



REAL: REtrieval-reAsoning and Logic-constructed Attention Behaviors for Long-Context KV Cache Compression

Mengjie Li[1] , Yuan Feng[2] , Xike Xie[3] , and William J. Song[1†]

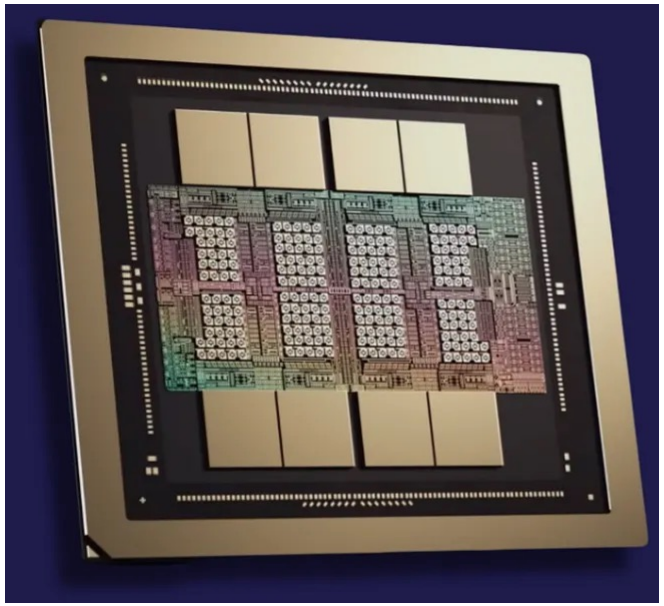
[1] School of Electrical and Electronic Engineering, Yonsei University
[2] School of Computer Science, University of Science and Technology of China
[3] School of Biomedical Engineering, University of Science and Technology of China

Content

- **Background:** Memory Bottleneck of KV Cache
- **Limitations of Existing Head-wise Methods:** Awareness of Success
- **Motivation:** Awareness of Both Success and Failure
- **REAL:** Attention Behavior Matrix & Head-wise KV Budget Allocation
- **Evaluation**
- **Conclusion**

Background: Memory Bottleneck of KV Cache

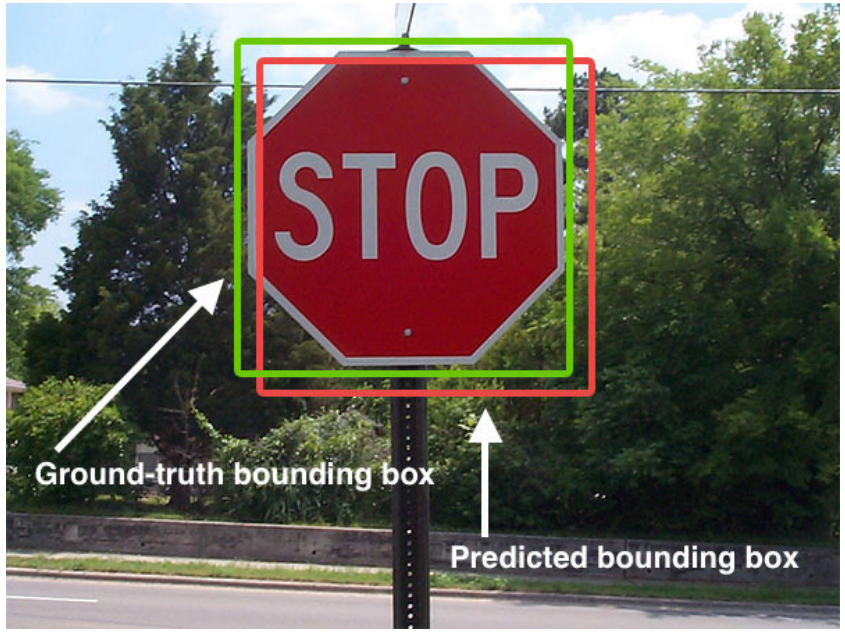
GPT-3



- **KV cache requirement: 1,208 GB**
  - For example, the OpenAI GPT-3 model
  - Batch size: 64
  - Sequence length: 4096
- **NVIDIA Rubin GPU HBM4: 288 GB**
  - The memory capacity is still insufficient.

Efficient KV cache management is essential.

Limitations of Existing Head-wise Methods: Awareness of Success



Prediction matches the ground truth.

Retrieval-based behavior:

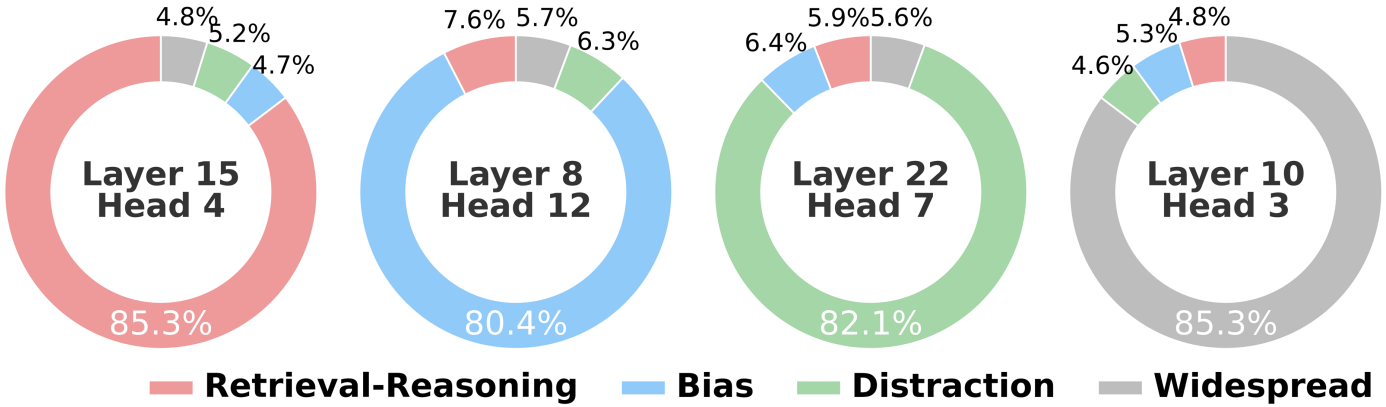
- DuoAttention
- RazorAttention
- HeadKV-R
- HeadKV-R2

Question:
What is the **favorite thing** of **the younger one** between **John** and **Mary**?

Needle:
John is 12 years old. Mary is 13 years old. (r part) Mary's favorite thing is to take a walk in Chaoyang Park and have a cup of Espresso in the evening. (c^1 part) John's favorite thing is to play basketball at the local gym and enjoy a smoothie afterward. (c^2 part)

[Ignorance]
The correct answer is derived from **the c^2 part** given **background r**, while the influence of **misleading c^1 is neglected**.

Motivation: Awareness of Both Success and Failure



Diverse head behaviors arise during inference.

	Question-related	Answer-related	Similar to answer source structure
Retrieval & reasoning	✓	✓	**N/A**
Bias	✓	✗	✓
Distraction	✗	✗	✓
Widespread	✗	✗	✗

How to Construct A Needle.

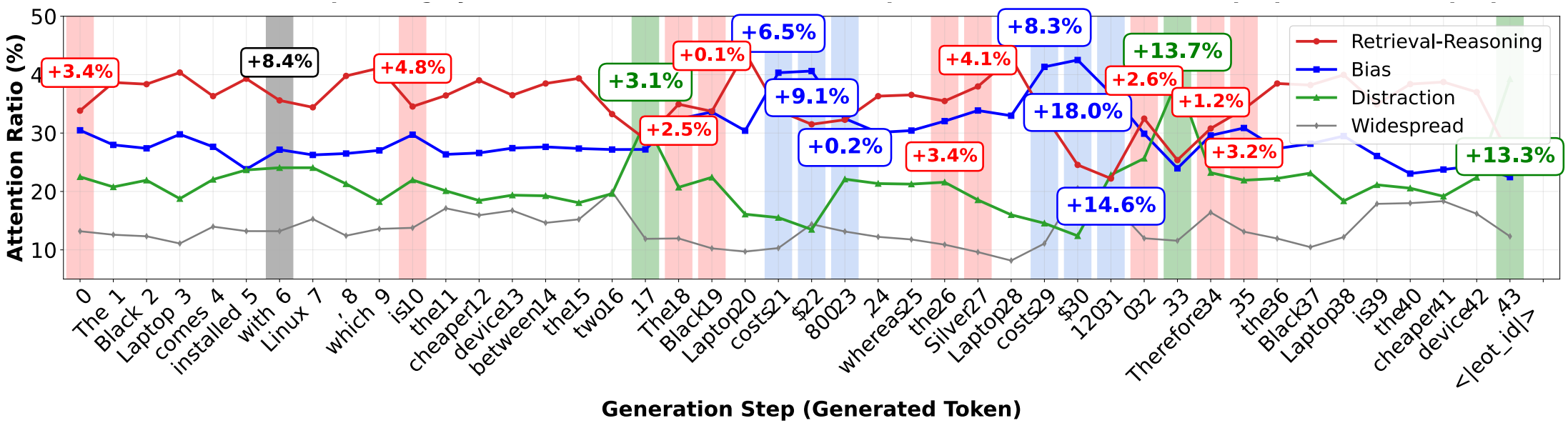
Question:
Which operating system comes installed on the cheaper device between the black laptop and the silver laptop?

Needle:
Reflecting the 2025 industry shift towards AI-integrated computing and neural processing, the silver laptop costs $1200, whereas the black laptop, targeting the essential productivity market without NPU acceleration, costs $800. The black laptop comes installed with Linux. The silver laptop comes installed with Windows. Microsoft Windows is known for its graphical user interfaces and broad hardware compatibility. The Red Tablet comes installed with Android. Google Android is based on the Linux kernel and designed primarily for mobile devices.

Attention Behavior:
Retrieval-Reasoning, Bias, Distraction, Widespread

Attention Behavior Matrix: Source

Question: Which operating system comes installed on the cheaper device between the black laptop and the silver laptop?
Needle: ... the silver laptop costs $1200 ... the black laptop costs $800 ... comes installed with Linux ... the red tablet comes installed with Android ...



Dominant attention behaviors affect model inference.

- **Bias causes misdirection:** At decisive steps 21, 22, 23 for "*cheaper*," the model bypasses retrieval-reasoning and instead extracts the price "*$800*" from its bias.
- **Distraction misses insights with nearly 70%:** At two steps, 17 and 43, within three dominant steps, the model overlooks meaningful information.

Bias → False Positive (FP) | Distraction → False Negative (FN)

Attention Behavior Matrix: Metric Calculation

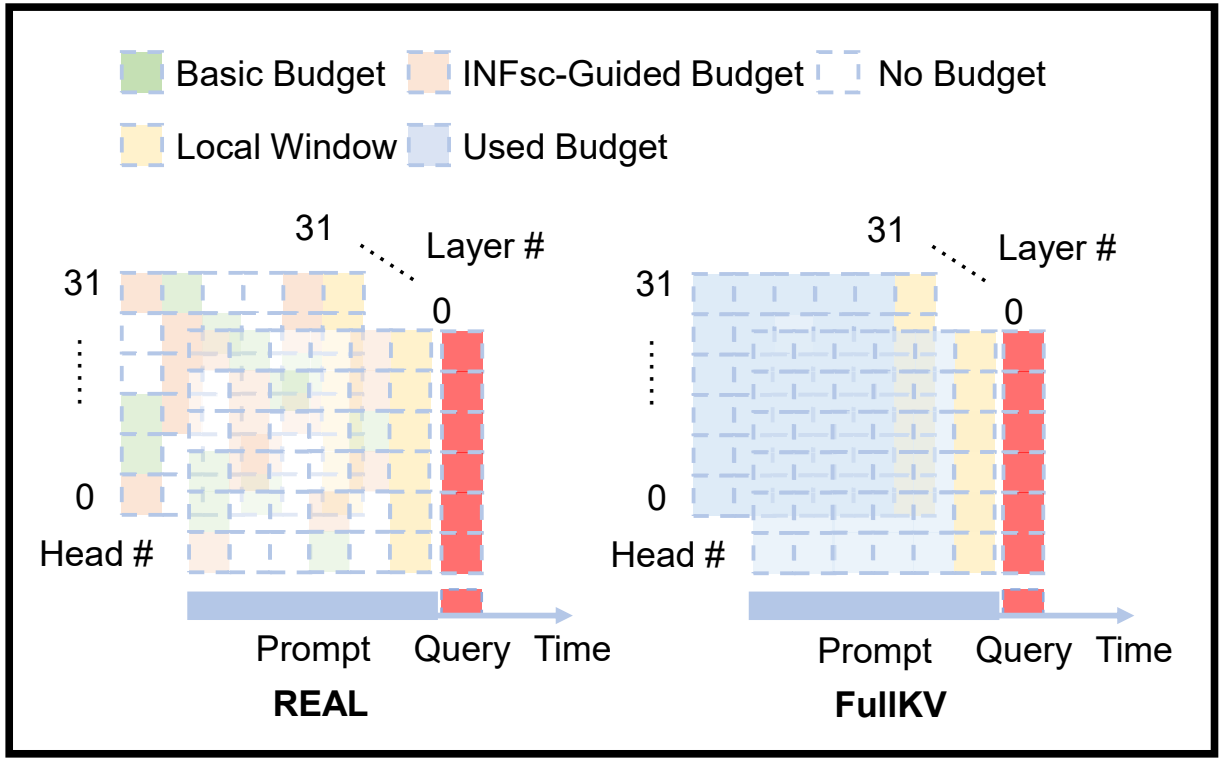
		Ground truth	
		Answer-related (P)	Non-answer-related (N)
Inference	Answer-related (P)	Retrieval-reasoning (TP)	Bias (FP)
	Non-answer-related (N)	Distraction (FN)	Widespread (TN)

$$RAsc = \frac{W_R}{W_R + W_D} \qquad LCsc = \frac{W_R}{W_R + W_B} \qquad INFsc = \frac{2 \cdot RA_{sc} \cdot LC_{sc}}{RA_{sc} + LC_{sc}}$$

The head attends more to retrieval-reasoning and less to bias and distraction.

→ Higher $INFsc$ → Larger KV budget

Head-wise KV Budget Allocation: Budget Amount



KV budget allocation illustration.

1. Basic budget $b_{base} = l\,r\,(1 - \frac{1}{\beta})$

2. Global budget pool $B_{total} = \frac{l\,r}{\beta}$ L H

3. Dynamic allocation $b_{i,j}^{dyn} = B_{total} \cdot H_{i,j}^{inf}$

4. $b_{i,j} = b_{base} + b_{i,j}^{dyn}$

5. $C_{i,j}$ = floor (max (0, $b_{i,j}$) + 0.5)

- Initially assigned cache ratio r for a head
- Sequence length l
- Predefined allocation ratio β

- Head (i, j)
- Layer count L, head count per layer H
- INF score $H_{i,j}^{inf}$

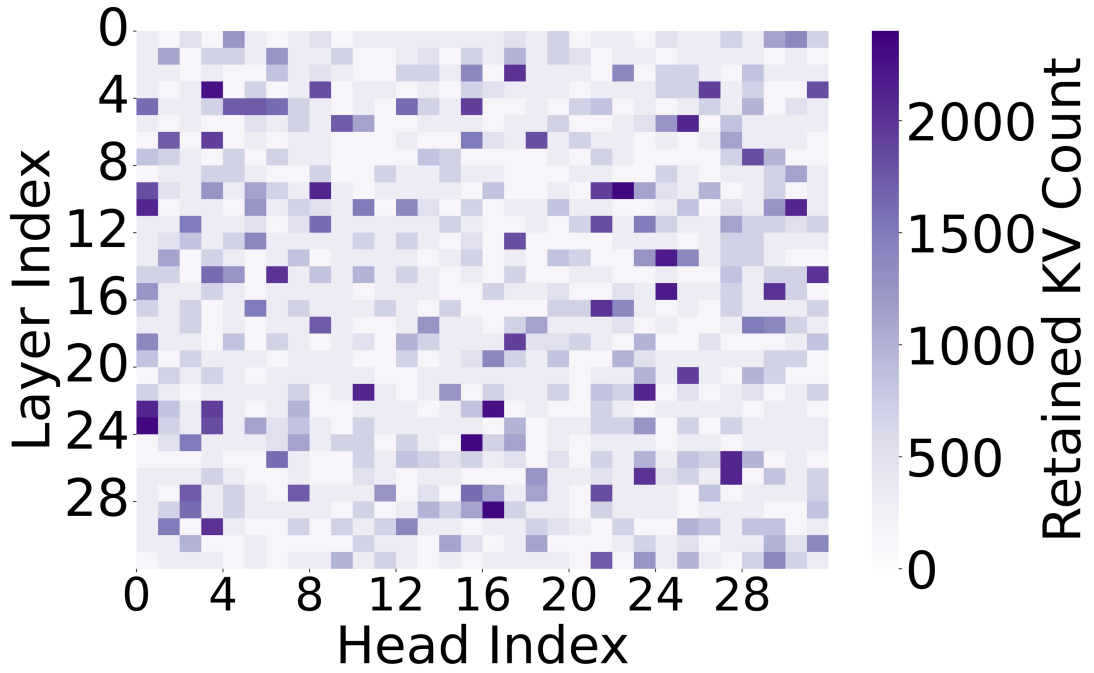
Head-wise KV Budget Allocation: How to Select Historical KV

- KV length $\leq$ the budget: Original K and V are returned unchanged
- KV length $>$ the budget:

$$S(Q_w, K_c) = Softmax(\frac{Q_w K_c^T}{\sqrt{d_k}})$$

$$KV_{res} = Gather(K_c, TopK(S, r)) + KV_w$$

  - Query window Q_w
  - Cached keys K_c
  - Compressed KV cache KV_{res}
  - KV window KV_w
  - Cache ratio r



Retained KV counts when token budget = 512.

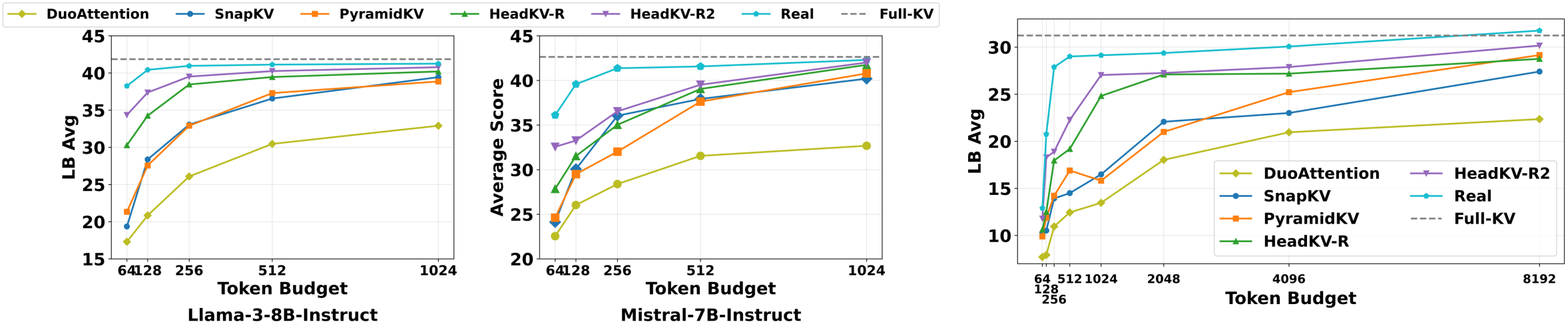
Experimental Setup

- **Base models:** Llama-3-8B-Instruct, Mistral-7B-Instruct, 123B Mistral-Large-Instruct-2411
- **27 datasets** from LongBench and LongBench v2
- **Baselines:** SnapKV, PyramidKV, DuoAttention, HeadKV-R, HeadKV-R2
- **Question settings:**
  - **Question-aware:** Questions are compressed alongside prefix context.
  - **Question-agnostic:** Only the prefix context is compressed w/o knowing future questions.

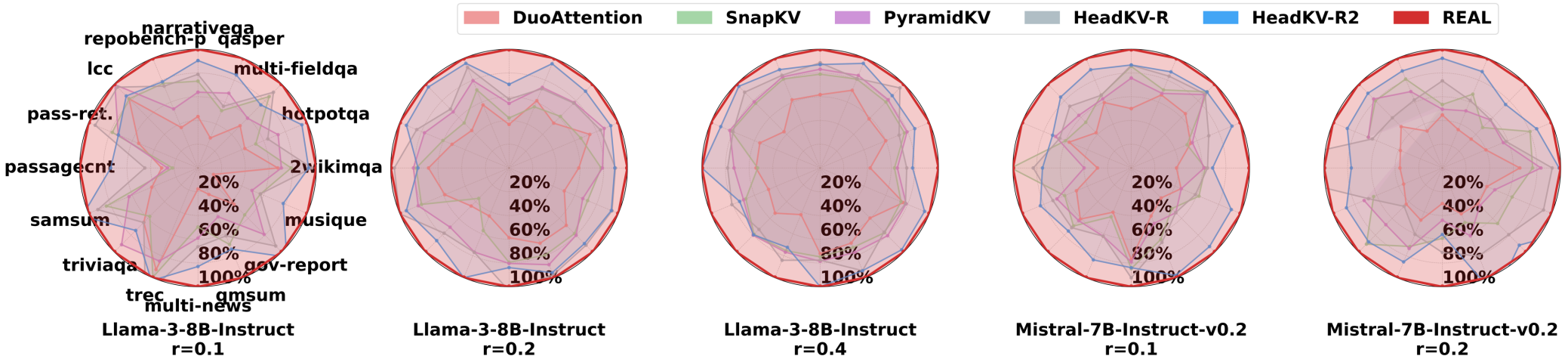
Experimental Setup

- **Token budget**
  - **Fixed token budgets** (e.g., 128, 4096, 8192 tokens) for **question-aware** compression.
  - Controlling the **cache ratios** (e.g., 10%, 20%, or 40% of the sequence length) for **question-agnostic** compression.

Evaluation Results: Main Results



Question-aware compression: ***REAL outperforms all baseline methods across token budgets (64-1024) on three models.***



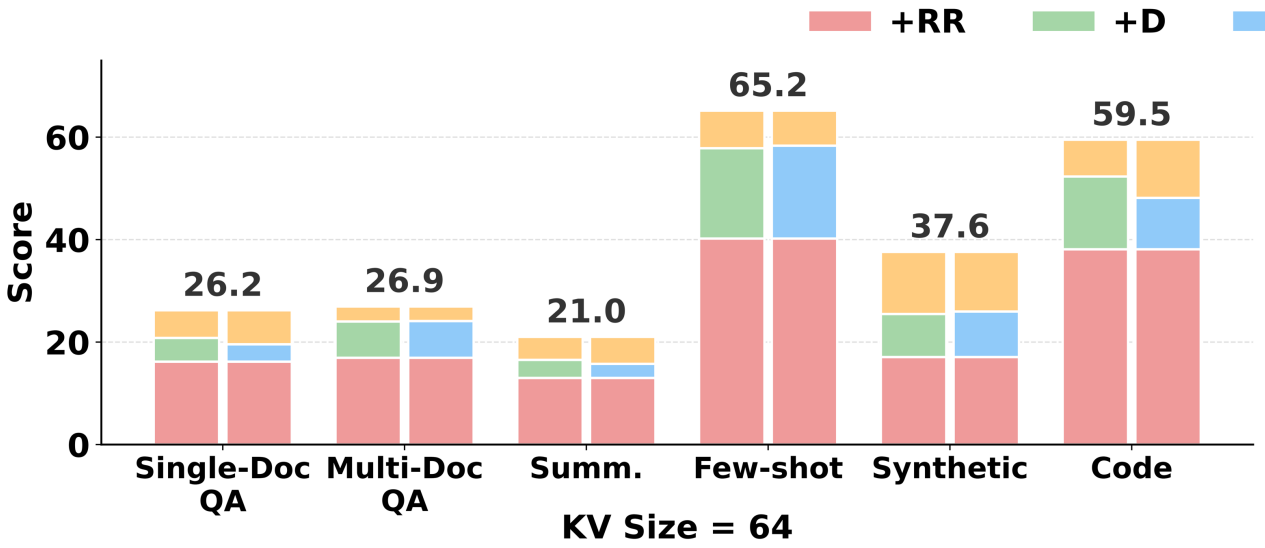
Question-agnostic compression: ***REAL outperforms all baseline methods across cache ratios (r=0.1, 0.2, 0.4) on Llama-3-8B-Instruct and Mistral-7B-Instruct.***

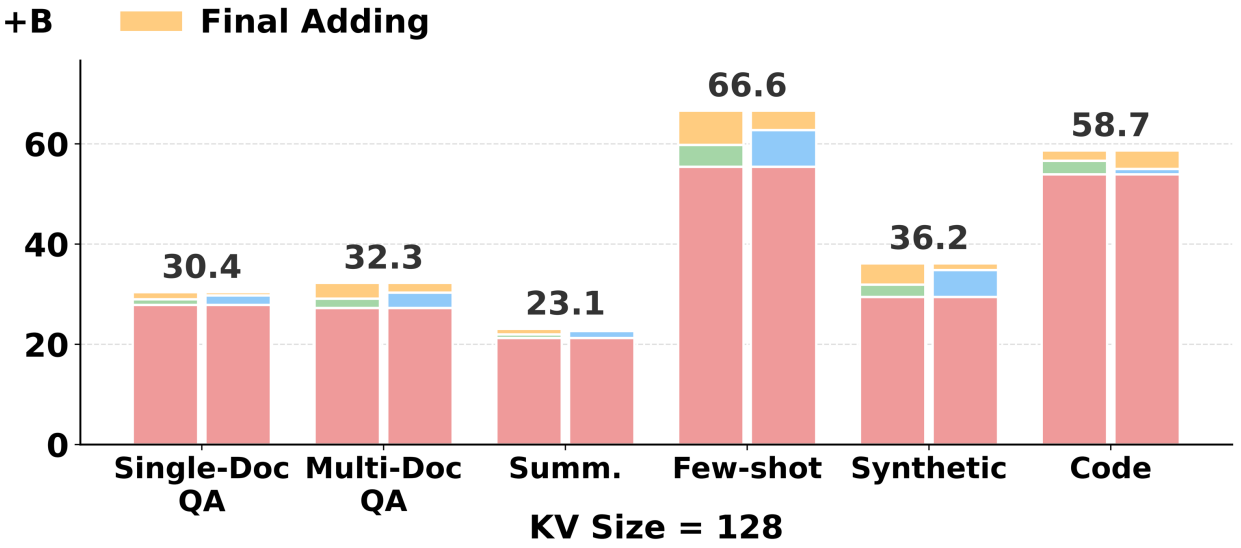
Evaluation Results: Catastrophe via Inverted Allocation

Logic	Inverted Prediction	Correct Answer
Entity	Poetry.	The Atlas Mountains.
Why?	Life is but a walking shadow.	Because she is in unrequited love with someone else.

Representative instances showing the model hallucinates.

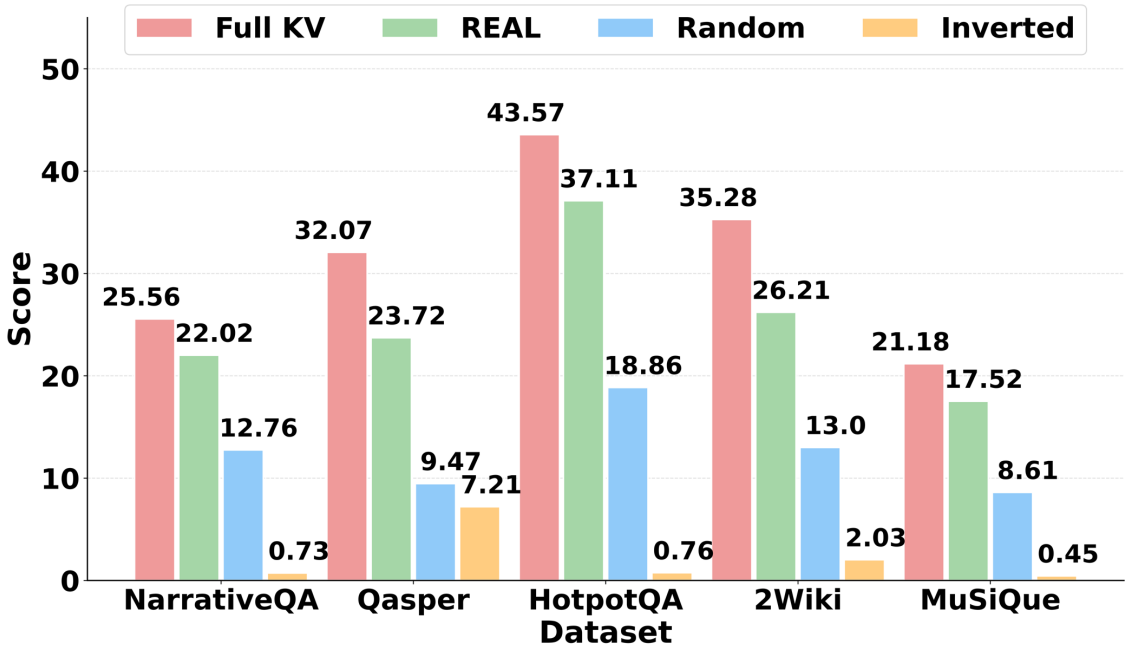
Evaluation Results: Ablation Study

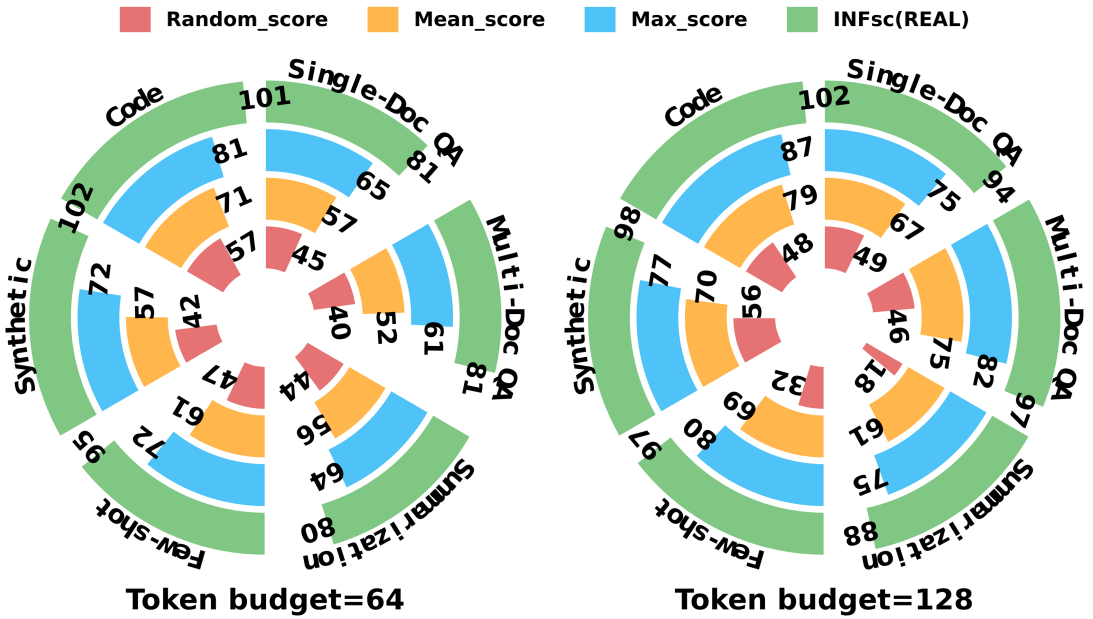




Attention behavior ablation: The adding performance validates the complementary information of the three components.

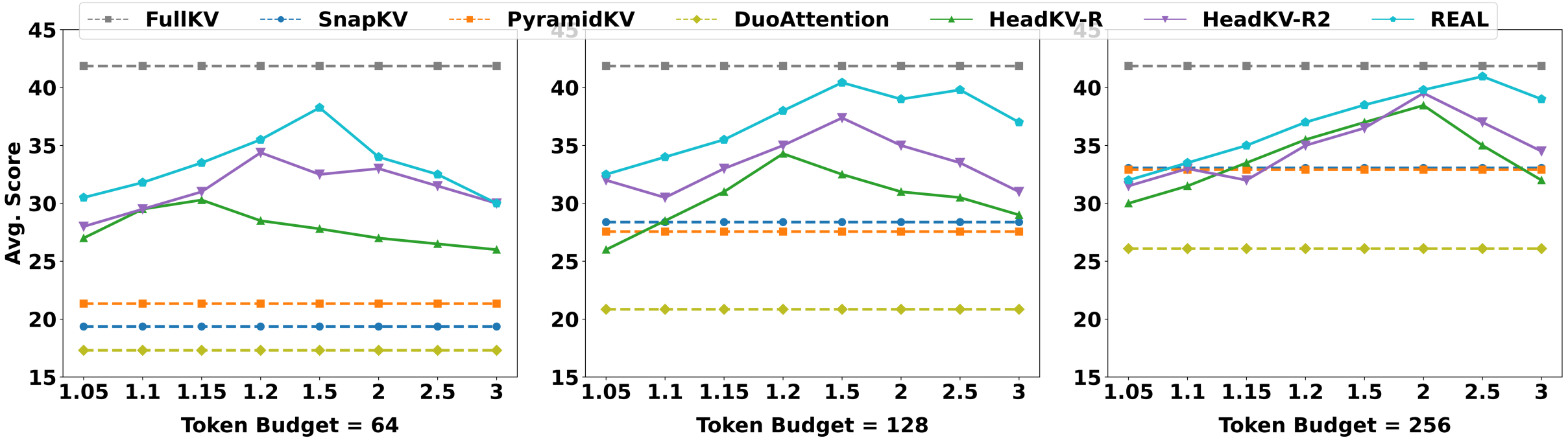
Evaluation Results: Ablation Study





Metric ablation: REAL shows efficient caching with awareness of both success and failure.

Evaluation Results: Hyperparameter Analysis



Hyperparameter sensitivity of the allocation parameter β across different token budgets on Llama-3-8B-Instruct.

Evaluation Results: Resource Analysis

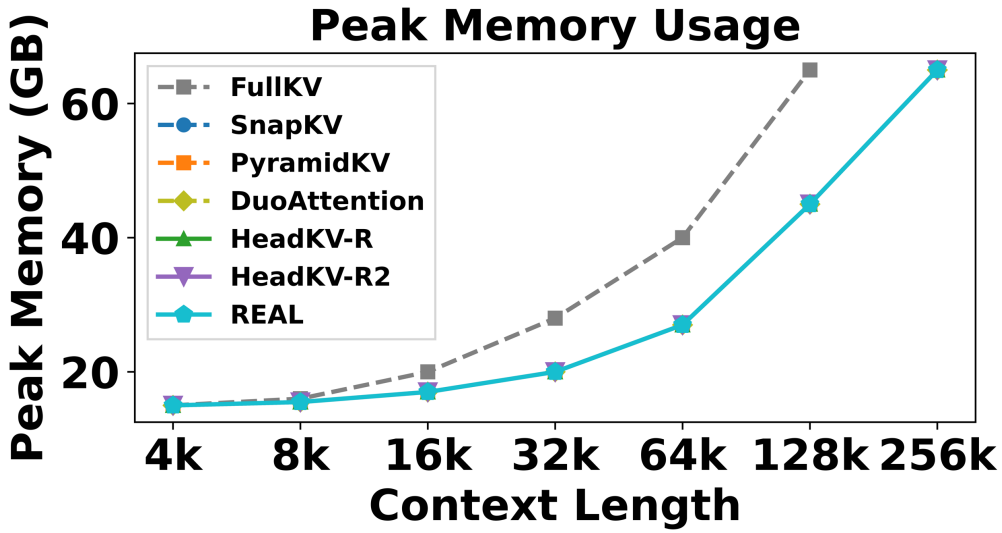
Method	Iter.1	Iter.2	Iter.3	Iter.4	Iter.5	Iter.6	Iter.7	Iter.8	Iter.9	Iter.10	Avg.
FullKV	4.39	4.72	4.18	4.63	4.29	4.51	4.37	4.89	4.11	4.46	4.45
DuoAttention	5.18	5.31	5.15	5.27	5.19	5.24	5.21	5.28	5.20	5.27	5.23
SnapKV	4.99	5.32	4.79	5.11	5.46	4.93	5.28	4.67	5.19	5.43	5.07
PyramidKV	4.78	5.34	4.91	5.26	4.97	5.18	4.85	5.11	4.93	5.69	5.02
HeadKV-R	4.73	4.95	4.81	4.92	4.68	4.90	4.79	4.99	4.84	4.88	4.86
HeadKV-R2	4.91	4.35	4.66	4.82	4.58	4.79	4.47	4.99	4.23	4.64	4.72
REAL	4.55	4.71	4.48	4.66	4.60	4.74	4.52	4.69	4.57	4.63	4.63

Latencies(/s) across 10 iterations for first-token generation on Mistral-7B-Instruct.

Evaluation Results: Resource Analysis

Model	Iter.1	Iter.2	Iter.3	Iter.4	Iter.5	Avg.
Llama-3-8B-Instruct	0.0247	0.0272	0.0258	0.0262	0.0263	0.02604
Mistral-7B-Instruct-v0.2	0.0270	0.0286	0.0267	0.0282	0.0288	0.02786

Latencies(/s) to build the attention behavior matrix: This one-time profiling cost is negligible.



REAL achieves comparable memory efficiency with baselines.

Conclusion

- Head-wise **successful and failed attention behavior** addresses the limitations of existing approaches.

- Dual strategy to allocate KV budget: **maximizing the signal-to-noise ratio** by strengthening valid reasoning pathways while inhibiting the noise arising from bias and distraction.

🔗 **Repository:** https://github.com/yonseicasl/real

📧 **Contact:** lemoji@yonsei.ac.kr