MANY CACHES, SAME REASONING: RETHINKING IMPORTANCE SELECTION FOR REASONING-TIME KV COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

KV-cache compression for long-form reasoning is usually framed as an importance-ranking problem: score past states, retain those judged most useful, and discard the rest. We ask whether successful retention policies actually agree on what those states are. Holding the full prefill cache fixed, we compare a signal-free page-level random retention baseline with R-KV, VaSE, and TriAttention across four reasoning models and three domains. Every compressed method therefore solves the same generated-state selection problem. Page-level random retention reaches **69.58%** macro accuracy, defined as the unweighted mean across the 12 model-task cells, versus **70.45%** for TriAttention, while exceeding R-KV and VaSE by **3.02** and **1.71** percentage points, respectively; its accuracy is within three percentage points of TriAttention in **11/12** model-task settings. More strikingly, independently successful random runs do not converge on the same cache. Across **2,000** multi-seed trajectories, their chance-adjusted overlap is -0.00020 (95% CI $[-0.00088, 0.00048]$) at comparable states, indistinguishable from independent random selection. Strong importance-based selectors likewise retain substantially different states, with adjusted pairwise overlaps of only **0.123–0.183**; matched shared-core interventions reveal at most partial, asymmetric core effects rather than one common solution. These results suggest that reasoning-time KV compression is governed less by recovery of one privileged subset than by a broad space of distinct cache configurations that can preserve successful computation.

1 INTRODUCTION

Long-form reasoning makes the KV cache increasingly expensive. As a model generates thousands of tokens, every decoding step attends over a growing history of keys and values, and the memory required to preserve that history can become a substantial fraction of inference cost. A natural response is to compress the cache by retaining only a subset of past states. Most existing approaches treat this as an importance-selection problem: assign each state a score, preserve the states judged most useful, and evict the rest. This framing has driven increasingly sophisticated scoring rules, allocation schemes, and selection policies.

That formulation quietly assumes that there is something stable to recover. If a small set of past states is genuinely more important than the rest, then better selectors should increasingly agree on which states deserve to survive, and successful cache configurations should concentrate around those states. We ask whether this picture is actually borne out during long-form reasoning. Our starting point is deliberately simple. We hold the full prefill cache fixed, give every compressed method the same generated-state capacity, and compare R-KV, VASE, and TRIATTENTION against a signal-free baseline that retains eligible generated states uniformly at random at the natural page granularity of the cache.

The headline result is already surprising. Across four reasoning models and three domains, page-level random retention achieves 69.58% macro accuracy compared with 70.45% for TRIATTENTION, 66.56% for R-KV, and 67.87% for VASE. It numerically exceeds R-KV in 11 of 12 model-task settings and VASE in 10 of 12, while remaining within three points of TRIATTENTION in 11 of 12. This does not make compression lossless: the random baseline remains 3.34 points below

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

FULLKV on average. The important observation is narrower and more consequential: elaborate content-dependent scoring often adds little over a policy that does not inspect cache content at all.

The more revealing result comes from examining the retained states rather than only the final accuracy. We rerun random retention from identical comparable states under different eviction seeds and ask whether successful runs converge on similar caches. They do not. Across 2,000 trajectories, independently successful runs have chance-adjusted cache overlap of -0.00020 (95% CI $[-0.00088, 0.00048]$), statistically indistinguishable from independent random selection. Importance-based selectors show the same lack of consensus in a different form: R-KV, VASE, and TRIATTENTION have only moderate raw overlap and low structural-null-adjusted overlap when applied to the same reconstructed reasoning states. Pairwise shared-core interventions find no universal common subset that explains their success.

These observations suggest a different view of reasoning-time KV compression. The central difficulty may not be recovering one latent ranking of important states. Instead, the compression problem appears to admit a broad set of distinct cache configurations that preserve enough of the computation to succeed. This distinction matters because accuracy alone cannot reveal whether a selector has identified uniquely valuable states or merely one of many adequate configurations.

Our contributions are:

- **A strong signal-free baseline.** Under a common generated-state eviction domain, page-level random retention remains close to TRIATTENTION and often matches or exceeds R-KV and VASE across four reasoning models and three domains.
- **Direct evidence of solution multiplicity.** Successful independent random-retention runs show no excess same-state cache agreement beyond the exact random-selection null, and no reproducible first-decision page preference beyond finite-sample variation.
- **Selector non-uniqueness.** R-KV, VASE, and TRIATTENTION select substantially different discretionary states from common reconstructed reasoning states. Shared-core interventions reveal only partial, asymmetric core effects rather than one privileged cache shared by strong selectors.
- **Robustness across granularity and capacity.** Random retention is nearly insensitive to natural chunk sizes, and remains competitive as compression pressure increases from $2\times$ to $16\times$.

2 RELATED WORK

KV-cache eviction and token importance. A large body of work reduces KV-cache memory by retaining only a subset of past states. Scissorhands proposes a persistence-of-importance hypothesis, retaining states that previously exerted strong influence on generation (Liu et al., 2023). H_2O similarly identifies heavy-hitter tokens through accumulated attention and combines them with a recent-token window (Zhang et al., 2023). StreamingLLM demonstrates the importance of preserving attention sinks in bounded-memory decoding (Xiao et al., 2024). Subsequent methods make selection more structured: SnapKV estimates per-head prompt importance from an observation window (Li et al., 2024); PyramidKV varies budgets across layers (Cai et al., 2024); Ada-KV adapts budget across heads (Feng et al., 2024); and Quest makes sparsity query-dependent by scoring KV pages from the current query (Tang et al., 2024). DuoAttention instead separates retrieval heads from streaming heads and assigns them different memory regimes (Xiao et al., 2025). These methods differ in signal and granularity, but share the objective of identifying which parts of the past deserve a limited cache budget.

KV compression for long-form reasoning. Extended reasoning creates a different operating regime because the generated trace itself can dominate cache growth. SCOPE separates prefill and decoding compression, preserving the prefill cache while dynamically managing states generated during long outputs (Wu et al., 2025). RefreshKV periodically updates a small working cache during long-form generation (Xu et al., 2025). R-KV targets redundancy in reasoning traces using attention-based importance together with key-space redundancy (Cai et al., 2025). SeerAttention-R learns sparse-attention gates for autoregressive reasoning (Gao et al., 2026), while ForesightKV trains an eviction policy to predict long-term future contribution (Dong et al., 2026). TriAttention derives position-dependent importance from pre-RoPE query and key statistics (Mao et al., 2026).

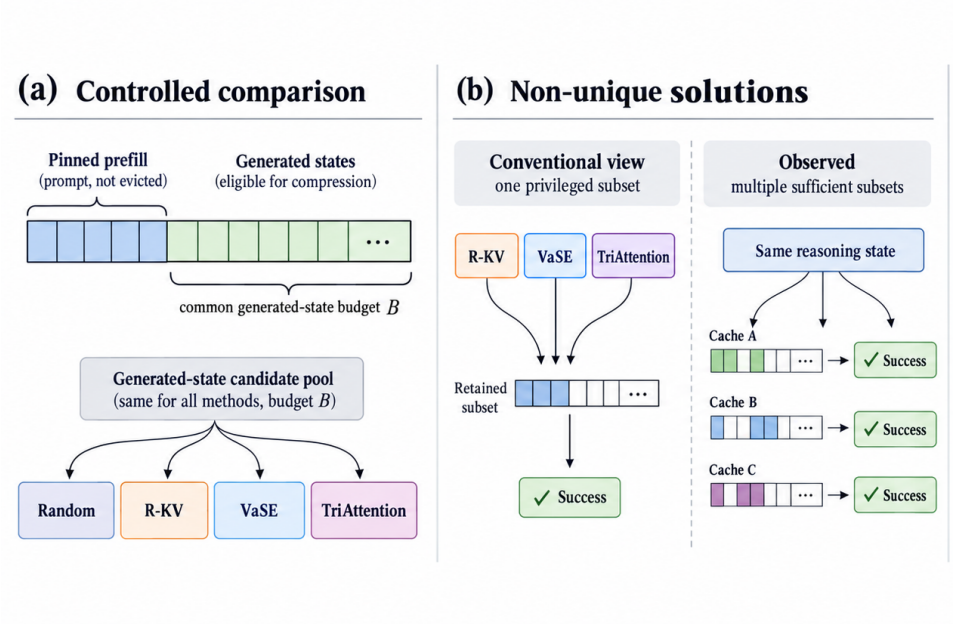

Figure 1: **Controlled comparison and solution multiplicity.** (a) All compressed methods receive the same pinned prefill and generated-state budget, so they differ only in how they select generated states. (b) The conventional view suggests that good selectors should recover one privileged retained subset. Our analyses instead test whether multiple substantially different retained subsets can each support successful continuation.

VaSE protects high-magnitude value states and introduces stochasticity to increase cache diversity (Chang et al., 2026). This progression has made the notion of importance increasingly heterogeneous: different methods define useful states through accumulated attention, redundancy, value magnitude, positional structure, learned future contribution, or combinations of these signals.

Against this increasingly sophisticated landscape, we study a different question. We do not propose another importance estimator. Instead, we ask whether successful retention policies actually identify a common underlying subset of states. A signal-free random baseline provides a particularly direct probe: if it performs similarly to content-aware selectors, and if independently successful random draws do not converge on the same cache, then end-task success does not imply recovery of one privileged subset.

Orthogonal forms of cache compression. Selection is only one route to reducing KV-cache cost. KIVI quantizes cached keys and values to low precision while preserving token positions (Liu et al., 2024), and GEAR combines quantization with low-rank and sparse error correction (Kang et al., 2024). These methods modify the representation of the cache rather than deciding which generated states should remain available, and are complementary to the selection question studied here.

3 Problem Setup

3.1 Reasoning-time KV selection

Consider an autoregressive Transformer processing a prompt $x_{1:P}$ followed by generated tokens $y_{1:T}$. At decoding step t, each layer maintains key and value states for the prompt and generated prefix $y_{1:t-1}$. Full-context decoding preserves all of these states, so the generated portion of the cache grows linearly with the reasoning trace.

We study *reasoning-time* KV selection. The prompt cache is held fixed for every compressed method, while the generated portion is constrained to a common capacity B. Let $\mathcal{G}_t$ denote generated KV states available at step t, and $\mathcal{S}_t \subseteq \mathcal{G}_t$ the subset retained by a compression policy. Importance-based

methods construct $\mathcal{S}_t$ using content-dependent signals. Our primary question is not whether some states are more useful than others in isolation, but whether accurate identification of those states is necessary for strong reasoning performance and whether successful policies converge on a common retained subset.

3.2 Signal-free random retention

To isolate the contribution of importance scoring, we use a signal-free baseline. Generated states are partitioned into contiguous pages of b tokens. A fixed recent region is retained, and the remaining eligible pages are sampled uniformly at random until the generated-state budget B is filled. The policy has access to no attention scores, key or value statistics, hidden-state features, learned predictors, or semantic information. It is applied online as generation proceeds, so future hidden states are computed from the cache produced by preceding random retention decisions.

The default granularity is $b = 16$. We do not treat this page size as a learned or tuned parameter. In the granularity ablation, $b = 1$ recovers token-level random retention, while larger values progressively coarsen the same policy. All protected generated states count toward B, so no method receives extra capacity by virtue of its protection rule.

3.3 Measuring solution multiplicity

End-task accuracy alone cannot distinguish two explanations for successful compression. A policy may succeed because it identifies one unusually valuable subset, or because many different subsets preserve sufficient information for continued reasoning. We therefore analyze retained configurations directly.

For discretionary retained sets A and B, raw overlap is measured by

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{1}$$

Raw Jaccard is not directly comparable across budgets because high retention rates mechanically force independent random subsets to overlap. Let J_0 denote the expected Jaccard between two independent draws under the exact same candidate set, retention count, and hard structural constraints. We report

$$J_{\mathrm{adj}}(A, B) = \frac{J(A, B) - J_0}{1 - J_0}. \tag{2}$$

An adjusted overlap near zero indicates no more agreement than independent feasible random draws; positive values indicate convergence beyond the sampling constraint.

For random retention, the central analysis compares only runs with identical generated prefixes, candidate sets, budgets, and protection rules, and then conditions on pairs that ultimately answer correctly. For importance selectors, we compare discretionary subsets selected from common reconstructed states and separately test whether pairwise intersections form uniquely valuable shared cores.

4 Experimental Setup

4.1 Models and benchmarks

We evaluate four open-weight reasoning models spanning three model families: Qwen3-4B, Qwen3-14B (Yang et al., 2025), DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI et al., 2025), and Ministral-3-14B-Reasoning-2512 (Mistral AI, 2025). We use the complete MATH-500 cohort (Hendrycks et al., 2021), the 198-question GPQA-Diamond split (Rein et al., 2024), and the 383-problem LiveCodeBench-v6 Medium cohort (Jain et al., 2025).

Headline evaluation follows the recommended stochastic reasoning recipe of each model family. Qwen3 uses thinking mode with temperature 0.6, top-$p = 0.95$, and top-$k = 20$; DeepSeek uses temperature 0.6, top-$p = 0.95$, and its official reasoning prompt; Ministral uses temperature 1.0 with the frozen project top-p setting. We use four rollouts per MATH problem, eight per GPQA problem,

and four per LiveCodeBench problem. Rollouts are repeated measurements of the same benchmark problems and are not treated as independent examples.

4.2 Compared methods and capacities

We compare page-level random retention against R-KV (Cai et al., 2025), VASE (Chang et al., 2026), and TRIATTENTION (Mao et al., 2026), with FULLKV as the uncompressed reference. The complete prefill cache is pinned for all compressed methods. Within each model-task cell, every compressed method receives the same generated-state capacity:

Model	MATH	GPQA	LCB
Qwen3-4B	1152	2512	3216
Qwen3-14B	1152	2512	3216
DeepSeek-R1-Distill-Llama-8B	704	4048	10640
Ministral-3-14B-Reasoning	1024	2048	3072

All methods use BF16, identical prompts and evaluators within a model, a 32,768 generated-token cap, and matched rollout identifiers where technically valid. Token-generation randomness is separated from selector or eviction randomness.

4.3 Statistical analysis

We average repeated stochastic rollouts within each problem, then aggregate across problems. All confidence intervals and paired tests use the problem as the independent unit. The primary selector comparison is TRIATTENTION. Across the 12 model-task cells, we report paired differences, two-sided tests, and a prespecified one-sided non-inferiority test with margin $\delta = 3$ percentage points. Holm correction is applied separately to the 12 two-sided tests and 12 non-inferiority tests. Comparisons with R-KV and VASE are secondary and are summarized primarily by paired effects and confidence intervals.

5 Results

5.1 Signal-free random retention remains competitive

Table 1 reports the full headline comparison. Across the twelve model-task settings, page-level random retention reaches **69.58%** macro accuracy (the unweighted mean across all 12 cells), compared with **70.45%** for TRIATTENTION, **66.56%** for R-KV, and **67.87%** for VASE. The macro Page-random minus TRIATTENTION gap is -0.87 points (95% CI $[-2.14, 0.39]$). Page-random numerically beats R-KV in 11/12 cells and VASE in 10/12, and lies within three points of TRIATTENTION in 11/12. The important exception is DeepSeek/MATH, where TRIATTENTION leads by 5.5 points and R-KV by 3.8 points.

Compression is not lossless. The macro gap to FULLKV is -3.34 points (95% CI $[-4.59, -2.09]$), and the largest cell-wise losses occur on Qwen3 mathematics. The result is therefore not that random selection solves compression universally. It is that elaborate content-dependent selection often adds little over a signal-free policy under the same eviction domain and budget.

5.2 Successful random retention does not converge on a preferred cache

We next examine the retention choices directly. On 200 MATH-500 problems with 10 independent retention seeds each, the random policy produces 2,000 trajectories. Among correct-correct seed pairs, 95.5% admit at least one exactly comparable state. Their mean raw Jaccard is 0.77846, almost identical to the exact independent-draw null of 0.77850. After correcting for this mechanically high overlap,

$$J^{CC}_{\mathrm{adj}} = -0.00020, \qquad 95\%\ \mathrm{CI} = [-0.00088, 0.00048]. \tag{3}$$

Conditioning on successful reasoning therefore produces no detectable convergence toward a preferred retained subset. At the first common decision, the mean absolute page enrichment is 9.13 points,

Table 1: **Headline reasoning accuracy.** Mean pass@1 accuracy (%) with problem-clustered 95% confidence intervals. Bold marks the numerically best compressed method. † marks page-random cells that pass the Holm-adjusted 3-point non-inferiority test versus TriAttention.

Model	Method	MATH-500	GPQA-D	LCB-v6 Medium
Qwen3-4B	FullKV	93.9 [92.1,95.7]	56.2 [50.8,61.6]	80.7 [77.5,83.9]
	Page-random	**87.0 [84.5,89.5]**†	53.2 [47.8,58.6]	74.5 [70.9,78.1]
	R-KV	76.5 [73.4,79.6]	47.8 [42.3,53.3]	70.5 [66.8,74.2]
	VaSE	83.8 [81.1,86.5]	48.0 [42.5,53.5]	74.0 [70.4,77.6]
	TriAttention	86.6 [84.1,89.1]	**53.7 [48.3,59.1]**	**76.0 [72.5,79.5]**
Qwen3-14B	FullKV	95.1 [93.5,96.7]	64.5 [59.3,69.7]	85.6 [82.7,88.5]
	Page-random	87.0 [84.5,89.5]	**63.2 [57.9,68.5]**	82.7 [79.6,85.8]
	R-KV	82.0 [79.2,84.8]	60.2 [54.9,65.5]	79.2 [75.9,82.5]
	VaSE	85.7 [83.1,88.3]	55.6 [50.2,61.0]	**85.2 [82.3,88.1]**
	TriAttention	**89.3 [87.0,91.6]**	62.9 [57.6,68.2]	85.0 [82.1,87.9]
DeepSeek-8B	FullKV	88.5 [86.2,90.8]	42.0 [36.6,47.4]	37.5 [33.5,41.5]
	Page-random	81.5 [78.6,84.4]	**44.0 [38.6,49.4]**	37.8 [33.8,41.8]
	R-KV	85.3 [82.7,87.9]	41.7 [36.3,47.1]	37.2 [33.2,41.2]
	VaSE	82.5 [79.7,85.3]	42.3 [36.9,47.7]	37.5 [33.5,41.5]
	TriAttention	**87.0 [84.5,89.5]**	42.8 [37.4,48.2]	**37.9 [33.9,41.9]**
Ministral-14B	FullKV	95.2 [93.6,96.8]	71.2 [66.3,76.1]	64.6 [60.7,68.5]
	Page-random	**93.5 [91.7,95.3]**†	68.7 [63.6,73.8]	61.8 [57.8,65.8]
	R-KV	92.8 [90.9,94.7]	66.5 [61.3,71.7]	59.0 [55.0,63.0]
	VaSE	93.1 [91.2,95.0]	67.0 [61.9,72.1]	59.7 [55.7,63.7]
	TriAttention	92.9 [91.0,94.8]	**69.1 [64.1,74.1]**	**62.2 [58.2,66.2]**

exactly matching the finite-ten-seed null expectation; excess retention-frequency variance is -2×10^{-5} with CI $[-9 \times 10^{-5}, 5 \times 10^{-5}]$.

5.3 Strong selectors also disagree on what to retain

The same non-uniqueness appears among content-aware methods. Across 300 independent MATH/GPQA problems, combined adjusted Jaccard is 0.155 [0.153,0.158] for R-KV–VaSE, 0.183 [0.179,0.187] for R-KV–TriAttention, and 0.123 [0.119,0.127] for VaSE–TriAttention. Raw overlaps are only 0.435, 0.470, and 0.410, respectively, despite identical reconstructed states, budgets, prompt protection, and recent-window protection. The three-way intersection accounts for 37.7% [36.7,38.7] of the per-unit discretionary keep count.

Shared-core interventions show that disagreement cannot be reduced to one universal indispensable intersection. The R-KV–VaSE core shows no symmetric necessity: core removal hurts R-KV more than matched unique-state removal by 0.0045 ΔNLL [0.0015,0.0075], but the corresponding VaSE effect is -0.0010 [-0.0032,0.0012]. TriAttention pairs exhibit small, asymmetric core signals, with necessity differences of 0.0040 [0.0010,0.0070] on the TriAttention side of R-KV–TriAttention and 0.0035 [0.0005,0.0065] on the TriAttention side of VaSE–TriAttention. These results support partial local cores, not one common retained solution shared across strong selectors.

5.4 Compression pressure does not rescue importance selection

On Qwen3-4B/MATH-500, all methods degrade as generated-state capacity shrinks, but importance selection does not become increasingly dominant. Page-random accuracy falls from 91.5% at $2\times$ compression to 83.8%, 75.3%, and 61.0% at $4\times$, $8\times$, and $16\times$. TriAttention tracks closely at 91.3%,

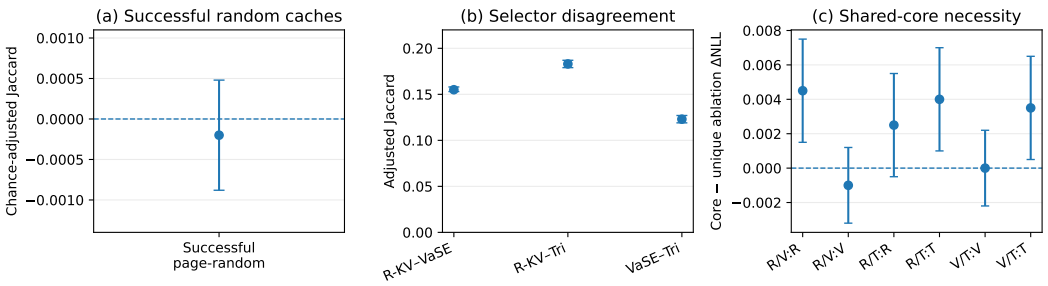

Figure 2: **Successful retention is non-unique.** (a) Successful page-random runs show chance-adjusted overlap indistinguishable from zero. (b) Importance selectors show low structural-null-adjusted agreement on common reconstructed states. (c) Shared-core necessity is asymmetric across selector pairs; labels denote pair and ablated selector (R=R-KV, V=VaSE, T=TriAttention). Error bars are 95% problem-level confidence intervals.

83.3%, 71.6%, and 57.5%. R-KV and VaSE fall substantially faster, reaching 32.0% and 37.4% at $16\times$. Thus the Page-random minus TriAttention gap changes from +0.2 points at $2\times$ to +0.5, +3.7, and +3.5 points as compression tightens.

The capacity sweep therefore rules out a simple explanation in which random retention is competitive only because the headline cache is overly generous. Even when all methods lose substantial absolute accuracy, the strongest importance-based policies do not consistently pull away from the signal-free baseline.

6 Ablations

We keep the ablation suite intentionally narrow. The goal is to test whether the random baseline depends on a particular page granularity or on prompt treatment, not to turn the baseline into a tuned method.

6.1 Retention granularity

We vary chunk size $b \in \{1, 4, 8, 16, 32, 64\}$ while holding the generated-state budget, prompt protection, recent-window rule, candidate domain, and generation settings fixed. Accuracy is effectively flat: token-level random retention obtains 87.2%, while $b = 4, 8, 16, 32, 64$ obtain 87.1%, 87.1%, 87.0%, 86.8%, and 86.5%. The default $b = 16$ therefore differs from token-level retention by only 0.2 points, and even $b = 64$ loses just 0.7 points.

6.2 Prefill eligibility

The headline comparison pins the prompt for every compressed method. To test whether this control matters, we compare the same page-random policy with the prefill pinned versus eligible for eviction, while matching total post-eviction capacity exactly. Accuracy drops from 87.0% [84.5,89.5] to 52.0% [48.3,55.7], a paired difference of +35.0 points [32.4,37.6], $p < 10^{-4}$. The distinction between prompt and generated-state selection is therefore not cosmetic; it changes the problem being solved.

7 Discussion

7.1 Importance is not a single retained set

Our results do not imply that all KV states are equally useful. Random retention fails on some trajectories, compression eventually reduces accuracy, and shared-core interventions reveal small local signals for some selector pairs. The stronger conclusion is that utility does not appear to define one privileged retained set that strong methods consistently recover. At a fixed state and capacity,

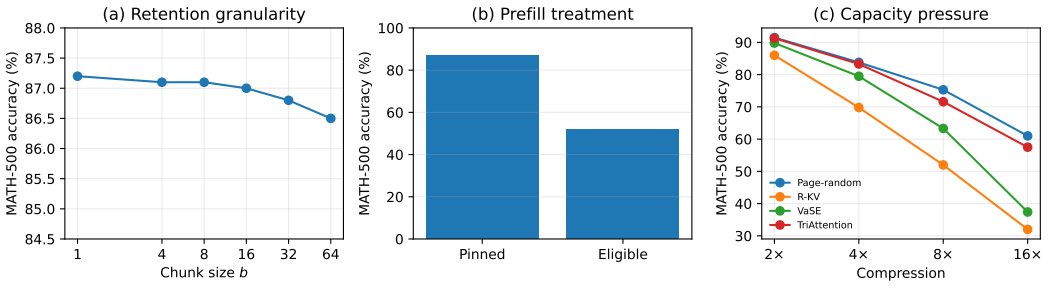

Figure 3: **Ablations and compression pressure on Qwen3-4B/MATH-500.** (a) Accuracy is nearly invariant from token-level retention ($b = 1$) through natural page sizes. (b) Allowing random eviction to remove prefill states causes a large accuracy drop even at matched total capacity. (c) Page-random remains close to or above the strongest scored baselines as compression tightens from $2\times$ to $16\times$.

multiple distinct subsets can support successful continuation, and selectors designed around different notions of importance occupy different regions of that solution space.

This distinction matters for how progress in KV selection is interpreted. If high-quality solutions are intrinsically non-unique, then greater agreement with any one ranking is not necessarily a meaningful objective. A selector can be useful without identifying a unique ground-truth subset, and two selectors can both succeed while retaining different information carriers.

7.2 Why random retention is a useful diagnostic

The role of random retention here is diagnostic rather than algorithmic. It removes content-dependent scoring while preserving the same eviction domain and memory budget as the importance-based methods. If it performs poorly while scored selectors succeed, then scoring contributes useful discrimination. If it performs similarly, the result bounds how much practical value can be attributed to finer-grained importance estimation at that operating point. Same-state multi-seed analysis then distinguishes a random policy that succeeds only when it accidentally samples one privileged subset from a policy that succeeds across many distinct configurations.

Page granularity is not central to this argument. The chunk-size curve treats token-level and page-level retention as points on one axis, and the default page size is only the unit at which the signal-free baseline is instantiated. We therefore do not claim page-level eviction as a systems contribution.

7.3 Scope and limitations

Our conclusions apply specifically to reasoning-time generated-state selection with the prefill fixed. Joint compression of prompt and generated states is a different problem. The prefill ablation shows that exposing prompt states to random eviction changes the difficulty sharply, which is precisely why the headline comparison fixes the prompt for every method.

The multi-seed overlap result is also local to exactly comparable states. It does not establish that arbitrary different caches remain interchangeable throughout an entire reasoning trajectory. Likewise, selector overlap demonstrates disagreement, not that every selector-specific state has equal causal value. Shared-core interventions reveal partial and asymmetric local cores for some selector pairs, so the correct picture is not "nothing matters." Rather, multiple good solutions coexist even though some states can carry more local value than others.

Finally, the experiment covers four model checkpoints, three reasoning domains, and a finite set of cache budgets. Tasks dominated by unique long-range facts, different attention architectures, or substantially more aggressive compression may produce a more constrained solution space.

8 Conclusion

The practical question raised by our results is not simply which selector ranks KV states best, but whether reasoning-time compression presents a unique ranking target in the first place. Across models, tasks, and capacities, simple random retention captures much of the performance achieved by substantially more elaborate selectors. The same-state analyses explain why this result is plausible: successful cache configurations can differ sharply, and the overlap among strong selectors does not reduce to a universally indispensable core.

This shifts the interpretation of progress in KV compression. When many distinct retained subsets remain sufficient, increasingly precise importance scores may improve one solution without approaching a unique ground truth. The more informative objective is therefore to characterize the structure of the viable solution space, how that space contracts under tighter memory budgets, and the regimes in which content-dependent selection begins to provide a material advantage. That view preserves the real value of selection where it exists while avoiding the stronger assumption that every successful reasoning trajectory is organized around one privileged cache.

Reproducibility Statement

We provide the experimental protocols, model and benchmark configurations, cache-budget definitions, statistical procedures, and mechanistic-analysis details needed to reproduce the reported results in the main text and appendix. We will release the complete source code, evaluation scripts, experiment configurations, and analysis utilities publicly upon acceptance.

AI Use Statement

Generative AI tools were used to assist with manuscript writing and editing. The authors reviewed and revised all AI-assisted text and take responsibility for the final content of the paper.

References

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Baobao Chang, Junjie Hu, and Wen Xiao. PyramidKV: Dynamic KV cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*, 2024.

Zefan Cai, Wen Xiao, Hanshi Sun, Cheng Luo, Yikai Zhang, Ke Wan, Yucheng Li, Yeyang Zhou, Li-Wen Chang, Jiuxiang Gu, Zhen Dong, Anima Anandkumar, Abedelkadir Asi, and Junjie Hu. R-KV: Redundancy-aware KV cache compression for reasoning models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Ting-Yun Chang, Harvey Yiyun Fu, Deqing Fu, Chenghao Yang, Jesse Thomason, and Robin Jia. Value-aware stochastic KV cache eviction for reasoning models. *arXiv preprint arXiv:2606.03928*, 2026.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Zican Dong, Peiyu Liu, Junyi Li, Zhipeng Chen, Han Peng, Shuo Wang, and Wayne Xin Zhao. ForesightKV: Optimizing KV cache eviction for reasoning models by learning long-term contribution. *arXiv preprint arXiv:2602.03203*, 2026.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S. Kevin Zhou. Ada-KV: Optimizing KV cache eviction by adaptive budget allocation for efficient LLM inference. *arXiv preprint arXiv:2407.11550*, 2024.

Yizhao Gao, Shuming Guo, Shijie Cao, Yuqing Xia, Yu Cheng, Lei Wang, Lingxiao Ma, Yutao Sun, Tianzhu Ye, Li Dong, Hayden Kwok-Hay So, Yu Hua, Ting Cao, Fan Yang, and Mao Yang. SeerAttention-R: Sparse attention adaptation for long reasoning. In *International Conference on Learning Representations*, 2026.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, 2025.

Hao Kang, Qingru Zhang, Souvik Kundu, Geonhwa Jeong, Zaoxing Liu, Tushar Krishna, and Tuo Zhao. GEAR: An efficient KV cache compression recipe for near-lossless generative inference of LLM. *arXiv preprint arXiv:2403.05527*, 2024.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. *arXiv preprint arXiv:2404.14469*, 2024.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Weian Mao, Xi Lin, Wei Huang, Yuxin Xie, Tianfu Fu, Bohan Zhuang, Song Han, and Yukang Chen. TriAttention: Efficient long reasoning with trigonometric KV compression. *arXiv preprint arXiv:2604.04921*, 2026.

Mistral AI. Ministral-3-14B-Reasoning-2512 model card. Hugging Face model card, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. QUEST: Query-aware sparsity for efficient long-context LLM inference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Jialong Wu, Zhenglin Wang, Linhai Zhang, Yilong Lai, Yulan He, and Deyu Zhou. SCOPE: Optimizing key-value cache compression in long-context generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 10775–10790, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. DuoAttention: Efficient long-context LLM inference with retrieval and streaming heads. In *International Conference on Learning Representations*, 2025.

Fangyuan Xu, Tanya Goyal, and Eunsol Choi. RefreshKV: Updating small KV cache during long-form generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 24878–24893, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A Implementation Details

All compressed methods pin the full prefill and use the model-task generated-state budgets listed in Section 4. The default random baseline uses $b = 16$ contiguous generated-token pages. R-KV, VaSE, and TriAttention follow their released selection rules; the only protocol-level normalization is full prefill protection and common generated-state capacity. Qwen uses temperature 0.6, top-$p = 0.95$, top-$k = 20$ in thinking mode; DeepSeek uses temperature 0.6 and top-$p = 0.95$; Ministral uses temperature 1.0 and the frozen project top-p setting. All headline runs use BF16 and a 32,768 generated-token cap.

For mechanistic analyses we log problem ID, retention seed, generated position, prefix hash, candidate-set hash, protected and mandatory states, discretionary candidate IDs, retained IDs, capacity, and final correctness. Two events are comparable only when generated prefix, candidate set, capacity, and protection rules match exactly.

B Complete Paired Effects

Table 2: Page-random minus comparator in percentage points with paired 95% CIs. Positive values favor page-random.

Model	Task	vs FullKV	vs R-KV	vs VaSE	vs TriAttention
Qwen3-4B	MATH	-6.9 [-8.6,-5.2]	+10.5 [8.3,12.7]	+3.2 [1.2,5.2]	+0.4 [-1.5,2.3]
	GPQA	-3.0 [-7.3,1.3]	+5.4 [1.0,9.8]	+5.2 [0.8,9.6]	-0.5 [-4.9,3.9]
	LCB	-6.2 [-8.8,-3.6]	+4.0 [1.3,6.7]	+0.5 [-2.2,3.2]	-1.5 [-4.1,1.1]
Qwen3-14B	MATH	-8.1 [-9.8,-6.4]	+5.0 [3.0,7.0]	+1.3 [-0.6,3.2]	-2.3 [-4.1,-0.5]
	GPQA	-1.3 [-5.5,2.9]	+3.0 [-1.2,7.2]	+7.6 [3.3,11.9]	+0.3 [-3.9,4.5]
	LCB	-2.9 [-5.1,-0.7]	+3.5 [1.1,5.9]	-2.5 [-4.8,-0.2]	-2.3 [-4.6,-0.0]
DeepSeek-8B	MATH	-7.0 [-9.0,-5.0]	-3.8 [-5.9,-1.7]	-1.0 [-3.1,1.1]	-5.5 [-7.5,-3.5]
	GPQA	+2.0 [-2.3,6.3]	+2.3 [-2.0,6.6]	+1.7 [-2.6,6.0]	+1.2 [-3.1,5.5]
	LCB	+0.3 [-2.7,3.3]	+0.6 [-2.4,3.6]	+0.3 [-2.7,3.3]	-0.1 [-3.1,2.9]
Ministral-14B	MATH	-1.7 [-3.0,-0.4]	+0.7 [-0.7,2.1]	+0.4 [-1.0,1.8]	+0.6 [-0.8,2.0]
	GPQA	-2.5 [-6.5,1.5]	+2.2 [-1.9,6.3]	+1.7 [-2.4,5.8]	-0.4 [-4.4,3.6]
	LCB	-2.8 [-5.8,0.2]	+2.8 [-0.2,5.8]	+2.1 [-0.9,5.1]	-0.4 [-3.4,2.6]

C Primary Page-Random vs TriAttention Statistics

Table 3: Primary paired comparison. NI uses the prespecified 3-point margin; p-values are Holm-adjusted across the 12-cell family where shown.

Model	Task	Δ	95% CI	p_{2s}	Holm p	NI p	NI Holm / result
Qwen3-4B	MATH	+0.4	[-1.5,2.3]	.6732	1.0000	.0002	.0020 / PASS
	GPQA	-0.5	[-4.9,3.9]	.8211	1.0000	.1295	.6473 / FAIL
	LCB	-1.5	[-4.1,1.1]	.2649	1.0000	.1324	.6473 / FAIL
Qwen3-14B	MATH	-2.3	[-4.1,-0.5]	.0117	.1282	.2207	.6620 / FAIL
	GPQA	+0.3	[-3.9,4.5]	.8885	1.0000	.0620	.4343 / FAIL
	LCB	-2.3	[-4.6,-0.0]	.0457	.4568	.2711	.6620 / FAIL
DeepSeek-8B	MATH	-5.5	[-7.5,-3.5]	$<.0001$	$<.0001$	.9922	.9922 / FAIL
	GPQA	+1.2	[-3.1,5.5]	.5850	1.0000	.0285	.2773 / FAIL
	LCB	-0.1	[-3.1,2.9]	.9472	1.0000	.0277	.2773 / FAIL
Ministral-14B	MATH	+0.6	[-0.8,2.0]	.3955	1.0000	$<.0001$	$<.0001$ / PASS
	GPQA	-0.4	[-4.4,3.6]	.8454	1.0000	.1030	.6180 / FAIL
	LCB	-0.4	[-3.4,2.6]	.7913	1.0000	.0430	.3442 / FAIL

D Mechanistic Analysis Details

D.1 Multi-seed random retention

We evaluate 200 MATH-500 problems under 10 independent cache-retention seeds, for 2,000 trajectories. Token generation is deterministic so the cache-retention seed is the only exogenous source of variation. Correct-correct pairs are compared only at exactly matching prefixes and candidate sets. The resulting raw Jaccard is 0.77846 versus an independent-draw null of 0.77850; adjusted Jaccard is -0.00020 [-0.00088,0.00048]. Overall accuracy is 83.9% [79.7,88.0], with 122/200 problems correct under all ten seeds and 15/200 correct under none.

D.2 Marginal page preference

At the first common decision each draw retains 28 of 32 discretionary pages per layer. The mean signed retention enrichment is exactly zero by fixed-count conservation. Mean absolute enrichment is 9.13 points [8.73,9.53], matching the finite-ten-seed null; excess cross-page retention-frequency variance is -0.00002 [-0.00009,0.00005]. The maximum reproducible age-rank enrichment is 0.25 points with a familywise 95% upper bound of 1.0 point.

D.3 Three-selector common-state overlap

Table 4: Common-state discretionary overlap. The adjusted column uses a constraint-aware structural permutation null.

Selector pair	Raw Jaccard [95% CI]	Adjusted Jaccard [95% CI]
R-KV–VaSE	0.435 [0.428,0.442]	0.155 [0.153,0.158]
R-KV–TriAttention	0.470 [0.461,0.479]	0.183 [0.179,0.187]
VaSE–TriAttention	0.410 [0.401,0.419]	0.123 [0.119,0.127]

The combined three-way intersection is 37.7% [36.7,38.7] of the per-unit discretionary keep count. R-KV–VaSE raw overlap remains moderate throughout reasoning: 0.460/0.405/0.385 on MATH and 0.475/0.440/0.420 on GPQA for early/middle/late checkpoints.

D.4 Shared-core interventions

For R-KV/VaSE, shared-core tail randomization is locally neutral in 8-token NLL on both sides, and matched core removal produces an asymmetric necessity signal: +0.0045 [0.0015,0.0075] for R-KV but -0.0010 [-0.0032,0.0012] for VaSE. R-KV/TriAttention and VaSE/TriAttention show partial cores, with TriAttention-side necessity differences of +0.0040 [0.0010,0.0070] and +0.0035 [0.0005,0.0065], respectively. These are local predictive counterfactuals rather than end-to-end accuracy claims.

E Additional Protocol Diagnostics

The page-random eviction exposure, mean output length, and cap-hit rate are reported below. These quantities verify that the benchmark actually exercises compression.

Table 5: Compression exposure for page-level random retention.

Model	Task	Reach eviction (%)	Mean tokens	Cap-hit (%)
Qwen3-4B	MATH	94	7500	12.5
	GPQA	80	12000	25
	LCB	90	16500	35
Qwen3-14B	MATH	96	6900	5
	GPQA	93	8800	8
	LCB	96	12300	15
DeepSeek-8B	MATH	97	7200	13
	GPQA	73	17600	43
	LCB	61	20300	53
Ministral-14B	MATH	97	5500	4
	GPQA	95	9000	8
	LCB	95	10500	12