COUNT COMPETITORS, NOT CONTEXT: WHEN KV-CACHE BUDGETS TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Every KV-cache eviction method must decide how many tokens to keep, and practice sets that budget as a fraction of the prompt. We show that the fraction is the wrong unit. The budget is set by competition: evidence survives eviction exactly when the budget covers the evidence plus the tokens the scorer ranks above it, and prompt length matters only through those competitors. Prior work misses this for two reasons. It evaluates budgets on prompts where length and competing content grow together, so their effects cannot be separated, and it treats keeping the evidence as enough to answer. Separating length from competition exposes costly errors in both directions: on long prompts with little competing content, a fixed fraction keeps about 100 times more cache than needed, while at a fixed length, added distractors raise the required budget more than 200-fold. A count calibrated once at short context forecasts budgets at up to 512K tokens, where a constant fraction fails. Keeping the evidence is also not enough: answers depend on keeping the rows that bind the evidence to the question. Compacting the cache to these budgets cuts KV memory by up to two orders of magnitude with no loss of accuracy. The rule for sizing a cache is simple: count competitors, not context.

1 INTRODUCTION

KV-cache memory bounds long-context serving. The cache grows linearly with the prompt: one 32K request to Qwen2.5-7B-Instruct-1M (Yang et al., 2025) stores 1,792 MiB of keys and values, and a 256K request needs eight times as much. Eviction methods therefore keep a subset of cached positions. They size that subset either as a fraction of the prompt or as a token count chosen at one setting (Li et al., 2024; Cai et al., 2025; Feng et al., 2025; Jégou et al., 2024). Neither choice comes with a rule for when the budget remains valid at another length or on other content.

That missing rule is costly in both directions. On literal retrieval, Qwen3.5-9B (Qwen Team, 2026) with negative-log-likelihood (NLL) scoring reaches half its full-cache accuracy with 24–28 retained positions at every length from 16K to 256K. A 1% budget keeps 164 positions at 16K and 2,621 at 256K; at the longest length it spends 100 times the memory the task needs. At a fixed 32K length, however, 512 distractor statements raise the requirement to 5,277 positions, sixteen times what the same 1% budget keeps. A fixed fraction is wasteful in one setting and insufficient in the other because it tracks neither quantity that sets the budget.

We identify that quantity. For any selector that ranks cache positions in a fixed order, a set S of evidence tokens survives top-N retention exactly when $N \geq R = |S| + C$, where C counts the non-evidence positions ranked ahead of the weakest member of S (Section 2). The identity is elementary; its consequence is not. Length has no direct role. Retention at a given budget transfers to a new length exactly when the distribution of R transfers, and a fixed fraction is the right unit only when competitors grow in proportion to length. Accuracy follows retention up to two correction terms, wrong answers despite retained evidence and correct answers without it, and we measure both on paired prompts rather than assume them away (Section 3).

We test each input to R with interventions that change it while holding the others fixed. The study spans 16K calibrations on 14 models, paired length sweeps up to 256K, controlled distractor and filler doses on three model families, paired retention–generation measurements, four structured benchmark suites, and physical cache compaction on dense and hybrid architectures. Our contributions are:

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

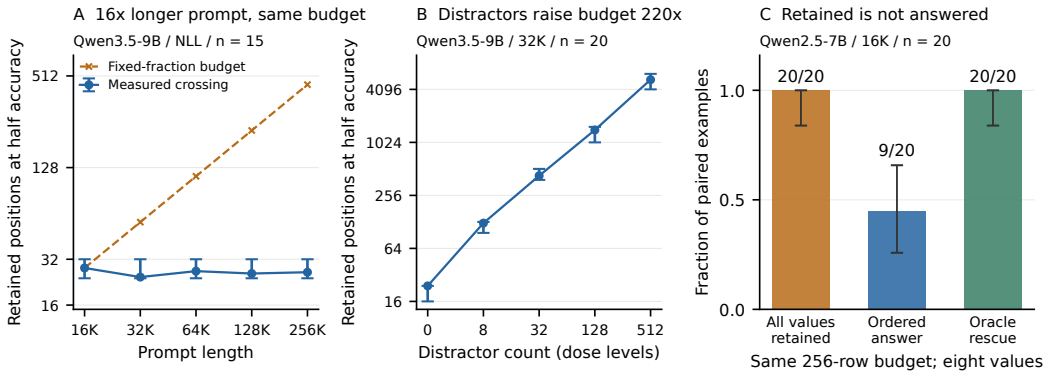

Figure 1: **Competitors, not length, set the budget, and retaining evidence does not guarantee an answer.** (A) Qwen3.5-9B NLL crossings stay at 24–28 retained positions from 16K to 256K; the dashed line scales the 16K crossing in proportion to length, as a fixed fraction would. (B) At a fixed 32K length, distractors raise the crossing from 24.0 to 5,277.4 positions. Bars in A and B mark the two tested budgets that bracket each crossing. (C) At $N = 256$, all eight annotated values survive on 20/20 prompts, but only 9/20 ordered answers are correct; retaining complete records and the question scores 20/20. Bars in C are 95% Wilson intervals (Wilson, 1927).

- **Competitors, not length, set the budget.** We derive the retention identity, a distribution-free finite-range bound, and a bound on how accuracy changes across lengths (Sections 2–3). At a fixed 32K length, distractors raise the NLL budget 220-fold on Qwen3.5-9B and 158-fold on Gemma-4-E4B. Adding filler instead leaves it flat: the long-to-short ratio is 1.00 on Qwen (8× filler) and 0.99 on Gemma (4×), where a fixed fraction predicts 8 and 4.

- **Counts transfer, and short-context calibrations forecast long-context budgets.** Three Qwen checkpoints hold 23–34 positions across a sixteenfold length range. The log–log slope of the budget against length is -0.011 on Qwen3.5-9B, where a fixed fraction predicts 1, and retained counts align accuracy curves better than fractions with every paired interval excluding zero. Committed forecasts land within a factor of two on 22 of 25 testable cells up to 512K; a constant fraction fitted post hoc to the same calibration lands on none of 23. The 14-model calibration maps where count transfer fails.

- **Retaining evidence is not answering.** All eight requested values survive on 20/20 prompts while only 9/20 answers are correct. On Llama-3.1-8B, the single requested value survives at every budget from 32 to 2,048 positions and the model answers 0 of 140 in order. Keeping complete key–value records and the question restores 20/20 on Qwen; records without the question score 0/20. Needle-style retention audits therefore overstate what a compressed cache can answer, and allocators must keep the rows that bind a value to its query.

- **Logical budgets become physical memory.** Compacting the cache after full 32K prefill reduces cache storage from 1,792 to 17.5–18.4 MiB on a dense model (97–102×) and from 1,049.5 to 92.5 MiB on a hybrid model (11×), with 20/20 answers correct per model. The hybrid result isolates the recurrent state and block granularity that a token count cannot reclaim.

The same framework explains where a literal-retrieval count stops transferring. On RULER, BABILong, InfiniteBench, and LongBench, the evidence set and its competitors change with the task, and so does the budget: on RULER needle retrieval, TOVA reaches half accuracy at 0.36% of the cache while NLL, the best literal-retrieval scorer, stays below it through 20% (Section 5.5). The practical rule follows directly. Report budgets as retained counts with the scorer, the full-cache score, and the competitor load; verify answers, not only retained evidence, before reusing a budget; and measure bytes after physical compaction.

2 EVIDENCE SURVIVES AT RANK $|S| + C$

A prompt has L candidate cache positions. A scorer orders them once, breaking ties by a fixed rule. Write $r(i)$ for the rank of position i and $K_N = \{i : r(i) \le N\}$ for the first N positions. Because the order does not depend on N, the retained sets are nested. This covers content scorers such as NLL, which keeps the prompt tokens the model finds most surprising (highest negative log-likelihood), and learned content predictors. Window, chunk, and per-head policies have their own containment events, treated below and in Appendix A.

Let S be a specified, nonempty evidence set and define

$$R = \max_{i \in S} r(i), \qquad C = \big|\{j \notin S : r(j) < R\}\big|. \tag{1}$$

Proposition 1 (Evidence retention)**.** *For fixed S and a fixed total ordering,*

$$S \subseteq K_N \iff R \le N, \qquad R = |S| + C. \tag{2}$$

Across prompts, the retention probability is therefore $F_R(N) = \Pr(R \le N)$.

Proof. The last-ranked member of S is retained exactly when $R \le N$. Among the first R positions, $|S|$ belong to S and C do not. □

From retention to answers. Let Y_N indicate a correct answer from the compressed cache, and let $E_N = \{R \le N\}$. Splitting correct answers by whether S survives gives

$$A(N) = \Pr(Y_N = 1) = F_R(N) - \Pr(E_N, Y_N = 0) + \Pr(E_N^c, Y_N = 1). \tag{3}$$

The first correction counts wrong answers despite retaining S; the second counts correct answers after losing part of S. Their sum bounds $|A(N) - F_R(N)|$, and accuracy equals the rank CDF when both vanish. Because the two terms can cancel, matching curves alone cannot confirm that S is what the model uses. We therefore measure retention and correctness on the same prompts (Section 5.4) and report full-cache accuracy with every compressed result.

Alternative and joint evidence. If any one of a family $\mathcal{S}$ of evidence sets suffices, retention first holds at

$$R_{\mathcal{S}} = \min_{S \in \mathcal{S}} \max_{i \in S} r(i). \tag{4}$$

Jointly required spans are merged into one set before taking the maximum. On natural tasks $\mathcal{S}$ is latent and must be measured rather than annotated after the fact, which is why structured tasks need their own calibration (Section 5.5).

Heads and layers. An attention selector may rank positions separately for each cache-owning head and layer. A required entry then survives only if $r_{\ell h}(i) \le N_{\ell h}$ on its route, and averaging ranks over routes changes the event. We analyze such selectors route by route (Appendix F).

3 COMPETITORS, NOT LENGTH, SET THE BUDGET

The identity $R = |S| + C$ splits the budget into evidence size and competition. For each outside position j, let I_j indicate that it ranks ahead of the last member of S. Then

$$\mathbb{E}[C \mid S, T] = \sum_{j \notin S} \Pr(I_j = 1 \mid S, T), \tag{5}$$

where T is the evidence-score threshold. The sum requires no independence assumption, and length enters only as its number of terms. Two regimes follow. When competition is tight, few outside positions ever outrank the evidence, R stays near $|S|$, and a retained count transfers across lengths. When every outside position outranks the evidence with the same probability p, $\mathbb{E}[C] = (L - |S|)p$ grows linearly and a fixed fraction near p is the right unit. The scorer and the content determine which regime holds; length alone cannot. An experiment that varies only L cannot tell the regimes apart, so our interventions vary competitors and filler separately. A flat budget over a finite range

establishes $Lp \ll 1$ on that range, not $p \propto 1/L$. Softmax normalization rescales every score by the same factor, preserves ranks, and so cannot by itself bound competition.

For evidence size m and integer $k \geq 0$, Markov's inequality gives a distribution-free finite-range bound:

$$\Pr(R > m + k) = \Pr(C \geq k + 1) \leq \frac{\mathbb{E}[C]}{k + 1}. \tag{6}$$

If $\mathbb{E}[C] < \varepsilon$, an m-position budget retains S with probability at least $1 - \varepsilon$. A reliability target q requires the quantile $F_R^{-1}(q)$ rather than the mean: stable competitor distributions make a count transferable when evidence size is bounded, and competitor counts concentrated near ρL make a fraction transferable.

Accuracy transfer. Let $\epsilon_L(N) = \Pr(E_N, Y_N = 0) + \Pr(E_N^c, Y_N = 1)$ at length L. Equation 3 yields

$$|A_{L_2}(N) - A_{L_1}(N)| \leq |F_{R,L_2}(N) - F_{R,L_1}(N)| + \epsilon_{L_1}(N) + \epsilon_{L_2}(N). \tag{7}$$

An accuracy target transfers when ranks are stable and retention predicts generation at both lengths. A lower bound needs only the retained-but-wrong term, $A_L(N) \geq F_{R,L}(N) - \Pr(E_N, Y_N = 0)$, and correct answers without every annotated row make it conservative. Section 5 tests each input: filler at a fixed task, distractors at a fixed length, and the retained-but-wrong term on paired prompts.

4 Experimental protocol

Execution. All experiments ran on eight NVIDIA B300 GPUs on one host. For answer quality, we prefill the full prompt, compress the KV cache after each layer's attention, and decode at the original token positions. The first generated token uses full-prefill logits and later tokens use the compressed cache. For physical memory, we finish full prefill and copy the selected KV entries into smaller tensors while keeping all other model state. Both protocols compute every representation from the full prompt, so neither removes prefill work. The artifact records model revisions, data snapshots, seeds, budgets, outputs, and cache measurements (Appendix B).

Models. The 14-model calibration covers Qwen2.5, Qwen3.5, Gemma-4, GLM-4.7, Phi-4, Llama-3.1, and Mistral-Nemo checkpoints (Yang et al., 2025; Qwen Team, 2026; Gemma Team, 2026; GLM Team et al., 2025; Abdin et al., 2024; Grattafiori et al., 2024; Mistral AI Team, 2024); Appendix B.1 lists every checkpoint.

Tasks, policies, and outcomes. The controlled task retrieves a ten-digit value from filler text. We test NLL, a separately trained four-feature content predictor (Appendix L), TOVA (Oren et al., 2024), SnapKV (Li et al., 2024) with a w-token observation window (SnapKV-w32), PyramidKV (Cai et al., 2025), and native structured baselines. We also test UnionKV, our content-plus-attention allocator (Section 6), in fixed-share, adaptive (UnionKV-v2), and count-first variants. Retrieval accuracy checks whether the normalized answer appears in the normalized response. LongBench QA (Bai et al., 2024) uses token F1; RULER (Hsieh et al., 2024) uses its official mean reference recall for multi-item tasks and any-reference matching for QA; BABILong (Kuratov et al., 2024) and InfiniteBench (Zhang et al., 2024) supply reasoning and natural-format retrieval. Query visibility is matched within every comparison.

Crossings and validity. The crossing N_{50} is the first budget at which measured accuracy reaches half the full-cache score, interpolated on the log-budget axis between the two adjacent tested budgets that bracket it. A cell is valid only when full-cache accuracy is at least 0.8; invalid cells are reported but never scored. Intervals come from paired bootstrap resampling of example identities across policies and budgets (Efron, 1979) with 2,000 resamples, and from Wilson intervals for single proportions.

Coordinate and forecast tests. To compare retained count N with fraction N/L, we interpolate each pair of adjacent-length accuracy curves over their shared transition and average the across-curve standard deviation; lower dispersion means that coordinate aligns the curves better. Appendix Figure 4 gives the 41-point grid and inclusion rule. Forecast error is $e = \log_2(N_{\text{measured}}/N_{\text{predicted}})$. We

Table 1: **Competitors move the NLL budget; filler does not.** Left: N_{50} in retained positions at a fixed 32K length as M distractor statements are added (20 paired prompts per dose; every full-cache arm scores 20/20). Right: ratio of long- to short-context N_{50} at fixed $M = 32$ with paired 95% intervals; a fixed fraction predicts the filler factor.

	Distractors M at 32K						Filler at $M = 32$	
Model	0	8	32	128	512	Fold	Filler	Ratio
Qwen3.5-9B	24.0	124.0	427.7	1,416.4	5,277.4	220×	8×	1.00 [0.93, 1.07]
Gemma-4-E4B	35.9	137.8	416.9	1,299.8	5,665.4	158×	4×	0.99 [0.90, 1.09]

report numerical error over crossed cells and factor-of-two coverage over all testable cells, counting bounded non-crossings as misses (Appendix D).

5 Experimental results

5.1 Competitors raise the budget 220-fold at fixed length; filler does not

Distractors. Each distractor statement pairs a non-matching key phrase with its own random ten-digit value, so its digits are as surprising to the model as the answer's and compete directly for NLL rank. At 32K on Qwen3.5-9B, the NLL crossing rises from 24.0 to 124.0, 427.7, 1,416.4, and 5,277.4 retained positions as we add $M = 0, 8, 32, 128, 512$ statements (Table 1). The full cache answers 20/20 at every dose: the distractors confuse the scorer, not the model. Length is fixed, so competition alone produces this 220-fold increase, and each added statement costs 10.05–12.66 retained positions between adjacent doses (Appendix Table 13). Gemma-4-E4B reproduces the curve on a different architecture, rising 158-fold from 35.9 to 5,665.4 positions (Appendix Table 15).

Filler. Holding $M = 32$ fixed while growing the prompt eightfold, from 8K to 64K, leaves the Qwen NLL crossing unchanged at a ratio of 1.00 ($[0.93, 1.07]$); SnapKV-$w32$ is also flat at 1.07 ($[0.80, 1.34]$). A native Gemma-4-E4B repeat over a fourfold range gives 0.99 ($[0.90, 1.09]$) for NLL. The same intervention identifies scorers whose competition does grow with filler, the second regime of Section 3: adaptive UnionKV-v2 rises 2.38× ($[1.52, 2.62]$) on Qwen and the learned content predictor rises 2.68× ($[2.32, 3.21]$) on Gemma (Appendix Table 16). On Llama-3.1-8B, unprotected NLL never reaches half accuracy through 8,192 positions at any dose, whereas UnionKV-v2 rises from 19.6 to 1,773.6 positions (Appendix Table 14). Whether a count transfers is thus a measurable property of the scorer–model pair, and these two interventions measure it directly.

5.2 Counts transfer across a sixteenfold length range

Qwen. We paired the same 15 literal-retrieval examples across five lengths (16K, 32K, 64K, 128K, and 256K) on Qwen3.5-9B and Qwen2.5-7B-Instruct-1M; every full-cache baseline scores 15/15. NLL crossings are 28.0, 24.4, 26.7, 25.8, and 26.3 retained positions on Qwen3.5-9B, which interleaves attention with recurrent layers, and 30.5, 31.4, 34.2, 31.4, and 29.8 on the dense Qwen2.5 model. A fixed fraction implies a log–log slope of 1 between budget and length; the measured slopes are -0.011 ($[-0.039, 0.009]$) and -0.007 ($[-0.040, 0.022]$). Across adjacent lengths, retained counts align accuracy curves better than fractions, reducing dispersion by 0.189 ($[0.134, 0.230]$) and 0.190 ($[0.161, 0.209]$). A four-length Qwen3.5-4B sweep from 16K to 128K replicates the result, with crossings of 23.4–26.8 and a dispersion reduction of 0.203 ($[0.153, 0.245]$). Appendices B.2–B.4 give every curve.

Beyond Qwen. GLM-4.7-Flash NLL crosses at 24.5, 37.3, 34.2, and 27.7 positions from 16K to 128K with 15/15 full baselines, and counts again reduce dispersion (0.070, $[0.035, 0.093]$), although several 128K curves dip after first crossing. Gemma-4-E4B NLL rises from 24.7 to 35.4 positions between 16K and 32K, and neither coordinate is preferred (-0.054, $[-0.130, 0.039]$; Appendix B.5). The 14-model 16K calibration shows the same spread at a single length: Qwen3.5-27B NLL crosses at 21.1 positions, Gemma-4-26B-A4B at 227.1, and Llama-3.1-8B not at all (Appendix B.1). A retained count belongs to a scorer on a model and must be calibrated per pair; once calibrated, it

transfers across length wherever competitors are stable. Earlier Gemma attention-policy rows used an unsupported query reconstruction and are excluded from all native-policy evidence (Appendix B).

5.3 Short-context calibrations forecast long-context budgets

The forecast registry contains 39 committed cells: 23 numerical crossings, two valid cells that stay below target beyond twice the predicted budget, and 14 unscoreable cells whose full-cache reference fails the validity gate, whose kernel cannot run, or whose budget ladder stops early (Appendix Table 8). On the 23 crossings, mean $|\log_2(N_{\text{measured}}/N_{\text{predicted}})|$ is 0.274, and 22/23 lie within a factor of two. A constant-fraction comparator fitted post hoc to the same calibration has mean error 2.698 and places none of the 23 within that band. Counting the two non-crossings as misses gives 22/25 testable cells within a factor of two, across six models in four families. The three misses are one SnapKV and two TOVA cells, all window or current-query policies, two of them on Llama-3.1-8B.

5.4 Retaining evidence is not answering

The paired diagnostic uses Qwen2.5-7B-Instruct-1M at 16K with $K \in \{1, 2, 4, 8\}$ requested ten-digit values, one or three copies per value, and 20 prompts per condition. It records value retention and ordered-answer correctness from the same NLL selection; every full-cache condition scores 20/20. For eight unique values at $N = 256$, every annotated value survives on 20/20 prompts, yet only 9/20 ordered answers are correct (95% Wilson interval $[0.26, 0.66]$), and accuracy stays at 8–10/20 through $N = 2,048$. The gap is starker on Llama-3.1-8B: the single requested value survives at every budget from 32 to 2,048 positions, yet ordered accuracy is 0/20 at each of those seven budgets, against 20/20 with the full cache. Protecting the first four prompt rows at $N = 128$ restores 13/20 (exact McNemar $p = 2.4 \times 10^{-4}$ (McNemar, 1947)). The error also runs in the other direction: three Qwen prompts with two requested values succeed at $N = 64$ without complete value retention. Value retention is neither sufficient nor necessary for an answer (Appendix G.6).

Forced selections at the same $N = 256$ and on the same prompts locate the missing information (Appendices G.7–G.8). Replacing all but the first or last row of each value with matched donors destroys every successful answer, a causal effect of the value rows ($p = 1.9 \times 10^{-6}$ for $K = 1$). Retaining all value rows yields 20/20, 17/20, and 9/20 correct answers for $K = 1, 4, 8$. For $K = 8$, adding the question rows raises accuracy to 18/20, keeping complete key–value records with the question gives 20/20, and keeping the records without the question gives 0/20. The failing ingredient is the binding between values, their keys, and the query. An allocator that protects only the values, which is exactly what needle-retention audits check, discards these rows.

5.5 Structured tasks need their own budgets

Structured benchmarks change both the evidence set and its competitors, and the best selector changes with them (Appendix M). On 16K RULER needle retrieval, Qwen3.5-9B scores 15/15 with the full cache; TOVA, SnapKV-$w32$, and UnionKV-v2 cross at 0.360%, 1.38%, and 3.16% of the cache, while NLL, the best literal-retrieval scorer, stays below half accuracy through 20%. On BABILong QA1 (full cache 18/20), SnapKV-$w32$ is above half accuracy already at 1%, UnionKV-v2 crosses at 7.07%, and NLL stays below half accuracy through 25%. On query-visible InfiniteBench key–value retrieval, SnapKV-$w32$ crosses at 3.35% on Qwen3.5-9B, while NLL and UnionKV-v2 stay below target through 10%. RULER QA2, BABILong QA2–QA3, and LongBench multi-hop QA (full-cache token F1 of 0.298–0.335) fail the full-cache validity gate and define no crossing. No single budget or scorer carries across these tasks. This is the situation in which the rank identity says a literal-retrieval count should not carry over: the evidence set S and its competitors C both change with the task.

6 Allocators must keep what binds the answer

The retention–generation gap has a direct consequence for allocator design: reserving slots for high-scoring content is not enough unless the allocator also keeps the rows that connect that content to the query. We test this with UnionKV, our allocator that combines content and attention scores (Appendix I). Its adaptive version, UnionKV-v2, first keeps designated opening and recent positions, following attention-sink practice (Xiao et al., 2024). It then estimates content demand from middle-

prompt NLL scores more than three standard deviations above the prompt mean, adds a 16-position margin, and fills the remaining slots with attention-selected rows.

On Qwen3.5-9B at 32K, NLL and UnionKV-v2 both answer 15/15 literal prompts with only 40 retained positions. On key–value binding prompts, which the full cache answers 15/15, UnionKV-v2 answers 0/15 at 1%, 5%, and 10% of the cache and 1/15 at 25%. Selecting contiguous groups solves the same prompts: at 5%, ChunkKV (Liu et al., 2025) with SnapKV-$w32$ scores 15/15, against 11/15 for SnapKV-$w32$ alone, at nearly identical realized cache lengths of 1,608 and 1,638 positions. All arms see the query.

Fixing the content reservation also repairs UnionKV. A count-first variant tuned on 20 development prompts reserves 32 content slots. On 15 held-out prompts, it keeps 15/15 literal answers at $N = 40$ and answers 11/15 binding prompts at 25%, against 1/15 for both UnionKV-v2 and a fixed-demand heuristic. It is the only tested UnionKV variant that holds both tasks: fixed-share UnionKV also reaches 11/15 on binding but answers 0/15 literal prompts at $N = 40$ (Appendix I.1).

Post-prefill row selection confirms that the right rows matter more than the number of rows (Appendix H.2). On 24 Qwen binding prompts answered by the full cache, attention-selected rows plus protected opening and recent positions answer 23/24 at 5%, while equal-count random rows answer 0/24. The attention set covers 99.7% of the annotated answer span, versus 5.3% for NLL selection, which answers 0/24. The same attention policy answers only 2/23 Phi-4 prompts and 3/24 Llama prompts that their full caches answer, so the rows a model needs are model-specific and must be calibrated per model. Native KVzip and AdaKV runs use their masked pipelines; because those masks leave the physical tensors allocated, Appendix J reports their answer quality and logical retention separately from memory.

7 Compaction turns retained counts into memory savings

Measured savings. We ran 20 paired 32K literal-retrieval requests on Qwen2.5-7B-Instruct-1M (dense attention) and Qwen3.5-9B (hybrid attention and recurrent state). At a logical budget of 256 positions, live cache storage falls from 1,792 to 17.5–18.4 MiB on the dense model (97–102×) and from 1,049.5 to 92.5 MiB on the hybrid (11×; Appendix Table 37). The full and compacted arms both answer all 20 requests correctly on both models. The measurement keeps the full prefill, the original positional coordinates, and every non-KV state tensor, and it gathers the retained entries into new tensors, so the savings are attributable to eviction alone. We report in-process Hugging Face cache bytes; native vLLM (Kwon et al., 2023) page-table eviction is a separate engineering step.

What a token count cannot reclaim. The hybrid result separates two floors. Its 25.5 MiB of recurrent and convolutional state is independent of the retained count, and compaction copies whole source blocks, so the hybrid keeps 2,144 physical prompt positions for 256 logical ones. For an ideal compacted cache with page size B, N retained prompt entries, and t generated entries,

$$M(N, t) = M_{\text{fixed}} + \sum_{\ell \in \mathcal{A}} b_\ell B \left\lceil \frac{N + t}{B} \right\rceil , \tag{8}$$

where b_ℓ is the KV bytes per token in attention layer ℓ and M_{fixed} includes recurrent state. Without compaction, scattered retained entries pin their original pages and a logical budget saves nothing. Compaction runs after prefill, so it reduces decode-time memory; peak prefill memory is unchanged.

Cost accounting. Unlike prompt-reconstruction emulation, which rebuilds a short prompt from retained tokens and so also removes prefill work, shifts positions, and truncates recurrent history, our protocol changes only the cache. We therefore draw no latency claims from reconstructed prompts. Appendix Table 38 reports synchronized selection, copy, and decode durations separately, and Appendix K.1 gives the decode-horizon page arithmetic.

8 Related work

Which tokens to keep. H2O and Scissorhands keep tokens that accumulate attention, and StreamingLLM keeps attention sinks and a recent window (Zhang et al., 2023; Liu et al., 2023;

Xiao et al., 2024). SnapKV scores the prompt with an observation window of recent queries; TOVA evicts the token with the lowest attention from the current query, averaged over a layer's heads; and Expected Attention scores each entry by its expected attention under the distribution of future queries (Li et al., 2024; Oren et al., 2024; Devoto et al., 2025). SirLLM and InfoKV select entries by token entropy or predictive information, close relatives of our NLL scorer, and IndexMem learns an eviction policy (Yao et al., 2024; Kai et al., 2026; Yang et al., 2026). KVzip scores a reusable, query-agnostic cache through context reconstruction (Kim et al., 2025). These methods improve the ranking. Our identity states what any fixed ranking implies for the budget and shows that the ranking's competitors, not the prompt length, set its size.

How to divide a budget. PyramidKV varies budgets across layers, AdaKV across heads, and FastGen profiles heads to assign each a policy (Cai et al., 2025; Feng et al., 2025; Ge et al., 2024). DuoAttention separates retrieval heads from streaming heads, and ChunkKV selects contiguous semantic chunks (Xiao et al., 2025; Liu et al., 2025). Quest retrieves query-relevant pages without evicting them (Tang et al., 2024). These methods take the total budget as given; we characterize what that total must be and when it transfers. We evaluate ChunkKV inside kvpress (Jégou et al., 2024) and FastGen through the community cold-compress implementation (Answer.AI, 2024).

How budgets are evaluated. Long-context benchmarks such as RULER, LongBench, BABILong, and InfiniteBench measure what a model can retrieve and reason over (Hsieh et al., 2024; Bai et al., 2024; Kuratov et al., 2024; Zhang et al., 2024). SCBench evaluates cache generation, compression, retrieval, and loading under reuse, and broader benchmarks compare compression families on accuracy and serving cost (Li et al., 2025; Yuan et al., 2024; Gao et al., 2025). These evaluations report compression at chosen fractions or counts. Our interventions supply the missing variable: they separate length from competing content, test whether a measured budget forecasts a new length, and check retained evidence, generated answers, and physical bytes on the same requests. Paged serving systems such as vLLM (Kwon et al., 2023) determine whether a logical budget becomes reclaimed memory, which Section 7 measures.

9 Conclusion

A KV-cache budget is set by the evidence a task needs and by the tokens that outrank it, not by the length of the prompt. This single identity explains why a fixed fraction spends 100 times the needed memory at 256K on literal retrieval yet falls sixteenfold short under dense distractors at 32K; why counts transfer across a sixteenfold length range where competition is stable; why short-context calibrations forecast long-context budgets within a factor of two on 22 of 25 testable cells; and why structured tasks need their own budgets. Retaining evidence is not answering, so allocators must keep the rows that bind a value to its query. Physical compaction converts small retained counts into 11–102$\times$ memory savings after full prefill.

Scope. The count results come from controlled literal retrieval with 15–20 paired prompts per cell, a design that makes exact rank audits possible; on natural tasks the evidence set is latent and must be measured. The retention identity applies to fixed, nested orderings, and window, chunk, and per-head policies have their own events. Intervals describe variation over the tested examples. Physical savings are in-process cache bytes on two models; native vLLM page-table eviction, end-to-end latency, and concurrent serving are the next engineering steps.

Practice. Report budgets as retained counts together with the scorer, the full-cache score, and the competitor load. Calibrate a count on the target scorer and model, stress it with added competitors rather than added length, and verify answers, not only retained evidence, before reusing it. Count memory savings only after physical compaction.

Reproducibility Statement

The accompanying artifact contains source code, exact configurations, model and data provenance, raw per-example records, environment specifications, analysis scripts, and build instructions for every table and figure. Appendix O maps the reported results to their source records.

AI ASSISTANCE

AI tools assisted with source auditing, experiment code, analysis, and manuscript drafting. The release records the resulting edits and their validation.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Answer.AI. Cold compress: A toolkit for benchmarking KV cache compression approaches. https://github.com/AnswerDotAI/cold-compress, 2024.

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

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

Alessio Devoto, Maximilian Jeblick, and Simon Jégou. Expected attention: KV cache compression by estimating attention from future queries distribution. *arXiv preprint arXiv:2510.00636*, 2025.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S. Kevin Zhou. Ada-KV: Optimizing KV cache eviction by adaptive budget allocation for efficient LLM inference. In *Advances in Neural Information Processing Systems*, 2025.

Wei Gao, Xinyu Zhou, Peng Sun, Tianwei Zhang, and Yonggang Wen. Rethinking key-value cache compression techniques for large language model serving. In *Proceedings of Machine Learning and Systems*, 2025.

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

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

GLM Team, Aohan Zeng, Xin Lv, Qinkai Zheng, et al. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling*, 2024.

Simon Jégou, Maximilian Jeblick, Alessio Devoto, et al. kvpress: LLM KV cache compression made easy. https://github.com/NVIDIA/kvpress, 2024.

Jushi Kai, Zhuiri Xiao, Alexandra Birch, and Zhouhan Lin. Information-aware KV cache compression for long reasoning. *arXiv preprint arXiv:2606.26875*, 2026.

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*, 2025.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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

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.

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.

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*, 2025.

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

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Mistral AI Team. Mistral NeMo. https://mistral.ai/news/mistral-nemo, 2024.

Robert G. Newcombe. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine*, 17(22):2635–2650, 1998.

Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. Transformers are multi-state RNNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18724–18741, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026.

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

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

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

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

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. Qwen2.5-1M technical report. *arXiv preprint arXiv:2501.15383*, 2025.

Xintong Yang, Hao Gu, Binxing Xu, Lujun Li, Bei Liu, Jiacheng Liu, Qiyuan Zhu, Yike Guo, and Sirui Han. IndexMem: Learned KV-cache eviction with latent memory for long-context LLM inference. In *International Conference on Machine Learning*, 2026.

Yao Yao, Zuchao Li, and Hai Zhao. SirLLM: Streaming infinite retentive LLM. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 2611–2624, 2024.

Jiayi Yuan, Hongyi Liu, Shaochen Zhong, Yu-Neng Chuang, Songchen Li, Guanchu Wang, Duy Le, Hongye Jin, Vipin Chaudhary, Zhaozhuo Xu, Zirui Liu, and Xia Hu. KV cache compression, but what must we give in return? A comprehensive benchmark of long context capable approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞Bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

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

A Formal scope of the rank statement

Let x be a prompt with L positions, and let $s_x(i)$ score position i. Break ties by a rule fixed before evaluation. Assume the scores do not change with budget, so the top-N retained sets $K_N(x)$ are nested. For a specified nonempty evidence set $S(x)$, define

$$R_S(x) = \max_{i \in S(x)} \operatorname{rank}_{s_x}(i).$$

Then, exactly,

$$S(x) \subseteq K_N(x) \quad \Longleftrightarrow \quad R_S(x) \le N.$$

This proposition concerns retention, not generation. If $Y_N \in \{0, 1\}$ denotes task correctness after retaining K_N, then

$$\Pr(Y_N = 1) = F_{R_S}(N) - \Pr(R_S \le N, Y_N = 0) + \Pr(R_S > N, Y_N = 1).$$

The two correction terms are, respectively, failures despite retaining the nominated set and successes after losing part of it. Equality with the rank CDF holds exactly when these two probabilities are equal. Sufficiency of the retained set together with necessity of every nominated member is one sufficient route to zero correction terms, but it is not necessary because nonzero terms can cancel. The experiments below measure both correction terms on literal retrieval.

Full-baseline normalization. The experimental estimator crosses at half the observed full-cache score, not at raw accuracy $1/2$. Let B denote correctness with the full cache and let $\alpha = \Pr(B) > 0$. Then

$$\frac{\Pr(Y_N = 1)}{\alpha} = \Pr(Y_N = 1 \mid B) + \frac{\Pr(Y_N = 1, B^c)}{\alpha}.$$

Only when the joint rescue term $\Pr(Y_N = 1, B^c)$ is zero does normalized accuracy equal accuracy conditional on B. Even then, a conditional rank-CDF interpretation additionally requires the retention–generation correction terms to be small or cancel within B, and it concerns $F_{R_S|B}$ rather than the unconditional rank distribution. Thus an unconditional median rank is not generally the half-full-baseline crossing; even $\alpha = 1$ leaves the behavioral correction terms. The registered analyses below keep all examples and the original half-full estimator; they do not condition on full-cache correctness.

The distinction matters in the E3 rerun. In a post-hoc paired diagnostic under the frozen 32-token horizon and answer lead-in, UnionKV-v2 at requested fractions 9.9% and 20% scores 14/15 versus a full-cache 11/15: 10 examples are correct in both arms, one only with full cache, four only after compression, and zero are incorrect in both arms. Normalized accuracy is therefore $14/11 = 1.273$, while accuracy conditional on full-cache correctness is $10/11 = 0.909$. At TOVA 0.076%, the compressed score is 12/15, with eight correct in both arms, three only with full cache, and four only after compression. The diagnostic leaves the registered crossing unchanged; it shows that the no-rescue condition fails in this cell. Analysis: `results/analysis/fresh_e3_full_baseline_rescue.`

`json` (SHA-256 `8421ba2f...`), generated by `src/analyze_full_baseline_rescue.py` (SHA-256 `e05348ba...`); the analyzer independently replays all 375 saved E3 scores.

If an example admits known alternative sufficient sets $\mathcal{S}(x)$, the first budget that contains at least one such set is summarized by $R_*(x) = \min_{S \in \mathcal{S}(x)} R_S(x)$. The family $\mathcal{S}(x)$ is operationally unknown on redundant, paraphrastic, and multi-hop tasks, and generation can still fail after retaining a set thought to be sufficient. A literal-span or minimum-copy audit is therefore a containment proxy, not an observed minimal sufficient set or an upper bound on the required budget.

For scale, even under independent identically distributed Bernoulli trials from a specified population and a policy fixed before sampling, 20 successes in 20 trials give the exact one-sided 95% Clopper–Pearson lower bound $0.05^{1/20} = 0.8609$, not 0.95. Fifty-nine all-success trials are required for that lower bound to exceed 0.95. Adaptive policy selection, multiple comparisons, or a shift from synthetic templates to deployment traffic requires additional accounting. Our template examples do not define a deployment population; the calculation illustrates the finite-sample limit.

For unique scores, $R_S = |S| + C_L$, where $C_L = \sum_{j \notin S} \mathbf{1}\{s_x(j) > T_S\}$ and $T_S = \min_{i \in S} s_x(i)$. In general, $\mathbb{E}[C_L \mid T_S]$ is a sum of position-specific exceedance probabilities; writing it as $(L - |S|)p_L$ requires equal conditional exceedance probabilities across background positions (conditional exchangeability is sufficient, but independence is unnecessary). Tightness, $C_L = O_p(1)$, keeps fixed interior rank quantiles bounded when evidence size is bounded; distributional stability across lengths makes them transferable. An extensive regime is $C_L/L \to \rho > 0$. For a stationary unnormalized scorer with fixed $p > 0$, $\mathbb{E}C_L = \Theta(L)$; a finite plateau may instead reflect $Lp \ll 1$ over the observed range. Softmax normalization alone does not stabilize ranks because common rescaling preserves order.

Window protection, chunk selection, per-head budgets, and layer-adaptive policies need separate treatment when their retained sets are not generated by one budget-independent total order. For such policies the exact event is the policy-specific vector-budget containment event; a rank computed from head/layer means is not equivalent to it.

Reading guide. Sections marked historical preserve measurements from earlier campaigns; their numbers can differ from the reruns reported in the main text, and their round IDs stay attached to the original predictions and verdicts. All main-text numbers come from the rerun sections, and all 174 registered rerun items are complete. Appendix B gives the execution protocol and calibration; length sweeps and historical forecasts follow. Appendix E.1 gives the distractor and filler interventions, and Appendix G.6 pairs retention with generation. Appendices H.2 and I.1 give row interventions and held-out allocation tests. Appendix K measures physical cache storage, and Appendix M reports task-specific boundaries.

Metric convention. Unless a subsection explicitly names ordered exact output, token F1, or an official task-specific metric, historical "accuracy" is normalized answer containment: the normalized generated response contains a normalized reference answer, which is weaker than equality with the full response. Rerun subsections state their outcome rule locally.

B Evaluation setup and calibration

Harness and eviction semantics. The historical grid uses prefill KV-compression hooks in kvpress: the full prompt is prefilled, and generation proceeds from the resulting compressed cache. Simple token scorers retain a top keep-fraction under one ordering. Windowed, adaptive, chunked, and per-head policies instead construct method-specific retained sets and need not be nested across budgets; their requested fraction is a logical budget, and the fixed-order proposition does not apply to them directly. Window methods are evaluated at *all tested* budgets, including budgets smaller than their observation window; that region is measured, not excluded.

KVzip and AdaKV are absent from that grid because their kvpress implementations perform *masked* eviction that takes effect only inside kvpress's own generation pipeline, which the raw-cache decode path does not use. Their returned cache lengths equal the full context, so a score from that path would measure the full context. Appendix J instead evaluates both methods through their native masked pipeline and reports logical and physical retention separately.

ExpectedAttention scores 0.00 below 1% and 0.13–0.47 at 25% in this harness, where every other tested method saturates. The cause, possibly an interaction between the partial-RoPE hook and its rotation-averaging step, is untested, so these rows are excluded from the historical fits of budget against length.

Historical model groups. The original fit set contains Qwen3.5-9B, Qwen2.5-7B-Instruct-1M, Qwen3.5-4B, Gemma-4-E4B, Phi-4, and GLM-4.7-Flash. Phi-4 is measured only at 8K and 16K. GLM entered the fit set after passing its round-1 holdout test; its historical grids use content scorers only because the bundled generic query reconstruction is incompatible. The held-out groups contain Mistral-Nemo in round 1 and Gemma-E2B, Qwen3.5-27B, Qwen3.5-MoE, and Llama-3.1-Instruct in round 2. Qwen3.5-2B, Phi-4-reasoning-plus, and Gemma-26B were subsequently added for 16K calibration only.

Execution status. All fourteen registered behavioral calibrations are complete. The GLM 16K rerun contains 1,095 outcomes and matches all 555 unaffected legacy control rows. The 32K, 64K, and 128K cohorts are also complete with 1,095 outcomes each and 15/15 full baselines. The GLM raw and manifest hashes begin `38640c64` and `661bf08b`; Table 2 gives every calibration disposition.

Execution environment. The historical grids used $n = 15$–25 per cell on eight NVIDIA B300 GPUs; an author-reported GPU failure changed some historical sample counts. Rerun cohorts use their complete registered example grids on the same eight B300 GPUs. The July fit constants come from an earlier software environment. The 2026-07-12 campaign used torch 2.11.0+cu130, transformers 5.8.0, and kvpress 0.4.0. Agreement across these campaigns is observed across environments, not a controlled software comparison.

MoE calibration. The native-query adapter for Qwen3.5-35B-A3B passed 60 bitwise layer/window checks and 15 actual generation arms before the registered 16K cohort ran. The complete cohort contains 1,095 outcomes, with full-cache accuracy 15/15. NLL crosses at 19.93 requested positions (16–24 rung; conditional 95% interval $[19.60, 21.10]$), v2 at 28.33 (24–32), InfoKV at 216.85 (192–256), and the learned scorer at 693.38 (512–1024). This is a one-length, within-family result. All 1,095 saved scores independently replay; raw and manifest hashes begin `befc7b0e` and `b6f2c1ca`. The frozen calibration summary is `results/analysis/qwen35_moe_completed_calibration_20260922.json`.

Attention backends. In the historical grid, $\leq$32K cells ran under PyTorch SDPA and $\geq$64K cells under FlashAttention-2: SDPA's math fallback materializes the masked attention matrix kvpress requires ($\sim$128 GiB at 64K), which OOMs. The historical Mistral-Nemo 16K calibration smoke was identical under SDPA and FlashAttention-2 with the same seeds; backend equivalence was checked on this model only. The original gemma-4 E-series runs were kernel-infeasible at $\geq$64K under this stack; gemma constants rest on 16–32K measurements plus rank audits.

Gemma-4 attention-selector fidelity. An audit against the native source shows that the bundled generic query reconstruction does not reproduce Gemma-4 attention: it omits the native query normalization and uses inverse-square-root head scaling where the native module uses scaling 1.0. On the 26B checkpoint it also infers eight KV heads from configuration although the global-attention cache has two. Historical Gemma rows for TOVA, SnapKV, PyramidKV, InfoKV, and UnionKV therefore reflect this unsupported reconstruction and are excluded from all native-policy evidence. Full-cache, NLL, and learned-content rows do not use that reconstruction.

A class-restricted native adapter passes its frozen actual-model gate on the exact E2B, E4B, and 26B revisions: 144 layer–window query/probability comparisons and complete prefill logits are bitwise equal, all six content-feature checks are finite, and all 21 generation arms satisfy replay, invocation, and cache-extent checks. This validates the adapter, not task accuracy. With this adapter, the E4B smoke test, E2B/E4B/26B 16K calibration, both E4B length cohorts, all five distractor-dose cohorts, and all three fixed-$M = 32$ filler executions are complete. The separate 32K filler repeat exactly matches its same-ID dose-primary outcomes; it contributes no additional independent prompts.

The E4B calibration contains 2,715 outcomes, a 15/15 full baseline, and exact agreement on all 555 unaffected legacy controls. Its NLL half-baseline crossing is 24.70 positions in the measured 24–32

bracket; native-adapter TOVA, SnapKV-$w32$, PyramidKV-$w64$, and UnionKV-v2 cross at 92.19, 134.65, 469.51, and 32.82, while SnapKV-$w1024$ is right-censored above 8,192. This calibration cell is distinct from the completed Gemma length and distractor interventions. E4B raw and manifest SHA-256 prefixes are `ae50806e` and `d29e9509`; gate completion, raw, and launch prefixes are `7a3d4014`, `d74d7723`, and `bed02341`.

gemma-4-E4B pressable layers. gemma-4 mixes attention types, and cache-owner topology differs by checkpoint. On E4B, 35 of 42 layers use sliding-window attention (window 512), whose KV caches are already truncated, so score-based eviction over the full prompt is undefined on them and the upstream kvpress Gemma convention excludes them from press hooks. The final 18 layers share KV with earlier layers and own no cache entry. E4B presses therefore target the four full-attention cache-owning layers, and reported E4B keep-fractions refer only to those caches; this four-layer count does not apply to E2B or 26B.

In the audited stock Transformers 5.8 execution, an owner layer publishes its cache tuple before its attention call returns. The post-forward press replaces cache attributes afterward, so later sharing layers read the full published K/V during that prefill. On the following decode step, the owner update starts from the compressed cache plus the new token and publishes that compressed state to sharing layers. This phase behavior is specific to this Transformers version. The runtime architecture audit checks the returned cache and one subsequent decode step.

B.1 Registered 16K calibration

Table 2 reports the two policies shared by the 14-model registry. All fourteen cohorts are complete: twelve contain 2,715 outcomes each, while GLM and Qwen3.5-MoE use their registered five-method grids of 1,095 outcomes each. All 34,770 saved scores reproduce from complete predictions and independently regenerated synthetic answers. Every cohort uses 15 examples; the full-cache score is 14/15 for Phi-4 and 15/15 for the other thirteen models. Each model keeps its registered methods and budget ladder; budgets are logical positions, not a common physical-memory budget.

The differences are substantial even within this fixed task. Llama has a 15/15 full baseline, yet unprotected NLL remains below half baseline through 8,192 positions; v2 crosses at 19.6 in the 16–24 bracket. This policy contrast recurs in its five-dose experiment in Appendix E.1. Qwen3.5-27B NLL crosses at 21.1 requested positions, whereas the Mistral-Nemo point estimate is 5,440.3. The latter comes from a nonmonotone curve, with a conditional bootstrap interval spanning 31.4–6,655.1; a single interpolated threshold would hide this instability. Phi-4 passes the observed 0.8 full-baseline gate, but 31 of its 2,000 paired resamples fail that gate, leaving 1,969 numerical crossings for each displayed policy. These comparisons characterize checkpoint–policy pairs on one prompt construction; they do not isolate architecture as the cause. The complete rung data, full-baseline intervals, and bootstrap state counts are in `results/analysis/calibration14.json`; `src/analyze_calibration14.py` regenerates the table from validated raw records.

Gemma-4-26B-A4B crosses at 227.1 requested positions for NLL and 216.0 for v2, both within the 192–256 measured bracket. Their conditional paired intervals are $[196.0, 287.4]$ and $[201.4, 226.7]$, with 2,000 numerical draws each. These requested budgets apply to the five pressed full-attention layers, not uniformly to all thirty cache-owning layers. At requested $N = 256$, both policies record a mean cache length of 13,696 across those thirty layers, consistent with twenty-five unchanged 16K caches and five 256-position caches. This unweighted mean is not a byte count, and a small requested crossing on this model is not whole-model memory compression. The complete calibration snapshot is `results/analysis/calibration14_completed_20260922.json`.

B.2 Qwen3.5-9B length cohorts

The analysis contains five complete Qwen3.5-9B lengths: 16,384, 32,768, 65,536, 131,072, and 262,144 tokens. Each length contains the same 15 paired examples per operating point, ten scorers, 18 compressed budgets plus full cache, and 2,715 outcome rows; every full-cache baseline is 15/15. The outcome at every point is the mean of 15 binary normalized-containment scores. Earlier two-endpoint and four-length snapshots are archived (paths below); the five-length statistic uses only the complete analysis.

Table 2: **Fresh registered 16K calibration boundaries.** Entries are N_{50} in requested retained positions with measured rung brackets. The second line is the conditional paired-bootstrap interval and number of numerical crossings among 2,000 resamples. Full baselines below the registered 0.8 gate remain explicit. Zero-width resampling intervals do not remove rung or population uncertainty. Realized cache-length summaries are retained in the JSON report.

Model	Full	NLL	UnionKV-v2
Qwen3.5-9B	15/15	28.0 [24,32] CI 27.0–29.5; 2000/2000	39.2 [32,48] CI 39.2–39.2; 2000/2000
Qwen2.5-7B-Instruct-1M	15/15	30.5 [24,32] CI 28.3–35.4; 2000/2000	39.2 [32,48] CI 39.2–39.2; 2000/2000
Qwen3.5-4B	15/15	26.8 [24,32] CI 24.4–28.9; 2000/2000	39.2 [32,48] CI 39.2–39.2; 2000/2000
Gemma-4-E4B	15/15	24.7 [24,32] CI 13.5–29.8; 2000/2000	32.8 [32,48] CI 22.4–37.3; 2000/2000
Phi-4	14/15	13.7 [8,16] CI 11.9–27.4; 1969/2000	21.3 [16,24] CI 19.9–24.4; 1969/2000
GLM-4.7-Flash	15/15	24.5 [24,32] CI 20.3–26.7; 2000/2000	27.4 [24,32] CI 26.3–27.7; 2000/2000
Mistral-Nemo	15/15	5440.3 [4096,6144] CI 31.4–6655.1; 2000/2000	76.4 [64,96] CI 65.6–88.0; 2000/2000
Gemma-4-E2B	15/15	20.0 [16,24] CI 18.2–22.2; 2000/2000	31.4 [24,32] CI 28.7–37.3; 2000/2000
Qwen3.5-27B	15/15	21.1 [16,24] CI 19.9–24.4; 2000/2000	30.5 [24,32] CI 28.7–35.4; 2000/2000
Qwen3.5-35B-A3B	15/15	19.9 [16,24] CI 19.6–21.1; 2000/2000	28.3 [24,32] CI 27.7–29.8; 2000/2000
Llama-3.1-8B-Instruct	15/15	> 8192 0/2000 crossed	19.6 [16,24] CI 19.6–19.6; 2000/2000
Qwen3.5-2B	15/15	24.4 [24,32] CI 21.1–26.7; 2000/2000	34.2 [32,48] CI 29.8–38.0; 2000/2000
Phi-4-reasoning-plus	15/15	11.6 [8,16] CI 11.3–12.3; 2000/2000	19.9 [16,24] CI 19.6–20.6; 2000/2000
Gemma-4-26B-A4B	15/15	227.1 [192,256] CI 196.0–287.4; 2000/2000	216.0 [192,256] CI 201.4–226.7; 2000/2000

Across all ten scorers, mean count-coordinate dispersion over the four adjacent registered-length pairs is smaller than fraction-coordinate dispersion. For NLL the means are 0.062 and 0.251; their paired difference is 0.189 with conditional percentile interval $[0.134, 0.230]$, defined in all 2,000 paired bootstrap resamples. Its five numerical crossings all lie in the measured 24–32 bracket. Resampling the 15 example IDs jointly across the full and compressed arms gives conditional crossing intervals $[27.0, 29.5]$, $[21.1, 26.7]$, $[25.2, 27.7]$, $[22.4, 27.1]$, and $[23.4, 27.4]$ in length order; all 2,000 resamples at every length have a valid baseline and a numerical crossing. The five-length log–log crossing slope is -0.011 with conditional interval $[-0.039, 0.009]$. These intervals describe the measured template distribution and do not adjust across scorers.

The frozen source for the five-length claims is `results/analysis/qwen35_completed_five_lengths_20260922T1002.json` (SHA-256 `01f4d0a4...`); it contains the exact raw, manifest, live-analysis, source-code, and figure hashes. The raw 16K, 32K, 64K, 128K, and 256K files have SHA-256 prefixes `e4d16c81`, `3a622ed7`, `a58328d0`, `7f13dd52`, and `b17af95a`. The frozen two-endpoint registry remains `results/analysis/qwen35_completed_endpoints_20260922T0823.json` (SHA-256 `e74a8bf9...`); exact copies of the provisional four-length analyses are `results/analysis/transfer_four_available_lengths_pre128k_20260922T0959.json` and `results/analysis/fresh_summary_four_available_lengths_pre128k_20260922T0959.json`.

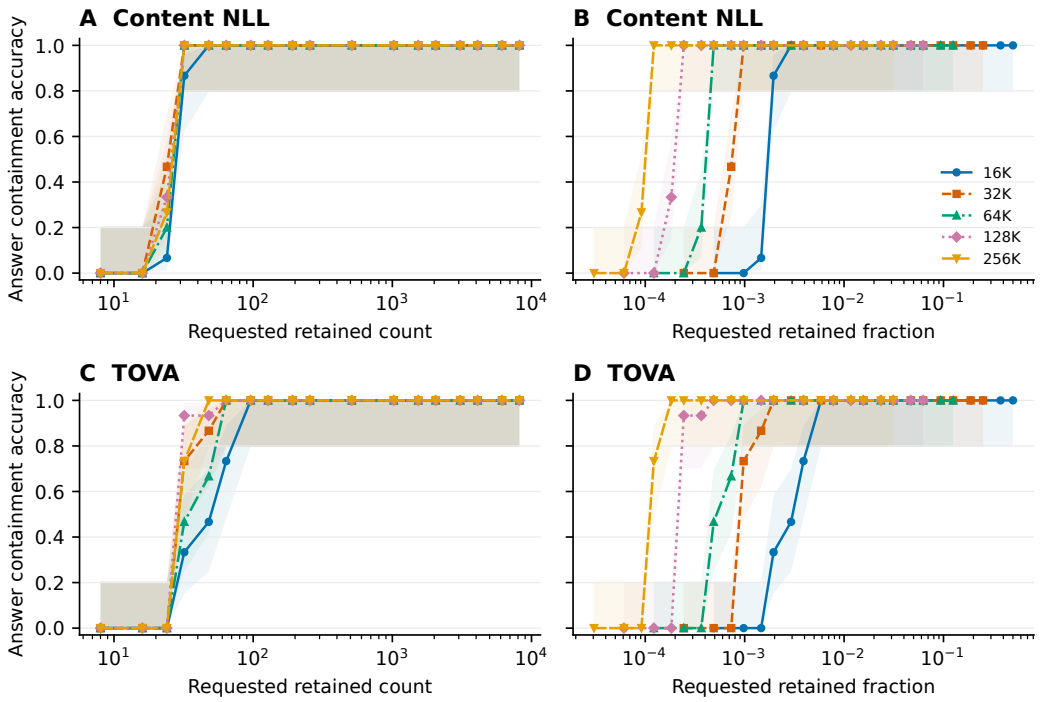

Figure 2: **Five completed Qwen3.5-9B length cohorts for NLL and TOVA.** Curves show the same binary containment outcomes against requested retained count (left) and requested fraction (right), with pointwise 95% Wilson intervals; color, marker, and line style redundantly encode length. At 16K, 32K, 64K, 128K, and 256K, NLL's half-accuracy crossings are 28.0, 24.4, 26.7, 25.8, and 26.3 positions; TOVA's are 49.8, 29.2, 34.2, 28.0, and 29.2. Requested counts are logical per-layer budgets, not physical allocated bytes.

The 256K run uses shared full-prefill capture and independent cache branches; 181 overlapping identities match the original loop's predictions and scores; 256K caches were not compared bitwise.

B.3 Dense Qwen2.5 five-length replication

The registered Qwen2.5-7B-Instruct-1M sweep is also complete at 16K, 32K, 64K, 128K, and 256K. Each length has 2,715 actual inference outcomes, the same 15 examples paired over operating points and lengths, and a 15/15 full-cache baseline. The separate 512K and 1M verification cohorts are complete with 12 and eight paired examples, respectively; neither is appended to this 15-example, ten-method estimand. Table 3 reports the NLL result; N_{50} is the first half-baseline crossing obtained by log-budget interpolation between measured rungs.

Table 3: **Dense-model NLL length sweep.** Counts are requested logical positions, not observed physical allocation. Intervals are conditional percentiles from 2,000 paired example bootstrap resamples; every resample has a valid baseline and a numerical crossing.

Length	N_{50}	measured bracket	conditional 95% interval
16K	30.50	24–32	[28.33, 35.41]
32K	31.43	24–32	[28.73, 36.41]
64K	34.24	32–48	[29.78, 37.99]
128K	31.43	24–32	[28.73, 36.41]
256K	29.78	24–32	[28.33, 34.24]

NLL's mean adjacent-pair dispersion is 0.031 in count coordinates and 0.221 in fraction coordinates. The paired difference is 0.190 with conditional interval $[0.161, 0.209]$, defined in all 2,000 resamples;

its log–log crossing slope is -0.0069 with interval $[-0.0404, 0.0217]$. Nine scorers have a defined count-versus-fraction comparison and all nine point estimates favor count, but only five intervals exclude zero. Several attention scorers have censored crossings, so we do not fit slopes after dropping those lengths. UnionKV-v2 crosses in the 32–48 rung at all five lengths, but its count-dispersion comparison is undefined because the observed step curves supply fewer than three transition-grid points. Its repeated interpolated value and zero-width empirical crossing intervals reflect identical measured steps and grid resolution, not an exact population threshold.

The frozen source is `results/analysis/qwen25_completed_five_lengths_20260922T1220.json` (SHA-256 `10955daa...`); it commits all five raw outputs and manifests, both analysis scripts, all ten scorer series, and all 50 per-length crossing records. This dense checkpoint extends the finite-range NLL plateau to a second architecture within the Qwen family, through 256K.

Separate 512K and 1M verification. The 512K cohort contains 444 outcomes and a 12/12 full-cache baseline. NLL crosses at 31 requested positions in the measured 21–31 bracket (conditional 95% interval $[27.2, 36.8]$); TOVA crosses at 128.0 in the 105–210 bracket, while SnapKV-w32 and InfoKV remain right-censored above 210. The nominal 105-position rung realizes 104 positions because the serialized fraction is floored by the selector; requested and realized counts are recorded separately.

At exactly 1,048,576 input tokens, all 200 outcomes complete, but the full-cache baseline is only 1/8 (Wilson 95% interval $[0.022, 0.471]$), below the registered 0.8 gate. NLL, TOVA, SnapKV-w32, InfoKV, and v2 score 0/8 at each tested count of 10, 31, 63, and 105. The learned scorer scores 4/8 at 63 and 105. With an invalid full baseline, the 1M cohort defines no N_{50}; the failure is specific to this task, prompt, and 32-token decoding protocol. All 644 saved scores independently replay under the legacy normalized-containment metric; the complete table and paired outcomes are in `results/analysis/qwen25_long_verification.json`.

B.4 QWEN3.5-4B FOUR-LENGTH REPLICATION

The third complete registered length sweep uses Qwen3.5-4B at 16K, 32K, 64K, and 128K. Each length contains 2,715 actual outcomes over the same 15 example IDs, and every full-cache baseline is 15/15. NLL crossings are 26.84, 25.79, 26.84, and 23.40 requested positions. The first three lie in the measured 24–32 bracket and the last in 16–24; their conditional paired 95% intervals are $[24.44, 28.93]$, $[22.43, 27.11]$, $[24.44, 28.88]$, and $[20.61, 26.84]$, with numerical crossings in every one of 2,000 resamples at each length.

Across the three adjacent registered length pairs, NLL dispersion is 0.047 in count coordinates and 0.250 in fraction coordinates. The paired difference is 0.203 with interval $[0.153, 0.245]$; its log–log crossing slope is -0.0536 with interval $[-0.1033, 0.0047]$. Both statistics are defined in all 2,000 paired resamples. Nine scorers have a defined count–fraction comparison on all three pairs, and each favors count with a conditional interval above zero. Adaptive v2 is a separate case: only the 64K–128K pair has enough transition points for the point comparison, and its interval conditions on 1,302 numerical resamples while 698 remain undefined. It therefore does not supply the same three-pair comparison.

The frozen registry `results/analysis/qwen35_4b_completed_four_lengths_20260922T1716.json` (SHA-256 `e33c86ec...`) commits all raw files, manifests, forty per-length crossings, and ten scorer-level transfer series. All 10,860 saved scores and all forty point crossings were independently replayed. These results add a smaller hybrid checkpoint within the Qwen family, through 128K.

B.5 REGISTERED LENGTH REPLICATIONS OUTSIDE QWEN

The Phi-4 sweep is complete at 8K and 16K, with 2,565 and 2,715 outcomes respectively and the same 15 example IDs. All 5,280 saved scores independently replay; the full-cache baselines are 15/15 and 14/15. NLL crossings are 19.60 and 13.72 requested positions, with measured brackets 16–24 and 8–16. At 8K every example changes from incorrect at 16 positions to correct at 24, yielding the same interpolated crossing in all 2,000 bootstrap draws; the zero-width interval reflects the rung grid,

not certainty. At 16K the curve is nonmonotone, and its conditional interval is $[11.93, 27.41]$, with 31 resamples failing the full-baseline gate.

All ten scorer comparisons favor count coordinates over fraction coordinates on this single length pair, with conditional intervals above zero. For NLL, however, the dispersion reduction is only 0.0128 with interval $[0.0052, 0.1093]$, and the descriptive two-point crossing slope is -0.515 with interval $[-0.716, 0.484]$. Both NLL statistics are numerical in 1,969 of 2,000 resamples. Curve alignment and a flat crossing are therefore separate properties. Table 4 reports all eight completed cohort identities and distinguishes methods absent from a registered grid.

The native-adapter Gemma-4-E4B length sweep contains 5,430 independently replayed outcomes, with 15/15 full-cache success at both 16K and 32K. NLL crosses at 24.7 and 35.4 requested positions, with measured brackets 24–32 and 32–48; v2 crosses at 32.8 and 41.6. For NLL, the fraction-minus-count dispersion difference is -0.0537, interval $[-0.1300, 0.0395]$, and the descriptive two-point slope is 0.520, interval $[0.270, 1.313]$. Both have 2,000 defined resamples. The crossing increases, while the dispersion comparison resolves neither coordinate as preferable.

Across Gemma's ten policies, five dispersion intervals are above zero; two additional positive point estimates have intervals containing zero, as do the three negative point estimates (NLL, the learned scorer, and v2). SnapKV-$w1024$ has no numerical point crossing at either length: its dispersion comparison is defined in 1,999 resamples, but only 110 give a numerical slope, so its slope is not estimable. The source is `results/analysis/gemma_completed_two_lengths_20260922.json`. These two-point, zero-distractor results stop at 32K; the twenty-example, fixed-$M = 32$ filler test is separate.

The complete GLM-4.7-Flash sweep tests 16K, 32K, 64K, and 128K with the same fifteen examples and a 15/15 full baseline at each length. NLL crossings are 24.5, 37.3, 34.2, and 27.7 requested positions; v2 gives 27.4, 26.3, 27.1, and 24.5. All four registered compressed policies have lower adjacent-pair dispersion in count coordinates, with paired intervals above zero. For NLL, the dispersion reduction is 0.0703, interval $[0.0352, 0.0935]$, and the descriptive log–log slope is 0.0412, interval $[-0.0535, 0.1627]$. V2's reduction is 0.1909, interval $[0.1307, 0.2124]$, while its slope is -0.0438, interval $[-0.0980, -0.0044]$. All these statistics have 2,000 defined resamples. Better count alignment thus coexists with small systematic crossing shifts. Several 128K curves are nonmonotone, so the first crossing remains descriptive. TOVA was not registered for this checkpoint.

The 128K execution preserves fifteen independent original examples, each with all 73 registered outcomes, and merges only after every shard passes its source and row-grid checks. The four-length snapshot is `results/analysis/glm_completed_four_lengths_20260922.json`. It contains 4,380 actual outcomes: 4,379 scores independently replay from complete original saved predictions, while one capped GLM32 score is corroborated by the separate full-output replay described below.

The analysis `src/analyze_other_length_replication.py` validates all eight registered cohort identities and recomputes crossings and paired uncertainty from raw outcomes. Scores are replayed from complete saved predictions, except for one GLM 32K InfoKV output (example 9, $N = 8$) that reaches the original 500-character logging cap. A later actual-model replay saves all 32 generated tokens and 747 decoded characters, matches the canonical prefix and score zero, and corroborates that recorded score. The report counts it separately from scores replayed from complete original strings. All eight cohort identities pass the analyzer's no-preview final mode: 15,089 original saved predictions independently replay, and one recorded score is separately corroborated, among 15,090 actual outcomes. A model contributes a transfer comparison only when its entire registered length set is complete and every full-cache baseline is valid; missing or invalid lengths are never dropped.

Logical positions vs. physical bytes. Reported keep-fractions count logical retained positions in the caches a press can act on, not KV bytes saved. The gap is widest on gemma-4-E4B, where four layers are pressable under the tested convention; E2B and 26B have different cache-owner topology. The byte measurements in Appendix K are architecture-specific for Qwen2.5-7B-Instruct-1M and Qwen3.5-9B. Count-based reporting should be accompanied by a per-architecture bytes audit before cross-model memory claims are made.

Table 4: **Registered length replications outside Qwen.** Entries are requested-position N_{50} values with measured rung brackets; the second line gives the conditional 95% paired-bootstrap interval and numerical crossings out of 2,000 resamples. Full baselines below 0.8 are invalid. Two-point sweeps describe only the tested factor-of-two range.

Model	Length	Full	NLL	TOVA	UnionKV-v2
Phi-4	8K	15/15	19.6 [16,24] CI 19.6–19.6; 2000/2000	36.3 [32,48] CI 30.5–41.0; 2000/2000	19.9 [16,24] CI 19.6–20.6; 2000/2000
	16K	14/15	13.7 [8,16] CI 11.9–27.4; 1969/2000	37.3 [32,48] CI 31.4–41.9; 1969/2000	21.3 [16,24] CI 19.9–24.4; 1969/2000
Gemma-4-E4B	16K	15/15	24.7 [24,32] CI 13.5–29.8; 2000/2000	92.2 [64,96] CI 70.8–109.3; 2000/2000	32.8 [32,48] CI 22.4–37.3; 2000/2000
	32K	15/15	35.4 [32,48] CI 30.1–38.0; 2000/2000	56.8 [48,64] CI 49.2–68.5; 2000/2000	41.6 [32,48] CI 38.4–48.9; 2000/2000
GLM-4.7-Flash	16K	15/15	24.5 [24,32] CI 20.3–26.7; 2000/2000	*not registered*	27.4 [24,32] CI 26.3–27.7; 2000/2000
	32K	15/15	37.3 [32,48] CI 21.4–38.6; 2000/2000	*not registered*	26.3 [24,32] CI 24.4–27.4; 2000/2000
	64K	15/15	34.2 [32,48] CI 28.7–37.3; 2000/2000	*not registered*	27.1 [24,32] CI 25.8–27.7; 2000/2000
	128K	15/15	27.7 [24,32] CI 22.2–34.2; 2000/2000	*not registered*	24.5 [24,32] CI 21.1–27.1; 2000/2000

C HISTORICAL COUNT-COORDINATE SUMMARIES AND RANK AUDITS

Table 5: **Historical constant-count pattern on the rounds-1 fit grid.** N_{50}: nominal requested count Lkf at which each method first reaches 50% of full-context accuracy (synthetic buried-needle `number_string`, position 0.5, log-interpolated crossing, n=15–25/cell). This coordinate is not an observed integer cache length. b: slope of $\log N_{50}$ vs. $\log L$; a constant fraction would give b=1. Bootstrap 95% CIs in the same nominal-count coordinate are released (`bootstrap_n50.md`); 56 of 225 CIs are kf-grid resolution-limited, and the Qwen2.5-1M window-method constants have 3–6×-wide CIs, so trailing digits are not significant.

Model	Method	16K	32K	64K	128K	256K	slope b
Qwen3.5-9B	NLL (content)	23.8	23.2	20.7	19.8	23.3	−0.029
	Cheap 4-feature	58.1	23.2	20.7	18.5	22.6	−0.305
	TOVA (w=1)	34.7	29.2	30.9	35.7	34.6	+0.028
	UnionKV-v1	51.8	47.5	48.9	45.0	82.9	+0.128
	SnapKV-w32	137.8	212.5	175.4	170.9	159.8	+0.011
	PyramidKV	252.7	342.2	222.4	389.4	—	+0.125
	SnapKV-w256	1199.4	997.4	835.3	—	—	−0.261
	InfoKV-style blend	171.1	154.7	125.5	105.5	87.2	−0.250
Qwen2.5-7B-1M	NLL (content)	26.3	27.6	28.1	41.4	30.8	+0.104
	Cheap 4-feature	51.8	46.3	42.5	41.4	38.6	−0.101
	TOVA (w=1)	97.4	48.9	50.5	47.7	73.9	−0.083
	UnionKV-v1	51.8	82.4	62.8	91.4	82.9	+0.151
	SnapKV-w32	919.5	707.4	1399.4	—	—	+0.303
	InfoKV-style blend	226.0	103.6	97.8	176.1	202.1	+0.044

Table 8 and the registry give the complete forecast accounting: 23 defined crossings, two testable no-crossing cells, and 14 unscoreable cells. Overview plots of these forecasts must read the registry's validity field and distinguish valid cells, such as the 512K measurements, from invalid ones, such as the 1M cells, which are excluded.

Additional historical observations. On InfiniteBench at 128K, the 5-digit `passkey` crosses 1.4–2× earlier than the 10-digit `number_string` for the tested content scorers (Qwen2.5-1M NLL: 43 vs. 61) and by larger factors for the tested window methods (SnapKV-w32: 83 vs. 490). At 16K, NLL crossings lie at 23–24 for the five measured Qwen3.5 sizes, while the gemma-4-26B-A4B-it

Table 6: **Historical count crossings on four further models (2026-07-12 campaign).** phi-4 is context-capped at 16K (measured 8–16K); gemma-4 E-series is kernel-infeasible at ≥64K; GLM runs the content scorer only. The archive records GLM and Qwen3.5-4B cells at ≥64K as round-2 predictions measured after its internal freeze. Two-point slopes for gemma and phi-4 are uninformative about length invariance; Gemma attention-policy entries use the unsupported query reconstruction and are excluded from native-policy evidence.

Model	Method	8K	16K	32K	64K	128K	slope b
Qwen3.5-4B	NLL (content)	—	23.5	23.2	21.4	21.4	−0.052
	Cheap 4-feature	—	44.8	23.2	21.4	20.7	−0.346
	TOVA (w=1)	—	51.8	39.4	44.8	39.7	−0.097
	UnionKV-v2	—	51.8	45.0	43.9	40.9	−0.106
	SnapKV-w32	—	168.3	134.1	131.1	137.7	−0.090
	InfoKV-style blend	—	120.7	105.6	—	—	−0.193
gemma-4-E4B-it	NLL (content)	—	22.8	40.0	kernel-		+0.811
	Cheap 4-feature	—	16.4	29.1	infeasible		+0.827
	TOVA (w=1)	—	126.0	90.7			−0.474
	UnionKV-v2	—	38.2	44.3			+0.214
	SnapKV-w32	—	148.9	76.3			−0.965
	InfoKV-style blend	—	143.3	55.5			−1.368
phi-4	NLL (content)	25.4	16.8	16K context cap			−0.596
	Cheap 4-feature	19.1	21.3				+0.157
	TOVA (w=1)	31.0	42.8				+0.465
	UnionKV-v2	25.9	23.1				−0.165
	SnapKV-w32	60.7	58.4				−0.056
	InfoKV-style blend	57.9	51.7				−0.163
GLM-4.7-Flash	NLL (content)	—	22.4	40.9	36.8	28.5	+0.089
	Cheap 4-feature	—	31.8	16.5	20.7	22.2	−0.123

crossing is about ten times larger (238.8), so this within-family alignment does not carry across families. Phi-4-reasoning-plus and base phi-4 have similar measured crossings (16.4 vs. 16.8). Window pooling is associated with deeper audited ranks in this grid, and position changes little on the synthetic position sweep (SnapKV-w32 worst-span rank 96–106 over fractional positions 0.1–0.9). These are associations in the measured cells.

Table 7: **Rank audit (round 1).** Unconditional median rank of the worst answer-span token under each scorer (128K, 50 synthetic examples per model) versus the measured half-full-baseline N_{50} range across L. Their proximity is an empirical heuristic, not the identity in Appendix A; imperfect full baselines and retention–generation mismatches separate the two quantities. The rank uses a token-mean approximation; SnapKV additionally evicts per KV-head with pooling, which the approximation does not capture—its measured constants sit above the audited rank by the pooling gap.

Model	Method	Median worst-span rank	Measured N_{50}	Gap
Qwen3.5-9B	NLL	26.0	20–24	≈ 1×
Qwen3.5-9B	TOVA	33.0	29–36	≈ 1×
Qwen3.5-9B	SnapKV-w32	105.0	138–213	1.3–2.0×
Qwen3.5-9B	SnapKV-w256	456.5	835–1199	1.8–2.6×
Qwen2.5-1M	NLL	31.0	26–41	≈ 1×
Qwen2.5-1M	TOVA	43.0	48–97	1.1–2.3×
Qwen2.5-1M	SnapKV-w32	153.5	707–1399	4.6–9.1×

The figure filters exact model identities before taking the median across available lengths, so no non-Qwen3.5 checkpoint enters the Qwen2.5 crossing map. Qwen2.5 NLL, TOVA, and SnapKV-w32 medians are 28.1, 50.5, and 919.5 positions; SnapKV-w256 has no Qwen2.5 crossing and is omitted.

Round-2 audits: eight new models. The 2026-07-12 campaign repeats the literal-span audit on every new model (36 rows each: 12 examples × 3 needle positions, at 16 or 32K). Unconditional

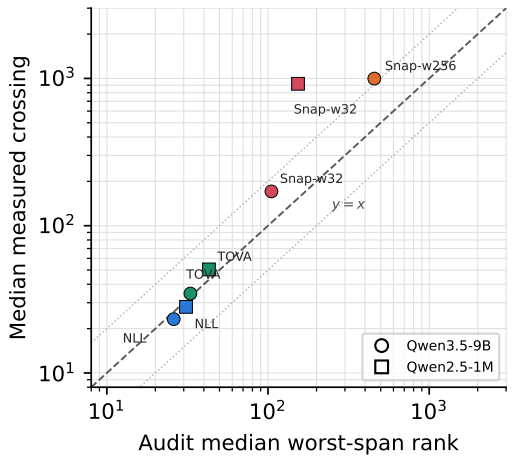

Figure 3: **Historical association between a literal-span rank proxy and measured crossing.** The proxy is the unconditional median, over ten base examples crossed with five needle positions, of the worst answer-span-token rank in a token-mean ordering at 128K; the measured quantity is the median of the available historical half-full-baseline crossings across lengths for that model and scorer. Circles are Qwen3.5-9B and squares Qwen2.5-1M. Content and point-query cells lie near the identity line in this audit; pooled per-head SnapKV lies above it. The proxy is not conditioned on full-cache correctness and does not reproduce the vector-budget policy, so the dotted $2\times$ band is descriptive.

median worst-span NLL ranks range from 14 to 41. Several half-full-baseline crossings are numerically close to these medians (gemma-4-E4B: 41 vs. 40.0; phi-4: 17 vs. 16.8; Qwen3.5-35B-A3B: 23 vs. 23), while SnapKV retains a substantial token-mean/per-route gap (Qwen3.5-27B: 123.5 vs. 232–276). Llama NLL is the counterexample: audit median 14 but no measurable crossing in the tested ladder. The median literal-span rank thus predicts the crossing in some cells but does not determine it; with an imperfect full baseline it also differs from the conditional rank quantity that normalized accuracy requires.

D HISTORICAL LENGTH FORECASTS AND HELD-OUT CALIBRATION

Disposition of every committed point prediction. Table 8 accounts for all 39 point predictions of rounds 1–2: 23 cells with defined crossings, two testable cells with no crossing, and 14 unscoreable cells. The reasons are mechanical: the 1M full-context reference falls below the registered validity gate under this task and harness, some keep-fraction ladders stop before a required crossing, and gemma-4 E-series cells at $\geq$64K are kernel-infeasible in this stack. Among the 23 defined crossings, 22 lie inside the frozen numerical band and one lies outside. Counting the two testable no-crossing cells as misses gives 22/25 under strict accounting; all three misses are window or point-query scorers, two on Llama-3.1. A post-hoc reanalysis of the same 23 crossed cells compares the frozen 16K count with two alternatives. The frozen count has mean $|e| = 0.274$ (median 0.152, maximum 1.634; 22/23 within a factor of two). A constant fraction fitted to the same 16K calibration has mean $|e| = 2.698$ (median 2.388, maximum 5.530) and places 0/23 within a factor of two. A leave-one-out power law fitted to the measured multi-length data has mean $|e| = 0.400$ on its 21 scoreable cells. The one-constant count therefore matches a curve fitted to the measured lengths, and the constant fraction never lands inside its own band. Source: `provenance/historical_results/reanalysis_20260713/REANALYSIS_SUMMARY.md`.

Round-1 notes. The 512K NLL crossing is 6.0×10^{-5} (31 tokens), versus a registered point of 5.3×10^{-5} fit at $\leq$256K. The 1M grid completed (n=12/cell), but uncompressed accuracy is 0.12–0.50, below the registered 0.8 validity floor; those cells are invalid for compression scoring under this task and harness. GLM-4.7-Flash moves from a 16K calibration crossing of 21 tokens to 29 at 128K, inside its band. Mistral-Nemo NLL is non-monotone at 16K and never crosses within

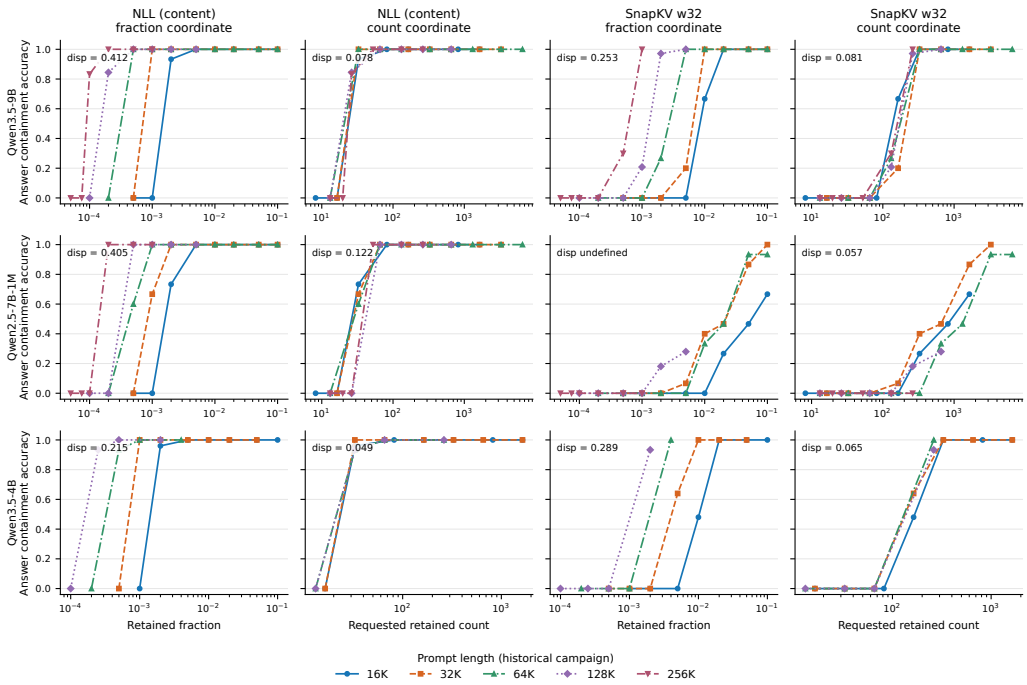

Figure 4: **Historical curve-collapse figure.** For coordinate $x \in \{\mathrm{kf}, \mathrm{kf}L\}$, each accuracy curve is linearly interpolated in $\log_{10} x$ on 41 equally spaced points over the common support. Points are retained when at least one curve lies strictly between 0.05 and 0.95; "disp" is the mean population standard deviation across lengths on those points and is undefined with fewer than three eligible points. The primary adjacent-pair statistic applies the same function to consecutive lengths and averages defined pairs; the all-length value shown here uses all lengths simultaneously and therefore has different support. The renderer `src/plot_historical_collapse.py` reuses the frozen estimator and adds marker and line-style encodings for length.

that historical ladder; UnionKV-v2 crosses while SnapKV-w32 scores 0.00 through the tested 1% budget.

Round-2 notes. The full 64K sweep passes 11/11 across five models in three families and four scorer classes, implied counts within $\times 1.4$ of frozen constants everywhere. The MoE passes at both lengths (implied 21/29 vs. frozen 23), a within-family architecture transfer. The single numerical-band failure is Llama SnapKV-w32 at 128K (implied 161, bootstrap CI $[125, 237]$, vs. frozen 52). Under strict accounting, Llama TOVA at 128K is a further no-crossing miss after testing to $11.4\times$ its predicted point. Together with Qwen2.5-1M TOVA at 512K, the strict record is 22/25 testable cells; the continuous error summary is conditional on the 23 cells with defined crossings. The three strict misses are attention scorers: one window-pooled SnapKV cell and two point-query TOVA cells. The frozen gemma-4-E2B predictions at $\geq$64K could not run because of kernel infeasibility and remain unscoreable.

Calibration detail. Rank-audit predictions (P-R2-2: unconditional 16–32K audit medians predicting 64–128K half-full-baseline crossings) land 17/19 inside registered bands, with mean $|e| = 0.58$. The audit median is an empirical predictor, not the rank identity. The two outside are Llama NLL at 128K ($9.4\times$ its audited rank—the registered special case "remains undefined or crosses only $\geq 4\times$ rank," and therefore a pass by that rule) and GLM NLL at 64K (a marginal $2.04\times$ miss).

Among four candidate scaling forms fit to all 30 multi-L series (constant, cL, cL^{β}, and $c + d \log L$), the constant wins AICc in 28/30 and leave-one-out error in 18/30; the power and log-additive forms win seven and five leave-one-out comparisons. Constant fraction wins no series under either criterion. Where the power form wins leave-one-out error, it mostly implies *declining* counts (learned-predictor

Table 8: **All 39 committed point predictions, by disposition** (rounds 1–2 pooled; produced by the frozen scorer, per-cell rows in `calibration.md`). The first two rows are the 23 crossed cells used by the historical continuous-error summary. Bound predictions (not point predictions) are scored separately: 3 PASS, 2 invalid cells at 1M, 1 insufficient.

Disposition	n	Cells
Scored, inside the frozen $\times 2$ band	22	Qwen3.5 (11), Llama-3.1 (6), GLM-4.7 (3), Qwen2.5-1M (2)
Scored, outside the band (failure)	1	Llama-3.1 SnapKV-$w32$ @128K (161 measured vs. 52 predicted)
No crossing at $\geq 2\times$ the predicted point	2	Qwen2.5-1M TOVA @512K (tested to $2.2\times$); Llama-3.1 TOVA @128K ($11.4\times$)
(counted as misses in the strict accounting: 22 of 25)		
Unscoreable: full-context reference invalid	2	Qwen2.5-1M NLL and cheap-predictor @1M (uncompressed 0.36)
Unscoreable: cell never ran	10	gemma-4-E2B (8) and E4B (2) at 64–128K, kernel-infeasible
Unscoreable: ladder coverage stops below the predicted point	2	Qwen3.5-27B SnapKV-$w32$ @128K; Mistral-Nemo NLL (its 16K calibration crossing is undefined)
Total committed	39	

Table 9: **Round-1 author-recorded freeze scoreboard.** The archive records predictions frozen at 2026-07-07T10:20:03Z (SHA-256 prefix `c7998dc3`) before its verification runs; fit set 16K–256K only; predictions are $2\times$ bands around $\mathrm{kf}^* = N^*/L$ with frozen constants; frozen estimator: 50%-of-full crossing by log-x interpolation, median over examples, $n \geq 10$ per cell. Verification: Qwen2.5-7B-1M, synthetic `number_string`.

ID	Frozen prediction	Measured	Verdict
P1a	NLL crosses in $[2.7, 10.7]\times 10^{-5}$ @512K (pt. 5.3×10^{-5})	6.0×10^{-5}	**PASS**
P1b	Cheap pred. crosses in $[3.9, 15.6]\times 10^{-5}$ @512K	5.4×10^{-5}	**PASS**
P1c	TOVA crosses in $[4.5, 18]\times 10^{-5}$ @512K	no crossing (0.25 @ 2×10^{-4})	**FAIL**
P1d	SnapKV-w32 ≤ 0.10 at all kf $\leq 2 \times 10^{-4}$ @512K	0.00 everywhere	**PASS**
P2a	NLL crosses in $[1.3, 5.3]\times 10^{-5}$ @1M	unscoreable: uncompressed accuracy 0.12–0.50 $<$ 0.8 validity floor at 1M (invalid full-cache baseline)	n/a
P2b	Cheap pred. crosses in $[2.0, 7.8]\times 10^{-5}$ @1M		
P2c	SnapKV-w32 ≤ 0.10 at all kf $\leq 10^{-4}$ @1M		
P2d	pooled $\lvert d\log N_{50}/d\log L\rvert \leq 0.25$ for NLL, 16K→1M		
P3a	GLM 128K content cliff within $[N_{50}^{16\mathrm{K}}/2,\ 2N_{50}^{16\mathrm{K}}]$	$21 \to 29 \in [10.5, 42]$	**PASS**
P3b	GLM SnapKV-w32 $\geq 4\times$ NLL count, or undefined	attention presses incompatible	excl.
P4a	Mistral 128K content cliff within calibration band	16K crossing undefined (peak 0.47)	n/a
P4b	Mistral SnapKV-w32 $\geq 4\times$ NLL, or undefined	SnapKV 0.00 at all kf $\leq 10^{-2}$	n/a
P5a	UnionKV-v2 within 0.05 of best single method, $\geq$6/8 cells	7/8 (miss at 26-token budget)	**PASS**
P5b	UnionKV-v2 crosses by kf $\leq 1.7 \times 10^{-4}$ @512K	7.7×10^{-5}	**PASS**
	…and by kf $\leq 8.6 \times 10^{-5}$ @1M	unscoreable (P2 validity rule)	n/a

β from -0.31 to -0.75), opposite in direction to constant-fraction scaling. Window-side exceptions remain: Llama SnapKV-w32 grows from 52 to 161 ($\beta \approx 0.46$), Qwen2.5-1M SnapKV-w32 is noisy and possibly growing (920/707/1399, with 3–6$\times$-wide intervals), and Qwen3.5 SnapKV-w32 is flat over the tested grid.

The released uncertainty uses 2,000 percentile resamples per cell. Of 225 defined-N_{50} intervals, 56 are zero-width under example resampling; these are resolution-limited estimates, not exact thresholds. The widest cells are retained and named (Qwen2.5-1M PyramidKV at 32K: $N_{50} = 3093$ $[512, 3119]$). Mean $|e|$ is 0.376 in round 1 and 0.258 in round 2; by family it is 0.223 for Qwen3.5 (11/11 in band), 0.265 for GLM, 0.350 for Qwen2.5-1M, and 0.336 for Llama (6/7).

Sampling intervals versus ladder resolution. The archived 291-cell bootstrap registry has 225 defined N_{50} values: 217 log-interpolated crossings and eight left-censored upper bounds. Of 56 zero-width bootstrap intervals, 50 still lie between distinct nominal requested-count rungs and six are left-censored at the smallest rung. The released rung audit therefore keeps each historical CI

Table 10: **Round-2 author-recorded freeze scoreboard** (verbatim from the mechanical scorer; recorded freeze 2026-07-12T10:08:44Z, SHA-256 prefix `62690dc9`). Archive timestamps place the 16K calibrations and 16–32K audits before the 64–128K verification files. "impl." is the measured crossing re-expressed as a count; "insuff.": kf coverage at scoring time below what the frozen estimator requires—reported as unscored, not as passes.

Model	L	Method	Pred. kf* ($\times 2$ band)	Measured	N_{50} impl./frozen	Verdict
Qwen3.5-27B	64K	NLL	3.51×10^{-4} $[1.8, 7.0]\times10^{-4}$	3.16×10^{-4}	21 / 23	**PASS**
Qwen3.5-27B	64K	SnapKV-w32	3.56×10^{-3} $[1.8, 7.1]\times10^{-3}$	4.21×10^{-3}	276 / 233	**PASS**
Qwen3.5-27B	64K	TOVA	4.12×10^{-4} $[2.1, 8.2]\times10^{-4}$	3.29×10^{-4}	22 / 27	**PASS**
Qwen3.5-27B	64K	UnionKV-v2	4.12×10^{-4} $[2.1, 8.2]\times10^{-4}$	3.65×10^{-4}	24 / 27	**PASS**
Qwen3.5-27B	128K	NLL	1.75×10^{-4} $[0.88, 3.5]\times10^{-4}$	1.58×10^{-4}	21 / 23	**PASS**
Qwen3.5-27B	128K	SnapKV-w32	1.78×10^{-3} $[0.89, 3.6]\times10^{-3}$	—	—	insuff.
Qwen3.5-27B	128K	TOVA	2.06×10^{-4} $[1.0, 4.1]\times10^{-4}$	1.99×10^{-4}	26 / 27	**PASS**
Qwen3.5-27B	128K	UnionKV-v2	2.06×10^{-4} $[1.0, 4.1]\times10^{-4}$	1.99×10^{-4}	26 / 27	**PASS**
Qwen3.5-35B-A3B	64K	NLL	3.51×10^{-4} $[1.8, 7.0]\times10^{-4}$	3.16×10^{-4}	21 / 23	**PASS**
Qwen3.5-35B-A3B	128K	NLL	1.75×10^{-4} $[0.88, 3.5]\times10^{-4}$	2.24×10^{-4}	29 / 23	**PASS**
Qwen3.5-4B	64K	NLL	4.27×10^{-4} $[2.1, 8.5]\times10^{-4}$	3.27×10^{-4}	21 / 28	**PASS**
Qwen3.5-4B	128K	NLL	2.14×10^{-4} $[1.1, 4.3]\times10^{-4}$	1.63×10^{-4}	21 / 28	**PASS**
Llama-3.1-8B	64K	Cheap pred.	3.36×10^{-4} $[1.7, 6.7]\times10^{-4}$	3.06×10^{-4}	20 / 22	**PASS**
Llama-3.1-8B	64K	SnapKV-w32	7.93×10^{-4} $[4.0, 15.9]\times10^{-4}$	7.46×10^{-4}	49 / 52	**PASS**
Llama-3.1-8B	64K	TOVA	3.51×10^{-4} $[1.8, 7.0]\times10^{-4}$	3.27×10^{-4}	21 / 23	**PASS**
Llama-3.1-8B	64K	UnionKV-v2	3.51×10^{-4} $[1.8, 7.0]\times10^{-4}$	3.16×10^{-4}	21 / 23	**PASS**
Llama-3.1-8B	128K	Cheap pred.	1.68×10^{-4} $[0.84, 3.4]\times10^{-4}$	1.58×10^{-4}	21 / 22	**PASS**
Llama-3.1-8B	128K	SnapKV-w32	3.97×10^{-4} $[2.0, 7.9]\times10^{-4}$	1.23×10^{-3}	161 / 52	**FAIL**
Llama-3.1-8B	128K	TOVA	1.75×10^{-4} $[0.88, 3.5]\times10^{-4}$	—	—	insuff.
Llama-3.1-8B	128K	UnionKV-v2	1.75×10^{-4} $[0.88, 3.5]\times10^{-4}$	1.58×10^{-4}	21 / 23	**PASS**
GLM-4.7-Flash	64K	NLL	4.58×10^{-4} $[2.3, 9.2]\times10^{-4}$	5.61×10^{-4}	37 / 30	**PASS**
GLM-4.7-Flash	128K	NLL	2.29×10^{-4} $[1.1, 4.6]\times10^{-4}$	1.99×10^{-4}	26 / 30	**PASS**

Total: 19 PASS, 1 FAIL, 2 insufficient (Qwen3.5-27B SnapKV-w32 at 128K; Llama TOVA at 128K).

unchanged and adds the adjacent Lkf bracket or upper bound; these fractional endpoints are not observed retained-entry counts. Raw `mean_cache_len` supplies a separate realized cache-length summary at every reported rung for all 225 registry cells and for 21/23 numerical prospective cells; the two 512K P1 cells store null cache lengths and retain nominal brackets only. Raw `method=none` rows reproduce the frozen full-baseline counts and means for all 225 cells and the full means for all 23 prospective cells. Two interrupted null-byte lines are recorded and excluded explicitly; subsequent valid rows still match every frozen per-rung sample count. The machine-readable audit is `results/historical_registry/historical_rung_brackets.json`; it uses existing records only.

Commitment ordering. Freeze times, hashes, and the ordering of freezes before verification runs are author-recorded; we refer to all rounds as author-recorded freezes. The 20 scored round-2 cells share a task template, estimator, and tokenizer families, so they are dependent rather than 20 independent replications.

E Distractor and filler interventions

Design and verdicts. The archive contains `PREREGISTRATION_OUTLIER.md`, SHA-256 `69481cdb...`, plus harness hash `87dfe93f...`, and records an internal freeze at 2026-07-13T00:57:21Z before its first cell. Task `number_string_decoys`: identical to `number_string` (needle byte-identical per example; $M{=}0$ token-identical to the original task, verified) plus M injected distractor statements with distinct non-matching key phrases, random 10-digit values, uniform random positions. Models: Qwen3.5-9B, Llama-3.1-8B, gemma-4-E4B; scorers: NLL, learned content predictor (`cheap_pred`), SnapKV-w32, UnionKV-v2 (+ uncompressed). P1/P3: dose-response, $M \in \{0, 8, 32, 128, 512\}$ at fixed $L{=}32768$; P2: filler control, fixed $M{=}32$, $L \in \{8\text{K}, 16\text{K}, 32\text{K}, 64\text{K}\}$ (gemma capped at 32K). All 23 registered jobs match the frozen design ($n{=}20$/cell, zero error rows, no exclusions); uncompressed accuracy is 1.00 on every cell even at $M{=}512$—the distractors confuse only the scorers. Estimator unchanged from round 2. The archived harness gives mechanical verdicts **P1 FAIL** (0/3, directional-high), **P2 PASS** (2/3: Qwen pass, gemma pass, Llama learned predictor at the exact 2.0 boundary), and **P3 FAIL** (1/3: Llama

Table 11: **Round-3 dose-response: N_{50} (tokens) vs. injected distractor count M at L=32768** (n=20/cell; full-context accuracy 1.00 everywhere). "$>X$": no 50% crossing within the registered kf ladder (X = keep-count at the largest measured budget); the frozen coverage rule scores these as failures and reports them as lower bounds. M=0 reproduces the filler-only constants. Llama NLL is its known noise-floor case (no crossing at any M; its content scorer here is the cheap predictor, as registered).

Model	Scorer	M=0	M=8	M=32	M=128	M=512
Qwen3.5-9B	NLL	23.2	103.6	455.0	1261.0	>3277
	Cheap pred.	23.2	529.5	>1638	>1638	>3277
	SnapKV-w32	199.7	824.1	1073.4	978.5	1073.4
	UnionKV-v2	36.8	103.6	463.4	471.9	481.6
Llama-3.1-8B	NLL	no crossing at any M (registered noise-floor case)				
	Cheap pred.	20.0	126.1	611.5	>1638	>3277
	SnapKV-w32	85.1	192.3	231.7	185.8	113.6
	UnionKV-v2	23.2	103.6	231.7	1036.2	2317.0
gemma-4-E4B	NLL	39.9	120.7	435.9	1061.5	>3277
	Cheap pred.	29.7	>1638	>1638	>1638	>3277
	SnapKV-w32	65.5	>1638	>1638	>1638	>3277
	UnionKV-v2	43.6	126.1	445.9	1036.2	1162.0

pass at 1.33×; Qwen fail at 5.4× saturating; gemma fail, collapse). Gemma attention-policy verdicts use the unsupported reconstruction; native-adapter Gemma results follow in Appendix E.1.

P2 filler-control values. At fixed M=32: Qwen NLL $481/494/455/455$ across 8K–64K ($\max/\min = 1.09$); gemma NLL $436/492/436$ across 8K–32K (1.13); Llama learned predictor $1227/819/611/875$ (2.01, non-monotone—the same noisy content scorer flagged in round 2).

Secondary findings. The learned predictor is highly distractor-sensitive in these cells: it loses its crossing at $M = 32$ on Qwen and at $M = 8$ on gemma. UnionKV-v2 remains near 470 tokens on Qwen from $M = 32$ through 512, while pure NLL exceeds the largest 3277-token rung; the unsupported Gemma UnionKV row reaches 1162 at $M = 512$; on Llama the crossing grows from 23 to 2317. Because v2 uses an adaptive outlier-demand branch rather than a fixed 50/50 split, its plateau is not a "content half" effect. The optional $M = 512$ ladder extension was not run, leaving those cells as lower bounds; coarse rungs limit crossing precision even though the registered directional failures are far outside their bands.

E.1 Competitor-dose and filler replication

The Qwen3.5-9B dose sweep tests $M \in \{0, 8, 32, 128, 512\}$ at 32K with 20 paired examples, four compressed policies, and 20 budgets per policy. Each dose contains 1,620 outcomes; every matched full-cache baseline succeeds on 20/20. The separate filler sweeps hold $M = 32$ and use 8K, 16K, and 64K; the existing 32K dose cohort supplies the fourth length without being counted twice. Full-cache scores at the three separate lengths are 20/20, 19/20, and 20/20. Every crossing uses half its own matched full-cache baseline. The budget ladder is wider than in the historical sweep above.

NLL's four signed adjacent secants range from 10.05 to 12.66 tokens per added distractor. Their sampling intervals and their measured-rung bounds answer different questions: resampling captures variation across the twenty prompts, whereas the rungs limit where a crossing was directly observed. The four contrasts reuse the same examples and are not independent model replications. SnapKV's measured crossings are nonmonotone over the dose grid, and the learned predictor remains right-censored at the two largest doses. Displacement rates differ across scorers.

For filler expansion, the NLL endpoint crossing ratio is 1.00, while adaptive UnionKV-v2 grows by 2.38× and the learned predictor falls to 0.56×. A policy that uses NLL features therefore need not inherit the NLL plateau: adaptive allocation, protected positions, and final candidate selection make v2 a different policy from a fixed NLL ordering. The experiment does not isolate which of these choices causes v2's shift, and it uses one distractor template.

Table 12: **Fresh Qwen3.5-9B dose response at 32K.** Entries are log-interpolated half-baseline crossings, with measured rung brackets in parentheses. Bounds remain censored. Each dose has 20 paired examples and full-cache accuracy 20/20.

Scorer	$M = 0$	$M = 8$	$M = 32$	$M = 128$	$M = 512$
NLL	24.0 (16–24)	124.0 (96–128)	427.7 (384–512)	1416.4 (1024–1536)	5277.4 (4096–6144)
Learned predictor	21.0 (16–24)	574.9 (512–768)	5016.6 (4096–6144)	> 8192	> 8192
SnapKV-w32	185.1 (128–192)	800.2 (768–1024)	1149.8 (1024–1536)	1024.0 (768–1024)	1095.6 (1024–1536)
UnionKV-v2	38.3 (32–48)	146.5 (128–192)	427.7 (384–512)	512.0 (384–512)	627.1 (512–768)

Table 13: **Fresh paired dose and filler contrasts.** Secants measure NLL tokens per added distractor at fixed 32K; ratios divide the 64K crossing by the 8K crossing at $M = 32$. Brackets are paired-example percentile intervals (2,000 resamples, seed 20260922); all reported contrasts are numerically defined in every resample.

Contrast	Scorer	Estimate	95% interval
Dose $0 \to 8$	NLL	12.50	[11.22, 14.66]
Dose $8 \to 32$	NLL	12.66	[11.60, 13.35]
Dose $32 \to 128$	NLL	10.30	[8.72, 12.33]
Dose $128 \to 512$	NLL	10.05	[7.65, 12.60]
Filler 8K $\to$ 64K	Learned predictor	0.561	[0.359, 0.838]
Filler 8K $\to$ 64K	NLL	1.000	[0.933, 1.072]
Filler 8K $\to$ 64K	SnapKV-w32	1.074	[0.801, 1.336]
Filler 8K $\to$ 64K	UnionKV-v2	2.384	[1.522, 2.624]

The immutable progress snapshot is `results/analysis/qwen35_completed_dose_filler_20260922T0932.json` (SHA-256 `829bca74...`). It records the eight raw-cohort hashes, the analysis hash, rung brackets, and every bootstrap censoring count. `src/analyze_competitors.py` regenerates the live analysis; `src/generate_competitor_table.py` verifies raw-file commitments and regenerates the Qwen tables from the live analysis (or the named snapshot via its `--input` option). No missing outcome is imputed.

The Llama sweep covers all five registered doses. Each dose has 1,620 outcomes, 20 paired examples, and a 20/20 full-cache baseline. Release-aware validation passes every raw/manifest pair, and independent replay reproduces all 8,100 stored normalized-containment scores. Natural NLL scores 0/20 at every measured budget through $N = 8192$, so all five crossings and every adjacent-dose bootstrap endpoint pair are right-censored. This is a boundary of the unprotected NLL policy, consistent with the anchor diagnostic in Appendix G.6.

Table 14: **Complete fresh Llama-3.1-8B-Instruct dose sweep at 32K.** Entries are N_{50} in requested retained positions, with measured rung brackets rather than confidence intervals. Bounds preserve censoring. Full-cache accuracy is 20/20 at every dose. Each completed cohort uses the same 20 paired example IDs.

Policy	$M = 0$	$M = 8$	$M = 32$	$M = 128$	$M = 512$
NLL	> 8192	> 8192	> 8192	> 8192	> 8192
Learned predictor	17.97 [16,24]	142.97 [128,192]	724.78 [512,768]	5367.28 [4096,6144]	> 8192
SnapKV-w32	71.86 [64,96]	198.24 [192,256]	192.00 [128,192]	163.25 [128,192]	114.91 [96,128]
UnionKV-v2	19.60 [16,24]	78.38 [64,96]	221.70 [192,256]	627.07 [512,768]	1773.62 [1536,2048]

The paired adjacent secants further separate the policies. For the learned scorer they are 15.63, 24.24, and 48.36 positions per distractor before the final contrast becomes undefined through censoring; the conditional 95% paired intervals are $[12.74, 17.58]$, $[19.56, 29.28]$, and $[34.43, 63.81]$, with 2,000,

2,000, and 1,999 numerical draws. UnionKV-v2's four secants decrease from 7.35 to 2.99. Their resampled intervals are zero-width because the binary curves and measured rung grid yield the same interpolation in every draw. SnapKV-$w32$ peaks at $M = 8$ and then declines: its last three secants are -0.260, -0.299, and -0.126. The first two intervals, $[-1.712, 3.483]$ and $[-1.008, 0.157]$, include zero; only the $M = 128 \to 512$ interval, $[-0.211, -0.062]$, excludes zero. The policies therefore have incompatible dose-response shapes.

The immutable five-dose snapshot is `results/analysis/llama_completed_five_doses_20260922T1543.json` (SHA-256 prefix `915bfc92`); it binds every raw and manifest digest, all operating-point aggregates, all 20 crossings, and all 16 paired adjacent contrasts. An earlier M0/M8 snapshot has SHA-256 prefix `f0a9fc4c`.

At fixed $M = 32$, the complete Llama filler sweep contains 6,400 outcomes over 8K, 16K, 32K, and 64K, with every full-cache baseline at 20/20. Independent replay reproduces every saved score. NLL remains right-censored above 6,144 positions at 8K and 8,192 at the other lengths. The learned scorer's crossings are 1,319.3, 845.3, 724.8, and 930.4, giving an endpoint ratio of 0.705 with paired interval $[0.573, 0.859]$; its largest-to-smallest crossing ratio is 1.820, so the endpoints conceal a decline followed by a rise. SnapKV-$w32$ instead has an endpoint ratio of 1.304, interval $[1.101, 1.689]$. UnionKV-v2 crosses at 221.7 at all four lengths, within the measured 192–256 bracket. All 2,000 paired draws give a ratio of one for v2 because the measured rung accuracies agree; this reflects the rung grid. The three numerical endpoint comparisons have 2,000 numerical draws each; NLL has none. The frozen source is `results/analysis/llama_completed_four_filler_lengths_20260922.json`.

The native-adapter Gemma-4-E4B sweep also contains all five doses, 8,100 outcomes, and a 20/20 full baseline at each dose. All saved predictions are shorter than the logging cap and independently reproduce their recorded scores. Table 15 separates the NLL response from the behavior of the other policies; these native-adapter results replace the historical Gemma attention rows.

Table 15: **Complete fresh Gemma-4-E4B dose sweep at 32K.** Entries are N_{50} in requested retained positions, with measured rung brackets rather than confidence intervals. Bounds preserve censoring. Full-cache accuracy is 20/20 at every dose. Each completed cohort uses the same 20 paired example IDs.

Policy	$M = 0$	$M = 8$	$M = 32$	$M = 128$	$M = 512$
NLL	35.93 [32,48]	137.79 [128,192]	416.90 [384,512]	1299.82 [1024,1536]	5665.43 [4096,6144]
Learned predictor	29.78 [24,32]	5367.28 [4096,6144]	5899.87 [4096,6144]	> 8192	> 8192
SnapKV-w32	83.06 [64,96]	512.00 [384,512]	1600.44 [1536,2048]	1723.32 [1536,2048]	1254.14 [1024,1536]
UnionKV-v2	41.07 [32,48]	150.54 [128,192]	341.99 [256,384]	740.21 [512,768]	1319.34 [1024,1536]

Gemma NLL crossings rise from 35.9 to 137.8, 416.9, 1,299.8, and 5,665.4 positions. The four adjacent secants are 12.73, 11.63, 9.20, and 11.37 positions per distractor, with paired 95% intervals $[10.76, 14.46]$, $[10.14, 12.63]$, $[8.75, 9.91]$, and $[9.83, 13.54]$; each has 2,000 numerical resamples. The rates are positive but unequal: the budget grows with dose without a single displacement constant.

The learned predictor instead jumps from 29.8 positions at $M = 0$ to 5,367.3 at $M = 8$, and is right-censored above 8,192 at $M = 128$ and 512. Its $M = 32 \to 128$ contrast has only four numerical bootstrap draws, so it has no stable slope. SnapKV-$w32$ peaks at the measured $M = 128$ point and then declines: the last interpolated secant is -1.22 with paired interval $[-2.51, -0.16]$, while its measured-rung bound is $[-2.67, 0]$. The grid bound includes zero even though the resampling interval does not. UnionKV-v2's adjacent rates decline from 13.68 to 1.51. Neither pooled attention nor the learned content score inherits the NLL response.

The immutable Gemma snapshot is `results/analysis/gemma_completed_five_doses_20260922.json` (SHA-256 prefix `fd38b8b2`). It commits all five raw/manifest pairs, all 20 crossings, all 16 adjacent contrasts, and the 8,100 saved-score checks. `src/generate_competitor_replication_tables.py` validates and renders the dose and filler tables.

At fixed $M = 32$, Gemma's three primary 8K/16K/32K filler cohorts have 20 paired examples and 20/20 full-cache baselines at every length. NLL crossings are 422.6, 419.5, and 416.9 positions. The endpoint ratio is 0.986, with paired interval $[0.904, 1.086]$. The learned scorer instead grows by $2.676\times$, interval $[2.321, 3.206]$, and SnapKV-$w32$ by $1.365\times$, interval $[1.126, 1.732]$. V2's point ratio is 1.371, but its interval $[0.939, 1.540]$ includes one. Each comparison has 2,000 numerical paired resamples. Over this fourfold range, which ends at 32K, the NLL count transfers approximately, while the learned scorer's budget grows. The separately registered 32K filler execution uses the same twenty IDs as the primary 32K dose execution. All 1,620 repeated rows match the primary predictions and scores exactly; they check reproducibility rather than enlarge the sample. The no-preview replication report validates all 24 registered dose/filler jobs without pooling these executions.

Table 16: **Complete fresh filler-length replication at fixed $M = 32$.** Ratios divide the longest-context crossing by the shortest-context crossing: 32K/8K for Gemma and 64K/8K for Qwen and Llama. Intervals are paired-example percentiles conditional on numerical endpoint crossings; the final column counts numerical resamples out of 2,000. A censored endpoint yields no point ratio. Each crossing uses half its own matched full-cache score; repeated Gemma 32K executions are not pooled.

Model	Policy	Ratio	Conditional 95% interval	Defined
Qwen3.5-9B	NLL	1.000	[0.933,1.072]	2000
	Learned predictor	0.561	[0.359,0.838]	2000
	SnapKV-w32	1.074	[0.801,1.336]	2000
	UnionKV-v2	2.384	[1.522,2.624]	2000
Gemma-4-E4B	NLL	0.986	[0.904,1.086]	2000
	Learned predictor	2.676	[2.321,3.206]	2000
	SnapKV-w32	1.365	[1.126,1.732]	2000
	UnionKV-v2	1.371	[0.939,1.540]	2000
Llama-3.1-8B-Instruct	NLL	–	–	0
	Learned predictor	0.705	[0.573,0.859]	2000
	SnapKV-w32	1.304	[1.101,1.689]	2000
	UnionKV-v2	1.000	[1.000,1.000]	2000

F HISTORICAL PREDICTIONS OF ROUTE-SPECIFIC RANK INFLATION

Table 17: **Historical route-tail statistic and registered classes.** T_{90} is the median over examples of the ratio between the 90th- and 50th-percentile per-route (layer×KV-head) worst-span-token rank, from one score-logging prefill per example at 16K and 32K. The registered threshold $T_{90} \geq 10$ separates the calibration pairs shown here (36–174 versus 1.5–5.4); subsequent verification tests the associated growth rule.

Model	Scorer	T_{90}^{16K}	T_{90}^{32K}	Class	Model	Scorer	T_{90}^{16K}	T_{90}^{32K}	Class
Llama-3.1-8B	TOVA	72.8	173.9	HEAVY	Qwen3.5-27B	TOVA	3.4	5.4	tight
Llama-3.1-8B	SnapKV	62.9	45.6	HEAVY	Qwen3.5-27B	SnapKV	2.3	3.8	tight
Qwen2.5-7B-1M	TOVA	79.4	63.9	HEAVY	Qwen3.5-4B	TOVA	1.6	1.5	tight
Qwen2.5-7B-1M	SnapKV	47.1	36.2	HEAVY	Qwen3.5-4B	SnapKV	2.3	2.2	tight
Qwen3.5-9B	TOVA	2.1	1.5	tight	gemma-4-E4B	TOVA	2.8	1.7	tight
Qwen3.5-9B	SnapKV	2.1	2.4	tight	gemma-4-E4B	SnapKV	3.5	1.6	tight

Route-level diagnostic. The head/layer-mean worst-span rank does not separate the calibration groups, whereas the per-route ratio in Table 17 does: four pairs labeled observed or suspected growth lie at 36–174 and eight comparison pairs at 1.5–5.4. This empty calibration gap motivated the threshold of 10. It is consistent with a heavy tail of weak retrieval routes in the former group. The prospective record below is reported separately from this retrospective separation.

Author-recorded freeze. The archive records `PREREGISTRATION_R4.md`, SHA-256 `5ca3a4c7...`, at 2026-07-13T06:15:38Z, with the frozen predictions JSON, audit manifest, rule

code, and scorer all hash-covered (`PREREG_R4_HASH.txt`); verification cells use n=15 under FlashAttention-2, full-context reference 1.00 everywhere, and the frozen round-1 estimator.

Frozen rules. Rule A ($\leq$32K audits only): $T_{90}(32\text{K}) \geq 10 \Rightarrow$ the pair's N_{50} inflates $> 2\times$ its 32K plateau at some feasible L; $< 10 \Rightarrow$ stays within $\times 2$ at all L (onset explicitly registered as *not* claimed predictable from $\leq$32K). Rule B (one audit prefill at the target L, roughly an order of magnitude cheaper than a kf ladder): a route-quantile magnitude estimate for SnapKV and a directional growth statistic G for TOVA.

Prospective scoreboard. The registered numerical bands contain V1, V2, V5a, and V5b and miss V3 and V4: 4/6. A separate direction check compares each measured count with its frozen reference and calls growth "HEAVY" only when the ratio exceeds two; this agrees in 5/6 cells. V3 is the shared miss (HEAVY predicted, measured/reference $= 1.21$). V4 has the predicted HEAVY direction (2.13) but lies 7% beyond its numerical band, so it is a direction hit and band miss. The magnitude estimators remain weak: errors range from $1.5\times$ under to $9.5\times$ over, and the Llama route-median underestimates by $6.8\times$. These two scoreboards answer different questions and are not pooled into one success rate.

G Evidence annotations across retrieval tasks

This appendix gives the full protocols and per-cell scoreboards of two historical prospective campaigns. Round 5 freezes literal-span rank predictions before eviction ladders on previously unmeasured real-data cells; round 6 extends the audit to multi-span, redundant, and QA settings. The paired diagnostic (Appendix G.6) shows that these annotations are retention proxies, not minimal sufficient sets.

G.1 Historical single-span forecasts on held-out retrieval cells

Cell selection and archive scan for prior ladders. Five (model, task, L) cells on InfiniteBench that an archive-wide scan found without prior eviction ladders (scan output `UNMEASURED_PROOF.txt`): Qwen3.5-9B `passkey`@32K, Qwen3.5-4B `number_string`@32K, phi-4 `passkey`@16K, Qwen3.5-27B `number_string`@32K (disclosed exception: 10 prior causal-intervention rows, no ladder of any kind), and Qwen3.5-2B `number_string`@32K. Full-context validity was gated at mean accuracy ≥ 0.8 (n=10) *before* registration; all five registered cells pass at 1.0 and re-verify at 1.0 with n=15. Two candidate cells failed the gate and were dropped as scope results, before any prediction was made: gemma-4-E4B-it `passkey`@16K (0.0; a degenerate "The pass key is:" loop) and Qwen3.5-4B `kv_retrieval`@32K (0.4). The `kv_retrieval`/structural arm therefore has no surviving cell in this campaign.

Two-stage freeze. Stage 1 (`PREREGISTRATION_REAL.md`, four cells): combined SHA-256 `dbf9182083d30bdbb75819018ff539b1efd58e63628d34a901d3b63ad4131b10`, frozen 2026-07-14T06:22:31Z. Stage 2 (`PREREGISTRATION_REAL_STAGE2.md`, the 27B cell, whose audit ran $\sim$4$\times$ slower): combined SHA-256 `4b816c55fade7fd316b932be8d508d585a4c4fcecbbec798313b281b82fc70ea`, recorded frozen at 06:28:42Z; internal logs place the first 27B eviction row at about 06:29:30Z. Both hashes cover the frozen predictions JSON and the scorer (`score_realcorpus.py`, identical to the frozen round-2 crossing estimator: mean score per kf, validity mean ≥ 0.8, crossing by log-x interpolation at half the full-context score).

Frozen prediction rules. Per (cell, scorer), from a 14-example rank audit (score logging only, no eviction): `nll_only` and TOVA predict N_{50} as the audit median worst-span-token rank with a $\times 2$ band; SnapKV-w32 predicts $1.75\times$ the median layer-mean rank with band $[1\times, 3\times]$ (documented per-head pooling slack); UnionKV-v2 takes the TOVA point (round-2 finding: the union constant tracks its structural component). Predicted crossing $\text{kf}^* = N_{50}^{\text{pred}}/L$. Also registered: a secondary in-range check (measured N_{50} inside the min–max audit worst-rank range) and a competitor-density statement—per-example $\rho = (\text{worst-span rank} - 1)/L$, with cells of larger ρ_{median} predicted to cross at proportionally larger kf.

Table 18: **Round-5 scoreboard, verbatim** (frozen bands; measured implied N_{50} = measured crossing $\times L$; verification ladders n=15/cell, SDPA, 2415 rows, 0 errors, full-context 1.0 on all five cells). **6 PASS / 14 FAIL**; method totals: UnionKV-v2 4/5, TOVA 2/5, `nll_only` 0/5, SnapKV-w32 0/5. Every FAIL is an over-prediction—the measured crossing is at *smaller* kf than the band. The secondary registered min–max check passes 2/20 (27B SnapKV, 4B UnionKV-v2).

Cell	Scorer	Pred. (band)	Meas.	Verdict	Cell	Scorer	Pred. (band)	Meas.	Verdict
9B pk 32K	NLL	399 [200, 798]	55	**FAIL**	2B ns 32K	NLL	573 [286, 1146]	88	**FAIL**
9B pk 32K	TOVA	49 [24, 98]	16	**FAIL**	2B ns 32K	TOVA	120 [60, 240]	88	**PASS**
9B pk 32K	SnapKV	186 [106, 320]	58	**FAIL**	2B ns 32K	SnapKV	693 [396, 1188]	111	**FAIL**
9B pk 32K	Union	49 [24, 98]	88	**PASS**	2B ns 32K	Union	120 [60, 240]	100	**PASS**
4B ns 32K	NLL	1594 [797, 3187]	100	**FAIL**	27B ns 32K	NLL	2342 [1171, 4685]	49	**FAIL**
4B ns 32K	TOVA	118 [59, 235]	82	**PASS**	27B ns 32K	TOVA	105 [52, 210]	48	**FAIL**
4B ns 32K	SnapKV	607 [347, 1041]	195	**FAIL**	27B ns 32K	SnapKV	442 [252, 758]	232	**FAIL**
4B ns 32K	Union	118 [59, 235]	104	**PASS**	27B ns 32K	Union	105 [52, 210]	58	**PASS**
phi-4 pk 16K	NLL	2386 [1193, 4771]	23	**FAIL**					
phi-4 pk 16K	TOVA	57 [28, 114]	16	**FAIL**					
phi-4 pk 16K	SnapKV	331 [189, 567]	52	**FAIL**					
phi-4 pk 16K	Union	57 [28, 114]	23	**FAIL**					

Direction of every miss. All 14 misses have the measured crossing at smaller kf (fewer tokens) than the registered band, by 2–100×. In these cells the whole-literal-span proxy is conservative relative to the observed 50%-of-full crossing.

Registered competitor-density verdicts. Spearman correlation of measured N_{50} against $\rho_{\text{median}} \cdot L$ across the five cells, per scorer: TOVA $+\mathbf{1.00}$—the audit statistic rank-orders all five real cells' measured crossings perfectly, though the absolute level runs ∼1.3–3× high; UnionKV-v2 $+0.60$; SnapKV-w32 $+0.50$; `nll_only` $-\mathbf{0.60}$—**falsified**. Real-text NLL outlier density (ρ_{median} 0.012–0.146, up to 100× the TOVA arm on the same tokens) does not set the content crossing: measured content N_{50} (23–100) sits near the synthetic constants (16–28) plus a small real-format increment, orders of magnitude below ρL.

Post-hoc diagnosis (labeled exploratory). Treating "audit worst-span rank ≤ kept count" as a binary predictor over all (example, kf) verification pairs gives accuracy 0.73–0.96 for TOVA/SnapKV but only 0.34–0.60 for NLL. NLL generations can be correct at keep counts 5–100× below the audited worst literal-span rank. Repeated adjacent needle statements and usable partial spans are plausible explanations; either way, the whole-span annotation does not measure a necessary set. The July real-versus-synthetic shifts were measured from ladders; this campaign shows that a single whole-span audit does not prospectively recover the real-format content crossing.

Artifacts. `results/realcorpus_prereg_20260714/`: `UNMEASURED_PROOF.txt` (never-measured scan), `validity/` (n=10 gate runs, pre-registration), `audit/` (five 14-example rank audits with per-head statistics), `predictions_frozen.json` / `predictions_frozen_stage2.json`, `PREREGISTRATION_REAL.md` + `PREREG_REAL_HASH.txt` and the stage-2 pair, `verify/` (10 ladder files), `score_realcorpus.py`, and `SCOREBOARD_FINAL.txt` (verbatim scorer output).

G.2 Historical multi-span, repeated-evidence, and QA forecasts

Historical hypothesis and freezes. The frozen round-6 hypothesis asserted that accuracy under top-N eviction equals the CDF of the worst required-token rank of a "minimal sufficient evidence set." The experiments did not observe such sets; they annotated literal spans, minimum-ranked copies, or answer variants. We report its verdicts as recorded and interpret the measured statistics as containment proxies under Appendix A. Its rank medians are unconditional, whereas its measured estimator crosses at half the observed full-context score; no conditional-rank adjustment is applied. The archive records three experiments as separately frozen before their first eviction row (all times 2026-07-14 UTC; first eviction row in parentheses): **E1** `PREREGISTRATION_E1.md`, hash `ee8eef4ccba9be53656128402efb73a34b7ba387aea042f0f74a38657bbc8565`, frozen 20:59:01Z (21:00:14Z); **E2** `PREREGISTRATION_E2.md`, hash `b0311aa9f6f6fc810b9fc1e56f3ffb6e004b697da74c6c3086828fd81e3bb4c5`, frozen 21:17:59Z

(21:21:18Z); **E3** `PREREGISTRATION_E3.md`, hash `55b61d65544ad84e8 2f3756f0c26a8399888dd30ca12f6fbc0c520634fcd6d55`, frozen 21:22:56Z (21:23:24Z; revision 2, superseding `136c8945...`, whose pooled-quantile kf ladder under-resolved both arms—statistics and predictions byte-identical, no eviction row in existence at either freeze). Estimator everywhere: the frozen round-2/round-5 crossing estimator (log-x interpolation at half the full-context score, $n \geq 10$ rows per point; validity gate mean full-context accuracy ≥ 0.8 for E1/E2, ≥ 0.6 registered for E3). 3,973 verification rows, 0 error rows, SDPA, bf16, n=15/cell; verbatim scoreboards in `SCOREBOARD_E1/E2/E3.txt`.

Table 19: **E1 scoreboard, verbatim** (multi-span synthetic; Qwen3.5-9B@32K, phi-4@16K; n=15/cell, all 8 validity gates 0.80–1.00). P1 compares a half-full-baseline crossing with an unconditional audit median of the max-over-K-spans worst-token rank; this is the frozen heuristic, not a normalized-accuracy identity. P1 passes when the crossing lies inside the $\times 2$ band. P2: predicted-CDF vs. measured-curve MAE ≤ 0.15 (shape Spearman in parentheses). Totals: **P1 15/24** ($K \leq 2$: 11/12), **P2 14/24** ($K \leq 2$: 11/12), and **P3 6/6**: Spearman(measured, predicted N_{50} across K) $= 1.00$ on every model×scorer. Eight of nine P1 misses are conservative (measured below band); the exception is phi-4 TOVA K=8 (1.2× past the band edge).

Model	K	Scorer	Pred.	Meas.	P1	P2 (MAE)	Model	K	Scorer	Pred.	Meas.	P1	P2 (MAE)
9B	1	NLL	39	30	**PASS**	**PASS** (.067)	phi-4	1	NLL	20	16	**PASS**	**PASS** (.046)
9B	1	TOVA	45	47	**PASS**	**PASS** (.048)	phi-4	1	TOVA	44	29	**PASS**	**PASS** (.099)
9B	1	Union	45	30	**PASS**	**PASS** (.124)	phi-4	1	Union	44	28	**PASS**	**PASS** (.118)
9B	2	NLL	75	55	**PASS**	**PASS** (.050)	phi-4	2	NLL	53	35	**PASS**	**PASS** (.138)
9B	2	TOVA	162	126	**PASS**	**PASS** (.033)	phi-4	2	TOVA	70	86	**PASS**	**PASS** (.123)
9B	2	Union	162	55	**FAIL**	**FAIL** (.200)	phi-4	2	Union	70	48	**PASS**	**PASS** (.135)
9B	4	NLL	209	69	**FAIL**	**FAIL** (.256)	phi-4	4	NLL	136	63	**FAIL**	**FAIL** (.331)
9B	4	TOVA	366	326	**PASS**	**PASS** (.022)	phi-4	4	TOVA	127	254	**PASS**	**FAIL** (.210)
9B	4	Union	366	95	**FAIL**	**FAIL** (.322)	phi-4	4	Union	127	63	**FAIL**	**FAIL** (.255)
9B	8	NLL	2533	235	**FAIL**	**FAIL** (.620)	phi-4	8	NLL	767	150	**FAIL**	**FAIL** (.276)
9B	8	TOVA	941	772	**PASS**	**PASS** (.053)	phi-4	8	TOVA	213	524	**FAIL**	**FAIL** (.152)
9B	8	Union	941	235	**FAIL**	**FAIL** (.353)	phi-4	8	Union	213	132	**PASS**	**PASS** (.114)

E1: multi-span synthetic—K required needles. Task: $K \in \{1, 2, 4, 8\}$ distinct key-phrase needles carrying distinct 10-digit values at uniform random positions; the question lists all keys. The historical metric checks normalized containment of every value without requiring the requested order. Its span annotation tokenizes a leading space together with each value, unlike the value-only annotation in the ordered-output diagnostic. The proxy requirement is the maximum worst-token rank over those annotated spans. Measured N_{50} rises monotonically with K on all six model–scorer arms (Spearman 1.00). At $K \leq 2$, proxy-CDF and measured curves have MAE 0.03–0.14, a close empirical agreement. At $K \geq 4$, point predictions are often strongly conservative, consistent with the model completing values from subsets of the nominated literal spans. Full values and original scoreboard remain in Table 19.

Table 20: **E2 part A (post-hoc, labeled): redundancy-aware re-audit of the five round-5 cells** against their already-measured crossings (`e2/POSTHOC_ANALYSIS.txt`). Conservatism = predicted/measured N_{50}; "removed" is the scorer's reported median log-conservatism removed by the statistic change.

Lane	Round-5 conservatism	Redundancy-aware	Retro ×2-band hits	Removed
`nll_only`	6.5–105.5×	**0.58–1.27**×	5/5 (was 0/5)	1.07 (all of it)
TOVA	1.4–3.5×	0.98–2.14×	3/5 (was 2/5)	0.56
SnapKV-w32	1.9–6.4×	1.66–3.18×	2/5 (was 0/5)	0.38

E2: redundancy-aware real-corpus audit. The audited proxy is the minimum, over stated answer copies, of the worst-token rank within a copy. Post hoc, this proxy reduces the content scorer's 6.5–105.5× discrepancy to 0.58–1.27×. Prospectively, the two NLL cells land near their predicted points (97 vs. 100; 46 vs. 42). The two 2B attention cells that the frozen scorer printed as insufficient are band misses, because their measured crossings exceed the upper band. This exposes the tradeoff directly: auditing a fixed literal copy was conservative in the observed sample, whereas taking the best-ranked copy can under-provision when the model uses another copy.

Table 21: **E2 part B (prospective historical record).** The frozen scorer printed 6 PASS / 0 FAIL / 2 INSUFFICIENT. Under consistent band accounting, the two Qwen3.5-2B TOVA/Union crossings (37 and 46) exceed the registered upper bound 36 and are upper-band misses, not unresolved low-end crossings. The artifact keeps both classifications.

Cell	Scorer	Meas. N_{50}	Pred. (band)	Verdict
Qwen3.5-9B `ns@16K`	`nll_only`	97	100 [50, 200]	**PASS**
Qwen3.5-9B `ns@16K`	TOVA	70	91 [46, 182]	**PASS**
Qwen3.5-9B `ns@16K`	SnapKV-w32	224	360 [206, 618]	**PASS**
Qwen3.5-9B `ns@16K`	UnionKV-v2	106	91 [46, 182]	**PASS**
Qwen3.5-2B `passkey@32K`	`nll_only`	46	42 [21, 84]	**PASS**
Qwen3.5-2B `passkey@32K`	TOVA	37	18 [9, 36]	**FAIL**
Qwen3.5-2B `passkey@32K`	SnapKV-w32	88	147 [84, 252]	**PASS**
Qwen3.5-2B `passkey@32K`	UnionKV-v2	46	18 [9, 36]	**FAIL**

E3: real multi-hop QA curve shape. RULER `qa_1`, Qwen3.5-9B at 16K (the mirror caps at 16,384 tokens, so the planned 32K cell ran at 16K). The proxy is the minimum, over gold-answer variants, of the worst in-variant rank; its per-example CDF is compared with the measured curve. Full-context mean is 0.733, n=15, with eight ladder points. Registered positive Spearman association occurs for NLL (0.81), UnionKV-v2 (0.79), and TOVA (0.09), though TOVA is nearly flat and provides little shape signal. MAE is 0.20–0.31. UnionKV-v2's half-full-baseline crossing is close to the unconditional proxy median; NLL reaches its measured half-full crossing about five times earlier. These are different estimands unless the normalization bridge above holds, and the E3 rerun shows that its no-rescue condition can fail (Appendix A). Correct answers when the annotated event is false are consistent with partial, paraphrastic, alternative, or contextually encoded evidence; these rows do not distinguish among them. The curves and correlations are descriptive.

Execution notes. Prompt encoding was fixed at the validity-gate stage before the freeze and is shared by all audits and verifications: Qwen3.5 receives an empty `<think></think>` block plus an answer lead-in, singular for K=1 ("The answer is") after a 9B gate failure on the plural form, and phi-4 a plain lead-in. Audit rows written by a stale earlier run are excluded: all E1 audits were rerun before the freeze. The 9B K=8 verification, interrupted at example 13 of 15, was completed for examples 12–14 with a `--sample-start` resume flag added to `ext_multispan.py` after the E1 freeze; the flag changes the hashed file but not task generation, encoding, scoring, or the seed-stable example IDs (final cell 465/465 rows, 0 duplicates, n=15 at every ladder point). E2 audits log scores only and cannot affect any scoreboard.

Verdict and artifacts. The historical data show perfect ordering of measured crossings across K in the tested multi-span arms and positive proxy-curve association in the tested real-QA arms. Point calibration weakens sharply for $K \geq 4$ and for paraphrastic QA. The nominated spans are containment proxies. The rank CDF tracks accuracy only to the extent that retained-but-incorrect and omitted-but-correct terms are small or cancel; these experiments show both close and poor calibration. Artifacts remain under `results/extension_20260714/`, with both the printed and the reclassified verdicts.

G.3 E1 MULTI-SPAN RERUN

The registered E1 rerun separates validity, rank audit, and verification into 24 jobs. All eight 10-example validity jobs, all eight 15-example rank audits, and all eight 15-example verification ladders are complete for Qwen3.5-9B and Phi-4 at $K \in \{1, 2, 4, 8\}$. The seven E2 rank-only audits and all three E3 QA stages follow separately; historical E2/E3 values are not mixed into them.

On Qwen, the NLL crossing rises from 29.6 at $K = 1$ to 55.1 and 68.5 at $K = 2$ and $K = 4$, while the annotated NLL-rank median rises from 39 to 75 and 205. At $K = 8$, the first tested rung, 235.2 nominal positions, already exceeds half-baseline accuracy, whereas the annotated median is 2,410. This left-censored contrast shows that the historical span annotation is conservative in this cell; the crossing lies somewhere below 235.2. Every Qwen matched full baseline is 15/15.

Table 22: **E1 multi-span verification.** The frozen outcome requires every value to match by raw, normalized, or digit-stripped containment, without an ordering requirement. "Full" is the matched verification baseline. Method entries are nominal requested counts—fixed actual prompt length times requested fraction—at the first half-full-baseline crossing; the smaller line is the adjacent measured-rung bracket, not a confidence interval or realized cache length. "Annotated NLL rank" is the unconditional median maximum rank across the K historical spans, each of which tokenizes a leading space together with its value. Comparing it with a half-full crossing is an empirical heuristic.

Model	K	Full	Annotated NLL rank	NLL	TOVA	UnionKV-v2
Qwen3.5-9B	1	15/15	39	29.6 [22.5,39.0]	47.3 [39.0,78.0]	30.2 [22.5,39.0]
Qwen3.5-9B	2	15/15	75	55.1 [40.5,75.0]	120.3 [75.0,150.0]	55.1 [40.5,75.0]
Qwen3.5-9B	4	15/15	205	68.5 [52.2,83.6]	326.1 [183.0,366.0]	77.3 [52.2,83.6]
Qwen3.5-9B	8	15/15	2410	≤ 235.2	759.3 [633.3,941.0]	≤ 235.2
Phi-4	1	13/15	20	16.2 [11.0,20.0]	29.1 [20.0,40.0]	28.3 [20.0,40.0]
Phi-4	2	12/15	53	35.0 [26.5,35.0]	86.1 [70.0,106.0]	47.8 [35.0,53.0]
Phi-4	4	12/15	136	63.2 [50.8,68.0]	254.0 [127.0,254.0]	63.2 [50.8,68.0]
Phi-4	8	15/15	767	150.3 [85.2,191.8]	523.9 [383.5,767.0]	131.6 [85.2,191.8]

The Phi-4 matched full baselines are 13/15, 12/15, 12/15, and 15/15 for $K = 1, 2, 4, 8$. Of 2,000 paired resamples, 252, 690, and 736 fall below the 0.8 full-baseline gate at $K = 1, 2, 4$; no $K = 8$ resample is invalid. All valid $K = 1$ resamples cross for every method. At $K = 2$, NLL and UnionKV-v2 cross in all 1,310 valid resamples; TOVA crosses in 1,295 and is right-censored in 15. At $K = 4$, NLL, TOVA, and v2 have 1,080/184, 1,068/196, and 1,214/50 crossing/right-censored resamples after the 736 invalid baselines. All 2,000 $K = 8$ resamples cross. Any percentile interval in the analysis is conditional on the numerical-crossing resamples and must be read with these dispositions. The measured Phi NLL crossings rise from 16.2 to 35.0, 63.2, and 150.3 nominal positions, while the annotated-rank medians rise from 20 to 53, 136, and 767; the widening $K = 8$ gap again shows that the annotation is not a behavioral threshold.

The fail-closed analyzer validates each registered manifest, output hash, exact row identity, binary score, paired example set, prompt length, rank range, and realized compressed-cache length. Validity and verification targets match exactly on their shared IDs. Audit and verification share validated generator source, registered arguments, deterministic seeds, IDs, context lengths, and answer prefixes; prompt hashes were not serialized, so prompt equality rests on these shared inputs. Of 2,955 verification rows, 2,913 saved predictions are shorter than the 500-character logging cap and independently reproduce the frozen scorer. The other 42 are capped—one Phi $K = 4$ row and 41 Phi $K = 8$ rows—so their recorded hits were computed from the full in-memory prediction but cannot be reconstructed from the saved string. Digit stripping adds partial hits by joining number boundaries in 11 uncapped Phi rows; it changes none of their binary all-values outcomes. The table therefore names the exact frozen metric. Analysis source, derived registry, and generated-table SHA-256 prefixes are `83886908`, `b8dea205`, and `5f48250c`; full cohort hashes and checks are in `docs/extensions_analysis.md`.

G.4 E2 REPEATED-EVIDENCE RANK AUDITS

The E2 rerun contains seven rank-audit cells and no generation ladder. Each cell has 14 examples with two complete occurrences of the answer, fixed actual length 16K or 32K, and three recorded score lanes. For each scorer the annotated threshold is the better-ranked copy's worst token, the minimum over the two copy-wise maxima. Table 23 reports medians; the final column compares the legacy first-copy NLL statistic with the better-copy statistic on the same examples.

The better-copy statistic changes the operational annotation substantially: the median NLL legacy/best ratio ranges from $6.54\times$ to $175.04\times$ across these cells. This confirms that a fixed-copy annotation can be highly conservative when the answer repeats. E2 runs no eviction or decode, so it does not test whether the better-ranked copy suffices; choosing a minimum can under-provision when generation

Table 23: **E2 repeated-evidence rank audits** (n=14/cell). NLL, TOVA, and SnapKV entries are median better-copy worst-token ranks. "Legacy/best" is the median per-example ratio for NLL. These are rank audits, not generation crossings.

Cell	NLL	TOVA	SnapKV	NLL legacy/best
Q3.5-9B / number / 16K	100.0	91.5	206.5	22.83×
Q3.5-2B / passkey / 32K	42.0	18.0	84.0	6.54×
Q3.5-4B / number / 32K	85.5	99.5	187.5	19.21×
Phi-4 / passkey / 16K	13.0	35.0	94.0	175.04×
Q3.5-27B / number / 32K	64.0	100.5	219.5	35.27×
Q3.5-9B / passkey / 32K	66.0	16.0	85.0	6.75×
Q3.5-2B / number / 32K	72.0	90.5	158.0	7.74×

uses another copy. All seven outputs contain one metadata and 14 audit rows, with no missing scorer lanes or attention errors. They pass release-aware manifest and exact-identity validation.

G.5 E3 QA SHAPE RERUN

The E3 rerun is a separately registered shape test, not the official-metric RULER campaign in Appendix M. It uses Qwen3.5-9B, the pinned RULER `qa_1` 16K source, an empty think block plus "The answer is" lead-in, at most 32 generated tokens, and legacy normalized any-reference containment. The 15 prompts contain 14,191–16,189 tokens, so retained fraction rather than cap times fraction is the primary coordinate. Every saved prediction is shorter than the 300-character logging cap, all saved references equal the complete source references, and all 375 scores independently reproduce from those strings.

The full baseline is 11/15. It passes E3's historically registered 0.6 validity gate but not the 0.8 gate used for the main prediction accounting and boundary tables. Table 24 gives accuracy counts at the eight requested fractions; every policy curve is nonmonotone.

Table 24: **E3 Qwen3.5-9B `qa_1` shape test** (n=15). Entries are normalized any-reference containment counts; full cache is 11/15. Percentages are requested retained fractions, not realized counts.

Policy	.049%	.076%	.16%	.57%	1.1%	9.9%	20%	79%
NLL	3	2	3	3	3	4	5	9
TOVA	8	12	13	12	11	12	12	12
UnionKV-v2	5	3	6	4	4	14	14	11

Against half of the E3-specific full baseline, the point curve gives an NLL crossing of 23.75% in the measured 20–79% bracket, a left-censored TOVA result at $\leq 0.049\%$, and a UnionKV-v2 crossing of 0.141% in the 0.076–0.16% bracket. Paired resampling exposes their instability. Of 2,000 resamples, 158 fail the 0.6 baseline gate. NLL has 1,713 numerical crossings, 118 left-censored and 11 right-censored draws; TOVA has 1,645 left-censored and only 197 numerical crossings; UnionKV-v2 has 1,111 crossings and 731 left-censored draws. The crossings are therefore unstable.

The separate audit records median annotated thresholds of 11,442 for NLL, 23 for TOVA, and 87 for SnapKV-w32. They are minimum-over-answer-occurrence containment proxies; UnionKV-v2 has no single matching audit lane. With nonmonotone curves, correct answers below the nominated thresholds, an E3-specific gate, and an answer lead-in, this rerun marks a boundary of rank prediction on multi-hop QA. All 34 registered extension jobs have validated outputs; E2 is rank-only and E3 is a protocol-specific shape test. Analysis source and derived-registry SHA-256 prefixes are `b43e978e` and `335e2df3`; the E3 verification raw SHA-256 is `55b1de33`.

G.6 PAIRED RETENTION–GENERATION DIAGNOSTIC

Design. We directly record the retention event and generation outcome on the same runs. Each exactly 16,384-token prompt contains $K \in \{1, 2, 4, 8\}$ distinct key–value facts; the redundant condition places three complete copies of each fact, while the unique condition places one. Unique and redundant examples are paired on keys, ten-digit values, and first-copy positions (seed 2026092201; $n = 20$ per condition). One uncompressed prefill of Qwen2.5-7B-Instruct-1M computes token NLL.

Table 25: **Qwen2.5-7B-Instruct-1M joint events at the first tested budget with complete annotated retention in 20/20 examples.** Y: ordered exact answer; Cont.: historical all-value containment. "Ret.–wrong" and "Omit.–correct" are observed joint events, not identified causal mechanisms.

Copies	K	N	E_N	Y_N	Cont.	Ret.–wrong	Omit.–correct
one	1	64	20	20	20	0	0
one	2	128	20	20	20	0	0
one	4	128	20	13	20	7	0
one	8	256	20	9	9	11	0
three	1	64	20	20	20	0	0
three	2	128	20	19	19	1	0
three	4	256	20	16	16	4	0
three	8	256	20	8	11	12	0

Sorting NLL with prompt position as a deterministic tie break fixes one budget-independent order. Separate kvpress prefills then physically retain the top $N \in \{16, 32, 64, 128, 256, 512, 1024, 2048\}$ rows and greedily decode. All 1,280 compressed rows have the requested physical cache length; the 160 paired full-cache baselines and all compressed rows completed without an error or exclusion.

For value v and copy c, let R_{vc} be the worst rank among the copy's value-token rows. The annotated event is

$$E_N = \left\{ \max_v \min_c R_{vc} \leq N \right\}.$$

This is an exact containment event for the annotated rows, not a claim that they are a sufficient set: it omits the key–value binding, the question, and information that full prefill may have written into other contextual KV entries. The primary outcome requires the generated sequence of isolated ten-digit codes to equal the requested sequence exactly and in order. We retain all-value substring containment as a secondary historical metric.

Results and interpretation. The single-value proxy is close to the behavioral outcome: in the unique condition E_N and ordered accuracy both move from 0/20 at $N = 16$ to 20/20 at $N = 64$, with one omitted-but-correct example at $N = 32$. For two unique values at $N = 64$, retention is 11/20 and accuracy is 14/20; the difference is three omitted-but-correct outcomes. Agreement breaks in the other direction as output complexity grows. With four unique values at $N = 128$, all annotated spans survive and all answers contain every value, yet only 13/20 put the codes in the requested order. With eight unique values at $N = 256$, retention is 20/20 and ordered accuracy is 9/20; accuracy remains 8–10/20 through $N = 2048$ although retention stays 20/20, while the paired full-cache condition is 20/20. Three-copy redundancy does not repair this gap: at $K = 8, N = 256$, retention is 20/20, ordered accuracy 8/20, and containment 11/20. These are failures of annotated-span sufficiency for this decoder and metric; the row interventions below locate the missing rows.

Second-model boundary and compatibility control. Under natural NLL selection on the verified Llama-3.1-8B-Instruct checkpoint, complete unique-value retention reaches 20/20 by $N = 32, 64, 64, 128$ for $K = 1, 2, 4, 8$, yet ordered accuracy is 0/20 for every unique condition through $N = 2048$. The full-cache ordered baselines are 20/20 for $K = 1, 2$ and 17/20 for $K = 4, 8$; all-value containment is 20/20 in every condition. A diagnostic-informed anchor control shows that this result is highly sensitive to selector–model compatibility: at unique $K = 1, N = 128$, natural NLL gives 0/20 ordered and 0/20 containment, forcing prompt rows 0–3 while filling the other 124 slots by the same NLL order gives 13/20 ordered and 20/20 containment, and also forcing the final 32 rows gives 15/20 and 20/20. Relative to natural NLL, the paired ordered gains are 13/20 with no losses and 15/20 with no losses (two-sided exact McNemar $p = 2.44 \times 10^{-4}$ and 6.10×10^{-5}); the 95% Wilson intervals are $[0.433, 0.819]$ and $[0.531, 0.888]$. The extra tail rows add two ordered successes over first-four protection, too few to isolate a tail effect. These policies were chosen after inspecting the diagnostic and are oracle controls, not a deployable selector. They show that complete value retention is insufficient for this configuration and that the outcome depends on retaining other prompt rows. The full derived registry SHA-256 is `514c75ce...`; control raw SHA-256 is `5bf08a68...`. All 60 control rows have exact physical budget 128 and no errors.

Annotation and audit trail. Value annotations exclude separator tokens. The raw pass annotated the complete output of `tokenizer(" "+value)`, and Qwen encodes the leading separator as standalone whitespace token 220, which is prompt syntax rather than answer evidence. The derivation removes only tokens whose individual decode is nonempty whitespace, then recomputes full-span, first-token, all-copy, and joint-event fields from the saved token IDs and per-token ranks. It changes 466 of 1,280 raw retention indicators; generation, frozen orderings, retained caches, and outputs are unchanged. Paper analyses read only `results/derived/evidence_diagnostic/qwen25_7b_16k_full.corrected.jsonl` (SHA-256 `8867b50c...`). Every derived row nests the original fields under `raw_annotation`; its metadata records the immutable raw file SHA-256 `6f49fdbb...`, tokenizer-file hashes, and derivation source hash. The runner, correction script, exact schema, and commands are documented in `docs/evidence_diagnostic_protocol.md`.

G.7 Diagnostic-informed value-row intervention

The completed Qwen diagnostic motivated a fixed-$N = 256$ follow-up on the same unique $K = 1, 4, 8$ prompts. This budget was chosen after the diagnostic showed that it retained every value row in 20/20 examples for all three conditions, so the follow-up is post hoc. Four arms retain no value rows, the first token row of each value, the last token row, or every value row. Each arm has exactly 256 positions and fills its remaining slots from the same stable NLL-ranked non-value pool. Thus the intervention concerns value-associated KV rows after full contextual prefill, not whether the input string or its information remains elsewhere in the cache.

Table 26: **Paired Qwen value-row intervention at** $N = 256$ ($n = 20$ per cell). Entries are ordered exact / all-value containment counts. The natural top-NLL anchor and whole-value arm produce byte-identical prediction strings on all 60 paired examples. "Partial" covers none, first-only, and last-only; all three have the displayed outcome. Wilson intervals refer to ordered accuracy.

K	Full cache	Natural NLL	Whole values	Partial (each)	95% CI, whole
1	20/20	20/20	20/20	0/0	[0.839, 1.000]
4	20/20	17/18	17/18	0/0	[0.640, 0.948]
8	20/20	9/9	9/9	0/0	[0.258, 0.658]

Every successful whole-value output becomes incorrect after replacing all but the first or last value rows with matched NLL donors: 20, 17, and 9 discordant pairs for $K = 1, 4, 8$. Two-sided exact McNemar values are 1.91×10^{-6}, 1.53×10^{-5}, and 3.91×10^{-3}, respectively; the zero-success partial-arm Wilson interval is $[0, 0.161]$. This establishes a causal effect of these row-set replacements in the measured cache states; it concerns row sets, not individual rows. Whole-value retention is still not sufficient: the whole arm itself fails on 3/20 $K = 4$ examples and 11/20 $K = 8$ examples despite the valid full-cache baseline. All 240 rows complete without error, have physical cache length 256, and log the exact forced, excluded, donor, and realized sets. Raw SHA-256 is `ddf1a6a9...`; summary SHA-256 is `670f35a0....`

G.8 Diagnostic-informed contextual-row follow-up

The incomplete ordered outputs after whole-value retention motivated a second paired Qwen follow-up at the same $N = 256$. It uses the unique $K = 4, 8$ prompts and compares natural NLL retention with three forced groups: every complete key–value record row; the question suffix plus every value row; and both complete records and the question suffix. All four arms retain every value row. The three interventions fill unused capacity from one stable NLL order that excludes the union of the named groups, and log every forced, donor, and realized position. The complete candidate prompt contains the question in every arm, but under this compression convention its KV rows can be evicted. The row groups are oracle annotations chosen after the diagnostic, not a deployable selector.

Forcing complete records while omitting the question suffix loses all 17 natural ordered successes at $K = 4$ and all 9 at $K = 8$, with no gains (two-sided exact McNemar $p = 1.53 \times 10^{-5}$ and 3.91×10^{-3}). The question-plus-values swap has two gains and one loss at $K = 4$ ($p = 1$), and ten gains and one loss at $K = 8$ ($p = 0.0117$). Forcing both named groups reaches 20/20: three gains and no losses at $K = 4$ ($p = 0.25$), and eleven gains and no losses at $K = 8$ ($p = 9.77 \times 10^{-4}$).

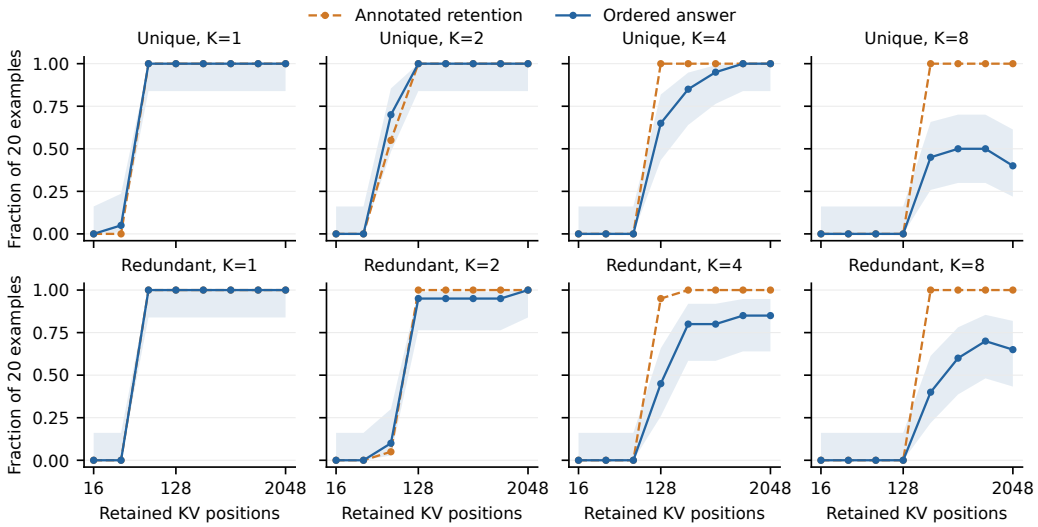

Figure 5: **Complete Qwen diagnostic ladders omitted from the main-text summary figure.** Orange is annotated-value retention and blue is ordered-answer accuracy for every unique/redundant $K \in \{1, 2, 4, 8\}$ condition ($n = 20$ paired examples per budget). In redundant conditions, retention requires at least one complete copy of every requested value. Shading is the pointwise 95% Wilson interval for ordered accuracy. Every paired full-cache baseline is 20/20. These are natural fixed-NLL policies; the post-diagnostic oracle swaps appear in Tables 26 and 27. Source registry: `qwen25_7b_16k_full.corrected.jsonl`, SHA-256 `8867b50c...`.

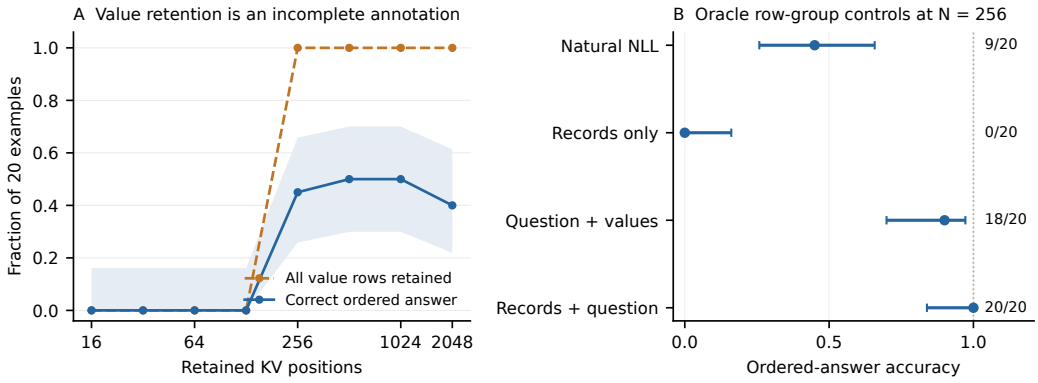

Figure 6: **Detailed Qwen retention and post-diagnostic rescue view.** Qwen2.5-7B-Instruct-1M, 16K prompts, eight unique required values, $n = 20$. Left: the natural NLL ordering retains all annotated value rows at $N = 256$, while ordered accuracy is 9/20. Right: named row-group interventions at the same 256-row budget; complete records and the question restore 20/20. These are oracle controls chosen after the diagnostic. The full cache scores 20/20. Shading and error bars are pointwise 95% Wilson intervals.

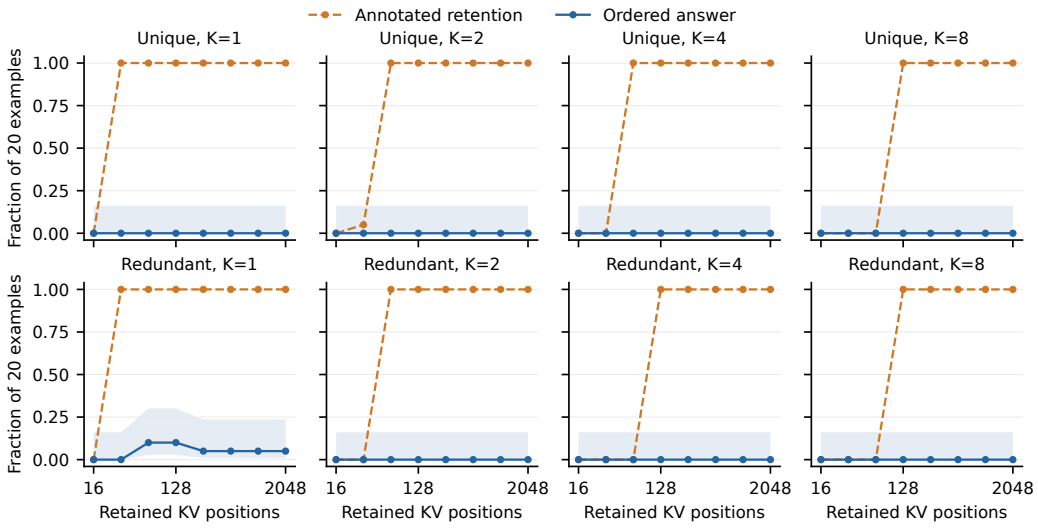

Figure 7: **Complete Llama diagnostic ladders.** The layout and metrics match Figure 5. Natural NLL retains the annotated value rows but produces almost no correct compressed outputs in these conditions. Full-cache ordered baselines are 20/20 for unique $K = 1, 2$ and 17/20 for unique $K = 4, 8$; all-value containment is 20/20. The anchor control reported in the text shows that fixed prompt-row protection rescues much of this failure. Source registry: `llama31_8b_16k_full.corrected.jsonl`, SHA-256 `514c75ce...`.

Table 27: **Qwen contextual-row swaps at** $N = 256$ ($n = 20$ paired prompts per cell). Entries are ordered exact / all-value containment counts. All paired full-cache baselines are 20/20.

K	Natural NLL	Complete records	Question + values	Records + question
4	17/18	0/0	18/19	20/20
8	9/9	0/0	18/20	20/20

Complete value or record retention therefore does not determine the answer, and swaps that include the question suffix substantially improve accuracy at $K = 8$. Each swap also removes an equal number of donor rows, so the contrasts compare row sets rather than isolate a single group. All 160 rows have exact physical cache length 256 and pass the identity, source-hash, common-donor, and realized-set checks. Raw SHA-256 is `9b7c3544...`; summary SHA-256 is `8adc8d65...`.

H Controlled KV-row deletion and rescue

Table 28: **Historical content-scale screen (E4).** Normalized answer-containment score after evicting N=14 tokens from the prefilled cache at 32K, by targeting rule (n=40/cell; random averaged over 3 seeds).

	`number_string`		`kv_retrieval`	
Condition (N=14 evicted)	Qwen3.5	Qwen2.5-1M	Qwen3.5	Qwen2.5-1M
Full cache (no eviction)	1.00	1.00	0.30	0.30
Evict ground-truth span	0.00	0.00	0.00	0.00
Evict top-NLL tokens	0.00	0.05	0.25	0.30
Evict top-attention tokens	0.45	0.80	0.25	0.30
Evict random (matched N)	1.00	1.00	0.28	0.28

Content-scale screen. Evicting the ∼14 highest-NLL tokens collapses `number_string` (1.00 → 0.00/0.05), while the recorded `kv_retrieval` scores change little. The `kv_retrieval` column is invalid: the historical decoder allowed only 32 output tokens, which truncates UUID answers (for example, an observed prediction ends at `5e6b7b90-710d-4953-9b18-3e96b2` before the full target suffix), so its 0.30 full-cache baseline is a protocol failure. Structural-task reruns use a 128-token output cap.

H.1 E4 INTERVENTION WITH A UUID-SAFE DECODE CAP

We repeated E4 on Qwen3.5-9B and Qwen2.5-7B-Instruct-1M with the full UUID-safe decode cap. Each task has 40 registered paired prompts. Full cache succeeds on 40/40 prompts in three model–task cells and 35/40 for Qwen2.5 KV retrieval; all intervention contrasts condition on the same full-cache-competent prompt IDs. The target arm deletes the matched literal-answer neighborhood padded by two token rows on each side, not the bare answer span. NLL and random arms broadcast one global deletion list to every layer and KV head. The attention arm instead selects a separate list in every layer and KV head. Thus N is matched per route, but the attention intervention is not a common global N-token set.

Table 29: **E4 outcomes at 32K with a UUID-safe decode cap.** "Gate" is unconditioned full-cache success among 40 registered prompts; every intervention denominator is the corresponding competent subset. The three random arms share prompts and arise from successive draws of one per-example PRNG stream, so they remain separate paired arms rather than a pooled sample.

Model	Task	N per route	Gate	Padded target	Top NLL	Top attention	Random 0/1/2
Qwen2.5	`number_string`	14	40/40	0/40	1/40	22/40	40 / 40 / 40 of 40
Qwen2.5	`kv_retrieval`	32–40	35/40	0/35	35/35	35/35	34 / 33 / 35 of 35
Qwen3.5	`number_string`	14	40/40	0/40	0/40	22/40	40 / 40 / 40 of 40
Qwen3.5	`kv_retrieval`	32–40	40/40	0/40	38/40	39/40	39 / 37 / 40 of 40

On `number_string`, padded-target deletion differs from full by -1.000 on both models (approximate paired 95% Newcombe interval $[-1.000, -0.876]$). Top-NLL differs by -0.975 $[-0.996, -0.839]$ on Qwen2.5 and -1.000 $[-1.000, -0.876]$ on Qwen3.5; top-attention differs by -0.450 $[-0.602, -0.282]$ on both. For Qwen2.5 `kv_retrieval`, the unconditioned full-cache gate is 35/40. Among those same 35 prompts, top-NLL and top-attention remain 35/35, compared with 34/35, 33/35, and 35/35 under the three random arms. The 35/35 reference is conditional on the gate, not a 40/40 full-cache baseline. Qwen3.5 KV results remain 38/40 and 39/40 for NLL and attention, compared with random 39/40, 37/40, and 40/40.

Neither cohort shows a clean content/structure split at this deletion count: per-route attention deletion harms literal retrieval (22/40 remain correct on each model), whereas deleting 32–40 attention-selected rows per route on KV retrieval resembles the random controls among full-cache-competent prompts. This E4 policy also differs from the crossover policy below, which averages attention scores across layer and head before choosing one global set; their outcomes are not interchangeable. A source-derived rounding audit evaluates kvpress's $T - \lfloor T(1 - N/T) \rfloor$ rule and obtains exactly N evictions for every row, a check on the rule, not runtime cache-index or byte telemetry. The fail-closed analysis is `src/analyze_causal_e4_corrected.py`; Qwen2.5/Qwen3.5 raw and derived SHA-256 prefixes are `0952743b`, `48c2faeb`, and `bf02e66d`. Deleting the padded target neighborhood breaks the answer on every full-cache-competent prompt in both tasks and models.

Frozen crossover design. The archive records `DESIGN.md`, SHA-256 `f02f6c02...`, as frozen at 2026-07-13T03:47Z before its evaluation decodes: evict $N = \lceil 0.01\,T \rceil$ tokens under nine conditions per (model, task)—content-targeted (top-NLL), structure-targeted (top SnapKV-w32, head/layer-mean), a *position-matched* random control for each, and rescue arms; sinks (4) and tail (64) always protected; n=24/cell; paired bootstrap (10,000 draws) percentile CIs; frozen interaction estimator $I_{\text{evict}} = [D(\text{c}, \text{ns}) - D(\text{c}, \text{kv})] - [D(\text{s}, \text{ns}) - D(\text{s}, \text{kv})]$, each D an accuracy delta against the matched control. The design used a full-cache `kv_retrieval` gate of 0.7. Its hard-coded 32-token generation cap is too short for UUID answers: Qwen3.5-9B often emits a correct UUID prefix that is truncated before the suffix, yielding the 8/24 baseline in Table 30. This low base rate is a protocol failure. The historical fallback selected phi-4; the reruns use 128 output tokens for `kv_retrieval` and retain 32 for the literal-number task.

Table 30: **Historical scorer-targeted crossover screen at the structural scale** ($N = \lceil 0.01\,T \rceil$ evicted: 164 tokens for phi-4 at 16K, 328 for the 32K models; n=24/cell, 0 errors). Scores are normalized answer containment. ns = `number_string`, kv = `kv_retrieval`. Bottom: interaction I_{evict} (negative = crossover direction; paired bootstrap 95% CI). phi-4 is the gate-selected historical headline; Llama-3.1 the historical cross-check; Qwen3.5-9B is unresolved at its low kv base rate (decode-cap artifact, see text).

	phi-4 (16K)		Llama-3.1-8B (32K)		Qwen3.5-9B (32K)	
Condition	ns	kv	ns	kv	ns	kv
Full cache	0.96	0.96	1.00	1.00	1.00	0.33
Evict content set	0.00	0.75	0.00	0.92	0.00	0.13
Evict structural set	0.00	0.00	0.00	0.00	0.00	0.00
Random (content-matched)	0.96	0.71	1.00	0.88	1.00	0.29
Random (struct-matched)	0.96	0.38	1.00	0.58	1.00	0.25
I_{evict} [95% CI]	$\mathbf{-0.42}$ $[-0.67, -0.13]$		-0.63 $[-0.96, -0.29]$		-0.08 $[-0.25, +0.08]$ (n.s.)	

Historical results. Under the original cap, the interaction is negative on phi-4 (-0.42) and Llama (-0.63), while Qwen3.5-9B is unresolved (-0.08). The Qwen value is uninterpretable because its full-cache UUID output is truncated; the 128-token reruns below replace it.

Historical rescue asymmetry. The historical rescue arms keep a failing 68-token base and reinsert only 164–328 selected tokens. Reinsertion restores the content task but not `kv_retrieval`. Because the structural baseline itself requires roughly 3–5% of context, these rescue arms are below its operating threshold and cannot distinguish distributed necessity from budget starvation. The intervention is an asymmetric functional dissociation, not a double dissociation: the structural set overlaps the literal content span, and the Qwen interaction includes zero. The rerun below repeats the design at larger budgets.

H.2 Crossover rerun across models and budgets

We repeated all 18 registered cells—three models, two tasks, and 1%, 5%, and 10% selection fractions—with 128 generated tokens for UUID retrieval and 32 for the literal-number task, fixed across all nine arms. Each cohort contains 24 examples, 216 outcome rows, 24 set-summary rows, and no errors or skips. Qwen and Llama pass the full-cache gate on 24/24 examples; phi-4 passes on 23/24, and its contrasts use those same 23 paired examples. Each contrast below compares paired binary outcomes conditional on that gate. Table 31 gives the Qwen 1%/5% effects in detail; Table 32 gives the complete cross-model UUID rescue result.

Table 31: **Qwen paired policy effects.** Δ_E is selected-set eviction minus its position-matched random eviction; Δ_R is selected-set rescue minus equal-count uniform-random rescue. Brackets are approximate 95% Newcombe method-8 paired score intervals (marginal Wilson intervals with a phi-correlation adjustment and no continuity correction) (Newcombe, 1998). The implementation sets phi to zero when a marginal variance is zero, as specified by the method. All gates are 24/24.

Task	Fraction	Δ_E(NLL)	Δ_E(attn.)	Δ_R(NLL)	Δ_R(attn.)
`number_string`	1%	-1.00 $[-1.00, -0.80]$	-1.00 $[-1.00, -0.80]$	$+1.00$ $[+0.80, +1.00]$	$+1.00$ $[+0.80, +1.00]$
`number_string`	5%	-1.00 $[-1.00, -0.80]$	-1.00 $[-1.00, -0.80]$	$+1.00$ $[+0.80, +1.00]$	$+1.00$ $[+0.80, +1.00]$
`kv_retrieval`	1%	-0.29 $[-0.48, -0.10]$	-0.58 $[-0.76, -0.34]$	0.00 $[-0.14, +0.14]$	$+0.04$ $[-0.10, +0.20]$
`kv_retrieval`	5%	-0.38 $[-0.56, -0.13]$	-1.00 $[-1.00, -0.80]$	0.00 $[-0.14, +0.14]$	$+0.96$ $[+0.75, +0.99]$

On `number_string`, both scorer-selected sets cover 100% of the annotated answer rows; their identical deletion and rescue effects therefore do not separate content and structure. On `kv_retrieval`, mean annotated answer-span coverage is 1.4% (NLL) versus 44.3% (attention) at 1%, and 5.3% versus 99.7% at 5%. Correspondingly, attention rescue rises from 1/24 at 1% to 23/24 at 5%, while NLL rescue and uniform-random rescue remain 0/24; the paired 5% attention-rescue effect is $+0.96$ with the interval shown above. Exact two-sided McNemar tests and discordant counts are retained in

the derived JSON as an independent finite-sample check. Intervals are approximate paired sampling intervals over the frozen examples.

Table 32: **UUID rescue cohorts across models.** All counts are conditional on full-cache competence (gate). "Coverage" is the mean fraction of annotated answer-span rows in the attention-selected set. The kept base, NLL rescue, and equal-count uniform-random rescue are 0 in every row. The Qwen rescue does not carry over to phi-4 or Llama under the same policy.

Model	Fraction	Gate	Attn. coverage	Attn. rescue	NLL/random rescue
Qwen3.5-9B	1%	24/24	0.443	1/24	0/24
	5%	24/24	0.997	23/24	0/24
	10%	24/24	1.000	24/24	0/24
phi-4	1%	23/24	0.252	0/23	0/23
	5%	23/24	0.564	2/23	0/23
	10%	23/24	0.761	6/23	0/23
Llama-3.1-8B-Instruct	1%	24/24	0.248	0/24	0/24
	5%	24/24	0.563	3/24	0/24
	10%	24/24	0.723	4/24	0/24

Across all nine literal-number cohorts, NLL and attention rescue succeed on every gate-eligible example while the kept base and random rescue succeed on none. The UUID result is architecture dependent: at 5%, attention rescue is 23/24 for Qwen, 2/23 for phi-4, and 3/24 for Llama, alongside attention-set answer coverage of 0.997, 0.564, and 0.563. Increasing the fraction to 10% raises phi-4 and Llama only to 6/23 and 4/24, respectively, while Qwen reaches 24/24. The rescue effect is therefore strong on Qwen and model-specific.

At adequate budget, the reruns show an asymmetric functional policy contrast on Qwen, so the failed low-budget historical rescue did not indicate nonlocalizability. The contrast does not separate independent content and structure mechanisms: the attention set retains far more answer-span rows, the selected sets overlap, and full prefill may write answer information into other contextual KV entries. The raw format records span-coverage and overlap summaries but not selected indices or allocated bytes, so it supports neither individual-row necessity nor physical-memory claims. The fail-closed analysis is `src/analyze_causal_crossover_corrected.py`, SHA-256 `2f6cff1e...`; the complete 18-cohort derived registry has SHA-256 `0cabacf9...`.

I CONTENT AND STRUCTURAL ALLOCATION WITH UNIONKV

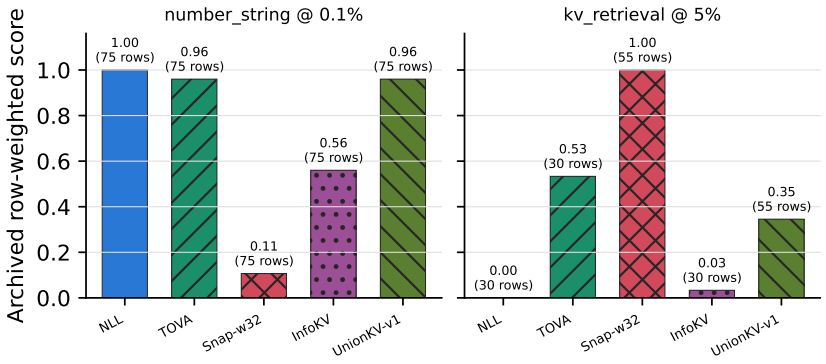

Figure 8: **Historical task-dependent policy scores (Qwen3.5-9B, 128K cap).** Left: `number_string` at 0.1% requested retention; each mean pools 45 synthetic and 30 InfiniteBench rows with actual lengths 125,358 or 131,072. Right: InfiniteBench `kv_retrieval` at 5%, with exactly 131,072 input tokens; NLL, TOVA, and InfoKV have 30 rows each, while SnapKV-w32 and v1 have 55. These are the archived row-weighted summaries: repeated suite/example identities remain, so the bars are descriptive rather than independent-sample estimates. The union arm is v1, not v2. The matched 32K comparison below gives the v2 evidence.

Table 33: **UnionKV-v2 at the content cliff** (128K synthetic `number_string`, n=20/cell). Accuracy vs. keep-fraction (retained count in parentheses). Bold: within 0.05 of the best single method (prereg P5a; 7/8 cells, miss at the 26-token budget where the allocator's seed-window and span-completion overhead exceeds the entire budget).

Model	Method	2×10^{-4} (26)	5×10^{-4} (66)	10^{-3} (131)	2×10^{-3} (262)
Qwen3.5-9B	NLL only	0.85	**1.00**	**1.00**	**1.00**
	InfoKV-style	0.00	0.00	0.85	**1.00**
	UnionKV-v1	0.00	0.85	**1.00**	**1.00**
	UnionKV-v2	0.00	**1.00**	**1.00**	**1.00**
Qwen2.5-7B-1M	NLL only	**0.00**	**1.00**	**1.00**	**1.00**
	InfoKV-style	**0.00**	0.00	0.10	**1.00**
	UnionKV-v1	**0.00**	0.05	**1.00**	**1.00**
	UnionKV-v2	**0.00**	**1.00**	**1.00**	**1.00**

Two-sidedness. At 0.1% of a 128K context, content methods are perfect or near-perfect (NLL 1.00, TOVA 0.96) and SnapKV-w32 scores 0.11; at 5% on `kv_retrieval` the order inverts (NLL 0.00, TOVA 0.53, SnapKV-w32 1.00). These cells motivate the historical Type I/Type II taxonomy (Appendix L). The InfoKV-style fixed blend scores 0.56 on the content cell at 0.1%, 0.10 on the Qwen2.5-1M content cell, and 0.03–0.20 on the tested `kv_retrieval` budgets.

The allocator. The historical UnionKV-v2 configuration is `g050_w8_s8`, but the harness dispatches `unionkv2` with `adaptive=True`. With NLL available, the adaptive branch reserves sinks and a capped recent window, then requests $\min(B, \max(8, \#\{z > 3\} + 16), L_{\text{mid}})$ content candidates, where B is the remaining budget. Structural candidates are added afterward by attention-seed dilation. The $\gamma = 0.50$ structural-first branch applies to v1 or the fallback, not to this v2 branch. Consequently, a 26-position v2 cache cannot be described as allocating exactly 13 positions to content. This prompt-dependent outlier heuristic is neither a calibrated count nor a guarantee that all content candidates survive final scoring when dilation overfills the union. Appendix I.1 therefore evaluates a count calibrated on disjoint development examples, with explicit accounting of protected positions, overlap, and final cardinality.

Results and scoping. Content side (Table 33): v2 matches the best single method in 7/8 registered cells; at 512K it crosses at 7.7×10^{-5}, inside the registered bound. Structural-side curves in this historical campaign are for v1: 0.35 at 5%, 0.80 at 10%, and 1.00 at 25%, while SnapKV reaches 1.00 by 5%. On held-out Mistral-Nemo, v2 crosses where its content component does not. The historical campaign therefore tested v2 on the content side only. The matched 32K comparison below evaluates the same adaptive-v2 configuration on both tasks and finds a structural weakness; the 128K v1 curves above belong to the separate historical campaign. V2's parameters were chosen by small sweeps.

Matched 32K binding comparison. The main-text selector comparison uses Qwen3.5-9B on 15 InfiniteBench `kv_retrieval` prompts truncated to 32,768 tokens, with the query visible to every arm and a 15/15 full-cache baseline. At requested fractions of 1%, 5%, 10%, and 25%, adaptive UnionKV-v2 scores 0, 0, 0, and 1 of 15; SnapKV-w32 scores 1, 11, 15, and 15; ChunkKV (chunk size 20) ranked by SnapKV-w32 scores 6, 15, 15, and 15; and ChunkKV with NLL scores and NLL alone score 0/15 at every budget. At 5%, the mean realized cache length is 1,608 for ChunkKV with SnapKV-w32 scores and 1,638 for SnapKV-w32 and UnionKV-v2. The paired literal-retrieval file gives NLL and UnionKV-v2 15/15 at $N = 40$. Raw records: `results/fresh_specialized/review_structured_definitive/qwen35_9b_kv_retrieval_L32768.jsonl` (SHA-256 `145678a9...`) and the matching `number_string` file.

I.1 Count-first allocation check

A disjoint 20-example development audit (example IDs 50–69) sets the content count to $c = 32$, the 90th-percentile recorded worst-value NLL rank. We then evaluate example IDs 0–14 without changing c. The strict implementation selects exactly N positions and gives the calibrated content

candidates priority over structural candidates. A fixed-demand heuristic instead asks the legacy adaptive adapter for demand 32, but its span overfill and final scoring can displace requested content candidates.

Table 34: **Qwen3.5-9B count-first check at 32K** ($n = 15$ paired examples). Entries are accuracy counts. Literal retrieval is shown at absolute budget $N = 40$; structural columns are requested fractions. All compressed arms realize the exact requested cache length. The fixed-share v1 arm uses observation window 32, whereas the other three use window 8, so its differences cannot be attributed solely to allocation.

Policy	Literal $N = 40$	Struct. 1%	5%	10%	25%
Full cache	15	15	15	15	15
Strict count-first $c32$, $w8$	15	0	1	2	11
Fixed-demand heuristic $c32$, $w8$	15	0	0	0	1
Adaptive v2 $w8$	15	0	0	0	1
Fixed-share v1 $w32$	0	0	0	2	11

The strict arm preserves the 15/15 literal result at $N = 40$ while improving the structural result over adaptive v2 at 5–25%; it still reaches only 11/15 at 25%. Fixed-share v1 also reaches 11/15 at 25%, but fails the 40-position literal cell and first reaches 15/15 literal accuracy at $N = 80$. Because v1 changes the observation window as well as allocation, the equality at the largest structural budget is descriptive rather than an allocator-only comparison. The fixed-demand heuristic is identical to adaptive v2 on every displayed cell, so asking the legacy adapter for an absolute demand does not protect content candidates through final selection. The strict arm's content-membership guarantee follows from the registered selector implementation; the per-example files log exact realized cache lengths but not the full selected index sets.

The literal and structural raw files contain 435 and 255 rows, respectively, with one metadata record and no errors or duplicate identities. Their SHA-256 values are `f290a07e...` and `73f2be79...`; the disjoint calibration file SHA-256 is `31bc85d8....` The fail-closed derived registry is `results/analysis/count_first.json` (SHA-256 `432957cf...`), generated by `src/analyze_count_first.py` (SHA-256 `fc49e66d...`). These results cover one model, one prompt construction per task, and one calibrated count.

J NATIVE-PIPELINE BASELINES: KVZIP AND ADAKV

Matched protocol. The native campaign embeds the query before selection for all three methods and uses the same prompts, decoding rule, and $n = 15$ examples at approximately 32K. Budgets are absolute logical retained slots per head and layer: 32, 64, 128, 256, 512, and about 1024. SnapKV physically gathers K/V tensors, so its physical sequence length equals its logical budget. KVzip and AdaKV use decode-effective masks: their tensor sequence length remains full, so they free no cache memory, and the artifact logs both mean logical retained slots and total masked slots. Each completed cohort has exactly 285 rows, zero recorded errors, valid 15/15 full-cache accuracy, and realized logical budgets matching the request. These rows are not pooled with the historical raw-core protocol.

Results. All four registered cohorts are complete. KVzip is strong on the two literal cohorts, including 15/15 on the dense model at 32 logical slots. On hybrid structural retrieval it remains 0/15 through about 1025 slots, while SnapKV reaches 7/15 and AdaKV 5/15 at that budget. On dense structural retrieval, whose recorded inputs contain 32,805–32,808 tokens, KVzip is 0/15 at all six budgets; SnapKV scores 1/15 at 256 and 512 and 0/15 elsewhere, while AdaKV scores 1/15 at 256, 512, and about 1025. The masked KVzip and AdaKV tensors keep about 32.8K entries each. These low accuracies hold up to the tested ceiling of 3.125%. The derived registry is `results/analysis/native_masked.json` (SHA-256 `104eb1f9...`); the dense structural raw and manifest SHA-256 prefixes are `de127201` and `bb8d6c37`.

Table 35: **Native-pipeline accuracy counts by logical budget** ($n = 15$). Columns are budgets 32/64/128/256/512/∼1024. KVzip and AdaKV mask logical slots while retaining full-sized physical tensors; SnapKV physically gathers.

Model	Task	Method	32	64	128	256	512	∼1024
Qwen3.5-9B	literal	KVzip	6	14	15	15	15	15
		SnapKV	0	0	0	11	14	15
		AdaKV	0	0	0	12	15	15
Qwen2.5-7B	literal	KVzip	15	15	15	15	15	15
		SnapKV	0	0	0	4	6	7
		AdaKV	0	0	0	4	3	8
Qwen3.5-9B	structural	KVzip	0	0	0	0	0	0
		SnapKV	0	0	0	0	1	7
		AdaKV	0	0	0	0	1	5
Qwen2.5-7B	structural	KVzip	0	0	0	0	0	0
		SnapKV	0	0	0	1	1	0
		AdaKV	0	0	0	1	1	1

J.1 Community cold-compression diagnostic

The paper-linked community implementation of a FastGen-inspired cold-compression baseline uses custom per-head policies; the FastGen authors' repository does not provide an implementation. We therefore name this result as a community implementation diagnostic rather than an authors-exact FastGen comparison. The runner retains the implementation's manual profiling probabilities, but uses the model's SDPA output for the forward pass and applies policies in a stable order. Each task has the same 20 paired 16K prompts in a full-cache arm and in seven recovery settings. Both full-cache arms score 20/20.

Table 36: **Community cold-compression results.** Accuracy is normalized answer containment. "Logical" is the median retained head–token-entry ratio relative to the paired full cache; it is not a physical-memory ratio. Every arm retains the same preallocated cache bytes as its full control.

	`number_string`		`kv_retrieval`	
Recovery	Accuracy	Logical	Accuracy	Logical
0.500	0/20	0.103	0/20	0.101
0.700	0/20	0.159	0/20	0.110
0.850	4/20	0.284	0/20	0.181
0.900	17/20	0.377	0/20	0.242
0.950	20/20	0.608	0/20	0.446
0.990	20/20	0.968	20/20	0.864
0.999	20/20	1.000	20/20	0.993

The two tasks separate sharply within this implementation: `number_string` first matches the full score at recovery 0.95, with median logical occupancy 0.608, while `kv_retrieval` remains 0/20 there and reaches 20/20 only at recovery 0.99, with median logical occupancy 0.864. These are seven measured settings, not estimates of a continuous threshold. A natural 0.999 KV arm is not an all-full integrity gate: its generated token IDs equal the paired full arm for 19/20 prompts even though both score 20/20. A separate forced-all-full control on the same 20 prompts matches the full arm exactly in generated IDs, first-token probabilities, prefill telemetry, and final telemetry.

The community runtime preallocates full-size K/V tensors. Median allocated cache storage remains 941,817,856 bytes for every `number_string` arm and 943,652,864 bytes for every `kv_retrieval` arm. Thus Table 36 measures logical policy occupancy and answer behavior, not physical memory reclamation. Its instrumented times include Python profiling, CPU mask construction, and telemetry and are not reported as latency or TTFT. Cohorts run with the implementation's original profiling forward path, which fails a full-equivalence integrity check, or with a 32-token KV horizon, which truncates UUID answers, are excluded. The two raw files contain 160 result rows each with no errors; their SHA-256 prefixes are `74cf7c0a` and `3c97f5b7`. The fail-closed summary is `results/fresh_specialized/cold_compress_corrected_summary.json` (SHA-256 `7a49430f...`), generated by `src/analyze_cold_corrected.py` (SHA-256 `d7139829...`).

K PHYSICAL POST-PREFILL CACHE EVICTION

Protocol. We run the systems experiment as an in-process Hugging Face cache intervention on one NVIDIA B300. Each arm prefills the complete 32,768-token prompt into a `DynamicCache`; no arm reconstructs or truncates the prompt. The selector reserves 16 head and 64 tail positions, then adds NLL-ranked seeds with ± 8 span dilation until the logical budget is exactly 256. The selector never reads the needle location. The evicted arm gathers the selected, already-computed K/V rows into new contiguous tensors and decodes with the retained keys' original RoPE transforms and explicit original absolute positions. The full and evicted arms use the same 20 prompts, greedy decoding, and 32-output horizon.

We report two physical granularities. Qwen3.5-9B uses source blocks of 528 positions, matching the historical vLLM hybrid block size; Qwen2.5-7B-Instruct-1M uses 16-position source blocks. Before gathering, every selected source block is expanded to all prompt positions that actually exist in that block. This is a block-constrained physical compaction in HF: the retained K/V tensors really shrink, but they are stored contiguously rather than through a vLLM block table. It therefore measures live cache tensor storage and copy cost, not native vLLM allocation or serving concurrency.

Table 37: **Paired 32K full-prefill measurements** (n=20 prompts per arm). Cache bytes count unique live tensor storages owned by the HF cache immediately after the intervention. "Match" is literal expected-answer substring match. The checked-in `summarize_systems_results.py` generates all displayed values from the contemporaneous current-source V2 records under `results/validation/systems_current_source_v2`; ranges are across prompts.

Model	Arm	Logical	Source blocks	Physical entries	Cache (MiB)	Match
Qwen3.5-9B hybrid	full	32,768	63	32,768	1,049.5	20/20
	evicted	256	5	2,144	92.5	20/20
Qwen2.5-7B dense	full	32,768	2,048	32,768	1,792.0	20/20
	evicted	256	20 [20,21]	320 [320,336]	17.5 [17.5,18.4]	20/20

Measured storage and architectural floor. On Qwen3.5-9B, all 20 selectors touch source blocks 0, 30, 31, 61, and 62. The last block contains only the final 32 prompt entries, hence $4\times528+32 = 2{,}144$ physically retained entries. Live cache storage falls from 1,100,480,512 to 96,993,280 bytes ($11.35\times$). The cache has eight full-attention layers, which are physically gathered, and 24 linear-attention layers. Hash, shape, dtype, device, and data-pointer checks show that all 48 recurrent/conv fields remain bitwise identical in the same storage. Those fields occupy 26,738,688 bytes (25.5 MiB) in this HF runtime and form a non-evictable floor. This floor is specific to the HF runtime and differs from the historical vLLM layout.

The dense Qwen2.5 model has 28 full-attention layers and no recurrent state. Its median 20 retained source blocks correspond to 320 K/V entries, reducing live cache storage from 1,879,048,192 to 18,350,080 bytes ($102.4\times$). The dense-model reduction thus holds as a post-prefill tensor intervention. Both arms pay for the complete prefill, so no comparable latency reduction follows.

Table 38: **Synchronized stage medians in seconds** from the paired runs. "Compute sum" is prefill + selector + physical gather through availability of the first output token. It excludes cache hashing, allocator snapshots, Python garbage collection, and reporting, so it is an instrumented compute-stage sum rather than client-observed TTFT. "Decode" is the following 31 one-token steps at a common horizon.

Model	Arm	Full prefill	Selector	Gather	Compute sum	Decode
Qwen3.5-9B	full	0.697	0	0	0.697	1.039
	evicted	0.689	0.130	0.125	0.948	1.040
Qwen2.5-7B	full	0.567	0	0	0.567	0.251
	evicted	0.554	0.091	0.096	0.748	0.253

The observed median times for the 31 following decode steps differ by $+0.001$ s on the hybrid model and $+0.002$ s on the dense model (evicted minus full); we make no inferential latency claim from

these 20 paired prompts. Selection and gathering add approximately 0.26 s on the hybrid model and 0.19 s on the dense model. The experiment measures cache storage and answer quality, not client TTFT, queueing, throughput, preemption, admission, or concurrent capacity.

These measurements come from one current-source V2 execution. The V2 protocol was frozen before execution to require unchanged physical-cache invariants, retrieval scores, and actual K/V growth, with historical strings kept separately. Against older saved records, three hybrid continuation texts differ, so a stricter V1 historical-text equality check fails; two further independent current-source loads agree on every tested identity and generated token sequence, and the cause of the historical difference is unknown. The derived summary is `results/analysis/systems_current_source_v2.json` (SHA-256 `ad8cc64e...`); dense, hybrid, and horizon raw SHA-256 prefixes are `d86df44c`, `56758a78`, and `2ff27e13`.

K.1 Output horizon and page arithmetic

The first emitted output is read from the final prefill logits and does not yet add a decode K/V entry. After H emitted outputs, the loop has materialized $H - 1$ decode entries. For an ideal compacted 256-entry prompt and a 528-entry allocator block, the analytic page count is therefore

$$P(H) = \left\lceil \frac{256 + (H - 1)}{528} \right\rceil .$$

This equation is a page-allocation projection. The HF measurement below uses token-contiguous tensors and grows continuously; it does not allocate vLLM pages.

Table 39: **Qwen3.5-9B decode-horizon probe after real 256-entry token compaction.** EOS is ignored to hold the requested horizon fixed. "Projected pages" applies the equation above; live bytes are measured HF cache tensor storage. Source: `qwen35_9b_32k_decode_horizons.json`.

Emitted outputs	Materialized decode K/V	Total attention entries	Projected pages	Live cache bytes
272	271	527	1	44,007,424
273	272	528	1	44,040,192
274	273	529	2	44,072,960
512	511	767	2	51,871,744

Each materialized Qwen3.5 decode entry adds exactly 32,768 bytes across the eight full-attention layers. The first two-page projected state is thus 274 emitted outputs, not 273. A server that reserves the next output slot ahead of materialization could cross earlier.

Architecture compatibility audit. We separately ran the same physical-gather primitive at 2K $\rightarrow$ 256 entries on 14 exact checkpoint revisions spanning dense attention, sliding-window attention, hybrid recurrent/attention caches, shared-KV layouts, and mixture-of-experts models. All 14 runs completed an original-position decode step; in every run all and only eligible full-attention K/V tensors shrank, while every non-K/V cache field retained the same shape, dtype, value hash, and storage pointer. The audit establishes compatibility of the intervention with those runtime cache layouts; it does not test selector quality, long-context stability, or throughput. Exact per-model layer classifications, hashes, and receipts are in `cache_architecture_audit_14models.jsonl` and its completion record.

K.2 Scope, engine feasibility, and historical provenance

The measured intervention gathers full-length attention K/V and preserves every non-K/V cache tensor. This is sufficient for the dense model and for the full-attention layers of the tested hybrid model. It assumes the Transformers cache representation exposes sequence-major K/V tensors, that retained keys keep their prefill-time positional transform, and that generated queries can be assigned their original absolute coordinates. The unique-storage byte count covers cache-owned tensors only; it is not total process HBM. The single-request protocol cannot be converted into a capacity claim by dividing an HBM pool by these bytes.

The audited vLLM 0.25.0 source has no supported operation that removes arbitrary interior blocks from a live request. A native experiment would require an engine patch that compacts or sparsely

addresses selected attention blocks, updates the active request's block table, returns the old blocks to the allocator, preserves the recurrent-state group, preserves logical positions, and invalidates affected prefix hashes. Only allocator telemetry after that operation could support concurrent-capacity and serving-latency claims.

Historical prompt-reconstruction serving. The archived vLLM 0.25.0 experiments submitted a newly reconstructed 256-token prompt to the server. They therefore removed almost all 32K prefill work, changed RoPE distances, and rebuilt Qwen3.5's recurrent state from the shortened sequence. Their throughput, TTFT, queue, and 66 MiB/888-request projections describe that reconstructed-prompt protocol plus analytic page arithmetic; they are not measurements of post-prefill eviction. The old tables and raw logs remain archived for provenance under `provenance/historical_results/serving_capacity_20260713` and `provenance/historical_results/serving_e2e_20260713`.

L SCORER DEFINITIONS AND LEARNED-PREDICTOR TRAINING

This appendix defines the historical two-type taxonomy used by the archived experiments and specifies the learned content predictor (`cheap_pred`) and its training.

Type I and Type II importance. The historical taxonomy distinguishes tokens flagged by their own content statistics from tokens selected through query–key attention structure. The experiments support an asymmetric functional distinction on two synthetic tasks, not a pure structural pathway: the historical rescue ran below the structural task's operating budget, the 5% attention rescue retains nearly the whole annotated answer span, the selected sets overlap, and the design does not manipulate latent mechanisms independently.

`cheap_pred`: inference features. At inference, four features per prompt token come from a single forward pass without backpropagation: (1) hidden-state activation norm at layer fraction 0.75; (2) token NLL under the model itself; (3) centroid/manifold distance from the sequence mean in activation space; (4) normalized position. The auxiliary MLP that consumes them is trained separately, as specified next.

Training labels and model. The predictor is trained on token-level *causal* records: perturb one sampled token position by an equal-norm move toward the sequence mean and record the induced output-KL—local, future-mean, and total-mean variants, with future-mean the KV-relevant one. Records span seven models across four families (Qwen3.5-2B/4B/9B, gemma-4-E2B/E4B, phi-4, GLM-4.7-Flash) and 11 datasets; ordering verified stable across layer fractions 0.50–0.90 and perturbation scales. A small MLP over the four standardized features is fit to the causal labels: on file-held-out splits it nearly matches the gradient-norm teacher on local labels (0.689/0.375 vs. 0.749/0.415 AUC/recall@keep-20%), exceeds it on future-mean (0.757/0.486 vs. 0.667/0.352), and matches total-effect AUC with higher recall. Generalization (future-mean): file-held-out 0.772/0.502, model-held-out 0.846/0.592, dataset-held-out 0.748/0.469. Ablations: NLL+position reaches 0.738 AUC, the three non-position features 0.643, position alone is *anti*-predictive at 0.308, so the predictor is not a recency heuristic. For its historical grid cells, eviction is top-N under the predicted score. Qwen2.5-1M, Mistral-Nemo, and Llama-3.1 are model-held-out for the predictor, so its round-2 passes on Llama are also a transfer result—read against its round-3 distractor fragility (Appendix E).

Predictor specification and reproducibility limit. The archived checkpoint is a $4 \to 16 \to 1$ MLP with GELU between its two affine layers. Inputs are `activation_norm_z`, `nll_z`, `manifold_z`, and `position_frac`, standardized again by the training-set mean and standard deviation stored in the checkpoint (standard deviations clamped below at 10^{-6}). The archived filename and current training-source defaults indicate an 80/20 file-group split with Python seed 47 and 350 full-batch epochs. No contemporaneous record of the exact historical invocation exists, so the optimizer trajectory is not bit-reproducible. The packaged future-mean corpus has 724,652 rows (578,698 train, 145,954 test; positive rates 0.2077/0.2079). A rerun of the stated configuration reproduced the split, normalization statistics, and row counts exactly and obtained AUC 0.770654 and top-20%-recall 0.500939, close to the archived 0.772250/0.502192; its weights were not bit-identical (maximum absolute parameter difference 0.35).

Predictor refit. The refit used AdamW with learning rate 3×10^{-3}, weight decay 10^{-3}, and positive-class-weighted binary cross entropy with logits, without early stopping. Code: `src/legacy/train_causal_salience_predictor.py`; archived metadata and weights: `data/predictor/cheap_future_file_split_full350_seed47.json` and its paired `.pt` file (SHA-256 `2801aecf...` and `46081af4...`); refit records: `results/predictor_refit.json` and `results/predictor_refit_comparison.json`. The checkpoint loads with `torch.load(..., weights_only=True)`.

M TASK-SPECIFIC RETRIEVAL AND REASONING BOUNDARIES

Dataset and metric of the historical rows. The legacy RULER boundary run, originally labeled 128K, loaded the `4096` configuration; recorded examples contain 2,435–4,259 tokens. Its local scorer marks a row correct when the normalized prediction contains any reference answer. That differs from the official NVIDIA RULER aggregation, which averages reference recall for multi-answer niah/vt/cwe/fwe tasks and uses any-reference matching for QA. Historical CWE, FWE, and VT rows require 10, 3, and 5 answers respectively; the saved logs retain only the first five answers and the first 500 prediction characters, so official aggregate scores cannot be reconstructed reliably from those truncated outputs.

The RULER reruns use the pinned dataset revision, the `16384` configuration, and a checked official-metric wrapper. Historical synthetic retrieval and InfiniteBench scores remain normalized answer containment rather than full-output equality; LongBench QA uses token F1.

M.1 OFFICIAL-METRIC RULER BOUNDARIES

The evaluation contains three Qwen3.5-9B cohorts with 15 paired examples, nine requested fractions from 0.05% through 20%, and complete saved predictions and references. The official metric is mean reference recall for `niah_multikey_1` and any-reference matching for the QA tasks. The effective output horizons are 64 tokens for NIAH and 32 for QA. Table entries are requested fractions; realized logical cache lengths are validated separately and no physical-memory claim follows.

The artifact retrospectively reconstructs all 45 inputs and checks their lengths, truncation modes, references, output caps, and target positions against all 1,665 saved outcomes. It supplies exact prompt and token-ID hashes for reproduction. Because these hashes were computed after the runs, the check establishes consistency with the recorded input fields.

Table 40: **Official-metric RULER boundaries** (n=15/cell). Entries are the first half-full-baseline crossing in requested retained percentage; the smaller line is the adjacent measured-rung bracket. > 20 is right-censored and ≤ 0.05 is left-censored. "Invalid full" means the observed full baseline is below the registered 0.8 gate. Actual tokens are recorded tokenizer lengths.

Task	Actual tokens	Full (%)	NLL	TOVA	SnapKV	UnionKV-v2
				0.36	1.38	3.16
niah_multikey_1	15,982–15,994	100.0	> 20	[0.2,0.5]	[1,2]	[2,5]
			0.1		0.2	0.2
qa_1	14,183–16,181	80.0	[0.05,0.1]	≤ 0.05	[0.1,0.2]	[0.1,0.2]
qa_2	15,729–16,384	73.3	invalid full	invalid full	invalid full	invalid full

On `niah_multikey_1`, the full baseline is 15/15. NLL remains below half-baseline accuracy through 20%; TOVA, SnapKV-w32, and UnionKV-v2 cross at 0.360%, 1.38%, and 3.16%, in measured brackets 0.2–0.5%, 1–2%, and 2–5%. All 2,000 paired bootstrap resamples reproduce each disposition.

For `qa_1`, the full baseline is exactly 12/15, the registered gate. The point crossings are at or below 0.2%, but all four observed curves are nonmonotone and 699/2,000 paired resamples fail the full-baseline gate. The remaining dispositions are also mixed: NLL has 821 crossed, 329 left-censored, and 151 right-censored resamples; TOVA has only 8 crossed and 1,293 left-censored; SnapKV has 1,257 crossed and 44 left-censored; v2 has 972 crossed and 329 left-censored. The sub-percent point estimates therefore describe this sample and ladder rather than a stable calibrated boundary.

The qa_2 full baseline is 11/15 and fails the gate, so no compressed crossing is reported. Ten of its 15 prompts use the frozen head–tail–suffix truncation at the 16,384-token cap; five are untruncated. More generally, the first generated token is computed from the final full-prefill logits before subsequent generation reads the compressed cache. The raw rows do not save generated token IDs, so this experiment cannot quantify how much any short QA score was determined before compressed-cache reuse. The saved QA strings commonly begin with formatting or thinking text.

Each raw cohort contains 555 exact registered rows and no errors. Every score and reference was independently recomputed from the complete saved record. The fail-closed analysis is `results/analysis/ruler_boundaries.json` (SHA-256 `fad438d8...`), generated by `src/analyze_benchmark_boundaries.py` (SHA-256 `7001ad61...`); the generated table SHA-256 is `35a346f8....`

M.2 BABILONG BOUNDARY

We evaluated Qwen2.5-7B-Instruct-1M on the official 16K BABILong QA1–QA3 sources, using 20 paired examples per task, a 20-token output horizon, and requested fractions of 1%, 5%, 10%, and 25%. Recorded prompt lengths are 14,226–16,373 tokens for QA1, 13,836–16,189 for QA2, and 14,514–15,955 for QA3. All 60 prompts re-encode exactly against the pinned source, none requires source truncation, and the official label scorer independently reproduces all 780 recorded scores. Table entries are normalized label accuracies, not literal-span retention rates or physical-memory measurements.

Table 41: **BABILong boundary** (n=20/task). Every compressed entry is the official-label success count at the requested retention fraction. Only QA1 passes the registered 0.8 full-cache gate; QA2 and QA3 are descriptive and do not define crossings.

Task	Policy	Full	1%	5%	10%	25%
QA1	NLL	18/20	0/20	0/20	0/20	4/20
QA1	SnapKV-w32	18/20	13/20	17/20	18/20	18/20
QA1	UnionKV-v2	18/20	1/20	6/20	12/20	17/20
QA2	NLL	12/20	0/20	0/20	0/20	1/20
QA2	SnapKV-w32	12/20	5/20	14/20	13/20	13/20
QA2	UnionKV-v2	12/20	0/20	0/20	5/20	9/20
QA3	NLL	3/20	0/20	0/20	0/20	3/20
QA3	SnapKV-w32	3/20	3/20	2/20	2/20	2/20
QA3	UnionKV-v2	3/20	0/20	2/20	4/20	3/20

QA1 has an 18/20 full baseline. SnapKV-w32 is already above half baseline at 1% (13/20), so its point boundary is left-censored at $\leq 1\%$; NLL is 4/20 at 25% and is right-censored at $> 25\%$. UnionKV-v2 crosses at 7.07%, between the measured 5% and 10% rungs. In 2,000 paired bootstrap resamples, 84 fail the full-baseline gate. Among the remainder, v2 crosses in all 1,916, with a conditional percentile interval of 4.26–11.40%. NLL has 24 crossings and 1,892 right-censored resamples; SnapKV has 40 crossings and 1,876 left-censored resamples. Conditional intervals for those rare numerical crossings do not replace the observed censoring dispositions.

QA2 and QA3 score only 12/20 and 3/20 with the full cache, below the competence gate. Their compressed counts are paired descriptions, including cases in which a compressed run corrects a full-cache error, and define no threshold. The first generated token comes from full-prefill logits before later tokens use the compressed cache. A separate exact full-cache replay records generated IDs and scores every prefix: the first token is "The" on QA1/QA2 and "Before" on QA3, and its official-label score is 0/20 for every task; the complete continuations exactly reproduce 18/20, 12/20, and 3/20. The observed successes therefore require later generated tokens, which read the compressed cache.

The analysis validates the exact 780-row arm grid, paired example IDs, source references, prompt construction, token counts, decode horizon, and uncapped saved strings. Its output is `results/analysis/babilong_boundary.json` (SHA-256 `fa825a6a...`), generated by `src/analyze_babilong_boundary.py` (SHA-256 `cbc1f38d...`); the generated table SHA-256 is `425bc1d0....` The 60-row first-token replay is `results/validation/babilong_first_token/qwen25_qa123_60.jsonl` (SHA-256 `cfbc766a...`), validated by `src/babilong_first_token_diagnostic.py` (SHA-256 `b6c1ff1b...`).

M.3 LONGBENCH BOUNDARY COHORTS

The LongBench evaluation contains all three registered Qwen3.5-9B cohorts, with 15 paired examples and nine requested fractions from 0.05% through 20%. Actual input lengths are 2,320–17,465 tokens for HotpotQA, 13,354–17,726 for MuSiQue, and 2,369–15,748 for 2WikiMQA. No input is source-truncated at the 32K cap. The frozen examples reduce the CLI output limit of 64 to an effective 52-token horizon. Scores use the registered legacy maximum-reference token F1, not an exact-answer metric or an official multi-reference aggregate.

Table 42: **LongBench boundary evaluation** (n=15/cell). Full is mean token F1. Every full baseline is below the registered 0.8 gate, so compressed curves are descriptive and no half-baseline crossing is reported. Actual tokens are recorded prompt lengths.

Task	Actual tokens	Full (%)	NLL	TOVA	SnapKV	UnionKV-v2
hotpotqa	2,320–17,465	33.5	invalid full	invalid full	invalid full	invalid full
musique	13,354–17,726	33.2	invalid full	invalid full	invalid full	invalid full
2wikimqa	2,369–15,748	29.8	invalid full	invalid full	invalid full	invalid full

Full-cache token F1 is 0.335 on HotpotQA, 0.332 on MuSiQue, and 0.298 on 2WikiMQA. Because all three baselines fail the competence gate, the analysis does not run crossing bootstrap inference or report an N_{50} for any method. All 1,665 registered rows are present and paired. Every saved prediction is shorter than the 500-character serialization cap, and all scores were independently recomputed from the pinned source references. These data describe low full-cache performance under the frozen sampling and horizon. The fail-closed analysis is `results/analysis/longbench_boundaries.json` (SHA-256 `33853ec3...`), generated by `src/analyze_benchmark_boundaries.py` (SHA-256 `7001ad61...`); the generated-table SHA-256 is `7b84881f....`

M.4 INFINITEBENCH BOUNDARY LADDERS

The boundary evaluation contains all 12 registered cohorts: four exact checkpoints crossed with InfiniteBench 10-digit, passkey, and KV-retrieval inputs. Every Qwen input has 32,768 tokens and every Phi-4 input has 16,384; each operating point uses the same 15 examples and nine requested retention fractions from 0.025% through 10%. Decode horizons come from the frozen examples and are 32 tokens for 10-digit, 26 for passkey, and 64 for KV retrieval. Scores are normalized answer containment. These are known-dataset boundary measurements, not prospective predictions.

The short 10-digit and passkey tasks cross below 1% in every valid displayed cell, but KV retrieval behaves differently. On Qwen3.5-9B, SnapKV-w32 crosses at 3.35% in the measured 2–5% bracket; paired example resampling gives a conditional 95% interval of $[2.37, 4.72]\%$, with a numerical crossing in 2,000/2,000 resamples. NLL and UnionKV-v2 remain above 10%, with all 2,000 resamples right-censored. On Qwen2.5-7B-Instruct-1M and Phi-4, all three KV-retrieval point estimates exceed 10%. SnapKV crosses in only 174/2,000 and 40/2,000 resamples, respectively, so its conditional interval is not reported as the uncertainty. The result is a model–task–policy boundary, not a common retained fraction.

Qwen3.5-4B KV retrieval fails the observed full-baseline gate at 8/15 and therefore has no compressed point estimate or bootstrap interval. Under the frozen 64-token decode protocol, the seven saved misses are unfinished `<think>` continuations without the UUID; the eight successes emit the UUID first. No longer-horizon decode was run. The fail-closed analysis is `results/analysis/real_tasks.json` (SHA-256 `91c1bbda...`); its 12 raw and manifest hashes, every measured curve, realized mean cache length, decode horizon, bootstrap censor count, and full-baseline gate are included. Analyzer SHA-256 is `6790162a...`; generated-table SHA-256 is `052873d3....`

The failure-mode ladder. Representative outputs on synthetic `number_string` (ground truth `8377530566`): *full span* (NLL at 131 retained tokens, 128K): "The special number is 8377530566."; *single-digit corruption* (SnapKV-w32 at $\mathrm{kf}=10^{-3}$): "8377530666"; *prefix truncation* (TOVA at 13 retained tokens: "837."; at 26: near-complete digits with deletions; at 65: perfect); *total miss* (span evicted): fluent confabulation ("The special number is 123456789.") with no marker of failure. In

Table 43: **InfiniteBench boundary evaluation** (n=15/cell). Entries are the first half-full-baseline crossing in requested retained fraction, expressed as a percentage; the smaller line is the adjacent measured percentage-rung bracket, not a confidence interval. > 10 is right-censored at the largest tested fraction. "Actual tokens" are recorded input lengths, not nominal cap–fraction products or physical cache bytes.

Model	Task	Actual tokens	Full	NLL	SnapKV	UnionKV-v2
Qwen3.5-9B	10-digit	32,768	15/15	0.282 [0.2,0.5]	0.655 [0.5,1]	0.306 [0.2,0.5]
Qwen3.5-9B	Passkey	32,768	15/15	0.168 [0.1,0.2]	0.168 [0.1,0.2]	0.233 [0.2,0.5]
Qwen3.5-9B	KV retrieval	32,768	15/15	> 10	3.35 [2,5]	> 10
Qwen2.5-7B Instruct-1M	10-digit	32,768	15/15	0.178 [0.1,0.2]	0.805 [0.5,1]	0.233 [0.2,0.5]
Qwen2.5-7B Instruct-1M	Passkey	32,768	15/15	0.141 [0.1,0.2]	0.251 [0.2,0.5]	0.141 [0.1,0.2]
Qwen2.5-7B Instruct-1M	KV retrieval	32,768	15/15	> 10	> 10	> 10
Qwen3.5-4B	10-digit	32,768	15/15	0.316 [0.2,0.5]	0.677 [0.5,1]	0.306 [0.2,0.5]
Qwen3.5-4B	Passkey	32,768	15/15	0.306 [0.2,0.5]	0.178 [0.1,0.2]	0.316 [0.2,0.5]
Qwen3.5-4B	KV retrieval	32,768	8/15	invalid full	invalid full	invalid full
Phi-4	10-digit	16,384	15/15	0.268 [0.2,0.5]	0.374 [0.2,0.5]	0.16 [0.1,0.2]
Phi-4	Passkey	16,384	15/15	0.138 [0.1,0.2]	0.316 [0.2,0.5]	0.141 [0.1,0.2]
Phi-4	KV retrieval	16,384	15/15	> 10	> 10	> 10

the causal-intervention logs, evicting the top-14 NLL tokens produces the same intermediate rungs (Qwen2.5-1M "87530566"; Qwen3.5 "837056")—partial span coverage, not target loss.

	`niah_multikey_1`		`qa_1`		`qa_2`	
Method	0.1%	1%	0.1%	1%	0.1%	1%
Full context	1.00		0.93		0.67	
NLL	0.00	0.00	0.20	0.20	0.20	0.20
Cheap pred.	0.00	0.00	0.13	0.33	0.07	0.27
TOVA	0.00	0.07	0.33	0.20	0.07	0.00
SnapKV-w32	0.00	0.00	0.20	0.27	0.20	0.07
SnapKV-w256	0.00	0.07	0.13	0.20	0.13	0.13
InfoKV-style	0.00	0.00	0.13	0.47	0.13	0.07
UnionKV-v1	0.00	0.00	0.13	0.47	0.13	0.07
UnionKV-v2	0.00	0.00	0.13	0.27	0.13	0.27

Table 44: Historical RULER boundary rows, Qwen3.5-9B (n=15/cell), originally labeled 128K; the loader used config `4096` and recorded inputs contain 2,435–4,259 tokens. Scores use local normalized any-reference containment, not the official RULER aggregate, and the truncated logs cannot recompute complete multi-answer recall. These rows are neither 128K nor official-metric evidence.

Real-task shifts. On InfiniteBench `number_string`, measured NLL crossings at 128K are larger than on the synthetic template (61 vs. 41 on Qwen2.5-1M; 86 vs. 20 on Qwen3.5-9B). This is consistent with a larger competing-outlier load, but the comparison changes format and content together and does not identify the cause. Crossings require task-family-specific measurement.

N HISTORICAL HELD-OUT CALIBRATION AND 512K VERIFICATION

O REPRODUCIBILITY AND HISTORICAL COMMITMENTS

The artifact includes the available grids, per-cell JSONL outputs, rank-audit records, campaign runners, frozen scoring scripts, reanalysis code, and figure-generation code. The included files and

	hotpotqa 0.1%	hotpotqa 1%	musique 0.1%
Full context	0.34	0.23	0.34
NLL	0.03	0.03	0.02
TOVA	0.10	0.21	0.22
SnapKV-w32	0.01	0.26	0.07
InfoKV-style	0.07	0.26	0.03
UnionKV-v2	0.03	0.07	0.02

Table 45: Historical LongBench multi-hop rows labeled 128K (actual inputs shorter; the rerun in Appendix M records LongBench input lengths), Qwen3.5-9B (n=15 at 0.1%; n=7–8 at 1%). Full-context scores are low, so these cells cannot validate a rank-CDF approximation for diffuse-evidence tasks.

Model	Method	5×10^{-4}	10^{-3}	2×10^{-3}	5×10^{-3}	10^{-2}
GLM-4.7-Flash	Full context	1.00	1.00	1.00	1.00	1.00
	NLL	0.00	0.20	1.00	1.00	1.00
Mistral-Nemo-2407	Full context	1.00	1.00	1.00	1.00	1.00
	NLL	0.00	0.00	0.20	0.47	0.07
	TOVA	0.00	0.00	0.00	0.40	0.67
	SnapKV-w32	0.00	0.00	0.00	0.00	0.00
	UnionKV-v2	0.00	0.00	0.00	0.73	0.87

Table 46: Held-out 16K calibration curves (accuracy vs. keep-fraction, n=15/cell). GLM: content scorer only; NLL crossing gives $N_{50} \approx 21$. Mistral: NLL is non-monotone (peak 0.47, never crosses); UnionKV-v2 crosses cleanly while SnapKV-w32 scores 0.00 throughout.

SHA-256 manifests let readers verify internal consistency and reproduce the reported scoring. Commitment times and their ordering are author-recorded. Round-2 verification used torch 2.11.0+cu130, transformers 5.8.0, kvpress 0.4.0, and flash-attn 2.8.3 at $\geq$64K. Gemma-4 rows from the unsupported query reconstruction are retained and labeled.

- **Round 1** (length extrapolation): `PREREGISTRATION.md`, SHA-256 `c7998dc3c37eb09419dc2a726b92d3428f71560394ce1a3b66f2cefc3277140c`, frozen 2026-07-07T10:20:03Z, with the fit-constant snapshot and frozen scorer `score_prereg.py`.
- **Round 2** (held-out models): `PREREGISTRATION_R2.md`, SHA-256 `62690dc9f7c04077915da6c602bd14daf033179c660da18d4e163876ec4d7c95`, frozen 2026-07-12T10:08:44Z; scorer `score_prereg_r2.py`.
- **Round 3** (distractor falsification): `PREREGISTRATION_OUTLIER.md`, SHA-256 `69481cdb37f09af486107dbd367b997943a83ab59e998644152b37b032aa3cbf`, plus harness hash `87dfe93f...`, recorded freeze 2026-07-13T00:57:21Z; scorer `score_outlier.py`; raw cells in `results/outlier_experiment_20260713/`.
- **Round 4** (inflation prediction): `PREREGISTRATION_R4.md`, SHA-256 `5ca3a4c7a710e71c5c4a08044dfd8e02d3d66e9124387c892ce6ce66ac8680c3`, frozen 2026-07-13T06:15:38Z, with frozen predictions JSON, audit and verification manifests, rule code, and scorer (`score_inflation_r4.py`) all hash-covered in `PREREG_R4_HASH.txt`.
- **Crossover design**: `DESIGN.md`, SHA-256 `f02f6c02be25fd380cba946eef5bb6e620ca3d24e29473fec613e811a9eb523f`, recorded freeze 2026-07-13T03:47Z; the phi-4 replication's model-choice gate documented before analysis.
- **Round 5** (real-corpus prospective prediction): `PREREGISTRATION_REAL.md`, combined SHA-256 `dbf9182083d30bdbb75819018ff539b1efd58e63628d34a901d3b63ad4131b10`, frozen 2026-07-14T06:22:31Z (four cells); `PREREGISTRATION_REAL_STAGE2.md`, combined SHA-256 `4b816c55fade7fd31`

Method	2×10^{-5}	4×10^{-5}	6×10^{-5}	10^{-4}	2×10^{-4}
Full context	1.00	1.00	1.00	1.00	1.00
NLL	0.00	0.00	0.50	1.00	0.92
Cheap pred.	0.00	0.00	0.67	1.00	1.00
TOVA	0.00	0.00	0.00	0.11	0.25
SnapKV-w32	0.00	0.00	0.00	0.00	0.00
InfoKV-style	0.00	0.00	0.00	0.00	0.00
UnionKV-v2	0.00	0.00	0.00	1.00	1.00

Table 47: 512K verification curves (Qwen2.5-7B-1M, n=6–12/cell; crossings scored by the frozen estimator). Retained counts at these keep-fractions: 10, 21, 31, 52, 105 tokens. Content methods cross at 28–31 tokens, matching their $\leq$256K constants; TOVA's constant has inflated past 105 (the P1c miss); SnapKV-w32 (frozen > 400) is correctly predicted to score ≤ 0.10 everywhere.

`6b932be8d508d585a4c4fcecbbec798313b281b82fc70ea`, frozen 06:28:42Z (27B cell, before its first eviction row); frozen scorer `score_realcorpus.py`; the archive's no-prior-measurement scan, validity gates, audits, ladders, and verbatim scoreboard in `results/realcorpus_prereg_20260714/` (Appendix G).

- **Round 6** (annotated-evidence extensions): three author-recorded per-experiment freezes, each recorded as hashed before that experiment's first eviction row (2026-07-14): `PREREGISTRATION_E1.md`, SHA-256 `ee8eef4ccba9be53656128402efb73a34b7ba387aea042f0f74a38657bbc8565`, 20:59:01Z; `PREREGISTRATION_E2.md`, SHA-256 `b0311aa9f6f6fc810b9fc1e56f3ffb6e004b697da74c6c3086828fd81e3bb4c5`, 21:17:59Z; `PREREGISTRATION_E3.md`, SHA-256 `55b61d65544ad84e82f3756f0c26a8399888dd30ca12f6fbc0c520634fcd6d55`, 21:22:56Z (revision 2 of `136c8945...`: ladder grid only, byte-identical statistics and predictions, both pre-eviction); frozen predictions JSON, scorers `score_e{1,2,3}.py`, verbatim scoreboards, execution notes, and the manifest `ARTIFACT_SHA256.txt` in `results/extension_20260714/` (Appendix G.2).
- **Studies**: native baselines (`results/native_baselines_20260713/`), historical prompt-reconstruction serving (`results/serving_capacity_20260713/` and `results/serving_e2e_20260713/`), statistical reanalysis (`scripts/reanalysis_20260713/`), and rank audits (`results/audit_20260712/`). Historical figure scripts are `scripts/figs_2027/fig1_collapse.py` and `scripts/figs_2027/fig3_serving.py`. Current figures state their scope, rerun or historical, in the caption.