

Reasoning Path Compression: Compressing Generation Trajectories for Efficient LLM Reasoning

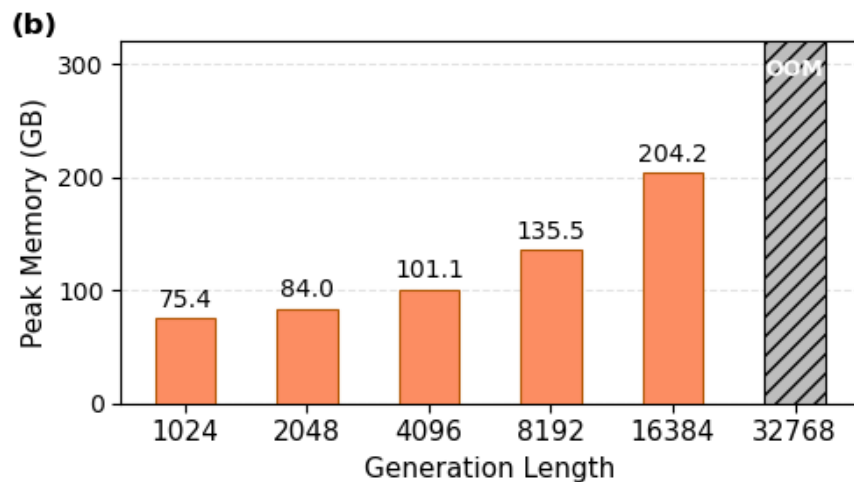
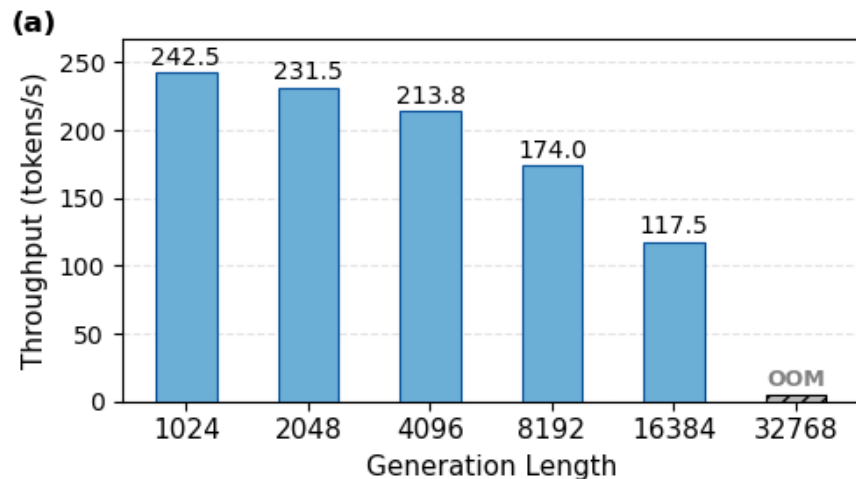
Jiwon Song¹, Dongwon Jo¹, Yulhwa Kim^{2*}, Jae-Joon Kim^{1*}

The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024.



¹Seoul National University, ²Sungkyunkwan University, *Corresponding Author

Motivation



Problem

Reasoning LLM = Long Reasoning Path → ***Huge KV Cache***

Reasoning LLMs generate > 50K tokens for one answer
KV cache → **Major memory and throughput bottleneck**

Goal

Compressing KV cache of generated tokens without retraining

Inference-time compression method for generated tokens

Observation: Semantic Sparsity

Question: Four identical spin-1/2 particles are placed in a one-dimensional infinite potential well of length L . Calculate the energies of the ground state, first excited state, and second excited state of the system in terms of energy E . (Here $E = \pi^2 \hbar^2 / 2mL^2$)

Short Input (~200 Tokens)

Reasoning LLM

Long Output (10,000+ Tokens)

Reasoning Path Example from QwQ-32B

[...] Since there are four particles, **the ground state should have the first two energy levels each filled with two particles. Wait, no, actually, each n level can hold two particles, so for four particles, that would be $n=1$ and $n=2$ each filled with two particles? Wait, no, $n=1$ can have two, $n=2$ can have two, so total four. So the ground state would have two particles in $n=1$ and two in $n=2$?**

Wait a second, but I need to be careful. Let me think again. [...] So for four particles, **the lowest energy configuration is to fill up $n=1$ (two particles) and $n=2$ (another two particles).** Therefore, the total energy would be $2E_1 + 2E_2$. Since $E_n = n^2E$, **so that would be $2(1^2E) + 2(2^2E) = 2E + 8E = 10E$.** Wait, but hold on, that seems like the total energy. Hmm, but let me confirm.

Wait, actually, each particle in the $n=1$ level contributes E , each [...] So total energy is $2E + 8E = 10E$. So the ground state energy is $10E$?

But wait, hold on. Let me check again. The energy levels for each particle are $E_n = n^2E$, so $n=1$ is E , $n=2$ is $4E$, $n=3$ is $9E$, etc. **So for two particles in $n=1$, their total contribution is $2E$. Two particles in $n=2$, their contribution is $2(4E) = 8E$.** So total $10E$. That seems right. [...]

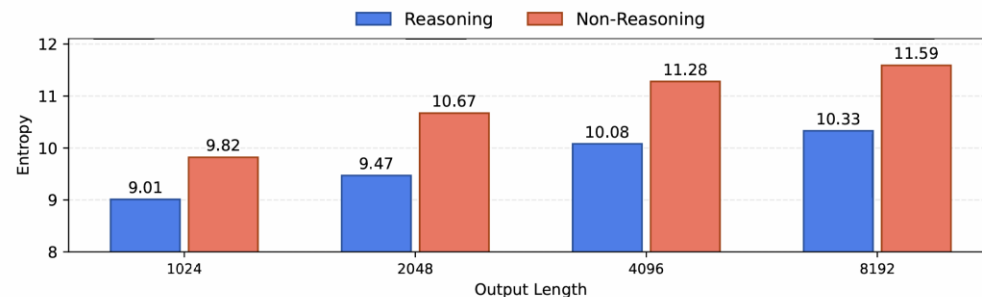
Reasoning paths contain redundant logic and self-checks

Low phrase-level entropy

Repetitive phrases and semantic overlap

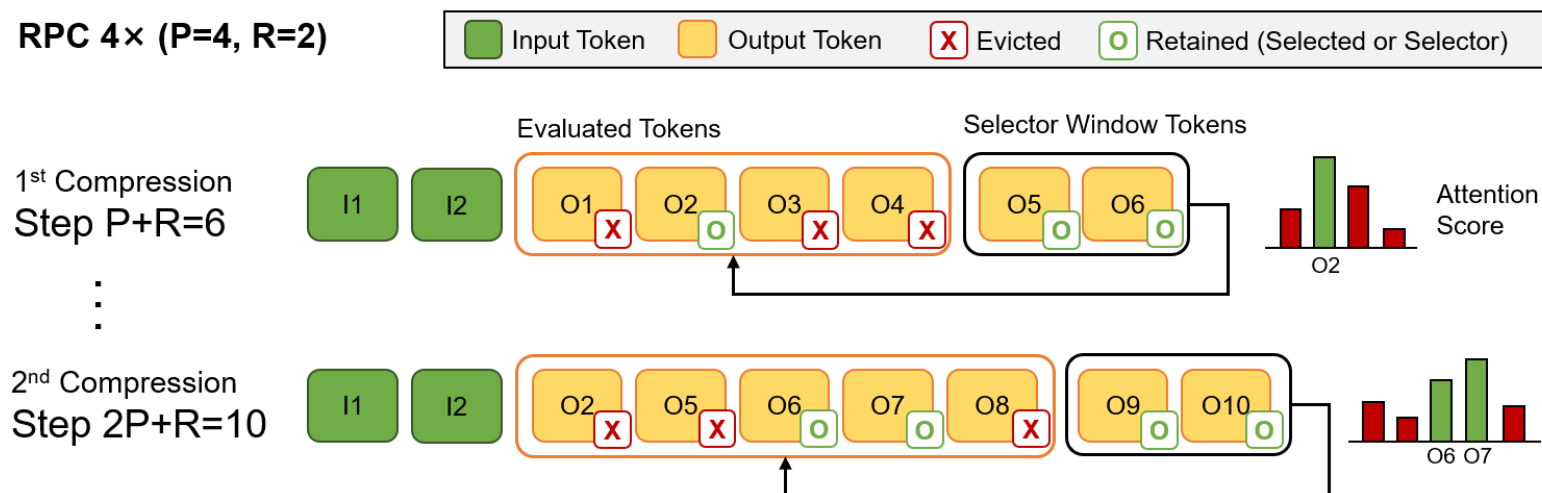
Define Semantic Sparsity

Enables aggressive compression of KV cache



Method: Reasoning Path Compression

RPC $4 \times (P=4, R=2)$

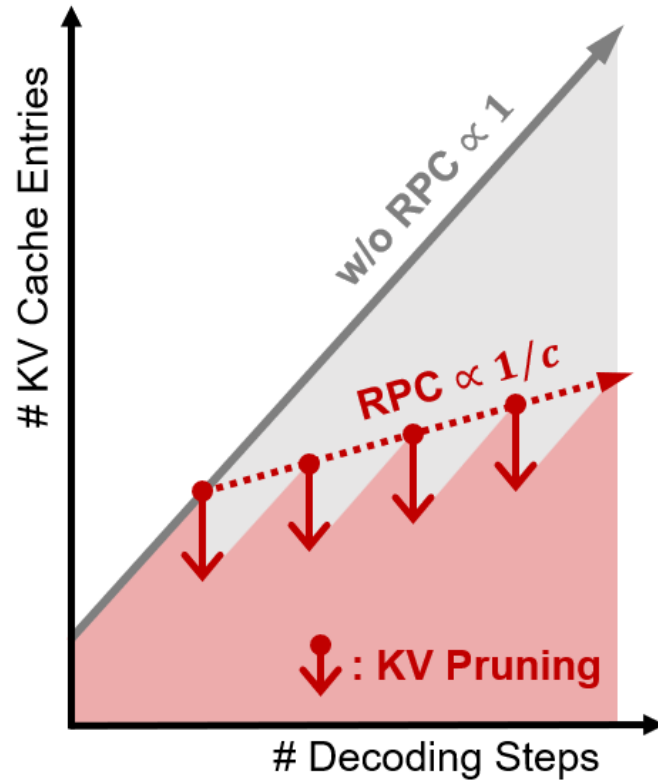


Reasoning Path Compression (RPC) = periodic KV cache compression during decoding

Uses attention-based importance from recent tokens (selector window)

Training-free / plug-in / model-agnostic

Periodic Compression Dynamics



Compression period P , # Selector window tokens R , Target ratio c

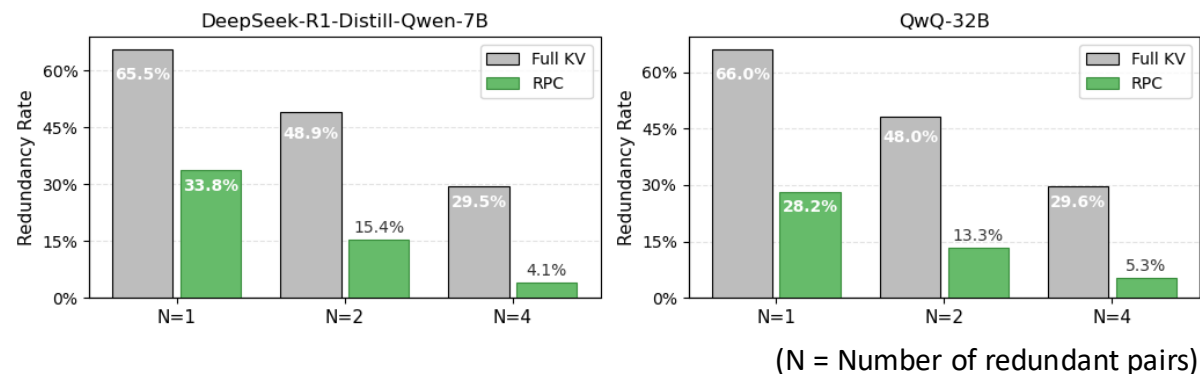
Compression triggered every P tokens

Query states of R select window tokens used for importance scoring

$NP/c + R$ KV entries retained after $NP + R$ decoding steps

Outdated, Unimportant tokens fade out \rightarrow Steady, compact context

Results: Redundancy Reduction and Accuracy



RPC reduces redundant sentence pairs

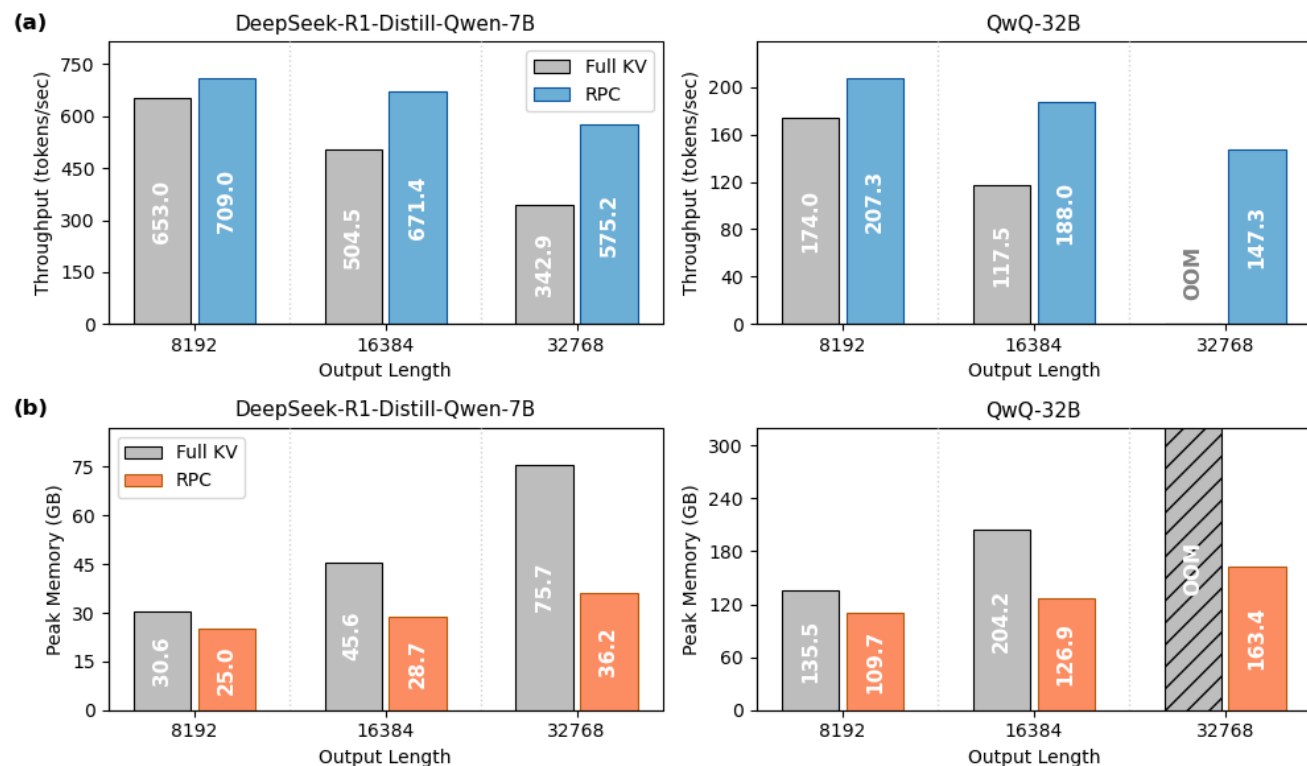
- Proportion of semantically redundant sentences decreased by over 50%
- Stronger effect with higher threshold (N=2,4)

Full KV cache level accuracy preserved

- Accuracy remains almost unchanged ($\leq 1.2\%$ drop on AIME 2024)
- Outperforms baselines by large margin

Method	DeepSeek-R1-Distill-Qwen-7B			QwQ-32B		
	AIME 2024 (pass@1)	LiveCodeBench (pass@1)	IFEval (pass@1)	AIME 2024 (pass@1)	LiveCodeBench (pass@1)	IFEval (pass@1)
Full KV Cache	55.5	37.6	55.1	79.5	63.4	83.9
H2O	42.5	22.5	51.8	75.0	54.2	74.3
TOVA	42.5	21.5	48.8	70.0	43.8	50.6
LightThinker	6.7	0.7	25.1	-	-	-
RPC ($P = 4096$)	52.9	35.9	56.6	78.3	62.2	82.6
RPC ($P = 1024$)	50.4	33.5	57.3	78.3	61.2	81.7

Results: Efficiency



Throughput improvement up to 1.6×
Gains amplify with model size and output length

Peak memory reduction up to 50%
Prevents OOM even for 32K + tokens reasoning

Conclusion & Takeaways

Key Idea

Reasoning LLMs often generate redundant reasoning paths

→ *Large KV Cache, Slow inference, High memory cost*

RPC exploits semantic sparsity

→ *Periodically removes low-importance KV entries w/o retraining and architecture change*

Main Results

4× KV Compression with redundancy reduction

Redundancy rate decreases by over 50%

Throughput ↑ 1.6×, memory ↓ >50% with accuracy drop ≤ 1.2% (AIME24, QwQ-32B)

RPC leverages semantic sparsity for faster, lighter, and scalable reasoning