FINITE-PRECISION STREAMING ATTENTION: SPACE BOUNDS AND THE ROLE OF RANDOMNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

We characterize the memory required for streaming attention on a prescribed finite input grid, charging state and arithmetic in bits under the error scale $\varepsilon \|\mathrm{softmax}(Kq)\|_2 \|V\|_F$. The lower bounds encode multi-bit symbols in binary occurrence frequencies: one terminal query recovers a selected symbol while respecting the input grid, stream length, and original token energy. Let $R = \ln T$ and let $M_{d,b}$ count unit-ball keys on $2^{-b}\mathbb{Z}^d$. For scalar values and $\varepsilon = T^{-\beta}$ with fixed $\beta > 0$, optimal space is $\Theta(R \min\{M_{d,b}, (R/\ln R)^d\})$ at fixed d and $b = O(\log T)$, and $\Theta(M_{d,b}R)$ at fixed b and polylogarithmic d. The arbitrary-state lower bounds match deterministic histograms or rounded moments, including under output feedback. For fixed d, b, B_0 and failure probability, uniformly over $T^{-B_0} \le \varepsilon \le .1$, randomized space is instead $\Theta(\ln(1/\varepsilon) + \ln\ln T)$ with or without feedback, while deterministic space is $\Theta(\ln T)$ in the charged-clock model. An all-prefix approximate-counting certificate establishes this separation with finite numerical workspace. The results concern bounded logits, with fixed-temperature and input-perturbation extensions. Finite checks certify decoding and active randomized updates; the measured configuration still favors exact histograms in total encoded state.

1 INTRODUCTION

How much information must a compressed attention state retain? Existing answers depend on input geometry and the required accuracy: autoregressive compression barriers use coordinatewise multiplicative error (Haris and Onak, 2025), whereas streaming attention results obtain precision-dependent bounds under an energy-normalized criterion (Kochetkova et al., 2025; Chen et al., 2026). We study a further distinction: *finite precision in a summary is not the same as a prescribed finite alphabet of inputs*. The latter restricts both the algorithm's inputs and the hard instances used to prove its memory requirements.

Binary frequencies retain high-accuracy information. Our main technical contribution is a finite frequency channel. An inverse response construction first specifies real-valued encoder intermediates, which need not lie on the input grid. We realize them instead through repeated keys with binary values. A single terminal query then recovers any selected multi-bit symbol, with explicit control of the total number of tokens, frequency-rounding error, finite decoder arithmetic, and the *actual binary stream's energy*. The resulting lower bound applies to arbitrary internal states, not just a chosen sketch or feature representation. Low precision per token therefore does not preclude a high information requirement at fine output accuracy.

Matching space laws. Write $R = \ln T$ and let $M_{d,b}$ be the number of legal unit-ball keys on the grid $2^{-b}\mathbb{Z}^d$. At fixed dimension, $b = O(\log T)$, and $\varepsilon = T^{-\beta}$, the optimal bit budget is $\Theta(R \min\{M_{d,b}, (R/\ln R)^d\})$: legal-key cardinality and analytic expansion degree are competing bottlenecks. At fixed grid precision and polylogarithmic dimension, the budget is $\Theta(M_{d,b}R)$. Explicitly conditioned key sets make the frequency channel match deterministic histograms or rounded moments. The contribution is the matching characterization under simultaneous input, length, and numerical constraints, rather than a new histogram or polynomial expansion.

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.

Randomization and feedback have different costs. At a fixed finite alphabet and $T^{-B_0} \le \varepsilon \le .1$, randomized memory is $\Theta(\ln(1/\varepsilon) + \ln\ln T)$, while deterministic memory remains $\Theta(\ln T)$ when no exact clock is supplied for free. Classical approximate counting supplies the randomized saving (Morris, 1978; Csűrös, 2009; Nelson and Yu, 2022); we prove its matching transfer to the original attention error, including finite workspace and an all-prefix guarantee valid under causal feedback. Thus feedback does not increase the matched randomized rate, whereas determinism can. Figure 1 summarizes these separate parameter slices.

Scope. The results isolate how input resolution, output accuracy, and randomness determine memory for bounded-logit attention. Their quantifiers are explicit, not jointly uniform in all parameters. Section 5 gives fixed-temperature and input-perturbation extensions; these do not establish optimality for floating-point KV caches or growing-temperature sharp attention.

2 Model and main results

Write $R = \ln T$, $E = \ln(1/\varepsilon)$, and take $T \ge 16$. For a nonnegative integer b, let

$$\mathcal{K}_{d,b} = B_2^d \cap (2^{-b}\mathbb{Z})^d, \qquad M_{d,b} = |\mathcal{K}_{d,b}|. \tag{1}$$

Keys and queries lie in $\mathcal{K}_{d,b}$. Main statements use *scalar* values $v_i \in [-1,1] \cap 2^{-b}\mathbb{Z}$; the high-accuracy vector extension is Corollary 8. The kernel is $e^{k^\top q}$. The integer b specifies an absolute grid spacing, not a floating-point word length. For each nonempty prefix set

$$\begin{aligned} s_i &= e^{k_i^\top q}, \quad Z = \sum_i s_i, \quad N = \sum_i s_i v_i, \\ U^2 &= \sum_i s_i^2, \quad F^2 = \sum_i v_i^2, \quad A = N/Z, \quad \Gamma = UF/Z. \end{aligned} \tag{2}$$

For every causal policy π, the required guarantee is

$$\mathbb{P}[\text{all actual queries satisfy } |\widehat{A} - A| \le \varepsilon\Gamma] \ge 1 - \delta. \tag{3}$$

There are at most T insertions and T queries, counted separately. Policies may use previous public answers and their own inputs, but not private state or unused coins. The reference is the *same realized KV stream*, not a second generative trajectory. If $F = 0$, the output must be exactly zero; cancellation of N does not imply $F = 0$.

Memory and clock convention. All retained state and temporary work are charged, including any stored parameters, counters, indices, and randomness. Outputs and input history are not readable external memory. The interface supplies (k, v) or q, *not* an uncharged exact event number; the algorithm need not validate the externally promised horizon. A coarse upper horizon $2^{\lceil \log_2 T \rceil}$ can be represented by its exponent. This conventional finite-state distinction is essential for sub-$\log T$ counting space. Output numbers may have any explicitly charged finite mantissa–exponent representation; requiring an $\Omega(\log T)$-bit counter or output register would trivially preclude that regime. Reference implementations that validate exact event counts use more memory than the counting construction.

The finite control is fixed independently of T; parameter-dependent information must be stored and charged. Fresh private random bits are an external randomness source in the theoretical model, while a reproducible pseudorandom backend is included in measured implementation state. Let $S_{\mathrm{obl}}, S_{\mathrm{ad}}, S_{\mathrm{det,ad}}$ denote optimal oblivious randomized, feedback randomized, and feedback deterministic space. Then $S_{\mathrm{obl}} \le S_{\mathrm{ad}} \le S_{\mathrm{det,ad}}$.

Theorem 1 (High accuracy, fixed dimension)**.** *Fix $d \ge 1$ and constants $\beta, C_b > 0$. For $\varepsilon = T^{-\beta}$, $0 \le b \le C_b \log_2 T$, and $0 < \delta \le .01$, all three optimal spaces are, for sufficiently large T,*

$$\Theta_{d,\beta,C_b}\left[R \min\left\{M_{d,b}, (R/\ln R)^d\right\}\right]. \tag{4}$$

The constants are uniform in allowed b, δ. The lower bound uses binary values, nonnegative keys/queries of norm at most $1/2$, and one terminal decoding query.

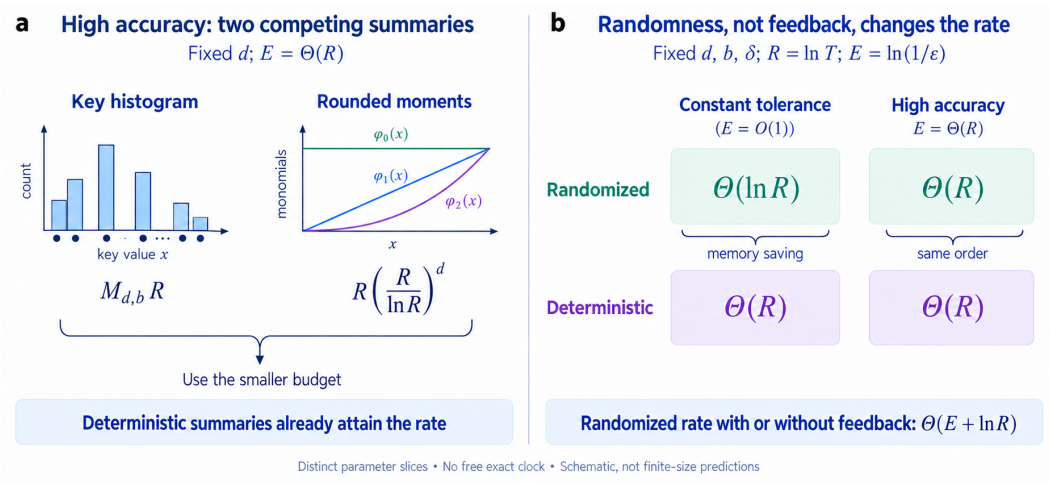


Figure 1: Theoretical map, not empirical curves. (a) At fixed d and $E = \Theta(R)$, legal-key cardinality and analytic degree compete; both rates are attained deterministically. (b) At fixed d, b, randomization saves memory when $E = o(R)$, including constant tolerance, while feedback does not change the randomized rate. The two panels are separate slices, each with its own quantifiers.

For fixed d, $M_{d,b} = \Theta_d(2^{bd})$. The two budgets meet at the asymptotic scale

$$b_{\text{crit}} = \log_2 \ln T - \log_2 \ln \ln T + O_{d,\beta}(1). \tag{5}$$

Equation (5) locates the crossover scale up to an unevaluated $O_{d,\beta}(1)$ term.

Theorem 2 (High accuracy, fixed grid precision)**.** *Fix $b \geq 0$, $\beta > 0$, and $A_0 \geq 0$. For $1 \leq d \leq R^{A_0}$, $\varepsilon = T^{-\beta}$, and $0 < \delta \leq .01$, all three spaces equal*

$$\Theta_{b,\beta,A_0}(M_{d,b}R). \tag{6}$$

Binary values and a single terminal decoding query suffice for the lower bound; keys may have negative coordinates.

Putting $s = 4^b$, every key has support at most s, and $M_{d,b} = 2^s\binom{d}{s} + O_b(d^{s-1})$ as $d \to \infty$. Thus the growing-dimension slice has bounded support at fixed b. The constants of Theorems 1 and 2 are *not* jointly uniform in d, b.

Theorem 3 (Accuracy-sensitive randomness at a fixed alphabet)**.** *Fix integers $d \geq 1$, $b \geq 0$, a constant $B_0 > 0$, and $\delta \in (0, .01]$. Uniformly for $T^{-B_0} \leq \varepsilon \leq .1$, under the charged-clock convention above,*

$$\boxed{S_{\text{obl}} = S_{\text{ad}} = \Theta_{d,b,B_0,\delta}(E + \ln \ln T), \qquad S_{\text{det,ad}} = \Theta_{d,b,B_0}(\ln T).} \tag{7}$$

The randomized lower bound uses only key/query 0 and binary values. Its $\ln \ln T$ component uses a fixed sequence of $O(\log T)$ decoding queries, rather than a single query.

At constant tolerance, randomized memory is $\Theta(\ln \ln T)$ but deterministic memory is $\Theta(\ln T)$. At $E = \Theta(R)$ they coincide in order. This theorem addresses arbitrary accuracy schedules in a *fixed-alphabet slice*, with an explicit dependence on δ. Its all-prefix counting proof, including finite numerical workspace, appears in Appendix E.

3 A FINITE FREQUENCY CHANNEL AND HIGH-ACCURACY MATCHING

The lower-bound construction must satisfy two constraints together: its inputs must lie on the prescribed grid, and its distinguishable responses must survive the allowed error on the original token stream. The following channel enforces both with a finite repetition budget.

Lemma 4 (Scalar frequency inverse channel)**.** *For public legal keys $k_1, \dots, k_n$, let $G_{ji} = e^{k_j^\top k_i}$ and suppose $\lambda_{\min}(G/n) \ge \kappa > 0$. Choose a public power of two B with $\sqrt{n}/\kappa \le B < 2\sqrt{n}/\kappa$. If integers $r, L \ge 1$ satisfy*

$$nr + 2 \le T, \qquad 64eB(\varepsilon + 1/r) \le 2^{-L}, \tag{8}$$

then every algorithm satisfying (3) *needs*

$$S \ge n[(1-\delta)L - h_2(\delta)] \text{ bits}. \tag{9}$$

The input values are binary and one terminal query reads the selected symbol.

Let X_i be independent L-bit symbols in $\{0, 2^{-L}, \dots, 1 - 2^{-L}\}$. With $H = G/n$, define

$$V^* = \tfrac{1}{2}\mathbf{1} + H^{-1}X/(2B) \in [0,1]^n. \tag{10}$$

This vector is an encoder intermediate, not a legal infinite-precision input. Insert each key r times, using a_i ones so that $|a_i/r - V_i^*| \le 1/r$. Add $(0,0)$ and $(0,1)$, charging both tokens. Before frequency rounding the response is

$$A^*(k_j) = \tfrac{1}{2} + \frac{rn}{2BZ'_j}X_j, \qquad Z'_j = r\sum_i G_{ji} + 2. \tag{11}$$

The denominator is public. For the *actual binary stream*,

$$U'^2_j = r\sum_i G_{ji}^2 + 2 \le e^2(nr+2), \qquad F'^2 = \sum_i a_i + 1 \le nr + 2. \tag{12}$$

The scaled algorithm error and frequency-rounding error total at most $8eB(\varepsilon + 1/r)$. Condition (8) leaves room for finite decoder arithmetic and nearest-symbol recovery. Alice transmits only the streaming state. Fano's inequality for each selected symbol and entropy subadditivity prove (9). In particular, the accounting charges every replicated token and uses the actual binary stream's energy (12). Appendix A supplies the complete proof and its vector form.

Conditioning, not uncharged amplification. For fixed d, take $a = 2^{-\lceil \log_2(2\sqrt{d}) \rceil}$ and a power of two p with $p \le a2^b$. The $(p+1)^d$ keys in $\{0, a/p, \dots, a\}^d$ obey

$$\lambda_{\min}(G/n) \ge [(p+1)^{-2}(16/a)^{-2p}(p!)^{-1}]^d. \tag{13}$$

A Lagrange coefficient estimate proves this bound. Taking a sufficiently small constant multiple of $\min\{R/\ln R, 2^b\}$ for p, $r = \lfloor (T-2)/n \rfloor$, and $L = \Theta_\beta(R)$ proves the lower part of Theorem 1; a single zero key handles very coarse grids.

For fixed b, let $\Delta = 2^{-b}$ and $s' = \min(4^b, d)$. Use all $n = 2^{s'}\binom{d}{s'} = \Theta_b(M_{d,b})$ keys with s' coordinates equal to $\pm\Delta$. Monomial features supported on exactly those coordinates, with exponents in $\{1,2\}$, split into tensor blocks of

$$W = \begin{pmatrix} \Delta & \Delta^2/\sqrt{2} \\ -\Delta & \Delta^2/\sqrt{2} \end{pmatrix}, \qquad G \succeq \Delta^{4s'} I, \quad \lambda_{\min}(G/n) \ge \Delta^{4s'}/n. \tag{14}$$

Thus the conditioning loss is only polynomial in R in Theorem 2, and the same frequency reduction applies. Full quantifier and length accounting is in Appendices B and D.

3.1 Deterministic summaries

A histogram stores each legal key's count and exact value sum. A moment summary stores rounded versions of

$$M_{\alpha,0} = \sum_i k_i^\alpha, \quad M_{\alpha,1} = \sum_i v_i k_i^\alpha, \qquad N_h(q) = \sum_{|\alpha| \le h} M_{\alpha,1} q^\alpha/\alpha!. \tag{15}$$

Choose $(h+1)! \ge 128e^2/\varepsilon$ and $D_h = \binom{d+h}{d}$. Taylor truncation respects the original energy directly:

$$|N - N_h| \le \frac{e}{(h+1)!}\sum_i |v_i| \le \frac{e^2}{(h+1)!}UF, \qquad |Z - Z_h| \le \frac{e^2}{(h+1)!}Z. \tag{16}$$

Each multiplication is rounded to P fractional bits, rather than preserving a bh-bit exact denominator. Appendix C gives an explicit $P = O(b + E + R + \ln D_h + \ln(h + 2))$ and the per-query arithmetic error $4D_h t(h + 2)2^{-P}$. This covers even a single nonzero value of size 2^{-b}. The normalization lemma is

$$|\widehat{Z} - Z| \le \eta Z, \quad |\widehat{N} - N| \le \eta UF \quad \Longrightarrow \quad |\widehat{N}/\widehat{Z} - N/Z| \le \frac{2\eta}{1-\eta}\Gamma. \tag{17}$$

The two scalar budgets are $O(M_{d,b}(b+E+R)+d(b+1))$ and $O_d(D_h(b+E+R+\ln D_h+\ln(h+2)))$. Selecting the smaller *before* the stream proves Theorem 1; the histogram proves Theorem 2. All errors are deterministic and pointwise, so output feedback cannot invalidate them.

4 Why accuracy changes the role of randomness

A finite alphabet permits a different representation: approximate the nonnegative counts, not each coordinate of an analytic expansion. This section proves the ideas behind Theorem 3; Appendix E supplies all numerical and state-space details.

A classical counter with an exponential all-prefix certificate. Use the binary floating-point counter of Csűrös (2009). With a power-of-two resolution $B_c = 2^w$, it stores (e, r), $0 \le r < B_c$, and decodes to

$$D(e, r) = (B_c + r)2^e - B_c.$$

The index $B_c e + r$ advances with probability 2^{-e}. A positive decoded jump is 2^e, so the centered process $D_u - u$ is a martingale and its conditional increment variance is at most D_u/B_c. A stopped Freedman argument (Tropp, 2011) gives, for $0 < \eta \le 1/4$,

$$\mathbb{P}[\exists u \le 2^H : |D_u - u| > \eta u] \le 2(H + 1)e^{-B_c\eta^2/6}. \tag{18}$$

Unlike a union over observed queries, this certificate covers *all intrinsic update times*. Use $C = 3M_{d,b}$ counters for key multiplicities and positive/negative integer value totals. Set $H = \lceil \log_2 T \rceil + b$, $\eta = \varepsilon/4$, and

$$B_c \ge 6\eta^{-2} \ln[2C(H + 1)/\delta]$$

rounded upward to a power of two. Assigning private coins to each counter makes a causal input policy only an interleaving of their already-uniform trajectories. The probability that any selected counter state is inaccurate is at most δ.

Signed values and charged arithmetic. Relative error η in the nonnegative statistics gives

$$|\widehat{N}_0 - N| \le \eta \sum_i s_i|v_i| \le \eta UF, \qquad |\widehat{Z}_0 - Z| \le \eta Z. \tag{19}$$

Thus cancellation needs no relative approximation of a signed sum. Dyadic transitions use fewer than two fresh bits in expectation. Direct decoding keeps the estimate in a finite mantissa–exponent representation, without expanding it into a $\Theta(\log T)$-bit integer or evaluating a long recurrence. Appendix E.3 gives a finite mantissa budget and a total-error certificate below $\varepsilon\Gamma$. All-zero sign counters return exact zero. Capping e at $H + 2$ charges worst-case memory without adding failure on the good event. Total state and workspace are $O(E + \ln \ln T)$ for fixed d, b, δ. The construction makes the finite constants explicit.

Two lower bounds and a deterministic obstruction. The frequency lemma with the sole key 0 gives $\Omega(\min\{E, R\})$ bits. A second reduction encodes one of $\Theta(\log T)$ possible initial all-one prefix lengths 16^j; Bob appends zeros and uses a fixed logarithmic sequence of queries to recover j. The original Γ leaves a constant decision gap, yielding $\Omega(\ln \ln T)$ bits. Since $E \le B_0 R$, their maximum is $\Omega_{B_0}(E + \ln \ln T)$.

For a deterministic algorithm with too few states, two short all-one prefixes reach the same state. Repeating the intervening cycle makes their lengths differ by a factor of at least four. An identical zero suffix then produces the same algorithmic output but exact means $1/2$ and at least $4/5$, whose allowed intervals are disjoint for $\varepsilon \le .1$. This forces $\Omega(T)$ states, or $\Omega(\log T)$ bits. It uses no free clock; with an externally supplied time index the pumping argument would not apply. Approximate counting supplies the classical saving; the result here is its matching transfer to the original attention error and causal feedback model.

5 SCOPE, TEMPERATURE, AND INPUT PERTURBATIONS

Fixed scale is not sharp attention. For $e^{\tau k^\top q}$ with any fixed positive rational τ, all three main rates remain valid, with constants depending on τ (Proposition 10). The bounded-logit restriction remains: $p_i/p_j \le e^{2\tau}$ and, at a prefix of length t, $p_i \le e^{2\tau}/t$. Thus fixed scale cannot maintain constant attention mass on one token as t grows; that requires a growing scale. At $\tau = 0$ keys become irrelevant and the key-dependent high-accuracy lower bound generally collapses.

A certified bridge to rounded inputs. Let original logits z_i and quantized logits $\widetilde{z}_i$ obey $\max_i |\widetilde{z}_i - z_i| \le a$, and suppose $\left\|\widetilde{V} - V\right\|_F \le \rho \|V\|_F$. If a compressor is valid on the rounded inputs with tolerance ε_c, then

$$\left\|\widehat{A} - A\right\|_2 \le \left[e^{2a}(1+\rho)(1+\varepsilon_c) - 1\right]\Gamma. \tag{20}$$

Proposition 11 proves this deterministic inequality on the same prefix. It separates input distortion from summary error. Floating-point underflow or arbitrary absolute value rounding need not satisfy a small relative ρ; a rounded key must still satisfy its compressor's domain assumptions. Group scales, outliers, nonuniform codebooks, and a floating exponent are not represented by replacing b with a format's total bit count.

6 CERTIFIED FINITE MECHANISMS AND MEMORY ACCOUNTING

Finite tests check the frequency decoder, nontrivial approximate updates, and full-state accounting. They use rational enclosures and the original-stream squared error test

$$(Z\widehat{A} - N)^2 \le \varepsilon^2 U^2 F^2. \tag{21}$$

Neither a floored Γ nor energy computed after averaging values is used. Proofs supply the universal guarantees; finite checks test the implementation and certificate interfaces.

Binary-frequency decoding. A nonempty finite instance of Lemma 4 uses

$$K = \{0,1\}, \quad b = 0, \quad r = 16384, \quad L = 2, \quad B = 8, \quad \varepsilon = 1/r, \quad T = 32770. \tag{22}$$

Here $G/2 \succeq I/4$ and $\Xi = 64eB(\varepsilon + 1/r)2^L = e/4 < 1$. The check recomputes all 16 messages and both targets using actual binary counts, certifying all 32 decoding cases over the entire permitted scalar output-error interval. The lemma gives about 3.80 bits for this two-key instance, so the check validates decoding rather than constant-level optimality. Appendix G also reports exact algebraic checks and short sequential tests of the deterministic reference implementations.

Active randomized counting. Fixed, query-feedback, and KV-feedback streams each use 2^{21} actual insertions and 32 queries. All 96 queries pass; the largest certified error ratio is 0.018288, or 1.83% of the allowed budget. Each run performs $4{,}194{,}304$ intrinsic counter updates, with 45.4% stochastic holds. These runs therefore exercise genuine approximation rather than an almost-exact counter. They share one reproducibility seed and do not estimate a worst-case failure probability. Appendix H gives the generators, activity counts, and numerical budgets.

Same-trace memory comparison. The dyadic summary and a minimal exact histogram (count plus signed sum per key) replay precisely the three saved active traces. The histogram does not generate a different feedback trajectory. Table 1 uses a common versioned codec that charges every instance field, including the complete pseudorandom backend. Allocation tracing covers construction, insertions, queries, serialization, and restoration. Six fresh-process midpoint continuations reproduce every suffix output and final encoding; replays and continuations are not additional independent samples.

The exact histogram is smaller both in encoded state (122 versus 2,720 bytes) and in measured Python allocations. Even the counter-field comparison favors it (135 versus 199 bits), so the difference is not only RNG overhead. This finite configuration does *not* demonstrate the asymptotic memory saving. Appendix I separates persistent encoding, object sizes, allocation peaks, sampled register liveness, and a reviewed numerical-workspace envelope.

Table 1: Same-trace resource audit at $d = 1, b = 0, T = 2^{21}, \varepsilon = .1$. Ranges are over three policies. The metrics are distinct and not additive. Traced peaks include driver and serialization/restore copies, but exclude preloaded traces and interpreter/native allocations.

Measure	Dyadic counting	Exact histogram
Counter-field capacity (bits)	199	135
Final full state (bytes)	2,720	122
Max. checkpoint state (bytes)	2,721	122
RNG in that encoding (bytes)	2,508	0
Final snapshot objects (bytes)	5,367	1,473
Max. tracked phase peak (bytes)	68,442–68,499	17,120–17,165

7 Related work

Autoregressive barriers and continuous-domain streaming. Haris and Onak (2025) directly study autoregressive KV compression, prove $\Omega(d \min\{T, e^d\})$ memory for their multiplicative guarantee, and analyze low-dimensional coverings in connection with SubGen (Zandieh et al., 2024). Their hard constructions exploit concentrated softmax weights. The guarantees differ: with uniform keys and one nonzero scalar value, $A = 1/T$ while $\Gamma = 1/\sqrt{T}$, so relative error and $\varepsilon\Gamma$ differ by $\sqrt{T}$. Kochetkova et al. (2025) introduce discrepancy-based streaming attention bounds under the latter scale. Chen et al. (2026) refine continuous norm-bounded precision/radius bounds using polynomial heads, coreset tails, and partitioning, and already account for finite-precision aggregated features. We instead constrain the input alphabet itself and obtain matching optimal bit costs on the stated parameter slices. Finite-precision aggregation does not by itself enforce that restriction on lower-bound inputs.

Static coresets, sketches, and response rank. Liberty et al. (2026) give all-query static coreset results *and* general-sketch bit lower bounds. Their absolute-error and query-radius assumptions differ from ours. Sui and Zhang (2026) and Opazo (2026) address normalized response or nonnegative-feature rank; representation-specific rank does not by itself characterize an arbitrary state. Static response packing can prove necessity but does not itself supply a finite-memory update procedure for changing KV inputs.

Numerical models and Gaussian entropy. KIVI quantizes keys per channel and values per token, with grouped quantization and a residual cache (Liu et al., 2024); KVQuant uses nonuniform datatypes and separate outlier handling (Hooper et al., 2024). These quantizers use scales and distributions absent from an absolute dyadic grid. The fine-grid growth $E^{d+1}/(\ln E)^d$ is also the known Gaussian-RKHS entropy order (Kühn, 2011; Steinwart and Fischer, 2021), consistent with $e^{x^\top y} = e^{\|x\|^2/2} e^{\|y\|^2/2} e^{-\|x-y\|^2/2}$. These function-class bounds do not directly provide legal length-T binary streams, state-only decoding under original-stream energy, or rounded online updates. Our frequency channel, conditioning estimates, and summaries supply these steps.

Counting and adaptive robustness. Approximate counters and their optimal space bounds are classical (Morris, 1978; Csűrös, 2009; Nelson and Yu, 2022). Our stopped-Freedman certificate makes the all-prefix constants explicit. Generic robust merge-and-reduce results exist (Braverman et al., 2021); here, high-accuracy correctness is deterministic, while lower-accuracy correctness uses a uniform event in intrinsic counter time.

8 Conclusion

Binary occurrence frequencies turn low-precision tokens into high-accuracy memory lower bounds. Explicit conditioning and rounded summaries match those bounds, identifying when legal-key cardinality or analytic degree is decisive. At a fixed alphabet, approximate counting separates randomization from determinism without an extra feedback cost. These conclusions hold on the three stated bounded-logit parameter slices.

AI USAGE STATEMENT

ChatGPT assisted substantially with problem formulation, theoretical exploration, proof development and checking, literature comparison, synthetic test design, implementation, interpretation, manuscript drafting and revision, and figure design. The theoretical illustration was generated with image_gen. Codex assisted with historical server experiments that are not used as evidence in this manuscript. AI assistance and finite tests do not constitute independent mathematical verification. The authors are responsible for reviewing the proofs, references, code, and final claims.

REPRODUCIBILITY STATEMENT

Appendices A–F give the proofs, numerical conventions, and scope of the three parameter-slice theorems. Appendices G–I document exact/interval checks, deterministic generators, saved outputs and hashes, full-state encoding, and six fresh-process continuation tests. The accompanying archive supplies reference engines, policies, seeds, recorded results, and table generators, including the minimal exact-histogram baseline. Source and environment manifests identify the recorded runs; implementation-dependent resource measurements should not be treated as measurements of a later source revision or a different environment. The finite tests do not replace the proofs or establish latency or finite-memory dominance.

ETHICS STATEMENT

The checks use synthetic finite-grid inputs, not human subjects or personal datasets. The theorems make no safety, fairness, text-quality, or deployment-efficiency claims for language models. The manuscript distinguishes mathematical guarantees, finite observations, parameter restrictions, and different memory measures, and retains the negative finite-memory comparison.

REFERENCES

Vladimir Braverman, Avinatan Hassidim, Yossi Matias, Mariano Schain, Sandeep Silwal, and Samson Zhou. Adversarial robustness of streaming algorithms through importance sampling. In *Advances in Neural Information Processing Systems*, 2021. https://arxiv.org/abs/2106.14952.

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

Miklós Csűrös. Approximate counting with a floating-point counter. *arXiv preprint arXiv:0904.3062*, 2009. https://arxiv.org/abs/0904.3062.

Themistoklis Haris and Krzysztof Onak. Compression barriers in autoregressive transformers. In *Proceedings of the Thirty-Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2757–2785, 2025. https://proceedings.mlr.press/v291/haris25a.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, 2024. https://arxiv.org/abs/2401.18079.

Thomas Kühn. Covering numbers of Gaussian reproducing kernel Hilbert spaces. *Journal of Complexity*, 27:489–499, 2011.

Ekaterina Kochetkova, Kshiteej Sheth, Insu Han, Amir Zandieh, and Michael Kapralov. Streaming attention approximation via discrepancy theory. In *Advances in Neural Information Processing Systems*, 2025. https://arxiv.org/abs/2502.07861. Version 3, March 24, 2026.

Edo Liberty, Alexandr Andoni, and Eldar Kleiner. Nearly optimal attention coresets. *arXiv preprint arXiv:2605.05602*, 2026. https://arxiv.org/abs/2605.05602v1.

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 International Conference on Machine Learning*, 2024. https://arxiv.org/abs/2402.02750.

Robert Morris. Counting large numbers of events in small registers. *Communications of the ACM*, 21(10):840–842, 1978. https://doi.org/10.1145/359619.359627.

Jelani Nelson and Huacheng Yu. Optimal bounds for approximate counting. In *Proceedings of the ACM Symposium on Principles of Database Systems*, 2022. https://doi.org/10.1145/3517804.3526225. https://arxiv.org/abs/2010.02116.

Vicente Opazo. Three tokens force exponential feature rank in nonnegative kernel attention. *arXiv preprint arXiv:2608.11427*, 2026. https://arxiv.org/abs/2608.11427v1.

Ingo Steinwart and Simon Fischer. A closer look at covering number bounds for Gaussian kernels. *Journal of Complexity*, 62:101513, 2021. https://doi.org/10.1016/j.jco.2020.101513.

Yuhe Sui and Jianing Zhang. The approximation rank of softmax attention: Sharp geometric laws and robust interaction dimension. *arXiv preprint arXiv:2608.28150*, 2026. https://arxiv.org/abs/2608.28150v1.

Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011. https://arxiv.org/abs/1101.3039.

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

A PROOF OF THE FREQUENCY INVERSE CHANNEL

We prove the following stronger vector version; Lemma 4 is its scalar specialization.

Lemma 5 (Vector frequency channel)**.** *Under the public-key and conditioning assumptions of Lemma 4, suppose*

$$nr + 2 \le T, \qquad 64eB\sqrt{d_v}(\varepsilon + 1/r) \le 2^{-L}. \tag{23}$$

For values in $\{0,1\}^{d_v}$ *and vector error* $\left\|\widehat{A} - A\right\|_2 \le \varepsilon\Gamma$*,*

$$S \ge nd_v[(1-\delta)L - h_2(\delta)].$$

Only one selected terminal query is needed.

A public object in the communication reduction is not free memory for the streaming algorithm: it is known to Alice and Bob independently of the encoded message. Alice transmits only the algorithm's state. The encoder and decoder need not be efficient for a communication lower bound, but all actual stream inputs must belong to the declared finite grid.

A.1 ENCODING AND LEGAL BINARY INPUTS

Set $H = G/n$, and let the entries of $X \in \mathbb{R}^{n\times d_v}$ be independent and uniform on $\{0, h, \dots, 1-h\}$, where $h = 2^{-L}$. For each column,

$$\left\|H^{-1}X_{:c}\right\|_2 \le \kappa^{-1}\left\|X_{:c}\right\|_2 \le \sqrt{n}/\kappa \le B.$$

Consequently $V^* = \frac{1}{2}\mathbf{1} + H^{-1}X/(2B)$ has all entries in $[0,1]$. Its entries need not be dyadic, and we do not insert them.

Compute each entry to absolute error at most $1/(8r)$, clip the result to $[0,1]$, and round its product with r to a nearest integer $a_{ic} \in \{0,\dots,r\}$. This guarantees

$$\left|\frac{a_{ic}}{r} - V^*_{ic}\right| \le \frac{1}{r}. \tag{24}$$

A positive, known lower bound on $\lambda_{\min}(H)$ makes the inverse computable to the prescribed accuracy. Approximating before rounding avoids assuming a procedure that exactly decides whether a transcendental entry lies on a half-integer boundary.

For each key k_i, Alice inserts r tokens. At its ℓth occurrence, value coordinate c is $\mathbf{1}\{\ell \le a_{ic}\}$. Thus every value is binary, the number of tokens is exactly nr, and the entire key sequence is public and independent of X. Different coordinates need not place their ones on different tokens. The legal vector alphabet is $\{0,1\}^{d_v}$, so these simultaneous coordinatewise choices are permitted.

Alice sends the streaming state to Bob. To decode a selected coordinate (j,c), Bob inserts the two public pairs $(0,\mathbf{0})$ and $(0,\mathbf{1})$ and asks $q = k_j$ at the second insertion. This uses $nr+2 \le T$ insertions. If every insertion requires a query, all other queries can be fixed to 0; their answers are ignored and the total query count still does not exceed T. In the separate-query interface, only the final query is needed.

A.2 SIGNAL, ACTUAL ENERGY, AND DECODING ERROR

Before frequency rounding,

$$rGV^* = \frac{r}{2}G\mathbf{1} + \frac{rn}{2B}X.$$

The two public tokens add 2 to the denominator and $\mathbf{1}$ to the numerator. For the selected query let $Z'_j = r\sum_i G_{ji} + 2$. Its ideal answer is the coordinatewise vector version of (11).

Let $\overline{V}_{ic} = a_{ic}/r$ and let A'_j denote the exact attention on the actual binary stream including the two public tokens. Bob forms

$$\widehat{X}_{j,:} = \frac{2BZ'_j}{rn}\left(\widehat{A}'_j - \tfrac{1}{2}\mathbf{1}\right).$$

The frequency perturbation in this quantity is

$$\frac{2B}{n}\left\|[G(\overline{V}-V^*)]_{j,:}\right\|_2 \le \frac{2B}{n}\sum_i G_{ji}\left\|\overline{V}_{i,:}-V^*_{i,:}\right\|_2 \tag{25}$$

$$\le 2eB\sqrt{d_v}/r. \tag{26}$$

Here $G_{ji} \le e$ follows from the unit-ball constraint.

Importantly, the original error scale uses the energy of the binary tokens, not of their per-key means. For these tokens,

$$U_j'^2 = r\sum_i G_{ji}^2 + 2 \le e^2(nr+2), \qquad F'^2 = \sum_{i,c} a_{ic} + d_v \le (nr+2)d_v.$$

On the event that the algorithm's answer satisfies (3), the scaled algorithmic error is therefore at most

$$\frac{2BZ_j'}{rn}\varepsilon\frac{U_j'F'}{Z_j'} \le 2eB\varepsilon\sqrt{d_v}\frac{nr+2}{nr} \le 6eB\varepsilon\sqrt{d_v}. \tag{27}$$

Combining (26) and (27) gives

$$\left\|\widehat{X}_{j,:}-X_{j,:}\right\|_2 \le 8eB\sqrt{d_v}(\varepsilon+1/r) \le h/8.$$

Bob can compute his public rescaling to another $h/8$ absolute error in the selected coordinate. Its total error is then less than $h/2$, so nearest-grid decoding recovers X_{jc}. All intermediate public constants are finite and computable; their precision cost is absorbed by the decoding computation, not secretly supplied to the streaming algorithm.

A.3 From one-symbol recovery to state information

For each possible selected symbol there is a legal terminal-query decoder with error probability at most δ. There is no need to execute all these decoders in one run. Let W be the state message. A randomized decoder can be replaced, for the purpose of the entropy bound, by the Bayes decoder based on W; its error is no larger. Fano's inequality then gives

$$H(X_{ic} \mid W) \le h_2(\delta) + \delta\log_2(2^L-1) \le h_2(\delta) + \delta L.$$

The nd_v symbols are independent, so conditional entropy subadditivity yields

$$S \ge H(W) \ge I(X;W) \tag{28}$$

$$\ge \sum_{i,c}\big[H(X_{ic}) - H(X_{ic} \mid W)\big] \tag{29}$$

$$\ge nd_v\big[(1-\delta)L - h_2(\delta)\big]. \tag{30}$$

One may regard a worst-case S-bit memory as an S-bit padded state. Its random generator's persistent information is part of that state. Bob's fresh randomness does not carry information about X. The same argument permits private or public algorithmic coins, with independent public coins conditioned upon throughout. This proves Lemma 5, and hence Lemma 4.

B Public key sets with explicit conditioning

The following two constructions supply all conditioning estimates used in the main theorems. In particular, the lower bounds do not rely on an unverified floating-point smallest eigenvalue.

B.1 A Cartesian dyadic grid

Lemma 6. *Fix $d \ge 1$, let $a = 2^{-\lceil\log_2(2\sqrt{d})\rceil}$ and $C_a = 16/a$, and let $p \ge 1$ be a power of two. For the $n = (p+1)^d$ keys in $\{0, a/p, \dots, a\}^d$, the matrix $G_{ij} = e^{k_i^\top k_j}$ satisfies*

$$\lambda_{\min}(G/n) \ge \kappa_{d,p} := \big[(p+1)^{-2}C_a^{-2p}(p!)^{-1}\big]^d.$$

Every key is nonnegative and has norm at most $1/2$. It is on the b-bit grid whenever $p \le a2^b$.

Proof. First consider one dimension, with $x_i = ai/p$ for $i = 0, \dots, p$. Write $V_{i\ell} = x_i^\ell$ for $0 \le \ell \le p$. The exponential feature expansion is a sum of positive semidefinite outer products, hence

$$G_1 = (e^{x_i x_j})_{i,j=0}^{p} \succeq V \operatorname{diag}(1/0!, \dots, 1/p!) V^\top.$$

The ith column of V^{-1} is the coefficient vector of the Lagrange polynomial

$$\ell_i(x) = \prod_{j \ne i} \frac{x - x_j}{x_i - x_j}.$$

The absolute denominator is $(a/p)^p i!(p-i)!$, and the sum of absolute numerator coefficients is at most $(1+a)^p$. Using $a \le 1$, $\binom{p}{i} \le 2^p$, and $p! \ge (p/e)^p$ gives

$$\|\text{coeff } \ell_i\|_1 \le \frac{(1+a)^p (p/a)^p \binom{p}{i}}{p!} \le (4e/a)^p \le C_a^p.$$

Consequently $\left\|V^{-1}\right\|_2 \le \left\|V^{-1}\right\|_F \le \sqrt{p+1} C_a^p$, and

$$\lambda_{\min}(G_1/(p+1)) \ge \frac{1}{(p+1)^2 C_a^{2p} p!}.$$

The d-dimensional exponential Gram matrix is $G_1^{\otimes d}$, which proves the bound. Finally, $a\sqrt{d} \le 1/2$, and a/p is an integer multiple of 2^{-b} under the stated power-of-two and grid conditions. □

For $p = 0$ we use the single key 0, with $n = 1$ and $\kappa = 1$. The reciprocal of the displayed bound is an explicit integer. Thus one can choose a conservative power-of-two B entirely from integer arithmetic.

B.2 Counting a fixed-precision alphabet

Let $\Delta = 2^{-b}$ and $s = 4^b$. Write a grid key as Δz for $z \in \mathbb{Z}^d$; legality means $\sum_j z_j^2 \le s$. Every nonzero coordinate contributes at least one, so the support has size at most s. If

$$C_{b,\ell} = \#\{(z_1, \dots, z_\ell) \in (\mathbb{Z} \setminus \{0\})^\ell : \textstyle\sum_j z_j^2 \le s\},$$

with $C_{b,0} = 1$, then exactly

$$M_{d,b} = \sum_{\ell=0}^{\min(d,s)} \binom{d}{\ell} C_{b,\ell}. \tag{31}$$

When $\ell = s$, all nonzero coordinates must be ± 1, so $C_{b,s} = 2^s$. For fixed b this is a polynomial in d of degree s once $d \ge s$, with leading term $2^s \binom{d}{s}$. This proves the alphabet asymptotic in the main text. It also gives an exact preflight count without listing the alphabet: the constants $C_{b,\ell}$ can be obtained by dynamic programming over the remaining integer squared-norm budget.

B.3 A sparse sign grid

Lemma 7. *Let $s' = \min(4^b, d)$ and use all $n = 2^{s'} \binom{d}{s'}$ keys with exactly s' nonzero coordinates, each equal to $\pm\Delta$. They belong to $\mathcal{K}_{d,b}$ and satisfy*

$$G \succeq \Delta^{4s'} I_n, \qquad \lambda_{\min}(G/n) \ge \Delta^{4s'}/n.$$

For fixed b, $n = \Theta_b(M_{d,b})$ uniformly over $d \ge 1$.

Proof. The squared norm is $s'\Delta^2 \le 1$. Expand the kernel as

$$e^{k^\top q} = \sum_{\alpha \in \mathbb{Z}^d_{\ge 0}} \frac{k^\alpha q^\alpha}{\alpha!}.$$

Retain only multi-indices whose support is an s'-element set and whose positive entries are all 1 or 2. Since each key has support exactly s', its retained features are nonzero only in the block for that same support. Each such feature block is $W^{\otimes s'}$, where

$$W = \begin{pmatrix} \Delta & \Delta^2/\sqrt{2} \\ -\Delta & \Delta^2/\sqrt{2} \end{pmatrix}.$$

The eigenvalues of $WW^\top$ are $2\Delta^2$ and Δ^4. Since $\Delta \le 1$, each tensor block has smallest squared singular value $\Delta^{4s'}$. All omitted features contribute positive semidefinite terms, establishing the Gram bound. For $d \ge s$, (31) shows that the ratio of $M_{d,b}$ to $2^s\binom{d}{s}$ is bounded above and below by positive constants depending only on b. There are only finitely many remaining $1 \le d < s$; adjusting these constants covers them as well. □

C Deterministic summaries and their bit budgets

We prove the upper bounds in the same numerical model as the lower bounds. The discussion keeps input precision b separate from arithmetic precision P. A nonempty prefix always satisfies

$$e^{-1}\sqrt{t} \le U \le e\sqrt{t}, \quad e^{-1}t \le Z \le et, \quad \|N\|_2 \le UF. \tag{32}$$

Furthermore, $U/Z = \|p(q)\|_2 \ge 1/\sqrt{t}$. If any value is nonzero, the input grid implies $F \ge 2^{-b}$, and hence

$$\Gamma \ge 2^{-b}/\sqrt{T}. \tag{33}$$

These bounds are analytic certificates; the algorithm need not query the original stream to evaluate F.

C.1 Normalization

Suppose $0 \le \eta < 1$, $|\widehat{Z} - Z| \le \eta Z$, and $\left\|\widehat{N} - N\right\|_2 \le \eta UF$. Then $\widehat{Z} \ge (1-\eta)Z > 0$, and

$$\left\|\widehat{N}/\widehat{Z} - N/Z\right\|_2 \le \frac{\left\|\widehat{N} - N\right\|_2}{\widehat{Z}} + \frac{\|N\|_2\,|Z - \widehat{Z}|}{Z\widehat{Z}} \tag{34}$$

$$\le \frac{2\eta}{1-\eta}\frac{UF}{Z}. \tag{35}$$

This proves (17) without assuming any sign condition on values. In particular, division is safe once the denominator certificate is established.

C.2 Exact key aggregation, including workspace

The vector space budget, including temporary arithmetic, is

$$O\big((d_v+1)M_{d,b}[b + E + \ln T + \ln(d_v+1)] + d(b+1)\big). \tag{36}$$

For each key in a fixed canonical enumeration of $\mathcal{K}_{d,b}$, store its count c_k and integer numerators of its coordinate sums w_k. They need $O(b + \log T)$ bits per field. Enumerating keys by a bounded-depth integer prefix search uses $O(d(b+1))$ bits; it need not save the key coordinate vector for every table entry. A simple implementation may search this enumeration to find an insertion's table position. Time efficiency is not required for the space claim, although the procedure is polynomial in the relevant alphabet representation size.

For a query, compute dyadic approximations $\widetilde{s}_k$ with $|\widetilde{s}_k - e^{k^\top q}| \le \zeta$, where $\zeta \le \varepsilon/(128e)$. Form $\widehat{Z} = \sum_k c_k\widetilde{s}_k$ and $\widehat{N} = \sum_k w_k\widetilde{s}_k$ using exact fixed-point accumulation at the chosen product scale. Since

$$\sum_k \|w_k\|_2 \le \sum_{i\le t}\|v_i\|_2 \le \sqrt{t}F,$$

one has

$$\left\|\widehat{N} - N\right\|_2 \le \zeta\sqrt{t}F \le e\zeta UF, \qquad |\widehat{Z} - Z| \le \zeta t \le e\zeta Z.$$

This bound depends on original token energy and remains valid when some aggregated sums cancel. A Taylor approximation with an explicit factorial remainder computes the exponential to the requested absolute accuracy using a number of guard bits and operations polynomial in $\ln(1/\zeta)$ and the input size. For example, on $[-1,1]$ the error after degree J is at most $3/(J+1)!$; accumulated evaluation rounding can be assigned an additional half of the ζ budget.

Use (17), and round the final coordinate outputs finely enough that their total Euclidean rounding error is below $\varepsilon 2^{-b}/(8\sqrt{T})$. This requires $O(b + \ln(1/\varepsilon) + \ln T + \ln(d_v+1))$ fractional bits. All-zero

values produce an exactly zero numerator. Arithmetic registers, loop counters, the current input/query, and enumeration stack are covered by (36). The algorithm does not need to retain either the original tokens or past answers.

C.3 Polynomial truncation

For a multi-index α, write $k^\alpha = \prod_j k_j^{\alpha_j}$ and $\alpha! = \prod_j \alpha_j!$. The multinomial formula gives

$$\sum_{\ell=0}^{h} \frac{(k^\top q)^\ell}{\ell!} = \sum_{|\alpha|\le h} \frac{k^\alpha q^\alpha}{\alpha!}.$$

There are $D_h = \binom{d+h}{d}$ such indices. The Taylor remainder at any $|k^\top q| \le 1$ is at most $e/(h+1)!$. Thus the unrounded moment representation satisfies

$$\|N - N_h\|_2 \le \frac{e\sqrt{t}F}{(h+1)!} \le \frac{e^2}{(h+1)!}UF, \qquad |Z - Z_h| \le \frac{e^2}{(h+1)!}Z.$$

With the choice in the main text these are at most $\varepsilon/128$ of their respective scales. This argument uses the actual F directly, so no value-amplitude partition is required for this upper bound.

C.4 Rounding at every multiplication

A sufficient working precision is

$$P = b + \left\lceil \log_2(1/\varepsilon) + \tfrac{1}{2}\log_2(Td_v) + \log_2(2^{14}D_h(h+2)) \right\rceil + 4. \tag{37}$$

Let $u = 2^{-P}$ with P given by (37); in particular $P \ge b$. Each real fixed-point quantity is an integer times u. Round the result of each multiplication to this grid, and accumulate these grid values as exact integers with $O(\log T)$ integer bits. An elementary telescoping bound for a product of factors of magnitude at most one gives an absolute product error at most mu after m multiplications. Inductively, a previous error is multiplied by a factor of magnitude at most one, and each new rounding adds at most u.

For k^α with one optional value factor, $m \le h+1$, so every stored moment differs from its exact target by at most $t(h+1)u$. To evaluate a query, form each coefficient $q^\alpha/\alpha!$ at precision P. Sequential products followed by the division by the positive integer $\alpha!$ give coefficient error at most $(h+2)u$. Its exact magnitude is at most one, and the stored moment magnitude is at most $2t$ for the displayed precision. Combining moment error, coefficient error, and the final product rounding, then summing the D_h terms, bounds arithmetic error per numerator coordinate or denominator by

$$E_{\mathrm{arith}} \le 4D_h t(h+2)u. \tag{38}$$

The Euclidean numerator error is at most $\sqrt{d_v}E_{\mathrm{arith}}$.

Our choice of P implies, conservatively,

$$u \le \frac{\varepsilon 2^{-b}}{2^{18}D_h(h+2)\sqrt{Td_v}}.$$

For $F > 0$, apply (32) and $F \ge 2^{-b}$. Since $t \le \sqrt{Tt}$,

$$\frac{\sqrt{d_v}E_{\mathrm{arith}}}{UF} \le 4eD_h(h+2)2^b\sqrt{td_v}\,u \le \frac{4e\varepsilon}{2^{18}} < \frac{\varepsilon}{128}.$$

Similarly $E_{\mathrm{arith}}/Z \le 4eD_h(h+2)u < \varepsilon/128$. Combining arithmetic and truncation gives common budget $\eta = \varepsilon/64$. For $0 < \varepsilon \le 1$, normalization in (17) then uses less than $\varepsilon/16$ of Γ. The final division rounded to the P-bit grid adds at most $\sqrt{d_v}u$, and (33) shows that this fits well inside the remaining allowance. For $F = 0$, every numerator update is exactly zero, so the output is exactly zero without dividing a nonzero rounding residue.

The finite parameter ε can be supplied as a computable tolerance: first choose a dyadic $\varepsilon_0 \in [\varepsilon/2, \varepsilon]$ and use conservative integer comparisons to choose h and P for ε_0. This avoids exact equality tests involving e^2 or an arbitrary real tolerance and changes only absolute constants. The supplied checks use rational tolerances.

Algorithm 1 Rounded-moment summary (all numeric state is fixed-point)

Require: Public $d, d_v, b, T, \varepsilon$; choose h, P as in Appendix C

Initialize $\widehat{M}_{\alpha,c} = 0$ for $|\alpha| \le h$, $c = 0, \dots, d_v$
procedure INSERT(k, v)
  Validate the integer-grid input and insertion budget
  for each α with $|\alpha| \le h$ **do**
    $z \leftarrow$ sequential product k^α, rounding each multiplication to P bits
    $\widehat{M}_{\alpha,0} \leftarrow \widehat{M}_{\alpha,0} + z$
    for $c = 1, \dots, d_v$ **do**
      $\widehat{M}_{\alpha,c} \leftarrow \widehat{M}_{\alpha,c} + \mathrm{round}_P(v_c z)$
    end for
  end for
end procedure
procedure QUERY(q)
  Validate the query and query budget; initialize $y_0, \dots, y_{d_v} \leftarrow 0$
  for each α with $|\alpha| \le h$ **do**
    $a \leftarrow \mathrm{round}_P(q^\alpha/\alpha!)$, using rounded sequential products
    for $c = 0, \dots, d_v$ **do**
      $y_c \leftarrow y_c + \mathrm{round}_P(a\widehat{M}_{\alpha,c})$
    end for
  end for
  return $\mathrm{round}_P((y_1, \dots, y_{d_v})/y_0)$
end procedure

C.5 MEMORY, RUNNING TIME, AND PATHWISE CORRECTNESS

For fixed d, the vector moment budget is

$$O\big((d_v+1)D_h[b + E + \ln T + \ln D_h + \ln(h+2) + \ln(d_v+1)]\big). \tag{39}$$

Each of the $(d_v+1)D_h$ moment entries uses $P + O(\log T)$ bits. Multi-indices may be generated sequentially; saving all exact high-degree dyadic products is unnecessary. A factorial uses $O(h\log(h+1)) = O(\ln(1/\varepsilon))$ bits at the chosen truncation order. Input registers, coefficients, counters, and temporary products fit within (39) for fixed d. Each insertion and query requires a polynomial number of bounded-precision integer operations in D_h, h, d_v and the input/work precision. This is an executability result, not a claim of low practical latency.

Every inequality above is deterministic for every sequence of legal inputs and every legal query. The proof never conditions on a randomly realized final trajectory, and never requires a future input to be independent of past approximation error. Hence each summary satisfies the entire feedback guarantee with probability one. Choosing one of the two summaries before the stream, according to public upper bounds, preserves this property. It does not imply that arbitrary midstream representation changes are free.

D COMPLETING THE MATCHING ASYMPTOTIC BOUNDS

We establish the vector extension as well as the scalar main theorems.

Corollary 8 (High-accuracy vector values)**.** *For $1 \le d_v \le R^V$ with fixed $V \ge 0$ and coordinatewise legal values in $[-1,1]^{d_v}$, multiply the rates in Theorems 1 and 2 by d_v. Their constants may additionally depend on V, and the error is $\left\|\widehat{A} - A\right\|_2 \le \varepsilon \left\|\mathrm{softmax}(Kq)\right\|_2 \left\|V\right\|_F$. All stated binary-value and key restrictions persist. No vector extension of Theorem 3 with growing d_v is claimed here.*

We spell out the quantifiers and sequence-length restriction; setting $d_v = 1$ proves the scalar statements.

D.1 Proof of Theorem 1

Fix d, β, C_b as in Theorem 1 and V as in Corollary 8 and put

$$E_0 = \min\{\beta, 1\}R, \qquad a = 2^{-\lceil \log_2(2\sqrt{d})\rceil}, \qquad C_a = 16/a.$$

Choose p as the largest power of two not exceeding

$$\min\left\{\frac{E_0}{64d\ln(E_0+e)},\ a2^b\right\}; \tag{40}$$

if this upper bound is below one, use $p = 0$. For $p \geq 1$, Lemma 6 gives $n = (p+1)^d$ legal keys and the explicit $\kappa_{d,p}$. For $p = 0$, take $n = 1$ and $\kappa = 1$. Choose B in the interval of Lemma 4, and set

$$r = \left\lfloor \frac{T-2}{n} \right\rfloor, \qquad L = \left\lfloor \frac{E_0}{4\ln 2} \right\rfloor.$$

For sufficiently large T, both are positive and $nr + 2 \leq T$.

The conditioning estimate gives

$$\ln(1/\kappa_{d,p}) = d\big[2\ln(p+1) + 2p\ln C_a + \ln(p!)\big] \leq E_0/8$$

for all sufficiently large T, uniformly over the allowed p. To see the uniformity, the bracket is increasing in p, while the analytic cap in (40) is $E_0/(64d\ln E_0)$ to first order; its leading contribution $dp\ln p$ is at most $E_0/64 + o(E_0)$. Smaller or zero p only improves this bound. Therefore

$$\ln B \leq E_0/8 + O_d(\ln R), \qquad \ln n = O_d(\ln R), \qquad 1/r \leq 2n/T$$

for sufficiently large T. Since $\varepsilon = T^{-\beta} \leq e^{-E_0}$ and $T^{-1} \leq e^{-E_0}$, and since $d_v \leq R^V$, we obtain

$$\ln\big[64eB\sqrt{d_v}(\varepsilon + 1/r)\big] \leq -7E_0/8 + O_{d,V}(\ln R). \tag{41}$$

In contrast, $\ln(2^{-L}) = -E_0/4 + O(1)$. The sufficient condition of Lemma 4 follows uniformly over b, d_v in the theorem. Since $\delta \leq 0.01$ and $L = \Theta_\beta(R)$, its entropy factor is $\Omega_\beta(R)$, uniformly over δ.

For completeness, grid-point counting in fixed dimension gives

$$c_d 2^{bd} \leq M_{d,b} \leq C_d 2^{bd} \qquad (b \geq 0). \tag{42}$$

The upper bound follows from the enclosing cube. A fixed smaller cube in the unit ball gives the lower bound once the grid is fine enough; the finitely many coarser values of b are covered by changing c_d. A largest power of two is within a factor two of its positive cap, so (40), including $p = 0$, gives

$$n = \Theta_{d,\beta}\left(\min\{M_{d,b}, (R/\ln R)^d\}\right).$$

When the cap is below one, $2^b < 1/a$ for all sufficiently large T, so $M_{d,b}$ is bounded by a constant depending only on d; the one-key construction supplies the matching constant number of information positions. The lower bound in (4) follows.

For the upper bound, (36) is $O_{d,\beta,C_b,V}(d_v M_{d,b} R)$. Stirling's estimate applied to the least factorial truncation order gives $h = \Theta_\beta(R/\ln R)$ and $D_h = \Theta_d((R/\ln R)^d)$. The bracket in (39) is $O_{d,\beta,C_b,V}(R)$, yielding the other branch. Public parameter selection takes the smaller. There is no need to enumerate an enormous alphabet just to select a method: for fixed d, the cube bound $(2^{b+1}+1)^d$ approximates $M_{d,b}$ within a constant depending only on d. Comparing this bound with the moment budget changes only that constant. Both algorithms are deterministic, so the lower bound for S_{obl} and upper bound for $S_{\mathrm{det,ad}}$ prove the result for all three spaces.

D.2 Proof of Theorem 2

Fix b, β, A, V. Let $\Delta = 2^{-b}$, $s' = \min(4^b, d)$, and use Lemma 7, with $n = 2^{s'}\binom{d}{s'}$. Its conditioning certificate gives

$$B < 2n^{3/2}\Delta^{-4s'}, \qquad n = \Theta_b(M_{d,b}).$$

As b is fixed and $d \leq R^A$, both n and B are bounded by fixed powers of R, with constants and exponents allowed to depend on b, A. In particular, $n < T - 2$ eventually, regardless of how

large these fixed powers are. Choose r, L as in the preceding proof, with $E_0 = \min(\beta, 1)R$. Now $\ln B, \ln n, \ln d_v = O_{b,A,V}(\ln R)$, so

$$\ln\bigl[64eB\sqrt{d_v}(\varepsilon + 1/r)\bigr] \le -E_0 + O_{b,A,V}(\ln R) < -E_0/4 + O(1).$$

Thus the frequency lemma yields $S = \Omega_{b,\beta,A,V}(nd_vR)$, proving the lower bound. The histogram upper bound is $O_{b,\beta,A,V}(d_vM_{d,b}R)$. Its enumeration and arithmetic workspace are covered because $M_{d,b} \ge 2d + 1$ and all other displayed logarithmic costs are $O(R)$. Again the upper bound is deterministic.

This proof is uniform over the displayed growing ranges of d and d_v, but not over b. In particular, it does not assert that the starting T is modest for large fixed b. Swapping the roles of fixed and growing parameters would change the theorem.

E Accuracy-sensitive space at a fixed alphabet

This section proves Theorem 3. Its constants may depend on the fixed d, b, B_0, δ. Approximate counting and its binary floating-point implementation are classical (Morris, 1978; Csűrös, 2009; Nelson and Yu, 2022). We use the counter of Csűrös (2009), not a newly invented counter. The upper-bound proof supplies a finite all-prefix certificate using the scalar Freedman inequality (Tropp, 2011), a signed attention reduction, and a numerical implementation without expanded counts.

E.1 A dyadic counter with a finite all-prefix certificate

Let $B_c = 2^w$ with an integer $w \ge 0$. A counter stores (e, r), initially $(0, 0)$, where $e \ge 0$ and $0 \le r < B_c$, and decodes to

$$D(e, r) = (B_c + r)2^e - B_c. \tag{43}$$

On a unit update, advance the integer index $B_ce + r$ by one with probability 2^{-e}; otherwise keep it unchanged. Advancing past $r = B_c - 1$ replaces (e, r) by $(e + 1, 0)$. The positive decoded jump is exactly 2^e, including at this boundary. The first B_c updates are deterministic and exact. This is the binary floating-point counter of Csűrös (2009).

Write D_u for the ideal decoded value after u intrinsic unit updates. Conditional on the current state,

$$\mathbb{E}[D_{u+1} - D_u \mid \mathcal{F}_u] = 1, \qquad \mathrm{Var}(D_{u+1} - D_u \mid \mathcal{F}_u) = 2^e - 1 \le D_u/B_c. \tag{44}$$

Consequently $M_u = D_u - u$ is a martingale, $\mathbb{E}D_u = u$, and $\mathbb{E}M_u^2 \le u(u-1)/(2B_c)$. The variance inequality is not asserted to be an equality for this counter.

Lemma 9 (Finite all-prefix certificate)**.** *For an integer $H \ge 1$ and $0 < \eta \le 1/4$,*

$$\mathbb{P}\bigl[\exists u \le 2^H : \ |D_u - u| > \eta u\bigr] \le 2(H + 1)\exp(-B_c\eta^2/6). \tag{45}$$

Proof. If $2^H \le B_c$ the event is empty. Otherwise let σ be the first relative-error violation and stop M at σ. No violation is possible before B_c updates. Consider a dyadic epoch $u \in [v, 2v]$ with $v \ge B_c$. Before the stopping update, $D_{i-1} \le (1 + \eta)(i - 1)$. Up to $2v$, increments of the stopped martingale therefore have absolute value at most

$$R_v = 2(1 + \eta)v/B_c,$$

and their predictable quadratic variation is at most

$$V_v = (1 + \eta)\sum_{i \le 2v}(i - 1)/B_c \le 2(1 + \eta)v^2/B_c.$$

The negative increment has magnitude at most one, which is covered by R_v because $v \ge B_c$. The bounds also cover the stopping increment, since its distribution is determined by the still-good preceding state; later increments vanish. Thus the bounds are deterministic for the stopped process, not assumptions about an uncapped trajectory.

If $\sigma \in [v, 2v]$, the stopped martingale attains absolute value greater than ηv by $2v$. Apply the maximal scalar Freedman bound to it and its negative:

$$\begin{aligned}\mathbb{P}(\sigma \in [v, 2v]) &\le 2\exp\left(-\frac{\eta^2 v^2}{2(V_v + R_v \eta v/3)}\right) \\ &\le 2\exp\left(-\frac{B_c \eta^2}{4(1+\eta)(1+\eta/3)}\right) \le 2\exp(-B_c\eta^2/6).\end{aligned}$$

There are at most $H+1$ relevant epochs. Summing the first-crossing events proves (45). □

E.2 Attention, uniform intrinsic time, and explicit parameters

Put $Q = 2^b$, $M = M_{d,b}$, and $C = 3M$. For every legal key keep three nonnegative counters: the number of occurrences, the positive integer value total, and the negative integer value total. A token with integer value numerator z increments the occurrence counter once and the appropriate sign counter $|z|$ times. Each counter receives at most QT unit updates. Set

$$H = \lceil \log_2 T \rceil + b, \qquad \eta = \varepsilon/4, \qquad B_c \ge \frac{6}{\eta^2} \ln \frac{2C(H+1)}{\delta}, \tag{46}$$

choosing B_c as a power of two. Lemma 9 and a union bound give simultaneous relative accuracy at every intrinsic prefix of every counter with probability at least $1 - \delta$.

An entirely rational, conservative choice avoids exact logarithm comparisons. Let $q = \lceil \log_2(2C(H+1)/\delta) \rceil$ and take B_c to be the least power of two at least $(21/5)q\eta^{-2}$. Since $\ln 2 < 7/10$, this implies (46). Rational tolerances can be rounded downward within a factor two before choosing parameters; the stored tolerance must have a charged, finite representation.

Why feedback does not invalidate this event. Assign independent private bit streams to the counters. Each complete counter path in its own unit-update index is defined independently of the policy that requests the next update. The good event quantifies over *every* such index, not just the policy's final counts. A causal policy selects an interleaving and stopping indices of these paths using previous answers. Every selected state remains good on the same event. Sampling the independent bits on demand is equivalent. No conditioning on a completed adaptive input sequence is used. Query evaluation is deterministic given the current counters and query, so no union over the query alphabet or number of queries is needed.

Let c_k, a_k^+, a_k^- be the true statistics and define the ideal counter responses

$$\widehat{Z}_0 = \sum_k e^{k\cdot q} \widehat{c}_k, \qquad \widehat{N}_0 = Q^{-1} \sum_k e^{k\cdot q}(\widehat{a}_k^+ - \widehat{a}_k^-).$$

On the common event,

$$|\widehat{Z}_0 - Z| \le \eta Z, \tag{47}$$

$$|\widehat{N}_0 - N| \le \eta Q^{-1} \sum_k e^{k\cdot q}(a_k^+ + a_k^-) = \eta \sum_i s_i |v_i| \le \eta U F. \tag{48}$$

This uses the original token energy and does not require relative accuracy of a cancelling signed sum. Zero sign totals have exactly zero states.

E.3 Charged finite arithmetic and worst-case memory

A cap that costs no additional failure probability. Cap e at $H+2$ and set one failure flag if the cap is reached. At the cap, $D(e, r) \ge B_c(2^{H+2} - 1) > 2 \cdot 2^H$, impossible on the relative-accuracy event within 2^H updates. Coupling the capped and uncapped paths therefore adds no failure probability. Fixed-width counter fields, including the flag but excluding shared parameters, occupy

$$C\big[w + \lceil \log_2(H+3) \rceil\big] + 1 \quad \text{bits.} \tag{49}$$

This is a counter-field budget, not a total-memory measurement.

Exact transitions and short decoding. Generate a 2^{-e}-probability event by requesting at most e independent fair bits, stopping immediately on the first one. This uses $\sum_{j=0}^{e-1} 2^{-j} < 2$ random bits in expectation and no large rational probability. The loop index uses $O(\log(H+3))$ bits. Decode (43) directly in a bounded mantissa–exponent format: if $e = 0$ the answer is r; otherwise subtract B_c from the representation $(B_c + r)2^e$. This subtraction is well conditioned, since the first term is at least $2B_c$. In particular, neither an integer shift of length e nor a recurrence of length $B_c e + r$ is required.

Arithmetic certificate. Use normalized binary mantissas with P bits, unbounded but explicitly encoded integer exponents, and

$$P = \max\left\{w + 2,\ \left\lceil \log_2 \frac{2^{14}(C+1)}{\varepsilon} \right\rceil + 2\right\}, \qquad u = 2^{2-P} \le \frac{\varepsilon}{2^{14}(C+1)}. \tag{50}$$

A positive multiplication, addition, or division can have relative rounding error at most u. For a subtraction use the absolute bound u times the sum of the operand magnitudes. Widely separated exponents are handled by discarding a below-resolution addend with this bound, not by materializing an arbitrarily long shifted integer. With normalized mantissas, an aligned addition needs a shift of at most $P + 4$; otherwise the smaller operand may be discarded. Intermediate integer widths are $O(P)$, apart from the charged fixed-input factors and exponent registers.

Here is a complete error budget. The integers $B_c + r$ and B_c are represented exactly since $P \ge w + 2$. Counter decoding has relative error at most $2u$. Compute each kernel value by the bounded fixed-point exponential routine at absolute error at most 2^{-P-8}, then round to the P-bit format; since the kernel lies in $[e^{-1}, e]$, its relative error is at most $2u$. The routine evaluates the Taylor terms by rounded recurrences at $P + O(\log(P+2))$ fractional bits; a factorial tail and at most a polynomial number of rounding errors supply this bound. It does not form growing exact rational denominators.

Compute numerator sign sums separately and the denominator positively. Each sum has at most M summands. Following an input term through decoding, coefficient evaluation, multiplication, and positive additions bounds its accumulated relative error by

$$(1+u)^{M+8} - 1 \le 8(C+1)u \le \varepsilon/2048.$$

For $M \ge 1$, $(M+8) \le 3(C+1)$; the displayed bound follows from $u \le \varepsilon/[2^{14}(C+1)]$ and $\varepsilon \le .1$. Zero terms are exactly zero. Put $W = \sum_i s_i|v_i| \le UF$. The sum of the two ideal sign magnitudes is at most $(1+\eta)W$. Final signed subtraction thus makes total numerator arithmetic error at most

$$\frac{\varepsilon}{32}(1+\eta)W;$$

the same conservative coefficient bounds denominator arithmetic relative to Z. Combining with (47)–(48), use

$$\gamma = \eta + \frac{\varepsilon}{32}(1+\eta) < \varepsilon/3$$

as a common total budget in $|\widehat{N} - N| \le \gamma UF$ and $|\widehat{Z} - Z| \le \gamma Z$. The normalization lemma gives error below $2\varepsilon\Gamma/[3(1-\varepsilon/3)]$. Final division has relative error at most u, so it adds at most $u[1 + 2\gamma/(1-\gamma)]\Gamma$. The sum is less than $3\varepsilon\Gamma/4$, hence below the required $\varepsilon\Gamma$. These estimates use neither a lower bound on F to require a huge output mantissa nor an assumption against cancellation. If $F = 0$, detect the zero sign states and return exact zero.

Total space and the clock. The counter-field expression (49) excludes shared parameters, keys, arithmetic workspace, and any seeded pseudorandom backend. These must be added in an implementation. In the mathematical fresh-bit model at fixed d, b, δ,

$$w = O(E + \log\log(H+2)), \quad P = O(E + \log\log(H+2)),$$

and exponent encoding takes $O(\log(H + w + P + 2))$ bits. The key list is fixed; it can instead be canonically enumerated. Store H, not T in binary, and no exact event clock. Thus the total state and workspace remain

$$O_{d,b,B_0,\delta}(E + \ln\ln T).$$

A parameter encoding can use a dyadic tolerance within $[\varepsilon/2, \varepsilon]$ and a similarly conservative failure budget. The horizon is externally promised, not validated by an uncharged counter. Each intrinsic

update uses fewer than two fresh bits in expectation. Including the early-stop loop and a possible residue increment, its expected arithmetic cost is $O(w + \log(H + 2))$ and a conservative worst-case cost is $O(w + H \log(H + 2))$; query evaluation is polynomial in the finite alphabet and charged precision. This proves the randomized upper bound in Theorem 3. The implementation uses a seeded generator for reproducibility, whose retained state is not part of the ideal independent-bit model and must be reported separately *and* included in measured totals.

E.4 The accuracy component of the randomized lower bound

Take the single public key 0, $n = 1$, $\kappa = B = 1$, and $r = T - 2$ in Lemma 4. For $L = \lfloor \min(E, \ln T)/(4 \ln 2) \rfloor$, condition (8) holds when $\min(E, \ln T)$ exceeds an absolute constant: its left side is at most a constant times $\exp[-\min(E, \ln T)]$, whereas 2^{-L} is of order $\exp[-\min(E, \ln T)/4]$. Fano gives $\Omega(\min(E, \ln T))$ bits. A one-bit instance covers bounded E. Since $E \le B_0 \ln T$, this is $\Omega_{B_0}(E)$. Only binary values and one terminal query are used.

E.5 The horizon component: distinguishing geometrically separated counts

Let m be the largest integer with $16^{m-1} \le T/4$; thus $m = \Theta(\ln T)$. Alice chooses J uniformly from $\{0, \dots, m-1\}$, inserts 16^J copies of $(0, 1)$, and transmits only the state. Bob inserts zero-valued tokens with key 0 until the *cumulative* number he has inserted is $z_i = 16^i$, for $i = 0, \dots, m-1$, querying 0 after each such stage. This suffix, including its query schedule, is independent of J and uses at most $T/4$ insertions and m queries.

At stage i, $A_i = 16^J/(16^J + 16^i)$ and $\Gamma_i = \sqrt{A_i} \le 1$. If $J \ge i$, then $A_i \ge 1/2$, so a valid answer is at least .4. If $J < i$, then $A_i \le 1/17$, so a valid answer is less than .16. Thresholding at $1/4$ locates the first low answer and recovers J; all high answers mean $J = m - 1$. The joint guarantee gives success at least $1 - \delta$. Fano therefore implies

$$S \ge (1 - \delta) \log_2 m - h_2(\delta) = \Omega(\ln \ln T).$$

This is a legitimate multiple-fixed-query lower bound. It uses the joint correctness requirement, not an assumption of independent per-query successes. It applies to oblivious streams because each complete input sequence is fixed before the algorithm's coins. Together with the accuracy lower bound, $\max\{E, \ln \ln T\} \ge (E + \ln \ln T)/2$ proves the randomized half of (7).

E.6 Why deterministic summaries still need logarithmic space

We use the classical finite-state cycling approach underlying deterministic counting lower bounds; see also Nelson and Yu (2022). Consider an algorithm with at most s persistent bits and at most $m = 2^s$ states at insertion boundaries, with its finite control included in the constant factor. If $m < T/100$ and T is sufficiently large, two all-one prefixes of lengths $1 \le a < b \le m + 1$ have the same state. Let $\ell = b - a$. Since the update is a time-homogeneous deterministic transition on input $(0, 1)$, the same state occurs after $a + k\ell$ ones for every $k \ge 0$ that fits the horizon.

Choose k minimally so $c = a + k\ell \ge 4a$. Then $c < 4a + \ell \le 5(m + 1)$. Append the same a zero-valued tokens to the prefixes of lengths a and c, and ask query 0. Both executions are legal, use at most $6(m + 1) < T$ insertions, and produce the same answer from the same state. Their exact means are $1/2$ and $c/(c + a) \ge 4/5$. Since $\Gamma \le 1$ and $\varepsilon \le .1$, their allowed intervals lie respectively inside $[.4, .6]$ and $[.7, 1.1]$, which are disjoint. This contradiction gives $s = \Omega(\ln T)$. A fixed finite-control factor changes only an additive constant.

This argument would not apply to a device supplied an uncharged exact time index: equal internal states at different times would then not determine equal future transitions. It is precisely why the clock convention is explicit. Conversely, the exact-key histogram uses $O_{d,b}(\ln T + E)$ bits, and $E \le B_0 \ln T$. It proves the matching deterministic upper bound and completes Theorem 3.

F Fixed temperatures and a perturbation transfer principle

The original unit-norm model implies $s_i/s_j \le e^2$. The following results strengthen the stated scope without claiming a sharp-attention phase diagram or identifying the dyadic model with floating-point hardware.

Proposition 10 (Any fixed positive temperature)**.** *Replace the kernel by $e^{\tau k\cdot q}$ for a fixed positive rational τ. Theorems 1, 2, and 3, and Corollary 8, retain their space orders, with constants additionally depending on τ. Their error criterion uses the softmax weights of this new kernel. The constants are not uniform as $\tau \to 0$ or $\tau \to \infty$.*

Proof. All kernel values lie in $[e^{-\tau}, e^{\tau}]$, so the frequency lemma replaces e by e^{τ}. In the Cartesian certificate, the first $p+1$ one-dimensional features have coefficients $\tau^{\ell}/\ell!$. For $0 \le \ell \le p$, these are bounded below by $\min(1,\tau)^p/p!$. The lower bound in (13) is therefore multiplied by $\min(1,\tau)^{dp}$. This contributes $O_{d,\tau}(p)$ to its logarithmic inverse-conditioning cost, below the $O_d(p\log p)$ term governing the asymptotic parameter choice.

In the fixed-grid construction, the selected monomials have total degree at most $2s'$. Their Gram contribution is bounded below by $\min(1,\tau)^{2s'}\Delta^{4s'}I$. At fixed b,τ, the same polynomial conditioning and frequency budget apply. The scalar counting lower bounds use key 0, so their exact responses do not change with τ.

The Taylor numerator and denominator bounds become

$$\|N - N_h\|_2 \le \frac{e^{2\tau}\tau^{h+1}}{(h+1)!}UF, \qquad |Z - Z_h| \le \frac{e^{2\tau}\tau^{h+1}}{(h+1)!}Z.$$

For fixed $\tau > 0$ the required h remains $\Theta_\tau(E/\ln E)$ as $E \to \infty$. Rounded evaluation requires only constant-τ modifications to its guards. Histograms and approximate counters likewise only change kernel-evaluation constants, not the orders of their space. At $\tau = 0$ the kernel is constant and geometric lower bounds can collapse; that case is deliberately excluded. □

Proposition 11 (Original-scale perturbation transfer)**.** *Consider original logits z_i and values V with $F = \|V\|_F > 0$. Let perturbed logits and values satisfy*

$$\max_i |\widetilde{z}_i - z_i| \le a, \qquad \left\|\widetilde{V} - V\right\|_F \le \rho F.$$

Suppose a compressor on the perturbed stream has error at most $\varepsilon_c\widetilde{\Gamma}$, where $\widetilde{\Gamma}$ is computed on that perturbed stream. Then its output obeys the following bound relative to the original exact attention and original scale:

$$\left\|\widehat{A} - A\right\|_2 \le \left[e^{2a}(1+\rho)(1+\varepsilon_c) - 1\right]\Gamma. \tag{51}$$

The statement is pointwise, so it also applies at every realized feedback prefix on a common success event. If $F = 0$, the corresponding exact-zero condition requires the value perturbation to preserve zero.

Proof. Let $p = \mathrm{softmax}(z)$ and $\widetilde{p} = \mathrm{softmax}(\widetilde{z})$. The logit perturbation bounds both the numerator ratio and normalizing-sum ratio, giving

$$e^{-2a}p_i \le \widetilde{p}_i \le e^{2a}p_i, \quad \|\widetilde{p} - p\|_2 \le (e^{2a}-1)\|p\|_2, \quad \|\widetilde{p}\|_2 \le e^{2a}\|p\|_2.$$

Consequently,

$$\begin{aligned}\left\|\widetilde{p}^{\top}\widetilde{V} - p^{\top}V\right\|_2 &\le \|\widetilde{p} - p\|_2 F + \|\widetilde{p}\|_2 \left\|\widetilde{V} - V\right\|_F \\ &\le [(e^{2a}-1) + e^{2a}\rho]\Gamma.\end{aligned}$$

Also $\widetilde{\Gamma} = \|\widetilde{p}\|_2\left\|\widetilde{V}\right\|_F \le e^{2a}(1+\rho)\Gamma$. Adding the compressor error proves (51). □

What this says about quantization. For unit-norm original keys and queries with Euclidean perturbations at most u, the logit $\tau k_i \cdot q$ changes by at most $a = \tau(2u + u^2)$. A componentwise relative value error at most u implies $\rho \le u$ (provided zeros remain zero and the stated relative bound actually holds). Such bounds can be inserted into Proposition 11. Underflow, clipping, group-dependent scales, and subnormal behavior require their actual error bounds; they are not automatically covered by a textbook relative-rounding assumption. The perturbed inputs must also satisfy the domain promised to their compressor. Enforcing this by a further projection costs an additional perturbation, not a free operation.

This proposition transfers a verified numerical error budget; it proves neither that FP16/BF16 has spacing 2^{-b} nor that its optimal memory is $M_{d,b} \log T$. Practical quantizers with per-channel/per-token scales and nonuniform levels (Liu et al., 2024; Hooper et al., 2024) have different alphabets and geometry. A growing τ, including temperature-dependent sharp retrieval, remains outside the fixed-parameter space laws.

G SELF-CONTAINED FINITE CHECKS

These short verification checks are separate from the active-counter panel in Appendix H. They require only the Python standard library and are regenerated by

```
python tests/check_revision.py
python tests/check_counter_implementation.py
```

The regression scripts write `results/revision_checks.json` and `results/local_traces.jsonl`, including configuration, exact integer inputs and outputs, source hashes, and environment information. The accompanying result files record the interpreter version for each rerun.

Table 2: Reproducible short-check coverage. Counts describe finite assertions or executed events, not samples proving an asymptotic theorem.

Check	Completed coverage
Counter expectation, variance, and probability mass	495 exact rational equalities; $K \in \{1, 2, 4, 8, 16\}$ and intrinsic counts 0 through 32
Geometric count decoder	144 exact rational decision-gap cases
Deterministic cycle obstruction	1,600 exact rational length/gap cases
Complete two-key frequency certificate	32 interval-certified message–target pairs; counts-only, no expanded tokens
Sequential deterministic reference checks	18 runs; 2,304 actual insertions and 576 actual queries
Finite-mantissa counter reference	12 seeded runs; 768 actual insertions and 96 actual queries

Counter checks. The complete finite probability distribution of the ideal Morris counter is propagated using rational transition probabilities $(K/(K+1))^J$. At every tested count, total mass, mean, and variance are checked as exact fractions against $\mathbb{E}Z_u = u$ and $\mathrm{Var}(Z_u) = u(u-1)/(2K)$ for the generalized-Morris regression counter. No Monte Carlo confidence interval is involved. The geometric-decoder and deterministic-pumping gaps are similarly checked using fractions. These checks support the algebraic implementation of proof ingredients; they do not validate a universal martingale inequality by experiment or constitute an implementation of the whole low-workspace randomized attention algorithm.

Frequency checks. For the codebook in (22), a degree-36 rational Taylor sum plus $3/37!$ encloses e. Interval inversion of the explicit two-by-two matrix produces legal binary counts. The code verifies the lower eigenvalue certificate via $(2e-5)/16 > 0$ and checks the entire allowable scalar output-error interval, including the true binary-token energy. All 32 message–target pairs remain within half a symbol spacing. The calculated upper enclosure of Ξ is approximately $0.6795704572 < 1$. Repetition is evaluated by exact counts, so zero actual tokens are credited to this check.

Sequential checks. The supplied histogram and rounded-moment reference implementations are run with $d = d_v = 1$, $T = 128$, $\varepsilon = 1/64$, and $b \in \{0, 1, 4\}$. Each implementation is run under fixed, query-feedback, and KV-feedback policies. Every run has 128 actual insertions and one query every four insertions. The policies are deterministic arithmetic functions of the event number and, where permitted, the previous exact algorithm output. That event number belongs to the external generator,

not an uncharged input to the compressed summary. Initial cases include all-zero values, cancellation, and a minimal nonzero value. No policy reads private algorithm state or an oracle result.

The external oracle encloses each kernel value by a rational Taylor bound and compares the *upper* squared residual with the *lower* right side of (21). All 576 queries are certified passes. The maximum certified normalized-error upper bound is approximately 7.864×10^{-5}; it is a finite result under conservative numerical parameters, not a tighter general error theorem. The original token energy is accumulated by the external evaluator and is never replaced by squared keywise means. Trace files preserve exact inputs and output fractions, enabling reruns without hidden notebook state.

Finite-mantissa randomized reference. The additional `CounterAttention` engine implements capped Morris indices, exact rational-probability transitions by repeated Bernoulli trials, and a bounded-mantissa/exponent arithmetic type. It stores neither an exact event clock nor exact expanded decoded counts. The oracle, not the engine, converts its output pair to a rational for checking. With $d = 1$, $b \in \{0, 1\}$, horizon 64, $\varepsilon = .1$, three input regimes, and two fixed test seeds, all 96 queries after 768 actual insertions are interval-certified passes. The maximum certified error ratio is about 2.297×10^{-11}. No cap is reached.

These strict constants choose K much larger than the small tested counts, so the counter behaves almost exactly on these examples. They do *not* exhibit finite-scale memory savings or estimate the universal failure probability. The seeded pseudorandom backend is for reproducible code checks; the theorem assumes independent random bits. Its backend state must be counted in any implementation memory measurement. The supplementary files include the engine, policies, seeds, exact output pairs, and query decisions, not just aggregate numbers.

Scope of the artifact. The deterministic reference engines retain exact horizon and query counters for validation and serialization. They are not the small-workspace counter construction. The generalized-Morris reference is a deliberately slow regression implementation; the dyadic implementation is used for Theorem 3. The active checks and complete-state comparison in Appendices H–I test nontrivial random transitions and report the measured absence of a total-memory advantage.

H FINITE-SCALE ACTIVATION AND COUNTER CALIBRATION

This appendix gives the finite parameters, exact checks, and active dyadic-counter streams for the construction in Appendix E. Complete scripts and result files are supplied. The generalized-Morris implementation in Appendix G is a separate regression baseline, not the algorithm used in Theorem 3. The active runs share one reproducibility seed and are not independent statistical evidence for a worst-case failure probability.

H.1 FINITE PARAMETERS AND A NEGATIVE MEMORY COMPARISON

For $d = 1, b = 0, \varepsilon = .1, \delta = .01, T = 2^{21}$, the rational parameter certificate selects

$$C = 9, \quad H = 21, \quad B_c = 2^{17}, \quad w = 17, \quad P = 23.$$

The dyadic counter fields in (49) occupy 199 bits. Three exact nonnegative statistics per key would occupy 198 bits at this horizon. More importantly, an exact scalar histogram needs only a count and a *signed* value sum per key, giving the smaller field budget

$$M[(\lceil \log_2 T \rceil + 1) + (\lceil \log_2 T \rceil + b + 2)] = 135 \text{ bits}$$

in this example. All these field-only comparisons exclude shared parameters, arithmetic, and any RNG backend. Thus the active-run configuration does *not* demonstrate a finite-scale memory advantage over exact histograms. This unfavorable comparison is retained rather than hidden by comparing against an unnecessarily redundant exact representation.

The file `results/finite_scale_checks.json` contains 24 parameter-only configurations for $b \in \{0, 1\}$, $\varepsilon \in \{.1, 1/32\}$, and horizon exponents $8, 10, 21, 32, 64, 256$. They charge zero executed tokens; large exponents are certificate calculations, not long-stream measurements. Full random-state and workspace costs must be added before choosing a practical representation.

H.2 Actual active-counter runs

We ran the dyadic implementation at the above parameters in fixed, query-feedback, and KV-feedback modes, with seed 104729. Every mode processes exactly 2^{21} tokens and queries every 2^{16} inserts. Keys are in $\{-1, 0, 1\}$ and scalar values are in $\{-1, 1\}$. The fixed schedule puts half the keys at zero and alternates the other two keys. Feedback modifies integer scheduling variables using only the previous public output; it never uses private counter states or oracle errors. The generator's event number is external and is not passed to the compressor.

Table 3: Actual dyadic-counter checks. A hold means a real unit update left the counter state unchanged. The last two columns are rounded upward.

Policy	Observed holds	Max. counter rel. error	Max. attention ratio
Fixed	1,903,086	< 0.003305	< 0.016181
Query feedback	1,903,086	< 0.003305	< 0.016357
KV feedback	1,903,442	< 0.004451	< 0.018288

Each run performs 4,194,304 intrinsic updates; all nonzero values require one sign-counter update as well as the occurrence update. Over three runs there are 6,291,456 primary insertions, 96 primary queries, 12,582,912 intrinsic updates, and 5,709,614 observed holds. Every intrinsic relative error is checked by exact integer arithmetic against $\eta = 1/40$; every query is checked with outward rational exponential intervals and the original squared residual inequality. No cap is reached. Fixed and query-feedback runs have the same KV inputs and counter randomness; the reported equality of their hold counts is expected, not an independent replication.

A midpoint snapshot is restored into a separate unobserved engine and continued on the same subsequent events. All outputs and final snapshots, including RNG state, match. This adds 3,145,728 inserts and 48 queries across the three continuations, reported separately rather than inflating primary coverage. This initial restoration is within the same process; the independent-process tests in Appendix I subsequently test both methods. The observer's exact update counts and ledger are external verification state; the active test is therefore not a low-memory process benchmark. Input hashes, seeds, the deterministic generator, exact output pairs, and interval decisions specify the inputs and outputs needed for regeneration.

These tests establish nontrivial approximation at a certified finite parameter setting. They do not establish universal probabilities, adversarial vulnerability of another method, a total-memory crossover, or an optimized implementation. Comparison with almost-exact short regressions should not optimize for the smallest observed error: all answers here pass while genuinely skipping millions of updates.

H.3 Exact checks and reproduction

The algebraic suite propagates exact distributions for $B_c \in \{1, 2, 4, 8, 16, 32\}$ through 64 intrinsic updates, checking normalization, expectation, and the variance upper bound: 1,170 rational assertions. It also checks 99 exponential enclosures, 60 synthetic mantissa-arithmetic cases, and one large-exponent register case. Synthetic registers include exponent 1000, but receive zero credited stream insertions. These finite checks do not prove Freedman's inequality or the all-prefix lemma; the proof is in Appendix E.1.

```
python tests/check_finite_scale.py
python tests/check_finite_scale.py --active --policy fixed
python tests/check_finite_scale.py --active \
  --policy query_feedback
python tests/check_finite_scale.py --active \
  --policy KV_feedback
```

The default active horizon is 21 and the seed is 104729. Results are written to `results/active_dyadic_*.json`; they include exact output pairs, a hash of regenerated input events, and continuation counts. Timing includes observers, interval checking, and the restore copy and is not an algorithmic latency measurement. The implementation uses a seeded Python RNG for reproducibility,

while the theorem assumes fresh independent random bits. Source hashes and environment manifests identify the recorded executions.

I COMPLETE-STATE RESOURCE AUDIT AND INDEPENDENT CONTINUATION

This audit closes the distinction between small counter fields and a complete executable state. It uses the certified active configuration in Appendix H, without changing B_c, P, ε, the seed, or the recorded inputs. An exact histogram stores only each key's count and signed value sum; an unnecessary third nonnegative total is not charged to that baseline. Its kernel/output fixed-point budget is $P = 33$ (the final pair has a 35-bit mantissa), whereas the dyadic engine uses $P = 23$. Both budgets are conservative and are chosen before execution; the claim of a minimal baseline refers to its sufficient statistics, not optimized arithmetic constants.

I.1 PAIRED TRACES AND ACTUAL EXECUTED WORK

For each of the three active policies, the source trace is defined by the published integer generator and 32 saved dyadic-summary answers. Both methods replay precisely that trace. In particular, the histogram's distinct answers do not change future inputs. The driver knows event numbers and saved public answers, but neither algorithm receives an event clock, oracle, or readable input history. Hashes of regenerated input events equal the original active-run hashes.

Each method–policy pair executes 2^{21} insertions and 32 queries: six replays, 12,582,912 insertions, and 192 queries in total. The dyadic outputs reproduce the active-run records exactly. An external rational-interval oracle, run outside the allocation measurement, certifies the exact histogram's outputs as well. These are resource replays of existing cases, not six independent new statistical samples.

At the midpoint each complete state is encoded to bytes. A fresh process receives those bytes and the externally generated suffix, not the original algorithm object. All six continuations reproduce all 16 subsequent answers and the final full encoding, including the dyadic PRNG state. The continuations add 6,291,456 insertions and 96 queries, excluded from the replay totals. The snapshot format is typed and versioned, checks lengths, ranges, metadata, and a checksum, and uses no `pickle` or executable deserialization. It is an internal audit format, not a security-hardened general-purpose interchange standard.

I.2 FOUR DISTINCT SPACE MEASUREMENTS

The first metric is the declared capacity of numeric counter fields, given by (49). It is 199 bits for the dyadic counters and failure flag. For the exact histogram, each of three keys has a 22-bit count and a 23-bit signed sum, giving 135 bits. These fields exclude parameters and arithmetic workspace.

The second is an actual canonical encoding of *every instance field*, including keys, derived parameters, containers, field names, and the complete seeded backend. Both implementations use the same format. At the final boundary, the dyadic encoding uses 89 parameter bytes, 20 key bytes, 89 numeric-state bytes, 2,508 RNG bytes, and 14 framing/checksum bytes, totaling 2,720. The exact encoding uses $45 + 20 + 43 + 0 + 14 = 122$ bytes. The largest state encoding at the four measured checkpoints is 2,721 bytes for dyadic counting and 122 for the exact histogram; the 199/135-bit field budgets, unlike a snapshot byte count, are fixed capacities. The RNG component includes its field name and codec tags; it is not just the backend's minimal information content. Shared class code and language-runtime objects are not serialized, and this codec is not information-optimal.

The third metric recursively counts reachable Python instance data, deduplicating shared objects and including the native `Random` instance. It excludes classes, modules, and the interpreter. The fourth measures the maximum of separately tracked Python allocation peaks for construction, insertion, query, serialization, and restore calls; tracing is enabled before construction. Preloaded source records and preexisting interpreter/native allocations are excluded; small driver structures and temporary serialization/restore copies are included. Thus neither the traced peak nor a line-boundary increment is claimed as pure algorithm working space. These four metrics are not additive.

Checkpoints cover the initial state, a publicly fixed activation-scale insertion 262,145, the midpoint, and the final state. The activation-scale checkpoint is not asserted to equal the first random hold under

Table 4: All six completed same-trace replays. State (B) denotes the full canonical persistent encoding in bytes; traced peak is the maximum of separately measured phases, not their sum. Error ratio is the certified upper bound divided by $\varepsilon\Gamma$.

Policy	Method	State (B)	Objects (B)	Peak (B)	Error ratio
Fixed	Dyadic	2,720	5,367	68,499	0.016180
Fixed	Exact	122	1,473	17,120	5×10^{-10}
Query feedback	Dyadic	2,720	5,367	68,442	0.016356
Query feedback	Exact	122	1,473	17,165	5×10^{-10}
KV feedback	Dyadic	2,720	5,367	68,489	0.018288
KV feedback	Exact	122	1,473	17,120	3.2×10^{-10}

every policy. Operation-phase allocation tracking covers the full sequence of updates and queries, not only these checkpoints. Uninstrumented driver transients are not an algorithmic peak certificate. Tracing overhead prevents using its elapsed times as latency benchmarks. The measurements support no finite total-memory advantage for randomized counting at these parameters.

I.3 Numeric workspace: observation versus a register certificate

A separate inspector samples algorithm-local numeric values at Python call, line, and return boundaries, excluding persistent fields and the external evaluator. It follows simultaneously live algorithm frames, rather than summing maxima from mutually exclusive calls. Across 24 method–policy–checkpoint tests, the maximal sampled additional numeric payload is 479 bits. This is *not* a bytecode-exact peak: expressions and native backend operations can allocate between samples.

We therefore report a second, independent bound for a finite-register realization of the inspected arithmetic. Put

$$g = \mathrm{bitlength}(P + 40) + 16, \qquad e_b = \mathrm{bitlength}(H + P + w + 32) + 2,$$

using $w = 0$ for the histogram. A conservative per-register width is

$$L_D = 2(P + 8 + g + 2b + 16) + e_b + 16, \qquad L_E = L_D + H + 2b.$$

Here L_D or L_E is evaluated with its own method's P. The rounded-exponential routine uses a factorial only until its threshold 2^{P+20} is exceeded, with a final overshoot of logarithmic width. Binary-float products/division and aligned addition need at most $2P + O(1)$ mantissa bits; large exponent gaps discard a below-resolution term instead of shifting by the exponent. Exact histogram products additionally need $H + 2b$ bits. This gives the displayed width bounds.

Inspecting the deepest relevant call chain (query, exponential or binary-float evaluation, then integer rounded division) allows at most 96 named/input numeric registers and 32 expression-result registers simultaneously. The envelope $128L_D$ or $128L_E$ includes these transients; it is deliberately conservative, not an observed peak and not a proof about the CPython allocator. Persistent parameters/state remain separately charged by the full encoding. For the particular 32-bit MT19937 backend, eight extra 32-bit scratch words cover a sequential implementation of its native update; its whole state is already in the persistent encoding. Serialization copies and Python control/container overhead are reflected in the distinct measured allocation metric, not hidden in this numerical-register model.

The inspector also evaluates a synthetic exponent-1000 counter without materializing a 1000-bit decoded integer. This is a register test with zero credited stream insertions. Eight contract tests cover full-state round trips, damaged encodings, exact-zero/cancellation behavior, absence of hidden

Table 5: Separate numeric measurements and reviewed finite-register envelopes at the tested configuration. The envelope columns are upper budgets, not measured savings or sums of observed frame peaks.

Metric (bits)	Dyadic	Exact
Sampled temporary numeric payload	479	479
Widest sampled explicit integer	53	88
Reviewed register width envelope	163	206
Reviewed numerical workspace envelope	20,864	26,368

clock/history fields, nontrivial random transitions, the finite Freedman constants, counter-field budgets, and original energy. Finite tests do not replace the analytic worst-case proof.

I.4 Reproduction

After regenerating the active traces, run

```
python tests/final_resource_audit.py
python tests/register_audit.py
python -m unittest discover -s tests -p test_final_contract.py
python scripts/build_final_tables.py
```

The six JSON records, exact byte snapshots, fresh-process continuation outputs, complete resource CSV, register observations, and environment/source hashes are included in `results/final_resource_audit/`. The tables are derived from these records without interpolated configurations. These commands require no remote server, hidden policies, or large-model downloads.

The tables report the supplied recorded executions. The editable archive retains subsequent integration changes to shared arithmetic/audit helpers; its source and environment manifests distinguish these revisions from the recorded measurement snapshot. Rerunning the supplied commands regenerates results in the current environment, rather than guaranteeing identical allocation peaks or register observations. Regenerated resource records and tables should be updated together, not mixed with the supplied rows.