldsymbol{\theta_2})$$

Shifting a block doesn't change output

Position re-encoding







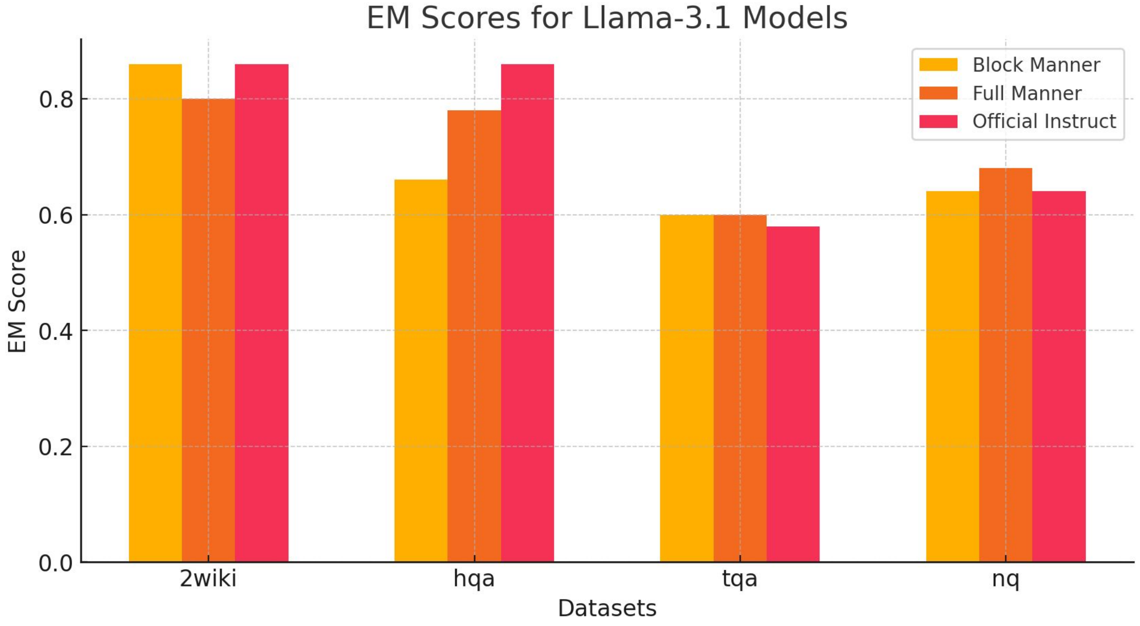
$$sim(K^{\theta_1}, Q^{\theta_2}) = f(K, Q, \boldsymbol{\theta_1} - \boldsymbol{\theta_2})$$

Shifting a block doesn't change output

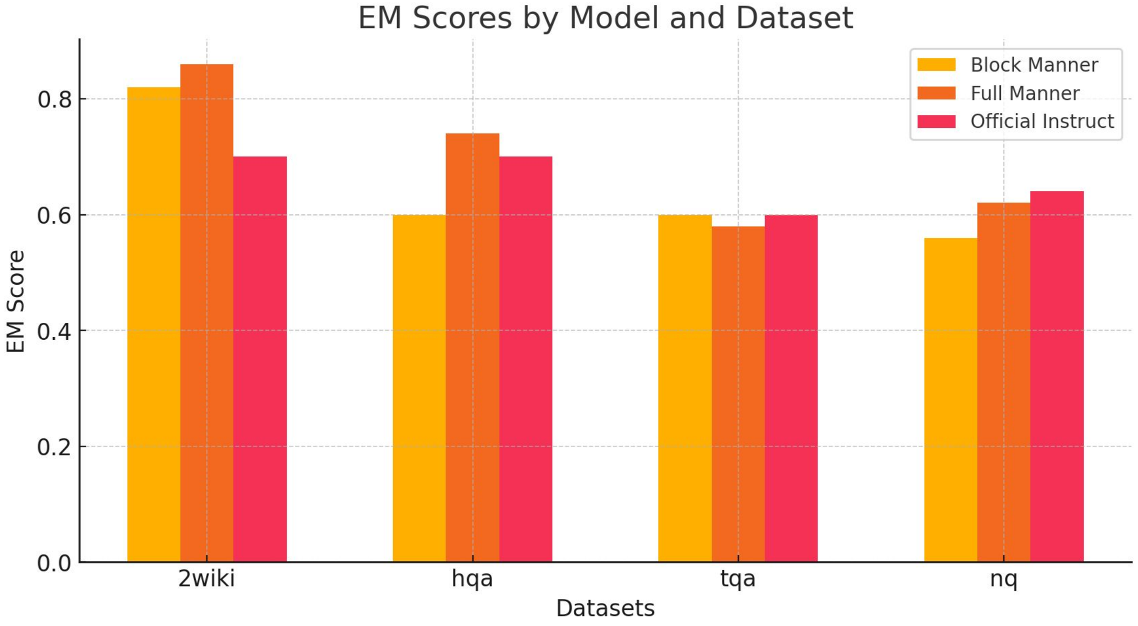
In actual implementation, we only need to **rotate keys** (not queries and values)

We could re-produce results

Llama 3.1- 8B

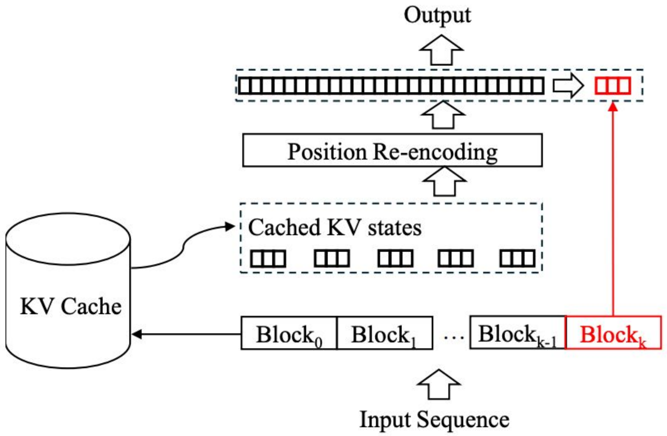


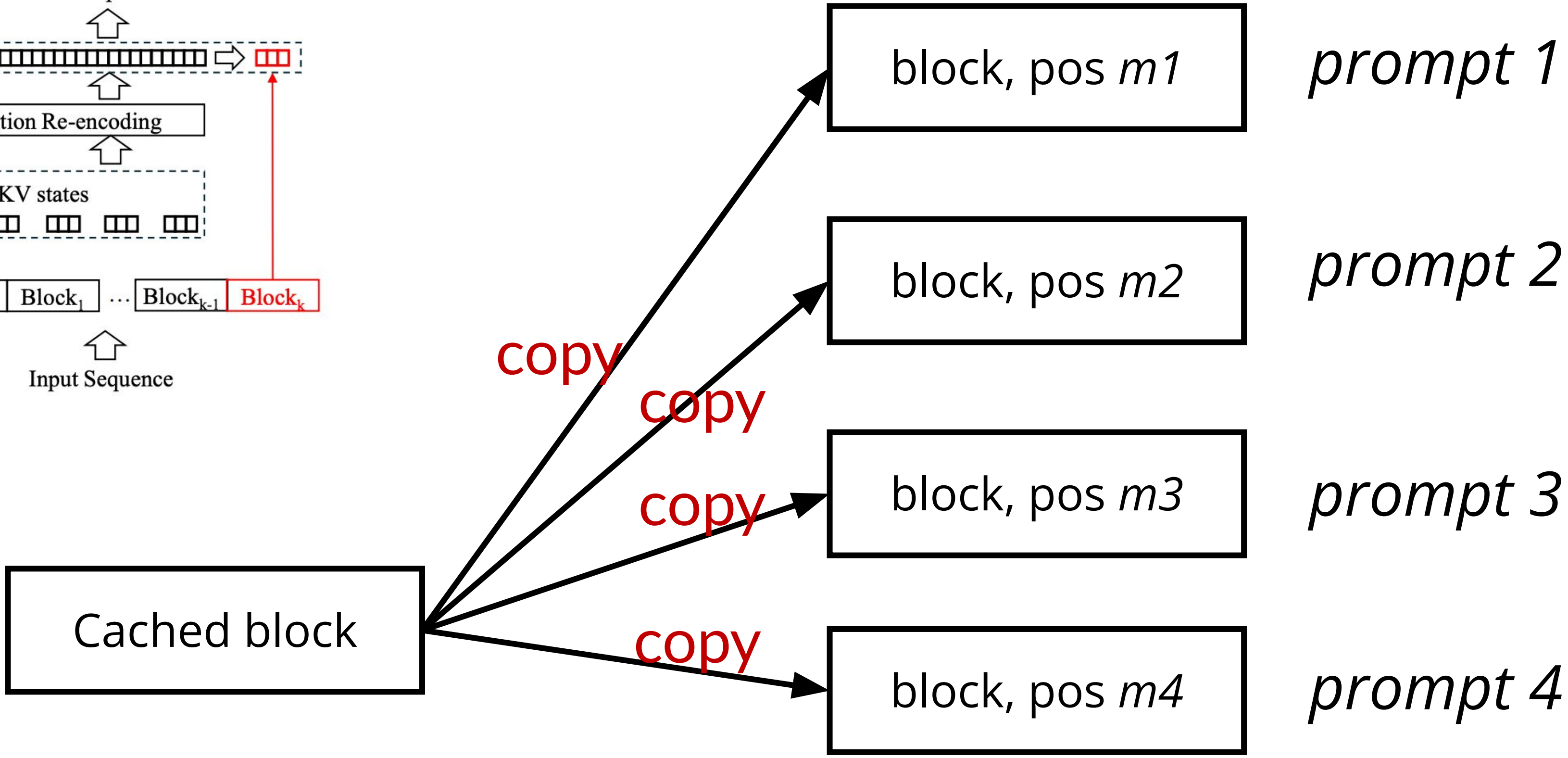
Llama 3.2 – 3B



Large model (70B parameters) could produce **similar accuracy without** fine-tuning

Limitation: **redundant** KV states for each cache hit





Our ongoing research: add positions ***on the fly,*** inside GPU registers (not DRAM)

Block diffusion (ICLR'25)

Autoregression: ✅ High quality ✅ Arbitrary-length ✅ KV caching ❌ Not parallelizable

There

Diffusion: ❌ Lower quality ❌ Fixed-length ❌ No KV caching ✅ Parallelizable

Block Diffusion: ✅ High quality ✅ Arbitrary-length ✅ KV caching ✅ Parallelizable

Block Diffusion Internal

Large Language Diffusion Models (LLaDA) : Sequential Sampling

Block 0	[MASK] [MASK] [MASK] [UNK]	Lily [MASK] [MASK] [UNK]	Lily can [MASK] [UNK]	Lily can run [UNK]
Block 1	[MASK] 2 [MASK] [MASK]	1 2 [MASK] [MASK]	1 2 kilometers [MASK]	1 2 kilometers per
Block 2	hour [MASK] [MASK] [MASK]	hour [MASK] [UNK] [MASK]	hour for [UNK] [MASK]	hour for [UNK] 4
Block 3	hours [MASK] [MASK] [MASK]	hours , [MASK] [MASK]	hours , so [MASK]	hours , so she
Block 4	runs [MASK] [MASK] [MASK]	runs [MASK] [MASK] of	runs a [MASK] of	runs a total of
Block 5	[UNK] [MASK] [MASK] [MASK]	[UNK] [MASK] 2 [MASK]	[UNK] 1 2 [MASK]	[UNK] 1 2 *
Block 6	[UNK] [MASK] [MASK] [MASK]	[UNK] 4 [MASK] [MASK]	[UNK] 4 = [MASK]	[UNK] 4 = [UNK]
Block 7	[MASK] 8 [MASK] [MASK]	4 8 [MASK] [MASK]	4 8 kilometers [MASK]	4 8 kilometers .
Block 8	[UNK] [MASK] [MASK] [MASK]	[UNK] After [MASK] [MASK]	[UNK] After that [MASK]	[UNK] After that ,
Block 9	she [MASK] [MASK] [MASK]	she runs [MASK] [MASK]	she runs [UNK] [MASK]	she runs [UNK] 6
Block 10	kilometers [MASK] [MASK] [MASK]	kilometers [MASK] hour [MASK]	kilometers per hour [MASK]	kilometers per hour for
Block 11	the [MASK] [MASK] [MASK]	the remaining [MASK] [MASK]	the remaining [UNK] [MASK]	the remaining [UNK] 4
Block 12	hours [MASK] [MASK] [MASK]	hours , [MASK] [MASK]	hours , [MASK] she	hours , so she
Block 13	runs [MASK] [MASK] [MASK]	runs [MASK] [MASK] of	runs [MASK] total of	runs a total of
Block 14	[UNK] [MASK] [MASK] [MASK]	[UNK] [MASK] [MASK] [UNK]	[UNK] [MASK] * [UNK]	[UNK] 6 * [UNK]
Block 15	[MASK] = [MASK] [MASK]	[MASK] = [UNK] [MASK]	4 = [UNK] [MASK]	4 = [UNK] 2
Block 16	[MASK] kilometers [MASK] [MASK]	4 kilometers [MASK] [MASK]	4 kilometers [MASK] [UNK]	4 kilometers . [UNK]
Block 17	Therefore [MASK] [MASK] [MASK]	Therefore , [MASK] [MASK]	Therefore , [MASK] can	Therefore , Lily can
Block 18	run [MASK] [MASK] [MASK]	run [MASK] [MASK] of	run [MASK] total of	run a total of
Block 19	[UNK] [MASK] [MASK] [MASK]	[UNK] 4 [MASK] [MASK]	[UNK] 4 8 [MASK]	[UNK] 4 8 +
Block 20	[UNK] [MASK] [MASK] [MASK]	[UNK] [MASK] 4 [MASK]	[UNK] 2 4 [MASK]	[UNK] 2 4 =
Block 21	[UNK] [MASK] [MASK] [MASK]	[UNK] [MASK] 2 [MASK]	[UNK] 7 2 [MASK]	[UNK] 7 2 kilometers
Block 22	in [MASK] [MASK] [MASK]	in [MASK] 8 [MASK]	in [UNK] 8 [MASK]	in [UNK] 8 hours
Block 23	. [MASK] [MASK] [MASK]	. [UNK] [MASK] [MASK]	. [UNK] The [MASK]	. [UNK] The final
Block 24	[MASK] [MASK] [UNK] [MASK]	[MASK] [MASK] [UNK] 7	[MASK] is [UNK] 7	result is [UNK] 7
Block 25	2 [MASK] [MASK] [MASK]	2 [MASK] [UNK] [MASK]	2 [MASK] [UNK] [UNK]	2 [UNK] [UNK] [UNK]
Block 26	[MASK] [UNK] [MASK] [MASK]	[MASK] [UNK] [UNK] [MASK]	[MASK] [UNK] [UNK] [UNK]	[UNK] [UNK] [UNK] [UNK]
Block 27	[MASK] [UNK] [MASK] [MASK]	[MASK] [UNK] [MASK] [UNK]	[MASK] [UNK] [UNK] [UNK]	[UNK] [UNK] [UNK] [UNK]
Block 28	[MASK] [MASK] [MASK] [UNK]	[MASK] [MASK] [UNK] [UNK]	[MASK] [UNK] [UNK] [UNK]	[UNK] [UNK] [UNK] [UNK]
Block 29	[MASK] [UNK] [MASK] [MASK]	[MASK] [UNK] [MASK] [UNK]	[MASK] [UNK] [UNK] [UNK]	[UNK] [UNK] [UNK] [UNK]
Block 30	[MASK] [UNK] [MASK] [MASK]	[UNK] [UNK] [MASK] [MASK]	[UNK] [UNK] [MASK] [UNK]	[UNK] [UNK] [UNK] [UNK]
Block 31	[MASK] [MASK] [MASK] [UNK]	[MASK] [MASK] [UNK] [UNK]	[MASK] [UNK] [UNK] [UNK]	[UNK] [UNK] [UNK] [UNK]

Prompt:

Lily can run 12 kilometers per hour for 4 hours. After that, she runs 6 kilometers per hour. How many kilometers can she run in 8 hours?

Final Answer:

Lily can run 1 2 kilometers per hour for 4 hours , so she runs a total of 1 2 * 4 = 4 8 kilometers . After that , she runs 6 kilometers per hour for the remaining 4 hours , so she runs a total of 6 * 4 = 2 4 kilometers . Therefore , Lily can run a total of 4 8 + 2 4 = 7 2 kilometers in 8 hours . The final result is 7 2

Parallel Block Diffusion

Large Language Diffusion Models (LLaDA) : Diagonal Sampling

Global Timestep: 34

Block_0	Block_1	Block_2	Block_3	Block_4	Block_5	Block_6	Block_7
[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]
Lily [MASK] [MASK] [UNK]	[MASK] 2 [MASK] [MASK]	hour [MASK] [MASK] [MASK]	hours [MASK] [MASK] [MASK]	can [MASK] [MASK] [MASK]	[MASK] [UNK] [MASK] [MASK]	[MASK] [UNK] [MASK] [MASK]	[MASK] [MASK] 8 [MASK]
Lily can [MASK] [UNK]	1 2 [MASK] [MASK]	hour for [MASK] [MASK]	hours , [MASK] [MASK]	can run [MASK] [MASK]	[MASK] [UNK] [MASK] 2	[MASK] [UNK] 4 [MASK]	[UNK] [MASK] 8 [MASK]
Lily can run [UNK]	1 2 [MASK] per	hour for [UNK] [MASK]	hours , [MASK] she	can run [MASK] total	of [UNK] [MASK] 2	[MASK] [UNK] 4 =	[UNK] 4 8 [MASK]
	1 2 kilometers per	hour for [UNK] 4	hours , so she	can run a total	of [UNK] 1 2	* [UNK] 4 =	[UNK] 4 8 kilometers

Block_8	Block_9	Block_10	Block_11	Block_12	Block_13	Block_14	Block_15
[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]
[MASK] [MASK] After [MASK]	[MASK] [MASK] runs [MASK]	[MASK] kilometers [MASK] [MASK]	for [MASK] [MASK] [MASK]	[MASK] hours [MASK] [MASK]	she [MASK] [MASK] [MASK]	[MASK] [MASK] [UNK] [MASK]	[MASK] [UNK] [MASK] [MASK]
[MASK] [UNK] After [MASK]	, [MASK] runs [MASK]	[MASK] kilometers per [MASK]	for [MASK] remaining [MASK]	4 hours [MASK] [MASK]	she [MASK] run [MASK]	[MASK] of [UNK] [MASK]	* [UNK] [MASK] [MASK]
. [UNK] After [MASK]	, [MASK] runs [UNK]	[MASK] kilometers per hour	for the remaining [MASK]	4 hours [MASK] so	she can run [MASK]	total of [UNK] [MASK]	* [UNK] [MASK] =
. [UNK] After that	, she runs [UNK]	6 kilometers per hour	for the remaining [UNK]	4 hours , so	she can run a	total of [UNK] 6	* [UNK] 4 =

Block_16	Block_17	Block_18	Block_19	Block_20	Block_21	Block_22	Block_23
[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]
[UNK] [MASK] [MASK] [MASK]	[MASK] [UNK] [MASK] [MASK]	[MASK] can [MASK] [MASK]	total [MASK] [MASK] [MASK]	8 [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] 7	2 [MASK] [MASK] [MASK]	[MASK] hours [MASK] [MASK]
[UNK] [MASK] [MASK] kilometers	. [UNK] [MASK] [MASK]	[MASK] can run [MASK]	total of [MASK] [MASK]	8 [MASK] [UNK] [MASK]	4 [MASK] [MASK] 7	2 kilometers [MASK] [MASK]	8 hours [MASK] [MASK]
[UNK] 2 [MASK] kilometers	. [UNK] [MASK] ,	[MASK] can run a	total of [UNK] [MASK]	8 [MASK] [UNK] 2	4 [MASK] [UNK] 7	2 kilometers in [MASK]	8 hours . [MASK]
[UNK] 2 4 kilometers	. [UNK] Therefore ,	Lily can run a	total of [UNK] 4	8 + [UNK] 2	4 = [UNK] 7	2 kilometers in [UNK]	8 hours . [UNK]

Block_24	Block_25	Block_26	Block_27	Block_28	Block_29	Block_30	Block_31
[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [MASK]
[MASK] [MASK] [MASK] [UNK]	7 [MASK] [MASK] [MASK]	[MASK] [MASK] [MASK] [UNK]	[MASK] [MASK] [UNK] [MASK]	[MASK] [MASK] [MASK] [UNK]	[MASK] [UNK] [MASK] [MASK]	[MASK] [MASK] [UNK] [MASK]	[MASK] [MASK] [MASK] [UNK]
[MASK] [MASK] : [UNK]	7 2 [MASK] [MASK]	[MASK] [MASK] [UNK] [UNK]	[MASK] [MASK] [UNK] [UNK]	[MASK] [MASK] [UNK] [UNK]	[UNK] [UNK] [MASK] [MASK]	[MASK] [UNK] [UNK] [MASK]	[MASK] [MASK] [UNK] [UNK]
[MASK] answer : [UNK]	7 2 [UNK] [MASK]	[MASK] [UNK] [UNK] [UNK]	[MASK] [UNK] [UNK] [UNK]	[MASK] [UNK] [UNK] [UNK]	[UNK] [UNK] [UNK] [MASK]	[UNK] [UNK] [UNK] [MASK]	[MASK] [UNK] [UNK] [UNK]
Final answer : [UNK]	7 2 [UNK] [UNK]	[UNK] [UNK] [UNK] [UNK]	[UNK] [UNK] [UNK] [UNK]	[UNK] [UNK] [UNK] [UNK]	[UNK] [UNK] [UNK] [UNK]	[UNK] [UNK] [UNK] [UNK]	[UNK] [UNK] [UNK] [UNK]

Prompt:

Summary

- RAG can be faster by bypassing cross-attention
- Block-attention can achieve good performance with fine-tuning
- Its position re-encoding can be memory inefficient

Questions?