ERROR CERTIFICATES FOR KV-CACHE EVICTION VIA RANDOMIZED DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Deterministic KV-cache eviction keeps the top-k tokens under an importance score and deletes the rest, and after the deletion the serving system cannot know what the eviction cost it on the current query. We replace the deterministic tail with Poisson sampling at known inclusion probabilities, which makes the eviction error's variance identifiable and turns a survey-sampling variance estimator over the retained set into a per-step error certificate at one extra scalar per retained token. We prove that no estimator computable from the information a deterministic, value-blind scheme retains can bound its own eviction error: evicted values can be altered so that everything retained is unchanged while the true attention-output error grows without bound. The certificate covers the realized attention error in 96.9–97.7% of 12,096 replay cells, in 98.1–99.7% on twelve further architectures, and in every answer of a deployed streaming assistant. A pre-registered study on LongBench at 6k and 16k tokens (about 74,000 generations) finds that the certificate does not predict task failure and output log-probability does; on the same tasks the certificate separates eviction-induced from inherent failures better than confidence under either sign (post hoc; AUC 0.70–0.76 against 0.53–0.57); a cheap evicted-set summary predicts a deterministic evictor's own induced failures within 3 points of the certificate on SnapKV, without an absolute scale. Gating recomputation on the certificate gains little over random gating at the same retrieval rate on real conversations and beats it on synthetic ones, and its variance decomposition localizes the lost block far better than raw attention (fact block in the top three in 79–93% against 13–59%). Randomization buys a conservative, evictor-independent error estimate and a localizer, not prediction.

1 INTRODUCTION

Long-context inference stores a key–value pair per token per layer, so the KV cache grows linearly with context and soon dominates memory. The standard remedy is eviction: score each cached token by an importance proxy, keep the top k, and delete the rest permanently. A large literature refines the score (Zhang et al., 2023; Li et al., 2024; Xiao et al., 2024b; Cai et al., 2025; Feng et al., 2025; Zhang et al., 2024; Li et al., 2026), evaluated by quality at matched budgets.

We ask a question the score race leaves open: after eviction, can the serving system know what the eviction cost it on the current query? For deterministic, value-blind selection the answer is no. Top-k keeps a set that is a function of scores, so conditioned on what is retained the evicted values are unconstrained: altering them changes the true attention output arbitrarily while every retained key, value, score, and downstream statistic stays bit-for-bit identical (Theorem 1). Any self-diagnostic computed from the retained state reads the same whether the eviction was harmless or destroyed the answer. We call this *silent failure* and measure it at scale: for streaming evictors, entropy and margin are no better than chance on their own eviction-induced failures, and only summaries of the evicted set or of its low-score boundary see part of them. Figure 1 shows one such failure beside its certified counterpart.

The escape comes from survey statistics, not from a better score: probability selection, unlike purposive selection, can assess its own error (Neyman, 1934). If the tail of the cache is evicted by Poisson sampling with inclusion probabilities π_i that the algorithm itself chooses, the design is known and design-based inference applies (Horvitz & Thompson, 1952; Yates & Grundy, 1953; Särndal et al., 1992): an inverse-probability (Hájek) correction removes the selection bias of the

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.

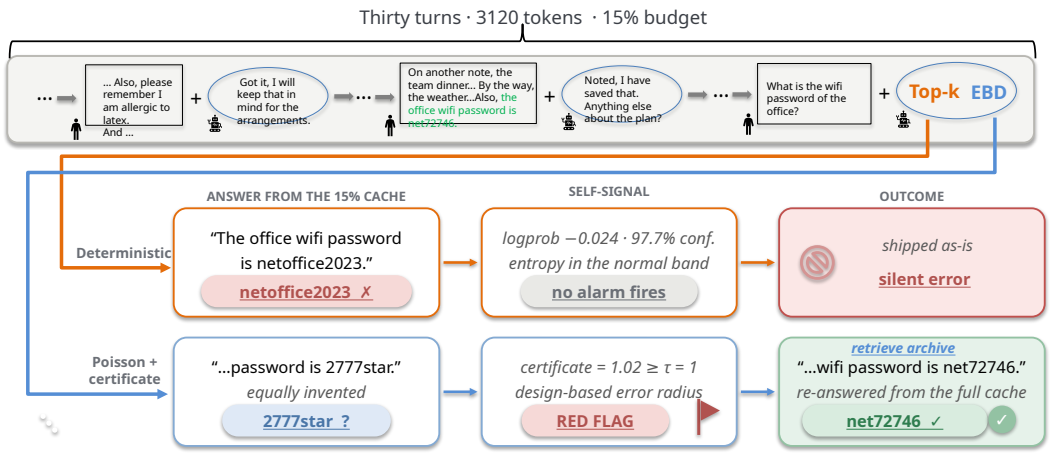

Figure 1: At a 15% budget, top-k invents the wifi password at 97.7% confidence while the certified design red-flags its own first answer and retrieves the truth (verbatim run, Qwen2.5-7B-Instruct, thirty-turn system of Section 7). Top: the timeline; turn 21 states the password (highlighted) and the question arrives ten turns later, after eviction under deterministic top-k and the certified Poisson design (EBD, eviction by design). Bottom: each design's answer, self-signal and outcome; aggregates in Table 12.

retained softmax, and a Horvitz–Thompson variance estimator, computable from the retained set alone, is first-order unbiased for the variance of the linearized error (Theorem 2). The correction is one logit offset, $\log(1/\pi_i)$, per retained tail token. An empirical-Bernstein radius (Howard et al., 2021; Waudby-Smith & Ramdas, 2024) turns the variance estimate into a per-step certificate, and an e-process construction (Ramdas et al., 2023; Vovk & Wang, 2021) extends it toward validity over the whole autoregressive trajectory (Section 3.4).

A pre-registered study, seven falsifiable claims with kill conditions committed before the runs, locates where the certificate pays (Section 6). Question-aware eviction at 25–50% budgets is nearly free on real tasks, output log-probability predicts failure better than any cache-side signal, and the damage regime lives in streaming and agent-memory settings, where history is compressed before the question is known. There the certificate answers what confidence cannot reliably answer, whose fault a failure is; a deterministic summary of the evicted set predicts induced failures nearly as well on SnapKV, and the certificate adds an absolute scale and, through its variance decomposition, a pointer to the lost block. Our contributions are the following.

- We prove that no estimator measurable with respect to the online information of a deterministic, value-blind eviction scheme can be accurate for its own attention-output error (Theorem 1), and realize it in a real model: permuting evicted values changes the true error of top-k by up to 31% while every online statistic stays bit-identical (Section 4).
- We give the design that makes the error's variance identifiable, certainty-plus-Poisson eviction with the Hájek correction as one logit offset and a first-order unbiased retained-set variance estimator (Theorem 2), whose per-step certificate covers the realized error in 96.9–97.7% of 12,096 replay cells at Spearman 0.94–0.98 and in 98.1–99.7% on twelve further architectures (Table 3).
- We measure silent failure on a panel of online self-signals over thousands of scored generations: entropy and margin of streaming evictors are no better than chance on their own eviction-induced failures, a query-time summary of the evicted set comes within 3 points of the certificate on SnapKV and 12 behind it on H2O, and the certificate reaches 0.67–0.75 (Section 6.3, Tables 10 and 19).
- A pre-registered study at 6k and 16k tokens (about 74,000 generations), agent-memory systems on five models from four families, real conversations, and a 32B replication locate the value and its limits: attribution at AUC 0.70–0.76 against 0.53–0.57 for confidence under its better sign, certificate-guided retrieval of 10% of the context at 0.70–0.87 against 0.18–0.28 for random blocks and 0.66–0.74 for a raw-attention ranking, and, on real conversations, gating little better than random gating at the same retrieval rate (Sections 6 and 7).

2 Related work

Deterministic eviction and compensation. H2O (Zhang et al., 2023), SnapKV (Li et al., 2024), TOVA (Oren et al., 2024), StreamingLLM (Xiao et al., 2024b), PyramidKV (Cai et al., 2025), and Ada-KV (Feng et al., 2025) select retained tokens deterministically from an importance proxy and differ in the proxy or the budget allocation; CAOTE (Goel et al., 2025) folds value information into the ranking, which improves which tokens go, not what the system knows afterwards (Remark 1). Merging methods such as CaM (Zhang et al., 2024) and MomentKV (Li et al., 2026) fold evicted mass into retained representatives, a post-hoc stratified ratio correction of the point estimate. None maintains an estimate of its own induced error; Theorem 1 covers value-blind designs, summaries that keep moments of the evicted set sit outside its premise, and Section 6.3 tests one. SparQ (Ribar et al., 2024) fetches keys selectively without eviction. The regime finding that question-aware eviction is nearly free and streaming eviction is not appears in SnapKV, SCBench (Li et al., 2025) and Yuan et al. (2024); our increment is the quantitative sweep from 6k to 125k tokens under an eviction-induced failure definition.

Randomized eviction and sparse attention with guarantees. MagicPIG (Chen et al., 2025) samples keys by locality-sensitive hashing and corrects with self-normalized importance sampling, a point estimator without a variance. Nexus sampling (Duong et al., 2026) evicts by reservoir sampling with an offline Horvitz–Thompson bound on retained utility, without an attention correction or an online error estimate. VASE (Chang et al., 2026) and Keyformer (Adnan et al., 2024) add stochasticity without probability bookkeeping. vAttention (Desai et al., 2026), HyperAttention and KDEformer (Han et al., 2024; Zandieh et al., 2023) bound sampled attention and Quest (Tang et al., 2024) selects pages, all over a resident cache, and offloading systems fetch from a host copy per query (Xiao et al., 2024a; Lee et al., 2024); a permanently evicted token returns only by such a transfer, and the certificate decides when one is due. A fixed-contract diagnostic (Zhang et al., 2026) studies when value-aware selection helps, without a randomized design. Concurrent work (Geh et al., 2026) formulates eviction as expectation estimation and corrects at decode time by self-normalized importance sampling, the same correction as our logit offset under a different sampling scheme; it estimates no error, gives no variance or coverage, and does not decide when to fall back.

Design-based inference and anytime validity. The estimator layer is classical, Horvitz–Thompson estimation, the Sen–Yates–Grundy variance, Poisson and rejective sampling, optimal allocation and the design effect (Horvitz & Thompson, 1952; Yates & Grundy, 1953; Hájek, 1964; Neyman, 1934; Kish, 1965; Särndal et al., 1992); the anytime layer is modern, confidence sequences, empirical-Bernstein bounds and e-values (Howard et al., 2021; Waudby-Smith & Ramdas, 2024; Ramdas et al., 2023; Vovk & Wang, 2021). Neither has been applied to permanent KV eviction.

3 Setup and theory

3.1 Objects

Fix one attention head at decode step t; everything extends by averaging over heads and layers. The cache is $\mathcal{C} = \{(k_i, v_i)\}_{i=1}^{n}$, the query is q_t, scores are $s_i = q_t^\top k_i/\sqrt{d}$, weights $a_i = e^{s_i}$, and the full-cache head output is the self-normalized mean

$$y_t \;=\; \frac{\sum_{i\in\mathcal{C}} a_i v_i}{\sum_{i\in\mathcal{C}} a_i}. \tag{1}$$

An *eviction mechanism* maps the prefill state to a retained set $\mathcal{S} \subseteq \mathcal{C}$; tokens outside $\mathcal{S}$ leave the device for good (a host copy costs a transfer; Section 7). The *online information* $\mathcal{F}_t$ available to the serving system at step t is the retained KV pairs, the design (any sampling probabilities used), the query history, and anything computed from them. Evicted values are not $\mathcal{F}_t$-measurable. The estimation target is the error $\|\hat{y}_t - y_t\|$ of whatever output $\hat{y}_t$ the system computes from $\mathcal{S}$.

3.2 Impossibility for deterministic eviction

Theorem 1 (Unidentifiability). *Let the eviction mechanism be deterministic and value-blind, meaning $\mathcal{S}$ is a function of the keys, scores, and query history only, and let it evict at least one token. Then*

for every $\mathcal{F}_t$-measurable estimator $\widehat{E}$ and every $M > 0$ there exist two cache configurations that generate identical online information $\mathcal{F}_t$ and whose true errors $\|\hat{y}_t - y_t\|$ differ by more than M. Consequently the eviction error is not identifiable from $\mathcal{F}_t$, uniformly over cache configurations.

Proof. Fix any configuration and let $\mathcal{E} = \mathcal{C} \setminus \mathcal{S}$ be the evicted set, which is nonempty whenever eviction occurs. Replace $\{v_i\}_{i\in\mathcal{E}}$ by $\{v_i + cu\}_{i\in\mathcal{E}}$ for a unit vector u and scalar c. Selection does not depend on values, so $\mathcal{S}$ is unchanged; retained pairs, scores, design, and query history are unchanged; hence $\mathcal{F}_t$ and with it $\widehat{E}$ and $\hat{y}_t$ are unchanged. The full output (1) shifts by $cu \cdot \sum_{i\in\mathcal{E}} a_i / \sum_{i\in\mathcal{C}} a_i$, nonzero and linear in c, so a large c makes the two errors differ by more than M. □

Remark 1 (Value-aware scores). If the score reads rotation-invariant summaries of each value, for instance its norm (Chang et al., 2026; Zhang et al., 2026), the construction survives in the bounded form of Appendix A: a common rotation of the evicted values preserves every summary while moving $\sum_{i\in\mathcal{E}} a_i v_i$ over a sphere whose radius the retained set cannot see.

3.3 Design: certainty plus Poisson tail, Hájek by logit offset

Randomized evictors keep no inclusion probabilities at decode time (Duong et al., 2026; Chang et al., 2026) or no variance layer (Chen et al., 2025); stored probabilities make a bias-corrected output, a variance estimator and a certificate computable from the retained set.

The mechanism keeps a *certainty layer* $\mathcal{A}$ (score top slice plus attention sinks and a recency window; $\pi_i = 1$) and applies to the tail $\mathcal{T} = \mathcal{C} \setminus \mathcal{A}$ independent Bernoulli retention with inclusion probabilities

$$\pi_i \;=\; \mathrm{clip}\left(m \cdot \frac{\mathrm{score}_i}{\sum_{j\in\mathcal{T}} \mathrm{score}_j},\; \varepsilon,\; 1 \right), \qquad i \in \mathcal{T}, \tag{2}$$

where m is the expected tail budget and $\varepsilon > 0$ a floor (Poisson sampling (Hájek, 1964)). The output on the retained set is the Hájek estimator

$$\hat{y}_t \;=\; \frac{\sum_{i\in\mathcal{S}} (a_i/\pi_i)\, v_i}{\sum_{i\in\mathcal{S}} a_i/\pi_i}, \tag{3}$$

which one line of code implements exactly: add $\log(1/\pi_i)$ to the retained logit and let the softmax denominator do the renormalization. No training, no new parameters, one extra scalar per retained tail token. The certainty layer consists of the protected positions (four attention sinks, thirty-two most recent tokens) and every token whose allocation in (2) hits the upper clip, the allocation being repeated with those tokens moved to $\mathcal{A}$ and m reduced by their count until none does (water-filling); $\varepsilon = 10^{-6}$ throughout.

Assumption 1 (Known design). *The probabilities π_i in (2) are chosen by the algorithm, stored, and bounded below by ε on the tail.*

Assumption 2 (Bounded weights). *Normalized weights are bounded: $a_i/(\pi_i \sum_{j\in\mathcal{C}} a_j) \le B/m$ for all $i \in \mathcal{T}$; the floor ε makes the bound finite and the certainty layer keeps B moderate. It is a regularity condition on decode–prefill alignment, not a guarantee (Appendix A).*

Nothing is assumed about the score: a bad one concentrates π on the wrong tokens and widens the certificate, but cannot bias the Horvitz–Thompson estimates, which rest on Assumption 1 alone.

Theorem 2 (Variance identifiability under Poisson design). *Write $N = \sum_{i\in\mathcal{C}} a_i$ and let $e_t^{\mathrm{lin}} = \frac{1}{N} \sum_{i\in\mathcal{T}} (\frac{I_i}{\pi_i} - 1)\, a_i (v_i - y_t)$ denote the first-order (linearized) error of (3), where I_i is the retention indicator, and write $\mathrm{Var}(e) = \mathbb{E}\|e - \mathbb{E}e\|^2$ for a vector e. Under Assumptions 1–2,*

$$\widehat{V}_t \;=\; \frac{1}{\widehat{N}^2} \sum_{i\in\mathcal{S}\cap\mathcal{T}} \frac{1-\pi_i}{\pi_i^2}\, a_i^2\, \|v_i - \hat{y}_t\|^2, \qquad \widehat{N} = \sum_{i\in\mathcal{S}} a_i/\pi_i,$$

satisfies $\mathbb{E}[\widehat{V}_t] = \mathrm{Var}(e_t^{\mathrm{lin}}) + O\big((B/m)^2 \max_{i\in\mathcal{T}} \|v_i - y_t\|^2\big)$, and $\widehat{V}_t$ is computable from the retained set alone. Under Poisson sampling the joint inclusion probabilities factor, so the Horvitz–Thompson variance estimator reduces to the single sum above, at cost $O(|\mathcal{S}\cap\mathcal{T}|)$ per head per step.

The proof (Appendix A) is the classical HT variance identity plus a Taylor expansion of the ratio; the $O(B^2/m^2)$ term is the price of plugging $\hat{y}_t$ and $\widehat{N}$ into the unknown y_t and N. The theorem concerns the linearized error of the per-layer attention output; Sections 5 to 7 measure its task-level meaning.

3.4 A CERTIFICATE WITH EMPIRICALLY VALIDATED COVERAGE

Per-step variance estimates become a running error certificate through the empirical-Bernstein construction. For step t define the radius

$$r_t = \frac{\sqrt{2\widehat{V}_t \log(1/\delta)} + b_t \log(1/\delta)}{\|\hat{y}_t\| + \epsilon_0}, \qquad b_t = \frac{\max_{i \in \mathcal{S} \cap \mathcal{T}} \sqrt{(1-\pi_i)/\pi_i^2}\, a_i \|v_i - \hat{y}_t\|}{\widehat{N}}, \tag{4}$$

the empirical-Bernstein bound shape (Howard et al., 2021; Waudby-Smith & Ramdas, 2024) instantiated with the design variance estimator of Theorem 2, a variance term plus a range term, both computable from the retained set, targeting the relative attention-output error at per-step confidence $1 - \delta$. The certificate in every task-level experiment is this radius at $\delta = 0.1$, averaged over four heads per layer and maximized over the first six decode steps; coverage is measured per unit in replay (Section 4) and in deployment, where the aggregate covers every answer and 97–99% of units (Table 14); Chebyshev's inequality with the true design variance gives a provable radius (Appendix A). Confidence level and probed steps are design constants fixed once for all experiments: they scale the radius and the operating point of the label-free rule, not the certificate's failure ranking (Table 6). The anytime extension is an e-process construction (Remark 2, Appendix A).

4 THE ESTIMATOR WORKS WHERE IT IS DEFINED

The retained-set variance estimator covers the true attention error, ranks it, and costs no accuracy, on thirteen architectures. We replay prefills of Qwen2.5-1.5B (Qwen et al., 2024) offline, evict at budgets $\{12.5\%, 25\%, 50\%\}$, and compare against the full-cache output at 12,096 (layer, head, query) cells. At $\delta = 0.1$ the certificate covers the realized error in 96.9%, 97.2%, and 97.7% of cells at the three budgets, and its Spearman correlation with the true error is 0.943, 0.965, and 0.979. The median certificate is roughly three times the median realized error at the 25% budget (0.0997 against 0.0317): loose as an absolute bound, strong as a ranking signal. Median relative error at 25% is 0.0317 for Poisson with Hájek, 0.0447 for top-k and 0.2386 for uniform sampling, the same ordering at every budget. Theorem 1 is realized in the same model: on each of six prompts, twenty permutations of the evicted values change the true error of top-k by up to 31% while every retained-set statistic of top-k is bit-identical, and the Poisson design's certificate covers its realized error under the same 120 permutations. The same replay on twelve further architectures gives coverage 0.981–0.997 and Spearman 0.973–0.988, with Poisson+Hájek's median error below top-k's on every model and Poisson without the offset above it (Table 3).

5 FROM ATTENTION ERROR TO TASK FAILURE: SYNTHETIC SUITES

The certificate transfers from attention error to task failure. On passcode retrieval with generation-time eviction (Qwen2.5-1.5B, 252 runs), with the question available before compression (*aware*) H2O succeeds fully and Poisson at 88.5%, and with compression before the question arrives (*stream*) both collapse to zero at 25% and 50%. On its own failures H2O's retained entropy scores AUC 0.405 (chance, $n = 48$) while the certificate scores 0.812 ($n = 192$; $z = 10.2$). Across four base models (Qwen2.5-1.5B/7B, Llama-3.1-8B (Grattafiori et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023)) and four RULER-style tasks (Hsieh et al., 2024) at 640 scored generations per cell, the certificate–failure AUC, pooling the two conditions, exceeds 0.5 in 16 of 16 cells (mean 0.836, range 0.65–0.97); within a condition, where both outcomes occur, it does so in 12 of 14, variable tracking being the exception (Table 17), and thresholding it gives a risk–coverage knob: error 0.229 at 30% coverage against 0.575 at full coverage. The exploration tax concentrates in key–value retrieval at the 25% budget, where spreading π over twenty keys costs 39–44 points on three of four models, and in needle retrieval on Llama and phi-4 in the stream condition (up to 43 points, Table 9), and is smaller elsewhere (Table 17); in the stream condition of this base-model suite every importance-based method, ours included, collapses (mean accuracy at most 10%). Budget allocation and compensation do not change this: PyramidKV (Cai et al., 2025) and a CaM-style merge (Zhang et al., 2024) average 0.050 and 0.056 against 0.047 for SnapKV at the 25% budget (four instruct models, four tasks) and cannot flag their own failures, while the gated system answers 0.434; at 50% the flag goes dark and the tax leaves it at 0.125 against 0.21–0.22 (Table 9).

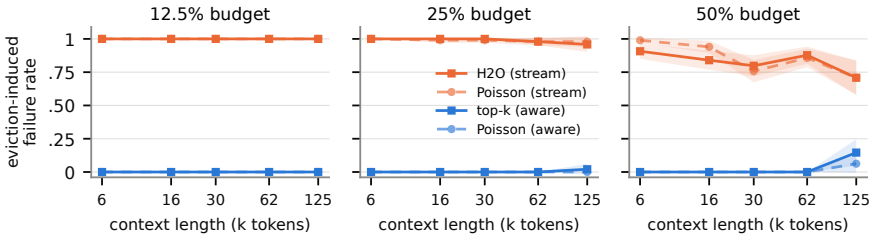

Figure 2: Streaming eviction is catastrophic at every context length from 6k to 125k tokens, and question-aware eviction is nearly free below 125k. Eviction-induced failure rate (among examples the full cache answers correctly) on exact-length synthetic retrieval, three budgets, YaRN beyond 32k; bands are 95% binomial intervals.

6 Pre-registered study on real workloads, at two scales

6.1 Design

Four instruction-tuned models (Qwen2.5-1.5B/7B-Instruct, Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3) run LongBench tasks (Bai et al., 2024) under a full grid: every example is answered by the full cache, by StreamingLLM-, H2O-, and SnapKV-style deterministic eviction, by question-aware top-k, and by Poisson eviction with the certificate (stream and aware; uncorrected softmax, re-run with the offset in Appendix C), at budgets $\{25\%, 50\%\}$ with 6k-token contexts (four LongBench tasks plus a synthetic needle anchor) and $\{12.5\%, 25\%, 50\%\}$ at 16k (HotpotQA, PassageRetrieval-en, MuSiQue, NarrativeQA), roughly 74,000 generations (Appendix C). Success is token-F1 ≥ 0.5 for QA and string match for retrieval and needle tasks; *eviction-induced* failure means the full-cache run answers correctly and the compressed run does not. Four claims with kill conditions (R1–R4) were fixed before any full shard ran and three scale hypotheses (S1–S3) before any 16k shard; Appendix D reproduces the registration and every verdict (Table 2).

6.2 Where eviction damages answers

Question-aware eviction is nearly free on real tasks, and the damage regime lives in streaming. At 16k, full-cache accuracy against Poisson at the 25% budget is 0.515 versus 0.516 on HotpotQA and 0.820 versus 0.815 on PassageRetrieval (Table 21); the eviction-induced failure rate in the aware condition, pooled over the four 16k tasks, is 6.2% at the 25% budget and 10.5% at the harshest 12.5% budget. Neither truncation (S1: 4.0% at 16k against 5.6% at 6k on shared tasks) nor memorization (context-dependent examples only: about one point; Appendix C) manufactures the finding.

Streaming eviction, where tokens are deleted before the question arrives, breaks 40–58% of answerable examples under H2O at the same budgets. On exact-length synthetic retrieval it destroys 70–100% at every context length from 6k to 125k tokens while the aware condition stays near zero below 125k (Figure 2); the full cache answers 97–100% up to 62k and 80% at 125k, the price of YaRN extrapolation, which conditioning on full-correct examples removes. At natural lengths of 19–23k the split is streaming 35–54% against aware 3–17%.

Two further pre-registered outcomes fix the division of labor. The randomized design costs nothing in the aware condition (top-k against Poisson within 0.7 points at 6k and 0.3 at 16k, S3) and 0–2.1 points against SnapKV in the stream condition (Table 16). And mean output log-probability predicts failure better than every cache-side signal, certificate included: overall failure 0.73–0.80 against 0.56–0.57, induced failure 0.782 against 0.778 at 6k (LongBench only) and 0.806 against 0.768 at 16k (paired difference -0.038, 95% CI $[-0.065, -0.010]$), and risk–coverage. The certificate's own pooled failure-prediction AUC on LongBench is 0.555 at 6k and 0.572 at 16k (R1, S2; Table 18) and 0.50–0.54 in the aware condition, because there is almost no eviction damage to see: failures on these tasks are overwhelmingly inherent. Selective answering should be gated on output confidence (Hendrycks & Gimpel, 2017; Geifman & El-Yaniv, 2017); a trust signal evaluated without this baseline overstates its case.

Table 1: Within each condition the certificate, which only a known randomized design provides, leads every deterministic self-signal, the strongest by five to six points, each scored on its own arm's eviction-induced failures; the streaming evictors' entropy and margin are no better than chance. Pre-registered 6k and 16k LongBench runs, 95% bootstrap intervals; Poisson is split by condition (pooled: 0.77).

self-signal	StreamingLLM	H2O	SnapKV	aware top-k	Poisson, stream (ours)	Poisson, aware (ours)
retained entropy	0.43 [0.40,0.46]	0.47 [0.44,0.50]	0.50 [0.47,0.53]	0.62 [0.57,0.68]	0.50 [0.47,0.52]	0.57 [0.53,0.61]
evicted score mass	–	0.67 [0.65,0.68]	0.67 [0.65,0.69]	0.59 [0.55,0.63]	–	–
keep-boundary margin	–	0.49 [0.49,0.50]	0.50 [0.49,0.50]	0.50 [0.49,0.50]	–	–
certificate	–	–	–	–	**0.72 [0.70,0.74]**	**0.68 [0.65,0.71]**

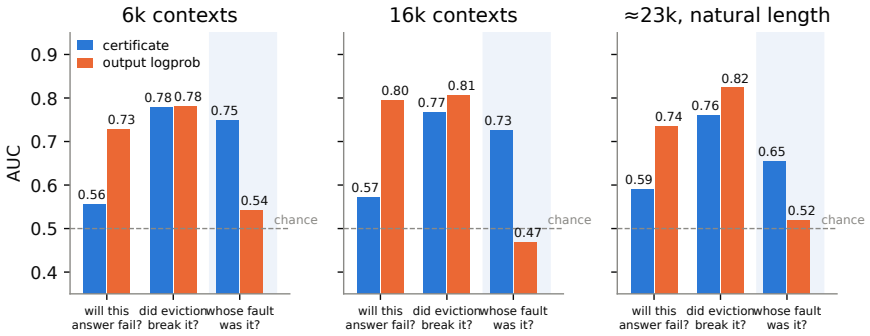

Figure 3: The certificate wins attribution at every scale while output confidence wins prediction. Three questions at 6k, 16k, and natural lengths near 23k tokens (MuSiQue and NarrativeQA), conditions pooled; the shaded column is attribution. Log-probability keeps its pre-registered sign (Table 11 uses its better sign).

6.3 The silent-failure panel

For streaming evictors, entropy and margin are no better than chance, mass-based signals see part of what the certificate sees, and a query-time summary of the evicted set sees most of it on SnapKV. Table 1 scores the pre-registered panel on each evictor's own induced failures, pooled over both scales: entropy 0.43–0.50 on StreamingLLM, H2O and SnapKV (0.62 for question-aware top-k), margin 0.49–0.50, evicted score mass 0.59–0.67 (0.59 for overall failure, Table 19). Since the theorem covers value-blind designs, we also gave H2O and SnapKV what a MomentKV-style method keeps, the count, mean key and mean value of the evicted set per kv head, and the query-time evicted-mass estimate $n_E \exp(q_t^\top \bar{k}_E/\sqrt{d})$ as a share (Table 10). Within cells it scores 0.645 on SnapKV against the certificate's 0.674 ($+3.0$, $[+0.1, +5.9]$) but 0.554 on H2O against 0.671 ($+11.7$), whose evicted set is too mixed for one mean key: the certificate leads on average, is evictor-independent and has a scale.

6.4 What survives: attribution, and scheduling

The certificate loses at predicting whether an answer is wrong and wins at a question output confidence cannot reliably answer: whose fault was it? Restricting to failures and asking each signal to separate eviction-induced from inherent ones, the certificate scores AUC 0.749 at 6k ($n = 7{,}924$ failures, 1,133 induced), 0.727 at 16k ($n = 11{,}717$, 1,776 induced), and 0.654 at the natural 23k-token lengths of MuSiQue and NarrativeQA ($n = 918$, 95 induced). Output log-probability, scored with whichever sign does better, reaches 0.542 and 0.531 at 6k and 16k (conditions pooled; Table 11 splits them), and its sign flips between suites and conditions: at 16k in the stream condition the induced failures are the more confident ones (Table 11), the pattern of Figure 1. The decline with length is mechanical: a fixed ratio leaves more resident tokens as context grows, so induced failures become rare (at the shared 25–50% budgets, 14% of failures at 6k, 11% at 16k and 8% at 23k) and borderline. The asymmetry is structural: confidence tracks difficulty whether or not the cache is at fault, while the certificate reads the sampling noise of the retained set and responds only to the compression channel. Triage against full-cache reruns confirms the verdicts: the tercile the certificate blames most recovers 2.4× (6k) and 1.35× (16k) as often as the one it exonerates (Appendix C).

A certificate-free attributor (two independent Poisson draws per query) reaches AUC 0.620–0.687 on label-assisted proxies, at double the cost and with no validity statement (Table 20).

Attribution converts into a scheduling product: under streaming compression, a quarter of queries may be re-run with the full cache. Under one global threshold, certificate gating captures 36% (31%) of the oracle's gain at the 12.5% (25%) budget, 1.7–1.8 times the gain of random gating and up to 2.1 times that of confidence gating (Figure 6, Table 20). At the 12.5% budget (16k) certificate gating recovers 0.33 against 0.30 for random and 0.28–0.29 for confidence gating ($+3.7$ points over random in each seed, intervals above $+2.3$); the lead over random holds at 25% and at 6k, with intervals excluding zero, and over four seeds (Appendix C). Confidence gating lands below random at the harsh budget because it spends re-runs on inherently hard examples; the certificate spends them on cache-damaged ones, which recomputation can save. At equal accuracy it re-runs 11.0–15.7% of queries where random gating needs 25%, and log-probability needs a per-task threshold to match (Table 7).

7 Deployment: certificate-gated agent memory

The rule $\mathrm{cert} \geq 1$ is the same everywhere and fires when the resident cache falls to a few hundred to a thousand tokens (225–308 at ten turns, 337–457 at thirty, 400–1,000 on real conversations); there the gated system answers where top-k fails, and so does any policy retrieving at the same rate: per-query selectivity is real on synthetic dialogues and small on real ones (Table 12).

Agent memory rots silently; the certificate is the smoke alarm. Fifty ten-turn dialogues embed one personal fact per user turn in natural chatter; after prefill (about 1,060 tokens) the cache is compressed to a 25% budget in the stream condition and four recall questions of varying age follow (Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct, 2,000 scored answers; Figure 5). The full cache answers 98.5%, so almost every compressed failure is eviction-induced. At the 25% budget facts one or two turns old survive at 46–61% under H2O and Poisson and facts three or more turns old at 17% or less, averaged over those ages, under every importance-based arm. On each arm's own failures the certificate reaches AUC 0.84 (Qwen) and 0.79 (Llama) against 0.72 and 0.66 for H2O's retained entropy, and where every entropy operating point needs labeled answers, the label-free rule $\mathrm{cert} \geq \tau = 1$ flags 99% of failures at a 49% false-alarm rate at the 15% budget on both models, and 54% of failures on Qwen and 28% on Llama at 25%, where the median certificate is near 1 (0.99, 0.91).

Thirty turns: the boundary tracks damage, not the ratio. Tripling the dialogue to thirty turns and thirty facts (3,160 tokens; fifty dialogues; six recall questions each) moves the boundary to a smaller ratio. At 25% (about 790 resident tokens) the median certificate reads 0.76 and the flag stays dark while both evictors collapse to 0.13–0.23: retrieval-type amnesia, which a per-step certificate does not measure. The heavy-damage regime returns below 15%: at 10% (about 320 resident tokens) the median certificate is 1.18, the flag fires on 95% of questions, and the gated system recovers 0.97 against 0.09 for top-k (AUC 0.82), 97% against 2% on facts stated 26 to 30 turns earlier. The boundary sits near 260 resident tokens for ten turns and 460 for thirty, not at a ratio.

The budget buys fewer retrievals, not accuracy. Sweeping the budget across the 10–30% grid (the fifty ten-turn dialogues; Figure 4), panel (a): gated accuracy falls from 98.5% at a 10% budget to 35% at 30% while raw Poisson and top-k stay at 11–19%, because the retrieval rate slides from 100% to 19%. Panel (b): the failure rate without retrieval barely moves (89% to 82%) while the gated system's silent-error rate (answers wrong and unflagged) rises from zero to 65%: at tight budgets failures are reconstruction distortion, which the certificate measures; at looser budgets they are retrieval-type amnesia, invisible to a per-step certificate by construction, while the ranking signal holds (failure AUC 0.92 to 0.80); Llama-3.1-8B reproduces every pattern. Panel (c): the median certificate crosses $\tau = 1$ at 25% on Qwen and 21% on Llama, where the red-flag rate passes one half; the exit from the heavy-damage regime is read off the certificate median (Table 5).

Real conversations. On LoCoMo (Maharana et al., 2024), ten real conversations of 11k–25k tokens, 699 questions per model, the full cache scores F1 0.297 (Qwen2.5-7B), 0.327 (Llama-3.1-8B) and 0.397 (Qwen2.5-32B). The heavy-damage regime sits at a resident cache of roughly 400–1,000 tokens, 2% of the history for Llama, 2–3% for Qwen2.5-7B and 5% for the 32B: there top-k keeps 0.06–0.09 and the gated system returns the full-cache score (0.297 on 7B, 0.316 on Llama, 0.390 on 32B),

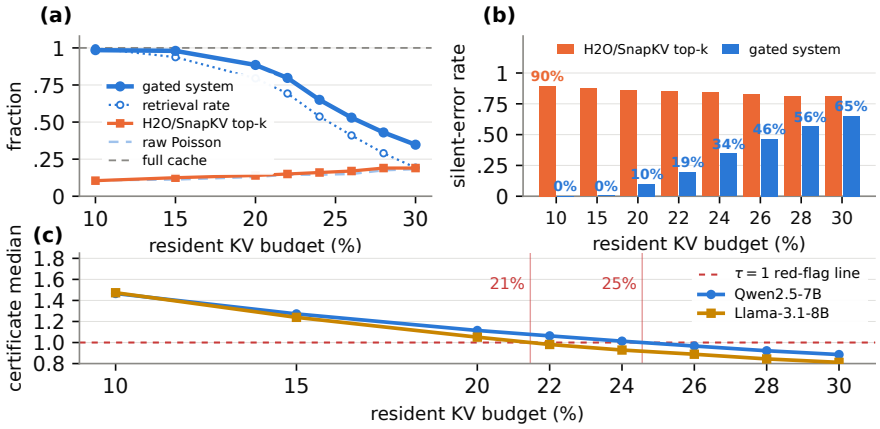

Figure 4: A larger budget buys the fraction of answers that need no retrieval, not accuracy, and the handover happens where the median certificate crosses $\tau = 1$. Merged 10–30% budget grid (fifty ten-turn dialogues, stream condition; Qwen2.5-7B in (a) and (b)): (a) gated accuracy and retrieval rate, with raw Poisson and top-k; (b) silent-error rate; (c) certificate medians for both models against $\tau = 1$.

because the flag fires on 96–100% of questions. That is the rate, not the selection: at the rule's own red-flag rate, random gating lands within 1.2 points of the certificate on every row, and at a matched 50% retrieval rate log-probability gating is level to 2.5 points better on Qwen and level on Llama (Table 12); the attribution edge over confidence is 0–11 points with overlapping intervals (Table 11). On real conversations the certificate supplies a label-free rate and little per-query selectivity.

Five models, four families, and a 32B replication. With the rule $\mathrm{cert} \geq 1$ fixed on every model, the ten-turn crossing sits at budgets of 21–29% but within 225–308 resident tokens, and at 10% every gated 7B–14B system answers 0.98–0.99 against 0.10–0.12 for top-k, as always retrieving would, since the flag fires on 98–100% of questions (Table 5). The 32B replication keeps the certificate's ranking within 0.10 of the 7B model (AUC 0.80/0.74/0.70 against 0.87/0.84/0.80 at 20/25/30%, Table 22).

Finer grains. Ranking evicted 64-token blocks by the variance contribution of their retained neighbors and retrieving 10% of the context recovers 0.70 accuracy on Qwen and 0.87 on Llama against 0.18–0.28 for random blocks (oracle 0.89–0.96). A ranking by the query's raw attention mass on the same retained neighbors, available to any evictor, recovers 0.66–0.74 but places the fact block in the top three in 13–59% against 79–93% for the variance ranking (Table 13): the red flag becomes a targeted retrieval, and the variance weighting aims it.

8 Discussion

Contribution and guidance. The estimator is classical; its online use under permanent eviction, the impossibility behind it and the measurements are new. Where the query is known before compression, eviction at 25–50% budgets is close to free; for answer-level trust, gate on confidence. In streaming and agent-memory settings the certificate is an attributor and localizer tied to the design; an evicted-set summary predicts induced failures nearly as well on SnapKV without a scale, and per-query gating pays clearly on synthetic dialogues and barely on real ones.

Limitations and future directions. Contexts reach 125k tokens, models 32B parameters and real tasks 25k; beyond these the trends are untested. The certificate concerns per-layer attention error, its task-level meaning is empirical, the anytime extension has two open ends (Appendix A), and in agent memory the gate's per-query selectivity is established only on synthetic dialogues. Budget ratios do not transfer across families (Table 5), the operating point follows δ and the probed steps (Table 6), and fixed-size designs (Hájek, 1964) are tested only in replay. The exploration tax is 0–2.1 points against SnapKV on LongBench and up to 44 on synthetic retrieval (Tables 9 and 17), and a 75% certainty layer recovers about half of the needle tax (Table 15). The unoptimized probe adds 0.2–0.3 s per answer and the unfused offset 12% to per-token latency; Appendix B lists the baselines and costs not measured.

ETHICS STATEMENT

This work studies the reliability of KV-cache compression in language-model serving. It involves no human subjects and no user data: the experiments use publicly released model checkpoints and public benchmarks (LongBench, LoCoMo) under their licenses, together with synthetic dialogues produced by the generators we release. The names, addresses, passwords and similar details in those dialogues are fabricated and correspond to no real person. The intended use is defensive: the certificate gives a serving system a conservative, evictor-independent estimate of the error its own compression introduced, and a red flag when that error is large, which reduces silent failures. Two limits of that framing are stated in the paper and we repeat them here. The certificate bounds the attention-output error, not task correctness, so it must not be read as a guarantee that an answer is right; Sections 6 and 7 report where it fails to predict task failure. The retrieval fallback trades computation for reliability, so raising the red-flag rate costs energy. The study used roughly 150 GPU-hours on single H100 or H200 GPUs. To support scrutiny, the seven main claims were pre-registered with kill conditions before the runs, the post hoc analyses are marked as such, and the logs, prompts and analysis scripts are released with the paper. We are not aware of any conflict of interest affecting this work.

REPRODUCIBILITY STATEMENT

All experiments run on single H100 or H200 GPUs (roughly 150 GPU-hours total). Per-run JSON logs (about 390,000 scored generations across the task suites, controls, vignettes and LoCoMo, and over 180,000 replay cells), the exact prompts, the analysis scripts that regenerate every number and figure from those logs, and the timestamped pre-registration files are packaged for release. Software, models and datasets are public (PyTorch 2.5.1, transformers 4.57.6; Qwen2.5, Qwen3, Llama-3.1, Mistral-7B, phi-4 and the base models of Table 3; LongBench and LoCoMo), and the synthetic generators are part of the release. Appendix C gives the harness, signal definitions, task sampling, scoring, and statistics; Appendix D reproduces the pre-registration record.

AI USE STATEMENT

We used generative AI throughout this project: Claude (Anthropic), mainly through the Claude Code agent.

Tasks with required disclosure. We used generative AI tools to help develop the conceptual framework, formulate mathematical claims, assist in writing the proofs, propose and refine hypotheses, design experiments and give feedback on the methodology, implement the method and the experiments, generate the synthetic dialogue datasets, reformat the LoCoMo and LongBench inputs into prompts, help interpret results, and assist with polishing the paper.

Tasks with recommended disclosure. We also used the LLM to suggest experimental parameters, write and edit table-generation scripts, build the demonstration library, polish the text of the paper and edit it for readability, and suggest candidate titles.

Review. The authors reviewed all AI-assisted work, including text, mathematics, code, analyses, figures and references. The attention offset and the certificate were tested against direct computation on reduced models, and the certificate's coverage was measured against full-cache computation (Section 4 and Table 14). Tables and figures report values computed by analysis scripts from the logged runs, and the numbers in the text are taken from those outputs. Every plot is drawn by a script from logged results; no figure or image was produced by an image-generation model and no plot was retouched. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Muhammad Adnan, Akhil Arunkumar, Gaurav Jain, Prashant J. Nair, Ilya Soloveychik, and Purushotham Kamath. Keyformer: KV cache reduction through key tokens selection for efficient generative inference. In *Proceedings of Machine Learning and Systems 6*, pp. 114–127, 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. doi: 10.18653/v1/2024.acl-long.172.

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.

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

Zhuoming Chen, Ranajoy Sadhukhan, Zihao Ye, Yang Zhou, Jianyu Zhang, Niklas Nolte, Yuandong Tian, Matthijs Douze, Léon Bottou, Zhihao Jia, and Beidi Chen. MagicPIG: LSH sampling for efficient LLM generation. In *International Conference on Learning Representations*, pp. 44169–44190, 2025.

Aditya Desai, Kumar Agrawal, Shuo Yang, Alejandro Cuadron Lafuente, Luis Gaspar Schroeder, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. vAttention: Verified sparse attention via sampling. In *International Conference on Learning Representations*, pp. 52253–52284, 2026.

Duc Duong, Hoang Anh Duy Le, Jianwen Xie, Anshumali Shrivastava, and Zhaozhuo Xu. Forget without compromise: Nexus sampling for streaming KV-cache eviction under fixed budgets. *arXiv preprint arXiv:2606.23961*, 2026.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S. Kevin Zhou. Ada-KV: Optimizing KV cache eviction by adaptive budget allocation for efficient LLM inference. In *Advances in Neural Information Processing Systems 38*, pp. 125462–125498, 2025. doi: 10.52202/085713-3773.

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*, pp. 22975–22988, 2024.

Renato Geh, Alex Chen, Daniel Israel, Aditya Grover, and Guy Van den Broeck. A probabilistic interpretation of KV cache eviction. *arXiv preprint arXiv:2608.28293*, 2026.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30*, 2017.

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

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jaroslav Hájek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523, 1964. doi: 10.1214/aoms/1177700375.

Insu Han, Rajesh Jayaram, Amin Karbasi, Vahab Mirrokni, David Woodruff, and Amir Zandieh. HyperAttention: Long-context attention in near-linear time. In *International Conference on Learning Representations*, pp. 39310–39325, 2024.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

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.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Wonbeom Lee, Jungi Lee, Junghwan Seo, and Jaewoong Sim. InfiniGen: Efficient generative inference of large language models with dynamic KV cache management. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pp. 155–172, 2024.

Yu Li, Binxu Li, and Tian Lan. MomentKV: Closing the directional gap in KV cache eviction for long-context inference. *arXiv preprint arXiv:2606.01563*, 2026.

Yucheng Li, Huiqiang Jiang, Qianhui Wu, Xufang Luo, Surin Ahn, Chengruidong Zhang, Amir 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*, pp. 66063–66093, 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 37*, pp. 22947–22970, 2024. doi: 10.52202/079017-0722.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–625, 1934. doi: 10.2307/2342192.

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. doi: 10.18653/v1/2024.emnlp-main.1043.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

Luka Ribar, Ivan Chelombiev, Luke Hudlass-Galley, Charlie Blake, Carlo Luschi, and Douglas Orr. SparQ attention: Bandwidth-efficient LLM inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42558–42583, 2024.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer, New York, 1992. doi: 10.1007/978-1-4612-4378-6.

Hanshi Sun, Li-Wen Chang, Wenlei Bao, Size Zheng, Ningxin Zheng, Xin Liu, Harry Dong, Yuejie Chi, and Beidi Chen. ShadowKV: KV cache in shadows for high-throughput long-context LLM inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57355–57373, 2025.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. QUEST: Query-aware sparsity for efficient long-context LLM inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47901–47911, 2024.

Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021. doi: 10.1214/20-AOS2020.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

Chaojun Xiao, Pengle Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. InfLLM: Training-free long-context extrapolation for LLMs with an efficient context memory. In *Advances in Neural Information Processing Systems 37*, pp. 119638–119661, 2024a. doi: 10.52202/079017-3801.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, pp. 21875–21895, 2024b.

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*, pp. 37228–37253, 2025.

F. Yates and P. M. Grundy. Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society: Series B (Methodological)*, 15(2): 253–261, 1953. doi: 10.1111/j.2517-6161.1953.tb00140.x.

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*, pp. 4623–4648, 2024. doi: 10.18653/v1/2024.findings-emnlp.266.

Amir Zandieh, Insu Han, Majid Daliri, and Amin Karbasi. KDEformer: Accelerating transformers via kernel density estimation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40605–40623, 2023.

Ruijie Zhang, Haozhe Liang, Da Chang, Li Hu, Fanqi Kong, Huaxiao Yin, and Yu Li. When does value-aware KV eviction help? A fixed-contract diagnostic for non-monotone cache compression. *arXiv preprint arXiv:2605.08234*, 2026.

Yuxin Zhang, Yuxuan Du, Gen Luo, Yunshan Zhong, Zhenyu Zhang, Shiwei Liu, and Rongrong Ji. CaM: Cache merging for memory-efficient LLMs inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58840–58850, 2024.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang "Atlas" Wang, and Beidi Chen. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems 36*, pp. 34661–34710, 2023. doi: 10.52202/075280-1506.

A PROOFS AND CONSTRUCTIONS

A.1 THEOREM 1 WITH BOUNDED VALUES

The proof shifts evicted values by cu with c unbounded. If values are norm-bounded, $\|v_i\| \le B_v$, replace the evicted values by $\|v_i\|u$ and, in a second configuration, by $-\|v_i\|u$ for a unit vector u: norms are preserved, every $\mathcal{F}_t$-measurable statistic is the same in both, and their full outputs (1) differ by $2\sum_{i\in\mathcal{E}} a_i\|v_i\| / \sum_{i\in\mathcal{C}} a_i$, at most $2B_v$ times the evicted attention mass. The error is then unidentifiable on an interval whose length is of that order and scales with the evicted mass rather than diverging, and no $\mathcal{F}_t$-measurable estimator can tell its two ends apart.

Scope of the theorem. The proof is short because the result is a statement about information, not a computation: it names what a serving system retains and shows an error that this information cannot bound. The argument is the KV-cache form of the distinction drawn by Neyman (1934) between purposive selection, which offers no internal basis for assessing its own error, and probability selection, which does; Theorem 2 is the second half of that distinction. The construction also applies to any single realization of the Poisson design: given the realized $\mathcal{S}$, shifting the evicted values leaves $\mathcal{F}_t$ unchanged. The identifiability that the known design restores is therefore that of the error's distribution over realizations, fixed by π, not of its realized value: the retained tail is a probability sample of the tail, and $\widehat{V}_t$ has the design variance as its expectation to first order. Every coverage statement in this paper is a statement over that distribution for a fixed cache configuration, which is what design-based inference means (Särndal et al., 1992); the empirical coverage figures average it over the configurations that the workloads produce.

A.2 THEOREM 2

Condition on the query q_t and the cache contents; randomness is only the Poisson indicators I_i, $i \in \mathcal{T}$, with known π_i (exact when q_t does not depend on the draw, as at the first decode step after it; later queries depend on it through the generated tokens, the second open end of the anytime construction below). Write $x_i = a_i(v_i - y_t)/N$ and note $\sum_{i\in\mathcal{C}} x_i = 0$ by (1), so the linearized error is $e_t^{\mathrm{lin}} = \sum_{i\in\mathcal{T}} (I_i/\pi_i - 1)\, x_i$. Independence across i gives

$$\mathrm{Var}\big(e_t^{\mathrm{lin}}\big) = \sum_{i\in\mathcal{T}} \frac{1-\pi_i}{\pi_i}\,\|x_i\|^2,$$

the Poisson-sampling case of the Horvitz–Thompson variance (Horvitz & Thompson, 1952; Särndal et al., 1992): joint inclusion probabilities factor, $\pi_{ij} = \pi_i\pi_j$, so the cross terms vanish. (The Sen–Yates–Grundy form (Yates & Grundy, 1953) presumes a fixed sample size and does not apply to a Poisson design.) The Horvitz–Thompson estimator of that population total,

$$\sum_{i\in\mathcal{S}\cap\mathcal{T}} \frac{1}{\pi_i}\cdot\frac{1-\pi_i}{\pi_i}\,\|x_i\|^2 = \sum_{i\in\mathcal{S}\cap\mathcal{T}} \frac{1-\pi_i}{\pi_i^2}\,\|x_i\|^2,$$

is unbiased for it (Horvitz & Thompson, 1952). $\widehat{V}_t$ replaces the unknown x_i by $a_i(v_i - \hat{y}_t)/\widehat{N}$. Under Assumption 2, $\mathrm{Var}(\widehat{N}/N) = \sum_{i\in\mathcal{T}}(1-\pi_i)\pi_i^{-1}(a_i/N)^2 \le B/m$ and likewise $\mathrm{Var}(e_t^{\mathrm{lin}}) \le (B/m)\max_i \|v_i - y_t\|^2$, so $\widehat{N}/N = 1 + O_p(\sqrt{B/m})$ and the first-order expansion of the Hájek ratio (Särndal et al., 1992) gives $\hat{y}_t - y_t = e_t^{\mathrm{lin}} + O_p(B/m)$ on the sample path. Substituting $\hat{y}_t$ and $\widehat{N}$ into the quadratic form perturbs $\widehat{V}_t$ by a relative $O_p(\sqrt{B/m})$; in expectation the cross terms between distinct units vanish, and each surviving term is bounded by B/m times the corresponding term of $\mathrm{Var}(e_t^{\mathrm{lin}})$, which gives the stated $O(B^2/m^2)$ remainder with values normalized to unit spread. □

On Assumption 2. The probabilities π_i are fixed at prefill from prefill scores, while a_i is the decode query's weight, so a query that attends heavily to a low-π tail token makes B large. The floor ε bounds $a_i/(\pi_i N)$ by $1/\varepsilon$, so the assumption always holds with some finite B; the certainty layer, which retains the head of the prefill score distribution with $\pi_i = 1$, is what keeps B moderate when decode queries align with prefill scores. The assumption is therefore a regularity condition on that alignment, and the coverage it yields is measured rather than proved (Sections 4 and 7).

A.3 THE ANYTIME CONSTRUCTION

Remark 2 (Anytime extension). The per-step radius (4) extends toward time-uniform validity by a standard route: an empirical-Bernstein e-process along decode steps within a head (Waudby-Smith & Ramdas, 2024), Ville's inequality (Ville, 1939; Howard et al., 2021) for the uniform-in-T statement, and arithmetic averaging of e-values across arbitrarily dependent heads, layers, and steps, which preserves validity (Vovk & Wang, 2021). When the tail indicators are redrawn at every step, coverage under this construction does not depend on when the certificate is read.

Fix a head and redraw the tail indicators at every step. The summands $\{(I_i/\pi_i - 1)\, x_i\}_{i\in\mathcal{T}}$ are then bounded by Assumption 2 and mean-zero given the past under the design, so the empirical-Bernstein construction of Waudby-Smith & Ramdas (2024) applied to the running sums yields a nonnegative supermartingale M_t with $\mathbb{E}[M_t] \le 1$; Ville's inequality (Ville, 1939) makes the resulting confidence sequence for the cumulative linearized error $\sum_{s\le t} e_s^{\mathrm{lin}}$ valid uniformly in t, and the boundary in (Howard et al., 2021) gives the stated radius shape (a variance term plus a range term, both computable from the retained set); a simultaneous per-step statement of the form (4) for all t needs in addition a union bound over steps or a per-step e-value. If the tail indicators are redrawn only every T_r steps, the increments are independent across blocks but share one draw within a block, and the later queries of a block depend on that draw, so the product runs over blocks only: the statement is then time-uniform across blocks but not within one, and with a single draw for the whole trajectory, as under permanent eviction, it reduces to a per-draw statement. For combination: if $E^{(1)}, \ldots, E^{(K)}$ are e-values for the per-(layer, head) nulls, arbitrarily dependent, then $\bar{E} = K^{-1}\sum_k E^{(k)}$ satisfies $\mathbb{E}[\bar{E}] \le 1$ under the intersection null, so thresholding $\bar{E}$ at $1/\delta$ is valid (Vovk & Wang, 2021); the linearization remainder of Theorem 2 adds the $O(B^2/m^2)$ slack. The two open ends are the reduction from the vector-valued linearized error to a scalar supermartingale, whose constants are not settled, and the reuse of one tail draw across steps, through which later queries depend on the draw, which the block structure handles conservatively. □

A.4 A PROVABLE RADIUS

With the design variance $V_t = \mathrm{Var}(e_t^{\mathrm{lin}})$ of Theorem 2, Chebyshev's inequality gives $P(\|e_t^{\mathrm{lin}}\| > \sqrt{V_t/\delta}) \le \delta$ under the design. The computable radius replaces V_t by the estimate $\widehat{V}_t$, unbiased to first order, a plug-in whose finite-sample validity is not proved (it is zero when no tail token survives, a case a deployment should treat as a red flag rather than as a clean bill; the expected tail size m kept it from occurring: none of the 183,532 certificates logged across our runs is zero); it is about $1/\sqrt{2\delta\log(1/\delta)}$ times the variance term of (4) (1.5× at $\delta = 0.1$) and has no range term, so it is narrower than (4) whenever the range term carries more than 32% of the radius, as in deployment (36–42%, Table 14). The empirical-Bernstein shape keeps the range term as a guard against a few heavily weighted tail tokens, and its coverage is measured (Section 4, Table 14). The range term b_t in (4) is likewise a sample quantity, the maximum over retained tail tokens; the population range is unobservable by Theorem 1, which is why (4) is an empirical bound shape with measured coverage rather than a distribution-free guarantee; the empirical-Bernstein inequality of Maurer & Pontil (2009) is proved for bounded i.i.d. terms with a finite-sample correction, whereas the Poisson tail terms are not identically distributed and (4) carries no such correction, and Appendix B states what the measured coverage does and does not establish.

A.5 WHERE EXISTING METHODS SIT

Remark 3 (Where existing methods sit). Every eviction scheme is a triple (design, estimator, variance handling). H2O-style top-k is a certainty-only design with a plug-in estimator and no variance layer; MagicPIG is sampling with self-normalized importance-sampling correction and no variance layer; Nexus and VASE are sampling designs without correction; CaM, MomentKV, and related merging methods are deterministic designs with a stratified ratio compensation of the point estimate. The certificate layer of this paper is orthogonal and can be attached to any known-probability design, including the LSH sampling probabilities of MagicPIG; we have not run that variant, and the paper's claim is for the variance and certificate layer, not for the sampling scheme underneath it. The stratified-ratio view suggests that merging error should scale with the within-stratum dispersion of evicted values rather than with evicted attention mass. We tested this in the replay framework with a

CaM-style merge, each evicted value folded into the retained token with the nearest key, weighted by attention mass: the merging error tracks evicted mass (Spearman 0.70–0.92 across four models and three budgets) and not dispersion (-0.07 to 0.10), because the merged representative inherits the retained key's score, so the score mismatch dominates the value dispersion. Merging removes under 8% of top-k's attention error, the Hájek-corrected Poisson design 15–50% at the same budgets (Table 8), and the merge produces no error estimate.

B SCOPE OF THE CLAIMS

This appendix collects in one place what each guarantee covers, which comparisons were not run, and the sample behind every number that appears in more than one place.

The certificate is conservative, not calibrated. At the nominal level $\delta = 0.1$ the radius (4) covers the realized attention error in 96.9–97.7% of replay cells, in 98.1–99.7% on the twelve further architectures and in 97–99% of units in deployment (Section 4, Tables 3 and 14), and its median is about three times the median realized error. The two facts say the same thing: the radius is a conservative envelope whose ordering of errors is faithful (Spearman 0.94–0.98 in replay, 0.77–0.93 in deployment), not a tight or calibrated bound. Its shape follows the empirical-Bernstein inequality, but its validity is measured rather than inherited (Appendix A): the range term is a sample maximum, the finite-sample correction of Maurer & Pontil (2009) is omitted, Theorem 2 holds up to an $O(B^2/m^2)$ remainder whose constant we do not track, and the e-process of Remark 2 is a construction that no experiment deploys. The deployed statistic is the mean over the probed units and the maximum over the first six decode steps of the per-unit radius, an aggregate that Theorem 2 does not cover. Table 6 shows that its failure ranking is stable across the implementation choices except max over units, while its operating point, the rate at which $\mathrm{cert} \geq 1$ fires, moves with δ and the number of probed steps. The choices $\delta = 0.1$, six probed steps and four evenly spaced heads per layer were fixed in the pre-registered LongBench suites before any vignette or LoCoMo run; the rule $\tau = 1$ was adopted after the first ten-turn vignette as the natural unit of a relative-error bound and held fixed for every later experiment (thirty turns, the budget sweep, the model families, LoCoMo, the 32B replication and the fine-grained probe). The certificate therefore fixes a red-flag rate without labels; it does not fix an operating point tuned to a task, and Section 7 reports where that rate is one (the heavy-damage regime) and where the flag goes dark.

Storage model. The theorems concern the state a serving system keeps on the device: $\mathcal{F}_t$ contains the retained KV pairs and nothing about the evicted values, which leave that state for good (Section 3). The deployment of Section 7 keeps the full KV cache, or the raw history, in host memory after prefill, as offloading systems do (Xiao et al., 2024a; Lee et al., 2024; Sun et al., 2025), and a red flag decides when to pay the host-to-device traffic or the re-prefill. Under that storage model other uses of the archive exist: per-query page retrieval (Tang et al., 2024), sampling over a resident cache with per-step guarantees (Desai et al., 2026), and CPU-side sampling (Chen et al., 2025). None of them was run as a baseline, and the end-to-end latency, throughput and transfer volume of the fallback were not measured; Section 7 measures whether the certificate selects the queries to re-run better than random or confidence gating at the same rate (Table 12), and on real conversations it gains at most 2.4 points over random and nothing over confidence. In the heavy-damage regime the flag fires on 95–100% of queries, so there the gated system is the always-retrieve system with a label-free trigger and carries its cost.

Relation to the closest methods. The paper claims neither the estimator nor the correction: the Horvitz–Thompson variance is classical, and the logit offset is the self-normalized importance-sampling correction that MagicPIG (Chen et al., 2025) and concurrent work (Geh et al., 2026) also apply. Against MagicPIG, the paper adds a variance layer and a certificate on a design that samples once after prefill. Against vAttention (Desai et al., 2026), which samples a resident cache at every step to meet an error target, here the unsampled tokens leave the device and the error bound is reported after the fact rather than enforced. Against Nexus and VASE (Duong et al., 2026; Chang et al., 2026), it keeps the inclusion probabilities that make both possible. Theorem 1 explains why a known design is needed, and Sections 6 and 7 measure where the resulting estimate pays and where it does not.

Baselines not run. Stronger answer-level uncertainty signals, semantic entropy (Kuhn et al., 2023) and a model's own probability that its answer is true (Kadavath et al., 2022), were not compared; the confidence baseline throughout is mean output log-probability under both signs. Head-level compression (Xiao et al., 2025; Ge et al., 2024) allocates budgets per head, while the certificate averages over a head subsample; its composition with head-wise designs is untested. Keyformer (Adnan et al., 2024) adds Gumbel noise to eviction scores, a randomized precedent that keeps no inclusion probabilities and therefore admits no certificate. HyperAttention (Han et al., 2024) and KDEformer (Zandieh et al., 2023) give sampling-based attention approximations with error guarantees over a resident cache; they do not evict. Nexus and VASE (Duong et al., 2026; Chang et al., 2026) are described from their papers and were not re-run.

Attribution needs labels. Eviction-induced failure is defined by a full-cache rerun, so attribution (Section 6.4, Table 11) is an offline diagnostic: it tells an operator how much of a failure rate the compression owns. The same full-cache labels score the pre-registered induced-failure claims R2 and R3; the attribution analysis among failures is post hoc. At serving time the certificate supplies the red-flag rate and the block localizer of Section 7 without labels.

Random and fixed budgets. Poisson sampling makes the retained-set size random, so at a nominal budget the Poisson arm holds more tokens than the deterministic arm on some queries and fewer on others; accuracy comparisons are at matched expected budgets. The fixed-size design of Appendix C removes the asymmetry but was tested only in replay.

Sample scopes of numbers reported in more than one place. (i) Two certificate AUCs on 6k induced failures pool the stream and aware conditions: 0.778 [0.760, 0.798] on the four LongBench tasks ($n = 5{,}576$ full-cache-correct Poisson answers; Section 6.2, R3), a same-sample comparison with log-probability (0.782), and 0.855 [0.84, 0.87] on the five-task pool that adds the needle anchor ($n = 8{,}680$), the figure R2 was scored on as registered. Pooling credits the certificate with telling the conditions apart; within a condition the five-task figure is 0.749 (stream) and 0.735 (aware), as Table 19 reports, log-probability leads in every condition, and R2's kill is triggered in the aware condition at 6k (Appendix D). (ii) The ten-turn certificate AUC at the 25% budget is 0.841 in Section 7 and Table 22 (online certificate of the certified vignette run, two Poisson seeds, 400 answers) and 0.802 in the deployed row of Table 6 (a separate trace-recording run of the same dialogues, one seed, 200 answers, certificate recomputed offline from the traces); the ablation compares implementation choices within one set of traces. (iii) The exploration tax, the accuracy of deterministic top-k minus that of the Poisson design, is 39–44 points on key–value retrieval (twenty keys) at the 25% budget on three of four base models in the aware condition of the pre-registered suite (Table 17); up to 43 points on one- and three-needle retrieval on Llama-3.1-8B-Instruct and phi-4 in the stream condition of the deterministic-baseline comparison, where it is at most 7 points on the two Qwen models (Table 9); up to 39 points on needle retrieval on the same two models in a separate certainty-layer run (Table 15); and 0–2.1 points on LongBench in the stream condition (Table 16). Section 8 quotes the LongBench figure and the synthetic maximum. (iv) Table 4 reports one Poisson seed; Tables 11 and 12 pool two at the 5–25% budgets and use one at 2% and 3%, as their captions state. (v) Figure 5 is drawn from the certified ten-turn run at the 25% budget, and its legend gives the per-model AUCs quoted in Section 7.

C Experimental details

Harness. All suites share one offline-replay harness. A prompt is prefilled once; importance scores are computed from accumulated attention (H2O-style, strided prefill queries), an observation window (SnapKV-style, last 64 prefill positions), or question-position queries (aware); each arm then receives its own compressed `DynamicCache` and generates greedily with per-step signal instrumentation. Protected positions (4 sinks, 32 recent) are never evicted. Poisson arms draw independent Bernoulli retention from (2) with the mass of the certainty layer chosen by the same budget accounting as the deterministic arms. The Hájek offset $+\log(1/\pi_i)$ enters the model as an additive attention bias on the retained keys. The pre-registered runs and the base-model synthetic suite (Table 17) were run with the uncorrected retained softmax and the certificate computed as the Hájek radius. The Poisson arms of both LongBench suites were then re-run with the offset on (four models, two budgets, two seeds, stream and aware; 12,800 LongBench generations per suite; two further seeds, 12,800 more per suite,

enter the attribution analysis (Table 11) and the four-seed check of the scheduling order in Section 6.4). At 16k: accuracy 0.239 and 0.340 at the 12.5% and 25% budgets against 0.243 and 0.344 uncorrected, certificate failure-prediction AUC 0.51–0.53 against 0.48–0.49 with log-probability at 0.75–0.76 (0.75–0.78 uncorrected), attribution AUC 0.65 against 0.65–0.67 (log-probability 0.38–0.39 against 0.40–0.42), and certificate gating at 32–33% of the oracle gain, 1.6–1.7 times random and 1.9 times confidence gating, against 31–36%, 1.7–1.8 and 1.8–2.1 uncorrected. In the 16k stream condition the attribution gap narrows from $+13.8$ to $+3.2$ points (Table 11), mostly because the certified rows lack the 50% budget: at the shared 12.5% and 25% budgets the pre-registered gap is $+7.6$ (certificate 0.666, log-probability 0.590), and the offset moves the two AUCs to 0.648 and 0.616; we have no mechanism for that remaining shift and report it as a limit. At 6k (four LongBench tasks, 25% and 50% budgets): accuracy 0.291 and 0.367 against 0.290 and 0.368, failure-prediction AUC 0.55–0.56 against 0.53 with log-probability at 0.71–0.73 in both, attribution AUC 0.65–0.74 against 0.64–0.72 (log-probability 0.49–0.50 against 0.49–0.52), and certificate gating $+3.3$ and $+2.9$ points over random against $+2.9$ and $+3.0$. The offset lowers the certificate's scale by 4–8% (median 0.72 against 0.75 at 6k, 0.85 against 0.92 at 16k), so only quantities that read the scale move; every ranking-based conclusion of Section 6 holds in the certified configuration, and a re-run of the synthetic suite without the offset (two instruct models, four tasks, 32 cells, independently drawn examples) gives Poisson accuracy 0.392 with the offset against 0.395 without, every cell within 0.04 except Llama single-needle retrieval at 50% in the stream condition (0.475 against 0.569). The agent-memory experiments of Section 7, whose red-flag rule depends on the certificate's scale, were run with the offset on. The certificate is (4) at $\delta = 0.1$ computed on four evenly spaced heads of every layer during the first six decode steps, averaged, then maximized over steps. The probability floor in (2) is $\varepsilon = 10^{-6}$ and the normalizer constant in (4) is $\epsilon_0 = 10^{-9}$. Poisson sampling leaves the retained-set size random. Fixed-size systematic sampling at the same inclusion probabilities (Hájek, 1964) keeps the budget exact; in the replay framework it keeps the certificate's coverage (0.980–0.999 against 0.981–0.997 for Poisson on four models, Table 8) and lowers the median error by 7–16%, at the price of joint inclusion probabilities that do not factor, so the single-sum estimator is then an approximation (the exact fixed-size form is Sen–Yates–Grundy (Yates & Grundy, 1953)) whose coverage is measured, not proved.

Signals. Retained-attention entropy: decode-step attention entropy over the retained set, normalized by $\log(\text{retained size})$, averaged as above. Evicted score mass: one minus the retained share of the arm's own importance mass at eviction time. Keep-boundary margin: minimum retained unprotected score minus maximum evicted score, in units of the score standard deviation. Output log-probability: mean generated-token log probability. All signals are online: computable at serving time by that arm.

Tasks and scoring. 6k suite: LongBench HotpotQA, 2WikiMQA, MultiFieldQA-en, and PassageRetrieval-en (100 examples each, deterministic shuffle), plus a three-passcode needle anchor; contexts middle-truncated to 6,000 tokens. 16k suite: HotpotQA and PassageRetrieval-en under the same sampling, plus MuSiQue and NarrativeQA; contexts to 16,000 tokens; budgets $\{12.5\%, 25\%, 50\%\}$. QA scored by LongBench token-F1 against all references (success F1 ≥ 0.5; the 0.3 threshold moves no verdict); retrieval by the first number in the answer and the needle anchor by the passcode's presence. Chat prompts split the template around the context so that stream arms compress before any question token exists.

No-context control. Each instruct model answers each task's questions with the context removed (same examples, same prompt scaffolding and scoring; 2,400 generations). No-context accuracy: 0.22–0.32 on HotpotQA and 2WikiMQA, 0.02–0.13 on MultiFieldQA-en, 0.07–0.13 on MuSiQue, 0.03–0.06 on NarrativeQA, 0.00–0.05 on PassageRetrieval-en. Of full-cache-correct examples, 75% (6k) and 84% (16k) are context-dependent. Restricted to those examples, aware-condition induced-failure rates are 11.5/6.9/2.3% for 16k Poisson at budgets 12.5/25/50% (unrestricted: 10.5/6.2/2.6%) and 9.5/4.4% for 6k Poisson at 25/50% (8.5/4.0%); streaming H2O rises from 43.8% to 47.3% at 16k/25%. The regime picture of Figure 2 is unchanged.

Agent-memory system and triage. The vignette generates fifty ten-turn dialogues from a pool of 32 fact templates (flight numbers, allergies, invoice numbers, door codes, contact names and others) embedded in shuffled natural chatter, one fact per user turn with a brief assistant acknowledgment; histories average 1,060 tokens. Compression uses the shared harness in the stream condition (H2O-

style prefill importance for the top-k arm, observation-window importance for the Poisson design), 25% budget, protected sinks and recency as elsewhere, two Poisson seeds; answers are scored by substring match of the fact value. The failures differ in kind: on the H2O arm at 25%, 39% (Qwen) and 46% (Llama) of them decline to answer by a phrase match ("The new SIM PIN is not provided in your latest message") and the rest state a wrong value ("The bike lock combination is 24-17-5"); recency-window failures are silent by construction. The budget sweep of Figure 4 reuses the same dialogue generator, seeds, arms, and scoring at budgets $\{20, 22, 24, 26, 28, 30\}\%$, merged for the operating picture with the 10% and 15% cells of a six-budget grid run under identical seeds (the shared 20% cell reproduces bit-identically across the two runs); the gated system is composed exactly from per-run records (greedy decoding makes the full-cache re-answer identical to the full arm) and the red-flag rule is fixed at $\tau = 1$ throughout; at budgets of 15% and below the near-zero silent rate is the gated system's distinctive property. On Llama-3.1-8B the sweep reproduces every pattern: failure rate flat from 89% to 80%, red-flag coverage of failures 100% to 8%, failure-median certificate 1.48 to 0.83. The thirty-turn variant draws thirty distinct kinds from the same pool (one per turn, no repeats), asks six recall questions per dialogue, and runs budgets $\{8, 10, 12, 15, 25\}\%$ with the same arms and scoring; the live transcripts (five seeds) are recorded at the 10% and 15% budgets, and Figure 1 shows the 15% run of one of them; in aggregate at that budget the gated system answers 0.465 against 0.451 for random retrieval at the rule's own rate and 0.078 without retrieval (Table 12), so the figure illustrates the mechanism, not the average outcome. The scale replication runs the identical generator, seeds, arms, and scoring on Qwen2.5-32B-Instruct at budgets $\{20, 25, 30\}\%$ (bf16, one H200). The fine-grained probe re-runs the 25% budget on the 7B model (the controls of Table 13 add Llama-3.1-8B), recording per-(layer, head) certificates over the first six decode steps and the per-retained-token variance contribution; on each red flag the question is re-answered four ways: full retrieval, evicted tokens of 64-token blocks ranked by retained-neighbor contribution up to 10% of the context, random blocks at the same budget, and an oracle that retrieves the block containing the queried fact value. A block's score is the sum of the per-token variance contributions $(1 - \pi_i)\pi_i^{-2} a_i^2 \|v_i - \hat{y}_t\|^2 / \widehat{N}^2$ of the retained tokens it contains, so a block with no retained token scores zero and ranks last; blocks are added in rank order, all their evicted tokens at once, until 10% of the context has been retrieved. A ranking that favors densely retained blocks, which contribute few evicted tokens each, therefore reaches further down its list before the budget is met: attention-mass retrieval recovers 0.62 accuracy on Qwen even on the 82 red flags where it ranks the fact block fourth or lower, while the top-three rate measures how sharply a ranking localizes. The archive is the full KV cache kept in host memory after prefill; a red flag re-answers from it or from the retrieved blocks, so retrieval costs host-to-device traffic rather than a re-prefill. The per-(layer, head) certificates of all 200 answers ship with the run records. Figure 5b sweeps each signal (certificate for Poisson, retained entropy for H2O) over its own range per model, flags answers above threshold, and averages the two models' curves at fixed false-alarm rate; the marked operating point applies the fixed rule $\mathrm{cert} \geq \tau = 1$ pooled over models, with no labels or tuning involved. The triage analysis splits Poisson stream failures into certificate terciles within each (model, task, budget) group, so certificate scales never mix, and measures the fraction whose full-cache run is correct. Among the third of stream failures the certificate blames most, 32.4% recover when rerun with the full cache; among the third it exonerates, 13.7% (6k suite, $n = 4{,}293$ failures, a 2.4× enrichment), while ranking the same failures by output confidence is uninformative (22.8% against 22.9%); at 16k the enrichment is 1.35× for the certificate against 1.11× for confidence.

Real conversations. LoCoMo (Maharana et al., 2024) provides ten two-person conversations of 19–32 dated sessions with 1,986 questions in five categories; we use categories 1–3 of the release, multi-hop, temporal and open-domain (699 questions). The transcript, with session headers and photo captions, is ingested as chat history under a system prompt that asks for value-only answers; each question is asked from the compressed views (recency, H2O, SnapKV-style top-k, Poisson with certificate) at budgets $\{5, 10, 15, 25\}\%$ (plus 2% and 3% on the 7B and 8B models) and from the full archive. Success is the reference contained in the answer or at least 75% of its tokens present (order-free, so date formats agree); the tables report mean token F1. The gated system is composed from the records as elsewhere.

Statistics. AUCs are Mann–Whitney; intervals are 95% cluster bootstraps (300–2,000 draws) resampling examples, or dialogues and conversations in the vignette and LoCoMo tables, so repeated measurements of one unit never inflate significance; Table 10 resamples model–task cells for its

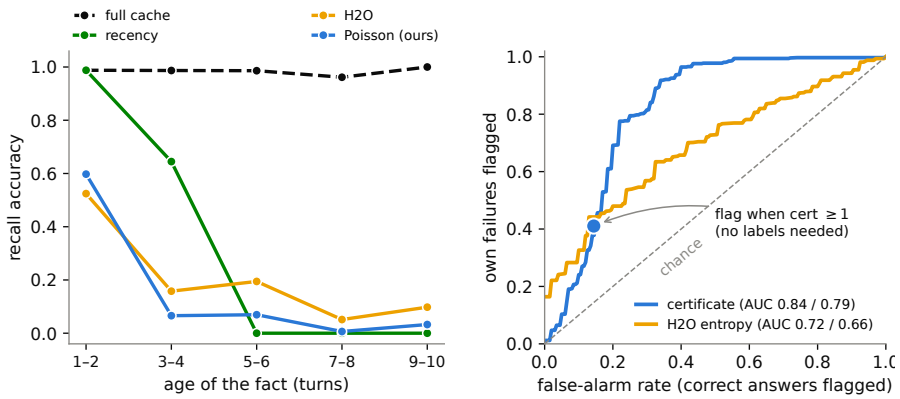

Figure 5: Under streaming compression to a 25% budget, facts three or more turns old are lost by every importance-based rule, and only the certificate flags the loss without labels. Two instruct models, 2,000 scored answers. (a) Recall by age of the fact. (b) Flagging one's own failures, model-averaged; the marked point is the label-free rule $\text{cert} \geq \tau = 1$.

per-cell differences. Paired signal comparisons bootstrap the AUC difference on shared examples. Pooled-across-model AUCs mix certificate scales across models and are therefore conservative; per-model values appear in Table 18. No correction for multiple comparisons is applied across the cells, budgets and settings reported; every interval is a per-comparison interval, and the pre-registered verdicts rest on pooled quantities fixed in advance rather than on the best cell.

D PRE-REGISTRATION RECORD

The registration file was written on 2026-07-22, after 4-example smoke runs and before the job array of full 6k shards was submitted, and its scale addendum before any 16k shard ran; both are timestamped in the released package alongside the Slurm submission records. Table 2 lists all seven claims with outcomes.

Table 2: The seven pre-registered claims and their outcomes. The prediction claims (R1, R3, R4, S2) failed. R2 survived as registered, with the Poisson arm pooled over conditions, and within the stream condition, but not within the aware condition at 6k (see below). The scale checks S1 and S3, which asked whether the regime finding was a truncation artifact or hid a tax at 16k, found neither. This is the division of labor reported in Section 6.

	Claim (abbreviated)	Kill condition	Verdict
R1	Certificate predicts own task failure on real tasks (deployable trust signal)	pooled AUC < 0.6	killed, both scales (0.555 at 6k, 0.572 at 16k)
R2	No deterministic self-signal sees its own eviction-induced failure	any signal $\geq$ cert lower CI	survived as registered (0.73 vs. 0.84); aware 6k killed
R3	Certificate $\geq$ output logprob on induced-failure prediction	logprob strictly dominates	not met (pooled tie at 6k; logprob wins otherwise)
R4	Certificate-gated $25\% \rightarrow 50\%$ escalation beats random at matched budget	cert $\leq$ random	killed (no headroom aware; $\approx$ random stream)
S1	6k "aware is free" is a truncation artifact; damage rises at 16k	rate rise $<$ 5pp confirms "free"	no rise: free at 16k (4.0% at 25%, shared tasks)
S2	Certificate AUC recovers (≥ 0.6) at 16k	< 0.6 = final	final kill (0.572)
S3	Exploration tax reappears at 12.5%	tax $<$ 3pp: no tax	no tax ($+0.3$pp)

R1 (certificate validity transfers): pooled LongBench certificate-to-own-failure AUC above 0.5 with CI excluding 0.5 in at least 3 of 4 models; kill if pooled AUC < 0.6. Outcome: per-model 0.552–0.581 (6k), all above 0.5; pooled 0.555, kill triggered; 16k pooled 0.572, kill confirmed final (S2).

R2 (silent failure is signal-general): every deterministic retained-set signal at AUC < 0.6 or CI overlapping 0.5; kill if any reaches the certificate's lower CI. Outcome as registered, with the Poisson arm pooled over the stream and aware conditions: overall-failure panel maxima 0.59 (evicted mass) against a certificate lower CI of 0.63; induced-failure panel maxima 0.73 (6k SnapKV evicted mass; 0.67 two-scale pooled) against certificate 0.855 (6k five-task pool including the needle anchor, lower CI 0.84); kill not triggered, though evicted mass exceeds the 0.6 line. Pooling also credits the certificate with telling the conditions apart, so we report the within-condition reading (Tables 1 and 19). In the stream condition the kill is not triggered: certificate 0.749 (lower CI 0.73) against 0.726 for SnapKV evicted mass on the five-task pool, 0.704 (lower CI 0.68) against 0.624 on the four LongBench tasks, and 0.72 (lower CI 0.70) against 0.67 on the two-scale pool. In the aware condition it is triggered at 6k: question-aware entropy reaches 0.699 against the certificate's lower CI of 0.693 on the five-task pool (on the LongBench tasks it falls just short, 0.6578 against 0.6583), and 0.580 against 0.526 on overall failures; it is not triggered at 16k (0.61 against 0.63) or on the two-scale pool (0.62 against 0.65). Silent failure is thus a property of streaming eviction, the scope in which the main text states it.

R3 (certificate at least matches output confidence on induced failures): kill if confidence strictly dominates. Outcome: 6k tie (-0.004, CI $[-0.031, +0.025]$); 16k confidence wins (-0.038, CI $[-0.065, -0.010]$), both with conditions pooled; within each condition confidence leads at both scales (0.76–0.85 against 0.67–0.73). Not met; prediction conceded to confidence.

R4 (certificate-gated escalation): kill if at or below random. Outcome: killed; aware has no headroom (fixed-budget accuracies 0.538 against 0.539), stream certificate-gating matches random, because a 50% stream cache is still broken. The recomputation analysis of Section 6.4 was designed after this autopsy and is post hoc.

S1–S3 as in Table 2: on the shared tasks the aware induced-failure rate at 16k/25% is 4.0% against 5.6% at 6k/25% (no rise; 12.5% reaches 7.3%; over all four 16k tasks 6.2% and 10.5%); certificate AUC 0.572 (< 0.6, final); aware tax $+0.3$pp at 12.5%.

Like the recomputation analysis, the per-seed replication, the exact random-gating expectations, the two-draw disagreement baseline, and the deployment experiments of Section 7 were added in revision and are post hoc; the pre-registered claims and kill conditions are exactly those listed above.

E Additional tables and figures

Table 3: The certificate is architecture-agnostic: on twelve models spanning multi-head and grouped-query attention (MHA, GQA), per-head and full-width QK-norm, fused qkv projections, logit soft-capping and mixture-of-experts feed-forward layers, coverage of the realized attention error at the 25% budget is 0.981–0.997 and the Spearman correlation between certificate and error is 0.973–0.988. Median relative attention error at the 25% budget: Poisson with the Hájek offset is below top-k on every model, and Poisson without the offset is above top-k on every model. Replay protocol of Section 4, 10,368–20,736 cells per model.

model	attention	L	coverage	Spearman	top-k	Poisson+Hájek	Poisson, no offset
SmolLM2-1.7B	MHA	24	0.984	0.978	0.034	**0.025**	0.043
Llama-3.2-3B	GQA	28	0.995	0.976	0.074	**0.039**	0.092
Qwen2.5-7B	GQA	28	0.981	0.973	0.049	**0.033**	0.063
Llama-3.1-8B	GQA	32	0.995	0.984	0.054	**0.032**	0.068
Mistral-7B-v0.3	GQA	32	0.997	0.984	0.068	**0.039**	0.085
Falcon3-7B	GQA	28	0.996	0.988	0.046	**0.027**	0.057
Qwen3-8B	GQA, QK-norm	36	0.995	0.980	0.059	**0.040**	0.074
OLMo-2-7B	MHA, QK-norm	32	0.997	0.975	0.179	**0.081**	0.210
Gemma-2-9B	GQA, soft-cap	42	0.993	0.983	0.033	**0.018**	0.042
phi-4	GQA, fused qkv	40	0.996	0.984	0.061	**0.043**	0.076
Qwen2.5-14B	GQA	48	0.995	0.981	0.062	**0.043**	0.080
Qwen1.5-MoE-A2.7B	MHA, MoE	24	0.993	0.982	0.041	**0.028**	0.051

Table 4: Real long-term conversations (LoCoMo, ten conversations of 11k–25k tokens, 699 multi-hop, temporal and open-domain questions, stream condition, mean token F1). The heavy-damage regime sits at a resident cache of roughly 400–1,000 tokens (2% of the history for Llama, 2–3% for Qwen2.5-7B, 5% for Qwen2.5-32B); there the certificate-gated system returns the full-cache score while top-k keeps a fifth to a third of it, and the 32B model raises the ceiling without changing the certificate's scale (median 1.23 and 1.25 at 5% for the two Qwen sizes). Above that point the flag goes dark and the residual failures are inherent or retrieval-type. The certificate separates eviction-induced from inherent failures better than output log-probability at every budget on every model under the pre-registered sign of log-probability; with its better sign the gap is 0–11 points with overlapping intervals (Table 11). One Poisson seed; Tables 11 and 12 pool two at 5–25%, hence their slightly different values.

model	budget	resident tokens	top-k	Poisson	gated ($\tau = 1$)	red-flag rate	silent	attribution AUC cert / logprob
Qwen2.5-7B-Instruct (10 conv.)	full	–	0.297					
	2%	424	0.092	0.068	**0.297**	1.00	0.00	0.58 / 0.48
	3%	630	0.090	0.093	**0.297**	1.00	0.00	0.54 / 0.52
	5%	1,053	0.110	0.089	**0.252**	0.79	0.20	0.54 / 0.50
	10%	2,110	0.130	0.140	**0.216**	0.48	0.48	0.55 / 0.51
	15%	3,164	0.153	0.155	**0.179**	0.14	0.80	0.53 / 0.48
	25%	5,308	0.192	0.206	**0.206**	0.00	0.92	0.58 / 0.44
Qwen2.5-32B-Instruct (10 conv.)	full	–	0.397					
	5%	1,051	0.090	0.131	**0.390**	0.96	0.04	0.55 / 0.48
	10%	2,110	0.157	0.181	**0.289**	0.52	0.44	0.55 / 0.45
	15%	3,163	0.194	0.209	**0.245**	0.18	0.76	0.54 / 0.46
	25%	5,308	0.265	0.255	**0.255**	0.00	0.89	0.60 / 0.49
Llama-3.1-8B-Instruct (10 conv.)	full	–	0.327					
	2%	427	0.063	0.067	**0.316**	0.97	0.03	0.55 / 0.52
	3%	630	0.072	0.088	**0.255**	0.68	0.30	0.54 / 0.46
	5%	1,043	0.086	0.098	**0.185**	0.33	0.64	0.55 / 0.48
	10%	2,100	0.121	0.159	**0.167**	0.02	0.92	0.55 / 0.47
	15%	3,173	0.157	0.183	**0.183**	0.00	0.93	0.57 / 0.47
	25%	5,287	0.194	0.227	**0.227**	0.00	0.91	0.59 / 0.55

Table 5: The red-flag boundary across model families in the certified configuration (ten-turn vignette, fifty dialogues, stream condition; thirty-turn where run). The budget at which the median certificate crosses $\tau = 1$ differs by family, but the resident cache at the crossing stays within a few hundred tokens; inside the heavy-damage regime (10%) the flag fires on 98–100% of questions and the gated system answers 0.98–0.99 against 0.10–0.12 for top-k on every model run at that budget (the 32B replication covers 20–30%).

model	$\tau = 1$ crossing (ten turns) budget	resident tokens	gated / top-k at 10%	red-flag rate at 10%	gated / top-k at 15%	AUC at 15%	crossing (thirty turns)
Qwen2.5-7B-Instruct	25%	264	0.98 / 0.10	1.00	0.98 / 0.12	0.92	14% (457 tokens)
Llama-3.1-8B-Instruct	21%	226	0.98 / 0.12	0.99	0.98 / 0.15	0.84	–
Qwen3-8B	21%	225	0.98 / 0.12	0.98	0.87 / 0.14	0.77	11% (345 tokens)
phi-4	22%	225	0.99 / 0.11	1.00	0.95 / 0.14	0.73	11% (337 tokens)
Qwen2.5-32B-Instruct	29%	308	– / –	–	– / –	–	–

Table 6: Implementation choices of the certificate, recomputed offline from per-unit traces of the ten-turn vignette (Qwen2.5-7B-Instruct, 200 answers per budget, offset on). The failure-ranking AUC is stable across confidence level, probed steps, head count and aggregation rule (0.77–0.82 at 25%, 0.86–0.91 at 15%), except max over units (0.40–0.47); the operating point of the label-free rule $\mathrm{cert} \geq 1$ moves with δ and the number of probed steps, which scale the radius, and max over units saturates the red flag. This is a separate trace-recording run of the same dialogues with its own Poisson draw, so its deployed-setting AUC (0.802 at 25%) differs from the online certificate of the certified vignette run reported in Section 7 and Table 22 (0.841); the comparisons in this table are within one set of traces (Appendix B).

	25% budget			15% budget		
setting	AUC	red-flag rate	gated acc.	AUC	red-flag rate	gated acc.
deployed ($\delta = 0.1$, 6 steps, 4 heads/layer, mean, max over steps)	0.802	0.47	0.575	0.877	0.95	1.000
$\delta = 0.05$	0.801	0.89	0.980	0.877	0.99	1.000
$\delta = 0.2$	0.803	0.03	0.165	0.878	0.48	0.575
1 decode step	0.815	0.01	0.150	0.872	0.76	0.840
3 decode steps	0.816	0.34	0.465	0.906	0.93	0.995
12 decode steps	0.788	0.64	0.735	0.893	0.95	1.000
1 head per layer	0.773	0.45	0.555	0.856	0.95	0.995
all 28 heads	0.782	0.47	0.585	0.873	0.94	0.990
median over units	0.810	0.20	0.325	0.881	0.87	0.960
max over units	0.403	1.00	1.000	0.465	1.00	1.000
mean over steps	0.797	0.10	0.235	0.893	0.89	0.975

Table 7: Iso-accuracy view of recomputation scheduling on the 16k LongBench suite (stream Poisson arms in the certified configuration, four models, four tasks, two seeds). With one threshold serving the mixed workload, the certificate reaches the accuracy of random gating at a quarter of the queries by re-running 11.0–15.7% of them, and log-probability gating needs more re-runs than random; with a separate threshold per model and task, log-probability gating catches up. The certificate is a relative error bound whose scale transfers across tasks; a confidence score's scale does not, so it needs per-task calibration to schedule well.

budget, thresholds	gate	5%	10%	15%	25%	50%	rate to match random at 25%
12.5%, one global threshold	certificate	**0.259**	**0.276**	**0.290**	**0.320**	**0.382**	15.7%
	log-probability	0.247	0.255	0.265	0.286	0.332	30.4%
	random	0.250	0.260	0.271	0.292	0.345	–
12.5%, per-task thresholds	certificate	0.256	**0.269**	0.283	0.308	0.364	19.5%
	log-probability	**0.257**	**0.269**	**0.284**	**0.313**	**0.374**	17.9%
	random	0.250	0.260	0.271	0.292	0.345	–
25%, one global threshold	certificate	**0.355**	**0.367**	**0.372**	**0.389**	**0.417**	11.0%
	log-probability	0.344	0.348	0.354	0.366	0.391	27.5%
	random	0.346	0.351	0.357	0.368	0.395	–
25%, per-task thresholds	certificate	**0.354**	**0.363**	**0.372**	0.385	0.415	12.8%
	log-probability	0.352	0.362	**0.372**	**0.391**	**0.422**	13.0%
	random	0.346	0.351	0.357	0.368	0.395	–

Table 8: Merging and fixed-size designs in the replay framework (six prompts, 48 probes per layer, every second layer, bf16). Merging error tracks the evicted attention mass rather than the within-stratum dispersion of evicted values that the stratified-ratio view of Remark 3 would predict, and the single-sum certificate keeps its coverage under fixed-size systematic sampling at matched inclusion probabilities.

model	budget	median relative error top-k	merge	Poisson	Spearman(merge error, ·) dispersion	evicted mass	coverage Poisson	fixed-size
Llama-3.1-8B-Instruct	50%	0.022	0.021	0.011	0.04	0.81	0.995	0.994
	25%	0.065	0.063	0.037	0.08	0.75	0.994	0.995
	12.5%	0.126	0.122	0.091	0.04	0.70	0.993	0.994
Qwen2.5-7B-Instruct	50%	0.014	0.013	0.008	−0.01	0.92	0.987	0.989
	25%	0.050	0.047	0.035	0.05	0.88	0.984	0.986
	12.5%	0.113	0.104	0.089	0.04	0.83	0.981	0.980
Qwen3-8B	50%	0.017	0.017	0.010	−0.07	0.87	0.993	0.996
	25%	0.059	0.056	0.041	−0.06	0.82	0.994	0.996
	12.5%	0.124	0.120	0.103	−0.06	0.76	0.991	0.993
phi-4	50%	0.019	0.018	0.011	0.06	0.88	0.997	0.999
	25%	0.061	0.060	0.043	0.10	0.82	0.996	0.997
	12.5%	0.125	0.125	0.106	0.07	0.76	0.995	0.996

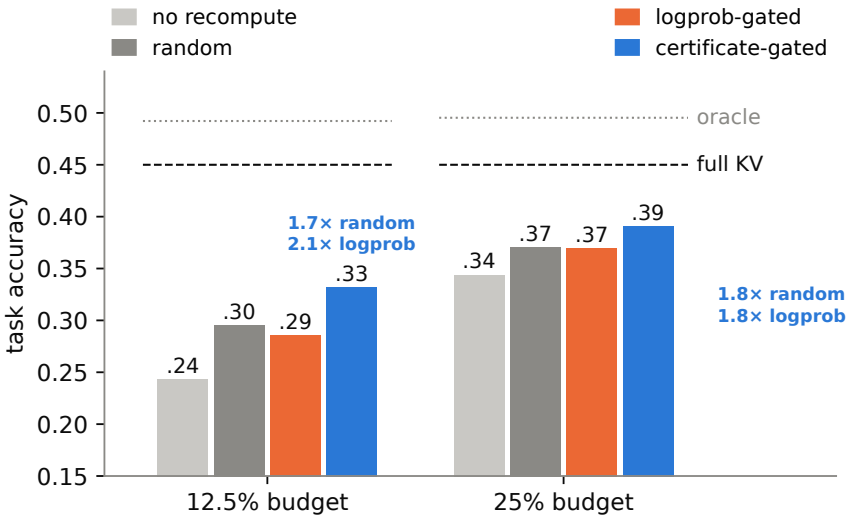

Figure 6: Certificate gating recovers more accuracy than random or confidence gating at matched compute. Streaming compression at 16k, 25% of queries re-run with the full cache; bars average the two Poisson seeds, and random gating is scored by its exact expectation. The oracle takes, per query, the better of the compressed and the full-cache answer (it re-runs exactly the queries the full cache fixes), so it can exceed the full-cache line.

Table 9: Deterministic baselines with budget allocation (PyramidKV, layer-wise budgets from 1.5× to 0.5× the mean) and compensation (CaM-style merging of evicted values into the nearest retained key) on the synthetic suite, stream condition, certified configuration. Bold marks the best compressed system in each row. Allocation and merging change which tokens go and what the retained values hold; neither produces an estimate of the eviction error, and the certificate-gated system leads at the 25% budget on average (0.43 against 0.05–0.06), though not on Llama's needle tasks.

model	task	budget	full	SnapKV	PyramidKV	CaM-merge	Poisson	gated ($\tau = 1$)
Llama-3.1-8B-Instruct	needle	25%	1.00	0.36	0.36	**0.41**	0.11	0.16
	needle	50%	1.00	**0.88**	0.85	**0.88**	0.57	0.57
	3 needles	25%	1.00	0.12	**0.15**	**0.15**	0.03	0.14
	3 needles	50%	1.00	0.70	**0.76**	0.74	0.39	0.39
	kv retrieval	25%	1.00	0.00	0.00	0.00	0.00	**0.09**
	kv retrieval	50%	1.00	0.03	**0.04**	0.01	0.01	0.01
	var. tracking	25%	0.10	**0.01**	0.00	**0.01**	0.00	0.00
	var. tracking	50%	0.10	0.01	0.01	0.03	**0.09**	**0.09**
Qwen2.5-7B-Instruct	needle	25%	1.00	0.00	0.00	0.00	0.00	**0.57**
	needle	50%	1.00	0.10	**0.12**	0.10	0.03	0.03
	3 needles	25%	1.00	0.00	0.00	0.00	0.00	**0.80**
	3 needles	50%	1.00	0.04	0.04	0.04	0.01	**0.06**
	kv retrieval	25%	0.97	0.00	0.00	0.00	0.00	**0.97**
	kv retrieval	50%	0.97	0.00	0.00	0.00	0.00	**0.07**
	var. tracking	25%	0.04	0.00	0.00	0.00	0.00	**0.03**
	var. tracking	50%	0.04	0.00	0.00	0.00	**0.01**	**0.01**
Qwen3-8B	needle	25%	1.00	0.01	0.04	0.01	0.00	**0.33**
	needle	50%	1.00	0.10	**0.25**	0.15	0.05	0.05
	3 needles	25%	1.00	0.00	0.00	0.00	0.00	**0.66**
	3 needles	50%	1.00	0.00	**0.06**	0.01	0.00	0.01
	kv retrieval	25%	1.00	0.00	0.00	0.00	0.00	**0.64**
	kv retrieval	50%	1.00	0.00	0.00	0.00	0.00	0.00
	var. tracking	25%	0.80	0.00	0.00	0.00	0.00	**0.32**
	var. tracking	50%	0.80	0.00	**0.04**	0.00	0.00	0.00
phi-4	needle	25%	1.00	0.17	0.23	0.20	0.09	**0.51**
	needle	50%	1.00	0.85	0.82	**0.89**	0.42	0.42
	3 needles	25%	1.00	0.00	0.01	0.03	0.04	**0.68**
	3 needles	50%	1.00	**0.46**	0.45	**0.46**	0.16	0.16
	kv retrieval	25%	1.00	0.00	0.00	0.00	0.00	**0.81**
	kv retrieval	50%	1.00	**0.04**	0.03	**0.04**	0.03	0.03
	var. tracking	25%	0.31	0.06	0.01	0.07	0.04	**0.23**
	var. tracking	50%	0.31	0.23	0.03	**0.25**	0.12	0.12
mean	all	25%	0.83	0.05	0.05	0.06	0.02	0.43
mean	all	50%	0.83	0.21	0.22	0.22	0.12	0.12

Table 10: A cheap evicted-set summary closes most of the induced-failure gap on SnapKV; the certificate still leads on average and does not depend on the evictor. This table scores the prediction of each arm's own eviction-induced failures among full-cache-correct examples; attribution among failures is the separate question of Table 11. Deterministic query-time proxies on the H2O and SnapKV arms (count, mean key and mean value of the evicted set per kv head give the evicted partition-function estimate $n_E \exp(q_t^\top \bar{k}_E/\sqrt{d})$ as a mass share and a moment residual; the low-score tail mass is the query's attention on the lowest-scoring quarter of retained tokens) against the certificate on the Poisson stream arm, AUC for the arm's own eviction-induced failures, four instruct models, 6k and 16k LongBench suites, 32 model–task cells, both budgets. Pooled: all cells together (rewards a signal whose scale transfers across models). Per-cell: mean AUC within (model, task, budget) cells and the paired difference to the certificate with a bootstrap over cells. Cells with fewer than five induced failures or five successes on either compared arm are excluded (2 of 64 on SnapKV, 5 on H2O); over H2O's 59 cells the certificate's per-cell mean is 0.671.

arm	signal	pooled AUC [95% CI]	per-cell mean	certificate − signal [CI]	cells won
SnapKV	evicted score mass (eviction time)	0.623 [0.60,0.64]	0.527	+14.7 [+11.8,+17.7]	54/62
	retained entropy	0.511 [0.48,0.54]	0.467	+20.7 [+18.0,+23.6]	59/62
	query-time evicted mass (moment summary)	0.614 [0.59,0.64]	0.645	+3.0 [+0.1,+5.9]	37/62
	moment residual proxy	0.593 [0.57,0.61]	0.641	+3.3 [+0.8,+5.9]	39/62
	low-score tail mass	0.620 [0.59,0.65]	0.609	+6.5 [+3.6,+9.6]	44/62
H2O	evicted score mass (eviction time)	0.632 [0.61,0.65]	0.503	+16.8 [+13.3,+20.1]	54/59
	retained entropy	0.475 [0.45,0.51]	0.477	+19.4 [+16.1,+22.3]	55/59
	query-time evicted mass (moment summary)	0.575 [0.55,0.60]	0.554	+11.7 [+8.7,+14.7]	50/59
	moment residual proxy	0.583 [0.56,0.61]	0.554	+11.7 [+8.7,+15.0]	47/59
	low-score tail mass	0.507 [0.48,0.53]	0.523	+14.8 [+11.6,+18.0]	51/59
Poisson	**certificate**	**0.705** [0.69,0.72]	**0.674**		

Table 11: Attribution with a polarity-agnostic confidence baseline. AUC for separating eviction-induced from inherent failures; log-probability is scored with whichever sign does better ("high": confident failures are more often induced, "low": the reverse), with 95% cluster-bootstrap intervals over examples. The certificate keeps a 13–22 point lead on the LongBench suites in the pooled comparison and 3–5 points on the certified 16k stream and aware subsets; on LoCoMo the gap is 0–11 points with overlapping intervals. The sign of the confidence signal flips between suites and conditions, which is itself the argument against using it for attribution. The 23k natural-length subset of Figure 3 is not tabulated. The (all) rows pool the stream and aware conditions, which raises the certificate's AUC more than log-probability's; the certificate leads in every within-condition row. Certified rows pool four Poisson seeds and at 16k cover the 12.5% and 25% budgets only; pre-registered rows pool two seeds over all budgets. LoCoMo rows pool two seeds at 5–25% and use one at 2% and 3%.

setting	failures	induced	certificate [CI]	log-probability, best sign [CI]	gap
6k pre-reg. (all)	7924	1133	0.749 [0.73,0.77]	0.542 (low) [0.51,0.57]	+20.7
6k pre-reg. (stream)	4293	958	0.677 [0.65,0.70]	0.506 (low) [0.47,0.54]	+17.1
6k pre-reg. (aware)	3631	175	0.702 [0.66,0.76]	0.589 (low) [0.54,0.65]	+11.3
16k pre-reg. (all)	11717	1776	0.727 [0.71,0.75]	0.531 (high) [0.51,0.56]	+19.6
16k pre-reg. (stream)	6400	1498	0.703 [0.68,0.72]	0.565 (high) [0.54,0.59]	+13.8
16k pre-reg. (aware)	5317	278	0.625 [0.59,0.66]	0.540 (low) [0.48,0.60]	+8.5
6k certified (all)	15930	2151	0.760 [0.75,0.77]	0.540 (low) [0.52,0.56]	+22.0
6k certified (stream)	8568	1801	0.682 [0.67,0.70]	0.504 (low) [0.48,0.53]	+17.8
6k certified (aware)	7362	350	0.731 [0.70,0.76]	0.595 (low) [0.56,0.64]	+13.6
16k certified (all)	16231	2962	0.704 [0.69,0.72]	0.570 (high) [0.55,0.59]	+13.4
16k certified (stream)	9077	2529	0.648 [0.63,0.66]	0.616 (high) [0.59,0.63]	+3.2
16k certified (aware)	7154	433	0.590 [0.56,0.62]	0.541 (low) [0.51,0.58]	+4.9
LoCoMo Qwen2.5-7B-Instruct 0.02	674	79	0.578 [0.52,0.63]	0.519 (high) [0.44,0.60]	+5.9
LoCoMo Qwen2.5-7B-Instruct 0.03	673	77	0.539 [0.48,0.58]	0.521 (low) [0.46,0.57]	+1.8
LoCoMo Qwen2.5-7B-Instruct 0.05	1344	148	0.542 [0.47,0.60]	0.531 (high) [0.49,0.57]	+1.1
LoCoMo Qwen2.5-7B-Instruct 0.1	1323	129	0.545 [0.46,0.61]	0.520 (low) [0.46,0.58]	+2.5
LoCoMo Qwen2.5-7B-Instruct 0.15	1318	128	0.562 [0.47,0.64]	0.519 (low) [0.46,0.58]	+4.3
LoCoMo Qwen2.5-7B-Instruct 0.25	1279	94	0.596 [0.51,0.68]	0.546 (high) [0.48,0.60]	+5.0
LoCoMo Llama-3.1-8B-Instruct 0.02	670	79	0.553 [0.49,0.60]	0.516 (low) [0.44,0.59]	+3.7
LoCoMo Llama-3.1-8B-Instruct 0.03	666	73	0.539 [0.48,0.59]	0.538 (high) [0.43,0.63]	+0.1
LoCoMo Llama-3.1-8B-Instruct 0.05	1338	150	0.581 [0.54,0.62]	0.513 (high) [0.46,0.56]	+6.8
LoCoMo Llama-3.1-8B-Instruct 0.1	1312	130	0.572 [0.51,0.64]	0.511 (high) [0.45,0.58]	+6.1
LoCoMo Llama-3.1-8B-Instruct 0.15	1291	115	0.597 [0.53,0.66]	0.531 (high) [0.46,0.60]	+6.6
LoCoMo Llama-3.1-8B-Instruct 0.25	1263	95	0.593 [0.50,0.67]	0.514 (low) [0.42,0.61]	+7.9
LoCoMo Qwen2.5-32B-Instruct 0.05	1322	244	0.547 [0.49,0.61]	0.542 (high) [0.49,0.59]	+0.5
LoCoMo Qwen2.5-32B-Instruct 0.1	1295	224	0.560 [0.49,0.63]	0.543 (high) [0.50,0.59]	+1.7
LoCoMo Qwen2.5-32B-Instruct 0.15	1278	209	0.567 [0.49,0.64]	0.522 (high) [0.46,0.58]	+4.5
LoCoMo Qwen2.5-32B-Instruct 0.25	1238	169	0.614 [0.53,0.69]	0.505 (high) [0.44,0.59]	+10.9

Table 12: On real conversations the certificate has little per-query selectivity: at a matched retrieval rate its gating is within 2.4 points of random gating, and output log-probability gating is as good or better; on the synthetic vignettes the certificate beats random by 2–12 points but log-probability still matches or beats it. The rule $\mathrm{cert} \geq 1$ chooses the rate without labels; the random column re-runs the same fraction of queries at random (expectation). Matched-rate columns: accuracy (F1 on LoCoMo) when the 50% of queries with the highest signal are re-run from the full archive, differences in points with cluster-bootstrap intervals over conversations or dialogues. Two Poisson seeds on LoCoMo at 5–25% and one at 2% and 3%; certified configuration throughout.

				rule $\mathrm{cert} \geq 1$			matched retrieval rate 50%		
setting	budget	no retrieval	always retrieve	red rate	gated	random at that rate	cert − random [CI]	cert − log-prob [CI]	log-prob
LoCoMo, Qwen2.5-7B-Instruct	2%	0.068	0.297	1.00	0.297	0.297	+1.1 [+0.2,+2.2]	+0.1 [-1.8,+2.2]	0.192
	3%	0.093	0.297	1.00	0.297	0.297	−0.2 [-1.2,+0.5]	−1.9 [-4.7,-0.3]	0.212
	5%	0.090	0.297	0.80	0.252	0.255	−0.1 [-1.2,+1.0]	−1.6 [-2.3,-0.4]	0.208
	10%	0.139	0.297	0.48	0.219	0.215	+0.4 [-0.5,+1.1]	−1.7 [-2.6,-0.6]	0.239
	15%	0.153	0.297	0.14	0.182	0.173	+0.3 [-0.7,+1.2]	−2.3 [-4.1,-0.5]	0.252
	25%	0.209	0.297	0.00	0.209	0.209	+0.8 [-0.3,+1.9]	−0.3 [-2.0,+1.1]	0.264
LoCoMo, Llama-3.1-8B-Instruct	2%	0.067	0.327	0.97	0.316	0.319	+1.5 [-0.2,+3.4]	−1.0 [-3.4,+3.2]	0.223
	3%	0.088	0.327	0.68	0.255	0.251	+1.1 [-0.1,+2.4]	−0.8 [-2.8,+2.0]	0.226
	5%	0.108	0.327	0.35	0.196	0.184	+0.7 [-0.6,+1.5]	−0.9 [-2.5,+0.4]	0.234
	10%	0.155	0.327	0.02	0.164	0.159	+1.2 [+0.1,+2.1]	+0.2 [-1.3,+1.4]	0.252
	15%	0.188	0.327	0.00	0.188	0.188	+2.4 [+1.2,+3.0]	+1.5 [-0.0,+2.5]	0.266
	25%	0.229	0.327	0.00	0.229	0.229	+1.5 [+0.8,+2.3]	−0.4 [-1.6,+0.9]	0.297
LoCoMo, Qwen2.5-32B-Instruct	5%	0.130	0.397	0.96	0.390	0.386	−0.8 [-2.0,+0.4]	−2.5 [-3.8,-1.2]	0.281
	10%	0.180	0.397	0.52	0.298	0.294	+0.3 [-1.0,+1.4]	−2.2 [-3.8,-1.1]	0.314
	15%	0.214	0.397	0.18	0.254	0.246	+0.3 [-1.2,+1.5]	−2.3 [-3.5,-1.0]	0.332
	25%	0.266	0.397	0.00	0.266	0.266	+1.0 [-0.2,+2.1]	−0.7 [-2.3,+0.7]	0.348
ten-turn, Qwen2.5-7B-Instruct	10%	0.107	0.985	1.00	0.985	0.983	+4.4 [+2.7,+6.3]	−1.0 [-2.5,+0.0]	0.600
	15%	0.113	0.985	0.94	0.980	0.930	+4.6 [+2.9,+6.6]	−0.5 [-1.8,+0.5]	0.600
	20%	0.130	0.985	0.80	0.885	0.810	+5.0 [+3.2,+6.6]	−1.5 [-3.4,-0.3]	0.623
	25%	0.150	0.985	0.48	0.600	0.551	+4.7 [+3.0,+7.2]	−3.2 [-4.5,-1.2]	0.647
	40%	0.240	0.985	0.02	0.258	0.257	+6.0 [+3.7,+9.0]	−5.0 [-7.8,-2.0]	0.723
	60%	0.398	0.985	0.00	0.398	0.398	+7.6 [+5.1,+10.3]	−6.5 [-9.0,-2.5]	0.833
	22%	0.145	0.985	0.69	0.797	0.727	+5.0 [+3.4,+7.0]	−2.7 [-4.5,-1.0]	0.642
	24%	0.145	0.985	0.54	0.650	0.597	+5.3 [+3.2,+7.5]	−2.5 [-4.2,-1.0]	0.642
	26%	0.150	0.985	0.41	0.530	0.492	+5.0 [+2.9,+6.9]	−2.5 [-4.9,-1.0]	0.642
	28%	0.175	0.985	0.29	0.430	0.410	+5.0 [+2.6,+7.1]	−3.5 [-6.0,-1.2]	0.665
	30%	0.182	0.985	0.19	0.347	0.337	+5.4 [+3.1,+7.5]	−3.5 [-6.0,-1.7]	0.672
ten-turn, Llama-3.1-8B-Instruct	10%	0.115	0.985	0.99	0.985	0.981	+4.5 [+2.4,+6.6]	−0.8 [-2.7,+0.8]	0.603
	15%	0.128	0.985	0.93	0.983	0.925	+4.1 [+2.3,+6.4]	−1.7 [-3.6,-0.5]	0.615
	20%	0.145	0.985	0.67	0.765	0.706	+4.3 [+2.7,+6.8]	−2.7 [-4.5,-0.5]	0.635
	25%	0.163	0.985	0.26	0.400	0.374	+4.9 [+2.7,+7.0]	−3.0 [-5.0,-1.3]	0.652
	40%	0.310	0.985	0.00	0.310	0.310	+6.2 [+3.9,+9.9]	−7.8 [-10.5,-4.7]	0.787
	60%	0.583	0.985	0.00	0.583	0.583	+12.1 [+8.8,+13.9]	−2.0 [-6.5,+0.5]	0.925
	22%	0.147	0.985	0.47	0.585	0.541	+4.1 [+2.3,+6.2]	−3.0 [-4.5,-1.0]	0.637
	24%	0.158	0.985	0.28	0.415	0.391	+4.9 [+2.8,+6.9]	−2.7 [-4.9,-1.5]	0.647
	26%	0.165	0.985	0.21	0.360	0.339	+5.0 [+3.0,+7.4]	−2.5 [-4.4,-1.0]	0.650
	28%	0.175	0.985	0.12	0.292	0.276	+5.0 [+3.0,+7.0]	−3.2 [-5.1,-1.5]	0.662
	30%	0.200	0.985	0.07	0.263	0.259	+4.2 [+1.8,+6.1]	−4.2 [-7.5,-2.2]	0.677
thirty-turn, Qwen2.5-7B-Instruct	15%	0.078	0.990	0.41	0.465	0.451	+2.1 [+0.5,+3.3]	−2.0 [-3.5,-1.0]	0.575
	25%	0.130	0.990	0.01	0.142	0.141	+3.3 [+1.8,+5.2]	−3.0 [-4.7,-1.2]	0.623
	8%	0.052	0.990	0.98	0.990	0.976	+2.2 [+1.1,+3.5]	−0.2 [-1.0,+0.7]	0.545
	10%	0.062	0.990	0.95	0.972	0.945	+2.4 [+1.2,+3.9]	−0.8 [-1.8,+0.3]	0.558
	12%	0.077	0.990	0.86	0.902	0.861	+1.8 [+0.6,+3.3]	−1.8 [-3.0,-0.3]	0.570

Table 13: The variance weighting is what localizes the lost block. Ranking evicted 64-token blocks on red-flagged answers of the ten-turn vignette (25% budget, same retained set): by the certificate's per-token variance contribution, by the query's raw attention mass on the same retained neighbors (computable by any evictor), by eviction-time importance, by recency, at random, and with the fact block placed first (oracle). Attention mass recovers most of the retrieval accuracy but ranks the fact block in the top three far less often; intervals are cluster bootstraps over dialogues.

		accuracy after retrieving 10% of the context						fact block in top 3		
model	red-flagged	certificate	attention mass	importance	recency	random	oracle	certificate	attention mass	importance
Llama-3.1-8B-Instruct	54	**0.87** [0.76,0.96]	0.74 [0.64,0.84]	0.19	0.19	0.28	0.96	**0.93**	0.59	0.06
Qwen2.5-7B-Instruct	94	**0.70** [0.62,0.79]	0.66 [0.57,0.75]	0.17	0.22	0.18	0.89	**0.79**	0.13	0.00

Table 14: Coverage holds in the deployed streaming regime. In a separate coverage run of the ten- and thirty-turn vignettes (fifty dialogues, four questions each, 200 answers per cell, question issued after compression), the realized relative attention-output error of the Hájek estimate against the full archive is computed for the query the compressed model actually issued, at every layer and four evenly spaced heads over the first six decode steps, and compared with the certificate radius: the deployed statistic (mean over units, max over steps) covers the matching error in every answer, per-unit coverage is 0.97–0.99, the radius is about three times the realized error, and the range term $b\log(1/\delta)$ carries 36–42% of it.

model	turns	budget	coverage, deployed	coverage, per unit	cert median	realized error median	cert / error	range-term share	Spearman
Llama-3.1-8B-Instruct	10	15%	1.000	0.989	1.24	0.43	2.9	0.41	0.87
Llama-3.1-8B-Instruct	10	25%	1.000	0.993	0.92	0.30	3.0	0.39	0.92
Llama-3.1-8B-Instruct	30	10%	1.000	0.993	1.08	0.36	3.0	0.39	0.82
Llama-3.1-8B-Instruct	30	15%	1.000	0.993	0.87	0.29	3.0	0.38	0.91
Llama-3.1-8B-Instruct	30	25%	1.000	0.994	0.65	0.21	3.1	0.36	0.93
Qwen2.5-7B-Instruct	10	15%	1.000	0.970	1.29	0.50	2.6	0.42	0.90
Qwen2.5-7B-Instruct	10	25%	1.000	0.977	1.00	0.36	2.8	0.40	0.92
Qwen2.5-7B-Instruct	30	10%	1.000	0.976	1.19	0.43	2.7	0.40	0.77
Qwen2.5-7B-Instruct	30	15%	1.000	0.980	0.98	0.34	2.8	0.39	0.86
Qwen2.5-7B-Instruct	30	25%	1.000	0.984	0.75	0.25	3.0	0.38	0.90

Table 15: The exploration tax is a design parameter. The certainty layer of the design is adaptive (water-filling sets $\pi_i = 1$ on the highest scores), so forcing 25–50% of the budget deterministic changes little; forcing 75% recovers about half of the tax at the 50% budget at the price of a thinner tail, a larger radius and a higher red-flag rate, which raises the gated accuracy at 25%. Synthetic suite, stream condition, four instruct models; a separate run from Table 9 with independently drawn examples, so shared cells differ.

				Poisson, certainty layer forced to				gated ($\tau = 1$)	
model	task	budget	SnapKV	adaptive	25%	50%	75%	adaptive	75%
Llama-3.1-8B-Instruct	kv retrieval	25%	0.00	0.01	0.01	0.01	0.01	0.15	0.59
	kv retrieval	50%	0.03	0.01	0.01	0.01	0.00	0.01	0.00
	needle	25%	0.41	0.11	0.11	0.12	0.24	0.14	0.47
	needle	50%	0.88	0.57	0.57	0.62	0.74	0.57	0.74
	3 needles	25%	0.05	0.07	0.07	0.07	0.07	0.23	0.41
	3 needles	50%	0.61	0.38	0.38	0.38	0.46	0.38	0.46
Qwen2.5-7B-Instruct	kv retrieval	25%	0.00	0.00	0.00	0.00	0.00	0.99	0.98
	kv retrieval	50%	0.00	0.00	0.00	0.00	0.00	0.10	0.42
	needle	25%	0.01	0.01	0.01	0.01	0.00	0.48	0.65
	needle	50%	0.06	0.06	0.06	0.04	0.05	0.08	0.11
	3 needles	25%	0.00	0.00	0.00	0.00	0.00	0.79	0.85
	3 needles	50%	0.05	0.03	0.03	0.03	0.03	0.11	0.26
Qwen3-8B	kv retrieval	25%	0.00	0.00	0.00	0.00	0.00	0.64	0.91
	kv retrieval	50%	0.00	0.00	0.00	0.00	0.00	0.00	0.01
	needle	25%	0.00	0.00	0.00	0.00	0.00	0.34	0.56
	needle	50%	0.09	0.02	0.02	0.02	0.01	0.02	0.01
	3 needles	25%	0.00	0.00	0.00	0.00	0.00	0.66	0.88
	3 needles	50%	0.00	0.01	0.01	0.01	0.01	0.01	0.04
phi-4	kv retrieval	25%	0.00	0.00	0.00	0.00	0.00	0.81	0.97
	kv retrieval	50%	0.03	0.01	0.01	0.01	0.01	0.01	0.02
	needle	25%	0.23	0.07	0.07	0.06	0.11	0.51	0.69
	needle	50%	0.84	0.45	0.45	0.48	0.69	0.45	0.69
	3 needles	25%	0.03	0.06	0.06	0.05	0.04	0.71	0.83
	3 needles	50%	0.45	0.22	0.22	0.21	0.32	0.22	0.33

Table 16: Head-to-head accuracy in the stream condition, where the question arrives after compression: Poisson eviction (certified configuration where re-run, pre-registered run otherwise) against the deterministic arms, means over four instruct models and four LongBench tasks per suite. The design costs 0–2.1 points against SnapKV here and nothing in the aware condition (Section 6.2); the needle-retrieval exploration tax of the synthetic suites is the exception (Tables 9 and 15).

suite, budget	full	StreamingLLM	H2O	SnapKV	Poisson	Poisson − SnapKV, points (cells won)
6k, 25%	0.436	0.211	0.313	0.292	0.291	-0.2 (6/16)
6k, 50%	0.436	0.271	0.367	0.367	0.367	+0.0 (8/16)
16k, 12.5%	0.450	0.162	0.231	0.260	0.239	-2.1 (4/16)
16k, 25%	0.450	0.217	0.298	0.343	0.340	-0.2 (9/16)
16k, 50%	0.450	0.284	0.404	0.419	0.412	-0.6 (6/16)

Table 17: Per-cell results of the pre-registered synthetic suite of Section 5 (four base models, four RULER-style tasks, 640 scored Poisson generations per cell, budgets 25% and 50%, stream and aware conditions pooled). AUC of the certificate for predicting the Poisson arm's own task failure with 95% cluster-bootstrap intervals over examples; the stream and aware columns give the AUC within one condition where both outcomes occur at least five times (– otherwise), since pooling also credits the certificate with telling the conditions apart, and nearly every stream answer fails. The last two columns give the aware-condition exploration tax (deterministic question-aware top-k minus Poisson, in points) at each budget. Pooled over all cells, abstaining on the 70% of answers with the highest certificate leaves an error rate of 0.229 against 0.575 at full coverage.

model	task	failures	successes	AUC, pooled [95% CI]	stream	aware	tax@25%	tax@50%
Llama-3.1-8B	kv retrieval	313	327	0.923 [0.90, 0.94]	0.71	–	+0.0	+0.0
Llama-3.1-8B	needle	259	381	0.835 [0.81, 0.86]	0.77	–	+0.0	+0.0
Llama-3.1-8B	3 needles	266	374	0.877 [0.85, 0.90]	0.72	–	+0.0	+0.0
Llama-3.1-8B	var. tracking	333	307	0.830 [0.79, 0.86]	0.75	0.47	+0.0	-0.6
Mistral-7B-v0.3	kv retrieval	391	249	0.826 [0.81, 0.85]	–	0.82	+42.5	+1.9
Mistral-7B-v0.3	needle	314	326	0.793 [0.78, 0.81]	0.86	–	+0.0	+0.0
Mistral-7B-v0.3	3 needles	320	320	0.859 [0.84, 0.88]	–	–	+1.9	+0.0
Mistral-7B-v0.3	var. tracking	410	230	0.722 [0.70, 0.75]	–	0.56	+3.2	-1.2
Qwen2.5-1.5B	kv retrieval	418	222	0.970 [0.96, 0.98]	–	0.90	+43.7	+5.0
Qwen2.5-1.5B	needle	334	306	0.866 [0.85, 0.88]	–	0.83	+8.8	+0.0
Qwen2.5-1.5B	3 needles	342	298	0.960 [0.95, 0.97]	–	0.89	+13.2	+0.0
Qwen2.5-1.5B	var. tracking	600	40	0.748 [0.69, 0.80]	–	0.54	-1.3	+0.0
Qwen2.5-7B	kv retrieval	387	253	0.851 [0.83, 0.87]	–	0.85	+39.4	+1.2
Qwen2.5-7B	needle	316	324	0.790 [0.78, 0.80]	–	–	+0.0	+0.0
Qwen2.5-7B	3 needles	320	320	0.879 [0.86, 0.90]	–	–	+0.6	+0.0
Qwen2.5-7B	var. tracking	564	76	0.650 [0.60, 0.70]	–	0.49	-2.6	-1.8

Table 18: R1 detail: certificate-to-own-failure AUC, LongBench only, pre-registered runs (uncorrected softmax, certificate as the post-hoc Hájek radius). Every model exceeds chance; no interval clears the pre-registered 0.6 line (one point estimate, 0.602, sits on it) and the pooled AUC stays below it. Cell-level: 24/32 cells above 0.5 at 6k, 30/32 at 16k.

Model	6k		16k	
	AUC	95% CI	AUC	95% CI
Qwen2.5-1.5B-Instruct	0.581	[0.555, 0.609]	0.593	[0.567, 0.621]
Qwen2.5-7B-Instruct	0.558	[0.535, 0.581]	0.602	[0.583, 0.622]
Llama-3.1-8B-Instruct	0.565	[0.545, 0.587]	0.582	[0.567, 0.599]
Mistral-7B-Instruct-v0.3	0.552	[0.532, 0.574]	0.596	[0.580, 0.613]
pooled	0.555	[0.543, 0.568]	0.572	[0.562, 0.582]

Table 19: The full self-signal panel on the pre-registered runs (uncorrected softmax, certificate as the post-hoc Hájek radius): AUC predicting the arm's own failure, with 95% cluster-bootstrap intervals (resampling examples); best signal per row in bold. The 6k pool holds all five tasks including the needle anchor, and the Poisson arm is split by condition like the deterministic arms; pooled over both conditions its certificate would read .645 and .855, partly for telling the conditions apart. On the four LongBench tasks alone, the sample of Section 6.2 and R3, the pooled certificate scores 0.778 [0.760, 0.798] on induced failures against 0.782 for log-probability. Output log-probability is best in every row, which is why prediction is conceded to it; within each condition, SnapKV's evicted score mass and question-aware entropy come within four points of the certificate on induced failures, and Table 10 adds the query-time summaries.

Arm	certificate (ours)	entropy	evicted mass	margin	output logprob (baseline)
Overall failure (6k suite, pooled)					
StreamingLLM	–	.472 [.45,.50]	–	–	**.822** [.80,.84]
H2O	–	.521 [.50,.54]	.590 [.58,.61]	.503 [.50,.51]	**.772** [.75,.79]
SnapKV	–	.538 [.51,.56]	.592 [.58,.61]	.502 [.50,.51]	**.771** [.75,.79]
aware top-k	–	.580 [.56,.61]	.502 [.49,.52]	.499 [.50,.50]	**.799** [.78,.82]
Poisson, stream	.618 [.60,.63]	.519 [.49,.54]	–	–	**.754** [.74,.77]
Poisson, aware	.542 [.53,.56]	.583 [.56,.61]	–	–	**.800** [.78,.82]
Eviction-induced failure (full-correct examples, 6k suite)					
StreamingLLM	–	.455 [.42,.49]	–	–	**.861** [.84,.88]
H2O	–	.501 [.47,.53]	.660 [.64,.68]	.500 [.49,.51]	**.831** [.81,.85]
SnapKV	–	.509 [.48,.54]	.726 [.71,.74]	.501 [.50,.51]	**.834** [.82,.85]
aware top-k	–	.699 [.61,.78]	.575 [.51,.65]	.493 [.48,.50]	**.852** [.81,.89]
Poisson, stream	.749 [.73,.76]	.477 [.45,.51]	–	–	**.801** [.78,.82]
Poisson, aware	.735 [.69,.77]	.654 [.58,.72]	–	–	**.867** [.83,.90]

Table 20: Under one global threshold, certificate gating recovers the most accuracy in every suite, budget, and seed. Gated recomputation (stream condition, LongBench tasks only, 25% re-run rate) with per-seed replication. Random gating is the exact expectation $0.25\,\mathrm{acc}_{\mathrm{full}} + 0.75\,\mathrm{acc}_{\mathrm{base}}$; intervals are cluster bootstrap over examples. The two-draw attributor (partner-draw success and the F1 gap between draws, label-assisted proxies) scores AUC 0.632 and 0.687 at 6k and 0.626 and 0.620 at 16k against the certificate's 0.749 and 0.727 from a single draw; recomputation gated on two-draw disagreement (fires on 15% of queries) reaches .314/.274/.373 in the three suite/budget rows and trails certificate gating at the matched rate (.331–.343/.289–.306/.376–.379) in every configuration. At natural 23k-token lengths the full cache answers 27%, the headroom shrinks to ten points, and at $n = 200$ no gating signal separates from random: budgeted recomputation is a product for regimes with headroom.

Suite/budget	seed	base	full	cert (ours)	logprob	random (baseline)	cert−random [95% CI]	cert−logprob [95% CI]
6k / 25%	0	.296	.436	**.358**	.346	.331	$+.027$ [$+.015, +.039$]	$+.013$ [$-.004, +.030$]
	1	.284	.436	**.352**	.338	.322	$+.030$ [$+.019, +.043$]	$+.014$ [$-.003, +.033$]
16k / 12.5%	0	.240	.450	**.329**	.281	.292	$+.037$ [$+.023, +.049$]	$+.049$ [$+.031, +.066$]
	1	.247	.450	**.334**	.291	.298	$+.037$ [$+.024, +.050$]	$+.044$ [$+.026, +.062$]
16k / 25%	0	.343	.450	**.389**	.372	.369	$+.019$ [$+.010, +.030$]	$+.016$ [$+.002, +.032$]
	1	.346	.450	**.394**	.367	.372	$+.022$ [$+.011, +.032$]	$+.027$ [$+.010, +.043$]

Table 21: Question-aware compression tracks the full cache and streaming compression does not. Accuracy by task at 16k, pooled over models (aware condition for aware/Poisson columns; H2O is the stream-condition baseline); best compressed arm per row in bold.

Task (16k)	full	H2O@25%	H2O@50%	aware@25%	Poisson@25%	Poisson@50%
HotpotQA	0.515	0.420	0.470	**0.530**	0.516	0.524
MuSiQue	0.260	0.180	0.215	0.258	**0.263**	0.254
NarrativeQA	0.205	0.147	0.195	**0.200**	0.198	0.196
PassageRetrieval	0.820	0.445	0.738	0.818	0.815	**0.821**

Table 22: The certificate survives a 4.6× scale-up: AUC drops by 0.07–0.10, certificate medians rise, the red flag fires more often (0.92/0.71/0.43 against 0.80/0.48/0.19), and gated accuracy improves at every budget. Qwen2.5-7B against Qwen2.5-32B on the ten-turn vignette, identical dialogues, seeds, and scoring, offset on; $\tau = 1$ is unchanged.

	AUC (failure)		gated accuracy		certificate median		red-flag rate	
budget	7B	32B	7B	32B	7B	32B	7B	32B
20%	0.866	0.800	0.885	0.975	1.11	1.24	0.80	0.92
25%	0.841	0.742	0.600	0.823	0.99	1.09	0.48	0.71
30%	0.799	0.695	0.347	0.578	0.89	0.97	0.19	0.43

Table 23: Certificate-guided retrieval of 10% of the context recovers 3.9× the accuracy of random retrieval at the same budget. The 94 red-flagged answers of the fine-grained probe (7B, 25% budget, offset on); evicted 64-token blocks ranked by the variance contribution of their retained neighbors, which places the block holding the queried fact first in 64% of cases and in the top three in 79% (of roughly seventeen blocks).

re-answer policy	retrieved	accuracy
full archive	100%	1.000
oracle (fact block first)	10%	0.894
contribution-guided blocks (ours)	10%	**0.702**
random blocks (baseline)	10%	0.181