UTILITY, COVERAGE, AND EXACT RESIDENCY CAN RANK KV-CACHE POLICIES DIFFERENTLY

Anonymous authors
Paper under double-blind review

ABSTRACT

What should a compressed KV cache preserve? The answer depends on the claim: behavior on a target query, retained annotated content, or intact original slots for a literal-residency claim. These are different targets, yet KV-cache evaluations often treat answer quality or token coverage as evidence for all three. FACT-KV aligns tested-query utility, original-position coverage, and exact annotated-unit residency on the same selected state. In a prospectively frozen 500-source Mistral/HotpotQA contract, cumulative retention improves F1 over recency by 0.220 and coverage by 0.086, yet reduces all-units completeness by 0.348. The conflict also occurs within sources: 70 show both higher F1 and lower completeness. An official HotpotQA re-score of the frozen predictions gives similar F1 ($+0.217$) and EM ($+0.198$) advantages. A placement reference reproduces much of the residency contrast, implicating retention geometry rather than semantic destruction. Capacity and granularity stress tests move exact residency systematically in the tested settings, without a corresponding monotone improvement in answer utility. A contract-adapted H2O audit, together with Llama, reuse, attribution, and repair audits, establishes the boundary: exact annotated-unit residency is a preservation-audit target, not a utility surrogate. Preservation claims should name the target, report its direct readout, and declare the layer/head aggregation rule.

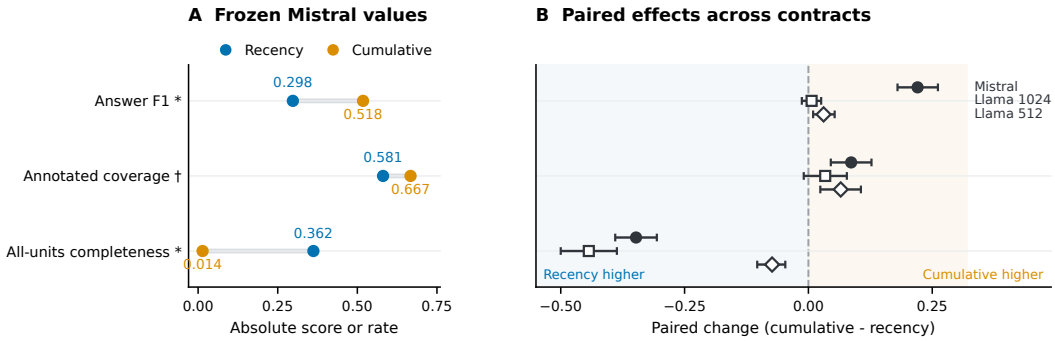

Figure 1: **Claim-specific readouts disagree under the same cache comparison. A:** Absolute values in the frozen Mistral contract. **B:** Cumulative-minus-recency effects for the same endpoints. The negative residency direction repeats in Llama, while its 1,024-token F1 effect is uncertain. Panel B uses direct labels: a filled circle denotes Mistral primary, a hollow square Llama-1,024 scope, and a hollow diamond Llama-512 sensitivity. * Frozen primary endpoint; † exploratory readout.

1 INTRODUCTION

A cache evaluation should specify what its preservation claim concerns: behavior on a tested query, retained annotated quantity, or literal availability of intact original units. These objectives need not agree: F1 measures the first and evidence-token coverage the second; neither measures whether an annotated unit remains intact. Scattered selection can cover many evidence positions while completing no unit, whereas a contiguous policy can retain fewer positions but preserve whole units. We call the latter property *exact annotated-unit residency*: a physical original-slot measurement, not a claim of semantic loss. The paper therefore studies a geometry-sensitive measurement mismatch: scat-

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.

tered token selection and contiguous retention can differ sharply under exact-span residency even when utility or retained annotated quantity favors the scattered policy.

This distinction changes what evidence a preservation claim requires. A compressor may be useful while fragmenting annotated units, and a policy that keeps more complete units need not answer better. Favorable F1 or coverage therefore cannot certify exact annotated-unit residency, while that residency cannot stand in for utility. FACT-KV makes these category errors testable by aligning all three readouts on the same source, selected state, and original-position trace. It audits preservation claims; it is not a benchmark of latency, throughput, or deployable memory efficiency.

Our frozen primary comparison uses Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) on 500 HotpotQA sources (Yang et al., 2018) at a 1,024-position budget. Relative to recency, cumulative-attention retention improves answer F1 by 0.220 and annotated-token coverage by 0.086, yet reduces all-units completeness by 0.348 (Figure 1). The conflict also appears within individual sources: 70 of 500 sources have both higher F1 and lower completeness under cumulative retention. A revision-stage frozen 300-source Llama-3.1-8B audit reproduces the residency ordering at the primary budget, but its F1 contrast is small and uncertain. Across these settings, the exact-residency ranking transfers more reliably than the behavioral ranking.

The practical consequence is a reporting contract, not a universal cache objective: systems may optimize tested-query utility, but exact annotated-unit residency claims also require a direct readout and explicit aggregation rule. This claim-to-readout contract organizes the measurement contribution and determines what each reported result can support.

Contributions. **1. Audit contract and tools.** FACT-KV aligns tested-query utility, retained annotated quantity, and exact annotated-unit residency with original-position traces, validators, and CPU replay. **2. Policy-ranking conflict.** A prospectively frozen panel establishes that better F1 and coverage can accompany fewer complete evidence units, including within the same sources. **3. Design stresses and boundaries.** Capacity and granularity stresses show that residency and utility respond differently to cache design; H2O, Llama, reuse, attribution, and repair audits delimit the claim's scope. **Evidence status.** The Mistral panel is the sole prospective confirmation; capacity, granularity, H2O, Llama, reuse, attribution, and repair provide scope or boundary evidence.

2 Claim-to-Readout Contract

Table 1 separates the preservation outcomes. Selectors are gold-invisible at inference time; annotations define evaluation units, except in explicitly labelled oracle interventions. On HotpotQA, units are annotated supporting-fact sentences, not paragraphs. The ordering is robust to alternative unit boundaries and within-unit thresholds; these are post-outcome checks, not alternative confirmatory endpoints (Appendix H.1).

Table 1: **Claim-specific reporting rules.** One favorable readout cannot substantiate the other preservation claims.

Claim and readout	Directly measures	Does not justify
Tested-query utility: F1/EM	Behavior under a declared execution contract	Literal annotated-unit residency
Retained quantity: original-position coverage	Retained annotated quantity	Intact-unit survival
Exact annotated-unit residency: complete-unit fraction	Fraction of annotated units retained intact under an explicit aggregation rule	Semantic necessity, causal use, or future utility
Exact annotated-unit residency: all-units completeness	Fraction of sources retaining every annotated unit intact	Graded damage or semantic sufficiency

Figure 2 gives a two-unit example in which equal token coverage accompanies different complete-unit fractions. For source s, let F_{si} be annotated unit i, $U_s = \bigcup_i F_{si}$, and K_s the retained original

positions after final selection. With m_s units,

$$C_s = \frac{|U_s \cap K_s|}{|U_s|}, \qquad S_{si} = \mathbf{1}[F_{si} \subseteq K_s], \qquad I_s = \frac{1}{m_s}\sum_i S_{si}, \qquad A_s = \prod_i S_{si}.$$

C_s measures retained position quantity, I_s measures unit-level residency, and A_s is the strict source-level claim-compliance endpoint and frozen primary residency endpoint. It is neither a semantic-preservation measure nor a continuous preservation score. For nonempty U_s and shared K_s, $A_s = \mathbf{1}[C_s = 1]$. Although this identity holds within a source, the dataset mean of C_s does not determine the fraction of sources satisfying $A_s = 1$. Readouts average equally over sources; annotations need not be minimal sufficient support.

Metric ladder. A_s is deliberately conjunctive, so we also report progressively less brittle unit-level, thresholded, stratified, and alternative-boundary variants. Their common negative direction is a post-outcome robustness result; the frozen endpoint remains unchanged. Thresholds and their provenance are given in Appendices H.1 and O.

Primary execution contract. Documents enter in chunks and undergo repeated KV compaction under a fixed 1,024-position budget. Both policies protect the first 16 and latest 128 resident positions. Cumulative retention sums last-layer attention over heads and the newly admitted chunk's queries; recency selects the same unprotected capacity by position. Original-coordinate maps are updated after every selection. After final selection, the last prompt token is recomputed against the selected cache before greedy decoding, aligning answer F1 and residency to the same selected state. Because retained KV states were contextualized before later eviction, absence of an original slot does not imply absence of all information previously mixed into surviving states; exact residency is therefore a physical-slot endpoint, not a representation-level information endpoint. Figure 2 summarizes this alignment. Appendix B specifies prompts, scoring, numerical checks, and the distinct contracts used by the native-method and reuse audits.

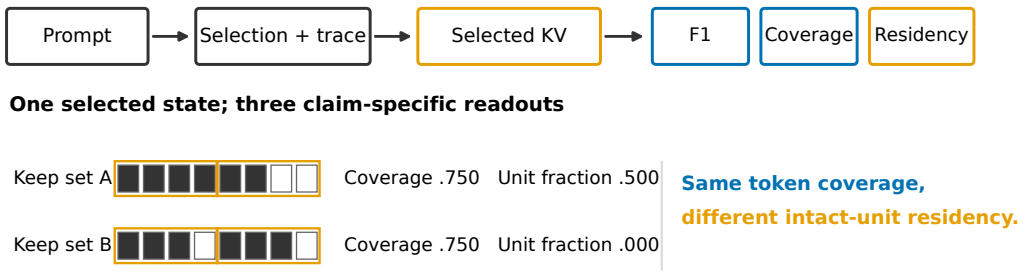

Figure 2: **One selected state supports three distinct audit questions.** Chunk admission and policy selection produce the final KV and original-position trace. Answer F1 uses the recomputed final prompt token; coverage and residency use the trace. The matched-coverage construction below shows that equal retained quantity can accompany unequal complete-unit residency.

3 Frozen Separation: Better Answers, Less Literal Residency

We compare recency and cumulative retention on 500 Mistral-7B-Instruct-v0.3/HotpotQA sources at a 1,024-position budget. The cohort, two primary endpoints, and decision rules were frozen before formal inference; engineering sources and formal sources are disjoint, and every source runs both policies. Intervals use 10,000 source-paired bootstrap draws. This controlled contrast separates readouts rather than ranking contemporary compressors. The frozen primary F1 uses the paper's token-overlap scorer. A post-outcome robustness analysis re-scores the unchanged predictions with the pinned official HotpotQA answer evaluator: cumulative minus recency is $+0.217$ $[0.175, 0.258]$ for F1 and $+0.198$ $[0.156, 0.242]$ for EM. Thus the behavioral direction and uncertainty conclusion are not artifacts of the custom normalization, while the frozen scorer remains the primary

endpoint. These are answer metrics under our generation contract, not leaderboard submissions (Appendix O).[1]

3.1 Utility and exact annotated-unit residency rank policies differently

Cumulative increases F1 from 0.298 to 0.518, while all-units completeness falls from 0.362 to 0.014. Table 2 reports the paired effects and the exploratory coverage and complete-unit readouts. The direction is unchanged under the legacy readout (Appendix B.3).

Table 2: Frozen 500-source comparison: cumulative minus recency, with paired 95% intervals. F1 and all-units completeness are frozen primary endpoints; coverage and complete-unit fraction are post-freeze exploratory summaries. Full-cache mean F1 is 0.578.

Readout	Recency	Cumulative	Difference [95% CI]
Answer F1*	0.298	0.518	$+\mathbf{0.220}$ $[\mathbf{0.180}, \mathbf{0.261}]$
Token coverage†	0.581	0.667	$+\mathbf{0.086}$ $[\mathbf{0.045}, \mathbf{0.127}]$
Complete-unit fraction†	0.583	0.130	$-\mathbf{0.453}$ $[-\mathbf{0.491}, -\mathbf{0.415}]$
All-units completeness*	0.362	0.014	$-\mathbf{0.348}$ $[-\mathbf{0.390}, -\mathbf{0.306}]$

3.2 The conflict occurs within sources

The predeclared event $\Delta\mathrm{F1} > 0$ and $\Delta A_s < 0$ occurs on 70/500 sources (14.0%; 95% interval 11.0%–17.2%). F1 improves on 241 sources, declines on 102 and ties on 157. Figure 3A includes all nine paired sign combinations: the joint event is not a prevalence estimate of a universal conflict, and the remaining sources need not show the opposite ordering. Source-level correlations and complete transition accounting appear in Appendix H and the artifact.

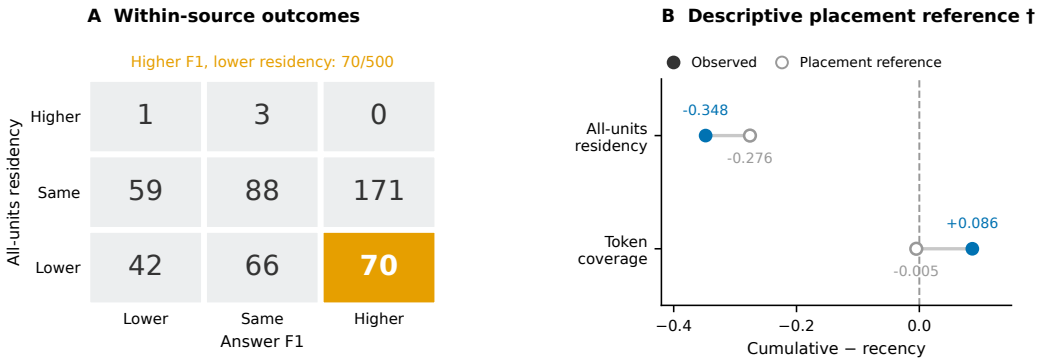

Figure 3: **Within-source conflict and a descriptive placement reference. A:** Seventy sources simultaneously have higher F1 and lower all-units completeness under cumulative retention. All nine cells contain all 500 sources. **B:** Solid circles are observed contrasts and open circles are position-blind placement-reference contrasts. The reference reproduces 79.3% of the residency gap but reverses the coverage direction. This is a descriptive gap ratio, not a causal explained fraction. Definitions and uncertainty are in Appendix D.4.2.

3.3 Placement reproduces much of the residency contrast

Holding observed keep sets fixed, the position-blind placement reference independently relocates every annotated unit as a contiguous block of the same length, with its start uniform over an estimated document extent; the observed keep set remains fixed. It preserves keep-set shape, unit count and lengths, but not observed positions, inter-unit spacing or overlap. This reference reproduces 79.3% of the observed all-units gap (-0.276 versus -0.348), but predicts coverage -0.005 rather than the observed $+0.086$. A start-decile-matched construction, which preserves coarse unit position, gives an all-units contrast of -0.351. The references implicate placement geometry without estimating causal mediation; neither preserves the full joint unit layout. Thus the result diagnoses an

[1] Official-coordinate sensitivity leaves the primary contrasts and 70/500 count unchanged (Appendix H.1).

exact-span audit failure of dispersed selection; it does not show that cumulative attention uniquely destroys semantically meaningful evidence. Protection-region, length, and threshold diagnostics leave the sign unchanged but are post-outcome checks, not protection or admission-schedule ablations (Appendices H, K, and H.1).

4 Design Stress Tests: Capacity and Granularity

We next organize two complementary stress tests around cache capacity and selection granularity (Figure 4). They are boundary evidence, not a second prospective confirmation: both were designed after the primary outcomes.

Capacity changes magnitude, not the category of conflict. On the same 500 Mistral sources, cumulative minus recency remains positive for F1 and negative for all-units residency at 512 positions ($+0.217$ versus -0.074) and 2,048 ($+0.030$ versus -0.098); all four 95% intervals exclude zero. At 2,048, however, 383/500 prompts fit without compaction. Capacity changes effect magnitude and exposure to compression, but not the observed sign conflict over the three tested budgets.

Coarser granularity increases complete-unit residency. If dispersed token selection limits exact residency, retaining scored positions in contiguous blocks should increase complete-unit survival. In a 300-source Mistral/HotpotQA panel, corresponding token-level and block arms share the scorer, budget, protections, and readout but differ in selection granularity; neither observes gold annotations (Appendix I.2). As width increases from 4 to 8 to 16, the complete-unit fraction gains rise descriptively from 0.136 to 0.206 to 0.261. F1 does not improve monotonically: its corresponding differences are -0.025, -0.005, and -0.011. Every residency interval is above zero, whereas every F1 interval crosses both zero and the -0.03 non-inferiority margin (Appendix J.1). Coarser selection therefore produces a descriptively monotone exact-residency response, but no tested width establishes F1 benefit, F1 harm, or the predeclared joint success rule.

Because the two arms induce different future attention trajectories, this intervention tests retention granularity as a policy change, not geometry as an isolated causal mediator. It does not establish a completed compressor design, and fixed-width blocks do not test sentence- or paragraph-aligned selection.

5 Boundary Tests: Residency Is Not a Policy Selector

We test whether higher completeness selects the better policy for a future query. In a separately frozen MuSiQue reuse experiment (Trivedi et al., 2022), Q0 states branch before decoding, Q0 answers do not enter Q1 history, and Q1 is answered without document re-prefill. Cumulative retention has zero all-units completeness versus 0.303 for recency, yet achieves higher Q1 F1: 0.419 versus 0.240, a paired difference of 0.179 $[0.114, 0.244]$ (Figure 5A). This reverses the predeclared directional hypothesis: higher exact annotated-unit residency does not identify the better future-query policy under this contract.

The oracle restoration inserts omitted Q1-support KV from an immutable full-prefill snapshot and improves F1 over a matched sham by 0.260 $[0.197, 0.327]$; every restored support paragraph contains a normalized reference-answer string. Corrected answer-excluded V2 is unresolved (0.022 $[-0.018, 0.060]$). Oracle restoration therefore establishes value for omitted answer-bearing states, not for answer-excluded context or completeness itself (Appendices I and E).

5.1 Attribution and repair do not collapse the boundary

The preceding F1 endpoint scores only the answer. A separately frozen 300-source joint-output panel therefore requires one generation to return an answer and supporting-sentence IDs. This contract differs from the primary panel, so the experiment is a boundary test rather than a second prospective confirmation (Appendix M).

The Answer-F1 contrast is $+0.010$ here, versus $+0.220$ in the primary panel. Under this new prompt, numerical contract, and document-disjoint cohort, the behavioral contrast is uncertain while

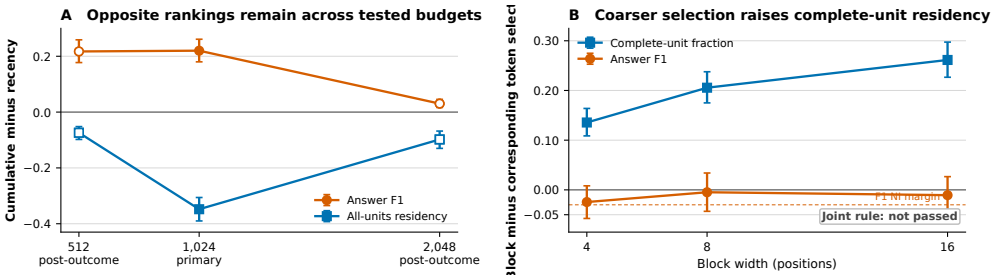

Figure 4: **Design knobs move exact annotated-unit residency and utility differently. A:** Cumulative minus recency across Mistral capacities. The filled 1,024-position estimates are prospective primary endpoints; open 512/2,048 markers are post-outcome sensitivities. **B:** Block minus corresponding token-level selection across post-outcome width sensitivities; width 16 is the original separately frozen structured intervention. All intervals are source-paired 95% intervals. The dashed line is the -0.03 F1 non-inferiority margin. Panel B shows marginal results. No width satisfies the predeclared joint coverage-equivalence, residency-superiority, and F1 non-inferiority rule.

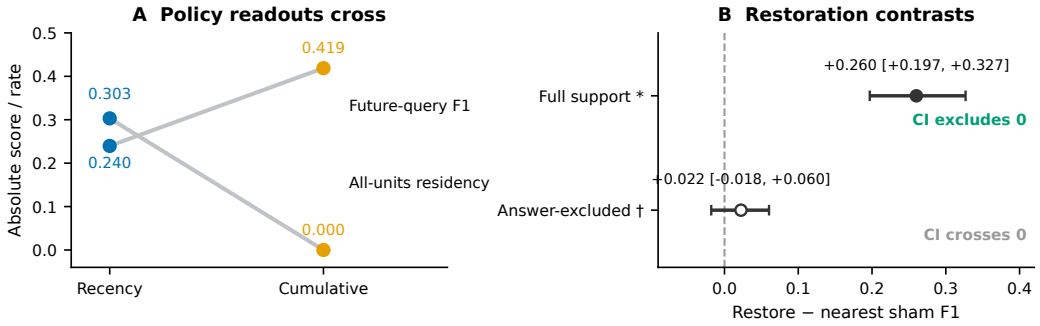

Figure 5: **Higher completeness need not imply higher utility. A:** Policy rankings reverse between future-query F1 and exact completeness. **B:** Full-support restoration has a CI excluding zero, while the corrected answer-excluded V2 CI crosses zero. * Frozen co-primary intervention; † corrected post-primary exploratory intervention. Intervals resample 85 overlap groups; they do not estimate a mediation effect of completeness.

Table 3: Revision-stage 300-source answer-and-attribution boundary test. Bracketed values are paired 95% bootstrap intervals for cumulative minus recency.

Policy	Answer F1	Joint F1	All-units residency
Recency	0.448	0.175	0.193
Cumulative	0.458	0.168	0.003
Cumulative − recency	$+0.010$ $[-0.044, 0.063]$	-0.0069 $[-0.044, 0.031]$	-0.190

the exact-residency ordering remains negative (-0.190). This is consistent with asymmetric transfer of exact-residency and behavioral rankings, but the contract and cohort changes prevent a clean transfer comparison.

The exact-residency separation is large, but the predeclared behavioral reversal is not established. We then formed offline, gold-informed repair plans for all 266 cumulative states admitting three equal-budget branches. Complete-unit restoration (R), scattered-evidence restoration that completes no previously incomplete unit (S), and a position-matched non-evidence sham (P) remove the same resident positions and restore exact admission-time KV states. R minus S Joint F1 is -0.0013 $[-0.0041, 0.0004]$; no advantage of the complete-unit restoration rule is established. The position-matched sham contrast is descriptive because the fixed testing sequence stopped at the failed policy reversal. The intervention changes token identity across branches and therefore does not isolate completeness as a causal mediator. Protocol, eligibility, residency checks and full results appear in Appendix M.

The R–S result does not show that complete and scattered repairs are semantically equivalent; both branches change token identity and contextual state. It also does not identify a causal mediator.

6 PORTABILITY AND RELEASED-METHOD SCOPE

The experiments have distinct evidential roles: primary findings, interventions, diagnostics, and scope audits. Four historical cohorts show the same F1/all-units direction under their original contracts, but none is another prospective confirmation (Appendix D, Table 10).

The revision-stage frozen Llama-3.1-8B scope audit provides a direct boundary on generalization (Table 4). At the primary 1,024-position budget, cumulative retention again sharply lowers all-units completeness relative to recency (-0.443 $[-0.500, -0.387]$), but its F1 difference is small and uncertain ($+0.006$ $[-0.013, 0.026]$). At the prespecified 512-position sensitivity budget, both a positive F1 contrast ($+0.031$ $[0.010, 0.053]$) and a negative residency contrast (-0.073 $[-0.103, -0.047]$) are observed. The prespecified primary decision is nevertheless structural-only replication: the residency ordering transfers to the new model and cohort, whereas the primary-budget behavioral ranking does not. Because model, runtime, and cohort change together, this audit does not identify architecture as the cause of the difference. Across the two 1,024-position model panels, the residency sign repeats while the F1 advantage is established only in the Mistral primary panel. This is a descriptive asymmetry in portability across the audited contracts, not a claim of universal residency invariance.

Table 4: Cross-model and budget scope (cumulative minus recency). The Llama 512 row is a prespecified sensitivity and cannot replace the 1,024-position primary decision. Coverage contrasts are descriptive in this compact table; their intervals and absolute means are in Appendix L.

Panel	Budget	ΔF1 [95% CI]	ΔC	ΔA [95% CI]	Interpretation
Mistral primary	1,024	$+0.220$ $[+0.180, +0.261]$	$+0.086$	-0.348 $[-0.390, -0.306]$	F1↑, residency↓
Llama scope	1,024	$+0.006$ $[-0.013, +0.026]$	$+0.034$	-0.443 $[-0.500, -0.387]$	Residency↓; F1 uncertain
Llama sensitivity	512	$+0.031$ $[+0.010, +0.053]$	$+0.065$	-0.073 $[-0.103, -0.047]$	F1↑, residency↓

6.1 A CONTRACT-ADAPTED H2O EXTENSION

Under the same 500-source, 1,024-position audit, contract-adapted H2O reproduces the F1–residency separation: relative to recency, F1 increases by 0.2289 $[0.1888, 0.2686]$, while the mean-head complete-unit residency contrast is -0.5052 $[-0.5387, -0.4720]$. The released Llama code does not directly support Mistral GQA or original-coordinate chunk admissions, so we froze a minimal compatibility adapter before the formal run: query-head attention mass is summed within each KV-head group, while H2O's cumulative score, irreversible top-k rule and 128-position recent window are retained. The adapter also protects the shared 16-position prefix. A 16/16 engineering gate verified exact full-cache reference tokens, trace-on/off equality, repeatability, coordinate validity and exact resident bytes. We therefore call this a *contract-adapted H2O GQA* audit, not the untouched upstream binary.

Table 5 reports the frozen result. Its coverage contrast is $+0.0262$ $[-0.0104, 0.0624]$. Thus H2O establishes the prespecified two-readout F1–residency separation, but not the stricter three-readout replication because the coverage interval includes zero (STRUCTURAL-ONLY under the frozen decision vocabulary). H2O's union all-units residency is 0.996, whereas its intersection is 0. Union, intersection, and mean-head therefore answer different physical questions: "exists somewhere," "exists everywhere," and typical-store residency. Their source distribution appears in Appendix O. Appendix N gives the protocol, paired intervals, amendments and stopped-run provenance.

Table 5: Matched 500-source audit at 1,024 resident positions per layer/KV head. H2O residency uses the frozen mean-head aggregation. Bytes are BF16 prompt-state K+V per source.

Policy	F1	Coverage	Mean-head unit fraction	Mean-head all units	MiB
Recency	0.298	0.581	0.583	0.362	128
Cumulative	0.518	0.667	0.130	0.014	128
Contract-adapted H2O GQA	0.526	0.607	0.078	0.005	128

Related retention and evaluation methods. Closest token and layer/head selectors include H2O, SnapKV and PyramidKV (Zhang et al., 2023; Li et al., 2024; Cai et al., 2025). ChunkKV, ShotKV, SABlock and SemantiCache make granularity or semantic structure explicit (Liu et al., 2025; 2026a; Chen et al., 2025; Wu et al., 2026), although their chunks need not match annotated units. Position-aware and learned-query selectors include DapQ, Judge Q and globally calibrated eviction (Tian et al., 2026; Liu et al., 2026b; Bui et al., 2026). KVCompose and Meta-Soft use composite or transformed states (Akulov et al., 2025; Luo et al., 2026), for which original-slot residency is incomplete without a representation fidelity endpoint. CacheGen addresses quantized cache streaming rather than token residency (Liu et al., 2024). KVDiagnosis, *Does Accuracy Equal Evidence?*, and an offline-cache quantization audit examine related gaps among behavior, evidence use and grounding (Qiu et al., 2026; Ai et al., 2026; Asad et al., 2026). These works broaden the native-method scope that remains unmeasured here; detailed comparisons appear in Appendix I.

Heterogeneous-cache scope. A post-outcome KVCache-Factory port audit gives union all-units completeness 1.0 and intersection 0.0 for both SnapKV and PyramidKV, with mean-head between them. This is aggregation sensitivity, not a matched-memory ranking (Appendix J.2). A stricter revision-stage SnapKV gate could not jointly verify official-reference equivalence, matched bytes and the primary readout, so we stopped before formal inference (Appendix P).

7 Reporting Rules and Scope

A preservation claim should name its target before choosing its readout (Table 1). Utility calls for behavioral evaluation; retained annotated quantity calls for original-position coverage; exact annotated-unit residency calls for its direct measurement. For heterogeneous caches, union means "exists anywhere," intersection means "in every audited store," and mean-head is a distributional summary. Transformed states need a separate fidelity endpoint.

Cache capacity, selection granularity, and layer/head aggregation should therefore be treated as part of the preservation contract, rather than implementation details hidden behind a single score.

Limitations. The prospective primary result covers one 7B model, task and budget; Llama does not transfer the primary-budget F1 advantage, and a second prospective task remains untested. Native-method and systems coverage is incomplete, including matched structured, learned-query, value-aware and transformed-state methods. The placement references omit the full joint layout, annotated units need not be minimal or sufficient, and the interventions do not isolate completeness as a causal mediator. Evidence-infeasible regimes, where the annotated union itself exceeds capacity, remain untested. We claim no semantic loss, deployment gain or universal compressor ranking.

8 Conclusion

FACT-KV separates tested-query utility, retained annotated quantity, and exact annotated-unit residency. The primary panel exposes their ranking conflict; design stresses and boundary audits show where it persists. Much of the exact-residency contrast is geometry-sensitive, not evidence of semantic destruction. The reporting rule is simple: name the target, measure it directly, and state the aggregation contract. Utility cannot certify intact residency, and residency cannot select the best utility policy.

Ethics Statement

This work evaluates cache-compression behaviour on public benchmarks (RULER, HotpotQA, Qasper and MuSiQue) (Hsieh et al., 2024; Yang et al., 2018; Dasigi et al., 2021; Trivedi et al., 2022) and openly released models. The new Qasper task extension was stopped at a pre-GPU feasibility gate. It uses existing public benchmark texts and synthetic planted facts and conducts no human-subject experiments. The benchmark texts are used under their released licences and carry whatever biases their sources contain, which this paper does not attempt to characterize. Structural absence alone does not establish factual unreliability. We release the audit so others can inspect the same distinction under their own execution contracts.

Reproducibility Statement

The anonymized supplementary ZIP submitted with the paper provides versioned traces, validators, scripts and frozen manifests linking estimates to source records, including the post-outcome official-score, threshold-ladder and H2O aggregation scripts. CPU replay verifies saved outputs rather than rerunning inference. Frozen protocols, post-outcome analyses and stopped branches are distinguished; Appendix G states the replay boundary.

AI Use Statement

Generative AI tools were used as assistive tools for brainstorming and critique of experimental designs, drafting and debugging analysis code, literature discovery, figure preparation, and editing and organizing manuscript text. Their suggestions, including candidate interpretations, were treated as unverified inputs. The authors approved all protocols and revisions, supervised experiment execution, and checked the reported statistics and citations against saved source records.

References

Mengting Ai, Jingrui He, and Yue Guo. Does accuracy equal evidence? reasoning faithfulness under KV cache compression. *arXiv preprint arXiv:2608.01631*, 2026.

Dmitry Akulov, Mohamed Sana, Antonio De Domenico, Tareq Si Salem, Nicola Piovesan, and Fadhel Ayed. KVCompose: Efficient structured KV cache compression with composite tokens. *arXiv preprint arXiv:2509.05165*, 2025.

Asaad Althoubi. DistillCache: KL-guided adaptive KV-cache eviction for memory-efficient LLM inference. *arXiv preprint arXiv:2608.08878*, 2026. URL https://arxiv.org/abs/2608.08878.

Atta Ul Asad, Ahsan Bilal, Muhammad Ali, Muhammad Haseeb, and Dean F. Hougen. Faithfulness is not free: Auditing offline KV-cache quantization in retrieval-augmented generation. *arXiv preprint arXiv:2608.30996*, 2026.

Changwoo Baek, Seungjun Shin, and Kyeongbo Kong. RestoreKV: Recovering full-cache behavior under aggressive query-agnostic KV cache eviction. *arXiv preprint arXiv:2608.01247*, 2026.

Ngoc Bui, Hieu Trung Nguyen, Arman Cohan, and Rex Ying. Make each token count: Towards improving long-context performance with KV cache eviction. *arXiv preprint arXiv:2605.09649*, 2026.

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. URL https://openreview.net/forum?id=ayi7qezU87.

Hong Chen, Xiang Liu, Yubo Gao, Yuxuan Fan, Bo Wang, Yuanlin Chu, Yuanguo Lin, and Xuming Hu. NestedKV: Nested memory routing for long-context KV cache compression. *arXiv preprint arXiv:2605.26678*, 2026a. URL https://arxiv.org/abs/2605.26678v2. Version 2.

Hong Chen, Yudong Zeng, Yongwei Huang, Zuhao Ouyang, Dongnan Zheng, Junyan Zhang, and Xuming Hu. TwinKV: A composable repair pass for KV cache eviction via pairwise key redundancy. *arXiv preprint arXiv:2608.27128*, 2026b.

Jinhan Chen, Jianchun Liu, Hongli Xu, Xianjun Gao, and Shilong Wang. SABlock: Semantic-aware KV cache eviction with adaptive compression block size. *arXiv preprint arXiv:2510.22556*, 2025.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

Raghavv Goel, Junyoung Park, Mukul Gagrani, Dalton Jones, Matthew Morse, Harper Langston, Mingu Lee, and Chris Lott. CAOTE: KV cache selection for LLMs via attention output error-based token eviction. *arXiv preprint arXiv:2504.14051*, 2025. URL https://arxiv.org/abs/2504.14051v6. Version 6.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy. arXiv:2404.06654.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Jang-Hyun Kim, Jinuk Kim, Sangwoo Kwon, Jae W. Lee, Sangdoo Yun, and Hyun Oh Song. KVzip: Query-agnostic KV cache compression with context reconstruction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5585.

Yucheng Li, Huiqiang Jiang, Qianhui Wu, Xufang Luo, Surin Ahn, Chengruidong Zhang, Amir H. Abdi, Dongsheng Li, Jianfeng Gao, Yuqing Yang, and Lili Qiu. SCBench: A KV cache-centric analysis of long-context methods. In *International Conference on Learning Representations*, 2025. arXiv:2412.10319.

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*, volume 37, 2024. doi: 10.52202/079017-0722.

Xiang Liu, Zhenheng Tang, Peijie Dong, Zeyu Li, Yue Liu, Bo Li, Xuming Hu, and Xiaowen Chu. ChunkKV: Semantic-preserving KV cache compression for efficient long-context LLM inference. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0966.

Xiang Liu, Zhenheng Tang, Hong Chen, Peijie Dong, Zeyu Li, Xiuze Zhou, Bo Li, Xuming Hu, and Xiaowen Chu. Semantic integrity matters: Benchmarking and preserving high-density reasoning in KV cache compression. In *International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=1fGD4sKs8j. arXiv:2502.01941.

Yijun Liu, Yixuan Wang, Yuzhuang Xu, Shiyu Ji, Yang Xu, Qingfu Zhu, and Wanxiang Che. Judge Q: Trainable queries for optimized information retention in KV cache eviction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026b.

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*, 2024. doi: 10.1145/3651890.3672274.

Wei Luo, Yi Huang, Songchen Ma, Huanyu Qu, Jiang Cai, and Mingkun Xu. Meta-soft: Leveraging composable meta-tokens for context-preserving KV cache compression. *arXiv preprint arXiv:2605.22337*, 2026.

Tho Mai and Joo-Young Kim. Reformulating KV cache eviction problem for long-context LLM inference. *arXiv preprint arXiv:2605.07234*, 2026. URL https://arxiv.org/abs/2605.07234.

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.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Shuvendu Roy, Mengyao Zhai, Hossein Hajimirsadeghi, and Golnoosh Samei. Coverage-driven KV cache eviction for efficient and improved inference of LLM. *arXiv preprint arXiv:2606.29563*, 2026. URL https://arxiv.org/abs/2606.29563.

Jack Shi and Jerry Gu. Trust the mass: Forced weights in KV-cache eviction. *arXiv preprint arXiv:2608.25230*, 2026.

Zhenxu Tian, Yi Su, Juntao Li, and Min Zhang. Where matters more than what: Decoding-aligned KV cache compression via position-aware pseudo queries. *arXiv preprint arXiv:2603.11564*, 2026.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Shunlong Wu, Hai Lin, Shaoshen Chen, Tingwei Lu, Yongqin Zeng, Shaoxiong Zhan, Hai-Tao Zheng, and Hong-Gee Kim. SemantiCache: Efficient KV cache compression via semantic chunking and clustered merging. *arXiv preprint arXiv:2603.14303*, 2026. URL https://arxiv.org/abs/2603.14303.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Ruijie Zhang, Haozhe Liang, Da Chang, Li Hu, Fanqi Kong, Huaxiao Yin, and Yu Li. When Does Value-Aware KV Eviction Help? A Fixed-Contract Diagnostic for Non-Monotone Cache Compression. *arXiv preprint arXiv:2605.08234*, 2026.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1506.

APPENDIX ROADMAP

The appendix retains material needed to assess a claim and moves development paths, hashes and verbose logs to the artifact README. The evidence chain is: positioning and prior work; primary protocol and original-coordinate validation (Appendices B–C); primary robustness and geometry (Appendices D, H); reuse and attribution (Appendix E); structured, Llama and joint/repair scope (Appendices I.2, L, M); released-method audits and stopped gates (Appendices J.2, N, P); and reproducibility boundaries (Appendix G). Each experiment retains its own contract and evidence status; cross-contract effects are not pooled, and a position budget does not imply equal bytes or serving cost.

Table 6: **Evidence status of the manuscript's claims.** Only the Mistral panel supplies the original prospective confirmation; "not established" is not evidence of equality.

Evidence level	Result	Interpretation
Confirmed primary result	The frozen Mistral panel has higher F1 and coverage but lower exact annotated-unit residency; a post-outcome official re-score preserves the behavioral direction	The three readouts rank the two policies differently under the primary contract
Scope/boundary evidence	Capacity and granularity stress the separation; Llama, H2O, reuse, attribution and repair delimit portability and use	Literal residency is neither free nor a universal utility selector
Unresolved	Native-method breadth, semantic sufficiency, second-task confirmation, and system efficiency	No universal compressor or deployment claim

A EXPANDED CLAIM-LEVEL POSITIONING

Measurement. KVDiagnosis combines original-position coverage, outputs and failure diagnostics, including attention boosts with sham controls (Qiu et al., 2026). Its ERR averages evidence-token retention over layer/head slots and ECov thresholds span coverage at 0.5; neither is the exact all-token unit indicator used here. *Does Accuracy Equal Evidence?* studies behavioral answer–evidence consistency and perturbation faithfulness rather than physical KV residency (Ai et al., 2026). FACT-KV's specific addition is the matched native-KV ranking of answer utility, evidence-token coverage and exact completeness.

Reuse and repair. SCBench and KVzip evaluate shared-context reuse (Li et al., 2025; Kim et al., 2025); RestoreKV and TwinKV study learned or budget-preserving restoration (Baek et al., 2026; Chen et al., 2026b); and Value-Aware separates support recovery from output-value ranking (Zhang et al., 2026). Our MuSiQue test is an annotation-aware oracle boundary experiment. It diagnoses omitted answer-bearing KV states and does not propose a repair method or establish completeness as a future-query policy selector.

Serving and implementation contract. Trust the Mass separates attention-output approximation from retained memory and query information in deployed comparisons (Shi & Gu, 2026). Our matched position budgets are experimental selector interventions under a declared harness, rather than comparisons among complete deployed systems at equal measured bytes. Original-position tracking, validation and mapping make the reported ordering auditable; the contribution is the empirical ordering, not the individual instrumentation components.

Coverage-aware and multi-scale retention. K-VEC improves token diversity across heads and layers (Roy et al., 2026); its coverage is not coverage of a gold evidence-token union. NestedKV combines global, block and sliding-window key anchors (Chen et al., 2026a). These are relevant native-method comparators, but neither result establishes the exact-unit ordering measured here. No native FACT-KV evaluation of either method is reported.

B DATA, PROMPTS, BUDGETS AND EXECUTION CONTRACTS

B.1 PANEL POPULATIONS AND SCORING

The Mistral 7B technical report (Jiang et al., 2023) introduces v0.1; we cite it as the model-family reference, while our runs use the Mistral-7B-Instruct-v0.3 checkpoint identified by the frozen manifests. Qwen runs use Qwen2.5-7B-Instruct (Qwen et al., 2024); exact model and inference-library revisions are recorded separately from the model-family citations.

A source is a document–question instance. Multiple Qasper instances can share a paper, so paper clusters, not questions treated as independent, are resampled. The native panels contain 240 instances per model, drawn under their frozen prompt/tokenization rules. HotpotQA retains 762 Qwen and 781 Mistral instances from the respective 800-source manifest pools. These are model-specific eligible populations, not a common-source comparison between models. All saved eligible sources enter the paired policy analysis; we do not select on F1 to form the native headline.

Table 10 consolidates cohort sizes, model/task identity, effect estimates, and evidence status. Absolute F1 is contract-specific and is not pooled across model-specific prompts or populations.

Native analyses join saved answers and F1 by source ID. Scores use the saved panel's accepted-reference and token-normalization contract. For the new post-selection and structured panels, the frozen scorer lowercases text, extracts `[a-z0-9]+` tokens, computes multiset token-overlap F1 and takes the maximum over accepted references. It does not remove articles and is not the official HotpotQA scorer; the paired results are contract-specific, not directly comparable to benchmark leaderboard F1. Complete annotated completeness uses all declared evidence units, not answer-token matching. A correct answer is not used to infer missing coordinates or evidence sufficiency. The native gate and integrity checks precede these paired summaries; original comparator outputs, including adverse and null effects, are retained in the machine-readable artifact.

Historical retained-token intervention panels use separate re-encoding contracts and are not evidence for the native-KV headline. Their populations, conditional gains, regressions and routing limits are summarized in the artifact's retained-token archive. Missing full-pool intervals are not synthesized from subgroup averages.

B.2 NATIVE PROMPT ADMISSION AND POLICY COMPARISON

All four native panels count 1024 retained original prompt positions, including the trailing question and protected tokens. Policies protect 16 prefix positions and 128 recent resident positions within that budget. Recency fills the remaining capacity from recent original positions; cumulative attention fills it by accumulated attention mass on resident positions. The tested cumulative signal sums last-layer attention over heads and current query positions during each admitted chunk and retains past mass for surviving positions. It is the documented intervention in these runners, not a claim of implementation equivalence to every released attention-based compressor. Qasper additionally saves item-token-max; HotpotQA saves key-norm but has no token-max arm.

The prompt contains the document followed by the current question and answer cue. Earlier chunks cannot see a question that has not arrived. Qasper admits 128-token chunks throughout; HotpotQA initially admits up to 1024 and then 128-token chunks. At overflow, the runner slices the current KV along its sequence dimension to the selected original positions, preserving their coordinate association. Comparators share sources, budget, arrival schedule, prompt and scoring within each panel. Panel-specific prompt formatting and generation caps remain part of the contract; they are not averaged away.

The HotpotQA runner asks for an answer based on the supplied passages without extra words, followed by passages, question and answer cue; it tokenizes this concatenated string with `add_special_tokens=False`, without a chat wrapper. Mistral/Qasper similarly requests a concise answer based only on the scientific article, followed by article, question and answer cue. Qwen/Qasper reconstructs its frozen prompt and checks alignment against the saved prompt hash. Exact strings and tokenizer choices are preserved in the four native runners included in the release, avoiding a new normalized prompt that would silently change the experiment.

Which state produces the first token? Each last prefill forward pass appends the final prompt chunk and saves its last-position logits. If that admission overflows the budget, final compaction happens *after* those logits were computed. The first answer token is selected from the saved logits. Later answer tokens consult the final compacted contextualized KV and append their own KV; decoding does not trigger further eviction. The original completeness endpoint describes the final prompt residency after compaction, before those generated positions are counted. Answer and final-residency measurements are therefore not perfectly time-aligned for the first token. Equal treatment across arms does not eliminate a possible dependence of effect magnitude on this contract. Appendix B.5 tests the available first-readout sensitivity without changing the original final-residency estimand.

B.3 Frozen new-source post-selection-readout validation

Before engineering, we froze the source-selection rule, exclusions, endpoints, decision categories and Qwen extension rule. We audited the available experiment records and version-controlled history for prior source exposure, starting from the earlier input-only exposure ledger. The audit excluded every located historical HotpotQA source ID, every candidate sharing a normalized document with an exposed source, and duplicate documents within the new set. From 3,710 candidates before tokenizer eligibility, a deterministic hash ordering selected 500 sources whose prompts exceed 1024 tokens and whose evidence aligns under both pinned tokenizers. No model outcome for any selected source was found in the audited project records before the design freeze. This is a claim about the audited project records, not model pretraining exposure or unavailable historical records.

The Mistral engineering suite used 16 disjoint historical sources and required all cells to finish, exact reproduction of saved legacy answers and scores, identical selected positions and residency across readout contracts, bitwise-identical non-readout KV, and identical full-cache first-token and complete sequences. After it passed, the final code, model revision `c170c708...`, analysis, and 500-source manifest were frozen before formal inference. Timestamps are project-local and are not a publicly certified registration.

Each source has five conditions: full cache under post-selection readout; recency and cumulative attention under both legacy and post-selection readouts. Post-selection readout first completes final KV selection, removes the retained final prompt token, recomputes that token at its original absolute position against the other retained contextualized KV, stores its new KV, and starts generation. This changes the readout token's representation and KV, not only its logits; all other retained KV remain bitwise identical. The primary population contains all 500 sources, without conditioning on full-cache correctness. Source-paired bootstrap intervals use 10,000 draws with seed 20260916.

Table 7: Frozen 500-source Mistral/HotpotQA validation. Cumulative minus recency; marginal paired 95% bootstrap intervals.

Readout	ΔF1	Δ all-units completeness	F1$\uparrow$ & completeness$\downarrow$
Legacy	$0.222\ [0.181, 0.264]$	$-0.348\ [-0.390, -0.306]$	73/500 (14.6%)
Post-selection	$0.220\ [0.180, 0.261]$	$-0.348\ [-0.390, -0.306]$	70/500 (14.0%)

The post-selection-minus-legacy policy-effect interaction is $-0.002\ [-0.016, 0.011]$. Post-selection recency and cumulative F1 are 0.298 and 0.518; all-units completeness is 36.2% and 1.4%. Full-cache F1 is 0.578 and exact match is 38.8%. On the 194 full-cache-exact sources, the primary F1 and completeness contrasts are $0.415\ [0.342, 0.488]$ and $-0.284\ [-0.350, -0.220]$; this is a predeclared secondary subset.

The frozen primary classification is `CONFIRMATORY_SUPPORT`: both primary intervals exclude zero in their specified directions. The result does not establish that evicted evidence causes an answer change. The earlier Qwen2.5-7B engineering suite preserved the full-cache first-token argmax on all 16 sources but changed the complete sequence on 7/16, violating the frozen equivalence gate. We therefore did not run that Qwen2.5-7B formal extension or amend the gate after observing the engineering result. The separately specified post-outcome 72B panel and its numerical-contract amendment are reported in Appendix J.3.

An independent CPU audit reconstructs all 2,500 source–condition records, re-scores raw answers and rebuilds evidence unions and completeness from keep sets, without importing the formal analyzer or inference core. It finds no missing, duplicate, failed, extra, scoring-mismatched or completeness-mismatched records. It reproduces ΔF1 0.220420, Δ completeness -0.348000 and

the readout interaction -0.001871. Independently seeded intervals are $[0.178, 0.262]$ for ΔF1 and $[-0.0155, 0.0107]$ for the interaction; the latter is not an equivalence test, since no margin was frozen.

B.4 Analysis status, pairing and uncertainty

Same-source joint events, alternative definitions, position strata and new completeness intervals are post-outcome analyses on the four historical panels. The 781-source manifest predates formal outcomes but includes pilot results; excluding the pilot does not establish a pre-pilot freeze. Detailed chronology and experiment-specific local frozen protocols remain in the artifact; Table 10 carries the corresponding evidence-status labels.

Native headline F1 and damaged–intact intervals retain the original panel analyzers. New accounting uses 10,000 paired bootstrap draws, seed 20260913, resampling whole Qasper papers or HotpotQA sources. Means weight sources: resampled sums are divided by the corresponding resampled source counts. The same draws preserve arm pairing and conditional contrasts. These are marginal 95% percentile intervals, not simultaneous discoveries. Empty bootstrap strata are omitted and counted, and very small strata are flagged; Qasper's single-source cumulative-intact groups are not persuasive reciprocal robustness tests. Qwen/Qasper original intervals are exploratory saved-analysis estimates from `results/qwen_qasper_same_contract_v1.json`. The separately seeded replay supplies sensitivity intervals, not exact regeneration of those archived intervals (Appendix G).

Original conditional studies retain their declared paired resampling units: papers for Qasper and source rows for HotpotQA. Conditional estimates are exploratory associations, not effects of inducing structural damage.

B.5 First-token readout alignment sensitivity

This exploratory replay preserves the historical final-residency estimand. For both 240-source Qasper panels, actual original-coordinate eviction traces reconstruct residency at the forward pass producing first-token logits by undoing only the final eviction. All sources pass count, coordinate and saved-keep-set checks; paired arms have equal pre-final capacity, which can exceed 1024. This is a measurement-time sensitivity, not another fixed-budget comparison. Historical HotpotQA lacks the required position traces. Saved answers lack generated token IDs, so an exact same-first-token subset cannot be certified by re-tokenizing text.

Table 8: Exploratory first-readout sensitivity on Qasper, $n = 240$ sources per model (149 Mistral/159 Qwen paper clusters). Cumulative minus recency; 10,000 paired paper-cluster bootstrap draws, seed 20260913, source-weighted means and marginal percentile 95% intervals. F1 intervals here use this exploratory replay, not the original primary analyzer. "Joint" is $\Delta\text{F1} > 0$ and a completeness decrease at the stated instant.

Model	Quantity	Estimate	95% interval
Mistral	Saved F1 change	0.0357	$[0.0106, 0.0607]$
	Pre-final completeness change	-0.2417	$[-0.2991, -0.1872]$
	Pre-final joint, 18/240	7.50%	$[4.31, 10.92]\%$
Qwen	Saved F1 change	0.1935	$[0.1530, 0.2360]$
	Pre-final completeness change	-0.2625	$[-0.3248, -0.2023]$
	Pre-final joint, 31/240	12.92%	$[8.75, 17.45]\%$

For Mistral, complete-source counts (recency/cumulative) change from 62/4 before the final eviction to 55/1 after it; for Qwen, from 65/2 to 59/1. The corresponding original post-final joint counts are 14 and 28. Thus the sign separation exists before the final eviction on both Qasper panels. This does not make all answer tokens time-aligned: later tokens read the final compacted KV, while the first token still uses saved earlier logits. Neither the pre-final nor post-final structural snapshot identifies the complete information used by the whole answer or its causal future utility. Panel provenance is retained in Appendix B.4; the artifact contains the replay specification, per-source records and analyzer.

B.6 Operational clarifications and endpoint choice

Exact cumulative signal. In the frozen post-selection runner, admission t contributes $\sum_h \sum_{q \in Q_t} \alpha_{q,p}^{(L,h)}$ to the score of each currently resident original position p, where Q_t contains the current chunk's queries and L is the last layer. The implementation converts that layer's attention tensor to float32 before summing heads and queries. It applies neither decay, length normalization nor value-norm weighting. The first admission contains up to 1,024 tokens and subsequent chunks up to 128. Consequently, accumulated mass depends on admission history; it is not a stationary estimate of token importance. One shared selected-position set is applied to every layer/head. Surviving positions retain their accumulated scores. Evicted coordinates may remain in the Python score dictionary, but are absent from the resident candidate list, receive no further attention updates and cannot be selected again. No cross-layer/head scorer ablation or multi-layer/value-aware sensitivity experiment was performed. See `postcomp_v1/core.py`, functions `prefill` and `select`.

Readout timing. Legacy native readout uses logits computed before the last compaction. The frozen primary post-selection arm instead removes and recomputes the final prompt token against the selected prefix, then appends its KV again; other retained states are unchanged. Thus the legacy first-token mismatch is not an unavoidable property of the primary arm. Neither contract re-encodes the document or erases information already mixed into retained states. The missing historical HotpotQA pre-final traces limit retrospective alignment analyses, not the recorded primary post-selection recomputation. No additional eviction occurs during primary answer decoding.

Why all units, and what it misses. The frozen A_s endpoint operationalizes the strict claim that *all* declared evidence units remain resident. It is a binary conjunction, not a continuous heavy-tailed measure or a token-necessity score. Any missing position can switch it off, including a minor miss in one long unit. Complete-unit fraction I_s exposes this loss of resolution: its recency / cumulative means are $0.583/0.130$, compared with $0.362/0.014$ for A_s. Reporting both preserves the frozen endpoint without making a post-outcome switch to a more favorable metric. The existing length-stratified analysis finds a fraction decrease in every third, but does not identify whether a particular missed position was necessary for the answer.

C Original-position accounting and validation

Each record joins an immutable source ID, annotation manifest, execution contract, eviction history, final original-position keep set and saved answer. Physical cache indices may change after compaction; only original prompt coordinates define residency. Overlapping annotation intervals are unioned for token coverage and remain separate for unit metrics. Future positions are not counted as losses, and generated positions are excluded from prompt residency. Annotations may be redundant or non-minimal and therefore do not become semantic-necessity labels.

Selectors receive only prompt/cache and model signals allowed by their arm; gold coordinates are used offline except in explicitly labelled oracle controls. Fail-closed replay checks unique sources and paired arms, prompt and manifest hashes, chronological admissions, valid and unique evictions, before/after cache counts, prompt/generated-position separation, interval identity, paired capacity, metric bounds and agreement between reconstructed keep sets and saved structural flags. A contradictory final keep set, missing arm, coordinate mismatch or changed input hash makes the dependent analysis unavailable. Tests cover overlaps, duplicate positions, sparse strata, final admission without eviction and last-eviction recovery. These checks validate the recorded structural state, not annotation adequacy or causal use. Exact schemas, commands, path manifests, validator logs and archived timing instrumentation are supplied in the artifact.

Table 9: Minimum record needed for each observable. Missing fields are not imputed from another panel.

Record	Required information	Supports
Source and contract	Stable IDs, prompt/tokenizer, budget, query visibility, admission and readout rule	Pairing and interpretation
Structural trace	Original-position evictions and final keep set	Coverage and exact residency
Annotation manifest	Original-coordinate unit intervals	Token union and unit metrics
Saved answer	Generated text, references and scorer	EM/F1 independent of residency

Table 10: Historical consistency grid and frozen primary cohort, cumulative minus recency at budget 1,024. Intervals are marginal paired 95% intervals. The four historical panels differ in provenance and are supporting evidence.

Task	Model	n	ΔF1	Δ completeness	Status
Qasper	Mistral	240	$0.036\ [0.012, 0.061]$	$-0.225\ [-0.283, -0.172]$	Exploratory discovery
Qasper	Qwen	240	$0.194\ [0.153, 0.237]$	$-0.242\ [-0.303, -0.184]$	Exploratory cross-model
HotpotQA	Qwen	762	$0.026\ [0.014, 0.037]$	$-0.366\ [-0.403, -0.329]$	Freeze chronology unverified
HotpotQA	Mistral	781	$0.241\ [0.209, 0.273]$	$-0.296\ [-0.332, -0.261]$	Post-pilot
HotpotQA	**Mistral**	**500**	$0.220\ [0.180, 0.261]$	$-0.348\ [-0.390, -0.306]$	**Frozen primary**

D Native source-level outcomes and robustness

D.1 Historical cohort estimates

D.2 Exploratory conditional analysis

Table 11 retains the damaged and intact effect estimates under each panel's original complete-envelope definition, alongside the pooled effects summarized in Section 3. Conditioning on policy-specific damage does not identify the effect of causing fragmentation.

Table 11: Native contextualized-KV panels, cumulative attention minus recency (1024-token budget). Δ completeness is the full-pool change in complete annotated-evidence completeness. Qasper panels are exploratory for this arm (the predeclared arm was token-max). Qwen/HotpotQA is a cross-task consistency panel with unverified freeze chronology; Mistral/HotpotQA is post-pilot supporting evidence. Intervals are paired 95% cluster-bootstrap intervals.

Panel	Status	n	Δ F1 pooled	Δ completeness	damaged Δ F1	intact Δ F1	damaged $-$ intact
Qasper / Mistral	expl.	240	$0.036\ [0.012, 0.061]$	-0.225	$0.061\ [0.031, 0.092]$	$-0.051\ [-0.102, -0.005]$	$0.112\ [0.052, 0.176]$
Qasper / Qwen	expl.	240	$0.194\ [0.153, 0.237]$	-0.242	$0.222\ [0.174, 0.273]$	$0.107\ [0.046, 0.176]$	$0.115\ [0.030, 0.197]$
HotpotQA / Qwen	unverified	762	$0.026\ [0.014, 0.037]$	-0.366	$0.049\ [0.036, 0.063]$	$-0.009\ [-0.028, 0.010]$	$0.058\ [0.035, 0.082]$
HotpotQA / Mistral	post-pilot	781	$0.241\ [0.209, 0.273]$	-0.296	$0.296\ [0.255, 0.336]$	$0.127\ [0.079, 0.174]$	$0.170\ [0.106, 0.232]$

D.3 Excluding engineering-pilot sources

This exploratory post-outcome sensitivity excludes all 16 source IDs in the saved Mistral/HotpotQA engineering pilot, without selecting on outcomes. The remaining 765 sources retain their original paired scores and completeness labels. With the existing source-paired bootstrap (10,000 draws, seed 20260913), cumulative attention minus recency gives F1 $0.239\ [0.207, 0.271]$, all-units completeness $-0.295\ [-0.331, -0.261]$, and the recency-damaged minus recency-intact F1-effect contrast $0.164\ [0.101, 0.227]$. Same-source F1 gain with completeness loss occurs on $78/765$ sources ($10.2\%\ [8.1\%, 12.4\%]$). Intervals are marginal percentile 95% intervals. Directions persist after exclusion; this does not establish a pre-pilot freeze or remove the possibility of pilot-informed design.

D.4 EVIDENCE-DEFINITION SENSITIVITY ON NATIVE TRACES

D.4.1 POST-FREEZE SENSITIVITY ON THE NEW 500-SOURCE COHORT

This analysis was specified after the frozen primary result was known. It recomputes the declared metrics directly from the frozen manifest's evidence units and the 1,000 saved recency/cumulative post-selection keep sets. It is exploratory and does not alter the frozen endpoint classification. All sources have nonempty units; checks find zero duplicate unit positions, duplicate or out-of-range retained positions, wrong-budget records, evidence-union mismatches or saved complete-flag mismatches.

For the all-units endpoint, 178 sources change from complete under recency to incomplete under cumulative attention, four change in the reverse direction, three are complete under both and 315 are incomplete under both. Among those 315, cumulative has higher evidence-token coverage on 255, equal coverage on three and lower coverage on 57. Table 2 reports the principal readouts; the artifact contains the complete metric ladder, unit-count strata, and validation checks.

D.4.2 POST-FREEZE PLACEMENT REFERENCE

The two policies differ in retention shape as well as in scoring: recency retains a contiguous suffix, cumulative attention retains scattered positions. A conjunctive completeness readout rewards contiguity by construction, so we asked how much of each contrast survives when the evidence locations are destroyed and only that shape is kept. Rules were fixed before any contrast was computed (`experiments/postcomp_v1/geometry_null_v1/SPEC.md`).

For each source and arm we hold the observed keep set fixed and re-place every annotated unit as a contiguous block of its own length, with start drawn uniformly over the document extent, independently per unit. This preserves unit count, unit lengths, the keep set and the estimated document length while changing evidence location, between-unit spacing, overlap and joint layout. Document extent is obtained by linear character-to-token interpolation from the span anchors; variants using the full prompt length and the largest observed evidence position are also reported.

Table 12: Position-blind placement reference on the frozen 500-source cohort. Reference/observed ratio is the reference gap divided by the observed gap. Gaps are cumulative minus recency.

Readout	Observed gap	Reference gap	Reference/observed
All-units complete	-0.348	-0.276	79.3%
Any-unit complete	-0.526	-0.651	123.8%
Complete-unit count	-1.108	-1.164	105.0%
Token-union coverage	0.086	-0.005	-5.8%

Figure 6A illustrates how equal retained capacity and coverage can coexist with different numbers of intact units in a constructed example; panel B summarizes the descriptive reference/observed ratios across readouts. Most of the all-units completeness contrast is reproduced after randomized placement under this specified extent, while the positive coverage contrast is not reproduced: under random placement the two policies have almost equal coverage (0.574 versus 0.579). Per arm, cumulative attention exceeds its own reference by 0.093 $[0.078, 0.108]$ in coverage and by 0.007 $[-0.002, 0.018]$ in all-units completeness, whereas recency exceeds its reference by 0.002 $[-0.029, 0.033]$ and 0.079 $[0.039, 0.119]$ respectively. Thus cumulative coverage exceeds this placement-reference expectation, recency coverage does not show a detectable departure and recency all-units completeness exceeds its reference expectation. The observed -0.348 gap, the -0.276 reference gap and the -0.072 $[-0.114, -0.030]$ residual are descriptive contrasts rather than a causal decomposition.

The full-prompt extent variant gives 82.8% and -7.6%. A third pre-specified variant bounded by the largest observed evidence position gives 41.0% and $+332\%$; its placement region is defined by the evidence locations themselves, so it partly preserves the quantity the reference removes and is the weakest of the three. `exact_audit.py` recomputes the three completeness nulls in closed form, without Monte Carlo and without shared code, and agrees with the simulation to within 0.0019 on every audited completeness readout.

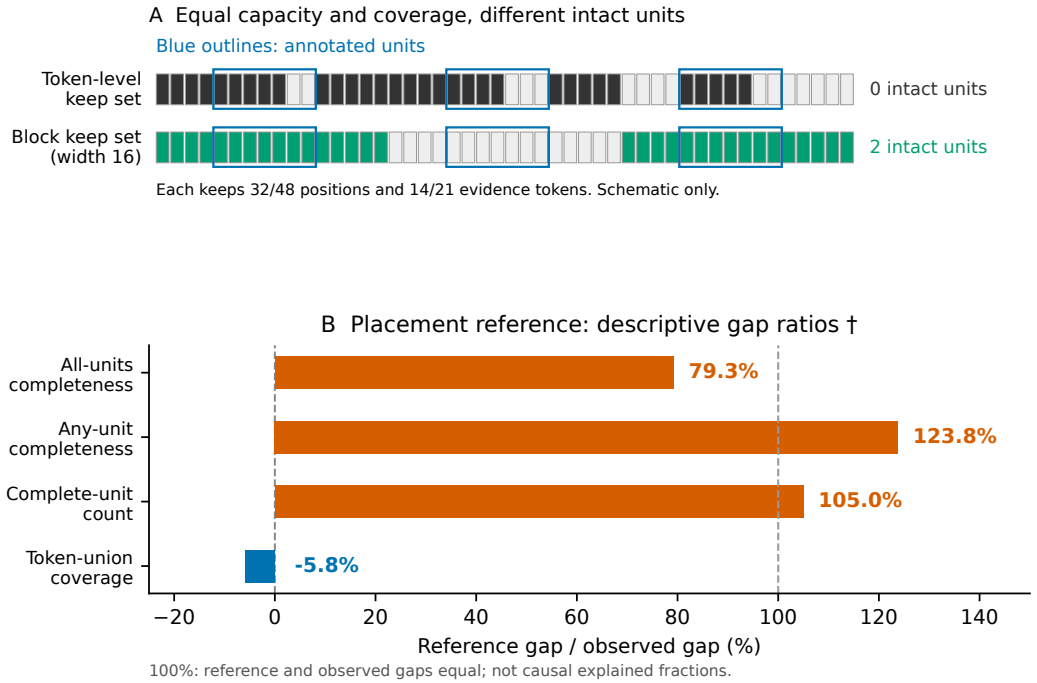

Figure 6: **Geometry illustration and full reference-ratio readout. A:** A constructed 48-position example: both keep sets retain 32 positions and 14 of 21 annotated positions, but complete zero versus two units. This is a schematic, not formal-cohort data or an achieved coverage-matching result. **B:** The descriptive reference/observed ratios from Table 12, not causal explained fractions. The dashed line denotes equal reference and observed gaps.

The continuous ratio and residual, rather than the internal 80% decision threshold, carry the interpretation. This post-outcome analysis changes no F1, coverage or completeness estimate and provides no causal attribution.

D.4.3 Post-freeze annotated-evidence headroom

We computed the union of every annotated evidence unit's original token positions for each frozen source. Evidence unions contain 32–398 tokens (median 92.5; mean 99.43), corresponding to a median 9.03% of the nominal 1,024-position budget. The selection contract also protects the first 16 and most recent 128 prompt positions. After unioning these mandatory positions with all annotated evidence, the required set has median size 230 and maximum 542. Hence all 500 sources can physically retain every annotated evidence position under the declared budget and protection rules, with at least 482 positions remaining in the worst case.

The saved-output check is post-primary. An independent implementation rebuilds unions from the frozen manifest; both audits verify coordinate alignment, protected positions, exact budgets and prompt/selector hashes. This establishes capacity headroom, not minimum sufficient support or F1 preservation by an oracle selector. Specifications and both outputs are in the artifact's `experiments/postcomp_v1/headroom_v1/` directory.

D.4.4 Post-freeze evidence-unit length sensitivity

Before inspecting arm-level completeness by length, we divided the 1,218 annotated units at manifest-only empirical tertiles: short (7–30 unique token positions), medium (31–45) and long (46–305). Within each eligible source and bin, we measure the fraction of units surviving completely, whether any unit survives completely and the number surviving. Point estimates weight eligible sources equally; intervals use 10,000 paired source-bootstrap draws.

Table 13: Exploratory unit-length sensitivity on saved post-selection records. All entries are cumulative attention minus recency with marginal 95% paired source-bootstrap intervals. Cutpoints are dataset-specific manifest tertiles.

Length (tokens)	Sources / units	Complete-unit fraction	Any complete unit	Complete-unit count
Short (7–30)	293 / 417	$-0.455\ [-0.526, -0.383]$	$-0.457\ [-0.536, -0.379]$	$-0.611\ [-0.713, -0.509]$
Medium (31–45)	304 / 404	$-0.450\ [-0.511, -0.389]$	$-0.470\ [-0.536, -0.405]$	$-0.605\ [-0.691, -0.523]$
Long (46–305)	294 / 397	$-0.465\ [-0.526, -0.404]$	$-0.507\ [-0.575, -0.439]$	$-0.650\ [-0.738, -0.565]$

The complete-unit fraction declines in every length third, including the shortest. Thus the contrast is not restricted to the longest annotated spans. This analysis was specified after the primary result was known; it does not control semantic difficulty or tokenizer choice, and every unit still uses an exact all-token completion rule. The producer, independent point-estimate audit and specification are included with the artifact.

D.4.5 Post-primary budget sensitivity

After the 1024-budget primary result was known, we specified an exploratory sensitivity on the same 500 sources at capacities 512 and 2048. Both policies use the post-selection readout contract. The primary denominator at each budget remains all 500 sources: all prompts overflow 512, while 117 overflow 2048 and 383 fit without compaction. Thus the 2048 all-source estimand includes many policy ties by design. The capacities bound final prompt residency, not peak memory or latency; the 512 arm still admits the initial 1024-token chunk before compaction.

Table 14: Exploratory budget sensitivity on the frozen 500-source cohort. Cumulative attention minus recency; brackets are marginal paired 95% source-bootstrap intervals. Coverage is over the annotated-token union.

Budget	Overflow	Readout	Recency	Cumulative	Change [95% CI]
512	500	F1	0.192	0.409	$0.217\ [0.177, 0.259]$
		Evidence-token coverage	0.291	0.331	$0.040\ [0.005, 0.074]$
		Any-unit completeness	0.478	0.134	$-0.344\ [-0.390, -0.300]$
		All-units completeness	0.074	0.000	$-0.074\ [-0.098, -0.052]$
2048	117	F1	0.551	0.582	$0.030\ [0.017, 0.046]$
		Evidence-token coverage	0.965	0.984	$0.019\ [0.007, 0.032]$
		Any-unit completeness	0.992	0.950	$-0.042\ [-0.062, -0.022]$
		All-units completeness	0.928	0.830	$-0.098\ [-0.130, -0.068]$

On the prespecified 117-source common-overflow subgroup, descriptive cumulative-minus-recency changes at 512/2048 are $0.213/0.129$ for F1, $0.072/0.081$ for evidence-token coverage, $-0.248/-0.179$ for any-unit completeness, and $-0.009/-0.419$ for all-units completeness. This subgroup describes a different population and does not replace the all-source analysis. Main and independent analyzers reproduce all reported values; complete transition and interaction tables remain in the artifact.

E Frozen MuSiQue reuse test

Design. This experiment is separate from the stopped historical follow-up branch. We formed Q0/Q1 pairs from official MuSiQue single-hop questions over distinct supporting paragraphs. Two automated, outcome-blind reviews and role-agnostic overlap grouping produced 221 sources in 85 groups. The frozen cohort contains 115 distinct Q1 strings and 221 contexts; 113 contexts had appeared in an earlier project experiment on a different composed question, but no present-arm outcomes had been observed.

Mistral-7B uses a 1,024-position Q0 prefill budget and a 928-position document cache reserving 96 positions for Q1. The cache branches before Q0 decoding; generated Q0 tokens never enter Q1 history, and Q1 never re-prefills the document. Restore and sham start from the same cumulative parent and perform equal-budget swaps. Restore inserts every missing position of the annotated Q1 support from an immutable full-prefill KV snapshot; sham inserts absent non-evidence states matched by original-coordinate distance. Every support paragraph contains a normalized reference-

answer string, making restoration an annotation-aware upper bound. All plan, cache, suffix, position and branch checks pass. Intervals use 10,000 paired draws over the 85 overlap groups.

Table 15: Frozen future-query results. The predicted policy reversal does not occur; full-support oracle restoration improves on its matched sham.

Status	Contrast	Estimate [95% CI]
Co-primary	Q1 cumulative − recency F1	0.179 $[0.114, 0.244]$
Co-primary	Full-support restore − matched sham F1	0.260 $[0.197, 0.327]$
Secondary	Q0 cumulative − recency F1	0.334 $[0.261, 0.407]$
Secondary	Q1-minus-Q0 policy interaction	-0.155 $[-0.247, -0.063]$

Absolute Q1 F1 is 0.240 for recency and 0.419 for cumulative despite complete-paragraph residency of 0.303 and 0.000. Restore scores 0.695 against 0.435 for matched sham, with 108 wins, 95 ties and 18 losses. Thus higher exact completeness does not select the better average Q1 policy, while some answer-bearing states omitted by cumulative remain useful when inserted.

Post-primary robustness. Four additional equal-budget shams start from the same parent, remove the same donor positions and exclude annotated evidence. Restore-minus-sham F1 is 0.283 $[0.220, 0.348]$ for a same-region sham and 0.263–0.273 across three fixed distance-band shams; all intervals exclude zero, while every sham-minus-parent interval includes zero. Full EM, win/tie/loss and intervention-size tables remain in the artifact. An EM normalization mismatch was corrected from saved generations after inference; no F1, arm, source or generation changed.

E.1 Correction and answer-excluded boundary

An earlier exploratory answer-span localization omitted title offsets. All derived answer/non-answer claims from that analysis are withdrawn. The frozen full-paragraph intervention and its multi-sham comparisons do not use this mapping and are unaffected.

A corrected post-primary V2 excludes every token overlapping a normalized gold-answer occurrence in the full titled paragraph. Restore minus nearest sham is 0.022 $[-0.018, 0.060]$; across five shams, only the same-region contrast excludes zero. The prespecified robustness rule fails. V2 therefore does not establish an independent benefit from answer-excluded support. The artifact retains the offset audit, all five comparisons and unsigned review sheets; no human semantic review is claimed.

Interpretation. The experiment establishes local functional relevance for an answer-bearing oracle replacement. It rejects a future-query policy reversal and leaves necessity, mediation and deployable restoration unresolved.

E.2 MuSiQue contrasts and sham robustness

The frozen contrasts and corrected answer-excluded result are reported in this appendix. Table 16 retains all five sham comparisons and their post-primary status.

Table 16: **Post-primary multi-sham robustness on the frozen cohort.** Intervals are marginal 95% intervals over 85 overlap groups. Every sham starts from the same parent, matches the restore swap count and budget, and excludes annotated evidence. The nearest-sham outcome preceded this robustness design; the other four sham rules were fixed before their outcomes.

Sham family	Restore − sham F1	Sham − base F1
Nearest position	0.260 $[0.197, 0.324]$	0.017 $[-0.013, 0.049]$
Same region	0.283 $[0.220, 0.348]$	-0.007 $[-0.035, 0.022]$
Distance band, seed 1	0.273 $[0.213, 0.335]$	0.004 $[-0.023, 0.031]$
Distance band, seed 2	0.268 $[0.206, 0.334]$	0.008 $[-0.023, 0.041]$
Distance band, seed 3	0.263 $[0.201, 0.328]$	0.013 $[-0.016, 0.046]$

F CLAIM-RELEVANT CORRECTIONS AND RESOLUTION

Table 17: Corrections and stopped extensions that change how a reported claim must be interpreted. Full development histories remain in the artifact.

Item	Finding	Consequence for claim
Qwen extension	Full-cache first-token argmax agreed on 16/16, but complete sequences changed on 7/16	Frozen gate not relaxed; no confirmatory second-model panel
ChunkKV comparison	Audited interface did not expose required original-coordinate traces	No native-method estimate; missing coverage, not algorithmic disqualification
Structured factorial	Coverage equivalence and F1 non-inferiority not established	Residency gain is supported; utility-preserving improvement is not
Answer-excluded restore	Title-offset error corrected; V2 benefit inconclusive	Earlier localization claims withdrawn; full-support result unaffected
Official support coordinates	Two identical-text spans mapped to wrong occurrences	Frozen records retained; corrected sensitivity in Appendix H.1

Qwen diagnostics distinguish admission-chunk from singleton readout evaluation, not batching across requests. Per-source logits, shape checks, retired serving experiments and earlier controlled branches are preserved in the development archive indexed in Appendix G.

The main-text stress matrix and claim-specific reporting rules summarize the current evidence status. This appendix retains only corrections and stopped branches that change how those claims must be interpreted.

G PORTABILITY, OVERHEAD AND REPRODUCIBILITY

G.1 AUDIT INTERFACE AND REPLAY BOUNDARY

Adapters emit source IDs, original positions, budgets, rounds and execution contracts; immutable manifests supply evidence intervals and grouping, while answer scores are joined separately. Hugging Face rolling-cache and block-diffusion harnesses implement this interface; new adapters must validate position mapping. No vLLM or paged-attention backend has passed an end-to-end FACT-KV equivalence check. For paged caches, an adapter would need to trace logical sequence positions through physical page allocation, reuse and compaction, including shared-prefix ownership; physical addresses alone are not stable evidence identities. This is a proposed requirement, not a validated serving implementation. CPU replay checks saved traces and scores; it does not regenerate inference or establish serving performance. No end-to-end latency, throughput, or memory advantage is claimed.

Artifact replay. The reviewer quickstart gives one command for saved-record replay of the primary table, within-source count, geometry reference, structured panel and MuSiQue results. It verifies packaged hashes and runs analyses in a separate working directory, preserving canonical evidence. Exact commands, paths and the inherited historical runner are documented in the artifact README; revision-stage exploratory analyses retain separate input hashes. Python dependencies are pinned in the release. CPU replay requires no GPU or network; floating-point comparisons allow absolute rounding differences up to 10^{-12} while counts and classifications must match. The primary bundle pins model and library revisions; historical panels support saved-score replay only. Exact paths, hashes, artifact categories, and development records appear once in the README.

G.2 A RELEASED-METHOD PORT AUDIT CONTRACT

The KVCache-Factory SnapKV/PyramidKV port implementation audit in Appendix J.2 follows six requirements: unchanged native selection, unchanged native readout, original-coordinate traces, layer/head identity, actual resident KV bytes, and traced/untraced output agreement. For heterogeneous keep sets $K_{s,\ell,h}$, declare whether a metric averages headwise results, unions positions present in any head, or intersects positions present in every head. Partial-layer traces support only partial-layer statements. Merged or synthesized states require separate provenance and fidelity endpoints;

cluster membership is not an original-slot keep set. This contract does not impose the primary harness's readout intervention, but native ChunkKV and KVzip comparisons remain unperformed.

Examples of claim-specific reporting. CAOTE combines attention and value information in an output-error eviction criterion (Goel et al., 2025); LaProx uses an output-aware formulation with projected values and inter-head interactions (Mai & Kim, 2026). DistillCache trains a policy with an output-distribution KL reward (Althoubi, 2026). Their fidelity objectives should be evaluated as such; they do not by themselves establish intact annotated units. Likewise KVzip's context-reconstruction scoring and K-VEC's cross-head/layer token coverage (Kim et al., 2025; Roy et al., 2026) do not substitute for an annotated-unit trace. These are conceptual mappings of measurement claims, not empirical FACT-KV audits or assertions that these methods fragment evidence.

H Exploratory post-outcome checks and evaluation limits

These saved-record analyses were specified during manuscript revision, after the primary outcomes were known. They are exploratory, do not alter frozen classifications and do not establish causality or cross-model generality. The supplement contains the specification, all source-level rows, input hashes and a CPU analyzer.

Budget and length sensitivity. On the same 500 primary sources, exploratory budgets 512/2,048 give F1 contrasts $+0.217/+0.030$ and all-units contrasts $-0.074/-0.098$. At 2,048, 383 prompts fit without compaction, so these averages include ties. Complete-unit fraction also declines in each mean-unit-length third, including 7–30-token units (-0.455). These are within-cohort sensitivities.

Scorer and benchmark-score boundaries. No matched-contract layer-averaged, value-weighted, decayed or admission-normalized cumulative scorer ablation has been run. Native-method comparisons change more than the scorer and cannot substitute for it. A post-outcome robustness analysis re-scores the unchanged frozen predictions with the pinned official HotpotQA evaluator. It preserves the behavioral contrast: cumulative minus recency is $+0.217$ $[0.175, 0.258]$ for F1 and $+0.198$ $[0.156, 0.242]$ for EM. This does not replace the frozen endpoint or turn the experiment into an official leaderboard submission because the generation contract remains study-specific. Keep sets and residency readouts are unchanged. The 70/500 within-source event retains its frozen-scorer definition and is not relabelled as an official-score joint count.

Source-level alignment. Post-outcome source-paired correlations are retained in the artifact rather than promoted as additional paper evidence. Positive within-policy association can coexist with opposite policy means; non-interchangeability does not imply orthogonality, and these correlations do not identify causal effects.

Joint-event subgroup boundary. The 70/500 joint event is not a subgroup analysis. Length and location sensitivities for residency, historical position strata and the partial correlations above do not establish whether this event concentrates by unit length, unit count or distance from the query. We do not report event-specific stratified rates for these factors.

Evidence grouping. We merge adjacent annotated units with at most 0, 8 or 32 intervening tokens, using the union of their positions without adding gap tokens. The 8/32 rules regroup 93/142 sources; the 0-gap rule changes none. Complete-unit fraction contrasts become $-0.4545/-0.4583$, versus -0.4534 originally. Merging all units within each source gives -0.3480. Union coverage, F1 and all-units completeness are invariant to these regroupings by construction; this is not independent evidence of semantic robustness. Alternative human annotations have not been evaluated. The official support-sentence coordinate check and paragraph expansion below use existing dataset annotations; they do not supply new human judgments. Shortest unique answer-bearing substrings have not been tested either: answer-conditioned trimming changes the annotated object and can remove relational support. It would require a separately specified audit, not relabeling current exact residency as semantic sufficiency.

Block widths and serving cost. The original calibration evaluated widths 8/16/32 on 32 sources, with coverage contrasts $0.02005/0.00716/0.03472$. Exploratory recomputation from the same

frozen keep sets gives complete-unit fraction contrasts $0.1536/0.2943/0.3672$. The original calibration did not decode answers; only width 16 received the original formal 300-source evaluation. The post-outcome extension in Appendix J.1 evaluates F1 at widths 4/8/16 on those same 300 sources; none establishes F1 non-inferiority. No adaptive-width or semantic-boundary-aligned selector result exists. No prefill/decode hybrid was tested; primary decoding performs no further eviction. Displacement of useful scattered tokens remains a hypothesis, not an identified mechanism or an estimated residency–utility frontier. Recorded per-source timers combine multiple arms and do not isolate selector, prefill or decode cost. We therefore report no arm-wise latency/FLOP advantage. The post-outcome KVCache-Factory SnapKV/PyramidKV port panel in Appendix J.2 supplies measured coverage under its own contract; native ChunkKV remains unmeasured. Synthetic adapter tests alone do not establish native model behavior.

Distance-stratified reference. A new exact descriptive reference keeps each unit's original start in the same tenth of full prompt coordinates and uniformly averages admissible contiguous starts within that bin, clipped to fit the original unit length. Observed keep sets remain fixed; unit placements are independent. Expected all-units completeness is 0.3597 for recency and 0.0087 for cumulative, a contrast of -0.3510. Coarse distance to the trailing query/readout is thus preserved, unlike the position-blind reference. The two references answer different conditional questions; neither identifies geometry's causal share. Both preserve each original unit length; neither preserves inter-unit spacing or the full overlap pattern. No jointly translated span configuration or selection-granularity-matched placement reference has been evaluated. A randomized-window compressor's utility is not tested. Randomly placed units may overlap, so marginal unit coverage is not relabeled as union coverage.

MuSiQue overlap boundary. No policy-wise cross-tabulation of Q1-support positions against Q0 retained and omitted regions was run. Small post-outcome overlap-group summaries and support position correlations remain in the artifact; they do not establish an interaction or explain the policy reversal.

Reuse integration and qualitative interpretation. A stage-wise adapter would need native original-position traces at each specified compression/reuse checkpoint, query-specific evidence annotations and utility readouts, with re-prefill, appends and state overwrites declared. Layer/head aggregation must also be explicit. MuSiQue is a restricted reuse test, not an implemented SCBench integration. For merged or reconstructed states, provenance does not certify exact annotated-unit residency; representation fidelity requires a separate endpoint. The corrected answer-excluded result does not identify why restoration lacks a stable benefit. Surviving anchors, redundancy and contextualized representations remain hypotheses; no qualitative case analysis here establishes one of them. Such an analysis would need declared case selection, source coordinates, inserted/excluded positions and paired generated answers, not an inferred reasoning trace.

H.1 Threshold and unit-boundary sensitivity

This post-outcome CPU analysis uses the same 500 paired keep sets and changes no generated answers or frozen decisions. Before computing these summaries, we recorded thresholds $\tau \in \{0.80, 0.90, 0.95, 0.99, 1.00\}$ and the source-average unit-survival endpoint

$$I_s(\tau) = m_s^{-1} \sum_i \mathbf{1}\left[|F_{si} \cap K_s|/|F_{si}| \geq \tau\right].$$

Integer count comparisons avoid floating-point threshold ambiguity. Intervals use 10,000 paired source resamples (seed 2026092101); they are marginal, not simultaneous bands. Figure 7 and Table 18 report every specified threshold.

Official annotation check and correction. The frozen units were obtained from HotpotQA supporting-fact sentence text, not whole paragraphs. Reconstructing the canonical title/sentence-ID mapping and the exact pinned-tokenizer prompt passed context and prompt-ID checks on all 500 sources, but exposed two incorrect text-occurrence coordinates among 1,218 units. The initial alignment check stopped; a documented amendment, made before aggregate sensitivity computation, added corrected official-sentence readouts. One frozen span selected a repeated title; another

selected an identical sentence in a different document. Text equality and token-coordinate checks alone had not detected incorrect annotation provenance.

All original records remain intact and all 500 sources enter both versions. Official-coordinate correction changes the coverage contrast from 0.086297 to 0.086040 and exact complete-unit fraction from -0.453367 to -0.454033. All-units contrast (-0.348), F1 contrast (0.220420) and the 70/500 joint count are unchanged. These are post-outcome corrections, not a new frozen confirmation. The artifact records both source IDs, original/corrected spans, canonical selected data, tokenizer, input hashes and per-source outputs.

Boundary definitions. The sentence analyses use 1,218 units. Paragraph expansion uses 1,000 units: for each official supporting document, take the union of all supplied sentence bodies, excluding repeated title prefixes and separator-only tokens. It adds non-support sentences and changes the evaluated evidence set; it is not a pure regrouping of annotated support. The all-support conjunction requires all official supporting-sentence positions. The exact contrasts are -0.454 for official sentences, -0.499 for expanded paragraphs and -0.348 for the all-support conjunction. None validates annotation minimality, semantic sufficiency or a preferred threshold for utility.

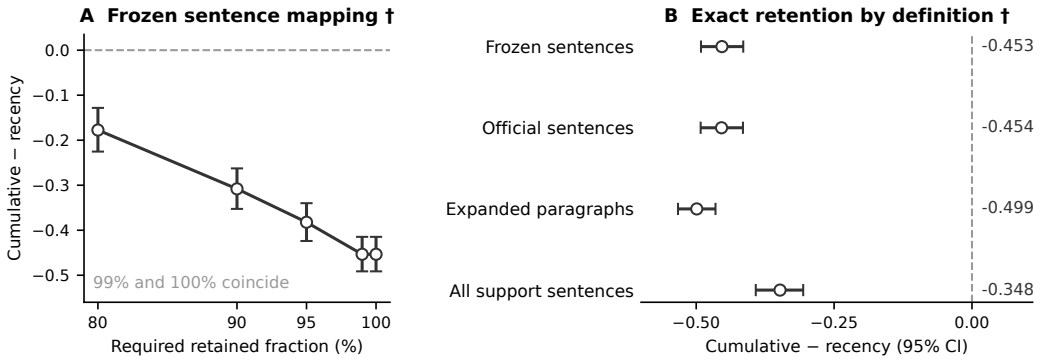

Figure 7: **Post-outcome threshold and boundary sensitivity.** A: Source-average unit survival under the original frozen sentence mapping; error bars are marginal paired 95% intervals. B: Exact-retention contrasts under each declared definition; all-support is a source-level conjunction, whereas the other rows average units within source. All rows use 500 sources. Hollow markers (†) denote exploratory analyses. Negative contrasts persist below 100%; this does not establish semantic validity.

Table 18: Cumulative-minus-recency unit-survival differences on all 500 sources. Official sentences correct two text-occurrence mappings; expanded paragraphs add non-support text. Full arm means and marginal intervals for every cell are in the artifact. All results are post-outcome exploratory.

Threshold	Frozen sentences [95% CI]	Official sentences	Expanded paragraphs
80%	-0.177 $[-0.225, -0.128]$	-0.179	-0.264
90%	-0.308 $[-0.353, -0.263]$	-0.309	-0.387
95%	-0.382 $[-0.424, -0.340]$	-0.384	-0.443
99%	-0.453 $[-0.491, -0.415]$	-0.454	-0.496
100%	-0.453 $[-0.491, -0.415]$	-0.454	-0.499

The negative difference is present at all five thresholds. For the frozen sentence mapping, 99% and 100% produce identical indicators in these saved records; their equality is not a general property of the thresholds. The trend is consistent with geometry-sensitive retention, not evidence that missing tokens are dispensable or that high exact residency improves answers.

Length weighting and omission location. A further post-outcome check weights each unit's threshold indicator by its length within source, then averages sources equally: $I_s^{(w)}(\tau) = \sum_i |F_{si}|\mathbf{1}[|F_{si} \cap K_s|/|F_{si}| \geq \tau] / \sum_i |F_{si}|$. Overlapping unit memberships may count more than once; this is not union coverage. On corrected official sentences, cumulative-minus-recency contrasts at 80/90/95/99/100% are $-0.168/-0.301/-0.376/-0.449/-0.449$. Marginal paired 95% intervals at 95% and 100% are $[-0.417, -0.333]$ and $[-0.487, -0.411]$, from 10,000 source resam-

ples (seed 2026092201). The artifact reports all thresholds and both coordinate mappings; changing weighting does not reverse these observed contrasts.

Of 1,218 official sentence units, cumulative leaves 1,063 incomplete: 734 omit interior positions only, 327 omit both boundary and interior positions, and two omit only a boundary position. Boundaries are the first and last token positions of each annotated unit. Recency leaves 508 units incomplete (0/507/1 in those categories). These pooled unit counts differ from source-average rates. The saved-record audit also reports every fixed length bin (1–16, 17–32, 33–64, 65+) and prompt-position quarter with its denominators. These geometric categories do not identify content/function words or cross-sentence connectors, and do not show which omissions caused an answer error. No linguistic necessity labels were inferred from token IDs.

I Additional comparisons and execution details

I.1 Related-work comparison

Table 19: Reported evaluation scope of the closest precedents. Entries describe the evidence used for the compared claims, rather than every capability of each system. "Behavioral" denotes fixed-trace output replay without a physical original-position residency trace; "method-specific" does not assert such a trace. "Three-readout ranking" means comparing policy rankings jointly across answer utility, annotated-token coverage and exact unit residency. A dash in that column means this specific comparison is not reported, not that paired policy evaluation is absent; KVDiagnosis includes paired evaluations and matched comparisons.

Work	Residency trace	Answer metric	Exact-unit residency	Three-readout ranking	Future-query reuse
KVDiagnosis	Original-position	✓	–	–	–
Does Accuracy Equal Evidence?	Behavioral	✓	–	–	–
SCBench / KVzip	–	✓	–	–	✓
RestoreKV / TwinKV	Method-specific	✓	–	–	–
ChunkKV / ShotKV / SABlock	Method-specific	✓	–	–	–
FACT-KV	Original-position	✓	✓	✓	✓

I.2 Structured-selector protocol and full outcomes

The structured comparison froze 16 engineering, 32 calibration and 300 formal Mistral/HotpotQA sources, document-disjoint from historical outcomes. Both arms used the same cumulative-attention scorer definition, original coordinates, a 1,024-position budget and post-selection readout. Token selection ranked individual positions; structured selection retained complete width-16 blocks. Gold annotations entered calibration and evaluation only, never either selector. Full-cache sequence equivalence passed on all 16 engineering sources.

Calibration selected width 16 from $\{8, 16, 32\}$ by paired coverage difference ($+0.02005$, $+0.00716$, $+0.03472$). Table 20 reports all formal readouts. Coverage's 90% interval $[-0.02026, 0.00721]$ misses the prespecified $[-0.02, 0.02]$ equivalence interval; F1's 95% lower bound -0.0483 misses the -0.03 non-inferiority margin. The joint success rule fails. This establishes a residency gain without establishing coverage equivalence or utility preservation; the machine labels do not establish material coverage mismatch or F1 inferiority. Different retained states may change later scores, so geometry is not isolated as a causal mediator. Equal resident capacity does not establish equal runtime or FLOPs.

Table 20: Frozen structured-selector factorial on 300 new Mistral/HotpotQA sources. Differences are structured minus token; the coverage interval is 90% and all other intervals are 95%.

Readout	Token	Structured	Difference [CI]
F1	0.5362	0.5255	-0.0107 $[-0.0483, 0.0264]$
Token-union coverage	0.6587	0.6521	-0.0065 $[-0.0203, 0.0072]$
Complete-unit fraction	0.1682	0.4296	$+0.2614$ $[0.2268, 0.2957]$
Any-unit completeness	0.3200	0.6833	$+0.3633$ $[0.3067, 0.4167]$
All-units completeness	0.0367	0.1767	$+0.1400$ $[0.1033, 0.1800]$

No native ChunkKV estimate is available because the audited interface did not expose the required coordinate trace. Valid instrumentation must preserve native selection/readout, verify output agreement and report actual layer/head memory. Our custom selector supports no ranking against released compressors.

J POST-OUTCOME SCOPE AUDITS

These reused-source audits test granularity, heterogeneous-cache aggregation and model scope. They were frozen before their own execution but after the primary outcomes and are not additional prospective confirmations.

J.1 STRUCTURED-RETENTION WIDTHS

All 300 structured-panel sources were rerun at widths 4, 8 and 16 against a corresponding token-level arm with the same nominal capacity. Independent replay checked membership, coordinates, protected positions, capacities, KV bytes, structural metrics and F1 for all 2,100 records. The ap-

Width	Δ complete-unit fraction	Δ F1	Δ coverage
4	$+0.1356\ [0.1086, 0.1638]$	$-0.0245\ [-0.0575, 0.0082]$	$+0.0014\ [-0.0050, 0.0078]$
8	$+0.2056\ [0.1750, 0.2377]$	$-0.0048\ [-0.0431, 0.0339]$	$+0.0046\ [-0.0051, 0.0139]$
16	$+0.2614\ [0.2266, 0.2973]$	$-0.0107\ [-0.0481, 0.0266]$	$-0.0068\ [-0.0202, 0.0068]$

Table 21: Block minus corresponding token-level control. Complete-unit and F1 intervals are 95%; coverage intervals are 90% (10,000 paired draws, seed 2026092202). All widths improve residency; none establishes F1 non-inferiority at -0.03. Marginal coverage equivalence at ± 0.02 holds for widths 4 and 8, but no width passes the joint rule.

proximate Bonferroni analysis retains the residency result at all widths and coverage equivalence only at width 4. The official-coordinate correction changes width-16 coverage from -0.006549 to -0.006753 (90% interval $[-0.020450, 0.006960]$); F1, residency and all decisions are unchanged.

J.2 NATIVE-METHOD AGGREGATION AUDIT

The native SnapKV and PyramidKV KVCache-Factory ports reuse all 500 primary source IDs under their own prefill/eager contract (Mistral-7B, BF16, transformers 4.44.2; observation window 32, pooling 5, nominal capacity 1,024, no merging). A 16-source gate verified traced/untraced outputs and KV equality; replay checked original coordinates, readouts and bytes. Mean-head averages metrics across layer/head sets; union uses positions present anywhere; intersection requires positions present everywhere. The audited layout stores 32 query heads per layer rather than the stock

Method	F1	Mean-head C	Mean-head I	Mean-head A	Union A	Intersect. A
Stock full	0.5768	1.0000	1.0000	1.0000	1.0000	1.0000
SnapKV	0.3493	0.6948	0.2927	0.0976	1.0000	0.0000
PyramidKV	0.3486	0.6896	0.3021	0.1058	1.0000	0.0000

Table 22: Native-contract means. Union coverage and unit fraction are also 1.0000; intersection coverage is 0.0039 and unit fraction is 0.0000 for both compressed methods. Mean-head, union and intersection answer different physical questions.

reference's 8 KV heads: logical prefill KV is 512.0/513.8 MiB for SnapKV/PyramidKV versus 225.1 MiB for stock, so this is not matched-memory ranking. Every unit has a complete copy in the union despite lower mean-head residency and zero intersection residency. Per-head fragmentation therefore does not imply global absence. Native ChunkKV remains unmeasured.

J.3 EXPLORATORY QWEN2.5-72B SCOPE AUDIT

This stratified 150-source panel uses the declared post-selection contract at budgets 1,024 and 2,048. The unchanged gate passed on 16 disjoint engineering sources under a documented FP32/no-TF32 amendment after BF16 sequence equality failed. At 1,024, full/recency/cumulative F1 is

0.1664/0.1257/0.1133; recency versus cumulative means are C 0.7013/0.7375, I 0.6957/0.1621 and A 0.5267/0.0267. At budget 2,048, 133/150 prompts fit without pruning and the F1 difference is

Readout	Cumulative − recency	95% interval
F1	-0.0124	$[-0.0501, 0.0231]$
C	$+0.0362$	$[-0.0291, 0.0999]$
I	-0.5336	$[-0.6014, -0.4667]$
A	-0.5000	$[-0.5800, -0.4200]$

Table 23: Budget-1,024 paired contrasts (10,000 stratified source-bootstrap draws, seed 2026092202). The panel establishes integrity divergence, not an F1 advantage or a controlled architecture effect.

$+0.0105$ $[0.0011, 0.0242]$. Full records, amendments and subgroup outputs are in the artifact.

K SAVED-TRACE UNIT DIAGNOSTICS

This post-outcome diagnostic uses the saved keep sets of the 500-source Mistral primary cohort and corrected official sentence coordinates; no new inference is performed. The 1,218 annotated units are grouped into fixed length bins (1–10, 11–20, 21–40 and more than 40 tokens) and ten equal-width position bins. Position is the unit's token-cell center divided by the whole prompt length. Within each bin, we average unit outcomes within source and then weight contributing sources equally. Thus the coverage readout here is mean unit coverage, not union token coverage C_s.

Intervals use 10,000 paired source-cluster bootstrap resamples (seed 2026092207); units are not treated as independent observations. All intervals are marginal and exploratory. For consistency with the main text, contrasts below are *cumulative minus recency*; they mechanically reverse the archived diagnostic's recency-minus-cumulative signs. Figure 8 shows the two stratified profiles.

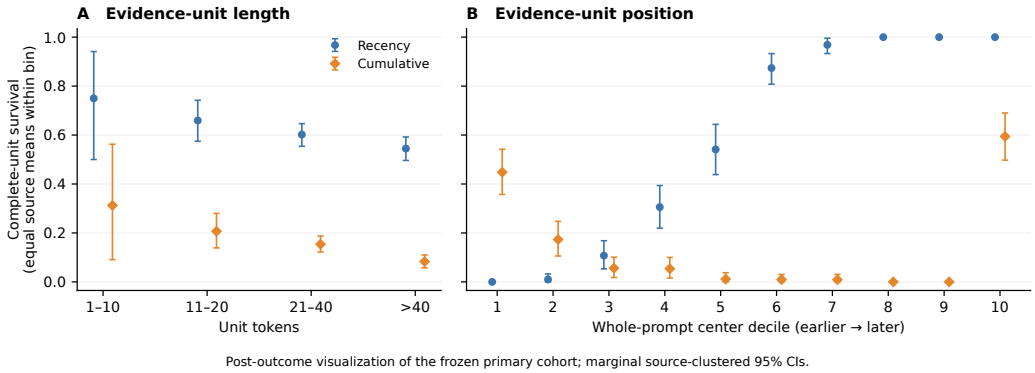

Figure 8: Post-outcome diagnostic on the frozen primary cohort. Source-balanced complete-unit survival by unit length (left) and normalized position (right), with source-clustered marginal 95% intervals. Position profiles are descriptive; they do not identify a causal explanation for the utility or residency contrasts. Bin counts remain in the artifact.

Cumulative-minus-recency complete-unit survival differs by -0.4375 $[-0.7692, -0.0667]$ for 1–10 tokens, -0.4526 $[-0.5714, -0.3276]$ for 11–20, -0.4477 $[-0.5053, -0.3895]$ for 21–40, and -0.4617 $[-0.5147, -0.4070]$ for longer units. The shortest bin contains only 16 sources and 16 units. The contrast is therefore observed outside the longest-unit bin, but these stratified comparisons do not isolate length from position or other source properties.

Position profiles differ in direction: in the first decile, survival is 0.0000 for recency and 0.4481 for cumulative, whereas in the eighth and ninth deciles it is 1.0000 and 0.0000, respectively. Recency consequently does not dominate every position bin. There are no single-unit sources. The two-unit group has 335 sources (670 units), with all-units completeness $0.3761/0.0209$ for recency/cumulative; the group with at least three units has 165 sources (548 units), with $0.3333/0.0000$. These summaries cannot determine how much of the overall contrast is caused by length, position or unit count.

Protection-overlap diagnostic. We also stratified the same 1,218 units by whether they intersect the first 16 protected positions, the most recent 128 protected positions, or neither, using that precedence to form exclusive groups. No unit intersects the prefix region; 98 units from 86 sources intersect the recent region, and the remaining 1,120 units occur in all 500 sources. Table 24 reports source-balanced unit-level outcomes. Recency's value of one in the recent stratum follows directly from its protection contract and is not an empirical mechanism claim.

Table 24: Post-outcome protection-overlap diagnostic. Contrasts are cumulative minus recency; intervals are marginal 95% source-cluster bootstrap intervals. Unit coverage here is the mean fraction retained within each unit.

Stratum	Sources	Units	Unit coverage Δ	Exact survival Δ
Recent-intersect	86	98	-0.1290 $[-0.1808, -0.0813]$	-0.3459 $[-0.4455, -0.2473]$
Neither	500	1,120	$+0.0811$ $[0.0370, 0.1259]$	-0.4703 $[-0.5116, -0.4276]$

Outside the recent protected region, cumulative retains a larger fraction of tokens within an average unit while recency retains more units exactly. This separation is consistent with the paper's distinction between token quantity and literal unit residency, but the strata remain descriptive and do not identify why either policy answers better. The artifact retains every bin's source and unit counts, both policies' unit coverage and survival, paired intervals, input hashes, and an independent CPU replay. Original-coordinate and corrected-coordinate headline checks are reported separately; the frozen primary estimates are not overwritten.

L LLAMA REVISION-STAGE SCOPE AUDIT

This revision-stage scope audit was designed after the Mistral outcomes were known. Its V2 execution contract was frozen at 2026-09-22 18:16:20 UTC, before any V2 model execution or outcome inspection. The artifact records the freeze manifest and protocol SHA-256 `3b00b3db00440050e04eee3bfbdfaca9a267f2cb0f4b35cf8413117afa0d2ee6`. Llama-3.1-8B-Instruct and its tokenizer were pinned to revision `0e9e39f249a16976918f6564b8830bc894c89659`. This is not another original prospective confirmation.

V1 stop and V2 boundary. V1 stopped because its frozen numerical-equivalence threshold failed despite exact generated-sequence and coordinate agreement; it produced no formal record. V2 was the single declared revision, and there is no V3. Before using fresh cohorts, V2 replaced that threshold with exact greedy-trajectory equality and retained logit differences as mandatory non-gating diagnostics. Model, prompt, tokenizer, precision, attention implementation, policies, budgets, endpoints and decision rules were unchanged. V1 remains archived as failed.

Fresh cohorts and frozen contract. Input-only eligibility yielded 1,445 candidates after excluding every enumerated prior source. We fixed mutually source- and document-disjoint sets of 16 engineering, 300 formal and 20 reserve sources. The runtime used one A100 80GB, FP32 eager attention and disabled TF32. Five arms covered full cache and recency/cumulative retention at budgets 1,024 and 512; 1,024 determined the primary decision and 512 was a prespecified sensitivity.

Engineering gate. Stock, chunked and traced full-cache paths had to agree exactly in prompt IDs, greedy trajectory, coordinates, answers, metrics and cache lengths. Compressed paths additionally required tracing-neutral state/cache hashes and valid budgets, protected positions and coordinate inverses. All 16 sources and 176 paths passed independent CPU validation. Logit diagnostics were finite and recorded for 3,858 aligned steps but did not determine GO/STOP; exact summaries remain in the artifact.

Formal execution and independent replay. The panel produced 1,500 records (300 sources by five arms). Initial containers whose GPU product string differed from the gate stopped before inference; only zero-record shards were retried under identical manifests, code and assignments, with all attempts linked in the artifact. Independent CPU replay verified every record hash, arm, tokenizer mapping, coordinate transform, budget, metric and runtime field before 10,000 paired-source bootstrap resamples (seed 2026092301).

Table 25: Llama revision-stage frozen scope audit. Differences are cumulative minus recency; intervals are marginal paired-source bootstrap 95% intervals.

Budget	Readout	Recency	Cumulative	Difference [95% CI]
1,024	Answer F1	0.059	0.065	$+0.006$ $[-0.013, +0.026]$
1,024	Token coverage	0.679	0.713	$+0.034$ $[-0.009, +0.078]$
1,024	Complete-unit fraction	0.670	0.120	-0.550 $[-0.594, -0.505]$
1,024	All-units completeness	0.460	0.017	-0.443 $[-0.500, -0.387]$
512	Answer F1	0.039	0.069	$+0.031$ $[+0.010, +0.053]$
512	Token coverage	0.287	0.352	$+0.065$ $[+0.024, +0.106]$
512	Complete-unit fraction	0.281	0.068	-0.213 $[-0.246, -0.180]$
512	All-units completeness	0.073	0.000	-0.073 $[-0.103, -0.047]$

Predeclared decision. A complete ranking-conflict replication required both the lower 95% bound for ΔF1 to exceed zero and the upper 95% bound for ΔA to be below zero at budget 1,024. The observed F1 interval crosses zero, whereas the all-units interval is wholly negative. The predeclared label is therefore `STRUCTURAL_ONLY_REPLICATION`. The 512-position sensitivity shows a positive F1 contrast and a negative integrity contrast, but it cannot replace the primary decision. Full-cache F1 is 0.083, so this panel also offers limited power for claims about utility. The supported conclusion is narrower: the recency–cumulative integrity separation transfers to this setting, while the primary-budget F1 advantage does not establish a cross-model replication.

M Official Joint-Output and Matched-Repair Boundary

Evidence level and chronology. This revision-stage extension was designed after the original primary panel. Its prompt and finite parser were developed only on 16 engineering sources. The final execution used FP32 eager attention with TF32 disabled after BF16 one-shot and chunked full-cache decoding diverged on three engineering cases. Under FP32, stock, chunked and attention-traced full-cache decoding matched token for token on all 16 cases. At `2026-09-23T04:14:00Z`, before the first calibration execution or outcome inspection, the protocol, parser, code, cohort manifests and repair planner were frozen under SHA-256 `a24d4e1ef5fab7513af39d823d3daef5b8508ae6e48a773544d6bb4908a368f1`. The subsequent formal freeze at `2026-09-23T04:27:58Z`, SHA-256 `ff4e9b3232d677c1d4545da4d05bee470db59f4adb771657482d3af41ace9b0a`, preceded all formal policy and repair outcomes.

The exposure audit scanned 14,288 working files, 15,523 archive members and 12,456 reachable Git blobs for recoverable HotpotQA source IDs and document titles. The resulting engineering (16), calibration (32), formal (300) and reserve (32) cohorts are mutually source- and document-disjoint and exclude all locally recoverable prior project exposures. This audit cannot enumerate remote-only or deleted uncommitted records and does not address model pretraining exposure.

Joint-output contract and calibration. Every passage sentence receives a reversible three-digit local ID in original dataset order. The final question and two-line output instruction occupy at most the protected recent 128 positions. The parser accepts a prospectively finite set of surface variants: case and whitespace differences, ASCII or full-width punctuation, bracketed IDs, and bare `E` plus integer IDs. It extracts IDs only from the support field, preserves out-of-catalog IDs as false positives, and never infers support from answer text, titles or gold annotations. Local scoring matches the official HotpotQA evaluator on 2,000 randomized differential fixtures.

The single 32-source full-cache calibration passed every frozen gate: 32/32 outputs contained both fields, 32/32 contained a valid support ID, mean Supporting-Fact F1 was 0.5061 and mean Joint F1 was 0.3039. An outcome-blind cumulative prefill found valid R/S/P repair plans for 29/32 states. No cumulative-versus-recency behavioral outcome was generated in calibration.

Table 26: Official HotpotQA metrics and literal-residency readouts on the frozen 300-source formal panel.

Policy	Answer EM	Answer F1	Support F1	Joint F1	Coverage	All units
Full	0.350	0.514	0.491	0.260	1.000	1.000
Recency	0.320	0.448	0.305	0.175	0.483	0.193
Cumulative	0.287	0.458	0.354	0.168	0.592	0.003
Width-16 block	0.297	0.448	0.363	0.184	0.578	0.110

The primary pairwise estimands are cumulative minus recency. Answer F1 is $+0.0102$ (two-sided 95% interval $[-0.0441, 0.0631]$; one-sided 95% lower bound -0.0357). Joint F1 is -0.00686 (two-sided interval $[-0.0438, 0.0306]$; one-sided upper bound $+0.0238$). The frozen conjunction therefore returns `POLICY_COMPLETE_NO_REVERSAL`. Answer-F1 win/tie/loss counts are 76/158/66; Joint-F1 counts are 96/115/89.

Offline matched repair. Among the 300 cumulative states, 266 admit every frozen repair branch. R restores the smallest missing bundle that completes one annotated sentence; S restores the same number of missing evidence positions without completing any previously incomplete sentence; P restores the same number of absent non-evidence document positions. All three remove one identical set of resident non-evidence positions. Restored tensors are exact admission-time KV snapshots, branches retain 1,024 positions, and the final readout token is recomputed before decoding. The median swap count is 7 (range 1–60). Every restored tensor passed bitwise equality against its saved snapshot.

Table 27: Repair outcomes on all 266 eligible cumulative states. Intervals are paired two-sided 95% bootstrap intervals. Because the policy reversal failed, all repair contrasts are descriptive under the frozen fixed sequence.

Branch	Answer F1	Support F1	Joint F1
Unchanged cumulative (C)	0.4574	0.3363	0.1625
Complete-unit restoration (R)	0.4759	0.3365	0.1660
Scattered-evidence restoration (S)	0.4706	0.3377	0.1674
Non-evidence sham (P)	0.4494	0.3301	0.1600
R − S			$-0.00133\ [-0.00409, 0.00036]$
R − P			$+0.00603\ [0.00131, 0.01189]$
R − C			$+0.00355\ [-0.00377, 0.01103]$

R versus S does not establish a benefit for the complete-unit restoration rule. R versus P is positive, but changes evidence content as well as placement and is descriptive after the testing sequence stopped. None of these comparisons holds token identity or contextual state fixed, so they do not estimate a controlled direct effect of completeness.

N CONTRACT-ADAPTED H2O GQA AUDIT

This separately frozen revision-stage extension asks whether a released heavy-hitter rule shows the utility–residency separation on the exact 500-source primary cohort. It is neither an original prospective confirmation nor a reproduction of H2O's published benchmark. The frozen H2O-minus-recency rule requires positive lower 95% bounds for F1 and coverage and a negative upper bound for complete-unit fraction; residency alone yields the label STRUCTURAL-ONLY. Intervals use 10,000 paired draws (seed 20260923).

Contract and validation. The pinned H2O implementation cannot directly apply its per-query-head mask to Mistral's 32-query-head/8-KV-head GQA cache. Our declared adapter sums attention mass within each four-query-head KV group, then applies the released per-head top-k rule. Each layer/KV head retains 880 heavy-hitter, 128 recent and 16 protected prefix positions. Selection follows each prompt-chunk admission; the last selected prompt token is recomputed before greedy decode. We therefore call this *contract-adapted H2O GQA*, not an unmodified upstream binary. Tokenizer compatibility and RoPE-table sizing amendments preceded outcomes and changed no scores, sources, metrics or decision rules. All 16 engineering sources passed output, repeat, coordinate, shape, protection and byte checks. Formal CPU replay verified all 500 prompts, 128,000 layer/head keep sets and reported aggregations. Exact commits, hashes, amendment records, hardware and validator logs are in the artifact.

Table 28: Contract-adapted H2O GQA paired differences with 95% intervals. H2O residency uses mean-head aggregation.

Readout	H2O − Recency	H2O − Cumulative
Answer F1	$+0.2289\ [0.1888, 0.2686]$	$+0.0084\ [-0.0143, 0.0306]$
Token coverage	$+0.0262\ [-0.0104, 0.0624]$	$-0.0601\ [-0.0676, -0.0527]$
Complete-unit fraction	$-0.5052\ [-0.5387, -0.4720]$	$-0.0518\ [-0.0651, -0.0392]$
All-units residency	$-0.3570\ [-0.3997, -0.3155]$	$-0.0090\ [-0.0189, -0.0011]$

H2O mean F1 is 0.5264, versus 0.2975 for recency and 0.5179 for cumulative; mean-head complete-unit fraction is 0.0780, versus 0.5832 and 0.1298. Its F1 and residency contrasts against recency exclude zero, but coverage does not. The frozen decision is therefore STRUCTURAL-ONLY, not a three-readout replication.

Table 29: Aggregation sensitivity for the same H2O cache. Union uses original positions present anywhere across 32 layers and 8 KV heads; intersection requires presence everywhere.

Aggregation	Coverage	Unit fraction	All units
Mean head	0.6069	0.0780	0.0050
Union	0.999967	0.9983	0.9960
Intersection	0.0721	0.0524	0.0000

The union contains 1,157–3,428 distinct positions per source (mean 1,800.6), so it is not a 1,024-slot policy arm. Near-perfect union residency means that almost every unit has one complete physical

copy somewhere; zero intersection does not imply global absence. The aggregation rule must therefore be declared.

An initial dispatch referenced a different 500-source manifest and was stopped when the prewritten paired join found zero ID overlap, before aggregate outcome calculation. The corrected manifest matches all primary IDs, prompt hashes and lengths; code and decision rules were unchanged. The stop report is archived. This audit covers one adapted rule, model, task and budget. It makes no latency, throughput, decode-time eviction or universal compressor-ranking claim.

O POST-OUTCOME SCORING, METRIC-LADDER, AND AGGREGATION AUDITS

This appendix reports CPU-only analyses requested after the frozen model outputs were known. They use the complete saved cohorts without selecting sources or rerunning inference. They are robustness and interpretation checks, not new prospective confirmations. Scripts, source-level tables, input hashes, and deterministic seeds are included in the artifact.

O.1 OFFICIAL HOTPOTQA ANSWER RE-SCORING

We pinned the answer functions in the official HotpotQA evaluator at upstream commit `3635853403a8735609ee997664e1528f4480762a`. Its normalization lowercases, removes punctuation and articles, and normalizes whitespace; its F1 also applies the official yes/no/noanswer rule. The re-score joins all 500 frozen source IDs to the canonical answer records, verifies within-source prompt hash agreement, and scores the unchanged post-selection predictions. Every source has exactly one canonical gold answer in this panel. Intervals use the same 10,000 paired source bootstrap design with seed 20260916.

Table 30: Post-outcome official HotpotQA answer re-score of the frozen 500-source predictions. The frozen custom-score endpoint remains primary.

Metric	Recency	Cumulative	Cumulative − recency [95% CI]
Official answer F1	.298	.515	$+.217$ $[+.175, +.258]$
Official answer EM	.146	.344	$+.198$ $[+.156, +.242]$

Full-cache official answer F1 is .579 and EM is .400. Cumulative beats recency on 226 sources for F1 and 119 for EM, with 102 and 20 losses, respectively. The result supports scorer robustness under this fixed generation contract; it does not turn the experiment into an official leaderboard submission or alter the prospective evidence level.

O.2 LITERAL-RESIDENCY METRIC LADDER AND STRATA

For each annotated unit, let $r_{si} = |F_{si} \cap K_s|/|F_{si}|$ and define $I_s(\tau) = m_s^{-1} \sum_i \mathbf{1}[r_{si} \geq \tau]$. The exact complete-unit fraction is $I_s(1)$. Figure 9 reports the full-cohort paired contrasts. Thresholds $\{.80, .90, .95, .99, 1.00\}$ were recorded before the original threshold summaries were computed; $\tau = .75$ was added later as an explicitly post-outcome descriptive sensitivity. All displayed threshold contrasts are negative; the 95% interval is below zero even at $\tau = .75$. Thus the result is not unique to a single-token exact-compliance discontinuity, although its magnitude grows as the required retained fraction approaches one.

Table 31: Selected full-cohort ladder results. All nonexact thresholds are post-outcome sensitivities.

Endpoint	Recency	Cumulative	Cumulative − recency [95% CI]
$I_s(.75)$	.587	.456	$-.131$ $[-.179, -.082]$
$I_s(.80)$	.586	.408	$-.177$ $[-.224, -.129]$
$I_s(.90)$	.585	.277	$-.308$ $[-.352, -.264]$
$I_s(.95)$	.584	.202	$-.382$ $[-.423, -.342]$
$I_s(.99)$	.583	.130	$-.453$ $[-.491, -.416]$
$I_s(1.00)$	.583	.130	$-.453$ $[-.491, -.416]$
Any exact unit	.792	.266	$-.526$ $[-.578, -.474]$

No source in the primary panel has only one annotated unit: 335 have two and 165 have three or more. At $\tau = .75$, the paired difference is $-.127$ $[-.183, -.069]$ for two-unit sources and $-.140$

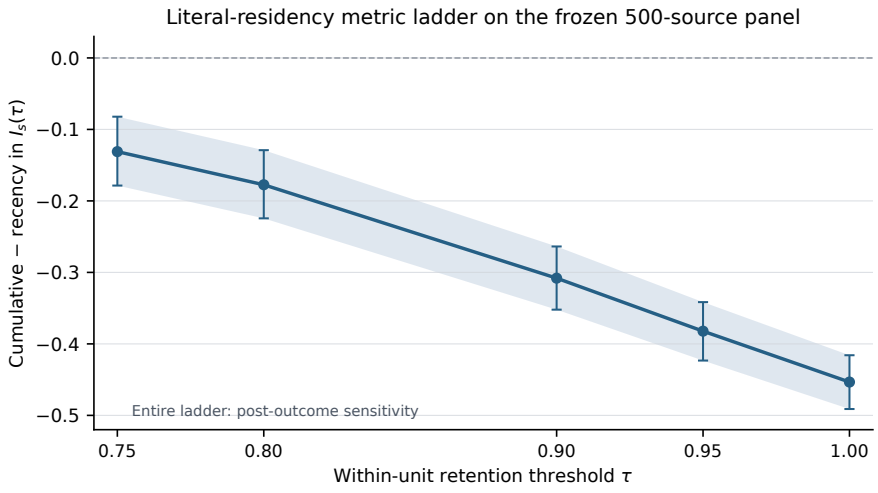

Figure 9: Post-outcome literal-residency ladder on the complete frozen 500-source panel. Points are cumulative-minus-recency means across sources; bars are paired 95% bootstrap intervals. $I_s(1)$ is the previously reported exploratory exact complete-unit fraction. The originally recorded grid was $\{.80, .90, .95, .99, 1.00\}$; $\tau = .75$ is an additional post-outcome descriptive sensitivity. None replaces the frozen all-units endpoint.

$[-.228, -.049]$ for sources with three or more units. Exact-unit differences are likewise negative in both strata. Unit-length quantiles show the same exact-residency direction; the lowest-threshold interval can cross zero in an individual length bin, so the analysis does not claim uniform effects at every threshold and length.

O.3 H2O AGGREGATION DISTRIBUTION

The saved H2O trace contains 32 layers by 8 KV-head stores for each of all 500 sources, each with exactly 1,024 original-coordinate positions. We recomputed every metric from those keep sets and verified the frozen compact summaries. Table 32 and Figure 10 show why the aggregation must be named: the table reports three summaries and the figure gives the source-level distribution.

Table 32: Post-outcome H2O aggregation audit. Intervals are 10,000-draw source-bootstrap 95% intervals.

Aggregation	Complete-unit fraction	All-units residency
Mean layer/head store	.078 [.065, .092]	.005 [.003, .008]
Union across stores	.998 [.996, 1.000]	.996 [.990, 1.000]
Intersection across stores	.052 [.040, .066]	.000 [.000, .000]

Union asks whether an intact physical copy exists anywhere; intersection asks whether every store retains one; and mean-head summarizes the distribution over fixed-capacity stores. None determines which store is functionally used during decoding, and none is a semantic-fidelity metric.

O.4 CONDITIONAL SECOND-TASK GATE

We audited whether a fresh 150–200-source Mistral/Qasper panel could be run by direct reuse. The outcome-free candidate file has 881 rows, including 641 source IDs disjoint from the prior 240-source panel. Data availability passes, but execution readiness does not: the reusable runner lacks the required trace-on/off equality gate and final-prompt-token recomputation after the last compaction, and no new disjoint manifest has been frozen. Prior Qasper policy outcomes are also known, so any later panel would be revision-stage scope evidence. Following the conditional plan, we stop before GPU inference. This is an engineering eligibility stop, not a task result.

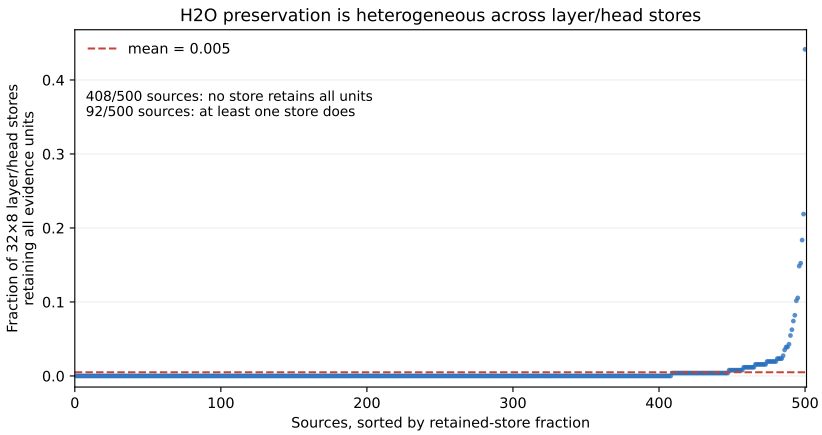

Figure 10: For each source, the fraction of H2O layer/head stores that retain all annotated units. Among 500 sources, 408 have a fraction of zero, 92 have a nonzero fraction below one, and none has a fraction of one. The union nonetheless succeeds on 498 sources because different stores retain different physical copies.

P RELEASED-METHOD FEASIBILITY GATE

Before launching a new matched-contract SnapKV panel, we ran a revision-stage, zero-GPU gate requiring a pinned official execution, exact layer/head coordinates, matched logical KV bytes, the primary post-selection readout, and 16/16 traced-versus-reference keep-set and output agreement. Only NATIVE-EQUIVALENCE FEASIBLE authorized formal inference.

The decision was CONTROLLED PORT ONLY; no new GPU panel was launched. The runner executes a hash-pinned KVCache-Factory path rather than the pinned official FasterDecoding/SnapKV path. Its 16-source check proves bitwise trace neutrality for the same port, not official keep-set or output equivalence. The port preserves exact coordinates for 32 layers by 32 query heads and therefore supports the aggregation analysis in Appendix J.2. It does not provide a matched comparison: at nominal capacity 1,024, logical prefill K+V averages 512.0 MiB versus 225.1 MiB for the stock full cache, and its first-token readout differs from the primary post-selection contract.