AUDITING KV-CACHE RETENTION WITH SAVED ATTENTION OUTPUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can saved attention outputs certify a retention decision after the original KV cache is released? We study the evidence needed for this task. For all group-retention outputs at one query, we attain the minimum native observation count at every Value width. Finite-error bounds quantify the stability of recovery from noisy records. With a retained reconstruction, an exact normalizer-residual identity constrains one shared physical Key–Value error using causal output records. It yields grid-free convex certificates for later queries, masks and output directions. These inputs may be selected after recording. Known appended rows enter without additional uncertainty. Checking fixed coefficients is linear in cache size at fixed record count. A sensitivity condition and same-code source witnesses establish strict information gain. At a fixed codec, frozen-model checks verify strict information gains and certify complete-head retention decisions. Matched-byte controls separately evaluate how evidence competes with finer quantization for storage. Together, the theory and checks quantify the decision value of saved outputs for a declared workload.

1 INTRODUCTION

Transformer decoders store earlier keys and values in a Key–Value (KV) cache for later attention calls. Compression and eviction reduce storage, but they also remove information before the next query is known. Once the original rows have been released, even checking a proposed deletion becomes difficult: the intact attention output is no longer available for comparison. A compressed cache still provides candidate outputs. Their errors depend on the missing source. This matters when one prefix serves several later queries or its retention budget changes after compression. For example, consider a 32-row prefix and a growing live suffix. The owner compresses the prefix and saves four causal attention outputs with their queries. It then releases the original prefix. A later query arrives. A rule now proposes retaining only the latest 16 old rows. Can the saved information bound the compressed proposal's error relative to the original intact cache? The bound must cover every compatible source and compare the same realized query and exact live suffix. The proposal can change with the query or memory budget. Recording must therefore precede the choice of the decision it supports.

Normalized attention outputs hide their total weights, so outputs from separate groups cannot in general be merged. A compressed reconstruction supplies more structure, but leaves uncertainty in both keys and values. Saved outputs constrain that uncertainty. The essential constraint is shared: every record must be explained by the same original cache. Our central question is therefore: *which future retention decisions are determined, or bounded, by a given saved record?* The answer depends on the workload. A record may determine all group unions at its own query, yet leave a later query ambiguous. A certified interval measures that remaining uncertainty. Each saved output excludes incompatible caches. The audit transfers these shared constraints to the later task and tests every remaining output against the chosen error budget.

Two workloads lead to different guarantees. For fixed-query group retention, we attain the minimum number of native output observations needed to recover every group union. For later queries and masks, an exact normalizer-residual identity turns saved outputs into constraints on the reconstruction error. It yields finite convex certificates without enumerating possible caches. A physical witness test and a sensitivity condition determine when the records strictly reduce ambiguity. Both results quantify how recorded outputs constrain the same unknown cache. Frozen-model checks connect

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.

interval contraction to certified retention decisions. Matched-byte comparisons then examine how a storage budget can be divided between evidence and codec precision.

2 Related work

KV-cache methods reduce storage through row selection and numerical quantization. H_2O balances recent tokens with heavy hitters and analyzes eviction under a dynamic submodular model (Zhang et al., 2023). FastGen profiles attention heads to choose their compression policies (Ge et al., 2024). SnapKV uses a prompt observation window to select positions for each head (Li et al., 2024). StreamingLLM preserves initial attention-sink tokens together with recent tokens to sustain streaming inference (Xiao et al., 2024). KVzip uses context reconstruction to score rows for reuse across future queries (Kim et al., 2025). Quantization trades numerical precision for storage. KIVI uses different quantization axes for Keys and Values (Liu et al., 2024b). KVQuant combines distribution-aware quantization with separate treatment of outliers (Hooper et al., 2024). GEAR supplements quantization with low-rank and sparse corrections to the reconstruction error (Kang et al., 2024). These methods can supply candidate caches when their retained cells and error bounds satisfy the audit's input contract. We bound a later attention output over all original caches compatible with the retained evidence. The codec determines physical uncertainty, and saved outputs impose additional constraints. Their storage cost motivates our comparison with finer quantization. The same saved outputs can certify candidates across future tasks.

Streaming summaries approximate attention from the observed K/V stream. SubGen uses clustered Keys and sampled Values (Zandieh et al., 2024). BalanceKV uses discrepancy-based compression with probabilistic error bounds (Kochetkova et al., 2025). Near-tight space bounds characterize streaming attention in specified norm and dimension regimes (Chen et al., 2026). Query-oblivious coresets seek prescribed accuracy over a whole query ball (Cohen, 2026). These works construct summaries under specified access and accuracy models. The retained codec and records determine an interval for each later query and mask. This interval supports the later retention test. The native-call count measures exact output observations at one fixed query. A streaming summary reads cache rows; our native interface observes specified attention calls. The physical certificate combines those observations with a retained reconstruction.

Transformer verification propagates perturbation bounds through attention and other network operations (Shi et al., 2020). Runtime-certified quantized attention bounds current-call errors and keeps originals in host memory for dense fallback (Calver, 2026). Our evaluator uses the retained reconstruction and records after the original prefix is released. ReST-KV selects evictions by output reconstruction and smoothing (An et al., 2026); such a proposal can be checked by our certificate. Historical output constraints couple the records through one shared, unknown cache. The resulting certificate concerns attention at the supplied state and query. It can therefore exploit evidence across attention calls. The construction uses normalization-aware merging from online softmax (Milakov & Gimelshein, 2018; Dao et al., 2022; Liu et al., 2024a; FlashInfer Contributors, 2025). Optimal recovery supplies the linear information benchmark (Micchelli & Rivlin, 1977; Plaskota, 1996; Ettehad & Foucart, 2021), and convex duality supplies support certificates (Boyd & Vandenberghe, 2004). These established tools explain the merge identities and convex checking steps. The attention-specific results are the attainable native observation counts and the finite normalizer-residual bounds conditioned on saved outputs.

3 Background and the retention audit

An attention head forms a weighted average of the values at visible positions. A new token supplies a query and adds its own key and value. Earlier tokens supply the cached rows. Keys determine the weights. Values supply the vectors being averaged. We study one head first.

At position i, write $k_i \in \mathbb{R}^{d_k}$ for the Key and $v_i \in \mathbb{R}^{d_v}$ for the Value. The matrices of these rows form $X = (K, V)$. For a query $q \in \mathbb{R}^{d_k}$, let I contain the old positions that may be removed. Let B contain the protected positions visible in this call. These can include the live suffix. A retention mask

$S \subseteq I$ keeps $S \cup B$, which must be nonempty. Attention is

$$Z_S(q) = \sum_{i \in S \cup B} e^{q^\top k_i / \sqrt{d_k}}, \qquad \alpha_i(q, S) = \frac{e^{q^\top k_i / \sqrt{d_k}}}{Z_S(q)},$$
$$y_S(q; X) = \sum_{i \in S \cup B} \alpha_i(q, S) v_i, \quad \sum_{i \in S \cup B} \alpha_i(q, S) = 1. \tag{1}$$

Thus deleting a row changes both the weighted numerator and its normalizer Z_S. The intact reference keeps $S = I$. In the running example, I has 32 old rows and the proposed mask has 16.

Before releasing the original prefix, the owner retains a compressed representation and selected outputs. A *codec* consists of this representation and its decoder; its reconstructed cache is $\bar{X}$. Quantization leaves a set of possible original values. For example, an unclipped exact nearest-grid reconstruction with spacing Δ gives $|x - \bar{x}| \le \Delta/2$ for a scalar coordinate x. The coordinate cells form $\mathcal{C}$. A verified radius R can further bound the joint reconstruction error. Known live rows have singleton cells.

Record e stores its query q_e, visible positions $T_e = S_e \cup B_e$, output $\widehat{y}_e$ and vector-error allowance ϵ_e. Write $y_e(X) = y_{S_e}(q_e; X)$ using that call's protected set B_e. Different causal records can have different visible prefixes. With $e = 1, \ldots, N_{\mathrm{rec}}$, the compatible cache sets are

$$\mathcal{P}_X = \{X \in \mathcal{C} : \|K - \bar{K}\|_F^2 + \|V - \bar{V}\|_F^2 \le R^2\},$$
$$\mathcal{P}_Y = \{X \in \mathcal{P}_X : \|y_e(X) - \widehat{y}_e\|_2 \le \epsilon_e \ \forall e\}. \tag{2}$$

Here vector norms are Euclidean and $\|\cdot\|_F$ is the Frobenius norm; matrix $\|\cdot\|_2$ below is the operator norm. The first set expresses codec uncertainty. The second intersects it with all saved-output constraints on the same cache. Cells, radii and error allowances are retained inputs; their acquisition contract is given in Appendix A. In the example, every candidate source must reproduce the four saved outputs within their error allowances. A source excluded by one record is excluded from every later audit, making the record bank reusable.

After recording, an evaluator receives a query, a mask and an output direction $a \in \mathbb{R}^{d_v}$ with $\|a\|_2 = 1$. The scalar task is $f_{q,S,a}(X) = a^\top y_S(q; X)$. It seeks an interval containing $f_{q,S,a}(X)$ for every $X \in \mathcal{P}_Y$, without access to the original prefix. Coordinate directions supply bounds $[L_j, U_j]$ on each component of the intact reference $y_I(q; X)$. Let $\widetilde{y}_S$ be the executed compressed-mask output. For an allowed head error τ, a sufficient retention test is

$$\left[\sum_{j=1}^{d_v} \max\{|\widetilde{y}_{S,j} - L_j|, |\widetilde{y}_{S,j} - U_j|\}^2\right]^{1/2} \le \tau. \tag{3}$$

Passing certifies the proposal for every compatible source cache. Failure to pass leaves it unresolved. An empty $\mathcal{P}_Y$ instead signals inconsistent inputs. The query, mask and direction may be chosen after the records; the live suffix is supplied exactly when evaluated.

4 Exact recovery at a recorded query

At a fixed query, future masks retain whole groups. Exact recovery is possible for this workload. Its cost depends on which quantities the recording interface preserves. Normalization combines each group's weighted mean with its relative attention mass. The observation count measures the output information needed to recover these quantities.

Output-only records save y_S. Normalization-aware records also save $\log Z_S$. Internal exports can supply group masses and means directly. PyTorch SDPA returns the output (PyTorch Contributors, 2025); FlashAttention can also return LSE (FlashAttention Contributors, 2026).

Fix $q \neq 0$ and a nonempty background B. Partition I into g nonempty groups $G_1, \ldots, G_g$ and set $G_0 = B$. Let P_i be group i's exponential mass and μ_i its weighted Value mean. Set $p_i = P_i / P_0$. A group mask now means $S \subseteq \{1, \ldots, g\}$:

$$y_S = \frac{\mu_0 + \sum_{i \in S} p_i \mu_i}{1 + \sum_{i \in S} p_i}. \tag{4}$$

The missing masses matter even for two groups. With $\mu_0 = 0$, the following two sources give the same singleton outputs:

$(p_1, p_2; \mu_1, \mu_2)$	$y_{\{1\}}$	$y_{\{2\}}$	$y_{\{1,2\}}$
$(1, 1; 2, 4)$	$2/2 = 1$	$4/2 = 2$	$6/3 = 2$
$(3, 1; 4/3, 4)$	$4/4 = 1$	$4/2 = 2$	$8/5$

Thus the singleton records alone leave the joint output ambiguous. Write $y_i = y_{\{i\}}$ and let y_{all} retain all groups. For $g > 1$, define $\omega_i = (1 + p_i)/(1 + \sum_j p_j)$, $\gamma_S = (|S| - 1)/(g - 1)$ and $\xi_{S,i} = \mathbf{1}\{i \in S\} - \gamma_S$. Put $\beta = (1 + \sum_i p_i)^{-1}$. The singleton and intact equations give

$$\sum_i (1 + p_i) y_i - (1 + \sum_i p_i) y_{\text{all}} = (g - 1)\mu_0,$$
$$\sum_i \omega_i - 1 = (g - 1)\beta, \qquad \sum_i \omega_i y_i - y_{\text{all}} = (g - 1)\beta\mu_0,$$
$$\beta\big(\mu_0 + \sum_{i \in S} p_i \mu_i\big) = \sum_{i \in S} \omega_i y_i - (|S| - 1)\beta\mu_0 = \sum_i \xi_{S,i}\omega_i y_i + \gamma_S y_{\text{all}}.$$

Division by the positive normalized mass yields

$$y_S = \frac{\sum_i \xi_{S,i}\omega_i y_i + \gamma_S y_{\text{all}}}{\xi_S^\top \omega + \gamma_S}, \qquad \xi_S^\top \omega + \gamma_S = \frac{1 + \sum_{i \in S} p_i}{1 + \sum_i p_i} > 0. \tag{5}$$

An extra proper union J supplies the linear system

$$M_J \omega = r_J, \qquad M_J = [\xi_{J,i}(y_i - y_J)]_i, \qquad r_J = \gamma_J(y_J - y_{\text{all}}). \tag{6}$$

Full column rank determines every union when $3 \le g \le d_v$. For instance, three singletons, intact and 12 can determine 13 and 23.

The lower bound needs independently varying masses and means. Use their positive-mass parameter space. Alternatively, assume an open summary chart U and a C^1 realization $\Xi : U \to \operatorname{int} \mathcal{P}_X$. The interior is relative to the mutable coordinates. Task injectivity means that the full union-output family distinguishes the chart's parameters.

Theorem 1 (Minimum output-only observations)**.** *For $g \ge 3$, $d_v \ge 1$ and $q \ne 0$, permit nonempty group masks and exact d_v-vector responses on this source class. Fixed mask families including intact recover every union outside an exceptional rank set with empty interior using*

$$N_{\text{known}} = g + \lceil g/d_v \rceil, \qquad N_{\text{hidden}} = g + 1 + \lceil g/d_v \rceil \tag{7}$$

calls when μ_0 is supplied and unknown, respectively. These counts are necessary on an open task-injective chart. The lower bound also permits deterministic adaptive calls with C^1 responses of width at most d_v. For $g = 1$, both counts are one. For $g = 2$, both are three.

The count separates g unknown masses from g or $g + 1$ vector means. Their dimensions are $D_{\text{known}} = g + gd_v$ and $D_{\text{hidden}} = g + (g+1)d_v$. On a constant-response manifold, the next adaptive action is fixed. The constant-rank theorem then gives, with response ranks $s_e \le d_v$,

$$\dim M_N = D - \sum_{e=1}^{N} s_e \ge D - Nd_v; \qquad \dim M_N = 0 \implies N \ge \lceil D/d_v \rceil.$$

Here M_N is a local set of sources with one common transcript. Task injectivity requires dimension zero for exact recovery. For attainability, let H be recovered groups, $p_H = \sum_{i \in H} p_i$, and append a block J of $r \le d_v$ groups. Its singletons and one union give

$$\sum_{i \in J} (y_i - y_{H \cup J}) p_i = (1 + p_H)(y_{H \cup J} - y_H) - \sum_{i \in J} y_i + r\mu_0.$$

Full column rank recovers the r new masses; singleton equations recover their means. Each such block costs $r + 1$ calls. Appendix B supplies the full-rank seeds and local rank selection. The construction attains the classical dimension bound for native output observations.

4.1 STRONGER RECORDS AND FINITE ERRORS

For $g > 1$, set $b_S = Z_S y_S$. Singleton and intact output–LSE pairs give

$$Z_0 = \frac{\sum_i Z_i - Z_{\text{all}}}{g-1}, \qquad b_0 = \frac{\sum_i b_i - b_{\text{all}}}{g-1}, \qquad (Z_{G_i}, b_{G_i}) = (Z_i - Z_0, b_i - b_0). \tag{8}$$

Thus $g+1$ pairs suffice without a rank condition. Appendix B.6 derives the finite-error bound for this merge. Recovered (p_i, μ_i) also evaluate nonnegative group gates w_i: replace p_i by $w_i p_i$ in (4).

Finite-error recovery depends on the conditioning of the recorded equations. Let $B_Y x = c_Y$ collect the raw output equations in $x = (p, w_0, \dots, w_g)$, with $w_0 = \mu_0, w_i = p_i \mu_i$. Its eth residual block is $(B_Y x - c_Y)_e = (1 + \sum_{i \in S_e} p_i)\widehat{y}_e - w_0 - \sum_{i \in S_e} w_i$. For a computed $\widehat{x}$, residual $r = B_Y \widehat{x} - c_Y$ and task $f_T(x) = (a_T^\top x + a_{T,0})/(1 + c_T^\top x)$, set $\widehat{f}_T = f_T(\widehat{x})$. If $1 + c_T^\top x \geq \underline{d}_T > 0$ and $B_Y^\top z = a_T - \widehat{f}_T c_T$, then

$$|f_T(x) - \widehat{f}_T| \leq \frac{\sum_e A_e \epsilon_e \|z_e\|_2 + |z^\top r|}{\underline{d}_T}, \qquad A_e = 1 + \sum_{i \in S_e} \overline{p}_i. \tag{9}$$

Here $\overline{p}_i$ is a prior mass upper bound. The coefficients z_e express the target through the recorded equations. Their norms measure noise amplification; the residual term charges an inexact solve. Appendix B.4 derives each step and the numerical-dual correction. It also gives a full-rank family whose noise constant diverges.

5 A PHYSICAL CERTIFICATE FOR LATER ATTENTION CALLS

The query and mask may change after recording. The target is an outer interval for the physical set $\mathcal{P}_Y$ in (2). This interval directly supports the retention test (3). Both the recorded and target outputs depend on one cache perturbation. Keeping the normalizer in their identities controls the nonlinear effect of Key errors when the query changes.

Set $\delta = (\Delta K, \Delta V) \in \mathcal{D} = \mathcal{P}_X - \bar{X}$ and $h_{\mathcal{D}}(w) = \sup_{\delta \in \mathcal{D}} w^\top \delta$. For a centered box of halfwidths H, its support is $\sum_j H_j |w_j|$. A joint ball can be retained through the support bound in Appendix C.

For record e, let $T_e = S_e \cup B_e$ and let $\bar{\alpha}_{ei}$ and $\bar{y}_e$ be its attention weights and output on $\bar{X}$. Write $u_{ei} = q_e^\top \Delta k_i / \sqrt{d_k}$. Choose cell bounds $|u_{ei}| \leq \eta_{ei}$ and $|c^\top \Delta v_i| \leq \omega_i(c)$, with $\omega_i(c) = \sum_j H^V_{ij} |c_j|$ for Value halfwidths H^V. Key halfwidths H^K give $\eta_{ei} = \sum_j |q_{ej}| H^K_{ij} / \sqrt{d_k}$. For any vector v, define

$$\mathcal{L}_e(v)\delta = \sum_{i \in T_e} \bar{\alpha}_{ei} [(\bar{v}_i - v) u_{ei} + \Delta v_i]. \tag{10}$$

The exact numerator identity is

$$d_e(\delta)(y_e - v) = \bar{y}_e - v + \mathcal{L}_e(v)\delta + R_e(v, \delta), \qquad d_e(\delta) = \sum_i \bar{\alpha}_{ei} e^{u_{ei}} > 0. \tag{11}$$

Expanding $e^{u_{ei}} = 1 + u_{ei} + (e^{u_{ei}} - 1 - u_{ei})$ gives the explicit remainder

$$R_e(v, \delta) = \sum_i \bar{\alpha}_{ei} \big[(e^{u_{ei}} - 1 - u_{ei})(\bar{v}_i - v) + (e^{u_{ei}} - 1) \Delta v_i \big].$$

Using $|e^u - 1| \leq e^\eta - 1$ and $|e^u - 1 - u| \leq e^\eta - 1 - \eta$ for $|u| \leq \eta$ gives

$$|c^\top R_e(v, \delta)| \leq E_e(c, v) := \sum_i \bar{\alpha}_{ei} [|c^\top (\bar{v}_i - v)| (e^{\eta_{ei}} - 1 - \eta_{ei}) + (e^{\eta_{ei}} - 1) \omega_i(c)]. \tag{12}$$

Define $d_e^\pm = \sum_i \bar{\alpha}_{ei} e^{\pm \eta_{ei}}$, $r_e = \widehat{y}_e - \bar{y}_e$ and $\mathfrak{e}_e(c) = \epsilon_e d_e^+ \|c\|_2 + E_e(c, \widehat{y}_e)$. At the received output $v = \widehat{y}_e$,

$$\begin{aligned} |c^\top (\mathcal{L}_e(\widehat{y}_e)\delta - r_e)| &= |d_e(\delta) c^\top (y_e - \widehat{y}_e) - c^\top R_e(\widehat{y}_e, \delta)| \\ &\leq \epsilon_e d_e^+ \|c\|_2 + E_e(c, \widehat{y}_e) = \mathfrak{e}_e(c). \end{aligned}$$

Thus all records constrain the same δ.

For a target $(q_\star, T_\star, a)$, define

$$\ell_\star(u)^\top \delta = \sum_i \bar{\alpha}_{\star i}[(a^\top \bar{v}_i - u)u_{\star i} + a^\top \Delta v_i],$$
$$E_\star(u) = \sum_i \bar{\alpha}_{\star i}[|a^\top \bar{v}_i - u|(e^{\eta_{\star i}} - 1 - \eta_{\star i}) + (e^{\eta_{\star i}} - 1)\omega_i(a)].$$

Theorem 2 (Grid-free physical certificate)**.** *For coefficients $c_e \in \mathbb{R}^{d_v}$, put $g(u,c) = \ell_\star(u) - \sum_e \mathcal{L}_e(\widehat{y}_e)^\top c_e$ and $\bar{f}_\star = a^\top \bar{y}_\star$. If*

$$\mathcal{B}_+(u,c) := \bar{f}_\star - u + \sum_e c_e^\top r_e + h_{\mathcal{D}}(g(u,c)) + E_\star(u) + \sum_e \mathfrak{e}_e(c_e) \le 0, \tag{13}$$

then $f_\star(X) \le u$ for every $X \in \mathcal{P}_Y$. Applying the formula to $-a$ gives a lower bound. The endpoint optimization is convex. Its ball-and-cell support formulation is an SOCP. The inclusion holds for every finite query, nonempty mask and unit direction, including choices made after the records. Known appended rows have zero uncertainty. Fixed-coefficient evaluation costs $O(N_{\mathrm{rec}} n(d_k + d_v))$ arithmetic operations.

Proof of the enclosure. Fix $X = \bar{X} + \delta \in \mathcal{P}_Y$. The target numerator expansion has remainder $\rho_\star$ with $|\rho_\star| \le E_\star(u)$. Substituting the definition of $g = g(u,c)$,

$$\begin{aligned} d_\star(\delta)(f_\star(X) - u) &= \bar{f}_\star - u + \ell_\star(u)^\top \delta + \rho_\star \\ &= \bar{f}_\star - u + \sum_e c_e^\top r_e + g^\top \delta + \sum_e c_e^\top (\mathcal{L}_e(\widehat{y}_e)\delta - r_e) + \rho_\star \\ &\le \bar{f}_\star - u + \sum_e c_e^\top r_e + h_{\mathcal{D}}(g) + \sum_e \mathfrak{e}_e(c_e) + E_\star(u) = \mathcal{B}_+(u,c) \le 0. \end{aligned}$$

Since $d_\star(\delta) > 0$, division yields $f_\star(X) \le u$. □

The coefficients transfer observed residuals to the target; the support term bounds unexplained sensitivity. Remaining terms charge record and nonlinear errors. Zero coefficients give the codec-only bound. Key-error squares and Key–Value products dominate the remainder, so coarse Key cells may prevent acceptance. Appendix C gives the conic representation, operation count and center–radius formula.

5.1 When saved evidence changes a decision

Let W_Y be a certified posterior width. For two verified sources $X^+, X^- \in \mathcal{P}_X$,

$$W_Y < |f_\star(X^+) - f_\star(X^-)| \tag{14}$$

proves strict contraction. The pair witnesses ambiguity in the codec prior $\mathcal{P}_X$ before conditioning on the records.

An input-side condition explains the gain. On a radius-R ball with $R > 0$, let j, b be the scalar target and record gradients. Let $e_\star, e_1$ bound their finite affine remainders. For nonzero gradients set $\rho = |j^\top b|/(\|j\|_2\|b\|_2)$ and $\tau = (\epsilon + e_1)/(R\|b\|_2)$. For $0 \le \tau < \rho$, put $\varphi = \sqrt{1-\rho^2}\sqrt{1-\tau^2} + \rho\tau$. A record-conditioned interval has radius at most $R\|j\|_2\varphi + e_\star$. The physical pair $\bar{X} \pm Rj/\|j\|_2$ has half-separation at least $R\|j\|_2 - e_\star$. Hence

$$R\|j\|_2(1 - \varphi) > 2e_\star \tag{15}$$

guarantees strict benefit for every compatible received value. Appendix C.4 derives the affine estimator and minimizes its coefficient cost explicitly. With cells, the stated source pair must be feasible. Otherwise use (14).

Later exact rows change reference weights but add no error coordinates. For multihead projection blocks W_h, certify each nonzero $W_h^\top a/\|W_h^\top a\|_2$, scale by $\|W_h^\top a\|_2$ and sum. This bounds the layer output at the supplied state without head independence.

6 EXACT VALUES AND PHYSICAL OUTER BOUNDS

The main certificate is an outer interval on the original physical set. If every record uses one query, background and group partition, a summary formulation provides another interval. Group masses and means reduce the optimization dimension. A constructive lift checks whether a summary endpoint corresponds to a physical cache, linking this reduction to endpoint exactness. Normalize group masses and moments by $\pi_i = P_i / \sum_h P_h$ and $m_i = \pi_i \mu_i$. Write $\Phi_q(X) = (\pi, m)$.

Proposition 3 (Computable summary outer prior)**.** *Coordinate cells and R give positive mass bounds $\pi_i^{\pm}, p_i^{\pm}$ and finite mean bounds $\mu_{ia}^{\pm}$. The compact convex set*

$$\Theta_{\text{prior}} = \{(\pi, m) : \sum_i \pi_i = 1,\ \pi_i^- \le \pi_i \le \pi_i^+,\ p_i^- \pi_0 \le \pi_i \le p_i^+ \pi_0,\ \mu_{ia}^- \pi_i \le m_{ia} \le \mu_{ia}^+ \pi_i\} \tag{16}$$

contains $\Phi_q(\mathcal{P}_X)$. The ratio constraints use $i \ge 1$. A verified mean ball adds $\|m_i - \pi_i \mu_i^c\|_2 \le r_i \pi_i$.

For example, let $[s_j^-, s_j^+]$ bound each score $q^\top k_j / \sqrt{d_k}$ using its cell and the radius R. Let $[v_{ja}^-, v_{ja}^+]$ bound each Value coordinate. Then

$$P_i^{\pm} = \sum_{j \in G_i} e^{s_j^{\pm}}, \qquad p_i^- = P_i^- / P_0^+, \qquad p_i^+ = P_i^+ / P_0^-,$$

$$\pi_i^- = \frac{P_i^-}{P_i^- + \sum_{h \ne i} P_h^+}, \qquad \pi_i^+ = \frac{P_i^+}{P_i^+ + \sum_{h \ne i} P_h^-}, \qquad [\mu_{ia}^-, \mu_{ia}^+] = [\min_{j \in G_i} v_{ja}^-, \max_{j \in G_i} v_{ja}^+].$$

Monotonicity gives the mass bounds. Each mean is a convex combination of its group's Values. Multiplying the resulting mean bounds by $\pi_i > 0$ gives the moment constraints. Intersecting these linear inequalities proves the forward inclusion in Proposition 3. Put $d_S = \pi_0 + \sum_{i \in S} \pi_i$ and $n_S = m_0 + \sum_{i \in S} m_i$. Same-query records give

$$\Theta_Y = \{(\pi, m) \in \Theta_{\text{prior}} : \|n_{S_e} - d_{S_e} \widehat{y}_e\|_2 \le \epsilon_e d_{S_e}\ \forall e\}.$$

For a target mask T, the positive denominator permits the substitutions (Charnes & Cooper, 1962)

$$t = d_T^{-1}, \quad (\widetilde{\pi}, \widetilde{m}) = t(\pi, m), \quad d_T(\widetilde{\pi}) = 1, \quad \sum_i \widetilde{\pi}_i = t, \quad a^\top n_T / d_T = a^\top n_T(\widetilde{m}).$$

Scaling gives $\pi_i^- t \le \widetilde{\pi}_i \le \pi_i^+ t$ and preserves the record and mean-ball cones. Conversely, $\widetilde{\pi} \ge 0$ and $d_T(\widetilde{\pi}) = 1$ give $t > 0$; division by t recovers every constraint. The LP or SOCP optimum is therefore the summary endpoint. Physical exactness requires a cache realizing it. Write $\theta = (\pi, m)$, $u_X = \sup_{X \in \mathcal{P}_Y} f_{q,S,a}(X)$ and $u_\Theta = \max_{\theta \in \Theta_Y} a^\top n_S / d_S$. Let $\theta^\star$ attain the latter maximum.

For a fixed background, construct a candidate cache as follows. Take a reference $X_0 \in \mathcal{P}_X$, $q \ne 0$, and allow all coordinates in the old groups to vary within their cells. The desired summary has $p_i = \pi_i / \pi_0$, $\mu_i = m_i / \pi_i$ and the fixed background mean. Let p_i^0, μ_i^0 be its reference ratios and means, and set $\vartheta_{ij} = e^{q^\top k_j^0 / \sqrt{d_k}} / P_i^0$, $\kappa_i = \sum_{j \in G_i} \vartheta_{ij}^2$. For $j \in G_i$, $i \ge 1$, define $X_\theta = X_0 + \Delta X$ by

$$\Delta k_j = \frac{\sqrt{d_k} q}{\|q\|_2^2} \log(p_i / p_i^0), \qquad \Delta v_j = \frac{\vartheta_{ij}}{\kappa_i} (\mu_i - \mu_i^0),$$

$$P_i(X_\theta) = P_i^0 (p_i / p_i^0) = P_0^0 p_i, \qquad \mu_i(X_\theta) - \mu_i^0 = \sum_j \vartheta_{ij} \Delta v_j = \mu_i - \mu_i^0.$$

The uniform score shift preserves each group's conditional weights. The background is unchanged, so $\Phi_q(X_\theta) = \theta$. Moreover,

$$\mathcal{C}_0(\theta) := \|\Delta X\|_F^2 = \frac{d_k}{\|q\|_2^2} \sum_{i=1}^{g} |G_i| \log^2(p_i / p_i^0) + \sum_{i=1}^{g} \frac{\|\mu_i - \mu_i^0\|_2^2}{\kappa_i},$$

$$\|X_\theta - \bar{X}\|_F \le \|X_0 - \bar{X}\|_F + \sqrt{\mathcal{C}_0(\theta)} \le R.$$

The last inequality is the radius feasibility test. If it and every coordinate-cell check pass at $\theta^\star \in \Theta_Y$, then

$$X_{\theta^\star} \in \mathcal{P}_Y, \qquad u_X \le u_\Theta = f_{q,S,a}(X_{\theta^\star}) \le u_X.$$

The lower endpoint follows with $-a$. Thus a feasible lift certifies equality at the corresponding summary endpoint. Appendix D gives the score bounds, mutable-background extension and full cost derivation.

Table 1: Exact recovery, physical enclosure and endpoint exactness. The native row assumes $g \geq 3$ and unknown μ_0.

Evidence / task	Mathematical guarantee	Required check
Native outputs, fixed q	Exact $y_S(q)$ for every $S \neq \varnothing$	$N_{\mathrm{rec}} = g + 1 + \lceil g/d_v \rceil$; $\epsilon_e = 0$, generic rank (Theorem 1)
Physical residual, later (q, S, a)	$f_{q,S,a}(\mathcal{P}_Y) \subseteq [L, U]$	(13) for a and $-a$
Summary cone, fixed q	$u_X \leq u_\Theta$; a lift gives $u_X = u_\Theta$	$\exists X^\star \in \mathcal{P}_Y : \Phi_q(X^\star) = \theta^\star$

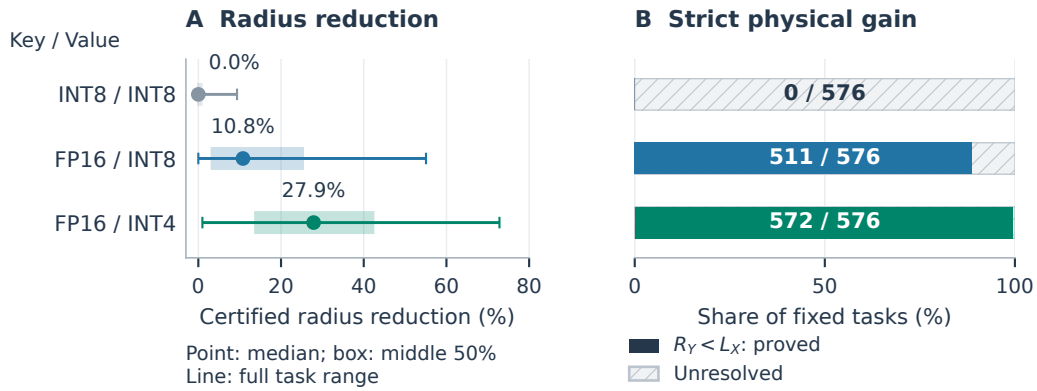

Figure 1: Smaller bounds versus proved information gain. R_X and R_Y are codec-only and record-conditioned radius upper bounds. L_X is a radius lower bound from a same-code physical pair. (A) Reduction $100(1 - R_Y/R_X)$: median, middle 50% and full range. (B) Strict gain requires $R_Y < L_X$; hatched tasks remain unresolved. Each precision uses all 576 fixed tasks. Record error is 10^{-3}; coefficients use the multi-candidate mode.

7 Evidence gain, storage and checking cost

The experiments follow three questions: when records tighten a fixed physical prior, which retention decisions this certifies, and how to allocate bytes between records and codec precision. Radius contraction, complete-head acceptance and matched-byte controls measure these effects. We use frozen SmolLM2-1.7B-Instruct caches at layer 23, head 0, with $d_k = d_v = 64$. Four controlled contexts cover prose, repeated text, code and numbers. Positions 0–31 form the uncertain prefix. Positions 32–47 supply actual queries and an exact live suffix. We save causal FP32 outputs at positions 32–35, with independently verified vector errors. The grid uses four output coordinates, twelve later queries and three masks: intact, recent 16 old positions, and top 16 reconstructed scores. This gives 144 tasks per context and precision. The certificate uses the reconstruction, cells and records.

Fixed coefficient tuples undergo independent outward checking. Figure 1 separates contraction from strict physical gain, witnessed by sources with identical codes and row scales. FP16 rounding cells bound the asymmetric Key uncertainty. Values use signed, power-of-two row quantizers. The appendix also reports an equal-precision min–max control at 8, 10 and 12 bits. The Key-precision comparison gives 0/576 strict-gain certificates for INT8/INT8 and 511/576 for FP16/INT8, consistent with the Key-dependent remainder. FP16/INT4 gives 572/576 in this multi-candidate mode.

For complete-head decisions, apply (3) to all 64 coordinates. Candidates keep 0, 8, 16, 24 or 32 old rows in recent or score order. Choose the fewest rows that pass. At $\tau = 0.5$, four records certify 23/48 queries versus 15/48 for the same codec without records. Of 17 improved decisions, fifteen gain certified evictions and two newly certify full retention, with zero eviction. Each has a same-code source violating its budget before the records are imposed, proving information necessity for that

guarantee. Figure 2 shows the full error-budget curve. Across these fixed queries, the attention-output error budget $\tau = 0.5$ is about 4.3–10.0% of the original head-output norm.

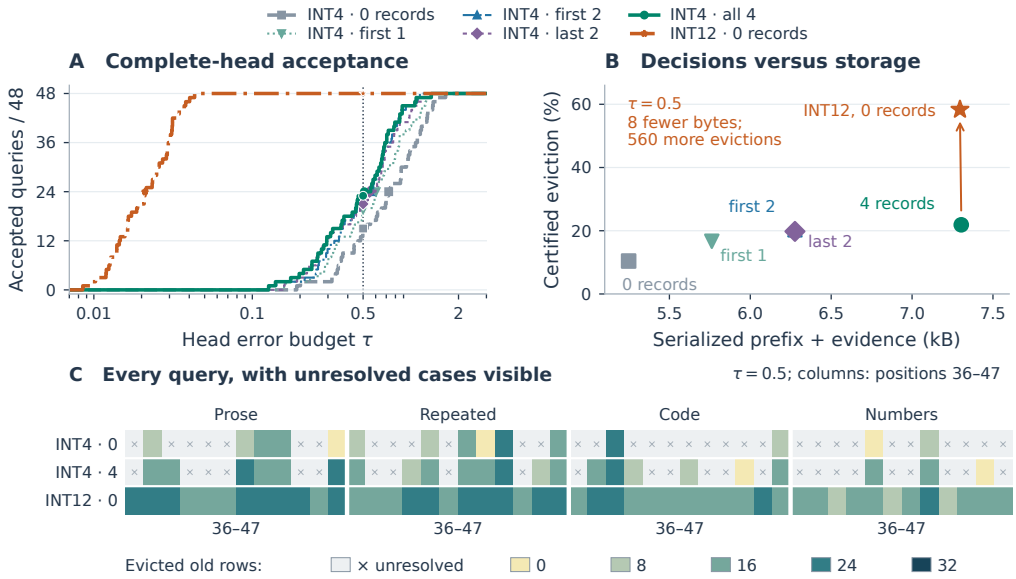

Figure 2: Evidence value and byte cost on all 48 queries. (A) Acceptance uses each query's smallest candidate error bound. (B) At $\tau = 0.5$, certified evictions are divided by $48 \cdot 32$; payloads cost 7304 bytes (INT4 plus four records) or 7296 (INT12). (C) Eviction counts at this budget, in chronological order within each context. A cross means unresolved; zero means accepted full retention. All panels use the same one-ray complete-head run.

Under a 7304-byte cap, INT12 without records certifies 48/48 queries at $\tau = 0.5$. INT4 with four records certifies 23/48 (Figure 2). Finer Values are the better tested byte allocation on these short prefixes. Each allocation defines its own codec-compatible source set.

With the INT4 codec fixed, saving the first one, two or four records gives 19, 22 or 23 accepts, versus 15 without records at $\tau = 0.5$. Saving only positions 34–35 gives 21 accepts. Appendix E.5 gives the byte accounting and protocol.

For 2048-row prefixes from the same four context families, all 128 scalar intervals pass outward checks. FP16/INT8 and FP16/INT4 give strict gains in 55/64 and 64/64 tasks. The offline checker takes about 0.4 seconds for a short-prefix complete head. Appendix E gives acquisition errors, task grids, CPU scaling timings, artifact sizes and threshold sweeps.

8 Conclusion and Future Work

Saved attention outputs can support retention decisions after the source prefix is released. Native mask counts give the exact-recovery cost for fixed-query group workloads. The physical residual certificate extends the same evidence to later queries and masks through a retained reconstruction. Its value depends on the workload and codec geometry. The frozen checks connect this information gain to certified retention and quantify the allocation of storage between output records and codec precision. A feasible cache lift further distinguishes an attainable physical endpoint from a summary outer bound.

Future work should divide a fixed byte budget between codec precision and output records for a declared workload. Record selection should reflect sensitivity, noise amplification and acquisition cost. Tighter Key remainders and cell-preserving lifts could resolve more decisions. Matched-byte comparisons and physical witnesses can measure how these choices change the set of certified decisions.

AI USE STATEMENT

Generative AI tools assisted manuscript drafting and revision, literature search and reference checking, mathematical formulation and proof writing, experimental design, and interpretation of results. They also assisted the implementation and debugging of experimental code and the preparation of figures. All AI-assisted content has been reviewed and verified by the authors. References were checked against primary sources. Numerical certificates were validated by separate checking routines, including directed interval arithmetic and serialized-data replay. The authors are responsible for the final manuscript, mathematical claims, code and results.

REPRODUCIBILITY STATEMENT

Section 3 specifies the recorded inputs, physical uncertainty set and retention criterion. Appendix A details the acquisition contract. Appendix B gives the native observation construction and lower bound, including finite-error recovery. Appendices C and D derive the physical certificate and the summary endpoint tests. Appendix E specifies the frozen model, contexts, queries, masks, quantizers, coefficient proposals and numerical checker. Appendix E.5 gives the serialized byte accounting, record ablations and complete-head decision protocol.

ETHICS STATEMENT

This work studies mathematical guarantees for attention-cache retention. The experiments use a frozen public language model and controlled text contexts. The evaluated guarantee concerns attention-output error at a specified query and state. Applications that store output records should treat them as derived model data and apply the access and deletion rules appropriate to their source inputs. The experiment design and interpretation retain all tested error budgets and both accepted and unresolved decisions. The storage comparison includes finer quantization within the same byte cap.

REFERENCES

Yongqi An, Chang Lu, Kuan Zhu, Tao Yu, Chaoyang Zhao, Hong Wu, Ming Tang, and Jinqiao Wang. ReST-KV: Robust KV cache eviction with layer-wise output reconstruction and spatial-temporal smoothing. In *International Conference on Learning Representations*, pp. 86943–86967, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8be9c134bb193d8bd3827d4df8488228-Abstract-Conference.html.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441.

Dean Calver. Runtime-certified bounded-error quantized attention, 2026. URL https://arxiv.org/abs/2605.20868v1. Version 1, May 20, 2026.

Abraham Charnes and William W. Cooper. Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 9(3–4):181–186, 1962. doi: 10.1002/nav.3800090303.

Justin Y. Chen, Ying Feng, Piotr Indyk, Michael Kapralov, Ekaterina Kochetkova, and Boris Prokhorov. Towards tight bounds for streaming attention, 2026. URL https://arxiv.org/abs/2606.07205v1.

Ofek I. Cohen. Query-oblivious coresets for softmax attention: Improved bounds and efficient constructions, 2026. URL https://arxiv.org/abs/2609.06327v2. Version 2, September 13, 2026.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359, 2022. doi: 10.52202/068431-1189. URL https://papers.nips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract-Conference.html.

Mahmood Ettehad and Simon Foucart. Instances of computational optimal recovery: Dealing with observation errors. *SIAM/ASA Journal on Uncertainty Quantification*, 9(4):1438–1456, 2021. doi: 10.1137/20M1328476.

FlashAttention Contributors. FlashAttention python interface, 2026. URL https://github.com/Dao-AILab/flash-attention/blob/00756db9d921da0846453283ddfbeb7457abd09b/flash_attn/flash_attn_interface.py. Source snapshot 00756db9d921da0846453283ddfbeb7457abd09b; accessed September 26, 2026.

FlashInfer Contributors. Attention states and recursive attention, 2025. URL https://github.com/flashinfer-ai/flashinfer/blob/8ddfd285e532e79a9691abda90565b8c31217a08/docs/tutorials/recursive_attention.rst. Source snapshot 8ddfd285e532e79a9691abda90565b8c31217a08; accessed September 26, 2026.

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. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. KVQuant: Towards 10 million context length LLM inference with KV cache quantization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 1270–1303, 2024. doi: 10.52202/079017-0040. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/028fcbcf85435d39a40c4d61b42c99a4-Abstract-Conference.html.

Hao Kang, Qingru Zhang, Souvik Kundu, Geonhwa Jeong, Zaoxing Liu, Tushar Krishna, and Tuo Zhao. GEAR: An efficient error reduction framework for KV cache compression in LLM inference. In *Proceedings of the 4th NeurIPS Efficient Natural Language and Speech Processing*

Workshop, volume 262 of *Proceedings of Machine Learning Research*, pp. 305–321, 2024. URL https://proceedings.mlr.press/v262/kang24a.html.

Jang-Hyun Kim, Jinuk Kim, Sangwoo Kwon, Jae W. Lee, Sangdoo Yun, and Hyun Oh Song. KVzip: Query-agnostic KV cache compression with context reconstruction. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167563–167591, 2025. doi: 10.52202/085713-5585. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4eaa4b8f2d08edb3f0af990d56134ea-Abstract-Conference.html.

Ekaterina Kochetkova, Kshiteej Jitesh Sheth, Insu Han, Amir Zandieh, and Michael Kapralov. Streaming attention approximation via discrepancy theory. In *Advances in Neural Information Processing Systems*, volume 38, pp. 20935–20972, 2025. doi: 10.52202/085713-0705. URL https://papers.neurips.cc/paper_files/paper/2025/hash/1de510838faddd72f79a7a0e9c564f0c-Abstract-Conference.html.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 22947–22970, 2024. doi: 10.52202/079017-0722. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/28ab418242603e0f7323e54185d19bde-Abstract-Conference.html.

Hao Liu, Matei Zaharia, and Pieter Abbeel. RingAttention with blockwise transformers for near-infinite context. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=WsRHpHH4s0.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32332–32344, 2024b. URL https://proceedings.mlr.press/v235/liu24bz.html.

Charles A. Micchelli and Theodore J. Rivlin. A survey of optimal recovery. In *Optimal Estimation in Approximation Theory*, pp. 1–54. Springer, 1977. doi: 10.1007/978-1-4684-2388-4_1.

Maxim Milakov and Natalia Gimelshein. Online normalizer calculation for softmax. *arXiv preprint arXiv:1805.02867*, 2018. URL https://arxiv.org/abs/1805.02867.

Leszek Plaskota. *Noisy Information and Computational Complexity*. Cambridge University Press, 1996. doi: 10.1017/CBO9780511600814.

PyTorch Contributors. torch.nn.functional.scaled_dot_product_attention. PyTorch 2.9 documentation, 2025. URL https://docs.pytorch.org/docs/2.9/generated/torch.nn.functional.scaled_dot_product_attention.html. Accessed September 26, 2026.

Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. Robustness verification for transformers. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJxwPJHFwS.

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, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5e5fd18f863cbe6d8ae392a93fd271c9-Abstract-Conference.html.

Amir Zandieh, Insu Han, Vahab Mirrokni, and Amin Karbasi. SubGen: Token generation in sublinear time and memory, 2024. URL https://arxiv.org/abs/2402.06082v1.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34661–34710, 2023. doi: 10.52202/075280-1506. URL https://papers.neurips.cc/paper_files/paper/2023/hash/6ceefa7b15572587b78ecfcebb2827f8-Abstract-Conference.html.

APPENDIX CONTENTS

The appendices follow the main argument: retained inputs, exact recovery, later-query certificates, physical endpoints, and experiments.

A Retained inputs and audit decisions . . . 14
A.1 Recorded inputs . . . 14
A.2 Quantization cells and record errors . . . 14
A.3 Decisions and storage accounting . . . 14

B Proof of the minimum native observation count . . . 14
B.1 Elimination of the hidden normalization . . . 15
B.2 A fixed construction at every Value width . . . 16
B.3 Generic rank and the adaptive lower bound . . . 17
B.4 Finite-error stability . . . 18
B.5 An ill-conditioned family and an alternative mask . . . 20
B.6 Normalization-aware records . . . 21

C Proof of the record-conditioned certificate . . . 22
C.1 Exact residual and received-record inequalities . . . 22
C.2 Target elimination, convexity and finite checking . . . 23
C.3 Post-selected queries, known rows and additional records . . . 24
C.4 Strict gain from gradient alignment . . . 24

D Physical outer bounds and when they are exact . . . 25
D.1 Constructing the summary prior . . . 26
D.2 Record constraints and fractional scaling . . . 26
D.3 An explicit lift and endpoint equality . . . 27

E Experimental protocols and checking costs . . . 28
E.1 Frozen caches, record acquisition and codecs . . . 28
E.2 Coefficient proposals and numerical checking . . . 28
E.3 Four-coordinate and complete-head retention decisions . . . 30
E.4 Long-prefix acquisition and serialized storage . . . 30
E.5 Matched-byte comparison and record ablations . . . 31

A RETAINED INPUTS AND AUDIT DECISIONS

A.1 RECORDED INPUTS

The cache owner has the original $X = (K, V)$ before compression. It saves a reconstruction $\bar{X}$, coordinate cells, an error radius and selected attention records. It then releases the original prefix. The evaluator receives

$$\bar{X},\ \mathcal{C},\ R, \quad (q_e, T_e, \widehat{y}_e, \epsilon_e)_{e=1}^{N_{\mathrm{rec}}}.$$

The fixed model and exact live suffix are shared inputs. A record query and its causal visible set refer to the computation made while the source was available. Ordinary causal recording needs no replay of a future target mask.

The output-only interface exports normalized vectors. The normalization-aware interface also exports LSE. Internal access may instead export masses and weighted means. The count theorem concerns the first interface. The physical residual certificate may use any of these records. The query metadata, visible sets and acquisition bounds are assumed correct.

A.2 QUANTIZATION CELLS AND RECORD ERRORS

For an unclipped nearest-grid codec with step Δ_j, and a reconstruction-rounding allowance e_j^{codec},

$$\begin{aligned} |x_j - \bar{x}_j| &\le |x_j - x_j^{\mathrm{grid}}| + |x_j^{\mathrm{grid}} - \bar{x}_j| \le \Delta_j/2 + e_j^{\mathrm{codec}} =: H_j, \\ \|X - \bar{X}\|_F^2 &= \sum_j |x_j - \bar{x}_j|^2 \le \sum_j H_j^2. \end{aligned}$$

A separate verified joint radius can be tighter. Clipping errors must be included if the source falls outside the grid. The bounds depend on the quantization rule and decoding error. For example, our row codec is distinct from the asymmetric axes in KIVI (Liu et al., 2024b). Likewise, $\epsilon_e = e_e^{\mathrm{acquisition}} + e_e^{\mathrm{storage}} + e_e^{\mathrm{communication}}$ is a deterministic vector-error allowance, not a standard deviation.

A.3 DECISIONS AND STORAGE ACCOUNTING

For a nonempty compatible set, let $f_- = \inf_{\mathcal{P}_Y} f$ and $f_+ = \sup_{\mathcal{P}_Y} f$. A computed $[L, U] \supseteq [f_-, f_+]$ and admissible interval I imply

$$[L,U] \subseteq I \Longrightarrow f(X) \in I \quad \forall X \in \mathcal{P}_Y, \qquad [L,U] \cap I = \varnothing \Longrightarrow f(X) \notin I \quad \forall X \in \mathcal{P}_Y.$$

Other cases remain unresolved. An empty $\mathcal{P}_Y$ signals inconsistent inputs. A feasible source outside I disproves universal acceptance.

The storage cost includes codec data, scales, queries, output vectors, mask identities and error metadata. The live suffix and model weights are charged equally on both sides. An exact-count optimum is a statement about native vector calls. It does not minimize bytes over arbitrary encodings. The matched-byte comparison in Appendix E.5 tests a separate storage-allocation question.

B PROOF OF THE MINIMUM NATIVE OBSERVATION COUNT

We prove Theorem 1. All masses are positive. Set $P = \sum_{i=1}^{g} p_i$, $\beta_0 = (1 + P)^{-1}$ and $\omega_i = (1 + p_i)\beta_0$.

B.1 Elimination of the hidden normalization

The singleton and intact equations give

$$
\begin{aligned}
(1+p_i)y_i &= \mu_0 + p_i\mu_i, \qquad (1+P)y_{\text{all}} = \mu_0 + \sum_i p_i\mu_i, \\
\sum_i (1+p_i)y_i - (1+P)y_{\text{all}} &= (g\mu_0 + \sum_i p_i\mu_i) - (\mu_0 + \sum_i p_i\mu_i) = (g-1)\mu_0, \\
\sum_i \omega_i - 1 &= \beta_0(g+P) - \beta_0(1+P) = (g-1)\beta_0, \\
(g-1)\beta_0\mu_0 &= \sum_i \omega_i y_i - y_{\text{all}}.
\end{aligned} \tag{B.1}
$$

For a nonempty group mask S,

$$
\begin{aligned}
\beta_0\left(\mu_0 + \sum_{i\in S} p_i\mu_i\right) &= \beta_0\left(\sum_{i\in S}(1+p_i)y_i - (|S|-1)\mu_0\right) \\
&= \sum_{i\in S}\omega_i y_i - \frac{|S|-1}{g-1}\left(\sum_i \omega_i y_i - y_{\text{all}}\right) \\
&= \sum_i \xi_{S,i}\omega_i y_i + \gamma_S y_{\text{all}}, \\
\beta_0\left(1 + \sum_{i\in S} p_i\right) &= \sum_{i\in S}\omega_i - (|S|-1)\beta_0 = \xi_S^\top\omega + \gamma_S > 0.
\end{aligned} \tag{B.2}
$$

Dividing these two identities proves (5). At an extra proper union J,

$$
\begin{aligned}
(\xi_J^\top\omega + \gamma_J)y_J &= \sum_i \xi_{J,i}\omega_i y_i + \gamma_J y_{\text{all}}, \\
\sum_i \xi_{J,i}(y_i - y_J)\omega_i &= \gamma_J(y_J - y_{\text{all}}) = r_J.
\end{aligned}
$$

This is the observed system $M_J\omega = r_J$.

Let $C = [y_i - \mu_0]_i$, $d_J = \xi_J^\top\omega + \gamma_J$ and $v_i = \omega_i\mathbf{1}\{i \in J\}/d_J$. Then

$$
\begin{aligned}
y_J - \mu_0 &= \frac{\sum_{i\in J}(1+p_i)(y_i-\mu_0)}{1+\sum_{i\in J}p_i} = Cv, \\
M_J &= [\xi_{J,i}\{(y_i-\mu_0) - (y_J-\mu_0)\}]_i = C(I - v\mathbf{1}^\top)\,\mathrm{diag}(\xi_J), \\
1 - \mathbf{1}^\top v &= \frac{d_J - \sum_{i\in J}\omega_i}{d_J} = -\frac{(|J|-1)\beta_0}{d_J} \neq 0.
\end{aligned}
$$

The matrix determinant identity used here is $\det(I - ab^\top) = 1 - b^\top a$. Since $1 < |J| < g$,

$$
\xi_{J,i} \neq 0 \quad \forall i, \qquad \det[(I - v\mathbf{1}^\top)\,\mathrm{diag}(\xi_J)] = -\frac{(|J|-1)\beta_0}{d_J}\prod_i \xi_{J,i} \neq 0.
$$

Thus $\mathrm{rank}\, C = g$ implies $\mathrm{rank}\, M_J = g$. The recovered quantities are

$$
\beta_0 = \frac{\sum_i \omega_i - 1}{g-1}, \quad p_i = \omega_i/\beta_0 - 1, \quad \mu_0 = \frac{\sum_i \omega_i y_i - y_{\text{all}}}{(g-1)\beta_0}, \quad \mu_i = \frac{(1+p_i)y_i - \mu_0}{p_i}.
$$

B.2 A FIXED CONSTRUCTION AT EVERY VALUE WIDTH

Suppose groups H have been recovered. Write $p_H = \sum_{i\in H} p_i$ and $y_H = (\mu_0 + \sum_{i\in H} p_i\mu_i)/(1 + p_H)$. For a new block J with $r = |J| \le d_v$, singleton equations yield

$$\begin{aligned}(1 + p_H + \sum_{i\in J} p_i)y_{H\cup J} &= (1 + p_H)y_H + \sum_{i\in J} p_i\mu_i \\ &= (1 + p_H)y_H + \sum_{i\in J}(1 + p_i)y_i - r\mu_0, \\ \sum_{i\in J}(y_i - y_{H\cup J})p_i &= (1 + p_H)(y_{H\cup J} - y_H) - \sum_{i\in J} y_i + r\mu_0. \end{aligned} \tag{B.3}$$

This needs r singleton calls and one union call.

For $d_v \ge 3$, take $s = \min(g, d_v)$ seed groups. Their s singletons, joint union and one proper union recover the seed. A simultaneous full-rank example sets all $p_i = 1$, $\mu_0 = 0$ and assigns singleton outputs $e_1, \dots, e_r$ to each new block. Each mean is twice its assigned basis vector. For $v = y_{H\cup J}$, a column dependence satisfies

$$\sum_{i=1}^{r} a_i(e_i - v) = 0 \implies a_j = v_j \sum_i a_i \ (j \le r), \qquad 0 = v_j \sum_i a_i \ (j > r).$$

If some $v_j > 0$ for $j > r$, these equations imply $\sum_i a_i = 0$ and then $a = 0$. Otherwise all means in the mixture have support in the first r coordinates. In that case

$$\sum_{j=1}^{r} v_j = \frac{\sum_{i\in H\cup J} p_i \sum_{j=1}^{r} \mu_{ij}}{1 + p_H + r} = \frac{2(p_H + r)}{1 + p_H + r},$$

$$\det(I - v_{1:r}\mathbf{1}^\top) = 1 - \sum_{j=1}^{r} v_j = \frac{1 - p_H - r}{1 + p_H + r} \ne 0,$$

because $p_H + r \ge s + 1 \ge 4$. The seed uses independent contrasts. Each appended block therefore has full column rank in the same $\mathbb{R}^{d_v}$. The number of calls is

$$s + 2 + (g - s) + \left\lceil \frac{g - s}{d_v} \right\rceil = g + 1 + \left\lceil \frac{g}{d_v} \right\rceil.$$

The final union is intact. Its output may be recorded first and used last.

For widths one and two, eliminate the Value means using singleton outputs. Every extra union gives

$$\sum_{i\in S}(y_S - y_i)p_i + (|S| - 1)\mu_0 = \sum_{i\in S} y_i - y_S. \tag{B.4}$$

At width one, use $p = (1, 2, 3)$ and $(\mu_0, \mu_1, \mu_2, \mu_3) = (0, 1, 2, 4)$. Rows for $12, 13, 23, 123$, multiplied by $12, 10, 3, 42$, give

$$C_1 = \begin{pmatrix} 9 & -1 & 0 & 12 \\ 21 & 0 & -4 & 10 \\ 0 & 4 & -1 & 3 \\ 81 & 46 & -24 & 84 \end{pmatrix}.$$

At width two and odd g, use $p = (1, 2, 3)$ and means $((0, 0), (1, 0), (0, 1), (1, 2))$. Take the two coordinates of rows $12, 13$ and the first coordinate of row 123. Multiplying them by $4, 6, 20, 10, 28$ gives

$$C_{2,o} = \begin{pmatrix} -1 & 1 & 0 & 4 & 0 \\ 3 & -1 & 0 & 0 & 6 \\ 6 & 0 & 1 & 20 & 0 \\ 12 & 0 & -3 & 0 & 10 \\ 2 & 16 & -5 & 56 & 0 \end{pmatrix}.$$

At width two and even g, take $p = (1, 2, 3, 4)$ and means $((0, 0), (1, 0), (0, 1), (1, 2), (2, -1))$. Rows $12, 34, 1234$, with both coordinates, scaled by $4, 6, 40, 20, 220, 330$, give

$$C_{2,e} = \begin{pmatrix} -1 & 1 & 0 & 0 & 4 & 0 \\ 3 & -1 & 0 & 0 & 0 & 6 \\ 0 & 0 & 25 & -9 & 40 & 0 \\ 0 & 0 & -25 & 21 & 0 & 20 \\ 130 & 240 & 75 & -112 & 660 & 0 \\ 120 & -100 & -375 & 384 & 0 & 990 \end{pmatrix}.$$

Gaussian elimination uses the update

$$C_{ij}^{(k+1)} = C_{ij}^{(k)} - \frac{C_{ik}^{(k)} C_{kj}^{(k)}}{C_{kk}^{(k)}} \quad (i, j > k).$$

For $C_{2,o}$, interchange rows four and five before the fourth pivot. The other two matrices need no row exchange. Thus $\det C = (-1)^s \prod_k C_{kk}^{(k)}$, where s counts row exchanges. The successive pivots are

	$C_{kk}^{(k)}$	$\det C$
C_1	$9,\ 7/3,\ 41/7,\ -6$	-738
$C_{2,o}$	$-1,\ 2,\ 1,\ -4,\ -80$	640
$C_{2,e}$	$-1,\ 2,\ 25,\ 12,\ -2630/3,\ 97680/263$	195360000

so each displayed reduced system has full column rank. Singleton equations then recover every mean.

At width one, append one group with ratio one and singleton t. For $\mu_H = p_H^{-1} \sum_{i \in H} p_i \mu_i$,

$$t - y_{H \cup \{i\}} = t - \frac{\mu_0 + p_H \mu_H + 2t - \mu_0}{p_H + 2} = \frac{p_H (t - \mu_H)}{p_H + 2} \neq 0$$

whenever $t \neq \mu_H$. The count is $7 + 2(g - 3) = 2g + 1$. At width two, append pairs with singleton outputs te_1, te_2 and unit ratios. The seed has $\mu_0 = 0$ and $p_H \geq 6$. Write $\mu_H = p_H^{-1} \sum_{i \in H} p_i \mu_i$. The coefficient matrix satisfies

$$v(t) = y_{H \cup J} = \frac{p_H \mu_H + 2t\mathbf{1}}{p_H + 3}, \qquad A(t) = tI - v(t)\mathbf{1}^\top,$$

$$\det(t^{-1} A(t)) = 1 - \mathbf{1}^\top v(t)/t = \frac{p_H - 1 - p_H \mathbf{1}^\top \mu_H / t}{p_H + 3} > 0 \quad \text{if } t > \frac{p_H |\mathbf{1}^\top \mu_H|}{p_H - 1}.$$

The odd and even counts are $6 + 3(g - 3)/2$ and $7 + 3(g - 4)/2$. Both equal $g + 1 + \lceil g/2 \rceil$.

If μ_0 is supplied and $d_v \geq 2$, a seed of $s = \min(g, d_v) \geq 2$ needs only singletons and their union:

$$y_H - \mu_0 = \sum_{i \in H} \omega_i (y_i - \mu_0),$$

$$\beta_0 = (\sum_{i \in H} \omega_i - 1)/(s - 1), \qquad p_i = \omega_i / \beta_0 - 1.$$

Independent columns $y_i - \mu_0$ recover ω. The block count is $s + 1 + (g - s) + \lceil (g - s)/d_v \rceil = g + \lceil g/d_v \rceil$. At width one, omit row 23 from the scalar seed and remove the known-μ_0 column in (B.4). The scaled matrix is

$$C_{1,k} = \begin{pmatrix} 9 & -1 & 0 \\ 21 & 0 & -4 \\ 81 & 46 & -24 \end{pmatrix}, \qquad \text{pivots } 9,\ 7/3,\ 492/7, \qquad \det C_{1,k} = 1476.$$

Six seed calls and two per appended group give $2g$ calls.

B.3 Generic rank and the adaptive lower bound

Write $w_i = p_i \mu_i$ and $w_0 = \mu_0$. Each vector record gives

$$\sum_{i \in S} y_S p_i - w_0 - \sum_{i \in S} w_i = -y_S. \tag{B.5}$$

The parameter dimensions are $D_{\text{hidden}} = g + (g+1)d_v$ and $D_{\text{known}} = g + gd_v$. Every entry of a fixed protocol is rational in these parameters. Its denominators are positive. Multiplying a full-rank minor by their product gives a polynomial P with $P(x_0) \neq 0$ at a constructed point. If P vanished on a neighborhood of x_*, every derivative there would vanish. Its finite Taylor formula would then give

$$P(x) = \sum_{|\alpha| \leq \deg P} \frac{\partial^\alpha P(x_*)}{\alpha!}(x - x_*)^\alpha = 0 \quad \forall x,$$

contradicting $P(x_0) \neq 0$. Hence its zero set has empty interior. Off this exceptional set the displayed inverse recovers all parameters. The full family of task outputs is therefore locally injective. A physical realization chart Ξ transfers this statement to the source class in the theorem. An explicit sufficient realization is given in Appendix D.

The constant-rank theorem applies at each adaptive step. If a C^1 map $F : M^d \to \mathbb{R}^r$ has constant derivative rank s near x, there are local coordinates in which

$$F(u_1, \ldots, u_d) = (u_1, \ldots, u_s, 0, \ldots, 0).$$

Consequently its local level set has dimension $d - s$. Start on a task-injective open chart M_0 of dimension D. Choose a point of maximal derivative rank s. A nonzero s-minor remains nonzero on a neighborhood U_0, so $s \leq \operatorname{rank} DF_1(x) \leq s$ for $x \in U_0$. The constant-rank coordinates therefore give a level set M_1. All sources on M_1 have the same first response. The next action is therefore fixed on M_1, even if the action rule is discontinuous. Repeating the coordinate statement gives

$$\begin{aligned} M_e &= \{x \in U_{e-1} \subset M_{e-1} : F_e(x) = y_e\}, \\ \dim M_e &= \dim M_{e-1} - s_e, \qquad 0 \leq s_e \leq d_v, \\ \dim M_N &= D - \sum_{e=1}^{N} s_e \geq D - Nd_v, \\ x, x' \in M_N &\Longrightarrow (y_1(x), \ldots, y_N(x)) = (y_1(x'), \ldots, y_N(x')). \end{aligned}$$

If $Nd_v < D$, choose distinct nearby $x, x' \in M_N$. Local task injectivity makes some task outputs different. The common transcript cannot recover both. Therefore

$$N \geq \lceil D/d_v \rceil = \begin{cases} g + 1 + \lceil g/d_v \rceil, & \mu_0 \text{ unknown}, \\ g + \lceil g/d_v \rceil, & \mu_0 \text{ supplied}. \end{cases}$$

The count is over exact response coordinates, not arbitrary stored bits.

For $g = 1$ there is one task and one call suffices. For $g = 2$ there are three tasks. With unknown μ_0, their map from the three means has, before positive row scaling, coefficient matrix

$$\begin{pmatrix} 1 & p_1 & 0 \\ 1 & 0 & p_2 \\ 1 & p_1 & p_2 \end{pmatrix}, \qquad \det = -p_1 p_2 \neq 0.$$

Its image contains an open subset of $\mathbb{R}^{3d_v}$, requiring three calls. With known μ_0 and $d_v \geq 2$, the singleton–union inverse has dimension $2d_v + 2 > 2d_v$. At $d_v = 1$, fix $\mu_0 = 0, p_2 = 2$ and evaluate the Jacobian with respect to (p_1, μ_1, μ_2) at $(1, 1, 3)$:

$$D(y_1, y_2, y_{12}) = \begin{pmatrix} 1/4 & 1/2 & 0 \\ 0 & 0 & 2/3 \\ -3/16 & 1/4 & 1/2 \end{pmatrix}, \qquad \det = -5/48 \neq 0.$$

This requires three scalar calls and completes the count proof.

B.4 Finite-error stability

Use $x = (p_1, \ldots, p_g, w_0, w_1, \ldots, w_g)$ in (B.5). Let B_Y and c_Y use the received outputs. A known w_0 is moved to the right side. Assume $0 < \underline{p}_i \leq p_i \leq \overline{p}_i$ on the declared prior. Set $A_e = 1 + \sum_{i \in S_e} \overline{p}_i$.

Proposition B.1 (Received-record stability). *Let $\widehat{x}$ have residual $r = B_Y\widehat{x} - c_Y$. Write $f_T(x) = (a_T^\top x + a_{T,0})/d_T(x)$, $d_T(x) = 1 + c_T^\top x \geq \underline{d}_T > 0$. Suppose $d_T(\widehat{x}) > 0$ and set $\widehat{f}_T = f_T(\widehat{x})$ and $v_T = a_T - \widehat{f}_T c_T$. If $B_Y^\top z = v_T$, then*

$$|f_T(x) - \widehat{f}_T| \leq \frac{\sum_e A_e \epsilon_e \|z_e\|_2 + |z^\top r|}{\underline{d}_T} \quad \textit{for every compatible } x. \tag{B.6}$$

For defect $h = v_T - B_Y^\top z$, add $\|h\|_2 \zeta_x$ to the numerator, where $\|x - \widehat{x}\|_2 \leq \zeta_x$ is verified. For the exact-output matrix B_0,

$$\|B_Y - B_0\|_2 \leq \beta_Y := \left(\sum_e |S_e| \epsilon_e^2\right)^{1/2}.$$

In particular, $\sigma_{\min}(B_0) > \beta_Y$ preserves full column rank.

Proof. For $\delta_e = \widehat{y}_e - y_{S_e}(x)$, the block residual is

$$\begin{aligned}
(B_Y x - c_Y)_e &= \widehat{y}_e(1 + \sum_{i \in S_e} p_i) - w_0 - \sum_{i \in S_e} w_i \\
&= (1 + \sum_{i \in S_e} p_i)(\widehat{y}_e - y_{S_e}(x)) = d_{S_e}(x)\delta_e, \\
\|(B_Y x - c_Y)_e\|_2 &\leq (1 + \sum_{i \in S_e} \overline{p}_i)\epsilon_e = A_e \epsilon_e.
\end{aligned}$$

Now expand the two target fractions:

$$\begin{aligned}
d_T(x)(f_T(x) - \widehat{f}_T) &= a_T^\top x + a_{T,0} - \widehat{f}_T(1 + c_T^\top x) \\
&= a_T^\top(x - \widehat{x}) - \widehat{f}_T c_T^\top(x - \widehat{x}) + \underbrace{a_T^\top \widehat{x} + a_{T,0} - \widehat{f}_T d_T(\widehat{x})}_{0} \\
&= (B_Y^\top z + h)^\top (x - \widehat{x}) \\
&= z^\top[(B_Y x - c_Y) - (B_Y \widehat{x} - c_Y)] + h^\top(x - \widehat{x}) \\
&= \sum_e z_e^\top d_{S_e}(x)\delta_e - z^\top r + h^\top(x - \widehat{x}), \\
|f_T(x) - \widehat{f}_T| &\leq \frac{\sum_e d_{S_e}(x)|z_e^\top \delta_e| + |z^\top r| + |h^\top(x - \widehat{x})|}{d_T(x)} \\
&\overset{\text{Cauchy–Schwarz}}{\leq} \frac{\sum_e A_e \epsilon_e \|z_e\|_2 + |z^\top r| + \|h\|_2 \zeta_x}{\underline{d}_T}.
\end{aligned}$$

Only the p_i columns change with the received outputs. Therefore

$$\begin{aligned}
\|B_Y - B_0\|_2^2 \leq \|B_Y - B_0\|_F^2 &= \sum_e \sum_{i \in S_e} \|\widehat{y}_e - y_e\|_2^2 \leq \sum_e |S_e| \epsilon_e^2 = \beta_Y^2, \\
\sigma_{\min}(B_Y) &= \min_{\|v\|_2 = 1} \|B_Y v\|_2 \\
&\geq \min_{\|v\|_2 = 1} [\|B_0 v\|_2 - \|(B_Y - B_0)v\|_2] \\
&\geq \sigma_{\min}(B_0) - \beta_Y > 0.
\end{aligned}$$

□

A received-data left inverse can bound ζ_x. Let $\|I - CB_Y\|_2 \le \eta_C < 1$ and $u = x - \widehat{x}$. Then

$$\begin{aligned} u &= C[(B_Y x - c_Y) - r] + (I - CB_Y)u, \\ (1-\eta_C)\|u\|_2 &\le \|C\|_2 \left(\|B_Y x - c_Y\|_2 + \|r\|_2\right) \\ &\le \|C\|_2 \left[\left(\sum_e A_e^2 \epsilon_e^2\right)^{1/2} + \|r\|_2\right], \\ \|u\|_2 &\le \frac{\|C\|_2}{1-\eta_C}\left[\left(\sum_e A_e^2 \epsilon_e^2\right)^{1/2} + \|r\|_2\right] =: \zeta_x. \end{aligned} \tag{B.7}$$

This bound uses one inverse of the full record matrix.

B.5 An ill-conditioned family and an alternative mask

Full rank alone does not bound the noise constant. Take three singleton groups, protected mass one and protected Value zero. Let $a = 1/32$, $|t| \le a$, $0 \le \sigma \le 1/4$ and

$$p = (1+4t, 1-4t, 1), \quad \mu_1 = \frac{1/2+t}{1/4+t} e_1, \quad \mu_2 = \frac{1/2-t}{1/4-t}(e_1 + \sigma e_2), \quad \mu_3 = 2e_3.$$

The group numerators are $p_1\mu_1 = (2+4t)e_1$, $p_2\mu_2 = (2-4t)(e_1+\sigma e_2)$ and $p_3\mu_3 = 2e_3$. Therefore

$$\begin{aligned} y_1 &= \frac{(2+4t)e_1}{2+4t} = e_1, \qquad y_2 = \frac{(2-4t)(e_1+\sigma e_2)}{2-4t} = e_1 + \sigma e_2, \\ y_3 &= 2e_3/2 = e_3, \\ y_{\text{all}} &= \frac{4e_1 + (2-4t)\sigma e_2 + 2e_3}{4} = e_1 + \sigma(1/2-t)e_2 + \tfrac{1}{2}e_3, \\ y_{12} &= \frac{4e_1 + (2-4t)\sigma e_2}{3}, \\ f(t) := e_1^\top y_{13} &= \frac{2+4t}{3+4t} = \frac{1/2+t}{3/4+t}, \\ f'(t) &= \frac{1/4}{(3/4+t)^2} > 0, \qquad f''(t) = -\frac{1/2}{(3/4+t)^3} < 0. \end{aligned}$$

For the five masks $1, 2, 3, 123, 12$, two sources admit a common transcript iff

$$\max\{\sigma, 4\sigma/3\}|t - t'| \le 2\epsilon \quad \Longleftrightarrow \quad |t - t'| \le 2u, \qquad u = \min\{a, 3\epsilon/(4\sigma)\}.$$

At $\sigma = 0$, put $u = a$. The midpoint transcript realizes every such pair. Concavity of f places the maximal separation at $t = -a, t' = -a + 2u$:

$$\begin{aligned} r_5(\sigma,\epsilon) &= \tfrac{1}{2}[f(-a+2u) - f(-a)] \\ &= \frac{(1/2 - a + 2u)(3/4 - a) - (1/2 - a)(3/4 - a + 2u)}{2(3/4-a)(3/4-a+2u)} \\ &= \frac{u/4}{(3/4-a)(3/4-a+2u)}. \end{aligned} \tag{B.8}$$

For any received transcript, the midpoint of its extreme f values attains half its range. The pair above therefore gives both bounds. At $\sigma > 0$, the singleton-contrast determinant is $\sigma \ne 0$, yet

$$\lim_{\epsilon \downarrow 0} \frac{r_5(\sigma,\epsilon)}{\epsilon} = \frac{3}{16\sigma(3/4-a)^2}, \qquad \lim_{\sigma \downarrow 0} r_5(\sigma,\epsilon) = \frac{8}{575} \quad (\epsilon > 0).$$

These sources fit one fixed unit ball. Use a query of norm $\sqrt{d_k}$, reference group means $2e_1, 2e_1, 2e_3$ and zero reference logits. The two logit changes are $\log(1+4t), \log(1-4t)$. For $|t| \le 1/32$,

$$\begin{aligned} |\log(1 \pm 4t)| &= \left|\int_0^{\pm 4t} \frac{ds}{1+s}\right| \le \frac{4|t|}{1-4|t|} \le 1/7, \\ \left|\frac{1/2 \pm t}{1/4 \pm t} - 2\right| &= \frac{|t|}{|1/4 \pm t|} \le 1/7, \qquad \left|\sigma \frac{1/2-t}{1/4-t}\right| \le 15\sigma/7, \\ \|\Delta K\|_F^2 + \|\Delta V\|_F^2 &\le \frac{4}{49} + \frac{225\sigma^2}{49} \le \frac{289}{784} < 1. \end{aligned}$$

The radius and center thus stay fixed as the rank becomes ill-conditioned.

An equally expensive mask can avoid this particular weak direction. The 23 output is $(v, \sigma v, 2-2v)$, where $v = (2-4t)/(3-4t) \in [15/23, 17/25]$ and $f = (4-5v)/(5-6v)$. Consequently

$$\|y_{23}(v) - y_{23}(v')\|_2 = \sqrt{5+\sigma^2}|v-v'|, \qquad \left|\frac{df}{dv}\right| = (5-6v)^{-2} \le (25/23)^2.$$

Every common transcript then has target diameter at most $2(25/23)^2\epsilon/\sqrt{5+\sigma^2}$. Its midpoint estimator has radius at most $(25/23)^2\epsilon/\sqrt{5+\sigma^2}$. This comparison concerns the displayed family, not every cache in its enclosing ball.

B.6 Normalization-aware records

Let $L_S = \log Z_S$ and $b_S = Z_S y_S$. A base-two LSE is first multiplied by $\log 2$. For $g > 1$,

$$\begin{aligned}
\sum_{i=1}^{g} Z_i - Z_{\text{all}} &= \sum_i (Z_0 + Z_{G_i}) - (Z_0 + \sum_i Z_{G_i}) = (g-1)Z_0, \\
\sum_{i=1}^{g} b_i - b_{\text{all}} &= \sum_i (b_0 + b_{G_i}) - (b_0 + \sum_i b_{G_i}) = (g-1)b_0, \\
Z_T &= \sum_{i\in T} Z_i - (|T|-1)Z_0 = \sum_i \xi_{T,i} Z_i + \gamma_T Z_{\text{all}}, \\
b_T &= \sum_i \xi_{T,i} b_i + \gamma_T b_{\text{all}}, \qquad y_T = b_T/Z_T.
\end{aligned} \tag{B.9}$$

Thus $g+1$ output–LSE pairs recover all unions without a contrast-rank test. If Z_0, b_0 are supplied, the g singleton pairs suffice. These are applications of normalization-aware attention merging (Milakov & Gimelshein, 2018; Dao et al., 2022; FlashInfer Contributors, 2025).

Suppose $|\widehat{L}_e - L_e| \le \delta_e$ and $\|\widehat{y}_e - y_e\|_2 \le \epsilon_e$. Put $\widehat{Z}_e = e^{\widehat{L}_e}$ and let $a_i = \xi_{T,i}$, $a_{\text{all}} = \gamma_T$. Then

$$\begin{aligned}
|Z_e - \widehat{Z}_e| &= \widehat{Z}_e |e^{L_e - \widehat{L}_e} - 1| \le \widehat{Z}_e (e^{\delta_e} - 1), \\
\|Z_e y_e - \widehat{Z}_e \widehat{y}_e\|_2 &= \|(Z_e - \widehat{Z}_e)\widehat{y}_e + Z_e(y_e - \widehat{y}_e)\|_2 \\
&\le \widehat{Z}_e[(e^{\delta_e} - 1)\|\widehat{y}_e\|_2 + e^{\delta_e}\epsilon_e].
\end{aligned}$$

Set $\widehat{Z}_T = \sum_e a_e \widehat{Z}_e$ and $\widehat{b}_T = \sum_e a_e \widehat{Z}_e \widehat{y}_e$. Summing the bounds yields

$$\begin{aligned}
|Z_T - \widehat{Z}_T| &\le \sum_e |a_e| \widehat{Z}_e (e^{\delta_e} - 1) =: E_Z, \\
\|b_T - \widehat{b}_T\|_2 &\le \sum_e |a_e| \widehat{Z}_e [(e^{\delta_e} - 1)\|\widehat{y}_e\|_2 + e^{\delta_e}\epsilon_e] =: E_b.
\end{aligned}$$

If $\widehat{Z}_T > E_Z$, let $\widehat{y}_T = \widehat{b}_T / \widehat{Z}_T$. Exact fraction subtraction gives

$$\begin{aligned}
y_T - \widehat{y}_T &= \frac{b_T - \widehat{b}_T + \widehat{y}_T(\widehat{Z}_T - Z_T)}{Z_T}, \\
\|y_T - \widehat{y}_T\|_2 &\le \frac{\|b_T - \widehat{b}_T\|_2 + \|\widehat{y}_T\|_2 |Z_T - \widehat{Z}_T|}{Z_T} \\
&\le \frac{E_b + \|\widehat{y}_T\|_2 E_Z}{\widehat{Z}_T - E_Z}.
\end{aligned} \tag{B.10}$$

All endpoints and the positivity test must be evaluated outward.

LSE records still leave unobserved query directions. Choose w with $q_e^\top w = 0$ for every recorded query, but $q_\star^\top w \neq 0$. Perturb one Key by tw. Then

$$q_e^\top (k_i + tw) = q_e^\top k_i, \qquad q_\star^\top (k_i + tw) = q_\star^\top k_i + t q_\star^\top w.$$

Every recorded output and LSE is unchanged. With unequal Values, the target can change. Stronger normalization access therefore does not eliminate the future-query question.

C PROOF OF THE RECORD-CONDITIONED CERTIFICATE

Fix the reconstruction and records, and put $\mathcal{D} = \mathcal{P}_X - \bar{X}$. Call e has query q_e and visible set $T_e \neq \varnothing$. Define

$$b_e = q_e/\sqrt{d_k}, \quad \bar{Z}_e = \sum_{i \in T_e} e^{b_e^\top \bar{k}_i}, \quad \bar{\alpha}_{ei} = e^{b_e^\top \bar{k}_i}/\bar{Z}_e, \quad \bar{y}_e = \sum_{i \in T_e} \bar{\alpha}_{ei} \bar{v}_i.$$

Sums over i use T_e, or $T_\star$ for a target. For every $\delta \in \mathcal{D}$, choose bounds

$$u_{ei} = b_e^\top \Delta k_i, \qquad |u_{ei}| \le \eta_{ei}, \qquad |c^\top \Delta v_i| \le \omega_i(c). \tag{C.1}$$

For centered enclosing cells with halfwidths H^K, H^V, one choice is $\eta_{ei} = \sum_j |b_{ej}| H^K_{ij}$ and $\omega_i(c) = \sum_j H^V_{ij} |c_j|$. A row-ball bound may further tighten these values. Known rows have zero halfwidths.

C.1 EXACT RESIDUAL AND RECEIVED-RECORD INEQUALITIES

Define

$$\mathcal{L}_e(v)\delta = \sum_i \bar{\alpha}_{ei}[(\bar{v}_i - v)u_{ei} + \Delta v_i], \qquad d_e(\delta) = \sum_i \bar{\alpha}_{ei} e^{u_{ei}}.$$

Starting from the actual attention computation gives

$$\begin{aligned}
y_e - v &= \frac{\sum_i e^{b_e^\top (\bar{k}_i + \Delta k_i)} (\bar{v}_i + \Delta v_i - v)}{\sum_i e^{b_e^\top (\bar{k}_i + \Delta k_i)}} \\
&= \frac{\sum_i \bar{\alpha}_{ei} e^{u_{ei}} (\bar{v}_i + \Delta v_i - v)}{d_e(\delta)}, \\
d_e(\delta)(y_e - v) &= \sum_i \bar{\alpha}_{ei}[(1 + u_{ei})(\bar{v}_i - v) + \Delta v_i] \\
&\quad + \sum_i \bar{\alpha}_{ei}[(e^{u_{ei}} - 1 - u_{ei})(\bar{v}_i - v) + (e^{u_{ei}} - 1)\Delta v_i] \\
&= \bar{y}_e - v + \mathcal{L}_e(v)\delta + R_e(v, \delta).
\end{aligned} \tag{C.2}$$

The second sum defines R_e. For $|u| \le \eta$, Taylor's integral identity gives

$$\begin{aligned}
0 \le e^u - 1 - u = u^2 \int_0^1 (1-t) e^{tu}\, dt \le \eta^2 \int_0^1 (1-t) e^{t\eta}\, dt = e^\eta - 1 - \eta =: r_{\exp}(\eta), \\
|e^u - 1| = \left| u \int_0^1 e^{tu}\, dt \right| \le \eta \int_0^1 e^{t\eta}\, dt = e^\eta - 1.
\end{aligned} \tag{C.3}$$

Both bounds vanish at $\eta = 0$. Consequently

$$\begin{aligned}
|c^\top R_e(v, \delta)| &\le \sum_i \bar{\alpha}_{ei}[|c^\top (\bar{v}_i - v)|\, |e^{u_{ei}} - 1 - u_{ei}| + |e^{u_{ei}} - 1|\, |c^\top \Delta v_i|] \\
&\overset{\text{(C.1),(C.3)}}{\le} \sum_i \bar{\alpha}_{ei}[|c^\top (\bar{v}_i - v)| r_{\exp}(\eta_{ei}) + (e^{\eta_{ei}} - 1)\omega_i(c)] = E_e(c, v),
\end{aligned} \tag{C.4}$$

$$0 < d_e^- := \sum_i \bar{\alpha}_{ei} e^{-\eta_{ei}} \le d_e(\delta) \le \sum_i \bar{\alpha}_{ei} e^{\eta_{ei}} =: d_e^+. \tag{C.5}$$

Set $v = \widehat{y}_e$, $r_e = \widehat{y}_e - \bar{y}_e$ and $\zeta_e = y_e - \widehat{y}_e$. Compatibility means $\|\zeta_e\|_2 \le \epsilon_e$. Then

$$\begin{aligned}
\mathcal{L}_e(\widehat{y}_e)\delta - r_e &= d_e(\delta)\zeta_e - R_e(\widehat{y}_e, \delta), \\
|c^\top [\mathcal{L}_e(\widehat{y}_e)\delta - r_e]| &\le d_e(\delta)|c^\top \zeta_e| + |c^\top R_e(\widehat{y}_e, \delta)| \\
&\overset{\text{Cauchy–Schwarz}}{\le} \epsilon_e d_e^+ \|c\|_2 + E_e(c, \widehat{y}_e) =: \mathfrak{e}_e(c).
\end{aligned} \tag{C.6}$$

The expansion point is the received output. It need not equal either the reconstructed output or the true output.

C.2 TARGET ELIMINATION, CONVEXITY AND FINITE CHECKING

For direction a, put $f_\star = a^\top y_\star$ and $\bar{f}_\star = a^\top \bar{y}_\star$. The scalar residual is

$$\begin{aligned} d_\star(\delta)(f_\star - u) &= \bar{f}_\star - u + \ell_\star(u)^\top \delta + \rho_\star(u,\delta), \\ \ell_\star(u)^\top \delta &= \sum_i \bar{\alpha}_{\star i}[(a^\top \bar{v}_i - u)u_{\star i} + a^\top \Delta v_i], \\ |\rho_\star(u,\delta)| &\le \sum_i \bar{\alpha}_{\star i}[|a^\top \bar{v}_i - u| r_{\exp}(\eta_{\star i}) + (e^{\eta_{\star i}} - 1)\omega_i(a)] = E_\star(u). \end{aligned}$$

For arbitrary $c = (c_e)_e$, define $g = \ell_\star(u) - \sum_e \mathcal{L}_e(\widehat{y}_e)^\top c_e$. For every $X \in \mathcal{P}_Y$,

$$\begin{aligned} d_\star(\delta)(f_\star - u) &= \bar{f}_\star - u + g^\top \delta + \sum_e c_e^\top \mathcal{L}_e(\widehat{y}_e)\delta + \rho_\star(u,\delta) \\ &= \bar{f}_\star - u + \sum_e c_e^\top r_e + g^\top \delta \\ &\quad + \sum_e c_e^\top [\mathcal{L}_e(\widehat{y}_e)\delta - r_e] + \rho_\star(u,\delta) \\ &\overset{(C.6)}{\le} \bar{f}_\star - u + \sum_e c_e^\top r_e + \underbrace{\sup_{\delta' \in \mathcal{D}} g^\top \delta'}_{h_{\mathcal{D}}(g)} + \sum_e \mathfrak{e}_e(c_e) + E_\star(u) \\ &= \mathcal{B}_+(u,c), \end{aligned}$$

$$\mathcal{B}_+(u,c) \le 0 \implies f_\star - u \le \mathcal{B}_+(u,c)/d_\star(\delta) \le 0. \tag{C.7}$$

The same chain for $-a$ yields $-f_\star \le u_-$. Thus $f_\star \in [-u_-, u_+]$, as asserted in Theorem 2.

For a fixed coefficient tuple, set

$$\widehat{f} = \bar{f}_\star + \sum_e c_e^\top r_e, \qquad g_c = \ell_\star(\widehat{f}) - \sum_e \mathcal{L}_e(\widehat{y}_e)^\top c_e.$$

The constant term in the residual is zero. Hence

$$\begin{aligned} d_\star(\delta)|f_\star - \widehat{f}| &\le |g_c^\top \delta| + \sum_e |c_e^\top [\mathcal{L}_e(\widehat{y}_e)\delta - r_e]| + |\rho_\star(\widehat{f},\delta)| \\ &\le \max\{h_{\mathcal{D}}(g_c), h_{\mathcal{D}}(-g_c)\} + \sum_e \mathfrak{e}_e(c_e) + E_\star(\widehat{f}), \\ |f_\star - \widehat{f}| &\le \frac{\max\{h_{\mathcal{D}}(g_c), h_{\mathcal{D}}(-g_c)\} + \sum_e \mathfrak{e}_e(c_e) + E_\star(\widehat{f})}{d_\star^-}. \end{aligned} \tag{C.8}$$

If the stored center is $\widetilde{f}$, the checker adds a bound on $|\widetilde{f} - \widehat{f}|$, since $|f_\star - \widetilde{f}| \le |f_\star - \widehat{f}| + |\widehat{f} - \widetilde{f}|$.

For $0 \le t \le 1$,

$$\begin{aligned} h_{\mathcal{D}}(tw + (1-t)w') &= \sup_{\delta \in \mathcal{D}}[tw^\top \delta + (1-t)w'^\top \delta] \\ &\le t h_{\mathcal{D}}(w) + (1-t)h_{\mathcal{D}}(w'). \end{aligned}$$

The map $(u,c) \mapsto g(u,c)$ is affine. Each error term is a positive sum of absolute affine functions or norms. Therefore

$$\mathcal{B}_+(t(u,c) + (1-t)(u',c')) \le t\mathcal{B}_+(u,c) + (1-t)\mathcal{B}_+(u',c').$$

This proves the asserted convexity.

For a box $[l,h]$, let $w_+ = \max(w,0)$ and $w_- = \min(w,0)$. Every split vector z and every $\delta \in [l,h] \cap R\mathbb{B}_2$ satisfy

$$\begin{aligned} w^\top \delta &= z^\top \delta + (w-z)^\top \delta \\ &\le \|z\|_2 \|\delta\|_2 + \sum_j [h_j(w_j - z_j)_+ + l_j(w_j - z_j)_-] \\ &\le R\|z\|_2 + \sum_j [h_j(w_j - z_j)_+ + l_j(w_j - z_j)_-]. \end{aligned} \tag{C.9}$$

Taking the supremum on the left proves a support bound for every z. Assume $A \cap B \neq \varnothing$, where $A = R\mathbb{B}_2$ and $B = [l, h]$. Split the common variable into $u \in A$ and $v \in B$ with $u = v$:

$$\begin{aligned} h_{A\cap B}(w) &= \max_{u\in A,\, v\in B,\, u=v} w^\top v, \\ \mathcal{L}(u, v, z) &= w^\top v + z^\top (u - v) = z^\top u + (w - z)^\top v, \\ \sup_{u\in A,\, v\in B} \mathcal{L}(u, v, z) &= h_A(z) + h_B(w - z) = R\|z\|_2 + h_B(w - z). \end{aligned}$$

Strong duality under the relative-interior condition (Boyd & Vandenberghe, 2004, Section 5.2.3) gives

$$\mathrm{ri}(R\mathbb{B}_2) \cap \mathrm{ri}[l, h] \neq \varnothing \implies h_{R\mathbb{B}_2\cap[l,h]}(w) = \min_z \{R\|z\|_2 + h_{[l,h]}(w - z)\}.$$

This condition guarantees equality and attainment. The upper bound (C.9) holds without it. Norm epigraphs are second-order cones. Absolute-value epigraphs are $t \geq x$ and $t \geq -x$. Substitution gives an SOCP with $O(N_{\mathrm{rec}}n + n(d_k + d_v) + N_{\mathrm{rec}}d_v)$ variables. The $N_{\mathrm{rec}}n$ term counts directional-remainder epigraphs. Fixed coefficients cost $O(N_{\mathrm{rec}}n(d_k + d_v))$ arithmetic operations. This is an evaluation cost, not an exact SOCP-solving or bit-complexity bound.

C.3 Post-selected queries, known rows and additional records

The quantifiers proved above are

$$\forall X \in \mathcal{P}_Y \;\; \forall q_\star \in \mathbb{R}^{d_k} \; \forall T_\star \neq \varnothing \;\; \forall a : \|a\|_2 = 1 \; \forall (u, c) : \quad \mathcal{B}_+(u, c) \leq 0 \Longrightarrow f_\star(X) \leq u.$$

Substituting functions $q_\star(Y), T_\star(Y), a(Y), c(Y)$ preserves this implication. No independence or union bound is used. For known appended rows, $\Delta k_i = \Delta v_i = 0$ implies $u_{ei} = \eta_{ei} = \omega_i(c) = 0$. These rows change the reference weights but add no residual error. Both computations use the same realized query and live suffix.

For head-projection blocks W_h, put $a_h = W_h^\top a$. An interval $[L_h, U_h]$ in direction $a_h/\|a_h\|_2$ gives

$$\sum_{h:a_h\neq 0} \|a_h\|_2 L_h \leq \sum_h a_h^\top y_h = a^\top \sum_h W_h y_h \leq \sum_{h:a_h\neq 0} \|a_h\|_2 U_h.$$

Shared Keys do not alter these termwise inequalities.

A saved LSE gives another strip. Define $F(t) = \log \sum_i \bar{\alpha}_{ei} e^{t u_{ei}}$ and $p_i(t) = \bar{\alpha}_{ei} e^{t u_{ei} - F(t)}$. The identity $\sum_i p_i(t)(u_{ei} - z)^2 = \sum_i p_i(t)(u_{ei} - F'(t))^2 + (z - F'(t))^2$ gives the minimum over z below. Differentiation and integration yield

$$\begin{aligned} F'(t) &= \sum_i p_i(t) u_{ei}, \\ F''(t) &= \sum_i p_i(t)(u_{ei} - F'(t))^2, \\ 0 \leq F''(t) &= \min_z \sum_i p_i(t)(u_{ei} - z)^2 \\ &\leq \sum_i p_i(t) \left(u_{ei} - \frac{u_{\max} + u_{\min}}{2} \right)^2 \leq \frac{(u_{\max} - u_{\min})^2}{4}, \\ L_e - \bar{L}_e - \sum_i \bar{\alpha}_{ei} u_{ei} &= F(1) - F(0) - F'(0) = \int_0^1 (1 - t) F''(t)\, dt \\ &\in [0, (u_{\max} - u_{\min})^2/8]. \end{aligned}$$

An observed LSE interval can thus be imposed with this explicit remainder.

C.4 Strict gain from gradient alignment

For $\operatorname{diam} f(\mathcal{A}) = \sup_{X, X' \in \mathcal{A}} |f(X) - f(X')|$, a verified width W_Y and verified sources $X^\pm \in \mathcal{P}_X$ give

$$\operatorname{diam} f(\mathcal{P}_Y) \leq W_Y < |f(X^+) - f(X^-)| \leq \operatorname{diam} f(\mathcal{P}_X). \tag{C.10}$$

The sources need not satisfy the newly imposed records.

For the input-side condition, suppose $\mathcal{D} = R\mathbb{B}_2$, $R > 0$, and

$$f(\bar{X}+\delta) = f_0 + j^\top\delta + r_\star(\delta), \qquad y(\bar{X}+\delta) = y_0 + b^\top\delta + r_1(\delta), \qquad |r_\star| \le e_\star, \quad |r_1| \le e_1.$$

The exact residual supplies these affine remainders. In particular, with $J = \ell_\star(\bar{f}_\star)$,

$$f_\star - \bar{f}_\star - J^\top\delta = \frac{\rho_\star(\bar{f}_\star,\delta) + (1-d_\star(\delta))J^\top\delta}{d_\star(\delta)},$$
$$|f_\star - \bar{f}_\star - J^\top\delta| \le \frac{E_\star(\bar{f}_\star) + R\|J\|_2 \max\{d_\star^+ - 1, 1 - d_\star^-\}}{d_\star^-} =: e_\star.$$

The scalar record uses the same calculation. Write $r = \widehat{y} - y_0$ and $s = \epsilon + e_1$. For every compatible received value and every $\lambda \in \mathbb{R}$,

$$\begin{aligned} |b^\top\delta - r| &\le |y - \widehat{y}| + |r_1| \le s, \\ |f - (f_0 + \lambda r)| &= |(j-\lambda b)^\top\delta + \lambda(b^\top\delta - r) + r_\star| \\ &\le R\|j - \lambda b\|_2 + s|\lambda| + e_\star. \end{aligned} \tag{C.11}$$

For $j, b \neq 0$, let $\rho = |j^\top b|/(\|j\|_2\|b\|_2)$, $\tau = s/(R\|b\|_2)$ and $\sigma = \operatorname{sign}(j^\top b)$. When $\rho > 0$, set $\lambda = \sigma\|j\|_2 x/\|b\|_2$. Then

$$R\|j - \lambda b\|_2 + s|\lambda| = R\|j\|_2 \underbrace{\left[\sqrt{(x-\rho)^2 + 1 - \rho^2} + \tau|x|\right]}_{F(x)}.$$

Replacing $x < 0$ by $|x|$ cannot increase this expression. For $0 \le \tau < \rho < 1$, the convex function on $x \ge 0$ satisfies

$$F'(x) = \frac{x-\rho}{\sqrt{(x-\rho)^2 + 1 - \rho^2}} + \tau,$$
$$F'(x) = 0 \iff x = x_* := \rho - \frac{\tau\sqrt{1-\rho^2}}{\sqrt{1-\tau^2}} > 0,$$
$$\begin{aligned} F(x_*) &= \frac{\sqrt{1-\rho^2}}{\sqrt{1-\tau^2}} + \rho\tau - \frac{\tau^2\sqrt{1-\rho^2}}{\sqrt{1-\tau^2}} \\ &= \sqrt{1-\rho^2}\sqrt{1-\tau^2} + \rho\tau = \varphi(\rho,\tau). \end{aligned}$$

At $\rho = 1$, the objective is $|1-x| + \tau x$, with minimum τ when $\tau < 1$. If $\tau \ge \rho$, $F'(0+) = \tau - \rho \ge 0$ and the minimum is $F(0) = 1$. For $\rho = 0$, the minimum is also attained at zero. For $s > 0$, $R\|j - \lambda b\|_2 + s|\lambda|$ is R times the objective in Ettehad & Foucart (2021, Theorem 4). Use a Euclidean model space, subspace $\{0\}$, radius R, and one observation with error bound s. The calculation above also covers $s = 0$. The attention remainder $e_\star$ is added separately. The bound (C.11) holds for every compatible received value r.

For $\delta_\pm = \pm Rj/\|j\|_2$,

$$\begin{aligned} f(\bar{X}+\delta_+) - f(\bar{X}+\delta_-) &= 2R\|j\|_2 + r_\star(\delta_+) - r_\star(\delta_-) \\ &\ge 2R\|j\|_2 - 2e_\star, \\ 2R\|j\|_2\varphi + 2e_\star < 2R\|j\|_2 - 2e_\star &\iff R\|j\|_2(1-\varphi) > 2e_\star. \end{aligned}$$

The last inequality and (C.11) prove strict gain for every compatible received value. With cells, $\bar{X} + \delta_\pm$ must also satisfy their coordinate bounds. Otherwise use the direct test (C.10).

D Physical outer bounds and when they are exact

The residual certificate works in the original cache coordinates. At a fixed query, group summaries give a second outer bound. Its endpoint equals the physical optimum when a compatible cache realizes it.

D.1 Constructing the summary prior

Fix q, put $b = q/\sqrt{d_k}$ and partition positions into $G_0 = B, G_1, \ldots, G_g$. Intersect each coordinate cell with its radius-R coordinate bounds. Write their endpoints as $k_{ja}^{\pm}, v_{ja}^{\pm}$. The following bounds are finite:

$$s_j^- = \max\{\bar{s}_j - R\|b\|_2,\ \sum_a \min(b_a k_{ja}^-, b_a k_{ja}^+)\},$$
$$s_j^+ = \min\{\bar{s}_j + R\|b\|_2,\ \sum_a \max(b_a k_{ja}^-, b_a k_{ja}^+)\}, \qquad \bar{s}_j = b^\top \bar{k}_j. \tag{D.1}$$

For every physical source,

$$|b^\top(k_j - \bar{k}_j)| \le \|b\|_2 \|k_j - \bar{k}_j\|_2 \le R\|b\|_2,$$
$$\sum_a \min(b_a k_{ja}^-, b_a k_{ja}^+) \le \sum_a b_a k_{ja} \le \sum_a \max(b_a k_{ja}^-, b_a k_{ja}^+),$$
$$s_j^- \le b^\top k_j \le s_j^+ \Longrightarrow P_i^- := \sum_{j\in G_i} e^{s_j^-} \le P_i := \sum_{j\in G_i} e^{b^\top k_j} \le P_i^+ := \sum_{j\in G_i} e^{s_j^+}.$$

Normalize by $\pi_i = P_i/\sum_h P_h$ and $m_i = \pi_i \mu_i$. For $p_i = P_i/P_0$,

$$p_i^- := P_i^-/P_0^+ \le p_i \le P_i^+/P_0^- =: p_i^+,$$
$$\pi_i^- := \frac{P_i^-}{P_i^- + \sum_{h\ne i} P_h^+} \le \frac{P_i}{P_i + \sum_{h\ne i} P_h} \le \frac{P_i^+}{P_i^+ + \sum_{h\ne i} P_h^-} =: \pi_i^+. \tag{D.2}$$

For the conditional weights $\vartheta_{ij} = e^{b^\top k_j}/P_i$, $\vartheta_{ij} \ge 0$ and $\sum_{j\in G_i} \vartheta_{ij} = 1$. Thus

$$\mu_{ia}^- := \min_{j\in G_i} v_{ja}^- \le \sum_{j\in G_i} \vartheta_{ij} v_{ja} = \mu_{ia} \le \max_{j\in G_i} v_{ja}^+ =: \mu_{ia}^+.$$

Each bound implies a linear summary constraint:

$$p_i^- \pi_0 \le \pi_i \le p_i^+ \pi_0, \qquad \mu_{ia}^- \pi_i \le m_{ia} \le \mu_{ia}^+ \pi_i, \qquad \sum_i \pi_i = 1, \quad \pi_i^- \le \pi_i \le \pi_i^+.$$

Their feasible set Θ_{prior} contains every physical summary. It is convex because all constraints are linear. It is closed and bounded because $\pi_i^- > 0$, $\pi_i \le 1$ and $|m_{ia}| \le \max(|\mu_{ia}^-|, |\mu_{ia}^+|)$. A separately verified mean ball adds the cone $\|m_i - \pi_i \mu_i^c\|_2 \le r_i \pi_i$. This proves Proposition 3 for the stated constants. Tighter valid coordinate or mean bounds can be substituted without changing the inclusion proof.

D.2 Record constraints and fractional scaling

Let $d_S = \pi_0 + \sum_{i\in S} \pi_i$ and $n_S = m_0 + \sum_{i\in S} m_i$. At this fixed query, each received record satisfies the equivalence

$$\|y_S - \widehat{y}_S\|_2 \le \epsilon_S \Longleftrightarrow \|n_S/d_S - \widehat{y}_S\|_2 \le \epsilon_S$$
$$\overset{d_S>0}{\Longleftrightarrow} \|n_S - d_S \widehat{y}_S\|_2 \le \epsilon_S d_S. \tag{D.3}$$

Intersecting these cones with Θ_{prior} gives Θ_Y. For a target $f = a^\top n_T/d_T$, set $t = 1/d_T$, $\widetilde{\pi} = t\pi$, $\widetilde{m} = tm$. Then

$$d_T(\widetilde{\pi}) = 1, \qquad \sum_i \widetilde{\pi}_i = t, \qquad \pi_i^- t \le \widetilde{\pi}_i \le \pi_i^+ t,$$
$$p_i^- \widetilde{\pi}_0 \le \widetilde{\pi}_i \le p_i^+ \widetilde{\pi}_0, \qquad \mu_{ia}^- \widetilde{\pi}_i \le \widetilde{m}_{ia} \le \mu_{ia}^+ \widetilde{\pi}_i,$$
$$\|n_S(\widetilde{m}) - d_S(\widetilde{\pi})\widehat{y}_S\|_2 \le \epsilon_S d_S(\widetilde{\pi}), \qquad f = a^\top n_T(\widetilde{m}). \tag{D.4}$$

The mean-ball cones scale in the same way. Conversely, $d_T(\widetilde{\pi}) = 1$ and $\widetilde{\pi}_i \ge 0$ imply $t > 0$. Dividing every constraint by t recovers a point of Θ_Y. The substitutions are inverse maps:

$$(\pi, m) \mapsto (t, t\pi, tm) \mapsto (\pi, m), \qquad (t, \widetilde{\pi}, \widetilde{m}) \mapsto (\widetilde{\pi}/t, \widetilde{m}/t) \mapsto (t, \widetilde{\pi}, \widetilde{m}).$$

This is the positive-denominator fractional transformation (Charnes & Cooper, 1962). Its LP or SOCP optimum is exact on Θ_Y. Physical equality requires the reverse inclusion below.

D.3 AN EXPLICIT LIFT AND ENDPOINT EQUALITY

Fix a reference X_0 and a nonzero query. Allow all coordinates in G_i, $i \geq 1$, to vary. Allow background Values to vary on a nonempty subset $J_0 \subseteq B$; keep background Keys and all other background coordinates fixed. For a supplied background mean, omit its mutable block. Define

$$p_i^0 = P_i^0/P_0^0, \quad \vartheta_{ij} = e^{q^\top k_j^0/\sqrt{d_k}}/P_i^0, \quad \kappa_i = \sum_{j\in G_i} \vartheta_{ij}^2, \quad \kappa_0 = \sum_{j\in J_0} \vartheta_{0j}^2.$$

For desired ratios $p_i > 0$ and means μ_i, use

$$\Delta k_j = \frac{\sqrt{d_k}q}{\|q\|_2^2}\log(p_i/p_i^0), \qquad \Delta v_j = \frac{\vartheta_{ij}}{\kappa_i}(\mu_i - \mu_i^0) \quad (j \in G_i). \tag{D.5}$$

On J_0, use the same Value formula with $i = 0$. All fixed coordinates have zero shift. For each editable group,

$$\begin{aligned} q^\top(k_j^0 + \Delta k_j)/\sqrt{d_k} &= q^\top k_j^0/\sqrt{d_k} + \log(p_i/p_i^0), \\ P_i(X_\theta) &= \sum_{j\in G_i} e^{q^\top k_j^0/\sqrt{d_k}}(p_i/p_i^0) = P_i^0(p_i/p_i^0) = P_0^0 p_i, \\ \frac{e^{q^\top(k_j^0+\Delta k_j)/\sqrt{d_k}}}{P_i(X_\theta)} &= \vartheta_{ij}, \\ \mu_i(X_\theta) - \mu_i^0 &= \sum_j \vartheta_{ij}\Delta v_j = \frac{\sum_j \vartheta_{ij}^2}{\kappa_i}(\mu_i - \mu_i^0) = \mu_i - \mu_i^0. \end{aligned}$$

The background uses the same last identity on J_0. Squaring the shifts gives

$$\begin{aligned} \|X_\theta - X_0\|_F^2 &= \sum_{i=1}^g \sum_{j\in G_i} \frac{d_k}{\|q\|_2^2}\log^2(p_i/p_i^0) + \sum_{i=0}^g \sum_{j\in J_i} \frac{\vartheta_{ij}^2}{\kappa_i^2}\|\mu_i - \mu_i^0\|_2^2 \\ &= \frac{d_k}{\|q\|_2^2}\sum_{i=1}^g |G_i|\log^2(p_i/p_i^0) + \sum_{i=0}^g \frac{\|\mu_i - \mu_i^0\|_2^2}{\kappa_i} =: \mathcal{C}_0(\theta), \end{aligned} \tag{D.6}$$

where $J_i = G_i$ for $i \geq 1$. The background term is omitted when its mean is supplied. Cauchy–Schwarz verifies the minimum Value cost for these fixed weights:

$$\|\mu_i - \mu_i^0\|_2^2 = \left\|\sum_{j\in J_i} \vartheta_{ij}\Delta v_j\right\|_2^2 \leq \kappa_i \sum_{j\in J_i} \|\Delta v_j\|_2^2.$$

Equality holds for the displayed shifts. Other Key changes may give a cheaper joint lift; no such optimality is required here.

Let X_0 have slack $\rho = R - \|X_0 - \bar{X}\|_F > 0$. For any $\theta \in \Theta_Y$ whose lift satisfies all coordinate cells,

$$\begin{aligned} \mathcal{C}_0(\theta) \leq \rho^2 \Longrightarrow \|X_\theta - \bar{X}\|_F &\leq \|X_0 - \bar{X}\|_F + \|X_\theta - X_0\|_F \\ &\leq R - \rho + \sqrt{\mathcal{C}_0(\theta)} \leq R, \\ \theta \in \Theta_Y, \quad \Phi_q(X_\theta) = \theta \Longrightarrow X_\theta &\in \mathcal{P}_Y. \end{aligned} \tag{D.7}$$

If $\theta_\pm$ attain the aggregate endpoints and both pass this test,

$$L_{\text{agg}} \leq L_{\text{phys}} \leq f(X_{\theta_-}) = L_{\text{agg}}, \qquad U_{\text{agg}} \geq U_{\text{phys}} \geq f(X_{\theta_+}) = U_{\text{agg}}.$$

Both equalities follow. If every $\theta \in \Theta_Y$ passes, the whole physical and aggregate summary sets coincide. Write $r_{\text{agg}} = (U_{\text{agg}} - L_{\text{agg}})/2$ and $r_{\text{phys}} = (U_{\text{phys}} - L_{\text{phys}})/2$. With feasible inner sources $X_\pm$ ordered so $f(X_-) \leq f(X_+)$,

$$0 \leq r_{\text{agg}} - r_{\text{phys}} \leq \frac{U_{\text{agg}} - L_{\text{agg}} - f(X_+) + f(X_-)}{2}. \tag{D.8}$$

A uniform lift test follows from a bound on the record matrix. Let x_0 be the reference moment vector and B_0 its exact-record matrix. Suppose $\sigma_{\min}(B_0) \geq \sigma_* > 0$, $p_i \in [\underline{p}_i, \overline{p}_i]$ for both source and reference, and $A_* = 1 + \sum_i \overline{p}_i$. If both source x and reference x_0 are within ϵ of the same received records, then

$$\begin{aligned}
\|y_e(x) - y_e(x_0)\|_2 &\leq \|y_e(x) - \widehat{y}_e\|_2 + \|\widehat{y}_e - y_e(x_0)\|_2 \leq 2\epsilon, \\
\|(B_0(x - x_0))_e\|_2 &= d_{S_e}(x)\|y_e(x_0) - y_e(x)\|_2 \leq 2A_*\epsilon, \\
\sigma_*\|x - x_0\|_2 &\leq \|B_0(x - x_0)\|_2 \leq 2\sqrt{N_{\mathrm{rec}}}A_*\epsilon, \\
\|x - x_0\|_2 &\leq \zeta := 2\sqrt{N_{\mathrm{rec}}}A_*\epsilon/\sigma_*.
\end{aligned}$$

The masses and means therefore satisfy

$$\begin{aligned}
|\log(p_i/p_i^0)| &= \left|\int_{p_i^0}^{p_i} \frac{ds}{s}\right| \leq |p_i - p_i^0|/\underline{p}_i \leq \zeta/\underline{p}_i, \\
\|\mu_i - \mu_i^0\|_2 &= \left\|\frac{w_i - w_i^0 - (p_i - p_i^0)\mu_i^0}{p_i}\right\|_2 \leq \frac{1 + \|\mu_i^0\|_2}{\underline{p}_i}\zeta, \qquad \|\mu_0 - \mu_0^0\|_2 \leq \zeta, \\
\mathcal{C}_0(\theta) &\leq \zeta^2\left[\frac{d_k}{\|q\|_2^2}\sum_i \frac{|G_i|}{\underline{p}_i^2} + \sum_i \frac{(1 + \|\mu_i^0\|_2)^2}{\kappa_i \underline{p}_i^2} + \kappa_0^{-1}\right].
\end{aligned}$$

If this last bound is at most ρ^2, all ball constraints pass. Coordinate cells must still be checked. On a pure ball with an interior reference, the lift is smooth in positive p_i and equals X_0 at the reference summaries. It supplies the local realization chart assumed by the count theorem.

E Experimental protocols and checking costs

E.1 Frozen caches, record acquisition and codecs

The frozen-array test uses four fixed contexts from SmolLM2-1.7B-Instruct, at layer 23 and head 0. Ordinary causal outputs at positions 32–35 are recomputed in binary32 and saved in that format. The error allowance is 10^{-3} for the mixed-precision study and 10^{-4} for the equal-precision study. Their full-vector acquisition errors are separately enclosed by directed interval arithmetic. Positions 0–31 are uncertain; positions 32 onward remain exact and each query sees only its causal prefix. Four fixed coordinate directions, twelve later queries and three declared masks produce 144 tasks per context and precision. All combinations are included.

The physical prior retains the coordinate quantization cells and their joint squared-error consequence. For IEEE binary16 Keys we initially use a symmetric halfwidth enclosing the two adjacent rounding half-gaps; at binade boundaries this is an outer box for the true asymmetric cell. For strict physical comparisons, witnesses use the smaller adjacent half-gap on both Key sides and are checked by an actual binary16 round trip. The round trip checks membership in the actual codec cell. Signed INT4/INT8 Value quantizers use a power-of-two step per row, large enough that no input is clipped. The alternative equal-precision study uses row min–max quantizers at 8, 10 and 12 bits, with an explicit binary64 reconstruction allowance. These two codec families define different physical priors and are evaluated separately.

E.2 Coefficient proposals and numerical checking

The coefficient proposal uses a 256×256 weighted Gram matrix assembled without a dense physical-coordinate Jacobian. It proposes a ridge solution and minimizes the actual convex certificate along that ray. A target-direction coordinate-descent proposal and the zero tuple are additional candidates. Global minimization is unnecessary for Theorem 2. The zero tuple also ensures that choosing from these candidates cannot worsen the same codec-only bound.

The independent checker reads stored dyadic inputs and coefficient tuples. It encloses reference weights, residuals, centers and errors; it neither imports the coefficient proposal nor invokes an optimizer. Transcendentals are evaluated with 65-digit interval exponentials. Binary64 sums include an explicit $\gamma_n = nu/(1 - nu)$ allowance and outward rounding. The theorem bounds every

compatible cache. Membership checks verify the physical witnesses used to measure information gain. The strict information-gain count compares a certified posterior width to an independently evaluated physical codec-only pair.

The equal-precision min–max study has 576 tasks at each bit width. At 8 bits it supplies no strict physical-gain certificate. At 10 and 12 bits the corresponding counts are 323 and 534. All 1728 target outputs lie inside the independently checked intervals. These are controlled precision comparisons on fixed arrays, rather than estimated success rates over language-model inputs.

The timing benchmark separates coefficient proposal, floating bound evaluation and outward verification. It uses synthetic arrays at 512, 2048, 8192 and 32768 positions with widths 64 and four vector records. The persistent Gram matrix stays 256×256. These synthetic-array tests measure computational scaling.

The fast implementation keeps the same mathematical certificate. It uses one Gram-based coefficient ray plus zero, and a vectorized outward exponential. Range reduction places arguments in $[-1/16, 1/16]$; degree-12 Horner evaluation has remainder at most $2(1/16)^{13}/13!$, followed by outward squaring. Small positive-argument remainders are evaluated after factoring their quadratic term. Checks of 5151 exponential intervals and 2076 remainder cases compare this implementation with 90-digit directed intervals. On the full 1728-task small-prefix grid, this faster proposal gives 550/576 strict physical comparisons for FP16/INT4 and 222/576 for FP16/INT8; INT8/INT8 gives zero. The FP16/INT4 median ratios by context are 0.803, 0.800, 0.754 and 0.666. These bound ratios and strict-gain counts use the same coefficient proposal as the timings below. All timings below use one CPU thread and include the actual outward target check, separately from coefficient construction.

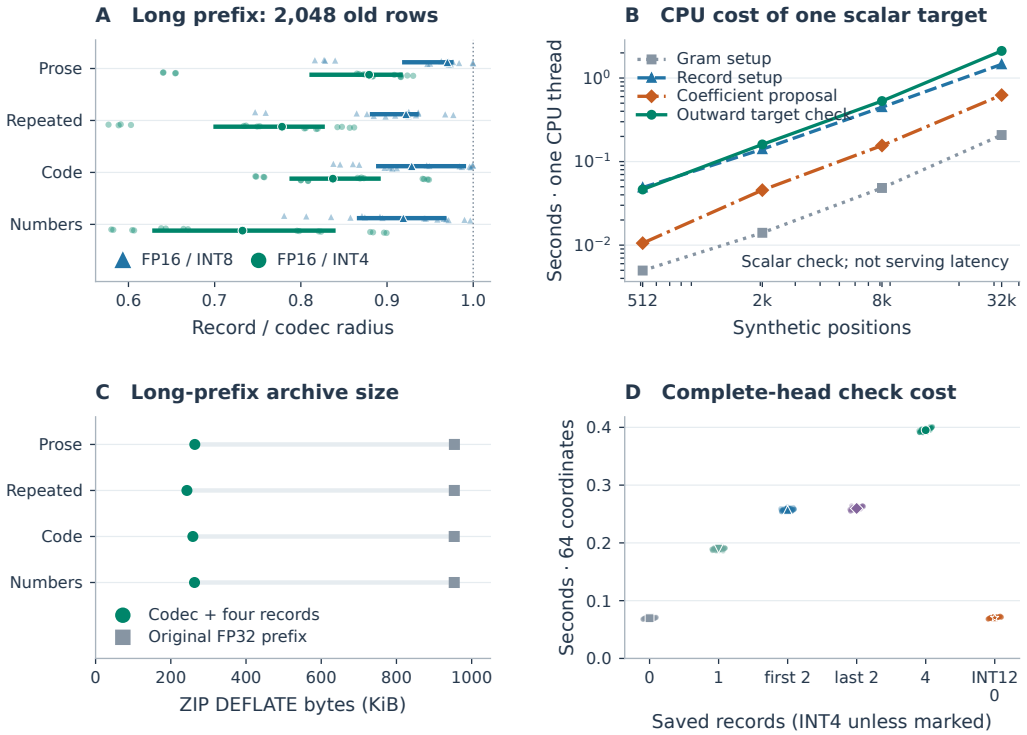

Figure 3: Long-prefix bounds, computation times and storage sizes. (A) All 128 long-prefix scalar tasks. Small points show tasks; large markers and thick segments show the median and interquartile range. (B) Synthetic-size timing separates reusable setup from per-target work. Each point is one single-thread measurement. (C) Both long-prefix artifacts use ZIP DEFLATE and omit the live suffix. (D) All 48 complete-head check times per scheme from the short-prefix matched-byte run, with median and interquartile range. The displayed spreads are descriptive, not confidence intervals.

E.3 FOUR-COORDINATE AND COMPLETE-HEAD RETENTION DECISIONS

For the four-readout decision control, the target is the original intact output, not the original output under the candidate mask. Candidate outputs use the retained reconstruction and exact suffix, calculated in binary32. The loss is maximum absolute error on head coordinates 0–3. The complete candidate family keeps 0, 8, 16, 24 or 32 old rows, in either recent or reconstructed-score order. Figure 4A shows every tested budget on all 48 future queries of the FP16/INT4 test. The loss and mask family are fixed before these comparisons. The necessity curve counts decisions for which a verified same-code physical source violates the selected mask's budget before record constraints are imposed. The witness rules out a codec-only guarantee for that mask and loss. Each accepted mask satisfies all four coordinate bounds simultaneously.

The full-head decision check uses all 64 original intact coordinate intervals at each of the 48 later queries, the same four records with vector error 10^{-3}, and the fast one-ray coefficient mode. Each candidate's executed binary32 output is compared with this original intact interval, and (3) combines the coordinate bounds into a Euclidean error bound. All 3072 coordinate certificates are checked. Figure 4B shows the complete-head fixed-budget sweep. A necessary witness is an original cache with the identical stored codes for which that candidate exceeds the head-norm budget; it proves that codec-only evidence cannot imply the selected guarantee. For adaptive power-of-two quantizers, these witnesses preserve the stored row exponent as well as the integer codes: nonzero rows satisfy $q_{\max}\Delta/2 < \max_j |v_j| \le q_{\max}\Delta$, where Δ is the saved step, while zero rows stay zero. This extra test prevents a candidate from causing the encoder to choose a different scale. It is also applied to the earlier exact-code witness experiments. Checking all decisions and witnesses takes 27.1 seconds. The median time for one 64-coordinate certificate is 0.483 seconds. Appendix E.5 reports timings for the matched-byte comparison. These full-head results concern one attention call. They have a different loss from the four-readout control and the longer-prefix scalar tests.

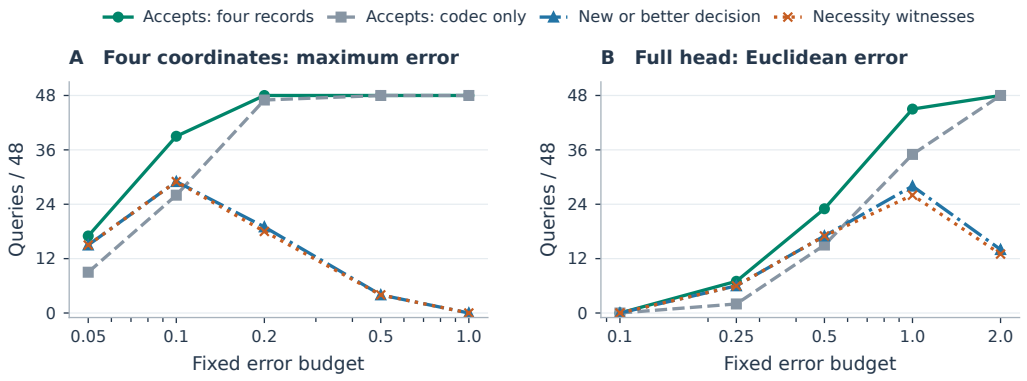

Figure 4: Complete fixed-budget sweeps for two different losses. Acceptance means that at least one of the ten candidate masks passes. "New or better" means that a previously unresolved query is certified, or that the chosen mask retains fewer rows than the codec-only choice. New certification can include full retention and zero eviction. A necessity witness is a same-code physical source that violates the selected record-supported guarantee before record constraints are imposed. Each panel includes all 48 fixed queries at every marked budget. Lines connect evaluated thresholds only.

E.4 LONG-PREFIX ACQUISITION AND SERIALIZED STORAGE

The larger-prefix test uses the archived SmolLM2-1.7B-Instruct checkpoint. Four forwards use the same four context families, each repeated to 2064 tokens. Layer 23 and head 0 are fixed in advance. FP32 eager attention with TF32 disabled supplies actual head outputs captured immediately before the output projection. Only its first four queries after the 2048-token prefix supply evidence. Later query indices 4, 8, 12 and 15, output coordinates 0 and 1, and intact or top-1024-old-position retention give 16 tasks per context and precision. The 16 real query vectors have rank 16. Reference Keys use binary16 and Values use INT8 or INT4; the 16 live positions remain exact. Every source witness

passes exact quantizer round trips, and each record has vector error allowance 10^{-3}. Figure 3A shows the full task distributions. Strict physical gains occur in 55/64 FP16/INT8 tasks and all 64 FP16/INT4 tasks. All native-record error bounds are below $3.85 \cdot 10^{-6}$, within the declared 10^{-3} allowance. All 128 independently checked intervals contain the target attention value. Coefficients use the fast one-ray mode. Across the eight context–precision conditions, median proposal times are 0.025–0.034 seconds. Median paired record-conditioned and codec-only outward checks take 0.112–0.145 seconds, after 0.124–0.217 seconds of reusable interval-record setup. Native-record errors are bounded from the captured FP32 values. Acquisition uses four model forwards without masked replays or training. The repeated contexts test long-prefix geometry on fixed inputs.

The serialized payload packs two signed INT4 Value codes per byte. It retains FP16 Keys, one signed power-of-two exponent per Value row, and four FP32 query–output pairs with format metadata. Decoding reproduces the tested reconstruction bit for bit. The payload omits the original prefix. Figure 3C compares the four payloads with their original prefixes, using ZIP DEFLATE for both. The uncompressed original prefix has 1048576 bytes. Both comparisons exclude the same model weights and exact live suffix. Original arrays used for validation are stored separately.

E.5 Matched-byte comparison and record ablations

The input is the same four short-prefix caches and all 48 future queries. The targets are the complete 64-dimensional intact outputs. All schemes keep FP16 Keys and the exact live suffix. Values use a signed integer code with a power-of-two scale per old row. The comparison changes the Value precision or the saved record bank. Each scheme uses its own reconstruction and coordinate cells. The physical prior therefore changes with the storage allocation.

The binary format has a 64-byte header, 4096 bytes of FP16 Keys, 64 bytes of signed row exponents, and packed Value codes. One stored record contains a two-byte causal index and two FP32 vectors of width 64. Its cost is $2 + 2 \cdot 64 \cdot 4 = 514$ bytes. Thus

$$B(b, m) = 64 + 4096 + 64 + 32 \cdot 64\, b/8 + 514m, \qquad B(4, 4) = 7304, \quad B(12, 0) = 7296.$$

Every binary payload is serialized and its integer codes are decoded back exactly. The two baselines lie within the same 7304-byte cap. Neither size uses lossless compression. Model weights and live rows are excluded equally.

For each source and scheme, the evaluator checks all 64 target-coordinate intervals. The certificate tuples are saved separately. Candidate masks retain 0, 8, 16, 24 or 32 old rows in recent or score order. Their executed FP32 outputs are compared with the original intact interval. For every coordinate j and every compatible cache,

$$|\widetilde{y}_{S,j} - y_j(X)| \le \max\{|\widetilde{y}_{S,j} - L_j|, |\widetilde{y}_{S,j} - U_j|\} =: b_j,$$
$$\|\widetilde{y}_S - y(X)\|_2^2 = \sum_j |\widetilde{y}_{S,j} - y_j(X)|^2 \le \sum_j b_j^2.$$

This proves the complete-head decision test in the main text. An unresolved query contributes zero to the total certified-eviction statistic.

Figure 2A shows the complete acceptance curves from the saved candidate bounds. All five preset budgets are included in this range. At $\tau = 0.5$, total certified evictions are 160, 256, 312, 304 and 336 rows for INT4 with zero, first one, first two, last two and all four records. INT12 without records certifies 896 evictions. Figure 2B divides these totals by $48 \cdot 32$. Panel C distinguishes unresolved queries from accepted zero-eviction decisions. INT12 certifies more evictions on these 48 queries. The earlier witnesses measure the effect of records at fixed INT4 precision; this comparison reallocates bytes between records and Value precision.

The first records are at positions 32–35. The last-two variant saves only positions 34–35. Every target is at positions 36–47, so no selected record uses a future target. The one-ray proposal is recomputed for each bank. More evidence can only shrink the exact compatible set. Its numerically proposed interval need not improve monotonically because the coefficient optimization is approximate.

Original intact head norms range from about 5.001 to 11.582, with median 6.993. The budget 0.5 is about 4.3–10.0% of this norm, with median 7.15%. The budget controls attention-output error. Downstream generation quality is not evaluated in this test.

Evaluating all six schemes takes 60.34 seconds on one CPU thread. Median 64-coordinate certificate times were 0.070 seconds without records, 0.190 with one, 0.257 with the first two, 0.260 with the last two, and 0.395 with four. The INT12 codec-only median was 0.070 seconds. These times include the target coefficient proposals and outward checks. Reusable record setup and candidate execution are separate.