Error Certificates for KV-Cache Eviction via Randomized Design

Peng Xie
Technical University of Munich
p.xie@tum.de

July 2026

Abstract

Deterministic KV-cache eviction keeps the top-k tokens under an importance score and deletes the rest. We prove that this design cannot know what it destroyed: evicted values can be altered so that everything the serving system retains is unchanged while the true attention-output error grows arbitrarily, so no serving-time estimator of that error is consistent. Randomized eviction restores identifiability. With a Poisson-sampled tail at known inclusion probabilities, one logit offset performs the Hájek correction inside the softmax, and a survey-sampling variance estimator over the retained set becomes a per-step error certificate with 0.97 empirical coverage at no accuracy cost. On real workloads we pre-registered seven claims and lost three: question-aware eviction at 25–50% budgets is nearly free; output log-probability predicts failure better than the certificate; certificate-gated budget escalation adds nothing. What survives is attribution: the certificate separates cache-induced from inherent failures (AUC 0.73–0.75, against 0.47–0.54 for output confidence) and schedules recomputation better than random or confidence gating. Randomization buys attribution, 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 quickly 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: accumulated attention [36], observation windows [21], recency and sinks [32], layer-wise budgets [2, 9], and merging with compensation [35, 20], each evaluated by quality at matched budgets.

This paper asks a question that the score race leaves unexamined: after eviction, can the system know how much the eviction cost it on the current query? For deterministic selection the answer is no, and not for want of a clever monitor. Deterministic top-k keeps a set that is a function of scores; 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 therefore returns the same reading in a world where the eviction was harmless and in a world where it destroyed the answer. We call this structural property *silent failure*, and we later measure its empirical form at scale: across thousands of scored generations on real tasks, the retained-attention entropy of H2O- and SnapKV-style eviction predicts their own eviction-induced failures at AUC 0.50, with intervals tight enough to pin it to chance.

The escape comes from survey statistics rather than from a better score. If the tail of the cache is evicted by Poisson sampling with inclusion probabilities π_i that the algorithm itself chooses, then the design is known, and the classical machinery of design-based inference applies [15, 27, 33, 26]: an inverse-probability (Hájek) correction removes the selection bias of the retained softmax, and a Sen–Yates–Grundy-type variance estimator, computable from the retained set alone, is unbiased for the variance of the linearized error (Theorem 2). The correction costs one scalar per retained tail token: adding $\log(1/\pi_i)$ to the retained logit makes the softmax denominator perform the Hájek renormalization. An empirical-Bernstein radius [16, 31] turns the variance estimate into a per-step certificate whose coverage we validate directly, and an e-process construction [25, 30] extends it toward validity that is uniform over the whole autoregressive trajectory (Section 3.4).

The theory says what randomization buys in principle. The experiments say what it buys in practice, and the two answers differ in an instructive way. We committed, in writing and before the runs, to seven falsifiable claims with kill conditions (Section 6). At the attention level everything holds: coverage above target, certificate–error correlation above 0.94, and no accuracy tax relative to top-k (Section 4). On synthetic long-context suites the certificate transfers to task level: 16 of 16 model–task cells show positive certificate–failure AUC (mean 0.836). On real workloads, three pre-registered claims died honestly. Question-aware eviction at 25–50% budgets is nearly free on LongBench tasks at both 6k and 16k contexts, so overall failure is dominated by task difficulty that no cache-side signal can see; mean output log-probability, a baseline the KV literature does not report, beats the certificate at predicting failure; and certificate-gated escalation from a 25% to a 50% budget helps nowhere.

What survives both scales is the claim we now put at the center. The certificate does not tell you *whether* the answer will be wrong; output confidence does that better and for free. The certificate tells you *why*: among failures it separates eviction-induced from inherent ones at AUC 0.75 and 0.73 at the two scales, where output confidence sits at or below chance, blind by construction because it conflates cache damage with task difficulty. Attribution is actionable where prediction is not: certificate-gated recomputation beats random and confidence gating at matched compute, and confidence gating can fall below random because it spends re-runs on examples the model cannot answer anyway (Figure 4, Section 6.4). In short: **randomization buys identifiability of the compression channel: attribution, not prediction.**

Our contributions:

1. **An impossibility theorem** (Section 3): no estimator measurable with respect to the online information of a deterministic eviction scheme is a consistent estimator of the induced attention-output error, with a two-line constructive proof and an extension to value-aware scores.

2. **A design and a certificate**: certainty-plus-Poisson eviction with the Hájek correction as a logit offset, an unbiased retained-set variance estimator, and a per-step error certificate with empirically validated coverage plus an e-process extension toward time-uniform validity; zero training, no new matrices, $O(|\text{tail}|)$ scalar work per step.

3. **The silent-failure phenomenon, measured**: a panel of four online self-signals for deterministic eviction, evaluated on thousands of scored generations per signal; entropy and margin signals sit at chance, evicted score mass reaches at most 0.73, and the certificate reaches 0.77–0.86 on the same yardstick (Figure 2).

4. **A two-scale, pre-registered real-task study** with all verdicts disclosed: question-aware eviction is nearly free at 25–50% budgets, so the compression-damage regime lives in streaming

settings; output confidence owns failure prediction; the certificate owns attribution and recomputation scheduling.

Everything runs on a single H100 or H200 per shard; the full evidence chain cost roughly 70 GPU-hours.

2 Related work

Deterministic eviction and compensation. H2O [36], SnapKV [21], TOVA [23], StreamingLLM [32], PyramidKV [2], and Ada-KV [9] select retained tokens deterministically from an importance proxy and differ in the proxy or the budget allocation. CAOTE [11] scores candidates by their deterministic contribution to the attention output, folding value information into the ranking; this improves which tokens go, not what the system knows afterwards, and Remark 1 covers such value-aware summaries. Merging methods such as CaM [35] and MomentKV [20] fold evicted mass into retained representatives, which in our framing is a post-hoc stratified ratio correction of the point estimate. None of these methods maintains an estimate of its own induced error, and Theorem 1 shows none can from the information they retain.

Randomized eviction without an inference layer. MagicPIG [4] samples keys by locality-sensitive hashing and corrects with self-normalized importance sampling; it is a point estimator without a variance estimate or a certificate. Nexus sampling [6] evicts by reservoir sampling and proves, in offline analysis, that a Horvitz–Thompson estimator of retained utility is unbiased; the inclusion probabilities do not correct the attention computation and no online error estimate is produced. VASE [3] adds stochasticity to protect large-magnitude values and to diversify retention, again without probability bookkeeping. These works show that randomness is entering the eviction literature for accuracy reasons; the present paper is about what known randomness makes identifiable.

Sparse attention with guarantees, full cache retained. vAttention [5] unifies top-k and sampling and gives per-step (ϵ, δ) guarantees, and Quest [28] selects pages query-adaptively; both keep the full KV resident and can revisit any token at any step. Permanent eviction is the harder information regime: once a token is deleted, no later step can recover it, and the inference must be about a quantity that can never be recomputed. A fixed-contract diagnostic for eviction [34] studies when value-aware selection helps but provides no randomized design, no retained-set variance estimator, and no impossibility result.

Design-based inference and anytime validity. The estimator layer is classical: Horvitz–Thompson estimation [15], the Sen–Yates–Grundy variance form [27, 33], Poisson and rejective sampling [13], optimal allocation [22], and the design effect [19], as consolidated in Särndal et al. [26]. The anytime layer is modern: time-uniform confidence sequences [16], empirical-Bernstein betting bounds [31], e-values and their combination [25, 30]. To our knowledge neither body of work has been applied to permanent KV eviction.

Selective prediction. Softmax confidence is a strong baseline for failure detection [14], and risk–coverage evaluation is standard [10]. The KV-compression literature does not report these baselines. We do, and they win the prediction axis; the certificate's value is orthogonal to them.

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 $\mathcal{S}$ are deleted permanently. The *online information* $\mathcal{F}_t$ available to the serving system at step t is the retained KV pairs, the design (any sampling probabilities the mechanism used), the query history, and every quantity computed from these. 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. 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 no $\mathcal{F}_t$-measurable estimator of the eviction error is consistent, 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 $cu \cdot \sum_{i\in\mathcal{E}} a_i / \sum_{i\in\mathcal{C}} a_i$, which is nonzero and scales linearly in c. Choosing c large makes the two errors differ by more than M. □

Remark 1 (Value-aware scores). If the score reads a finite vector of scalar summaries of each value, for instance its norm [3, 34], the construction survives with summary-preserving perturbations: rotations of evicted values preserve norms while moving $\sum_{i\in\mathcal{E}} a_i v_i$ freely on a sphere whose radius the retained set cannot see. Tracking finitely many moments of the evicted set shrinks, but never closes, the free directions. Section 4 shows the constructive version: permuting evicted values in a real model changes true error by a factor of 64 while every online statistic of the deterministic method is bit-identical.

The theorem is a worst-case statement about consistency. On natural data a deterministic method's own signals may correlate with damage; whether they do is an empirical question, which Section 6 answers with a powered panel: entropy- and margin-type signals sit at chance; the one partial exception, evicted score mass, reaches AUC 0.73 at its best and still trails the certificate by more than ten points on the same yardstick.

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

The mechanism keeps a *certainty set* $\mathcal{A}$ (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 [13]). 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, the same mechanism by which MagicPIG folds its importance-sampling correction into attention [4], here serving a known-π design rather than a point estimate. No training, no new parameters, one extra scalar per retained tail token. Throughout the experiments $\varepsilon = 10^{-6}$, and the certainty layer consists of the protected positions (four attention sinks, thirty-two most recent tokens) together with every token whose proportional allocation in (2) hits the upper clip: the head of the score distribution acquires $\pi_i = 1$ without a separate threshold. Poisson sampling leaves the retained-set size random; fixed-size alternatives (rejective or conditional Poisson sampling [13]) keep the budget exact at the price of non-factoring joint inclusion probabilities and a double-sum variance estimator, and we have not evaluated them.

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 $i \in \mathcal{T}$, guaranteed constructively by the floor ε and the certainty layer, which absorbs the head of the score distribution.*

Nothing is assumed about the quality of the score. This is the validity–efficiency separation of design-based inference: a bad score concentrates π on the wrong tokens, which inflates the variance and therefore widens the certificate, but it cannot bias the coverage, because validity rests on Assumption 1 alone.

Theorem 2 (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. 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 $\mathbb{E}[\widehat{V}_t] = \mathrm{Var}(e_t^{\mathrm{lin}}) + O(B^2/m^2)$, and $\widehat{V}_t$ is computable from the retained set alone. Under Poisson sampling the Sen–Yates–Grundy double sum collapses to the single sum above, so the cost is $O(|\mathcal{S}\cap\mathcal{T}|)$ 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, and we state it rather than hide it. The theorem deliberately concerns the linearized error of the per-layer attention output. It does not claim unbiased recovery of the full-cache output, nor a bound that crosses LayerNorm, residual streams, and autoregressive sampling; the task-level meaning of the certificate is an empirical question that Sections 5 and 6 answer.

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 [16, 31] 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 deployed in every experiment is exactly this radius at $\delta = 0.1$, averaged over a head and layer subsample and maximized over the first six decode steps. We state its guarantee at the level it has earned: the radius is a design-derived statistic whose finite-sample coverage we measure directly (Section 4), not a theorem, because the reduction from the vector-valued linearized error to a scalar supermartingale involves choices whose constants we have not settled.

Remark 2 (Anytime extension). The per-step radius extends toward time-uniform validity by a standard route: an empirical-Bernstein e-process along decode steps within a head [31], Ville's inequality [29, 16] for the uniform-in-T statement, and arithmetic averaging of e-values across arbitrarily dependent heads, layers, and steps, which preserves validity [30]. Under this construction, coverage does not depend on when the certificate is read: the system may consult it at every step, act on it, and stop early. Appendix A records the construction and its two open ends, the vector-to-scalar reduction and the reuse of one tail draw across steps (block structure handles the latter conservatively). We present this as a construction with a sketch rather than a theorem, and rest the deployed certificate's validity on the measured coverage.

Remark 3 (One law, two architectures). The bias of the uncorrected retained softmax is, to first order, $\mathrm{Cov}_p(a, v)/\mathbb{E}_p[a]$ under the retention distribution: a size-biased covariance term. The same algebra drives degree bias in message-passing graph networks, where the friendship paradox [8] makes high-degree neighbors over-represented, and the Eom–Jo sign criterion [7] predicts opposite intervention directions in the two systems: graph hubs are over-counted and need down-weighting, while attention sinks drain value mass and need retention. A companion manuscript develops the graph side; the law is narrative context here and carries no load in the proofs.

Remark 4 (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 could be attached to any known-probability design. One consequence of the stratified-ratio view is testable: merging error should scale with the within-stratum dispersion of evicted values rather than with evicted attention mass. We state this as a prediction and leave the controlled test to future work.

4 The estimator works where it is defined

The first experimental question is internal validity: does the retained-set variance estimator track the true error of the attention output, at the object the theory defines? We replay prefills of Qwen2.5-1.5B [24] offline, evict at budgets $\{12.5\%, 25\%, 50\%\}$, and compare against the full-cache output at 12,096 (layer, head, query) cells, so the true error is exactly computable. Three pre-registered checks, with pass lines fixed in advance.

Coverage (pass line 0.85). At $\delta = 0.1$ the certificate covers the realized error in 96.9%, 97.2%, and 97.7% of cells at the three budgets: valid and conservative. The conservatism is quantified rather than hidden: the median certificate is roughly three times the median realized error at the 25% budget (0.0997 against 0.0317), so the certificate is loose as an absolute bound and strong as a ranking signal, which is the property the rest of the paper uses.

Correlation (pass line 0.3). Spearman correlation between the certificate and the true error is 0.943, 0.965, and 0.979 across budgets.

Accuracy at equal budget. The randomized design does not pay an accuracy tax at the attention level; it collects one. Median relative error at the 25% budget is 0.0317 for Poisson-with-Hájek against 0.0447 for top-k and 0.2386 for uniform sampling, and the ordering is the same at every budget. Half of this gap is the Hájek correction itself; ablating it (sampling without the logit offset) forfeits the gain.

The impossibility, constructively. Applying six random permutations to the evicted values realizes Theorem 1 in a real model: the true error of the top-k output ranges from 0.014 to 0.898 across permutations, a factor of 64, while every retained-set statistic of the top-k method is bit-identical across all six worlds. The certificate, which the design makes possible, covers the realized error in all permutations.

5 From attention error to task failure: synthetic suites

The certificate is defined on attention outputs; tasks are what users see. Two synthetic suites bridge the gap under conditions where eviction is known to bite.

Needle retrieval. On passcode retrieval with generation-time eviction (Qwen2.5-1.5B, 252 runs) the regime matters more than the method: with the question available before compression (*aware*), H2O succeeds fully and Poisson succeeds at 88.5%; with compression before the question arrives (*stream*), both collapse to zero at these budgets. The asymmetry is in self-knowledge. Asked to predict its own failures, H2O's retained entropy scores AUC 0.405; with 24 failures and 24 successes this is statistically indistinguishable from chance ($z = -1.15$), and we flag that the informative claim is "no better than chance," not "below chance." The certificate scores 0.812 ($n = 192$; $z = 10.2$), with a median value of 0.99 on failures against 0.57 on successes: a red light that turns on.

A four-task benchmark. Across four base models (Qwen2.5-1.5B/7B, Llama-3.1-8B [12], Mistral-7B-v0.3 [18]) and four RULER-style tasks [17] at 640 scored generations per cell, the certificate–failure AUC is positive in 16 of 16 cells, each 95% interval excluding 0.5, mean 0.836, range 0.65–0.97. Thresholding the certificate yields a usable risk–coverage knob: error 0.229 at 30% coverage against 0.575 at full coverage. Two honest entries from the same suite: the aware-condition exploration tax concentrates in one cell type, multi-needle retrieval at the 25% budget, where spreading π over twenty needles costs 39–44 points on three of four models, and is near zero elsewhere; and all importance-based methods, ours included, collapse in the stream condition at these budgets, where only recency windows retain the answer by luck.

Synthetic suites are engineered so that eviction destroys information. Whether real workloads put a system in that regime is exactly the question the next section pre-registers.

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 [1] 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 online certificate (two seeds, stream and aware conditions), at

	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 (max 0.73 vs. cert 0.855; entropy at chance)
R3	Certificate $\geq$ output logprob on induced-failure prediction	logprob strictly dominates	not met (tie at 6k; logprob wins at 16k)
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	no rise: free at 16k (4.0% at 25%)
S2	Certificate AUC recovers (≥ 0.6) at 16k	< 0.6 = final	**final kill** (0.572)
S3	Exploration tax reappears at 12.5%	tax < 3pp confirms	no tax (+0.3pp)

Table 1: All seven pre-registered claims and their outcomes. Three died; what survived is the attribution result of Section 6.4.

budgets $\{25\%, 50\%\}$ with 6k-token contexts (tasks: HotpotQA, 2WikiMQA, MultiFieldQA-en, PassageRetrieval-en, plus a synthetic needle anchor) and $\{12.5\%, 25\%, 50\%\}$ with 16k-token contexts (HotpotQA, PassageRetrieval-en matched to the 6k sampling, plus MuSiQue and NarrativeQA), roughly 74,000 generations in total. Every compressed run logs a panel of online self-signals: retained-attention entropy, evicted score mass, keep-boundary margin, mean output log-probability, and, for Poisson arms, the certificate. Success is token-F1 ≥ 0.5 for QA (raw F1 logged; the 0.3 threshold moves no verdict), exact match for retrieval and needle tasks. *Eviction-induced* failure means the full-cache run answers correctly and the compressed run does not.

Before any full shard ran we fixed four claims with kill conditions (R1–R4), and before any 16k shard ran, three scale hypotheses (S1–S3). Table 1 lists all seven with outcomes; the pre-registration files are reproduced in Appendix C.

6.2 Why the prediction story died: the damage is not there

The certificate's pooled failure-prediction AUC on LongBench is 0.555 at 6k and 0.572 at 16k, under the pre-registered kill line of 0.6 both times, and the decomposition explains why. In the aware condition the AUC is 0.500 at 6k: chance. The reason is visible in the accuracy grid: question-aware eviction barely damages anything. 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; 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 (Figure 1). Failures on these tasks are overwhelmingly inherent, the model simply cannot answer, and a cache-side signal is correct not to see them. S1 rules out the truncation explanation: on the two tasks shared with the 6k suite under identical example sampling, moving from 6k windows, which cut the median HotpotQA context by half, to 16k windows, which fit it entirely, leaves the rate at 4.0% at the 25% budget against 5.6% at 6k (7.3% at 12.5%): no rise. A no-context control rules out the memorization explanation: answering with the context removed succeeds on 0–32% of examples depending on task and model, 75–84% of full-correct examples are context-dependent in the sense that the same model fails them without the context, and restricting the induced-failure rates to those context-dependent examples moves them by about one point (16k aware Poisson: 6.2% to 6.9% at the 25% budget, 10.5% to 11.5% at 12.5%; Appendix B), so memorized answers do not manufacture the free-compression finding.

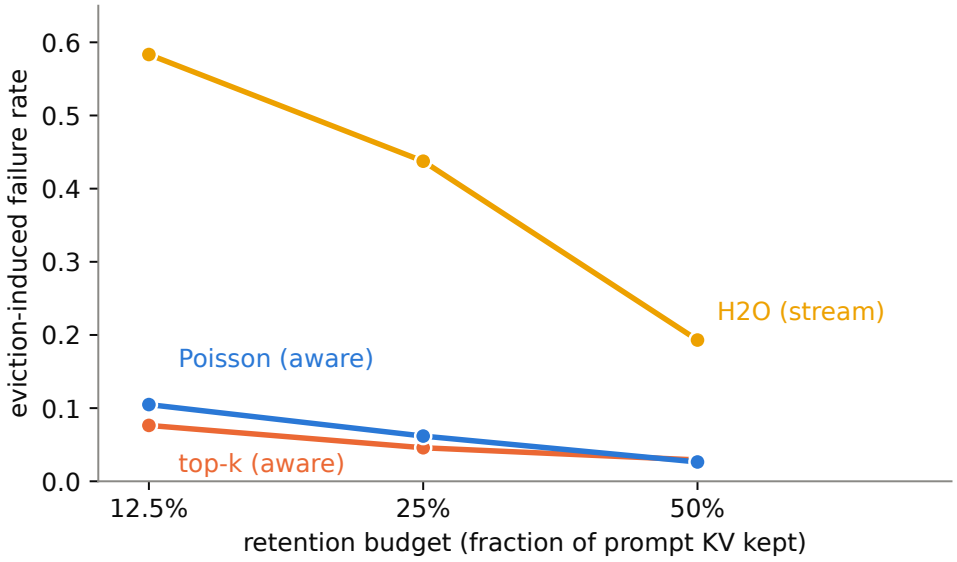

Figure 1: Eviction-induced failure rate (among examples the full cache answers correctly) at 16k contexts, pooled over four LongBench tasks and four models. Question-aware eviction is nearly free even at a 12.5% budget; the damage regime is streaming eviction, where the question is not yet known when tokens are deleted.

The damage regime exists, but it lives elsewhere: streaming eviction, where tokens are deleted before the question arrives, breaks 40–58% of answerable examples on the same tasks (Figure 1). This is the setting of multi-turn assistants and agent memory, where history is compressed before future queries are known, and it is where every result in the rest of this section concentrates.

Two further pre-registered outcomes complete the honest picture. The randomized design costs nothing on real tasks: the aware-condition gap between deterministic top-k and Poisson is at most 0.7 points at 6k and 0.3 points at 16k, at every budget including 12.5% (S3), so the multi-needle tax of Section 5 does not generalize beyond its synthetic construction. And mean output log-probability predicts failure better than every cache-side signal, certificate included, on every axis we pre-registered: overall failure (0.73–0.80 against 0.56–0.57 LongBench-pooled), induced failure (0.782 against 0.778 at 6k; 0.806 against 0.768 at 16k, paired difference -0.038, 95% CI $[-0.065, -0.010]$), and risk–coverage. Selective answering should be gated on output confidence [14, 10], not on the certificate, and a KV-compression paper that evaluates a trust signal without this baseline overstates its case.

6.3 The silent-failure panel

R2 is the claim the theory stakes out, and it survives with texture. Figure 2 shows the panel: every self-signal available to a deterministic evictor, scored on predicting that evictor's own induced failures, pooled over both scales (several thousand scored generations per row). Retained-attention entropy, the signal a practitioner would reach for first, sits at 0.43–0.51 for StreamingLLM, H2O, and SnapKV (two-scale pooled values in Figure 2; 6k-suite values with intervals in Table 3): the powered version of the needle anecdote, now with confidence intervals that close the question. Keep-boundary margin sits at 0.49–0.50 everywhere. Evicted score mass is the honest exception: it carries partial signal (0.59–0.73 across arms and suites), consistent with Theorem 1, which bounds

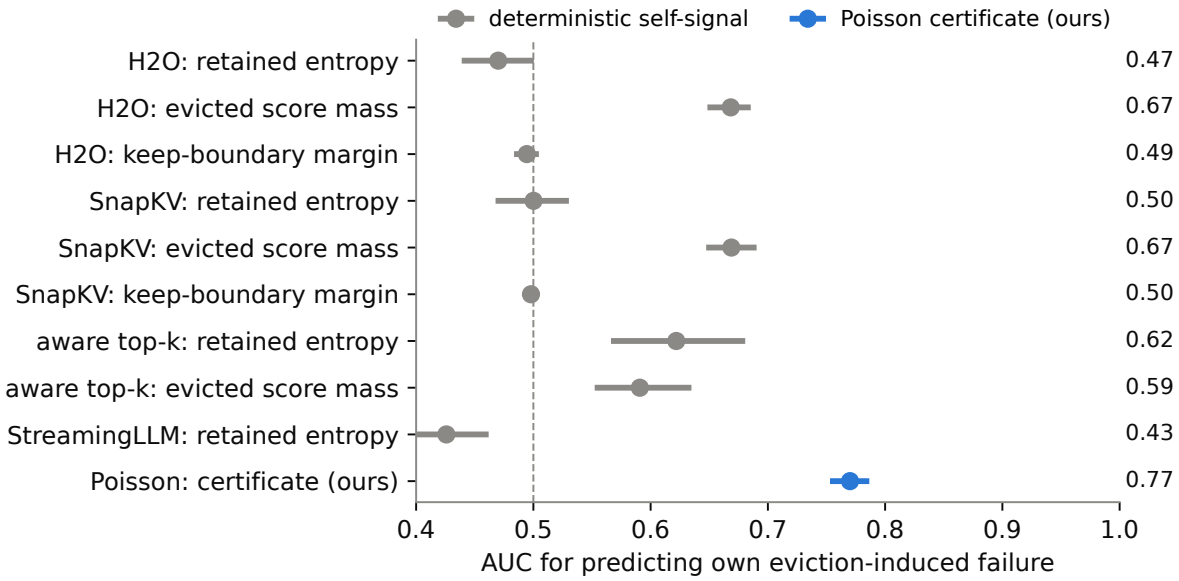

Figure 2: The self-signal panel on eviction-induced failures, pooled over 6k and 16k LongBench suites, with 95% cluster-bootstrap intervals (resampling examples). Deterministic entropy and margin signals sit at chance; evicted mass is partially sighted; the certificate, which only a known randomized design can provide, leads by ten points.

what is identifiable in the worst case rather than on natural data. It still trails the certificate by ten points, and its information washes out for overall failure prediction (0.59 pooled), where task difficulty dominates.

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 pose: *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) and 0.727 at 16k ($n = 11{,}717$, 1,776 induced); output log-probability scores 0.542 and 0.469 (Figure 3). The asymmetry is structural, not incidental: low confidence flags hard examples 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. This is Theorem 2 doing its exact job at task level: the design identifies the error of the channel the design controls, and nothing else.

A certificate-free attributor suggests itself: run the query twice with independent Poisson draws and compare. The two-seed grid evaluates it. Generated text is not logged, so we score the label-assisted proxies, whether the partner draw succeeded and the F1 gap between draws, which can only flatter the approach. Among failures they attribute at 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, and recomputation gated on draw disagreement (which fires on 15% of queries) trails certificate gating at the matched rate in every configuration. Running twice also doubles generation cost and carries no validity statement. The certificate is therefore the strongest online attributor we measured and the only one whose validity is tied to the design; evicted score mass (0.59–0.73, Figure 2 and Table 3) and draw disagreement are partial substitutes, not replacements.

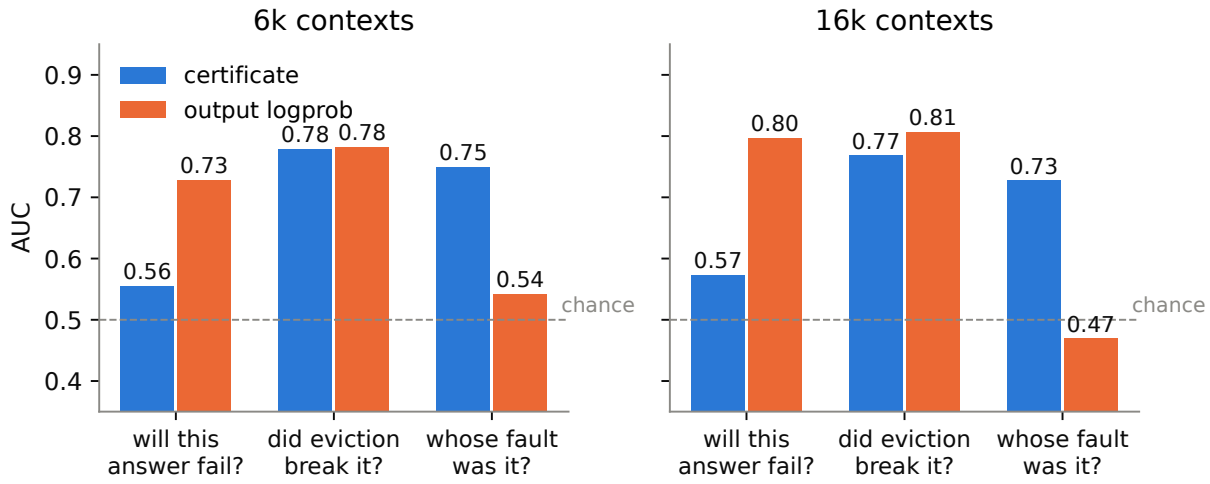

Figure 3: Certificate against output log-probability on three questions, LongBench-pooled at each scale. Output confidence wins failure prediction (left, center); the certificate wins attribution by a wide margin (right), where confidence sits at or below chance.

Attribution converts directly into a scheduling product. Consider serving under streaming compression with a recomputation budget: a quarter of queries may be re-run with the full cache. Both Poisson seeds replicate the result independently, and random gating is scored by its exact expectation rather than a single draw (Table 4). At the 12.5% budget (16k), certificate gating recovers accuracy 0.33 against 0.29 for random and 0.28–0.29 for confidence gating; the per-seed advantage over random is +3.7 points with 95% intervals $[+2.3, +5.0]$, and over confidence +4.4 to +4.9 points, intervals excluding zero. At the 25% budget the ordering is unchanged (+1.9 to +2.2 over random, intervals excluding zero; Figure 4), and the 6k suite reproduces it (+2.7 to +3.0 over random, intervals excluding zero; the margin over confidence there, +1.3 to +1.4, does not separate from zero). Confidence gating adds little and lands *below* random at the harsh budget because it spends re-runs on inherently hard examples, which recomputation cannot save; the certificate spends them on cache-damaged examples, which it can. This analysis was not pre-registered; it was designed after R4's autopsy showed that escalation to a larger compressed budget fails because a 50% stream cache is still broken, and we label it accordingly.

6.5 Cost

The prototype computes the certificate in Python, per layer, on a head subsample, during the first six decode steps. Median decode time is 0.043 s/token against 0.023 for deterministic eviction arms and 0.015 for the full cache: an overhead factor of roughly two in this instrumentation, far from the $O(|\text{tail}|)$ scalar cost the design admits, and we report the measured number rather than the asymptotic promise. A fused implementation is engineering, not research, and until it exists the certificate should be priced at the measured overhead.

7 Discussion

Guidance by regime. The two-scale study supports three concrete recommendations. First, for prefill-style compression where the query is known, deterministic or randomized eviction at

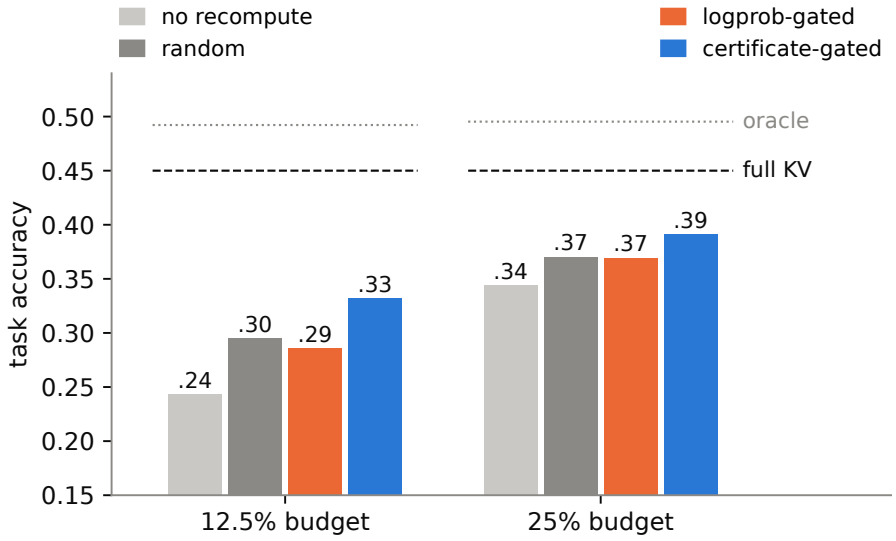

Figure 4: Certificate-gated recomputation under streaming compression at 16k, re-running 25% of queries with the full cache; bars average the two Poisson seeds and random gating is its exact expectation. The certificate beats random at both budgets (per-seed intervals excluding zero, Table 4); confidence gating adds little and falls below random at 12.5% because it cannot tell cache damage from task difficulty.

25–50% budgets is close to free on LongBench-type workloads, randomization costs nothing, and no certificate is needed; monitoring belongs at harsher budgets. Second, for answer-level trust, gate on output confidence; every cache-side signal, ours included, loses that comparison. Third, for streaming and agent-memory settings, where history is compressed before queries are known and roughly half of answerable queries can silently break, deterministic methods cannot see the damage (Figure 2), and the certificate is the strongest online attributor we measured and the only one whose validity is tied to the design, at a cost of one extra scalar per retained tail token and, in this prototype, a twofold decode overhead.

What randomization buys. The theory promised identifiability; the experiments locate its value. Randomizing the tail did not make eviction more accurate on real tasks (it is not worse either), and it did not yield a better failure predictor than free output confidence. It made one quantity estimable that deterministic designs provably cannot estimate, the error injected by the compression channel itself, and the two places that quantity is worth money, damage monitoring and recomputation scheduling in streaming regimes, are exactly the places the experiments certify.

Limitations. Contexts reach 16k tokens and models 8B parameters; longer contexts and larger models could move the damage-onset curve, although the 6k-to-16k direction moved it down, not up. The stream condition approximates multi-turn memory with single-turn tasks whose question arrives after compression; agent benchmarks with genuine multi-turn structure are the right next test. Success thresholds (F1 ≥ 0.5) are a choice; the pre-registered sensitivity check at 0.3 moves no verdict. The merging prediction of Section 3 is stated, not tested. The certificate concerns per-layer attention error, and its task-level meaning is established empirically, not by a bound that crosses the network; the anytime extension is a construction with a sketch, and the deployed per-step

certificate rests on measured coverage. The head and layer subsample, the six-step window, and the normalizer floor ϵ_0 are unablated implementation choices, and fixed-size sampling designs are untested. Certificate overhead is measured at a factor of two in an unoptimized prototype.

Acknowledgments

Experiments presented in this work were carried out using the CIT-TUM-HN cluster at TUM Campus Heilbronn.

Reproducibility

All experiments run on single H100 or H200 GPUs (roughly 70 GPU-hours total). Per-run JSON logs (about 98,000 scored generations across the task suites and controls, plus 12,096 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. Models and datasets are public (Qwen2.5, Llama-3.1, Mistral-7B; LongBench, RULER-style generators). Large language models assisted with experiment code, analysis scripting, and drafting; every reported number is regenerated by the released scripts from the released logs, and every citation was verified against the cited source.

References

[1] 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 ACL*, 2024. arXiv:2308.14508.

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

[3] 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.

[4] 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*, 2025. arXiv:2410.16179.

[5] Aditya Desai, Kumar Krishna Agrawal, et al. vAttention: Verified sparse attention. *arXiv preprint arXiv:2510.05688*, 2025.

[6] 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.

[7] Young-Ho Eom and Hang-Hyun Jo. Generalized friendship paradox in complex networks: The case of scientific collaboration. *Scientific Reports*, 4:4603, 2014.

[8] Scott L. Feld. Why your friends have more friends than you do. *American Journal of Sociology*, 96(6):1464–1477, 1991.

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

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

[11] 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.

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

[13] Jaroslav Hájek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35(4):1491–1523, 1964.

[14] 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. arXiv:1610.02136.

[15] Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[16] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics*, 49(2):1055–1080, 2021.

[17] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. arXiv:2404.06654.

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

[19] Leslie Kish. *Survey Sampling*. Wiley, 1965.

[20] 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.

[21] 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*, 2024. arXiv:2404.14469.

[22] Jerzy Neyman. On the two different aspects of the representative method. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934.

[23] Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. Transformers are multi-state RNNs. In *Proceedings of EMNLP*, 2024. arXiv:2401.06104.

[24] Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[25] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

[26] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer, 1992.

[27] Amode R. Sen. On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5:119–127, 1953.

[28] Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. Quest: Query-aware sparsity for efficient long-context LLM inference. In *International Conference on Machine Learning*, 2024. arXiv:2406.10774.

[29] Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, 1939.

[30] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *Annals of Statistics*, 49(3):1736–1754, 2021.

[31] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B*, 86(1):1–27, 2024.

[32] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. arXiv:2309.17453.

[33] Frank Yates and P. Michael Grundy. Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society: Series B*, 15(2): 253–261, 1953.

[34] 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.

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

[36] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems 36*, 2023. arXiv:2306.14048.

A Proofs

A.1 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. 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 $e_t^{\text{lin}} = \sum_{i\in\mathcal{T}}(I_i/\pi_i - 1)\, x_i$. Independence across i gives

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

the Poisson-sampling (single-sum) case of the Sen–Yates–Grundy form [27, 33]: joint inclusion probabilities factor, $\pi_{ij} = \pi_i \pi_j$, so the double sum vanishes. 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 [15]. $\widehat{V}_t$ replaces the unknown x_i by $a_i(v_i - \hat{y}_t)/\widehat{N}$. A first-order expansion of the Hájek ratio [26] gives $\hat{y}_t - y_t = e_t^{\text{lin}} + O_p(B^2/m^2)$ and $\widehat{N}/N = 1 + O_p(B/m)$ under Assumption 2, since each summand contributes at most B/m after normalization and the tail sample has expected size m. Propagating both substitutions through the quadratic form perturbs the expectation by $O(B^2/m^2)$, which is the stated remainder. □

A.2 The anytime construction of Remark 2 (sketch)

This records the construction and its open ends; it is a sketch, not a proof, and the deployed certificate rests on the measured coverage of Section 4. Fix a head. For each t the summands $\{(I_i/\pi_i - 1)\, x_i\}_{i \in \mathcal{T}}$ are bounded by Assumption 2 and mean-zero given the past, so the empirical-Bernstein supermartingale of Waudby-Smith and Ramdas [31] applied to the running sums yields a process M_t with $\mathbb{E}[M_t] \le 1$ under the null that the realized error stays within (4); Ville's inequality [29] converts M_t into the time-uniform statement, and the boundary in [16] gives the stated radius shape (a variance term plus a range term, both computable from the retained set). If tail indicators are redrawn every T_r steps the increments are independent across blocks; if the same draw is reused, increments within a block are identical and the product is taken over blocks, which only loosens the bound. 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 [30]; the linearization remainder of Theorem 2 adds the $O(B^2/m^2)$ slack. □

B 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, and apply the $+\log(1/\pi_i)$ logit offset. The certificate is (4) at $\delta = 0.1$ computed on every fourth head 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 $\epsilon_0 = 10^{-9}$.

Signals. Retained-attention entropy: decode-step attention entropy over the retained set, normalized by log(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 and needle by exact match. 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 1 is unchanged.

Statistics. AUCs are Mann–Whitney; intervals are 95% cluster bootstrap resampling examples (500 draws), so repeated measurements of one example never inflate significance. 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 2.

C Pre-registration record

The registration file was written on 2026-07-22 before any full 6k shard completed, and its scale addendum before any 16k shard ran; both are timestamped in the released package alongside the Slurm submission records. Abbreviated here:

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: overall-failure panel maxima 0.59 (evicted mass); induced-failure panel maxima 0.73 (6k SnapKV evicted mass; 0.67 two-scale pooled), against certificate 0.855 (6k, lower CI 0.84); kill not triggered under either reading; the partial visibility of evicted mass is reported.

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 $[-0.065, -0.010]$). 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. The recomputation analysis of Section 6.4 is post hoc.

S1–S3 as in Table 1: induced-failure rate at 16k/25% is 4.0% against 5.6% at 6k/25% (no rise; 12.5% reaches 7.3%); certificate AUC 0.572 (< 0.6, final); aware tax +0.3pp at 12.5%.

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

D Additional tables

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 2: R1 detail: certificate-to-own-failure AUC, LongBench only. Every model exceeds chance; none reaches the pre-registered 0.6 line. Cell-level: 24/32 cells above 0.5 at 6k, 30/32 at 16k.

Arm	certificate	entropy	evicted mass	margin	output logprob
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	.645 [.63,.66]	.527 [.50,.55]	–	–	.794 [.78,.81]
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	.855 [.84,.87]	.446 [.43,.47]	–	–	.848 [.83,.86]

Table 3: The full self-signal panel: AUC predicting the arm's own failure, with 95% cluster-bootstrap intervals (resampling examples). Output log-probability is strong everywhere, which is why the paper concedes prediction to it; no retained-set signal of a deterministic arm approaches the certificate on induced failures.

Suite/budget	seed	base	full	cert	logprob	random	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 4: 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. 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.

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 5: Accuracy by task at 16k, pooled over models (aware condition for aware/Poisson columns; H2O is the stream condition). Question-aware compression tracks the full cache; streaming compression does not.