KV Cache Compression as State Substitution: When Does KV Compression Yield a Reliable Replacement for the Full Cache?

Anonymous authors
Paper under double-blind review

Abstract

Long-context autoregressive inference is increasingly constrained by KV caches, whose memory footprint grows with sequence length, motivating compression for memory-efficient inference. However, when compression replaces the full cache with a smaller one that continued decoding must use and update, matching the current attention output is insufficient in general: two cache states can agree now yet diverge after the same append because softmax hides normalization mass. We characterize when current-output error can nevertheless control cache-extension error. For unchanged-row selection, this control depends on omitted attention mass and the separation between retained and omitted value means. Cancellation is the failure boundary: substantial omitted mass can remain nearly invisible in the current output. Away from it, current error controls the hidden mass exposed by an append, yielding a sharp, attainable one-append guarantee. The same control extends perturbatively to more general compressed caches that remain close, in response and normalization mass, to an unchanged-row selection on the same support. Fresh Qwen and Mistral evaluations support this mechanism: transition-aware scoring improves pairwise post-extension concordance over current-output scoring by .320 and .211, respectively, with 95% confidence intervals excluding zero in both model families. CaM and KeepKV test perturbative transfer at different scales. Finally, we characterize exactly when these headwise guarantees survive fixed output projection. Together, the results show how the geometry of compressed caches can make current-output error informative for reliable KV-cache replacement despite its global insufficiency.

1. Introduction

Long-context autoregressive inference is constrained by KV-cache memory, which grows with sequence length (Kwon et al., 2023; Nawrot et al., 2024). Existing methods reduce this cost through retention, budget allocation, sparse access, merging, and related state transformations (Liu et al., 2023; Zhang et al., 2023; Ge et al., 2024; Xiao et al., 2024; Li et al., 2024; Yang et al., 2024; Cai et al., 2025; Nawrot et al., 2024; Zhang et al., 2024). For methods that replace the full cache with a reduced one, continued decoding uses and updates the replacement, raising a reliability question beyond memory reduction. Query-family risk tracks numerator/denominator responses over future queries (Haverbeck et al., 2026); here we ask when current attention-output error controls the error after a common append.

The difficulty appears already in one head. For a fixed query, let S be the softmax normalization mass, N the unnormalized weighted-value sum, and $A = N/S$ the attention output. Appending KV rows changes S and N before renormalization, so states can share A yet differ in S and diverge after the same append. Exact attention kernels likewise preserve normalization scales when composing partial results (Dao et al., 2022; Ye et al., 2025). Thus current attention-output agreement is not preserved by append in general; whether this matters depends on the geometry of compressed caches.

The obstruction concerns the resulting cache state, not the compressor's objective. Different constructions may recover through different constraints; unchanged-row selection provides a tractable reference geometry. If an unchanged-row selector omits attention mass p, let δ be current-output

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.

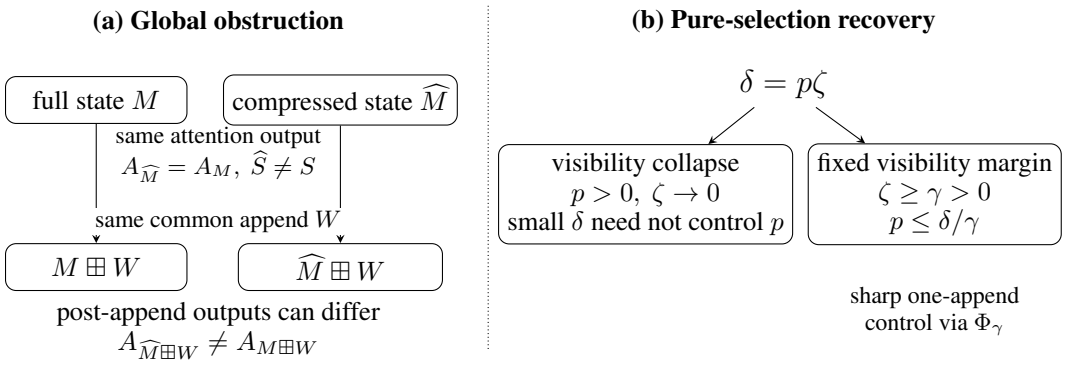


Figure 1: Replacement-state geometry at a fixed query. **Left:** M and $\widehat{M}$ are the full and compressed KV states; A_M is the attention output, S and $\widehat{S}$ their normalization masses, and W the common appended KV group; $\boxplus W$ denotes appending that group. Equal attention outputs can therefore hide different mass and separate after the same append. **Right:** for pure selection, p is omitted attention mass, ζ the retained–omitted conditional-value separation normalized by the value-set diameter D, and δ the similarly normalized current-output error, so $\delta = p\zeta$. The margin γ lower-bounds ζ, and Φ_γ is the sharp one-append continuity modulus.

error normalized by the admissible value-set diameter and ζ the retained–omitted conditional-value separation under the same normalization. Prior fixed-row analysis gives exactly $\delta = p\zeta$ and develops CertKV (Yang & Li, 2026). Here, we use the same identity for a transition question: when does a compressed cache remain a reliable replacement after a common append? Near cancellation, small δ can hide substantial omitted mass; away from cancellation, it controls hidden mass and yields a sharp one-append modulus.

Beyond selection, bounded response and log normalization-mass defects transfer the one-append guarantee perturbatively from an unchanged source-support reference; selection is the zero-defect case. Fresh CaM and KeepKV test this regime. A final multihead result identifies the output-projection boundary.

Contributions.

- **Sharp transition recovery.** We prove current-output agreement need not be preserved by a common append and, using $\delta = p\zeta$, derive the optimal fixed-margin modulus away from cancellation.
- **Constructor-level perturbative transfer.** Support-relative response and normalization defects extend the modulus beyond selection, which is recovered at zero defect (Theorem 2).
- **Fresh replacement-state evidence.** Output-matched tests show that transition-aware coefficients add endpoint discrimination beyond current-output error; CaM/KeepKV test constructor transfer, and two frozen native-continuation populations test whether common-write difficulty survives self-generated continuation.
- **Output-projection boundary.** Headwise radius bounds compose through fixed output projection; universal finite-scale nonexpansion can fail, with exact iff characterization.

Figure 1 summarizes the obstruction-to-recovery path developed in Sections 2–3.

2. State substitution under a common append

We analyze one common append in one head; the same query q probes both states and is not assumed to be the next autoregressive query. For a KV state $M = \{(k_i, v_i)\}_{i=1}^n$, define

$$S_M(q) = \sum_i e^{q^\top k_i}, \qquad N_M(q) = \sum_i e^{q^\top k_i} v_i, \qquad A_M(q) = \frac{N_M(q)}{S_M(q)}.$$

For a replacement $M \mapsto \widehat{M}$, define the hidden normalization mismatch

$$\eta_{\mathrm{rep}}(q) = \log \frac{S_{\widehat{M}}(q)}{S_M(q)}.$$

The attention output $A_M(q)$ does not identify $S_M(q)$ and $N_M(q)$ separately.

Now append the same KV group to both states. If its unnormalized mass is $G = xS_M(q)$ and its conditional value is v, the full and compressed post-append outputs are convex mixtures of their pre-append outputs and v, but with generally different mixture weights. Writing

$$\lambda = \frac{1}{1+x}, \qquad \widehat{\lambda} = \frac{e^{\eta_{\mathrm{rep}}}}{e^{\eta_{\mathrm{rep}}} + x},$$

direct subtraction gives

$$A^+_{\widehat{M}} - A^+_M = \widehat{\lambda}\big(A_{\widehat{M}} - A_M\big) + (\widehat{\lambda} - \lambda)(A_M - v). \tag{1}$$

The first term propagates visible pre-append output error. The second depends on a normalization coordinate that the current output does not identify.

Proposition 1 (Current-output agreement need not be preserved by append)**.** *There exist KV states $M, \widehat{M}$ and a common append such that $A_{\widehat{M}}(q) = A_M(q)$ before the append but $A^+_{\widehat{M}}(q) \neq A^+_M(q)$ afterward. Hence equality of the current attention output is not a right congruence under append.*

For example, take two unit-mass source rows with value zero and retain one. Both outputs are zero. A common unit-mass append with value one changes the full output to $1/3$ and the selected output to $1/2$. Thus append is not well defined on the equality classes induced by the attention output.

Pure selection links normalization mass to the source rows that survive. We next ask when that constraint makes small current-output error control append error.

3. Sharp transition recovery from selection geometry

Consider a pure-selection state $\widehat{M}$ obtained by retaining rows R and omitting rows O without modifying the surviving keys or values. For a fixed query, write

$$S = S_R + S_O, \qquad p = \frac{S_O}{S}, \qquad A_R = \frac{N_R}{S_R}, \qquad A_O = \frac{N_O}{S_O}.$$

The full output decomposes as

$$A = (1-p)A_R + pA_O,$$

whereas the selection candidate has $\widehat{A} = A_R$. Therefore

$$\widehat{A} - A = p(A_R - A_O). \tag{2}$$

The omitted mass p measures how much attention probability the selector removes; the separation between A_R and A_O measures how visible that omission is for the queried response.

Fix a convex admissible value set of diameter $D > 0$ containing every source value and every appended value in the declared one-append contract. All norms below are Euclidean. Define

$$\delta = \frac{\|\widehat{A} - A\|}{D}, \qquad \zeta = \frac{\|A_R - A_O\|}{D}.$$

The fixed-row identity (Yang & Li, 2026, Theorem 3.7) gives

$$\delta = p\zeta. \tag{3}$$

Thus small δ can reflect either small omitted mass p or cancellation $\zeta \approx 0$.

THE CANCELLATION BOUNDARY

For $0 < p < 1$, exact instantaneous agreement occurs precisely when $A_R = A_O$. On this cancellation locus the candidate still has mass $S_{\widehat{M}} = (1-p)S$. If a common append has relative mass $x = G/S$ and conditional value v, direct subtraction yields

$$\widehat{A}^+ - A^+ = \frac{p\big((1-p)(A_R - A_O) + x(v - A_O)\big)}{(1+x)(1-p+x)}. \tag{4}$$

Thus on $A_R = A_O$,

$$\widehat{A}^+ - A^+ = \frac{px(v - A_O)}{(1+x)(1-p+x)},$$

which is nonzero whenever $x > 0$ and $v \neq A_O$. The global obstruction of Proposition 1 therefore occurs inside pure selection at a concrete ill-conditioned boundary.

The scalar factor in the last display is also exactly the hidden coefficient $|\widehat{\lambda} - \lambda|$ for every pure-selection state, because $e^{\eta_{\mathrm{rep}}} = 1 - p$. For $0 \le r < 1$, define

$$h(r) = \frac{1 - \sqrt{1-r}}{1 + \sqrt{1-r}}.$$

Maximizing the coefficient over $x \ge 0$ gives

$$\sup_{x \ge 0} \frac{px}{(1+x)(1-p+x)} = \frac{1 - \sqrt{1-p}}{1 + \sqrt{1-p}} = h(p), \tag{5}$$

attained at $x = \sqrt{1-p}$. This is the shifted-sigmoid envelope $\tanh(|\eta_{\mathrm{rep}}|/4)$ with $\eta_{\mathrm{rep}} = \log(1-p)$ (Yang & Li, 2026, Appendix C.1). Both vanishing and overwhelming common writes hide this particular exposure; the largest effect occurs at an intermediate matched mass scale.

A SHARP REPLACEMENT BOUND

Away from cancellation, $\delta = p\zeta$ controls omitted mass. To retain the information in the transition itself, write $\rho = \|v - A_O\|/D \le 1$ and $\delta_+ = \|\widehat{A}^+ - A^+\|/D$. The triangle inequality applied to Equation (4) gives

$$\delta_+ \le B_{\mathrm{local}} := \frac{p((1-p)\zeta + x\rho)}{(1+x)(1-p+x)} \le B_{\mathrm{coeff}}(p, \delta, x) := \frac{(1-p)\delta + px}{(1+x)(1-p+x)}. \tag{6}$$

The second bound needs omitted mass and append mass, but no appended-value contrast or vector orientation. Keeping the shared denominator is important: bounding the propagated output error by δ discards its contraction.

Proposition 2 (Sharp fixed-mass envelope)**.** *For feasible $0 \le u \le p \le 1$, define*

$$\Psi(u, p) := \sup_{x \ge 0} \frac{(1-p)u + px}{(1+x)(1-p+x)}. \tag{7}$$

Then $B_{\mathrm{coeff}}(p, u, x) \le \Psi(u, p)$ for every $x \ge 0$, and the bound is attained or approached. Its branchwise closed form and endpoint conventions are given in Appendix A.3.

Let Γ_0 be a finite collection of pre-append responses. Let Γ_1 contain those responses and, for each query, the response after a declared common append. Write d_{Γ_j} for the largest response discrepancy in that collection.

Theorem 1 (Sharp recovery on a fixed-margin selection family)**.** *Suppose an unchanged-row selection family satisfies $\zeta(q) \ge \gamma$ whenever its omitted mass is nonzero, for $0 < \gamma \le 1$. If $u = d_{\Gamma_0}/D < \gamma$, then*

$$\frac{d_{\Gamma_1}}{D} \le \Phi_\gamma(u) := \max_{u \le p \le u/\gamma} \Psi(u, p) = \max_{x \ge 0} \frac{(1-r)u + rx}{(1+x)(1-r+x)}, \qquad r = \frac{u}{\gamma}. \tag{8}$$

This bound is sharp over diameter-bounded pure-selection states and admissible common appends, and $\Phi_\gamma(u) \to 0$ as $u \to 0$.

For $u > 0$, with $r = u/\gamma$, define $a_\gamma(u) = [r(2-r) - 2(1-r)u]/r^2$. The maximum has the closed form

$$\Phi_\gamma(u) = \begin{cases} u, & 2\gamma \ge 1 + u, \\ \big[a_\gamma(u) + \sqrt{a_\gamma(u)^2 - 1}\big]^{-1}, & 2\gamma < 1 + u, \end{cases} \qquad \Phi_\gamma(0) = 0. \tag{9}$$

Why this bound is sharp. At a point with $\delta = u$, diameter normalization and $\zeta \geq \gamma$ give the joint constraint $u \leq p \leq u/\gamma$. For fixed u, x,

$$\frac{\partial B_{\text{coeff}}}{\partial p} = \frac{x(1+x-u)}{(1+x)(1-p+x)^2} \geq 0.$$

The worst mass is therefore $p = r$. Maximizing over the remaining append scale gives Equation (9). In the interior case, collinear values $A_O = 0$, $A_R = \gamma D$, and $v = D$, with $p = r$ and the maximizing append mass, attain equality. When $2\gamma \geq 1 + u$, the pre-append discrepancy already attains the maximum u. Appendix A.3 proves the finite-contract statement, including points with $\delta < u$.

The simpler bound

$$\Phi_\gamma(u) \leq \omega_\gamma(u) := u + h(u/\gamma) \tag{10}$$

follows from Equation (1) and monotonicity of h. Here h is the sharp hidden-coefficient envelope, whereas ω_γ relaxes the full transition bound. For example, $\Phi_{.10}(.025) = .07861$ and $\omega_{.10}(.025) = .09680$. A fixed visibility margin is sufficient for uniform recovery; it is not necessary along every convergent sequence: $p = \zeta = t$ also gives vanishing append error, although $\zeta \to 0$.

Here γ is the selection family's margin from cancellation; Section 4 tests this conditioning on compressed caches.

Transfer beyond unchanged-row selection

Let C be any constructed state and let R be the unchanged-row state on the constructor's declared source support. At the same query define

$$\xi = \frac{\|A_C - A_R\|}{D}, \qquad \eta = \log \frac{S_C}{S_R}, \qquad \delta_C = \frac{\|A_C - A\|}{D}.$$

The support state still obeys the selection decomposition with omitted mass p and visibility ζ. Addition and subtraction give $A_C - A = p(A_R - A_O) + (A_C - A_R)$, so on $\zeta \geq \gamma$,

$$p \leq \frac{\delta_C + \xi}{\gamma}. \tag{11}$$

Theorem 2 (Approximate-factorization transfer)**.** *Suppose $\zeta \geq \gamma > 0$ and $q := \delta_C + \xi < \gamma$. After any admissible common append,*

$$\frac{\|A_C^+ - A^+\|}{D} \leq \Phi_\gamma(q) + \xi + \tanh\left(\frac{|\eta|}{4}\right). \tag{12}$$

If the constructor preserves the support keys, weights, and cardinality, then $\eta = 0$. Pure selection has $\xi = \eta = 0$ and recovers Theorem 1.

This is a sufficient transfer principle, not a characterization of all reliable compression; richer states may represent omitted information directly. Appendix A.4 gives the proof and exact normalization envelope.

Where exposure can remain

Equation (4) also identifies the variables that can defeat the restricted regime. Let

$$\rho = \frac{\|v - A_O\|}{D}, \qquad c = \cos \angle(A_R - A_O, v - A_O).$$

Then

$$\delta_+^2 = \frac{p^2\big((1-p)^2\zeta^2 + x^2\rho^2 + 2x(1-p)\zeta\rho c\big)}{(1+x)^2(1-p+x)^2}. \tag{13}$$

Large transition exposure therefore requires more than omitted mass alone. It is favored by appreciable p, low instantaneous visibility ζ, sufficient future contrast ρ, an effective append scale x, and an orientation that does not cancel those two value directions. This characterization yields predictions about the states on which practical selectors should fail.

Model	Φ cov.	Ψ cov.	B_{coeff} cov.	B_{local} cov.	Local recall
Qwen	.284	.452	.583	.698	.962
Mistral	.325	.470	.616	.708	.957

Table 1: Equal-method sound-bound ladder on 243,712 fresh-prompt natural-budget points at error tolerance .025. The uniform Φ bound also requires the fixed visibility margin $\zeta \geq .10$. Ψ observes omitted mass but maximizes over append scale; B_{coeff} observes both, and B_{local} adds local value contrast. Coverage is over the full population, recall over realized-safe points; issued decisions have precision one. The Ψ decomposition is a secondary analysis of the preregistered fresh population.

4. Do compressed caches satisfy sufficient recovery conditions?

We test whether compressed caches satisfy $\zeta \geq \gamma$ and small support-relative constructor defects in preregistered audits. Prompt is the sampling unit; an independent auditor recomputes all quantities and gates from frozen captures.

Compressed caches relative to the failure boundary

We first profile compressed caches produced by SnapKV, PyramidKV, and StreamingLLM on 64 Gemma-3-4B-IT (Gemma Team, 2025) prompts, with seven interactions from the uncompressed FullKV reference and four nested old-row budgets. The exact visible-plus-hidden decomposition is empirically tight: at the registered natural support budget, the median ratio of realized post-append error to the triangle bound induced by Equation (1) is 0.80. Yet among points with normalized current-output error at most 0.025, the fraction for which the normalized second term in Equation (1) (the latent exposure) is at least 0.025 is only 0.65% for SnapKV, 0.24% for PyramidKV, and 0.53% for StreamingLLM; every upper 95% interval endpoint is below 0.009. These rates fall more than an order of magnitude below the preregistered prevalence threshold.

The hidden channel is therefore tight but usually small; the remaining question is what geometry couples visible and latent error in these compressed caches.

Low current-output error usually accompanies low omitted mass

At the natural Gemma budget, the prompt-balanced fractions of low-current-error points with $p \leq \zeta$ are $.856$ for SnapKV, $.874$ for PyramidKV, and $.789$ for StreamingLLM. Increasing nested support budgets improves current-output error mainly through reduced omitted mass: the mass contribution exceeds the visibility contribution on $.773$–$.988$ of improving points, depending on method and budget transition. Thus typical agreement is associated with removing little query-relevant mass.

Against cardinality-matched random and fixed-workload-oracle supports, define recovery as the fraction of the random-to-oracle omitted-mass gap closed by the method. On Gemma, SnapKV/PyramidKV/StreamingLLM recover $.650/.624/.020$; StreamingLLM's interval crosses zero. The same geometry can therefore arise from different constructor–model–workload alignments. Appendix C.2 gives the controls and the exact budget-attribution identity.

Cross-family geometry is method- and model-dependent

We prospectively test these predictions on 64 LongBench rows (Bai et al., 2024) disjoint from the Gemma discovery population under Qwen3-8B (Yang et al., 2025) and Mistral-7B (Jiang et al., 2023). Each model independently tokenizes and crops the same rows. We reconstruct SnapKV, PyramidKV, StreamingLLM, and a prespecified H2O-prefill support at four budgets; the later dynamic H2O score update is outside the common-append transition.

Small current-output error remains predominantly mass-controlled in both fresh families. At the natural budgets, mass-dominated low-current-error fractions are 0.776–0.803 across Qwen methods. On Mistral, SnapKV, PyramidKV, and H2O lie between 0.847 and 0.863; StreamingLLM is a weak boundary at 0.515. Mass dominance holds for every method in both models; across adjacent budgets it spans 0.893–0.997 on Qwen and 0.657–0.997 on Mistral. Fixed-margin coverage is reported below.

The preregistered selected-tail analysis remains unresolved for Mistral–SnapKV (four exposed events; minimum ten). Because conditioning on low $\delta = p\zeta$ can mechanically place high-p events below a visibility cutoff, the other tail results are not treated as independent mechanism evidence. Cross-family tail localization therefore remains unresolved.

The matched controls also sharpen the interpretation of method dependence. On Qwen/Mistral, natural-budget fixed-workload-oracle recovery is respectively $0.868/0.892$ for SnapKV, $0.843/0.860$ for PyramidKV, $0.722/0.772$ for StreamingLLM, and $0.893/0.929$ for H2O. StreamingLLM thus moves from a Gemma near-null to strong positive alignment on both fresh model families. Future-mass alignment is not a property of the constructor name alone; it depends on the support rule as realized under a model and workload.

What transition information tightens the bound?

The derived calibration covers 974,848 Qwen/Mistral points at four budgets. At each probe let $u = p\zeta_c$, $e_+ = \|\widehat{A}^+ - A^+\|/D$, with ζ_c and ρ_c denoting the contrasts normalized by a single diameter over the source and all seven append values. This is the normalization used in Theorem 1; its construction is specified in Appendix B. A certificate is issued when its bound is at most ϵ; a uniform one also requires $G_\gamma := \{\zeta_c \geq \gamma,\ u < \gamma\}$.

On the previously observed population, replacing ω by Φ increases natural-budget coverage from 16.5–34.4% to 18.9–38.7%, a gain of 2.5–4.7 percentage points within individual cells. The local bound covers 56.6–77.6% with 92.8–97.0% safe-state recall. Because Φ_γ is sharp for the uniform interface that observes only current error and the fixed visibility margin, its remaining slack cannot be removed by a tighter bound using the same information. The progression from Φ_γ to Ψ, the coefficient bound, and the local bound therefore measures the value of revealing omitted mass, append scale, and local value geometry.

On the preregistered fresh population, the Ψ decomposition inserts Proposition 2 into the hierarchy:

$$e_+ \leq B_{\text{local}} \leq B_{\text{coeff}} \leq \Psi(u,p), \qquad \Psi(u,p) \leq \Phi_\gamma(u) \leq \omega_\gamma(u) \quad \text{on } G_\gamma.$$

As Table 1 shows, observing realized omitted mass raises equal-method coverage by $.168/.145$ on Qwen/Mistral; observing append scale raises it by another $.132/.145$. For the scale increment, paired 95% prompt-bootstrap lower endpoints are $.128/.141$.

On the common slice $u \leq .025, \zeta_c \geq .10$, observing p accounts for $.658/.610$ of the sequential bound-gap tightening and then observing x for $.342/.390$; these are order-specific bound differences, not causal shares. Local value geometry raises full-population coverage to $.698/.708$ and safe-state recall to $.962/.957$. Sensitivity and baseline comparisons are in Appendix C.1.

Does transition information distinguish attention-output-matched caches?

A tighter bound need not improve endpoint ordering. We therefore preregister a fresh-prompt test of whether the coefficient score separates unsafe from safe endpoints at nearly equal current-output error. It uses 64 new LongBench rows (32 each from 2WikiMultihopQA and Qasper) shared across Qwen3-8B and Mistral-7B-Instruct-v0.3; all development, discovery, and calibration rows are excluded by outcome-blind hash selection.

Each prompt is cropped to its first and last 128 model-native tokens. We capture query heads 0 and 16 in every layer and seven FullKV-greedy common prefixes. The same four selectors retain a 32-row recent window and use natural old-row budgets of 36 on Qwen and 32 on Mistral, with PyramidKV's registered layer schedule. This gives 243,712 local measurements across the 64 physical prompts. The endpoint is fidelity on these 256-token attention states rather than an end-to-end LongBench task endpoint.

The untrained scores are $S_{\text{read}} = u$ and $S_{\text{coeff}} = B_{\text{coeff}}(p,u,x)$. At $\epsilon = .025$, form all pairs among the seven writes within the same model, method, budget, prompt, and attention unit. Both members must satisfy $u \leq \epsilon$, their endpoint labels $Y = \mathbf{1}[e_+ > \epsilon]$ must differ, and $|u_i - u_j| \leq .10\epsilon = .0025$. Concordance is one when the unsafe member has the higher score, one half for a tie, and zero otherwise. Average pairs within prompt, eligible prompts within method, and then the four methods equally within model. We call the resulting quantity the equal-method average. Uncertainty uses a

paired, dataset-stratified prompt bootstrap with 20,000 replicates; points and pairs are not independent sampling units.

Model	Method	Prompts	Pairs	Current-output score	Transition-aware score	Difference [95% CI]
Qwen	SnapKV	15	26	.690	.970	.280 [.075, .520]
	PyramidKV	34	114	.533	.852	.319 [.206, .436]
	StreamingLLM	60	287	.471	.928	.458 [.368, .540]
	H_2O	32	66	.615	.840	.225 [.070, .390]
Mistral	SnapKV	8	10	.688	.750	.063 [$-.500$, .600]
	PyramidKV	20	25	.600	.825	.225 [.048, .429]
	StreamingLLM	53	540	.674	.873	.198 [.122, .275]
	H_2O	13	19	.564	.923	.359 [.143, .577]
Qwen	Equal-method avg.					.320 [.236, .411]
Mistral	Equal-method avg.					.211 [.040, .382]

Table 2: Fresh-prompt attention-output-matched concordance at the registered primary tolerance and caliper. Prompt and pair counts describe eligibility, not independent trials. Both equal-method intervals exclude zero. Mistral–SnapKV remains unresolved.

As Table 2 shows, the coefficient score improves equal-method concordance by .3205 on Qwen (95% CI $[.2364, .4114]$) and .2112 on Mistral ($[.0405, .3824]$); both intervals exclude zero. Seven of eight method intervals exclude zero. Mistral–SnapKV supplies only eight eligible prompts and ten pairs; its interval $[-.50, .60]$ leaves the method-level effect unresolved. Unsafe endpoints are rare even within the low-current-error subset, ranging from 0.096% to 2.592% across cells. Across the preregistered tolerance/caliper grid, all equal-method intervals remain positive. Sparse-cell eligibility and valid bootstrap counts are reported in Appendix C.4.

On these fresh prompts, the sharp uniform bound also raises coverage over ω by 2.7–4.7 percentage points at the primary tolerance. These results establish additional discrimination under the measured common-append contract. The scores use FullKV-derived coordinates, including future append mass; the experiment measures diagnostic discrimination on realized states.

After matching both (u, p), only StreamingLLM has registered support in both models; adding x raises concordance by .314 $[.112, .500]$ and .343 $[.199, .489]$. Other method cells remain sparse, precluding a cross-method claim.

Fresh transfer beyond unchanged-row selection

Theorem 2, its constants, constructors, and gates are frozen before model loading on 32 previously unused prompts per model. Public CaM changes only stored values on its selected support, so $\eta = 0$ exactly; it retains .9119 $[.9084, .9153]$ and .9642 $[.9593, .9687]$ of support-only bound coverage on Qwen/Mistral. Larger-defect KeepKV independently retains $.5684/.8258$. Removing its mass term causes a .01397 maximum soundness violation on fresh Qwen, while the complete bound has none. Selection is outcome-blind; this is fresh evidence in both model families. Appendix B.5 gives full coverage, defects, protocol, and provenance.

Native continuation. Because the theorem compares a common append, we test whether relative difficulty survives native continuation. In the frozen 14-prompt/model audit, common-write/native errors remain strongly ranked ($\rho_s = .9648/.9033$), but certificate/first-divergence direction is mixed ($.5131/-.0736$). An independent 32-prompt/model HotpotQA/MuSiQue replication passes the directional gate in both models: certificate/lifetime correlations are $.2196/.1418$ and common/native-error correlations are $.9377/.9384$. Lifetime-association intervals still cross zero, whereas both common/native-ordering intervals are strictly positive. Thus relative difficulty replicates without a closed-loop lifetime certificate (Appendix B.6).

5. THE OUTPUT-PROJECTION BOUNDARY

Sections 2–4 establish one-head common-append recovery under restricted geometry. We next ask when headwise radius bounds remain universally nonexpansive after fixed output projection.

Grouped-query attention and head-heterogeneous cache policies make headwise composition operational in modern decoders (Ainslie et al., 2023; Ge et al., 2024; Xiao et al., 2025; Tang et al., 2025); KQ-SVD already propagates per-head approximation terms through W_O (Lesens et al., 2026). Write z_h for a possible post-append error at head h, b_h for its valid radius bound, and $W_O = [W_{O,1}, \dots, W_{O,H}]$. The balls $\|z_h\|_2 \le b_h$ compose continuously through W_O (Corollary 1), but finite-scale universal nonexpansion holds iff $W_{O,i}^\top W_{O,j} = 0$ for $i \neq j$ and $W_{O,h}^\top W_{O,h} = c_h I$ with $c_h \ge 0$ (Proposition 4). In the retrospective real-W_O audit, a pointwise tighter valid bound can rank endpoints worse: the visible joint direction gives output-matched concordance increments of $-.1785/-.2630$ on Qwen/Mistral (Table 7). Independent radius balls are likewise too coarse for realized-stable endpoints; Appendix A.5 gives the fixed-direction, intrinsic-Gram, and Reynolds results.

Thus output projection changes the guarantee: constructor geometry can recover one-head transition stability, but generic cross-head mixing need not preserve finite-scale nonexpansion. Pure selection is the zero-defect reference, controlled (ξ, η) extend the sufficient class, and Proposition 4 marks the architectural boundary.

RELATED WORK

KV-cache compression spans selection/eviction, budget allocation, sparse access, merging/fusion, learned or recurrent compression, and quantization/coding (Li et al., 2025). These mechanisms need not share a scoring objective, and sparse access need not construct a smaller cache. We analyze unchanged-row selection exactly and support-anchored constructors only under controlled response and normalization defects; Appendix D maps the broader systems literature.

The closest theory studies overlapping quantities under different contracts. Yang & Li (2026) derive the fixed-row retained–omitted factorization and develop CertKV; we use the same identity for sharp common-append recovery. Haverbeck et al. (2026) treat compression as replacement and track query-indexed numerator/denominator response profiles for future-query minimax risk; our local contract asks when current-output error controls one declared append. Streaming approximation and worst-case lower bounds concern sequence approximation or space complexity (Kochetkova et al., 2025; Haris & Onak, 2025). ResKV reconstructs omitted numerator/denominator statistics that a normalized output can hide (Zhan et al., 2026). KQ-SVD propagates per-head approximation through W_O (Lesens et al., 2026); our fixed-W_O result gives the exact finite-scale nonexpansion boundary.

LIMITATIONS

The theory is local to finite common-append endpoints: sharpness uses unchanged-row selection, transfer uses measured support-relative defects, the fixed-W_O lift is structural, and FullKV-derived coordinates are retrospective. Fresh common-write evidence spans two model families and 256-token states; both native-continuation audits use one compression event and 32 tokens. The replication matches the registered lifetime direction, but its intervals still cross zero. Mistral–SnapKV remains unresolved, and the strict cross-family tail-localization standard was not met. The empirical target is replacement-state transition behavior, not downstream utility or runtime. Context-length scaling, repeated recompression, and third-family confirmation remain outside scope.

CONCLUSION

Current-output agreement need not compose because append exposes hidden mass. Away from cancellation, selection yields sharp one-append control, and support-relative defects transfer it to support-anchored constructors. Fresh output-matched tests favor transition-aware scoring; the multihead result marks a second composition boundary. These are sufficient recovery conditions, not necessary ones: reliable replacement is transition-relative and contract-dependent.

Reproducibility Statement

Appendices A and B give proofs and frozen protocols. An anonymized supplementary artifact accompanies the submission; it contains compact result tables, prompt-level contributions, version hashes, fixed PRNG seeds, and standalone audit scripts that reconstruct the released summaries and gates independently of the primary analysis pipeline. Raw tensor captures are not included in that package; their tensor-level provenance is checked by the separate raw-capture audit described in Appendix B.4. Exact version bindings needed for replay are recorded in the supplementary artifact.

AI Use Disclosure

The authors conceived the research questions and core framing and directed the proof strategy and experimental program. Generative AI tools assisted formulation and cross-checking of theoretical claims and proofs, experiment-design feedback, implementation and code auditing, literature search and analysis, cross-checking of result interpretations, and manuscript editing. The authors reviewed and verified all mathematical claims, code, results, citations, and text and take full responsibility for the final content.

References

Muhammad Adnan, Akhil Arunkumar, Gaurav Jain, Prashant J. Nair, Ilya Soloveychik, and Purushotham Kamath. Keyformer: KV cache reduction through key tokens selection for efficient generative inference. In *Proceedings of Machine Learning and Systems*, 2024.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, 2023. doi: 10.18653/v1/2023.emnlp-main.298.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137. Association for Computational Linguistics, 2024.

Ruisi Cai, Yuandong Tian, Zhangyang Wang, and Beidi Chen. LoCoCo: Dropping in convolutions for long context compression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and Wen Xiao. PyramidKV: Dynamic KV cache compression based on pyramidal information funneling. In *Second Conference on Language Modeling*, 2025.

Alex Chen, Renato Geh, Aditya Grover, Guy Van den Broeck, and Daniel Mingyi Israel. The pitfalls of KV cache compression. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract-Conference.html.

Harry Dong, Xinyu Yang, Zhenyu Zhang, Zhangyang Wang, Yuejie Chi, and Beidi Chen. Get more with LESS: Synthesizing recurrence with KV cache compression for efficient LLM inference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

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. In *Advances in Neural Information Processing Systems*, 2025.

Yuan Feng, Junlin Lv, Haoyu Guo, Yukun Cao, S. Kevin Zhou, and Xike Xie. CriticalKV: Optimizing KV cache eviction from an output perturbation perspective. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Yu Fu, Zefan Cai, Abedelkadir Asi, Wayne Xiong, Yue Dong, and Wen Xiao. Not all heads matter: A head-level KV cache compression method with integrated retrieval and reasoning. In *International Conference on Learning Representations*, 2025.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *International Conference on Learning Representations*, 2024.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Themistoklis Haris and Krzysztof Onak. Compression barriers in autoregressive transformers. In *Proceedings of the 38th Conference on Learning Theory*, 2025.

Lukas Haverbeck, Carmen Amo Alonso, Andres Felipe Posada-Moreno, Sebastian Trimpe, and Marco Pavone. The risk of KV cache compression. *arXiv preprint arXiv:2607.01520*, 2026. URL https://arxiv.org/abs/2607.01520.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. KVQuant: Towards 10 million context length LLM inference with KV cache quantization. In *Advances in Neural Information Processing Systems*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Ekaterina Kochetkova, Kshiteej Jitesh Sheth, Insu Han, Amir Zandieh, and Michael Kapralov. Streaming attention approximation via discrepancy theory. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Damien Lesens, Beheshteh T. Rakhshan, and Guillaume Rabusseau. KQ-SVD: Compressing the KV cache with provable guarantees on attention fidelity. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, 2026.

Haoyang Li, Yiming Li, Anxin Tian, Tianhao Tang, Zhanchao Xu, Xuejia Chen, Nicole Hu, Wei Dong, Qing Li, and Lei Chen. A survey on large language model acceleration based on KV cache management. *Transactions on Machine Learning Research*, 2025.

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. In *Advances in Neural Information Processing Systems*, 2024.

Akide Liu, Jing Liu, Zizheng Pan, Yefei He, Gholamreza Haffari, and Bohan Zhuang. MiniCache: KV cache compression in depth dimension for large language models. In *Advances in Neural Information Processing Systems*, 2024a.

Guangda Liu, Chengwei Li, Jieru Zhao, Chenqi Zhang, and Minyi Guo. ClusterKV: Manipulating LLM KV cache in semantic space for recallable compression. In *2025 62nd ACM/IEEE Design Automation Conference*, pp. 1–7, 2025. doi: 10.1109/DAC63849.2025.11132479.

Yuhan Liu, Hanchen Li, Yihua Cheng, Siddhant Ray, Yuyang Huang, Qizheng Zhang, Kuntai Du, Jiayi Yao, Shan Lu, Ganesh Ananthanarayanan, Michael Maire, Henry Hoffmann, Ari Holtzman, and Junchen Jiang. CacheGen: KV cache compression and streaming for fast large language model serving. In *Proceedings of the ACM SIGCOMM 2024 Conference*, pp. 38–56, 2024b. doi: 10.1145/3651890.3672274.

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*, 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*, 2024c.

Daming Luo, Christy Liang, and Junyu Xuan. How query visibility changes KV-cache compression rankings: A matched-budget audit. *arXiv preprint arXiv:2607.11942*, 2026. URL https://arxiv.org/abs/2607.11942.

Piotr Nawrot, Adrian Łańcucki, Marcin Chochowski, David Tarjan, and Edoardo M. Ponti. Dynamic memory compression: Retrofitting LLMs for accelerated inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 37396–37412, 2024.

Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. Transformers are multi-state RNNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Junjie Peng, You Wu, Haoyi Wu, Jialong Han, Xiaohua Xie, Kewei Tu, and Jianhuang Lai. GRKV: Global regression for training-free KV cache compression in long-context LLMs. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026.

Chen Qiu, Ziwu Liu, Chao Fei, Guozhong Li, and Panos Kalnis. KVDiagnosis: A diagnostic benchmark for KV-cache compression in long-context language models. *arXiv preprint arXiv:2608.09412*, 2026. URL https://arxiv.org/abs/2608.09412.

Luka Ribar, Ivan Chelombiev, Luke Hudlass-Galley, Charlie Blake, Carlo Luschi, and Douglas Orr. SparQ attention: Bandwidth-efficient LLM inference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Yiqun Shen, Song Yuan, Zhengze Zhang, Xiaoliang Wang, Daxin Jiang, and Cam-Tu Nguyen. LAVa: Layer-wise KV cache eviction with dynamic budget allocation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13672–13692. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.737. URL https://aclanthology.org/2025.findings-emnlp.737/.

Hanlin Tang, Yang Lin, Jing Lin, Qingsen Han, Danning Ke, Shikuan Hong, Yiwu Yao, and Gongyi Wang. RazorAttention: Efficient KV cache compression through retrieval heads. In *International Conference on Learning Representations*, 2025.

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.

Yuxuan Tian, Zihan Wang, Yebo Peng, Aomufei Yuan, Zhiming Wang, Bairen Yi, Xin Liu, Yong Cui, and Tong Yang. KeepKV: Achieving periodic lossless KV cache compression for efficient LLM inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

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.

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. URL https://arxiv.org/abs/2505.09388.

Chiwun Yang and Xiaoyu Li. Beyond sparse weights: When is attention compressible? *arXiv preprint arXiv:2608.21541*, 2026. URL https://arxiv.org/abs/2608.21541v1. Version 1.

Dongjie Yang, Xiaodong Han, Yan Gao, Yao Hu, Shilin Zhang, and Hai Zhao. PyramidInfer: Pyramid KV cache compression for high-throughput LLM inference. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3258–3270, 2024. doi: 10.18653/v1/2024.findings-acl.195.

Zihao Ye, Lequn Chen, Ruihang Lai, Wuwei Lin, Yineng Zhang, Stephanie Wang, Tianqi Chen, Baris Kasikci, Vinod Grover, Arvind Krishnamurthy, and Luis Ceze. FlashInfer: Efficient and customizable attention engine for LLM inference serving. In *Proceedings of Machine Learning and Systems*, volume 7, 2025. URL https://arxiv.org/abs/2501.01005.

Yuhang Zhan, Lisi Chen, and Shuo Shang. ResKV: Reconstructing omitted attention contributions for fixed-budget KV cache compression. *arXiv preprint arXiv:2607.29591*, 2026. URL https://arxiv.org/abs/2607.29591.

Yuxin Zhang, Yuxuan Du, Gen Luo, Yunshan Zhong, Zhenyu Zhang, Shiwei Liu, and Rongrong Ji. CaM: Cache merging for memory-efficient LLMs inference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

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*, 2023.

Appendix A. Proofs for replacement-state geometry

A.1. Conventions and one-append coefficient

Fix a query q and suppress its argument. For a full state M and candidate $\widehat{M}$, write

$$S = S_M, \qquad \widehat{S} = S_{\widehat{M}}, \qquad A = A_M, \qquad \widehat{A} = A_{\widehat{M}},$$

and

$$\eta_{\mathrm{rep}} = \log(\widehat{S}/S).$$

A common appended group has mass $G = xS$ with $x \geq 0$ and conditional value v. The post-append outputs are

$$A^+ = \frac{A + xv}{1 + x}, \qquad \widehat{A}^+ = \frac{e^{\eta_{\mathrm{rep}}}\widehat{A} + xv}{e^{\eta_{\mathrm{rep}}} + x}.$$

Hence, with

$$\lambda = \frac{1}{1 + x}, \qquad \widehat{\lambda} = \frac{e^{\eta_{\mathrm{rep}}}}{e^{\eta_{\mathrm{rep}}} + x},$$

we have

$$\widehat{A}^+ - A^+ = \widehat{\lambda}(\widehat{A} - A) + (\widehat{\lambda} - \lambda)(A - v). \tag{14}$$

The hidden coefficient is

$$|\widehat{\lambda} - \lambda| = \frac{x|e^{\eta_{\mathrm{rep}}} - 1|}{(1 + x)(e^{\eta_{\mathrm{rep}}} + x)}. \tag{15}$$

For $\eta_{\mathrm{rep}} \neq 0$, differentiation in x shows that the unique positive stationary point is $x = e^{\eta_{\mathrm{rep}}/2}$ and is the global maximum on $[0, \infty)$. For $\eta_{\mathrm{rep}} = 0$, the coefficient is identically zero. Therefore

$$\sup_{x \geq 0} |\widehat{\lambda} - \lambda| = \tanh\left(\frac{|\eta_{\mathrm{rep}}|}{4}\right). \tag{16}$$

Both endpoints $x = 0$ and $x \to \infty$ give zero coefficient.

A.2. PURE-SELECTION IDENTITIES

Let R and O partition the source rows, with R nonempty. For finite keys and queries all softmax weights are strictly positive, so $S_O = 0$ iff $O = \varnothing$. When $O = \varnothing$, the selection candidate equals the full state and all replacement errors vanish. All formulas involving A_O or ζ below are therefore stated for $O \neq \varnothing$; the empty-omitted case is included separately as the exact zero-error case.

For $O \neq \varnothing$, define

$$p = \frac{S_O}{S} \in (0,1), \qquad A_R = \frac{N_R}{S_R}, \qquad A_O = \frac{N_O}{S_O}.$$

Since

$$N = S_R A_R + S_O A_O,$$

we obtain

$$A = (1-p)A_R + pA_O.$$

The pure-selection candidate has $\widehat{S} = S_R = (1-p)S$ and $\widehat{A} = A_R$, so

$$\widehat{A} - A = p(A_R - A_O), \qquad \eta_{\mathrm{rep}} = \log(1-p). \tag{17}$$

For a diameter $D > 0$ and

$$\delta = \|\widehat{A} - A\|/D, \qquad \zeta = \|A_R - A_O\|/D,$$

this gives $\delta = p\zeta$.

For $0 \leq r < 1$, set

$$h(r) = \frac{1 - \sqrt{1-r}}{1 + \sqrt{1-r}}.$$

Now append a common group of relative mass $x = G/S$. Direct subtraction gives

$$\widehat{A}^+ - A^+ = \frac{(1-p)A_R + xv}{1-p+x} - \frac{(1-p)A_R + pA_O + xv}{1+x} \tag{18}$$

$$= \frac{p\big((1-p)(A_R - A_O) + x(v - A_O)\big)}{(1+x)(1-p+x)}. \tag{19}$$

On the cancellation locus $A_R = A_O$ this reduces to

$$\widehat{A}^+ - A^+ = \frac{px(v - A_O)}{(1+x)(1-p+x)}.$$

The scalar coefficient has derivative proportional to $(1-p) - x^2$, hence is maximized at $x = \sqrt{1-p}$, where

$$\sup_{x \geq 0} \frac{px}{(1+x)(1-p+x)} = \frac{1 - \sqrt{1-p}}{1 + \sqrt{1-p}} = h(p). \tag{20}$$

The same identity follows by substituting $\eta_{\mathrm{rep}} = \log(1-p)$ into Equation (16); thus the cancellation calculation and the general hidden-coefficient envelope agree. Moreover,

$$h'(r) = \frac{1}{\sqrt{1-r}\,(1+\sqrt{1-r})^2} > 0,$$

so replacing pointwise p by any valid upper bound preserves the inequality. The envelope maximization is attained at $x = \sqrt{1-p}$, while the substitution $p \leq u/\gamma$ is tight when the conditioning inequalities are equalities.

Finally let

$$\rho = \|v - A_O\|/D, \qquad c = \cos \angle(A_R - A_O, v - A_O),$$

with the cross term defined as zero if either vector vanishes. Squaring Equation (19) and normalizing by D^2 yields

$$\delta_+^2 = \frac{p^2\big((1-p)^2\zeta^2 + x^2\rho^2 + 2x(1-p)\zeta\rho c\big)}{(1+x)^2(1-p+x)^2}. \tag{21}$$

A.3. SHARP FIXED-MASS AND FIXED-MARGIN BOUNDS

Proof of Proposition 2. Put $r = 1 - p$. For $f(x) = (ru + px)/[(1 + x)(r + x)]$, the sign of $f'(x)$ is the sign of

$$r[p - u(1 + r)] - 2rux - px^2.$$

This polynomial decreases on $x \geq 0$. If $p \leq u(1 + r) = (2 - p)u$, the maximum is $f(0) = u$. Otherwise its unique positive root is

$$x_* = \frac{\sqrt{(1-p)(p-u)(p-(1-p)u)} - (1-p)u}{p}, \qquad \Psi(u,p) = \frac{p}{2 - p + 2x_*}.$$

The continuous endpoint conventions are $\Psi(u, 0) = u$ and $\Psi(u, 1) = 1$. Since the optimization ranges over exactly the unobserved append scale, the envelope is sharp at the declared (u, p) information level.

Proof of Theorem 1. The diameter of a set and its convex hull coincide. Thus every conditional output formed from the source and admissible appended values has pairwise distance at most D, and $\zeta, \rho \leq 1$. An empty omitted support has zero error and can be handled separately. For a nonempty omitted support, Equation (19) and the triangle inequality give

$$\delta_+ \leq B_{\text{local}} \leq B_{\text{coeff}}(p, \delta, x).$$

For $0 \leq u < 1$, this last expression is nondecreasing in each of δ and p when δ is replaced by u:

$$\frac{\partial B_{\text{coeff}}(p, u, x)}{\partial p} = \frac{x(1 + x - u)}{(1 + x)(1 - p + x)^2} \geq 0.$$

For every query in Γ_0, $\delta(q) \leq u$ and $p(q) \leq \delta(q)/\gamma \leq r = u/\gamma < 1$. Consequently

$$\delta_+(q) \leq B_{\text{coeff}}(p(q), u, x) \leq F(r, u, x) := \frac{(1 - r)u + rx}{(1 + x)(1 - r + x)}.$$

Since $F(r, u, 0) = u$, its maximum also bounds every pre-append output. Taking the maximum over Γ_1 proves Equation (8).

Maximization and the boundary case. For $u > 0$, put $b = 1 - r > 0$. The sign of $\partial_x F$ is the sign of

$$b[r - u(2 - r)] - 2bux - rx^2.$$

This expression is decreasing for $x \geq 0$. Its value at zero is positive exactly when $2\gamma < 1 + u$. Otherwise F is nonincreasing and its maximum is $F(r, u, 0) = u$. In the interior case, the unique positive critical point is

$$x_* = \frac{\sqrt{(1-r)(r-u)(r-u+ru)} - (1-r)u}{r}. \tag{22}$$

The quadratic $M(1 + x)(1 - r + x) - [(1 - r)u + rx]$ is tangent to zero at x_* when $M = \max_x F$. Setting its discriminant to zero yields

$$r^2M^2 - 2[r(2 - r) - 2(1 - r)u]M + r^2 = 0.$$

The smaller root is Equation (9); its reciprocal form avoids subtracting two nearly equal quantities.

Attainability. In the interior case take one omitted value at 0, one retained value at γD, and one appended value at D on the real line. Give the source rows positive unnormalized masses r and $1 - r$, and the append mass x_*. These masses can be realized by scalar logits at a nonzero query. Then $\zeta = \gamma$, $\rho = 1$, $\delta = r\gamma = u$, both value directions align, and every inequality above is an equality. In the boundary case the pre-append endpoint realizes u; an empty append has the same error, and nonempty appends can approach it as $x \downarrow 0$. This proves sharpness over the stated class, without requiring a particular captured model or finite workload to realize the extremizer.

Finally, Equation (14) gives $\delta_+ \leq \delta + h(p)$, and h is increasing. Thus $\Phi_\gamma(u) \leq u + h(u/\gamma) \to 0$. The sequence $p = \zeta = t$ satisfies $\sup_x \delta_+ \leq t^2 + h(t) \to 0$, showing why a positive fixed margin is a sufficient uniform condition rather than a necessary condition for each stable sequence.

A.4. Approximate-factorization transfer

Let C be a constructed state and R the unchanged-row state on its declared support. Put $s = A_C - A_R$ and $\xi = \|s\|/D$. Since $A_R - A = p(A_R - A_O)$,

$$A_C - A = p(A_R - A_O) + s.$$

If $\zeta = \|A_R - A_O\|/D \geq \gamma$, the reverse triangle inequality gives

$$p\gamma \leq p\zeta \leq \delta_C + \xi, \qquad p \leq (\delta_C + \xi)/\gamma.$$

This information-level bound is sharp by taking s antiparallel to $p(A_R - A_O)$ with prescribed norm.

Write $\eta = \log(S_C/S_R)$ and append common mass G with conditional value v. With

$$\lambda_R = \frac{S_R}{S_R + G}, \qquad \lambda_C = \frac{S_C}{S_C + G},$$

direct subtraction yields

$$A_C^+ - A_R^+ = \lambda_C s + (\lambda_C - \lambda_R)(A_R - v).$$

The diameter assumption gives $\|A_R - v\| \leq D$, hence

$$\frac{\|A_C^+ - A_R^+\|}{D} \leq \xi + |\lambda_C - \lambda_R|.$$

Writing $t = G/S_R$ and maximizing over $t \geq 0$ gives exactly

$$\sup_t |\lambda_C - \lambda_R| = \tanh(|\eta|/4),$$

with maximizer $t = e^{\eta/2}$ after the symmetric reparameterization. The support-only state remains a pure-selection state on the same omitted partition, with current error $p\zeta \leq q = \delta_C + \xi$ and $p \leq q/\gamma < 1$. Since $B_{\mathrm{coeff}}(p, u, x)$ is nondecreasing in both u and p, these inequalities imply $B_{\mathrm{coeff}}(p, p\zeta, x) \leq B_{\mathrm{coeff}}(q/\gamma, q, x)$ for every $x \geq 0$. Taking the supremum over x and using Theorem 1 gives

$$\frac{\|A_R^+ - A^+\|}{D} \leq \Phi_\gamma(q).$$

Adding the displayed constructor-to-support transport proves Equation (12). No joint sharpness claim is made in (δ_C, ξ, η).

A.5. Multihead composition and representation-family scope

Corollary 1 (GQA/MHA composition)**.** *For finitely many query heads, including shared-KV GQA heads, suppose each candidate is an unchanged-row selection satisfying Theorem 1. For fixed headwise diameters, positive visibility margins, and output projection W_O, the all-head discrepancy tends to zero whenever every normalized headwise instantaneous discrepancy tends to zero.*

Let $h = 1, \ldots, H$ index query heads, including heads sharing a KV head under GQA. If unchanged-row selection gives the headwise radius $b_h = D_h \Phi_{\gamma_h}(u_h)$, write the fixed output projection in block columns $W_{O,h}$ and define

$$\mathcal{C}_{W_O}(b) := \sup_{\|z_h\|_2 \leq b_h} \left\| \sum_{h=1}^{H} W_{O,h} z_h \right\|_2 .$$

Norm duality gives the exact support-function form

$$\mathcal{C}_{W_O}(b) = \sup_{\|y\|_2 = 1} \sum_{h=1}^{H} b_h \|W_{O,h}^\top y\|_2. \tag{23}$$

Thus $\mathcal{C}_{W_O}$ is the robust radius determined by independent head balls, not a spectral relaxation. Linearity and Corollary 1's headwise hypotheses give

$$\left\| \sum_h W_{O,h} \Delta_h^+ \right\|_2 \leq \mathcal{C}_{W_O}(b) \leq \min\left\{ \sum_h \|W_{O,h}\|_2 b_h,\ \|W_O\|_2 \Big(\sum_h b_h^2 \Big)^{1/2} \right\}. \tag{24}$$

For fixed finite H, D_h, and W_O, both displayed relaxations vanish when every $u_h \to 0$. Sharing a KV head changes neither the per-query-head softmax decomposition nor $\delta_h = p_h \zeta_h$; it only restricts the jointly feasible directions, which can make the independent-ball relaxation loose.

Proposition 3 (Headwise attenuation after output projection)**.** *Let $W_O = [W_{O,1}, \dots, W_{O,H}]$, $M = W_O^\top W_O$, and $A_a = \mathrm{diag}(a_1 I, \dots, a_H I)$ with $a_h \in [0,1]$. For fixed a, the inequality*

$$\|W_O A_a R\|_2 \le \|W_O R\|_2 \qquad \text{for every } R$$

holds if and only if $M - A_a^\top M A_a \succeq 0$. It holds for every R and every independent block attenuation $a \in [0,1]^H$ if and only if

$$W_{O,i}^\top W_{O,j} = 0 \qquad (i \neq j).$$

The fixed-a statement follows by squaring the norms. Pairwise block orthogonality is sufficient because both squared norms then split into a sum over heads. Conversely, choose $A_a = P_i$, the projection onto head i. Positive semidefiniteness of $M - P_i M P_i$ is necessary, and its ii diagonal block is zero. A positive-semidefinite block matrix with a zero diagonal block has every incident off-diagonal block zero: otherwise its quadratic form on $tu + v$ has a nonzero term linear in t and is negative for one sign of large t. Hence $M_{ij} = W_{O,i}^\top W_{O,j} = 0$ for every $j \neq i$; repeat over i. By continuity the characterization is unchanged if all coefficients are restricted to the open interval $(0,1)$. Nonorthogonal output blocks therefore preclude a universal headwise-contraction-to-endpoint-contraction theorem, even before adding the semantic innovation channel. This does not assert expansion for the coupled coefficients and residual directions realized by a particular GQA workload.

Proposition 4 (Head-radius and intrinsic-Gram boundaries)**.** *Let $W_O = [W_{O,1}, \dots, W_{O,H}]$, with each block acting on $\mathbb{R}^d$.*

(i) Head radii. *The implication*

$$\|z_h\|_2 \le \|r_h\|_2 \ (h = 1, \dots, H) \quad \Longrightarrow \quad \left\| \sum_h W_{O,h} z_h \right\|_2 \le \left\| \sum_h W_{O,h} r_h \right\|_2$$

holds for every r, z if and only if

$$W_{O,h}^\top W_{O,h} = c_h I \quad (c_h \ge 0), \qquad W_{O,i}^\top W_{O,j} = 0 \quad (i \neq j).$$

(ii) Full intrinsic Gram. *For row vectors $z_h^\top$ stacked as $Z \in \mathbb{R}^{H \times d}$, the energy $T_W(Z) = \| \sum_h W_{O,h} z_h \|_2^2$ is determined by $K = ZZ^\top$ for every Z if and only if*

$$W_{O,i}^\top W_{O,j} = c_{ij} I \qquad \text{for every } i, j.$$

In general it has the exact decomposition

$$T_W(Z) = \underbrace{\frac{1}{d} \sum_{i,j} K_{ij} \, \mathrm{tr}(W_{O,i}^\top W_{O,j})}_{\mu_{\mathrm{Haar}}(K, W_O)} + \underbrace{\sum_{i,j} z_i^\top \left(W_{O,i}^\top W_{O,j} - \frac{\mathrm{tr}(W_{O,i}^\top W_{O,j})}{d} I \right) z_j}_{R_{\mathrm{contract}}(Z, W_O)}. \tag{25}$$

The first term is the common-right-orthogonal Haar average at fixed K; this group average makes no distributional assumption about realized states.

For part (i), consider one head while setting all others to zero. Equal-norm exchange in both directions forces $\|W_{O,h} x\|_2$ to depend only on $\|x\|_2$, hence $W_{O,h}^\top W_{O,h} = c_h I$. For $i \neq j$, write $B = W_{O,i}^\top W_{O,j}$. If $B \neq 0$, choose x, y with $x^\top B y < 0$, set $r_i = x$, $r_j = ty$, $z_i = x$, and $z_j = 0$. For sufficiently small $t > 0$, the negative linear cross term makes $\|W_{O,i} r_i + W_{O,j} r_j\|_2^2 < \|W_{O,i} z_i\|_2^2$, contradicting the implication. The stated conditions are sufficient because both endpoint energies then equal weighted sums of squared head norms.

For part (ii), Z and ZU have the same Gram for every $U \in O(d)$. Universal Gram sufficiency therefore requires every bilinear block $W_{O,i}^\top W_{O,j}$ to be invariant under conjugation by $O(d)$; the commutant of the standard orthogonal representation consists exactly of scalar multiples of I. The converse is immediate. Splitting each block into its scalar trace component and traceless remainder gives Equation (25); Haar conjugation is the Reynolds projection onto the scalar component.

Proposition 5 (Noninheritance under synthesis)**.** *The selection factorization and certificate do not extend to arbitrary merging, synthesis, or reweighting states: an equal-output synthesized merge can alter normalization scale, which a common append then exposes.*

For Proposition 5, let two source rows have the same conditional value a and total unnormalized mass S. Replace them by one synthesized row with value a and mass λS, where $\lambda > 1$. Both current outputs equal a. After appending mass $G > 0$ with value $v \neq a$,

$$\widehat{A}^{+} - A^{+} = \frac{SG(\lambda - 1)(a - v)}{(\lambda S + G)(S + G)}, \tag{26}$$

which is nonzero. The witness is realizable: take scalar $q = 1$, two source keys $k_1 = k_2 = 0$ so $S = 2$, synthesized key $\widehat{k} = \log(2\lambda)$, and append key $k_{+} = \log G$. Since $\lambda S > S$, no unchanged subset of the source rows generates the candidate, so selection's p and ζ are not canonical. A mass-preserving or otherwise controlled merge can avoid this witness; the proposition denies automatic certificate transfer, not merge reliability.

A.6. Exact nested-budget attribution

For two nested pure-selection supports, write their normalized pre-append errors as

$$\delta_0 = p_0 \zeta_0, \qquad \delta_1 = p_1 \zeta_1.$$

The symmetric product decomposition

$$\delta_0 - \delta_1 = (p_0 - p_1)\frac{\zeta_0 + \zeta_1}{2} + (\zeta_0 - \zeta_1)\frac{p_0 + p_1}{2} \tag{27}$$

is exact by expansion. For nested supports $p_1 \leq p_0$, so the first term is nonnegative. The empirical budget attribution in the main text uses this identity only on strictly improving points; no causal regression model is introduced.

Appendix B. Capture and statistical protocol

B.1. Populations and common appends

The Gemma discovery population and the first 64-row Qwen/Mistral population support the historical prevalence, geometry, support-control, and four-budget calibration results. The sharper uniform-bound comparison on those Qwen/Mistral captures is a derived analysis of observed outcomes. The new confirmation instead selects 64 previously unused physical rows from LongBench's 2WikiMultihopQA and Qasper datasets, 32 per dataset. From physical rows 0–199 in each dataset, remove the union of all development, Gemma discovery, and earlier Qwen/Mistral rows. Order the remaining rows by a frozen SHA-256 function of a selection seed, dataset, and raw-row hash. The two models use the same selected physical rows, tokenized separately and cropped to the first 128 and last 128 tokens.

For each model and prompt, materialize 256 FullKV source rows. Observe query heads 0 and 16 in every layer, paired with the corresponding grouped-query KV head: 72 attention units for Qwen and 64 for Mistral. Seven greedy FullKV decoding steps supply query/key/value triples. At probe q_t, both source states are evaluated at that same query, then the identical captured prefix of appended rows is included in each state. The prefix is a single append group for this comparison. Thus later prefixes are measured against the original compressed source state; the experiment does not propagate a compressed model's own queries or errors between endpoints.

All four selection constructors preserve a 32-row recent window. The registered natural old-row budget is 36 for Qwen and 32 for Mistral. SnapKV, StreamingLLM, and H_2O use that old-row count in each unit; PyramidKV uses its frozen layer-dependent allocation. H_2O refers to prefill heavy-hitter support frozen before these appends. For the fresh test, point counts are $64 \cdot 72 \cdot 7 \cdot 4 = 129{,}024$ on Qwen and $64 \cdot 64 \cdot 7 \cdot 4 = 114{,}688$ on Mistral.

B.2. Normalization and score information

For each prompt and attention unit, fix

$$D = \max_{a,b \in \mathcal{V}} \|a - b\|_2, \qquad \mathcal{V} = \{\text{256 source values and seven captured append values}\}.$$

The same prefix-independent diameter applies to every support in the unit and bounds its convex-hull diameter. Because D is a maximum diameter, an extreme captured value can lower all normalized errors. A post-hoc leave-one-extreme audit defines D_{-1} by deleting the farthest-pair endpoint whose removal most reduces the remaining diameter, with all endpoint numerators unchanged. Median D/D_{-1} is $1.038/1.034$ and q99 is $1.535/1.610$ on Qwen/Mistral; at $\epsilon = .025$, about 1–2% of low-error-set membership and endpoint-safety labels flip, always toward leaving the low-error set or becoming unsafe, while equal-method concordance improvement remains .291 $[.190, .389]$ and .268 $[.146, .404]$. Thus sensitivity is real, but this fixed perturbation leaves the two-model discrimination conclusion unchanged and does not replace D or the registered analysis. A zero-diameter unit has identically zero response error. With the query-dependent conditional values A_R, A_O, v, the captured coordinates are

$$\zeta_c = \frac{\|A_R - A_O\|_2}{D}, \quad \rho_c = \frac{\|v - A_O\|_2}{D}, \quad u = p\zeta_c, \quad e_+ = \frac{\|\widehat{A}^+ - A^+\|_2}{D}.$$

The subscript c marks full-contract rather than source-only normalization; the theorem uses $\max_{\Gamma_0} u$. Logits include model attention scaling and positional encoding, and retained mass $1 - p$ is summed directly to avoid subtractive loss.

The bound ladder uses progressively different measured information:

$$\begin{aligned}
&S_{\text{read}} = u, && \text{instantaneous normalized error,}\\
&S_{\text{uniform}} = \Phi_\gamma(u), && \text{also the declared visibility margin,}\\
&S_{\text{mass}} = \Psi(u, p), && \text{also omitted mass,}\\
&S_{\text{coeff}} = \frac{(1-p)u + px}{(1+x)(1-p+x)}, && \text{also append/source mass ratio,}\\
&S_{\text{local}} = \frac{p((1-p)\zeta_c + x\rho_c)}{(1+x)(1-p+x)}, && \text{also appended–omitted value contrast.}
\end{aligned}$$

The local score omits the angle between the two value directions. The endpoint uses that angle through Equation (13). The preregistered fresh matched comparison uses S_{read} and S_{coeff}; the Ψ decomposition is retrospective. The coefficient score therefore does not reconstruct the endpoint. However, these coordinates and D use the full source and captured future; freezing a score before collecting outcomes does not make its inputs available to an online compressed decoder.

A point is certified when the chosen bound is at most ϵ. Φ_γ and ω_γ are evaluated only on $G_\gamma = \{\zeta_c \geq \gamma,\ u < \gamma\}$. Coverage divides the number certified by all points in the model–method cell. Safe-state recall divides it by all points with $e_+ \leq \epsilon$. Non-certification is abstention. The uniform comparison fixes $\gamma = .10$ and $\epsilon \in \{.01, .025, .05\}$.

B.3. Matched concordance and uncertainty

For fixed model m, method k, and prompt j, collect all unordered pairs of writes from the same attention unit that meet the primary matching rule in the fresh-prompt comparison. Let P_{mkj} be that pair set, and let $C_S(a, b)$ be concordance for score S, using the unsafe member as the positive endpoint. The prompt contribution is

$$d_{mkj} = \frac{1}{|P_{mkj}|} \sum_{(a,b)\in P_{mkj}} \big[C_{\text{coeff}}(a, b) - C_{\text{read}}(a, b)\big].$$

It is undefined when no eligible pair exists. Let $J_{mk} = \{j : |P_{mkj}| > 0\}$. The registered model estimand is

$$\Delta_m = \frac{1}{4}\sum_{k=1}^{4} \frac{1}{|J_{mk}|} \sum_{j\in J_{mk}} d_{mkj}.$$

The order matters: averaging available methods within each prompt first would weight methods differently when their eligible prompt sets differ. All manuscript equal-method averages use the displayed equal-method estimand.

The bootstrap resamples 32 prompts with replacement within each dataset, using the same sampled prompt indices for all four methods. For each draw, recompute each method mean over its eligible sampled prompts and then average the four method means. A draw is undefined if a method has

no eligible sampled prompt; it is excluded from percentile intervals and its exclusion is reported. These percentile intervals are conditional on all four method means being defined in the resample. At the primary $.025/.10$ setting, $0/20{,}000$ Qwen and $2/20{,}000$ Mistral draws are excluded; sparse robustness cells can condition more strongly, so the sparse-robustness table reports every valid-draw count. The interval uses 20,000 draws with fixed integer PRNG seed `20260916`. The primary gate requires a strictly positive lower endpoint of the paired 95% interval in each model separately. Method-level intervals, low-current-error unsafe prevalence, and eligible prompt/pair counts remain visible alongside the equal-method averages.

B.4. Reproducibility and aggregation correction

The frozen protocol specifies population selection, exclusions, constructors, budgets, scores, matching, aggregation, uncertainty, and reporting before attention outcomes are computed. A separately implemented raw-capture audit reconstructs supports, coordinates, endpoint labels, and summaries directly from frozen inputs rather than from derived analysis outputs. The released compact artifact reports that reconstruction; raw tensor captures are not included in the public package.

The first exported fresh-test equal-method average accidentally averaged methods within prompt. A subsequent correction restored the protocol's prompt-within-method, then equal-method order in both analysis and independent audit. It changed the equal-method summaries but neither raw captures nor method-level results. A public table of 4,608 prompt contributions and a standalone auditor permit exact reconstruction of all 18 equal-method tolerance/caliper summaries and their valid bootstrap counts. That check concerns aggregation of released contributions; the separate raw audit establishes their tensor-level provenance.

The later Ψ analysis reuses the immutable calibration point archive. Its first model-level decomposition row pooled selected slice points but carried an equal-method label. The corrected artifact preserves that pooled-slice estimand and separately computes method-first equal-weight estimates and paired prompt bootstraps. The two mass-share summaries differ by at most $.0013$; the full-population coverage and matching outputs are byte-identical to the initial audited result.

B.5. Fresh approximate-factorization confirmation

The transfer theorem, $\gamma = .10$, tolerance $.025$, public-CaM primary, KeepKV secondary, prompt-level estimands, and pass gates were frozen before either model was loaded. The population contains 32 previously unused physical LongBench rows per model, 16 from 2WikiMultihopQA and 16 from Qasper, selected by a deterministic, outcome-blind SHA-256 ordering after excluding rows used in the prior all-head architecture-control population and a separate prior GQA-composition population. Qwen3-8B and Mistral-7B use the same physical rows but tokenize them independently.

All layers, all 32 query heads, seven FullKV common-prefix writes, natural old-row budgets, and four frozen public-CaM seeds are included. Public CaM and its support comparator have identical keys, unit weights, and cardinality, so $\eta = 0$ pointwise. KeepKV changes keys and effective multiplicities and therefore retains the normalization term. The prompt is the uncertainty unit; 95% intervals use 20,000 dataset-stratified prompt bootstraps with a fixed PRNG seed recorded in the anonymized supplementary artifact. An independently implemented auditor reconstructs every prompt contribution and gate from the frozen outputs. Exact version bindings are recorded in the anonymized supplementary artifact.

B.6. Long-context native-continuation stress tests

We denote Spearman rank correlation by ρ_s, reserving ρ for the appended-value contrast in Section 3. The original protocol was frozen before outcomes and finalized only after an independent audit; exact version bindings are recorded in the anonymized supplementary artifact. It keeps the two primary model families fixed, uses the longest available prompt prefix up to 8192 model-native tokens, compresses once after prefill, and generates 32 greedy tokens without recompression. The primary CaM/support state keeps the same absolute natural budget: 36/32 old rows plus a 32-row recent window on Qwen/Mistral, i.e., 68/64 rows before native appends. A full 8192-token prefix therefore forms an intentionally aggressive sub-1% retention stress test rather than a proportionally scaled budget. Three paths share the same initial physical prompt: FullKV, a common-write control

driven by FullKV continuation writes, and a native path whose compressed state generates its own queries, keys, values, and tokens. The contrast between the latter two measures the cost of leaving the common-append contract.

After the registered exclusions, all 14 remaining physical rows per model (seven 2WikiMultihopQA and seven Qasper) form the frozen original population; no observed row is replaced. Public CaM is the primary constructor. The two-model directional gate requires both a positive association between frozen certificate occupancy C and first-divergence time T_{div} and preservation of the common-write/native error ordering. As frozen, this is a sign gate on the point estimates; bootstrap intervals calibrate uncertainty and are not pass/fail criteria. F1@32 is the registered multiset token-overlap F1 between native and FullKV 32-token continuations; it is a secondary diagnostic, not a primary gate coordinate.

Model	Cert. occ.	T_{div}	F1@32	$\rho_s(C, T_{\mathrm{div}})$	$\rho_s(E_c, E_n)$	$E_c/E_n/\Delta$
Qwen3-8B	.3042	9.14	.5067	.5131	.9648	.2902/.4165/.1264
Mistral-7B	.2609	5.50	.4448	$-.0736$	.9033	.4022/.5791/.1769

Table 3: Original frozen public-CaM native-continuation audit. E_c and E_n are mean aligned common-write and native errors; $\Delta = E_n - E_c$. The 95% prompt-bootstrap intervals for $\rho_s(E_c, E_n)$ are $[.8643, 1.0000]$ and $[.7472, .9822]$ on Qwen/Mistral. The corresponding intervals for $\rho_s(C, T_{\mathrm{div}})$ are $[-.1769, .8934]$ and $[-.5660, .4939]$.

Qwen satisfies both registered directions, whereas Mistral preserves the common/native ordering but not the certificate-to-first-divergence direction. The original two-model gate therefore fails and remains visible rather than being pooled away.

Independent cross-dataset replication. Before model execution, a second population was frozen and later finalized only after an independent audit. It uses 32 physical prompts per model—16 SHA-ranked HotpotQA rows and 16 SHA-ranked MuSiQue rows—with the same physical rows for both models, the same one-compression 32-token execution, and 20,000 dataset-stratified prompt bootstraps. No observed row is refilled. An independent auditor rebuilds the primary correlations and intervals directly from hash-verified prompt shards and matches the finalized summary exactly.

Model	Cert. occ.	T_{div}	F1@32	$\rho_s(C, T_{\mathrm{div}})$	$\rho_s(E_c, E_n)$	$E_c/E_n/\Delta$
Qwen3-8B	.2681	7.52	.5132	.2196	.9377	.3391/.4534/.1144
Mistral-7B	.2651	11.14	.5698	.1418	.9384	.3106/.4011/.0905

Table 4: Independent HotpotQA/MuSiQue replication. The 95% dataset-stratified bootstrap intervals for $\rho_s(E_c, E_n)$ are $[.8463, .9749]$ and $[.8437, .9745]$ on Qwen/Mistral; those for $\rho_s(C, T_{\mathrm{div}})$ are $[-.1486, .5567]$ and $[-.2410, .4763]$.

Both models pass the frozen directional gate in the independent replication. The common/native ordering intervals are entirely positive, whereas the certificate-to-lifetime intervals still cross zero. Thus the fresh population recovers the registered direction and independently confirms ordering persistence, but does not turn certificate occupancy into a high-confidence or universal closed-loop lifetime predictor.

An exploratory post-hoc horizon profile, added only after the registered analysis had been finalized, keeps $\rho_s(E_c, E_n) > .91$ at every measured horizon $h \in \{1, 2, 4, 8, 16, 32\}$ in both models, while the native–common error excess grows with horizon. This profile is descriptive and does not modify the preregistered gate. Across both frozen populations, the supported conclusion is therefore ordering persistence under one 32-token native continuation, not repeated-compression safety, a closed-loop certificate, or an online lifetime predictor.

Appendix C. Supplementary empirical results

Here $u = p\zeta_c$ is current-output error, $e_+ = \|\widehat{A}^+ - A^+\|/D$ is realized common-append error, and $G_\gamma = \{\zeta_c \geq \gamma, u < \gamma\}$. Coverage is certified/all points; safe-state recall is certified/realized-safe points.

C.1. Original certificate operating-point sensitivity

The original full-population calibration froze its tolerance grid before endpoint computation. A derived audit extracts the natural-support-budget slice at the registered visibility margin $\gamma = .10$; $.025$ remains the registered primary. Each range is over the eight Qwen/Mistral model–method cells. Table 5 reports the operating-point sensitivity.

ϵ	ω cov.	ω cov. $\mid G_{.10}$	Local cov.	Local recall	Always-safe prec.
.010	.085–.201	.138–.242	.334–.510	.913–.952	.356–.535
.025	.165–.344	.258–.414	.566–.776	.928–.970	.597–.802
.050	.249–.492	.402–.592	.770–.925	.960–.986	.802–.941

Table 5: Tolerance sensitivity of the certificate comparison. Every nonempty certificate has precision one, and local coverage exceeds uniform coverage in all 24 cells. Always-safe is a classifier baseline whose precision equals the realized-safe rate, not a certificate; always-abstain is sound with zero recall. These slices test numerical robustness; they do not assign task-utility meaning to ϵ.

C.2. Population geometry and the selected-tail correction

At the natural Gemma budget, a low-current-error point is called mass-dominated when $p \le \zeta$, so omitted mass is the smaller factor in $\delta = p\zeta$. The prompt-balanced mass-dominated fractions are 0.856 for SnapKV, 0.874 for PyramidKV, and 0.789 for StreamingLLM. Conversely, the preregistered high-mass/low-visibility diagnostic occupies only 2.6%, 0.8%, and 9.7% of the corresponding low-current-error populations. Thus instantaneous agreement is normally obtained by removing little query-relevant mass rather than by hiding a large omission through value cancellation.

The initially striking exposed-tail localization does not supply an independent test. Of 122 registered Gemma events, 121 satisfy $p \ge .025/q_{.25}$; the product identity and low-current-error condition then force $\zeta \le q_{.25}$ before the post-append endpoint is examined. Removing these mechanically contained points leaves one residual event and no resolved enrichment. Full-population quartiles still show that low visibility can carry a heavier extreme tail, but typical exposure is not monotone in ζ. We therefore retain cancellation as the sharp structural boundary while dropping the earlier tail p-value as mechanism evidence.

Increasing support budget moves the states along the same coordinate. For two nested budgets, the exact product identity

$$\delta_0 - \delta_1 = \underbrace{(p_0 - p_1)\frac{\zeta_0 + \zeta_1}{2}}_{M} + \underbrace{(\zeta_0 - \zeta_1)\frac{p_0 + p_1}{2}}_{Z}$$

separates omitted-mass and visibility contributions without a regression model. Across all adjacent budget transitions, the fraction of improving points with $M > Z$ is 0.927–0.988 for SnapKV, 0.902–0.967 for PyramidKV, and 0.773–0.844 for StreamingLLM. The practical budget trajectory is therefore primarily $p \downarrow$, not improved instantaneous cancellation.

C.3. Matched support controls

Small realized p does not by itself show that a selector is better than a support-size counterfactual: the recent window or the workload may already make old-row mass small. We therefore compare every public support with the exact expectation of a uniformly random old-row support of the same unit-level cardinality, and with a fixed workload oracle that chooses the support minimizing mean omitted mass over the seven registered future queries. Oracle recovery is the fraction of the matched-random-to-oracle omitted-mass gap closed by the public support.

On Gemma, SnapKV and PyramidKV beat matched random at every budget. At the natural budget their omitted masses are 0.195 and 0.222, compared with matched-random values 0.416 and 0.424, and they recover 0.650 and 0.624 of the available fixed-workload-oracle advantage. StreamingLLM is a useful boundary: its natural-budget recovery is 0.020 with a confidence interval crossing zero, and its two registered datasets have opposite signs. The state-geometry criterion is therefore more general than a single account of how every selector reaches it.

For a fixed support indicator r_i with $\sum_i r_i = k$ and registered workload mean future mass μ_i, the exact gain over cardinality-matched random is

$$\Delta_R = \sum_i (r_i - k/n)\mu_i = n \operatorname{Cov}_i(r_i, \mu_i).$$

This coordinate separates support/workload alignment from the downstream omitted mass itself. Practical methods need not optimize the future workload oracle for this alignment to be measured. Figure 2 visualizes this constructor–model–workload dependence across the registered populations.

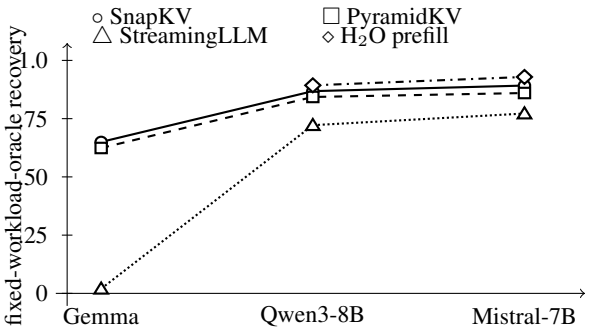


Figure 2: Natural-budget fixed-workload-oracle recovery. SnapKV and PyramidKV show strong support/workload alignment across all three registered populations, whereas StreamingLLM moves from near-null on Gemma to strong positive alignment on the fresh Qwen/Mistral populations; its Gemma bootstrap interval crosses zero. H_2O-prefill arm was added prospectively only in the fresh confirmation. Alignment is therefore a property of the realized constructor–model–workload combination, not of a method name alone.

C.4. Fresh-prompt robustness grid

The frozen grid varies $\epsilon \in \{.01, .025, .05\}$ and the matching caliper fraction $\ell \in \{.05, .10, .20\}$, with $|u_i - u_j| \leq \ell\epsilon$. These are overlapping analyses of the same prompts. Sparse cells can leave many bootstrap draws undefined, so the number of valid draws is reported for every interval. The $.025/.10$ setting remains primary. Table 6 reports the full tolerance/caliper grid.

Model	ϵ	ℓ	Difference	95% CI	Valid / 20,000
Qwen	.010	.05	.321	[.102, .489]	19,162
Qwen	.010	.10	.236	[.075, .387]	19,165
Qwen	.010	.20	.166	[.060, .268]	19,723
Qwen	.025	.05	.329	[.225, .439]	20,000
Qwen	.025	.10	.320	[.236, .411]	20,000
Qwen	.025	.20	.274	[.201, .347]	20,000
Qwen	.050	.05	.353	[.141, .589]	19,131
Qwen	.050	.10	.235	[.097, .340]	19,131
Qwen	.050	.20	.211	[.150, .271]	19,894
Mistral	.010	.05	.319	[.189, .463]	12,627
Mistral	.010	.10	.308	[.170, .498]	19,071
Mistral	.010	.20	.209	[.105, .322]	19,873
Mistral	.025	.05	.434	[.232, .603]	19,178
Mistral	.025	.10	.211	[.040, .382]	19,998
Mistral	.025	.20	.164	[.025, .308]	20,000
Mistral	.050	.05	.528	[.385, .688]	12,600
Mistral	.050	.10	.501	[.405, .602]	17,303
Mistral	.050	.20	.242	[.132, .359]	19,967

Table 6: Corrected equal-method concordance differences on the fresh population. Positive intervals across the grid are a robustness check, not 18 independent replications.

C.5. Retrospective architecture and representation controls

These controls use a separate all-head capture archive of a different, non-overlapping set of 64 physical prompts. The earlier calibration audit measures two query heads per layer and matches safe–unsafe one-write endpoints at low headwise current-output error. The subsequent all-head controls instead use all 32 query heads, eight shared KV groups, the real output projection, and pairs matched on continuous projected block error. A cache-repeat control copies each of the eight KV-group caches to 32 query-head caches before applying the public CaM prefill update. The derived protocols were frozen before their outcomes, but none is a new population confirmation.

For the GQA block audit, let $R_h^0 = \widehat{A}_h - A_h$, $T_h = v_h - A_{O,h}$, $r_h = 1 - p_h$, and define the exact headwise channels

$$V_h = \frac{r_h R_h^0}{(1 + x_h)(r_h + x_h)}, \qquad H_h = \frac{p_h x_h T_h}{(1 + x_h)(r_h + x_h)}.$$

Thus the one-write residual is $V_h + H_h$. Concatenate the headwise channels as V, H, partition W_O into blocks $W_{O,h}$, and let $\sigma_h = \|W_{O,h}\|_2$. For the same prompt and layer, D_{block} is the RMS norm of the seven prompt-only and seven post-write FullKV projected outputs. It is constructor-independent and differs from the single-head value diameter D, so the two normalized .025 thresholds are not numerically interchangeable. Starting from the sum of headwise spectral radii, we successively retain the real projected direction of the visible channel, the hidden channel, or both:

$$\begin{aligned}
B_{\text{sep}} &= \sum_h \sigma_h(\|V_h\| + \|H_h\|)/D_{\text{block}}, \\
B_{\text{prejoint}} &= (\|W_O V\| + \sum_h \sigma_h \|H_h\|)/D_{\text{block}}, \\
B_{\text{postjoint}} &= (\sum_h \sigma_h \|V_h\| + \|W_O H\|)/D_{\text{block}}, \\
B_{\text{channel}} &= (\|W_O V\| + \|W_O H\|)/D_{\text{block}}.
\end{aligned}$$

All obeyed the registered pointwise hierarchy, including $E_+ \le B_{\text{channel}}$. Here E_+ is the output-projected block error normalized by D_{block}, distinct from the single-head e_+ normalized by value-set diameter D in Appendix B.2. Table 7 reports bound-minus-output concordance on output-matched pairs. Retaining only visible joint direction produced strongly negative concordance increments. Indeed, B_{prejoint} is pointwise no larger than B_{sep} yet ranks endpoints substantially worse, so tighter bounds need not be better scores. Retaining both channels' joint directions was positive, but uses nearly the full directional transition: median E_+/B_{channel} was .841–.875. This localizes why the structural Corollary 1 need not yield a useful low-information block certificate; it does not rule out every such certificate. The registered five-percent coverage clause was infeasible because only .012–.577% of points were realized-safe at the block tolerance. The negative concordance, rather than this protocol-design failure, is the informative endpoint.

An exact-radius follow-up asks whether the information interface, rather than a particular relaxation, is already too coarse. At every realized-stable endpoint it sets $b_h = \|V_h + H_h\|_2$ and constructs explicit alternatives inside $\prod_h \{z_h : \|z_h\|_2 \le b_h\}$. Such an alternative exceeds the realized pre-write endpoint norm for 97.71–99.07% of stable Qwen endpoints and 99.40–99.65% of stable Mistral endpoints across methods; every prompt-bootstrap lower endpoint exceeds 97.44%. Thus any certificate sound over the complete product balls must abstain on those points. The witnesses need not be reachable under shared GQA dynamics, so this is insufficiency of the independent-radii interface, not evidence that realized endpoints expand.

Model	B_{sep}	B_{prejoint}	$B_{\text{postjoint}}$	B_{channel}
Qwen	.0573	$-.\mathbf{1785}$	.0637	.1707
Mistral	$-.0644$	$-.\mathbf{2630}$	$-.0235$	.0711

Table 7: Equal-method concordance increments in the retrospective GQA block audit. For B_{prejoint}, the 95% prompt-bootstrap intervals are $[-.1893, -.1682]$ and $[-.2709, -.2548]$; for B_{channel} they are $[.1615, .1796]$ and $[.0617, .0803]$.

A separate sequential audit reuses 32-prompt capture bundles from a prior GQA-composition audit; those captures were outcome-seen before this derived decomposition. For Z equal to the realized hidden-channel head tuple, Equation (25) is reconstructed pointwise. Table 8 shows that the noninvariant contract term is systematic rather than merely algebraically possible. The dominant positive contribution is already within each head's relation to its own output block; missing cross-GQA-group Gram relations are secondary. Thus the full intrinsic Gram captures the main energy scale but not the finite-scale contract response.

Model	Diagonal	Same-group cross	Cross-group	Total R_{contract}
Qwen	.1555	$-.0040$	.0206	.1722 [.1349, .2146]
Mistral	.1357	$-.0173$	.0290	.1474 [.1377, .1571]

Table 8: Equal-method means of the exact Reynolds residual sectors, each normalized pointwise by the block-Gram Haar mean. Intervals are dataset-stratified, prompt-clustered 95% intervals with 20,000 replicates.

The representation control transcribes the public CaM prefill update and compares it with an unchanged-row arm using exactly CaM's selected keys. Keys, support, cardinality, queries, seven common-prefix writes, and the real W_O are shared; only CaM's synthesized values differ. External key/value parity was exact on the frozen check. For any common horizon query, let $A_{\mathrm{CaM},h}$ and $A_{\mathrm{supp},h}$ evaluate the two pre-extension states, whose common key set has mass $S_{R,h}$, and let the shared appended prefix have mass G_h. Then

$$A^{+}_{\mathrm{CaM},h} - A^{+}_{\mathrm{supp},h} = \frac{S_{R,h}}{S_{R,h} + G_h} \left(A_{\mathrm{CaM},h} - A_{\mathrm{supp},h}\right).$$

The between-arm effect is therefore transported value redistribution, not a hidden mass mismatch. As Table 9 shows, CaM adds a small positive mean block error in both models, while the registered growth contrast is negative, consistent with dilution along this changing-query path rather than accumulation.

Model	Mean gap [95% CI]	Relative gap	Growth [95% CI]
Qwen	.003173 [.003014, .003329]	1.96%	$-.001507$ [$-.001703$, $-.001320$]
Mistral	.002131 [.002028, .002232]	1.04%	$-.000577$ [$-.000649$, $-.000504$]

Table 9: CaM merge-minus-support-only normalized projected block error. Mean gap averages writes 1–6; growth averages the change from write 0. All 64 prompt-average gaps per model are positive, whereas no prompt has positive growth.

A separate one-write control on the all-head archive gives KeepKV-minus-own-support gaps of $-.02171$ $[-.02379, -.01959]$ on Qwen and $-.02825$ $[-.02949, -.02691]$ on Mistral: its merging improves on retaining exactly its chosen rows. CaM and KeepKV therefore exhibit constructor-specific representation effects with opposite signs. The constructed counterexample in Proposition 5, rather than either empirical control, establishes noninheritance of the selection certificate.

The all-head controls use all layers and 64 prompts per model; the later contract-residual audit uses 32 prompts and eight registered layers per model. Reported intervals use 20,000 dataset-stratified prompt-bootstrap replicates. These controls concern architecture-level transition behavior; classic equal-Q/K/V-head MHA behavior, online prediction, repeated recompression, free rollout, end-to-end task utility, runtime effects, and generic merge-constructor behavior lie outside their empirical contract.

Appendix D. Broader related-work map

Field-facing mechanism map. Following recent KV-cache surveys, we keep selection, budget allocation, merging, quantization, and low-rank or learned compression separate (Li et al., 2025). Sparse access is listed separately because it can reduce query-time work without constructing a smaller replacement cache. Table 10 overlays the replacement-state question on these families; the overlay is descriptive and does not imply a shared recovery mechanism.

Table 10: Field-facing KV-cache mechanisms viewed through the replacement-state question. The engineering families follow distinctions summarized by Li et al. (2025); sparse access is split out here because it need not materialize a smaller replacement cache. The final column is our replacement-state overlay.

Field-facing mechanism	Representative engineering route	Reduced-state view	Relation to the present theory
Selection / eviction / retention	Scissorhands, H2O, TOVA, Keyformer, StreamingLLM, SnapKV, PyramidInfer/PyramidKV, FastGen, and RazorAttention (Liu et al., 2023; Zhang et al., 2023; Oren et al., 2024; Adnan et al., 2024; Xiao et al., 2024; Li et al., 2024; Yang et al., 2024; Cai et al., 2025; Ge et al., 2024; Tang et al., 2025).	Retains a subset of native rows; importance, recency, sink, or related scores determine what persists.	Unchanged-row selection is the exactly analyzed reference geometry; recovery is controlled through omitted mass and its visibility.
Budget / head–layer allocation	HeadKV, Ada-KV, and DuoAttention redistribute retention across heads, layers, or roles (Fu et al., 2025; Feng et al., 2025; Xiao et al., 2025).	Changes where retention capacity is spent and is often composed with a row-selection rule.	Allocation is orthogonal to the local recovery law; the theorem concerns the reduced state produced in each declared unit.
Sparse access / retrieval	QUEST, SparQ, and ClusterKV query only a subset or indexed view (Tang et al., 2024; Ribar et al., 2024; Liu et al., 2025).	Can reduce attention-time access while retaining a broader backing cache.	Not necessarily a state-substitution mechanism; the replacement-state contract applies when a smaller replacement cache is actually constructed.
Merging / fusion / aggregation	CaM, MiniCache, KeepKV, and LoCoCo move or aggregate information into a modified cache (Zhang et al., 2024; Liu et al., 2024a; Tian et al., 2026; Cai et al., 2024).	Changes retained values, support-relative mass, or aggregate state rather than merely deleting native rows.	Support-anchored cases inherit the one-append guarantee when the measured (ξ, η) defects are controlled; generic merging is not automatically certified.
Low-rank / recurrent / learned compression	LESS and DMC use auxiliary, recurrent, or learned compressed representations (Dong et al., 2024; Nawrot et al., 2024).	A richer compressed representation may carry history that is absent from an unchanged-row subset.	It faces the same transition-sufficiency question but may recover through mechanisms outside the present selection certificate.
Quantization / coding	KIVI, KVQuant, and CacheGen change the numerical storage or coding of K/V (Liu et al., 2024c; Hooper et al., 2024; Liu et al., 2024b).	Broadly preserves support while perturbing or coding the stored numerical representation.	The stored state still faces the replacement contract, but the present selection theorem does not directly certify numerical recoding.

Orthogonal criteria and diagnostics. Error criteria, allocation objectives, and empirical diagnostics cut across these mechanism families and provide complementary measurements rather than a single recovery law (Feng et al., 2025; 2026; Peng et al., 2026; Shen et al., 2025; Chen et al., 2026; Qiu et al., 2026; Luo et al., 2026). Our restricted recovery likewise complements worst-case linear-space barriers for arbitrary autoregressive attention states (Haris & Onak, 2025).