ThinKV: Thought-Adaptive KV Cache Compression For Efficient Reasoning Models

Akshat Ramachandran[1], Marina Neseem[2], Charbel Sakr[2], Rangharajan Venkatesan[2], Brucek Khailany[2] and Tushar Krishna[1]

[1]Georgia Institute of Technology, [2]NVIDIA Research

My Contact Details

Link to Paper

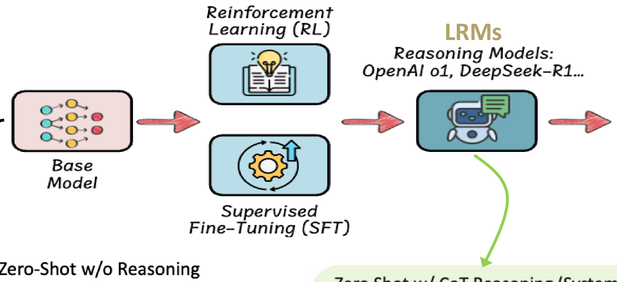
ICLR International Conference On Learning Representations

Reasoning models improve capability but introduce significant memory and efficiency challenges

ThinKV decomposes CoT into distinct thought types and thought-adaptively performs quantization and eviction

ThinKV establishes a new pareto frontier

What are Reasoning Models ?



Zero-Shot w/o Reasoning

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 ✗

Zero Shot w/ CoT Reasoning (System 2)

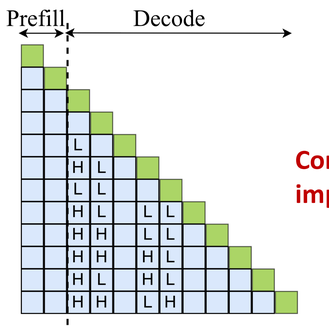
Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

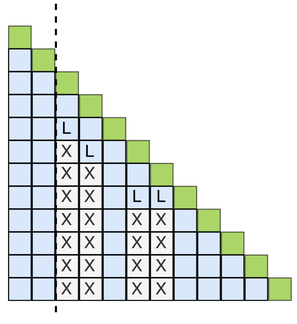
Unique Challenges of LRMs

Long output context generation

Model	Weights (GB)	KV Cache (GB)
DeepSeek-LLaMA-8B	16	64
DeepSeek-Qwen-32B	64	128
DeepSeek-LLaMA-70B	140	160



Convoluted token importance

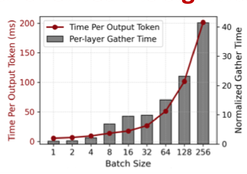


Internal memory fragmentation

Block 0	A	B	C	D
Block 1	E	F	G	H
Block 2	I	J	K	L
Block 3	M	N	O	P

	A	~~B~~	C	D
	E	~~F~~	~~G~~	H
	I	J	K	L
	M	~~N~~	O	P

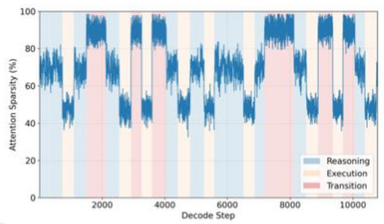
Gather overhead is significant



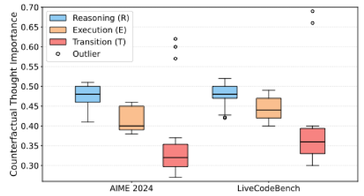
ThinKV's Key Observations

LRM chain-of-thought can be decomposed into 3 distinct thought types:

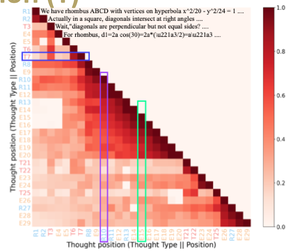
Reasoning (R), Execution (E) and Transition (T)



Degree of attention sparsity enables thought decomposition



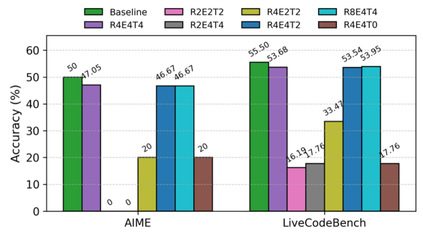
Thought importance: R >= E > T



With each T thought, prior thoughts diminish in influence

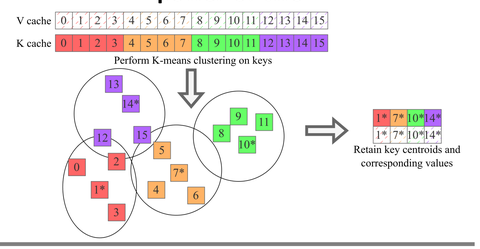
Think Before you Quantize, Think Before you Evict

Group Quantization, g=16, sf = 8-bit (E4M3)

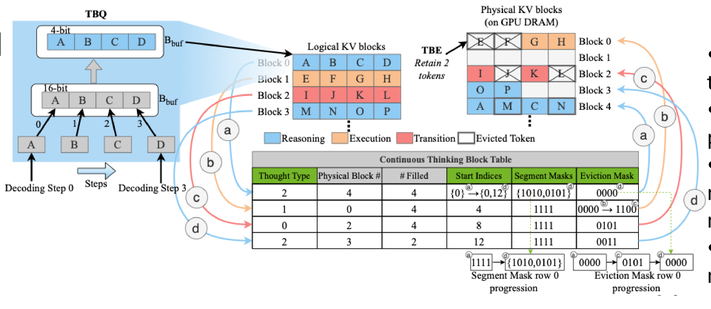


R, E @ NVFP4 and T @ Ternary

At the end of every T thought, progressively anneal previous thought segments to nearest power-of-two



Continuous Thinking Kernel: Extending vLLM



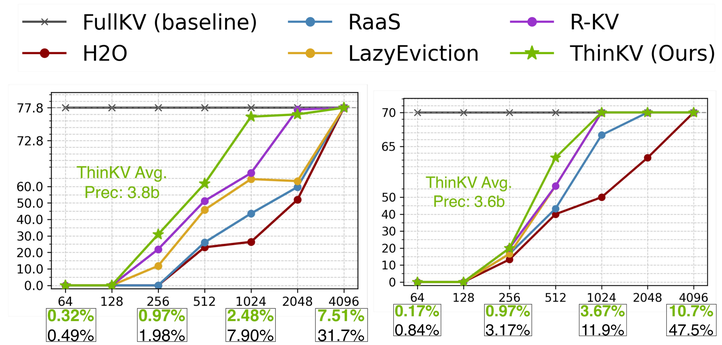
Continuous Thinking Block Table

Thought Type	Physical Block #	# Filled	Start Indices	Segment Masks	Eviction Mask
2	4	4	{0} → {0,12}	{1010,0101}	0000
1	0	4	4	1111	0000 → 1100
0	2	4	8	1111	0101
2	3	2	12	1111	0011

New Block Table Fields

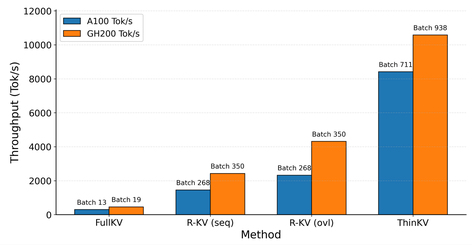
- **Thought type:** Thought type of tokens in a block.
- **Start indices:** Records the start position of the thought segment.
- **Segment masks:** If there are multiple start indices, the segment mask is a bit vector.
- **Eviction mask:** A bit vector marking positions of tokens evicted.

Accuracy and Performance Analysis

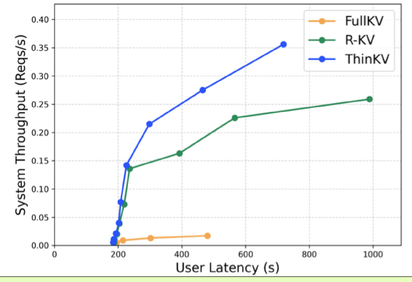


GPT-OSS-20B on LiveCodeBench | **DeepSeek-R1-LLaMA-70B on AIME**

ThinKV achieves near lossless accuracy with < 3.67% of FullKV memory



ThinKV supports 3X higher batch size and up to 5.8X higher TPS



ThinKV achieves 38% higher reqs/s and 27% lower latency compared to R-KV