LOCAL TESTS, LONG HORIZONS: WHAT MEMORY-COMPACTION TESTS IDENTIFY

Anonymous authors
Paper under double-blind review

ABSTRACT

Short reset tests can be reused to assess repeated memory updates, but a precise local test need not identify long-horizon risk. We study which output statistics a test must retain, which uncertainty a computable bound actually covers, and when the resulting test is affordable. Complete marginal tables can admit nearly unit stationary risk ambiguity. A support-aware continuation criterion characterizes exact identification for all initial states, while quantitative examples show why small first-horizon ambiguity is not a deployment-horizon guarantee. Across 330 geometries, certified frozen-kernel brackets reveal discrepancies that a two-step diagnostic misses. We then select among candidate output features using simultaneously calibrated observations, without supplying the generator's influential feature. Selection improves some matched-budget certification decisions but loses at the smallest budget because of its statistical correction. Established random-set and robust dynamic programming (DP) methods provide propagation; we make their information, temporal, and sampling relaxations explicit. The supported operating regime is cheap resettable software with declared small-block structure, not unrestricted language-model compaction.

1 INTRODUCTION

Suppose an agent remembers "Use configuration B only after test C passes." A compaction drops the condition. Later updates receive an already-corrupted instruction and may preserve it, spread its consequences, or repair it. Testing only the original clean instruction cannot establish how those later updates behave. Persistent memory systems repeatedly replace accumulated context with bounded representations (Packer et al., 2023; Xu et al., 2025; Chhikara et al., 2025; Wu et al., 2025), making this feedback a natural reliability question.

A direct test of H updates costs H calls. One-step tests with independent resets can instead estimate responses to different memories and reuse those responses across horizons. But there are two distinct compression decisions: the deployed system compresses its memory, and the evaluator compresses its observations into statistics. Testing every input does not compensate for discarding an output event that controls future errors. A test may therefore estimate the wrong summary very precisely.

This paper asks three questions. What retained information identifies the risk of a *frozen* update rule? Can a computable bound distinguish information loss from statistical uncertainty? Which assumptions make local testing worth its cost? The answers depend on the assumptions and guide a practical choice: before spending more calls, identify whether the obstacle is an inadequate statistic, a permissive transition model, or simply too few observations.

The first answer is an exact continuation criterion, including known structural transition zeros. Its sufficient direction is established value-equivalence reasoning (Poupart & Boutilier, 2002; Grimm et al., 2020); the converse exposes which row directions a test leaves unresolved. Exact non-identification is not automatically important. We therefore quantify stationary intervals, give a sharp long-horizon marginal counterexample, and test the first-unresolved-horizon diagnostic at deployment horizons. A small two-step width can coexist with much larger 128-step uncertainty, and zero two-step width does not rule out a later gap.

The second answer is a transparent account of the bound. Our count-tail upper expectations are specializations of existing random-set methods (Utkin & Destercke, 2009), composed by existing

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.

robust dynamic programming (Iyengar, 2005; Nilim & El Ghaoui, 2005). Their worst-case process can change transition rules with time even when the actual compactor is frozen. We define both semantics and distinguish them from discarded lower constraints and finite-sample uncertainty. The split between information and temporal effects depends on the chosen relaxation path; it is not an intrinsic causal attribution.

The third answer is an assumption–cost comparison and a data-selected feature experiment. The evaluator chooses among all candidate record bits under one simultaneous confidence event. The generator's influential bit is hidden, rather than supplied as the answer. The procedure can help, but its multiplicity correction can also outweigh its benefit. Likewise, concrete-input enumeration is exponential; known count dynamics or a verified low-dimensional mechanism can sharply reduce calls. These stronger models require independent justification.

The contribution is a diagnosis of *what is identifiable, at which horizon, and at what cost*, supported by exact constructions, numerical enclosures, and controlled tests. We make the limits visible rather than claiming a universally superior algorithm. The experiments use synthetic processes. No natural-language compactor has a validated error-state abstraction here.

2 Related work

Memory compression includes recurrent summaries, gist tokens, prompt selection, and archive rereading (Rae et al., 2019; Mu et al., 2023; Jiang et al., 2023; Lee et al., 2024). The long-term conversational memory benchmark LoCoMo and LongMemEval evaluate conversational memory (Maharana et al., 2024; Wu et al., 2024). Repeated-compaction rate–distortion analysis (Colaco & Lahjouji, 2026) and paired closed-loop evaluation (Min et al., 2026) are especially close motivations. We do not introduce self-induced input shift (Ross et al., 2011) or repeated-compaction evaluation.

Value-directed compression and value equivalence preserve appropriate continuation functions (Poupart & Boutilier, 2002; Grimm et al., 2020). Random-set expectations (Utkin & Destercke, 2009), uncertain Markov chains (Givan et al., 2000; de Cooman et al., 2009), and robust dynamic programming (DP) (Iyengar, 2005; Nilim & El Ghaoui, 2005) supply our mathematical machinery. Stationary versus time-varying uncertainty is also established (Campos et al., 2004; Suilen et al., 2024). Our role is to connect these objects to retained test information, supported identification, and measurable certification gaps, rather than to relabel the machinery as new. Statistical model checking and state abstraction are the broader setting (Ashok et al., 2019; Givan et al., 2003; Ferns et al., 2004; Zacchia Lun et al., 2018). Appendix A maps the remaining probability and calibration ingredients to their sources.

3 Model and test interface

A fixed rule updates memory by $M_{t+1} = C(M_t, O_{t+1})$ with at most B retained records. The evaluator knows the synthetic truth; the compactor does not receive a correctness flag. A finite analysis state $E_t \in \mathcal{E}$ describes errors or missing records. For a closed k-record block, $\mathcal{E} = \{0,1\}^k$; for a hard-capacity cache, only physically possible states are included. Unless otherwise stated, the frozen-kernel analysis assumes this state is sufficient and its transition law is stationary.

For row-stochastic K, initial law μ, and loss vector $L : \mathcal{E} \to [0,1]$, define

$$R_H(K) = \mu K^H L. \tag{1}$$

A uniform record query has $L(e) = |e|/k$. Query selection is independent of the hidden error state and does not change the update. Endpoint risk differs from the probability of any error over a trajectory: a system can recover often and still fail at least once with high probability.

A reset test supplies an input state and observes one output. Calls are independent conditional on a predetermined input design. The *retained* test statistics need not be independent when extracted from the same output. In the exact-moment analysis, F is a matrix of output features, including $\mathbf{1}$, and $M = KF$ is known. Compatible frozen kernels are

$$\mathcal{C}_{M,S} = \{P \geq 0 : P\mathbf{1} = \mathbf{1},\ PF = M,\ P(e,y) = 0 \text{ if } y \notin S_e\}.$$

Here S_e is an independently declared transition support. A zero observation count does not establish a structural zero.

For sampled count tests retain an ordered output statistic $\phi : \mathcal{E} \to \{0, \dots, k\}$ and its tails

$$T_e(\ell) = \Pr\{\phi(E_{t+1}) \geq \ell \mid E_t = e\}, \qquad 1 \leq \ell \leq k. \tag{2}$$

The test constructs U with $\Pr_{\mathcal{D}}(T \leq U) \geq 1 - \delta$. Conditional on that event, propagation gives risk bounds over deployments. Thus δ concerns repeated tests, whereas R_H concerns repeated deployments. "Not certified" does not mean "unreliable." Reference kernels are reserved for the evaluator; the sampled procedures receive observations, declared supports, and candidate features.

4 Which uncertainty matters over long horizons?

4.1 Complete local accuracy can miss a nearly unit risk gap

Theorem 1 (Sharp ambiguity from discarded dependence)**.** *Let $d \geq 2$, $0 < \epsilon \leq (d-1)/d$, and $\epsilon \leq b \leq 1$. Suppose every coordinate has the complete conditional marginal table*

$$\Pr(E_{t+1,i} = 1 \mid E_t = e) = \epsilon + (b - \epsilon) \prod_{j=1}^{d} e_j. \tag{3}$$

Starting clean, the exact range of uniform-query risk at every $H \geq 1$ is

$$\left[\epsilon,\ \frac{\epsilon\{1 - (b-\epsilon)^H\}}{1 - b + \epsilon}\right]. \tag{4}$$

Both endpoints are attained by stationary joint kernels. At $b = 1$ and $H \gg 1/\epsilon$, the gap approaches $1 - \epsilon$.

The upper kernel uses a common shock to make all records wrong together; the lower avoids that joint event while preserving each marginal. More samples of those same marginal expectations cannot separate them. Even retaining *every proper output marginal* can leave a gap approaching $1/2$ (Theorem 19). These results motivate retaining consequential joint events rather than claiming that short tests themselves are impossible. Full-vector observations distinguish the constructions. Proofs are in Appendices J.11 and K.1.

4.2 Identification with known transition zeros

Let $N = |\mathcal{E}|$. Use the effective supports of $\mathcal{C}_{M,S}$: remove any coordinates that all feasible rows force to zero. Define

$$\mathcal{V}_S = \{v \in \mathbb{R}^N : \ v|_{S_e} \in \mathrm{col}(F|_{S_e}) \text{ for every } e\}.$$

Different rows may use different feature coefficients. This is the appropriate tangent-space condition on a supported model class.

Theorem 2 (Supported frozen-risk identification)**.** *If K lies in the relative interior of $\mathcal{C}_{KF,S}$, all compatible vectors $P^h L$, $1 \leq h \leq H$, agree if and only if $K^j L \in \mathcal{V}_S$ for every $0 \leq j < H$. At the first failure there are two relative-interior compatible frozen kernels with different next-step risks from at least one initial state.*

A represented continuation has the same expectation in every compatible row, proving sufficiency by induction. At a failed row, a small perturbation orthogonal to its supported feature columns preserves all moments and changes that continuation. Appendix B gives the proof and explains why necessity for *one fixed* initial law needs extra reachability assumptions. Under full support, $\mathcal{V}_S = \mathrm{col}(F)$ and the fixed-law theorem in Appendix D is recovered. Its small interior witness is a qualitative construction, not our numerical measure of risk ambiguity.

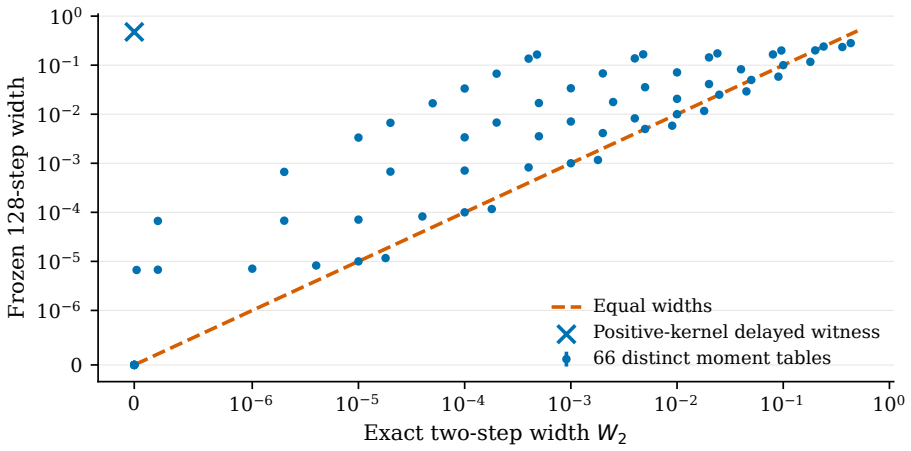

Figure 1: A first width need not describe deployment uncertainty. Dots represent all 66 distinct count tables in the 330-reference grid; vertical intervals are certified stationary brackets, not sampling error bars. A separate positive-kernel example has $W_2 = 0$ but a stationary witnessed gap 0.48 at $H = 128$.

4.3 A FIRST UNRESOLVED HORIZON IS NOT A DEPLOYMENT GUARANTEE

At the first horizon where a shared continuation v_j is not represented, its next expectation is still a linear row problem. If exact output-class laws $q_e(c)$ are known for classes C_c, the stationary width is

$$W_{j+1} = \sum_{e,c} \mu_e q_e(c) \left[\max_{y \in C_c \cap S_e} v_j(y) - \min_{y \in C_c \cap S_e} v_j(y) \right]. \tag{5}$$

Zero-mass classes contribute zero. Both extrema are attained by frozen kernels (Proposition 12). It measures the variation in future loss inside each indistinguishable output class, weighted by reachable class mass. It does not compound itself into a long-horizon bound.

We evaluate the 330 four-state geometries in Appendix I.2 at horizons 2, 8, 32, 128, 256, 512, and 1024. For $H = 128$, exact rational stationary limits, outward-rounded contraction bounds, and feasible frozen witnesses enclose every stationary width to within $6.98\,10^{-5}$ (Appendix B.3). There are 66 distinct count tables and five reference orientations per table. The descriptive rank correlation between W_2 and W_{128} is 0.87, but the ratio over positive widths ranges from 0.65 to 342.5. The first width shows a trend but does not reliably predict the later width.

Why an early diagnostic can miss later behavior. The following pair makes the failure concrete. With states $00, 10, 01, 11$, let

$$A_0 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad B_0 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Put $A_\zeta = (1-\zeta)A_0 + \zeta J$ and $B_\zeta = (1-\zeta)B_0 + \zeta J$, where every entry of J is $1/4$. For $0 < \zeta < 1$ both kernels are strictly positive and have the same complete count law in each row. Their common one-step continuation is count-measurable, so the entire compatible class has zero two-step width. Later continuation values need not be count-measurable: which one-error state is visited controls whether errors can reach the all-wrong state. At $\zeta = .001$, the fixed pair itself witnesses the 128-step separation below. No optimizer, untested input, or temporal adversary is needed for this example.

There is no universal multiplicative repair. Appendix B.2 gives two strictly positive frozen kernels with identical complete count laws and zero two-step ambiguity, but risks 0.52 and $1 - 1.87 \times 10^{-3}$ at step 128. In this grid, the support-aware two-step upper bound additionally certifies $110/330$ cases at target $.51$ that the generic residual bound does not; all four residual targets give $405/1320$

additional decisions. Those are explicitly *two-step* results. Appendix B.3 describes the stationary brackets used for the long-horizon comparisons.

What an approximate diagnostic must retain. For exact moments $PF = KF$, write $K^j L = Fa_j + r_j$. Telescoping and stochastic nonexpansivity give

$$\|P^H L - K^H L\|_\infty \le \sum_{j=0}^{H-1} \operatorname{span}(r_j).$$

If moments differ by at most η per entry, add $\eta\|a_j\|_1$ at each step. Thus approximate closure can support useful certification, but one must inspect the continuation residuals relevant to the requested horizon, not just the first one. Large coefficients can amplify sampling error even when residuals are small. This remains a reference-model diagnostic; a sampled evaluator cannot assume an unvalidated K or its residuals. Our confidence-region procedure below takes a different route: it propagates every model consistent with its sampled information.

5 What a computable upper bound actually covers

5.1 An explicit uncertainty class and a standard backup

For tail limits U, define the row class before using it:

$$\mathcal{Q}_e(U) = \{p \in \Delta(\mathcal{E}) : \; p(\phi \ge \ell) \le U_e(\ell),\; 1 \le \ell \le k\}.$$

A declared support can be added as $p(y) = 0$ outside S_e. Put $\bar{U}_e(\ell) = \min_{m \le \ell} U_e(m)$, $\bar{U}_e(0) = 1$, $\bar{U}_e(k+1) = 0$, and $\pi_{e,c} = \bar{U}_e(c) - \bar{U}_e(c+1)$. The masses $\pi_{e,c}$ sum to one. They assign probability to nested sets $A_c = \{y : \phi(y) \le c\}$, not to particular output orientations.

Proposition 3 (Nested-set upper expectation)**.** *When $A_0 \neq \varnothing$, the worst continuation over $\mathcal{Q}_e(U)$ is*

$$(\mathcal{B}_U v)(e) = \sum_{c=0}^{k} \pi_{e,c} \max_{y \in A_c} v(y). \tag{6}$$

The recursion $V_0 = L$, $V_{h+1} = \mathcal{B}_U V_h$ is exact for compatible time-varying row laws and conservative for every compatible frozen kernel. For increasing count loss it is no larger than the count-only safeguard using the same confidence table.

A small example distinguishes counts from upper tails. For two records, suppose $U_e(1) = .3$, $U_e(2) = .1$, and a continuation has values $(0, 1, .1, .2)$ on $00, 10, 01, 11$. The set masses are $(.7, .2, .1)$, so the upper expectation is .3: both nontrivial sets can put their mass on 10. Exact count probabilities $(.7, .2, .1)$ would instead force the last mass onto 11, giving .22. Thus the one-sided relaxation can enlarge a bound even when all output counts have been measured without sampling error. It is a restriction on retained information, not a failure of the optimizer. A supported version intersects each set with the allowed row support; empty low-count sets require the corresponding supported row optimization rather than the displayed unpruned formula.

This is the finite nested-random-set representation of upper expectation (Utkin & Destercke, 2009), followed by rectangular robust dynamic programming (DP) (Iyengar, 2005; Nilim & El Ghaoui, 2005). Its operational interpretation is to draw a set index c and choose its most harmful output. A different output can be chosen at every remaining horizon. Exact two-sided count laws instead assign each count class its known mass.

5.2 The relaxation lattice, with all definitions

Let $\mathcal{C}_=(T)$ contain frozen kernels with exactly the count law specified by T, and let $\mathcal{C}_\le(T)$ retain only its upper inequalities. For $a \in \{=, \le\}$ define

$$S_a(T) = \sup_{P \in \mathcal{C}_a(T)} \mu P^H L, \qquad D_a(T) = \sup_{P_1, \dots, P_H \in \mathcal{C}_a(T)} \mu P_1 \cdots P_H L.$$

The second expression is the time-varying, row-rectangular interpretation. On the test event $T \leq U$,

$$R_H \leq S_{=}(T) \leq S_{\leq}(T) \leq D_{\leq}(T) \leq D_{\leq}(U). \tag{7}$$

Along this chosen path the increments measure discarded orientation, dropped lower constraints, temporal relaxation, and sampling. The alternative path goes through $D_{=}(T)$; it need not be ordered with $S_{\leq}(T)$. Therefore attribution to information versus time is path dependent. Appendix B.4 gives the symmetric two-order average, which is a convention rather than a unique causal split.

The temporal gap is substantial even with exact two-sided counts. On states $00, 10, 01, 11$, require deterministic next counts $(1, 1, 2, 0)$ and start at 00. At $H = 4$, uniform-query stationary risk is $[1 - (1 - x)y^2]/2$ for compatible orientation parameters $x, y \in [0, 1]$. Thus $S_{=}(T) = 1/2$, while a time-varying sequence attains $D_{=}(T) = 1$. Appendix E proves this and keeps the less informative one-bit upper-tail example as a secondary illustration. Stationary/nonstationary semantics are established (Campos et al., 2004); these exact values clarify which class a test bound optimizes.

6 Finite data, feature choice, and the operating regime

Fixed-design binomial endpoints cover Rk row/threshold parameters by a union bound (Clopper & Pearson, 1934). Different statistics on a shared output can be dependent. All deployment horizons then reuse the same confidence event. Sequential input selection or test stopping instead requires time-uniform confidence sequences (Howard et al., 2021); repeated inspection of a fixed-sample interval is insufficient. Coverage does not make the allocation heuristic optimal: the adaptive scheduler is slower than balanced allocation on merge (Appendix K.9).

Proposition 4 (Selection among tested output features)**.** *Fix m candidate binary functions before collecting data. Give every count tail and every candidate mean in each row a binomial upper endpoint at level $\delta/[R(k + m)]$. For each candidate, intersect its mean cap with the common count-tail region and propagate its robust bound. Any data-dependent choice among these candidate bounds, including their minimum at a requested horizon, is valid with confidence $1 - \delta$, simultaneously over deployment horizons.*

One union event covers all candidates and all rows; backward induction then covers all horizons. The proof is in Appendix C. This straightforward simultaneous-inference argument closes the selection gap without claiming a novel confidence method. It permits choosing a candidate after seeing data, but not constructing an unrestricted new feature from those same observations. The extra bit constraint is evaluated by the standard one-multiplier dual in Appendix I.3.

The sampled procedure, without a reference kernel. For candidate b_j, let V_{ej} be the upper endpoint of its output mean and intersect the count region with that extra constraint:

$$\mathcal{Q}_e^{(j)}(U, V) = \{p \in \mathcal{Q}_e(U) : \textstyle\sum_y p(y) b_j(y) \leq V_{ej}\}.$$

Set $v_0^{(j)} = L$ and recursively maximize $pv_h^{(j)}$ over this row region to obtain $v_{h+1}^{(j)}(e)$. Report

$$\widehat{R}_H = \min_j \mu v_H^{(j)}.$$

On the single simultaneous event, every candidate is conservative, so the data-dependent minimum is conservative too. There is no evaluator transition matrix in these steps. A test first resets the specified inputs, then extracts the count histogram and candidate bit totals from each full output, constructs the confidence endpoints, and propagates the resulting regions. Unseen input rows receive the unconstrained upper value one. The procedure cannot select its feature dictionary freely after observing outcomes: that would change which confidence family had been protected.

This separates validity from statistical value. Adding a candidate shrinks its feasible row region *at a fixed confidence table*, but requiring all candidate means to be valid makes that table more conservative. A comparison using the native count baseline must pay this correction, as our experiment does. A fixed-bit control using the same corrected table separates feature selection from merely changing calibration. The computation scales with the number of candidate backups as well as with the concrete input count; the sampled information is shared, the optimization work is not.

Table 1: Assumptions, calls, and mean bound at $H = 128$, exact risk 3.45×10^{-3}. The pooled companion has the same mean-risk curve but a different, exchangeable transition law; pooling the hidden-bit family would be invalid. The two-input control uses stronger mechanism knowledge. These are alternative operating contracts, not a universal ranking.

Declared model and retained output	Reset calls	Upper bound
Concrete inputs; count tails, $k = 12$	2,097,152	0.05
Same inputs; data-selected bit	2,097,152	0.04
Count-lumpable monotone companion; 13 inputs	32,768	8.36×10^{-3}
Shared nonnegative affine mean; two inputs	8,192	6.05×10^{-3}

Affordability requires declared structure. An unrestricted k-bit row test uses $R = 2^k$ input configurations. Generic exact identification can require $\min(H, 2^k - 1)$ nonconstant output statistics (Corollary 6). Neither result describes an affordable black-box large language model (LLM) procedure. Even the sufficient uniform Hoeffding rate $O(RH^2 \log(Rk/\delta)/\varepsilon^2)$ reaches roughly 10^{13} calls at the stated 12-bit, 128-step, .01-precision scale. It is a pessimistic sufficient rate, not a lower bound, but the measured multi-million-call experiments are themselves too expensive if each call is an LLM update.

Table 1 makes the useful regimes explicit. A small declared block, a verified reward-preserving finite abstraction, or a known closed mean mechanism can make testing manageable when resets are cheap. These assumptions are not totally ordered; there is no universal "weakest" one. Calibration cannot discover arbitrary state sufficiency from a few favorable trajectories. The quantity to optimize is the risk bound at the affordable budget, not exact identification at any cost. The generic uniform perturbation analysis has sufficient rate $O(2^k H^2 \log(2^k k/\delta)/\varepsilon^2)$. At $k = 12$, $H = 128$, $\varepsilon = .01$, and $\delta = .05$, the expression before unspecified constants is of order 10^{13} calls. This worst-case sufficient rate is not a necessity bound, but it is not a practical prescription either. The measured experiments use much smaller budgets and report the bounds they actually obtain. A structural contract must reduce tested inputs, propagated uncertainty, or both before an expensive compactor becomes an appropriate target.

A bounded mismatch contract. If the next-output tail conditional on every deployment history ending in e is at most $T_e(\ell) + \eta_e$, propagate $\min(1, U_e + \eta_e)$ (Theorem 9). The general bound $H\eta \operatorname{span}(L)$ is a sensitivity summary, not the numerical guarantee used in the experiments. At $H = 512, g = .01$ it is vacuous; the useful result comes from directly composing inflated tail regions. For the cache-reuse wrapper, an independent pre-output gate with acceptance probability at most g supplies $U_e^g = U_e + g(1 - U_e)$ without estimating a discrepancy kernel. The unchanged ordinary branch and the extra snapshot storage are part of that contract (Appendix I.5).

7 Experiments and what they support

7.1 Data-selected augmentation, including the correction cost

The hidden-bit family has $k = 4, 8, 12$ and $L(e) = |e|/k$. The evaluator draws an influential index J; its next-bit mean is

$$(KL)(e) = .003 + .05e_J + .08L(e), \qquad (Ke_J)(e) = (KL)(e). \tag{8}$$

The evaluator receives no J, coefficients, or transition matrix. It evaluates all k bit features from the same reset outputs, using Proposition 4. The candidate family contains a sufficient statistic by construction; the tested question is whether corrected selection can find useful information without being told which bit it is. Exact closure explains the oracle endpoint, but does not guarantee a finite-budget improvement.

We use 24 test datasets for each of 12 block-size/allocation combinations, totaling 288 datasets with independent reset sampling within each design. The 24 hidden-index assignments per block size are reused across budgets. Every output contributes to both its count histogram and all bit totals. Efficient joint sufficient-count simulation charges the corresponding primitive resets rather

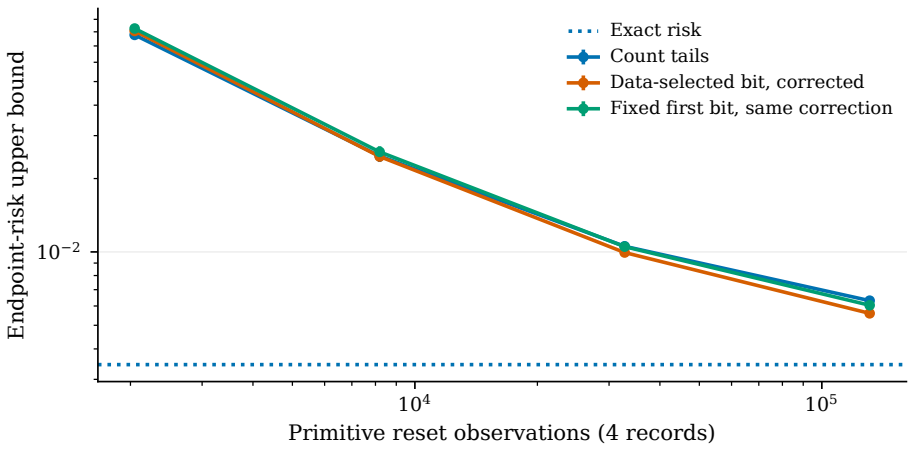

Figure 2: A hidden feature is selected from data under simultaneous correction, rather than supplied by the evaluator. At the smallest four-record budget, selection worsens the bound; at larger budgets it helps. Error bars describe mean variability across 24 tests, not simultaneous confidence bands.

Table 2: All data-selected-feature designs at $H = 128$. Means use 24 tests per row; each design selects from k candidates under a simultaneous correction. The first-row loss is not hidden by presenting only threshold changes. Exact risk is 3.45×10^{-3} in every row.

k	Reset calls	Count-only	Selected bit	Count minus selected
4	2,048	0.08	0.08	-2.93×10^{-3}
4	8,192	0.02	0.02	1.40×10^{-4}
4	32,768	0.01	9.94×10^{-3}	6.00×10^{-4}
4	131,072	6.31×10^{-3}	5.60×10^{-3}	7.10×10^{-4}
8	32,768	0.12	0.11	4.71×10^{-3}
8	131,072	0.04	0.03	4.24×10^{-3}
8	524,288	0.01	0.01	3.02×10^{-3}
8	2,097,152	8.67×10^{-3}	6.16×10^{-3}	2.51×10^{-3}
12	524,288	0.16	0.14	0.02
12	2,097,152	0.05	0.04	9.61×10^{-3}
12	8,388,608	0.02	0.01	6.46×10^{-3}
12	33,554,432	0.01	6.66×10^{-3}	4.96×10^{-3}

than treating one multinomial draw as one update. All budgets and targets are fixed before examining these results. Reported results are rounded; decisions use unrounded values.

At $H = 128$, selection identifies the hidden bit in all 288 tests. Yet at four records and 2,048 calls, the mean selected bound is 8.06×10^{-2}, worse than native count-only 7.77×10^{-2}: the multiplicity cost exceeds the gain. Target .08 is certified in $12/24$ versus $17/24$ tests. At 32,768 calls, the means are 9.94×10^{-3} and 0.01, and the .01 target improves from $9/24$ to $13/24$. The paired mean improvement is 6.03×10^{-4}, with descriptive 95-percent interval $[5.16 \times 10^{-4}, 6.90 \times 10^{-4}]$.

At 12 records and 2,097,152 calls, means improve from 0.05 to 0.04, changing target .05 decisions from $14/24$ to $24/24$. At 33,554,432 calls, they improve from 0.01 to 6.66×10^{-3}, changing target .01 decisions from $0/24$ to $24/24$. Those gains exceed selection cost but occur at impractical LLM-call budgets. Retaining all candidate outcomes is free only relative to the observed full-vector interface; feature storage and propagation computation still increase. Table 2 also gives every mean-bound design, including the smallest-budget loss.

7.2 Statistical slack and direct testing remain strong competitors

We use the learned dependent-output benchmark as a check on whether the method always improves the bound: 18 separately fitted joint selectors, six per corruption regime, with five independent tests per model. At 2,048 calls and $H = 256$, the low-corruption sampled row-tail bound is 0.25, exact risk 0.05, and oracle-tail bound 0.07. Thus 0.19 of 0.20 slack—about 92 percent—is sampling excess. The propagation-only improvement over the count safeguard is 5.90×10^{-3}, not an explanation for the bulk of the gap. Appendix K.10 describes the learned benchmark and its evaluation protocol.

Full-joint likelihood and cell-box regions change rank with budget; a richer retained output law is not automatically statistically worse. In the capacity task, range-aware direct testing is tighter at $H = 1024$ and local reuse is tighter at $H = 4096$. Consequently richer statistics are useful only if their information gain survives calibration cost. When statistical slack dominates, an affordable direct test or a defensible reduced model can be preferable to more elaborate propagation. The full capacity and likelihood controls remain in the appendix rather than competing for main-text space.

8 Discussion

The framework is useful when its assumptions and budget are clear. It identifies missing continuation information, provides exact supported intervals at the first unresolved step, and can guide corrected selection among declared candidate statistics. It also says when those diagnostics are insufficient: a first width is not a long-horizon bound; a time-varying optimum is not a frozen optimum; and millions of resets are not a practical LLM test.

The supported criterion identifies which admissible row perturbations can change risk when transition zeros are known. The stationary enclosures measure that uncertainty at deployment horizons, and corrected feature selection tests a testing choice without supplying the influential bit. Together, these results turn an exact-identification statement into a set of quantitative checks, rather than a promise that every richer test is better. For unrestricted unknown compactors, learning such an abstraction remains outside the demonstrated results. Deployment therefore requires a justified finite interface or small-block contract, with direct testing retained as a serious alternative.

AI USE STATEMENT

In this work, we used generative AI tools to draft sections of the paper and polish writing. Additionally, we used generative AI tools for editing code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shipra Agrawal, Yichuan Ding, Amin Saberi, and Yinyu Ye. Price of correlations in stochastic optimization. *Operations Research*, 60(1):150–162, 2012. doi: 10.1287/opre.1110.1011. URL https://pubsonline.informs.org/doi/10.1287/opre.1110.1011.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://gsbpreserve.stanford.edu/view/32366/learn-then-test-calibrating-predictive-algorithms-to-achieve-risk-control.

Pranav Ashok, Jan Křetínský, and Maximilian Weininger. PAC statistical model checking for Markov decision processes and stochastic games. In *Computer Aided Verification*, volume 11561 of *LNCS*, pp. 497–519, 2019. doi: 10.1007/978-3-030-25540-4_29. URL https://doi.org/10.1007/978-3-030-25540-4_29.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441. URL https://web.stanford.edu/~boyd/cvxbook/.

Marcilia Andrade Campos, Gracaliz Pereira Dimuro, Antonio Carlos da Rocha Costa, and Vladik Kreinovich. Computing 2-step predictions for interval-valued finite stationary Markov chains. Technical Report UTEP-CS-03-20a, University of Texas at El Paso, 2004. URL https://scholarworks.utep.edu/cs_techrep/382/.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory, 2025. URL https://arxiv.org/abs/2504.19413.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://doi.org/10.1093/biomet/26.4.404.

Ashwin Gerard Colaco and Nada Lahjouji. What to keep, what to forget: A rate–distortion view of memory compaction in LLMs and agents, 2026. URL https://arxiv.org/abs/2607.08032.

Gert de Cooman, Filip Hermans, and Erik Quaeghebeur. Imprecise Markov chains and their limit behavior. *Probability in the Engineering and Informational Sciences*, 23(4):597–635, 2009. doi: 10.1017/S0269964809990039. URL https://biblio.ugent.be/publication/498502.

Zihao Deng, Siddartha Devic, and Brendan Juba. Polynomial time reinforcement learning in factored state MDPs with linear value functions. In *International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 11280–11304, 2022. URL https://proceedings.mlr.press/v151/deng22c.html.

Norman Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Uncertainty in Artificial Intelligence*, pp. 162–169, 2004. URL https://arxiv.org/abs/1207.4114.

Robert Givan, Sonia Leach, and Thomas Dean. Bounded-parameter Markov decision processes. *Artificial Intelligence*, 122(1–2):71–109, 2000. doi: 10.1016/S0004-3702(00)00047-3. URL https://www.sciencedirect.com/science/article/pii/S0004370200000473.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003. doi: 10.1016/S0004-3702(02)00376-4. URL https://www.sciencedirect.com/science/article/pii/S0004370202003764.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5541–5552, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3bb585ea00014b0e3ebe4c6dd165a358-Abstract.html.

Christian Hensel, Sebastian Junges, Joost-Pieter Katoen, Tim Quatmann, and Matthias Volk. The probabilistic model checker Storm. *International Journal on Software Tools for Technology Transfer*, 24(4):589–610, 2022. doi: 10.1007/s10009-021-00633-z. URL https://www.stormchecker.org/publications.html.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991. URL https://arxiv.org/abs/1810.08240.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005. doi: 10.1287/moor.1040.0129. URL https://pubsonline.informs.org/doi/10.1287/moor.1040.0129.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376, 2023. doi: 10.18653/v1/2023.emnlp-main.825. URL https://aclanthology.org/2023.emnlp-main.825/.

Marta Kwiatkowska, Gethin Norman, and David Parker. PRISM 4.0: Verification of probabilistic real-time systems. In *Computer Aided Verification*, volume 6806 of *LNCS*, pp. 585–591, 2011. doi: 10.1007/978-3-642-22110-1_47. URL https://eprints.gla.ac.uk/56412/.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. A human-inspired reading agent with gist memory of very long contexts. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26396–26415, 2024. URL https://proceedings.mlr.press/v235/lee24c.html.

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

William A. Massey. Stochastic orderings for Markov processes on partially ordered spaces. *Mathematics of Operations Research*, 12(2):350–367, 1987. doi: 10.1287/moor.12.2.350. URL https://pubsonline.informs.org/doi/10.1287/moor.12.2.350.

Guanghui Min, Liang Wu, Mayank Darbari, Chen Chen, and Liangjie Hong. Toward reliable context compression for long-horizon agents: An empirical study of execution instability, 2026. URL https://arxiv.org/abs/2608.06503.

Jesse Mu, Xiang Lisa Li, and Noah Goodman. Learning to compress prompts with gist tokens, 2023. URL https://arxiv.org/abs/2304.08467.

Joseph Naor and Moni Naor. Small-bias probability spaces: Efficient constructions and applications. *SIAM Journal on Computing*, 22(4):838–856, 1993. doi: 10.1137/0222053. URL https://epubs.siam.org/doi/10.1137/0222053.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216. URL https://people.eecs.berkeley.edu/~elghaoui/pubs_rob_mdp.html.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems, 2023. URL https://arxiv.org/abs/2310.08560.

Jim Pitman and Wenpin Tang. Tree formulas, mean first passage times and Kemeny's constant of a Markov chain. *Bernoulli*, 24(3):1942–1972, 2018. doi: 10.3150/16-BEJ916. URL https://arxiv.org/abs/1603.09017.

Pascal Poupart and Craig Boutilier. Value-directed compression of POMDPs. In *Advances in Neural Information Processing Systems*, volume 15, 2002. URL https://papers.neurips.cc/paper_files/paper/2002/hash/14ea0d5b0cf49525d1866cb1e95ada5d-Abstract.html.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling, 2019. URL https://arxiv.org/abs/1911.05507.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Daniel Rudolf and Nikolaus Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, 24(4A):2610–2639, 2018. doi: 10.3150/17-BEJ938. URL https://research.tilburguniversity.edu/en/publications/perturbation-theory-for-markov-chains-via-wasserstein-distance/.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pp. 307–318. Princeton University Press, 1953. doi: 10.1515/9781400881970-018.

Marnix Suilen, Thom Badings, Eline M. Bovy, David Parker, and Nils Jansen. Robust Markov decision processes: A place where AI and formal methods meet, 2024. URL https://arxiv.org/abs/2411.11451.

Lev V. Utkin and Sébastien Destercke. Computing expectations with continuous p-boxes: univariate case. *International Journal of Approximate Reasoning*, 50(5):778–798, 2009. doi: 10.1016/j.ijar.2009.02.004. URL https://arxiv.org/abs/0906.1260.

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

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory, 2024. URL https://arxiv.org/abs/2410.10813.

Xixi Wu, Kuan Li, Yida Zhao, Liwen Zhang, Litu Ou, Huifeng Yin, Zhongwang Zhang, Xinmiao Yu, Dingchu Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Minhao Cheng, Shuai Wang, Hong Cheng, and Jingren Zhou. ReSum: Unlocking long-horizon search intelligence via context summarization, 2025. URL https://arxiv.org/abs/2509.13313.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents, 2025. URL https://arxiv.org/abs/2502.12110.

Yuriy Zacchia Lun, Jack Wheatley, Alessandro D'Innocenzo, and Alessandro Abate. Approximate abstractions of Markov chains with interval decision processes. *IFAC-PapersOnLine*, 51(16):91–96, 2018. doi: 10.1016/j.ifacol.2018.08.016. URL https://www.sciencedirect.com/science/article/pii/S2405896318311297.

A Detailed positioning and a reading guide

The continuation-preservation direction is established value-directed compression and value-equivalence reasoning (Poupart & Boutilier, 2002; Grimm et al., 2020). Theorem 2 supplies a relative-interior converse on declared row supports, for the vector of risks over all initial states. Appendix D gives the full-support fixed-law statement and its small-perturbation witness; the witness need not provide a useful lower bound near support boundaries. Proposition 12 gives supported extrema at the first unresolved horizon, using ordinary rowwise linear optimization. The long-horizon analysis shows that this first width is not a deployment-horizon bound. The finite-family selection experiment tests an operational consequence with simultaneous correction, not a general feature-discovery algorithm. The mathematical tools are elementary or established; the claimed contribution is the characterization of retained test information together with quantitatively checked limits, data-selected interventions, and a complete accounting of when the procedure is useful.

The input-specific expectation is *not* a new mathematical primitive: it is a finite nested-random-set upper expectation, exactly the weighted-supremum construction treated by Utkin & Destercke (2009). Its subsequent recursion is established rectangular robust DP (Nilim & El Ghaoui, 2005; Iyengar, 2005). Proving the specialization, attainment, and its comparison to a particular same-data safeguard does not make those general ingredients new. Our relaxation chain, analytic gap examples, finite-sample oracle diagnostics, and mismatch accounting explain what these computations mean for a frozen compactor. The existence of a stationarity gap itself predates the paper (Campos et al., 2004; de Cooman et al., 2009; Suilen et al., 2024).

Memory systems are motivation, not experimental baselines here. MemGPT manages a hierarchy of context and external memory (Packer et al., 2023); A-MEM links and evolves memory items (Xu et al., 2025); Mem0 extracts and consolidates persistent memories (Chhikara et al., 2025). ReSum uses recurring summaries for long-horizon search (Wu et al., 2025). Compressive Transformers compress older representations (Rae et al., 2019); gist tokens and LLMLingua compress prompts using learned or selection-based mechanisms (Mu et al., 2023; Jiang et al., 2023). A reading agent can revisit source material after consulting a gist (Lee et al., 2024). Access to an archive changes the information available for repair and must be charged or disclosed. Our experiments without an application programming interface (API) do not reproduce these systems or claim to outperform them.

LoCoMo and LongMemEval evaluate downstream memory behavior on conversational histories (Maharana et al., 2024; Wu et al., 2024). The rate–distortion perspective already treats repeated compaction explicitly (Colaco & Lahjouji, 2026), and the paired-continuation framework evaluates compaction through paired continuations rather than surface similarity alone (Min et al., 2026). Our difference is a finite-sample statement about an explicitly restricted stochastic process. A synthetic theorem does not supersede an empirical language-agent benchmark; they support different claims. Likewise, the feedback of earlier outputs into later inputs is not new: it is central in sequential prediction and imitation learning (Ross et al., 2011).

Robust dynamic programming and probabilistic verification. Bounded-parameter Markov decision processes (MDPs) represent transition uncertainty (Givan et al., 2000). Robust dynamic programming exploits rowwise or rectangular uncertainty sets (Iyengar, 2005; Nilim & El Ghaoui, 2005). Our input-specific tail update is the nested-random-set upper expectation for a particular collection of output events (Utkin & Destercke, 2009). Its horizon-wise sharpness follows the standard rectangular Bellman principle; it is not generally sharp among stationary kernels satisfying all record-level constraints. The ordered theorem instead constructs one greatest kernel because its additional stochastic monotonicity assumption permits a pathwise comparison, an established method (Massey, 1987).

Probably approximately correct (PAC) statistical model checking provides high-confidence reachability analysis of unknown probabilistic systems (Ashok et al., 2019). The probabilistic symbolic model checker (PRISM) and Storm are established verification systems (Kwiatkowska et al., 2011; Hensel et al., 2022). We implement tiny exact models and sampled test bounds directly for controlled comparisons; we do not report experiments run in either tool. A generic empirical transition matrix plus interval propagation would not by itself establish a distinct contribution. The relevant

choices here are smaller retained output statistics, input-pattern separation, and solved information gaps.

State abstraction and structural assumptions. State equivalence in MDPs formalizes when different concrete states can share an abstract future law (Givan et al., 2003). Bisimulation metrics quantify approximate rather than exact equivalence (Ferns et al., 2004); interval decision-process abstractions provide conservative approximations of Markov chains (Zacchia Lun et al., 2018). Count-lumpability is justified only for the declared symmetric generators. The input-specific method drops count-lumpability but still assumes transfer of the pattern-conditional output-count law. It does not infer a sufficient abstraction from finite data. Factored-state learning can exploit locality without conditional-output independence under different value-function assumptions (Deng et al., 2022); our explicit small-block method makes no comparable polynomial-in-global-state theorem.

Optimization, coverage, and confidence. The marginal cap is a finite linear program (LP) and uses standard duality (Boyd & Vandenberghe, 2004). Worst-case expectations under constrained marginals and the cost of ignoring dependence have a substantial literature (Agrawal et al., 2012). Parity supplies a classical limited-independence example (Naor & Naor, 1993); we use it to compute an absorbing system's exact long-horizon risk range, not as a new random-variable construction.

Our fixed-design calibration uses Clopper–Pearson (CP) endpoints (Clopper & Pearson, 1934); a simpler sufficient error rate follows from Hoeffding concentration (Hoeffding, 1963). The perturbation argument tracks a nonlinear cap's estimation error rather than the full exact dynamics, while contraction-based transition perturbation bounds are established (Rudolf & Schweizer, 2018). Post-training risk calibration and careful multiplicity control also underlie Learn then Test (Angelopoulos et al., 2025). For output-dependent testing, beta-mixture confidence sequences and betting-based inference are established (Howard et al., 2021; Waudby-Smith & Ramdas, 2024). Our two-clock statement connects a simultaneous response-table event to every deployment horizon; it is not a new martingale inequality or an optimal adaptive allocation result.

How to read the remaining appendices. Appendix D proves frozen moment identification. Appendix E isolates the relaxation gaps and explains feasible stationary brackets. Appendices F–H cover bounded mismatch, likelihood-region calibration, and the record-capacity limit. Subsequent appendices preserve the baseline marginal, parity, ordered-count, and sequential-confidence proofs, along with the exact learned compactor specification. Appendix I gives the quantitative, augmentation, and independently gated-mismatch bridges. Appendix L contains extended diagnostics.

B Supported identification and deployment-horizon diagnostics

B.1 Relative-interior version of the continuation criterion

Let $\mathcal{C}_{M,S} = \{P \geq 0 : P\mathbf{1} = \mathbf{1}, PF = M,\ P(e,y) = 0 \text{ for } y \notin S_e\}$ be nonempty. Remove any coordinates forced to zero by all feasible rows. Thus S_e is the effective allowed support, not a set inferred by declaring unobserved transitions impossible. Relative interior then means $K(e,y) > 0$ for all $y \in S_e$ and $K(e,y) = 0$ otherwise. Define

$$\mathcal{V}_S = \{v :\ v|_{S_e} \in \mathrm{col}(F|_{S_e}) \text{ for every input } e\}.$$

The representing coefficients may differ between rows. In particular, $\mathrm{col}(F) \subseteq \mathcal{V}_S$, with equality under full support; structural zeros can make the inclusion strict.

Proof of Theorem 2. For every $v \in \mathcal{V}_S$ there are coefficients a_e with $v|_{S_e} = F|_{S_e} a_e$. Hence $(Pv)(e) = M_e a_e$ is the same for every compatible P. Induct on j, starting with $v_0 = L$. If $v_j = K^j L \in \mathcal{V}_S$ for $j < H$, then $P^{j+1}L = Pv_j = Kv_j = v_{j+1}$ for every compatible P. This proves vector equality for every initial state.

Conversely, suppose j is the first failed membership, at row e. Project $v_j|_{S_e}$ onto the orthogonal complement of $\mathrm{col}(F|_{S_e})$, obtaining $r \neq 0$. Extend r by zero outside S_e. The constant feature gives $r^\top \mathbf{1} = 0$, and $r^\top F = 0$. For a sufficiently small $a > 0$, replacing row e of K by $K_e \pm a r^\top$ stays in the relative interior. All earlier vector continuations are shared by the sufficient direction, but the

next risk from state e differs by $2a\|r\|_2^2$. Therefore vector identification fails at $j+1$. The argument uses only the tangent space of the active feasible face, not unallowed transitions. It is the standard perturbation idea specialized to this inverse problem. □

For a *fixed* initial law, supported necessity does not follow: a failed row may be unreachable. For example, take an absorbing state 0 and a separate two-state component, test only normalization, and assign losses 0 and 1 in the second component. All risks from state 0 are identified, but not the risk vector. We state the supported theorem for all initial states and keep the full-support fixed-law corollary separate. These qualifications avoid converting a sufficient condition for a fixed deployment into a false necessity claim.

B.2 A ZERO FIRST WIDTH CAN HIDE LATER AMBIGUITY

Consider the following two stationary kernels on $00, 10, 01, 11$, with $L = (0, 1/2, 1/2, 1)^\top$:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Let $J = \mathbf{1}\mathbf{1}^\top/4$ and $A_\zeta = (1-\zeta)A + \zeta J$, $B_\zeta = (1-\zeta)B + \zeta J$. Every proper positive $\zeta < 1$ gives strictly positive matrices. Their complete count-class tables agree. Moreover $A_\zeta L = B_\zeta L = (1/2, 1/2, 1/2, 1-\zeta/2)^\top$, which is count-measurable, so *every* compatible kernel agrees through two steps. Thus $W_2 = 0$, not merely a zero gap for this pair. At $\zeta = .001$, direct rational matrix powers give risks 0.52 and $1 - 1.87 \times 10^{-3}$ at $H = 128$, a stationary witnessed gap 0.48. This disproves any universal multiplicative bound $W_H \le c_H W_2$. It also holds away from structural zeros. The matrices and powers are disclosed, and no optimizer is required. In longer-delay constructions the same issue can be postponed to subsequent horizons; we do not use W_2 as a deployment-horizon certificate.

B.3 CERTIFIED STATIONARY BRACKETS FOR ALL 330 GEOMETRIES

For the family in Eq. (24), every count-compatible matrix has independent row split parameters $x_e \in [0, 1]$:

$$P_x(e, \cdot) = (q_e(0),\ q_e(1)x_e,\ q_e(1)(1-x_e),\ q_e(2)).$$

The optional declared zero at $b = 1$ fixes $x_0 = 1$; otherwise there are four free splits. There are only 66 distinct unrestricted count tables in the 330 reference geometries. The five imbalances change the reference kernel and residual bound, not that test table. We retain all 330 rows for comparisons, and do not attach sampling p-values to their deterministic correlations.

Stationary limits without an unverified global optimizer. All matrices in this family have a unique stationary distribution. Every row has positive probability of reaching each outer state, and the outer states have self-loops; any unreachable inner state is transient. By the Markov chain tree formula (Pitman & Tang, 2018), invariant risk is a ratio of two multiaffine polynomials in the row splits, with strictly positive denominator. With the other coordinates fixed it is linear-fractional in one split and attains its minimum and maximum at an endpoint (or is constant). Moving coordinates successively to endpoints proves that the invariant extrema occur at the 16 corners, or eight corners with $x_0 = 1$. Each stationary corner distribution is solved using exact rational Gaussian elimination. A separate occupancy-measure LP verifies the implementation on test cases; the proof does not depend on that LP solver.

From invariant risk to finite-horizon risk. Write $\tau(P) = \frac{1}{2}\max_{i,j}\sum_y |P(i,y) - P(j,y)|$. If $\tau(P^4) \le \alpha < 1$ for every compatible frozen P, then

$$|\mu P^H L - \rho_P L| \le \alpha^{\lfloor H/4 \rfloor}\,\mathrm{span}(L),$$

by the usual total-variation contraction. To certify a uniform α, express each entry of P_x^4 in a tensor Bernstein basis of degree four in each split. Its coefficients are computed from the 5^4 tensor grid by an exact-rational interpolation transform with outward-rounded interval evaluation. A polynomial is

bounded by its Bernstein coefficients because the basis is nonnegative and sums to one. Apply this to every difference $P_x^4(i, A) - P_x^4(j, A)$ for all state pairs and all 16 output subsets A. The largest coefficient upper endpoint bounds $\tau(P_x^4)$ simultaneously for the entire split cube. This is a small polynomial-enclosure calculation, not a new general Markov contraction theorem.

The invariant extrema plus this uniform remainder give finite-horizon *outer* extrema. Evaluating the corner kernels at the requested horizon gives feasible stationary *inner* witnesses. Independently outward-rounded exact-count rectangular envelopes can further narrow the outer interval. Thus, writing extrema m_H, M_H, we report

$$m_H^{\text{out}} \le m_H \le m_H^{\text{wit}}, \quad M_H^{\text{wit}} \le M_H \le M_H^{\text{out}}.$$

The stationary width lies between $M_H^{\text{wit}} - m_H^{\text{wit}}$ and $M_H^{\text{out}} - m_H^{\text{out}}$. We do not label finite-horizon corner search an exact optimizer. At $H = 128$, the largest gap between these width brackets over all 330 unrestricted geometries is below $6.98\,10^{-5}$; at larger reported horizons it is smaller. The analysis evaluates both support conventions and all horizons $2, 8, 32, 128, 256, 512, 1024$. The two-step and deployment-horizon comparisons reuse each reference kernel from Appendix I.2.

B.4 Relaxation attribution is a convention, not a unique causal split

Let $a = S_=(T)$, $b = S_\le(T)$, $c = D_=(T)$, and $d = D_\le(T)$. The two information/temporal paths assign $(b - a, d - b)$ and $(d - c, c - a)$ to the information/temporal increments. Their difference is the interaction $d - b - c + a$. There is no canonical order induced by the certificate. If a symmetric numerical attribution is desired, average both orders:

$$\Phi_{\text{one-sided}} = \tfrac{1}{2}[(b - a) + (d - c)], \qquad \Phi_{\text{temporal}} = \tfrac{1}{2}[(c - a) + (d - b)].$$

Both are nonnegative by set inclusion and sum to $d - a$. This is the two-player Shapley average (Shapley, 1953), not an intrinsic causal quantity. For the one-bit $H = 2$ example, $(a, b, c, d) = (0, 1/4, 0, 1)$. The first path assigns $(1/4, 3/4)$, the other $(1, 0)$, and the symmetric average $(5/8, 3/8)$. The comparison changes no valid upper bound.

C Choosing a tested feature from data

Proof of Proposition 4. There are Rk count-tail parameters and Rm means of fixed candidate functions $b_1, \dots, b_m : \mathcal{E} \to \{0, 1\}$. A fixed design with independent resets permits one-sided binomial endpoints at level $\delta/[R(k + m)]$. A union bound gives a single event of probability at least $1 - \delta$ covering every parameter; no independence between statistics computed from the same output is needed. On this event the true row law belongs to each candidate's region. Backward conditional expectation yields every candidate bound at every deployment horizon. Thus both selecting a candidate from the data and taking the minimum of candidate bounds at a requested horizon preserve the same confidence event. This does not cover candidates themselves invented from these observations, uncorrected selection across distinct confidence constructions, or arbitrary optional stopping of the reset count. A declared finite family is required. Confidence sequences supply the separate extension to optional test stopping. □

A hidden influential record, not a supplied answer. The evaluator chooses J uniformly among k records using a seed independent of the test, and sets $p(e) = .003 + .05e_J + .08L(e)$. Conditional on e, the k output bits are independent Bernoulli($p(e)$). Neither J nor this function is passed to the certifier. The certifier receives only the input-indexed count histograms and candidate bit totals from the same outputs. It evaluates all k candidate bounds and chooses the smallest from the initially correct state. This is selection within a predeclared bit family; the generator still guarantees that this family contains a sufficient candidate. It is not discovery of arbitrary nonlinear memory features.

At $k = 4, 8, 12$ and $n = 128, 512, 2048, 8192$ resets per concrete row we use 24 independently generated test datasets, with the hidden index independently sampled across replicates. The horizons are $2, 8, 32, 128$ and the targets are $.005, .01, .02, .05, .08$. Hidden-index assignments are reused across budgets for each block size and replicate; within-budget comparisons are paired. The native count baseline spends $\delta/(Rk)$ per tail; the selected-feature family spends $\delta/(2Rk)$ per tail and per candidate bit. We also retain the count-only bound from the shared, more conservative tail table. Therefore an improvement over the native baseline must exceed the multiple-comparison cost. Each reported paired interval uses the independent test datasets, not the candidate features as replicates.

Exact same-output sufficient statistics. The generator draws the multinomial histogram of output counts. Conditional on a count c, its c one-bits form a uniform subset. It assigns bits sequentially by splitting the remaining groups with binomial probability $r/(k-j)$, where r ones remain to be placed before bit j. The resulting count histogram and all bit totals have exactly the law of observing the same n full vectors; candidate bits are not sampled as independent test batches. The identity $\sum_j S_j = \sum_c cN_c$ is checked on every dataset. Computational compression does not reduce the charged primitive-call budget.

Assumption–cost controls. Count pooling is *invalid* for the hidden-J family. For an explicit exchangeable companion $p(e) = .003 + .13L(e)$, a declared count-lumpable monotone kernel needs only $k+1$ reset rows. We run 100 tests at budgets $2{,}048, 8{,}192, 32{,}768, 2{,}097{,}152$, distributing calls as evenly as possible among the 13 counts. Its initially correct true risk equals that of the hidden-J family, but it is a different, more symmetric process. The two-input control assumes a shared nonnegative affine *mean* mechanism, not count-lumpability, and estimates its unknown numerical parameters. These assumption sets are not ordered universally. Table 1 distinguishes these assumptions; pooling is applied only to the exchangeable companion.

D WHAT EXACT OUTCOME MOMENTS IDENTIFY FOR A FROZEN COMPACTOR

This result concerns an *inverse* question. A transition kernel K is fixed, but a test retains only the rowwise expectations KF of selected outcome functions. Which finite-horizon risks are then common to every compatible frozen kernel? Preservation of continuation spaces is established in value-directed compression and value equivalence (Poupart & Boutilier, 2002; Grimm et al., 2020). The sufficiency argument below is that familiar argument. We also need its converse for an interior kernel, an explicit stationary separating pair, and a dimension statement. The theorem does not claim a new principle of Krylov subspaces, nor does it identify a useful sufficient state from data.

Theorem 5 (Exact frozen-risk identification)**.** *Let K be strictly positive and stochastic, F include $\mathbf{1}$, and μ be any fixed initial law. All $\mu P^h L$, $1 \le h \le H$, agree across $P \in \mathcal{C}_{KF}$ if and only if*

$$L, KL, \dots, K^{H-1}L \in \operatorname{col}(F). \tag{9}$$

If j is the first failed membership, two positive compatible stationary kernels agree through j but differ at $j+1$. For $r = K^jL - \Pi_F K^jL$, $\mu_e > 0$, and $0 < a < \min_y K(e,y)/\|r\|_\infty$, an attainable gap is $2a\mu_e\|r\|_2^2$.

Proof of Theorem 5. Write $\mathcal{V} = \operatorname{col}(F)$ and $v_j = K^jL$. The constant vector belongs to $\mathcal{V}$, so the condition $PF = KF$ includes row normalization whenever P is constrained to be nonnegative. Suppose first that $v_j \in \mathcal{V}$ for $0 \le j < H$. For any compatible P, induction gives $P^jL = v_j$: if $v_j = Fa_j$, then

$$P^{j+1}L = PFa_j = KFa_j = Kv_j = v_{j+1}.$$

Consequently every vector P^hL, and hence every specified initial-law risk μP^hL, agrees with its value under K through H. This part does not require positivity.

For necessity, let $j < H$ be the first index with $v_j \notin \mathcal{V}$. Orthogonally project onto $\mathcal{V}$ and put $r = v_j - \Pi_{\mathcal{V}} v_j$. Then $r \neq 0$, $r^\top F = 0$, $r^\top \mathbf{1} = 0$, and $r^\top v_j = \|r\|_2^2$. Choose a state e with $\mu_e > 0$, and any

$$0 < a < \frac{\min_y K(e,y)}{\|r\|_\infty}.$$

Strict positivity of K makes such a choice possible. Define two *stationary* kernels

$$P^\pm = K \pm a\,\mathbf{e}_e r^\top. \tag{10}$$

They are nonnegative, have row sum one, and satisfy $P^\pm F = KF$. The sufficiency induction through j uses only the earlier vectors and gives $(P^\pm)^j L = v_j$. At the next step,

$$\mu(P^+)^{j+1}L - \mu(P^-)^{j+1}L = 2a\mu_e\|r\|_2^2 > 0. \tag{11}$$

This contradicts identification even for the one fixed initial law μ. Choosing $a = \min_y K(e,y)/(2\|r\|_\infty)$ gives the explicit attainable gap $\mu_e \min_y K(e,y)\|r\|_2^2/\|r\|_\infty$. The result is about a pair of frozen kernels; neither hidden time variation nor a one-sided confidence region is used. □

What happens on the boundary? Positivity is essential for the converse. For example, if the tested next-count law from a state is concentrated on a singleton concrete outcome, support alone determines that row. A signed null-space perturbation may then leave the nonnegative simplex. The theorem makes no necessity claim at such boundary kernels. It is also an *exact-identification* result. A nonzero witness can be arbitrarily small near an invariant subspace; failure of exact identification does not imply that every conservative bound is useless.

Corollary 6 (How many exact outcome statistics are sufficient?)**.** *At an interior kernel, the minimum number of nonconstant linear outcome statistics identifying all risks through H is*

$$\operatorname{rank}[\mathbf{1}, L, KL, \ldots, K^{H-1}L] - 1. \tag{12}$$

For fixed nonconstant L and almost every positive stochastic K, this number is $\min(H, N-1)$. Thus a k-bit block can require $\min(H, 2^k - 1)$ such statistics, even with noiseless tests.

Corollary 7 (A discrete-summary obstruction at two steps)**.** *Fix a nonconstant loss L. For almost every positive K, no noninjective deterministic output summary $\psi : \mathcal{E} \to \{1, \ldots, m\}$ identifies all risks through horizon two, even if the exact conditional distribution of $\psi(E_{t+1})$ is known at every input. This holds even when the partition is chosen with knowledge of K.*

Proof of Corollary 6. By the theorem, a sufficient tested space must contain $\operatorname{span}\{\mathbf{1}, L, KL, \ldots, K^{H-1}L\}$, and that space itself suffices. Excluding the known constant therefore gives the stated minimum dimension. This is an oracle characterization at a specified kernel, not an algorithm that finds the space without learning about that kernel.

For the generic statement, fix a nonconstant $L \in \mathbb{R}^N$. The rank of the displayed matrix is at most $\min(N, H+1)$. It suffices to exhibit one strictly positive stochastic matrix attaining that rank: a nonzero maximal minor is then a nonzero polynomial in the free row entries of K, whose zero set has Lebesgue measure zero in the row-stochastic affine space. Let $J = \mathbf{1}\mathbf{1}^\top/N$. Choose an orthonormal basis $b_1, \ldots, b_{N-1}$ of $\mathbf{1}^\perp$ such that the projection of L onto every b_i is nonzero. Such bases exist because the centered L is nonzero, and each zero projection excludes only a proper subset of the orthogonal group. Choose distinct real λ_i and sufficiently small positive ϵ, and set

$$K = J + \epsilon \sum_{i=1}^{N-1} \lambda_i b_i b_i^\top.$$

Its row sums are one and all entries are positive when ϵ is small enough. After projecting away the constant coordinate, the columns $L, KL, \ldots, K^{H-1}L$ form a Vandermonde system with nonzero row multipliers and distinct eigenvalues $\epsilon\lambda_i$. Their projected rank is $\min(H, N-1)$. Including $\mathbf{1}$ gives $\min(H+1, N)$, proving the claim. For $N = 2^k$, the required number of nonconstant functions is generically $\min(H, 2^k - 1)$; it is exponential in k only once the range of required horizons is comparably long. This is not a sample-complexity lower bound for approximate risk bounds. □

Proof of Corollary 7 and the count-transcript consequence. For fixed nonconstant L, every equality $(KL)_e = (KL)_f$ with $e \neq f$ is a nontrivial linear equation on the row-stochastic affine space: the two rows can independently change their expectations of L. Thus, outside a finite union of measure-zero sets, the entries of KL are pairwise distinct. For any noninjective partition ψ, let F_ψ contain its output-class indicators. If $L \notin \operatorname{col}(F_\psi)$, Theorem 5 gives failure at horizon one. If it belongs, KL cannot belong because it is not constant on any merged class; the same theorem gives failure at horizon two. One generic pairwise-distinctness event works for *all* partitions, including partitions selected after K is known.

For count features, equality $P^+F = P^-F$ means equality of the entire next-count distribution, not just its mean. At every selected reset row, the retained observation law is identical under the two kernels. Induction over adaptive row choices and internal evaluator randomization proves equality of the full count-transcript distributions. Let their risks be $r_- < \tau < r_+$ and let $U(\mathcal{D})$ have uniform upper coverage at least $1 - \delta$. Under the high-risk kernel, $\mathbb{P}\{U \leq \tau\} \leq \delta$. The same transcript distribution gives exactly the same probability under the low-risk kernel. This argument allows any number of count-based reset queries and count-dependent stopping; it does not apply to an evaluator that uses discarded output orientation to choose later inputs or encode additional information in its retained transcript. For general real-valued moments, equal expectations need not imply equal observation laws, which is why this transcript conclusion is restricted to complete class/count distributions. □

An approximate version and its limitation. Suppose $\|(P-K)F\|_{\max} \le \eta$ for two frozen kernels, and write $K^j L = Fa_j + r_j$. Because the difference between expectations of r_j under two probability laws is at most $\mathrm{span}(r_j)$,

$$\|(P-K)K^j L\|_\infty \le \eta\|a_j\|_1 + \mathrm{span}(r_j).$$

The telescoping identity for $P^H - K^H$ and stochastic sup-norm nonexpansivity imply

$$\|P^H L - K^H L\|_\infty \le \sum_{j=0}^{H-1} \{\eta\|a_j\|_1 + \mathrm{span}(r_j)\}.$$

Thus approximate invariance can suffice for useful finite-horizon accuracy. Poorly conditioned feature coefficients can, however, amplify moment estimation error. We use this as interpretation, not as a claim of an optimal finite-sample algorithm.

E Separating four sources of conservatism

Fix an actual frozen kernel K, initial law μ, bounded terminal loss L, and exact row count laws $q_e(c)$. Write $T_e(\ell) = \sum_{c \ge \ell} q_e(c)$. Let $S_=(T)$ be the supremum of $\mu P^H L$ over *stationary* kernels with these exact count laws, and $S_\le(T)$ the stationary supremum using only upper tails. Let $D_\le(T) = \mu \mathcal{B}_T^H L$ be the rectangular value, and define $D_\le(U)$ analogously for a finite-sample upper table U. On the confidence event $U \ge T$,

$$\mu K^H L \le S_=(T) \le S_\le(T) \le D_\le(T) \le D_\le(U). \tag{13}$$

Each inclusion follows by enlarging the admissible model class. The four consecutive increments respectively measure discarded output orientation, discarded lower-tail constraints, relaxing a frozen law to time-varying choices, and finite-sample upper uncertainty. This is a diagnostic decomposition, not a method for globally solving the stationary optimization. The time-varying exact-count value $D_=(T)$ is separately computable as

$$(\mathcal{B}_= v)(e) = \sum_c q_e(c) \max_{|y|=c} v(y).$$

It lies above $S_=(T)$ and below $D_\le(T)$, but need not be ordered with $S_\le(T)$.

E.1 A one-record frozen gap at every horizon

Proposition 8 (A quantitative frozen–rectangular separation)**.** *Start a one-record process at zero, with exact tested upper tails $T_0 = 1, T_1 = 0$ and terminal loss $L(e) = e$. Its rectangular bound is one at every H. Compatible frozen kernels have parameter $x \in [0,1]$ and risk*

$$r_H(x) = \frac{x}{1+x}\{1 - (-x)^H\}. \tag{14}$$

At $H = 2$, the frozen worst case is $1/4$, so the temporal relaxation alone adds $3/4$. At odd horizons the gap is zero; along even horizons the frozen optimum tends to $1/2$.

Proof of Proposition 8. Use states $0, 1$, loss $(0, 1)$, initial state zero, and upper tails $T(0) = 1, T(1) = 0$. Every compatible stationary matrix is

$$P_x = \begin{pmatrix} 1-x & x \\ 1 & 0 \end{pmatrix}, \qquad 0 \le x \le 1.$$

The risk recurrence is $r_{t+1} = x(1 - r_t)$, $r_0 = 0$, so

$$r_H(x) = \frac{x}{1+x}\{1 - (-x)^H\}. \tag{15}$$

A rectangular process stays at zero until the last step and then sets $x = 1$, attaining one for every H. If H is odd, the frozen choice $x = 1$ also attains one. For even H, differentiating (15) gives a derivative whose sign is that of $1 - (H+1)x^H - Hx^{H+1}$. This is strictly decreasing from one to a negative value, so its unique zero is the global maximum. At $H = 2$, $r_2(x) = x(1-x)$ has maximum $1/4$. For even H, the maxima tend to $1/2$: they are at most $x/(1+x) \le 1/2$, while for any fixed $x < 1$ their limit is $x/(1+x)$, which approaches $1/2$ as $x \uparrow 1$. The upper table can be the *exact* tail table of P_1; its true even-horizon risk is zero. There is only one state per count, hence no discarded orientation. At $H = 2$, the four gaps in (13) are $0, 1/4, 3/4, 0$. □

The general distinction between stationary and nonstationary interval chains, including two-step gaps, predates this paper (Campos et al., 2004; de Cooman et al., 2009; Suilen et al., 2024). This example supplies explicit fractions and all-horizon behavior for the test semantics; it is not the first observation of the distinction.

E.2 An exact-count gap without one-sided uncertainty

On states $a = 00, b = 10, c = 01, d = 11$, take $L = (0, 1/2, 1/2, 1)$ and start at a. Rows a, b always output count one, row c count two, and row d count zero. Every compatible stationary kernel has the form

$$P_{x,y} = \begin{pmatrix} 0 & 1-x & x & 0 \\ 0 & 1-y & y & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \qquad x, y \in [0, 1]. \tag{16}$$

Direct multiplication yields

$$(P_{x,y}^4 L)(a) = \{1 - (1-x)y^2\}/2.$$

Its global stationary maximum is $1/2$. A time-varying exact-count process follows a, b, b, c, d and attains one. No confidence interval or relaxation from equalities to inequalities is involved. The compatible nominal kernel $x = 0, y = 1$ follows a four-cycle and has risk zero at horizon four. An exhaustive numerical grid checks the polynomial but is not used to prove the global optimum.

E.3 Orientation and one-sided tails can each matter alone

For an orientation-only example, use the same states and loss at $H = 2$, with the actual deterministic map $a \mapsto c, b \mapsto d, c \mapsto a, d \mapsto d$. True risk is zero. The exact count constraints still permit $a \mapsto b$, producing risk one under a frozen map. Thus orientation alone can account for the entire unit gap. For a one-sided-only example, use $a \mapsto d, b \mapsto d, c \mapsto a, d \mapsto a$. The exact count laws determine all rows, so the true and exact-count stationary risks are zero at $H = 2$. Upper tails permit $a \mapsto b$ instead of d, followed by $b \mapsto d$, attaining one under a frozen map. Orientation and temporal relaxation contribute no gap in this example.

E.4 A residual identity and feasible stationary witnesses

Let $V_h = \mathcal{B}_U^h L$ and let P be any fixed compatible kernel. Its Bellman slacks are $d_h = V_h - PV_{h-1} \geq 0$. Telescoping gives the exact identity

$$\mu V_H - \mu P^H L = \sum_{t=0}^{H-1} \mu P^t d_{H-t}. \tag{17}$$

The right side is nonnegative term by term. Compactness and continuity imply that the stationary supremum is attained; the rectangular gap vanishes exactly when some stationary P makes every positively occupied state–time slack zero. In particular, a single row law that maximizes all needed continuation values is sufficient. This is the standard Bellman telescoping argument applied as a stationarity diagnostic, not a new performance-difference identity.

For the learned three-bit benchmark, we enumerate nine maps that choose one destination per exact count, and 224 maps that choose one destination from each nested count ball. With exact count probabilities $q_e(c)$, each map $c \mapsto y_c$ defines the frozen row law $P(e, y) = \sum_c q_e(c)\mathbf{1}\{y = y_c\}$. Exact-count maps require $|y_c| = c$; upper-tail maps require $|y_c| \leq c$. Every enumerated matrix is feasible and reused unchanged at all times. The maximum over these maps and the nominal matrix is a *lower witness* for the stationary supremum; the rectangular value is an upper bound. This enumeration does not cover every stationary kernel, so the reported brackets are not advertised as globally optimized stationary values. Numerical precision is reported separately from mathematical attainment.

F CONSERVATIVE TEST-TO-DEPLOYMENT MISMATCH

Theorem 9 (Externally bounded response mismatch)**.** *Suppose the tested tails are T and, for every deployment history ending at e,*

$$\mathbb{P}(\phi(E_{t+1}) \geq \ell \mid \mathcal{F}_t) \leq T_e(\ell) + \eta_e. \tag{18}$$

Here the nonnegative η_e are externally justified bounds. On $T \leq U$, propagate $\widetilde{U}_e(\ell) = \min(1, U_e(\ell) + \eta_e)$. Then $\mu\mathcal{B}_{\widetilde{U}}^H L$ bounds endpoint loss even with history-dependent deployment responses. For uniform inflation η,

$$0 \leq \mu(\mathcal{B}_{U+\eta}^H - \mathcal{B}_U^H)L \leq \eta \sum_{j=0}^{H-1} \operatorname{span}(\mathcal{B}_U^j L) \leq H\eta \operatorname{span}(L). \tag{19}$$

A separate terminal-loss mismatch of at most ξ is handled by adding ξ and clipping at one.

Proof of Theorem 9. On the test event $T_e(\ell) \leq U_e(\ell)$, the assumed full-history conditional inequality gives

$$\mathbb{P}(\phi(E_{t+1}) \geq \ell \mid \mathcal{F}_t) \leq \min\{1, U_{E_t}(\ell) + \eta_{E_t}\} = \widetilde{U}_{E_t}(\ell).$$

Thus every history-dependent output law belongs to the inflated row uncertainty set. Backward conditional expectation proves the endpoint bound exactly as for rectangular DP. If the actual conditional query failure probability is at most $L(E_H) + \xi$, add ξ and clip at one. Neither argument requires the abstract state to be exactly Markov or the actual law to be stationary; it requires the displayed uniform conditional mismatch bound. Without an externally justified η, this does not solve state insufficiency.

For the perturbation statement, uniform inflation by η raises every tightened tail by between zero and η. The increments $a_\ell - a_{\ell-1}$ are nonnegative and sum to at most $\operatorname{span}(v)$, so

$$0 \leq \mathcal{B}_{U+\eta} v - \mathcal{B}_U v \leq \eta \operatorname{span}(v)\mathbf{1}.$$

Both operators preserve constants, are monotone, and are nonexpansive in sup norm. Comparing the two recursions yields

$$\|\mathcal{B}_{U+\eta}^H L - \mathcal{B}_U^H L\|_\infty \leq \eta \sum_{j=0}^{H-1} \operatorname{span}(\mathcal{B}_U^j L) \leq H\eta \operatorname{span}(L).$$

The same reasoning applies when $0 \leq U - T \leq \epsilon$. It controls statistical or mismatch excess above an oracle *rectangular* bound, not all slack above true risk. A small full-transition total-variation bound implies the tail bound, but marginal count bounds do not in general bound full-transition total variation. □

A sufficient test budget, not an optimal rate. For R concrete inputs and k tail thresholds, n independent resets per row, and $a = \sqrt{\log(2Rk/\delta)/(2n)}$, simultaneous two-sided Hoeffding bounds give $|\widehat{T} - T| \leq a$. Then $U = \min(1, \widehat{T} + a)$ has $0 \leq U - T \leq 2a$ on the event. Therefore $n \geq 2H^2 \log(2Rk/\delta)/\varepsilon^2$ suffices for certificate excess at most ε when $\operatorname{span}(L) \leq 1$. If the oracle spans decay, their sum replaces H in this calculation. Total calls are Rn, and full pattern coverage has $R = 2^k$. This loose sufficient bound is included to make the block-size dependence explicit; the CP bounds used in experiments can be much tighter for rare events.

G FULL-DISTRIBUTION LIKELIHOOD CONFIDENCE REGIONS

Likelihood and entropy uncertainty sets are established robust-MDP tools (Nilim & El Ghaoui, 2005). To avoid attributing every difference from a joint cell-box baseline to the retained information, we also use a coupled full-distribution likelihood region. Calibration is finite-sample rather than a chi-square approximation.

Let a fixed input row have n independent outcomes on s states, counts c_y, and empirical probabilities $\widehat{p}_y = c_y/n$. Choose a fixed Dirichlet $(a, \ldots, a)$ mixing distribution, $a = 1/2$. The probability of the ordered outcome sequence under that mixture is

$$m(c) = \frac{\Gamma(sa)}{\Gamma(n + sa)} \prod_y \frac{\Gamma(c_y + a)}{\Gamma(a)}.$$

For the true row distribution p, the ratio $m(c)/\prod_y p_y^{c_y}$ has expectation at most one, including boundary distributions (the missing mass corresponds to impossible sequences). Markov's inequality and a union bound over R input rows give simultaneous coverage by

$$\mathcal{C}_e = \{p : \mathrm{KL}(\widehat{p}_e \| p) \le \rho_e\}, \quad \rho_e = \frac{\log L(\widehat{p}_e) - \log m(c_e) + \log(R/\delta)}{n_e}. \tag{20}$$

Here KL denotes Kullback–Leibler divergence. For $n_e = 0$, use the full simplex. This is a frequentist likelihood-ratio confidence region, not a posterior credible region. The same construction on $k+1$ output counts gives the count-likelihood baseline. The full-joint and count-likelihood comparisons therefore change the observed outcome alphabet while using the same calibration principle.

For finite radius and nonconstant continuation vector v, the row robust expectation has the dual form

$$\sup_{p:\mathrm{KL}(q\|p)\le\rho} p^\top v = \inf_{\nu \ge \max_y v_y} \left\{ \nu - \exp\left(\sum_{y:q_y>0} q_y \log(\nu - v_y) - \rho \right) \right\}. \tag{21}$$

For $\rho > 0$, strict feasibility at $p = q$ on its support permits convex duality. The Lagrangian maximizer on observed states is proportional to $q_y/(\nu - v_y)$. Minimizing over the KL multiplier gives the exponential expression; at the boundary $\nu = \max v$, any remaining probability may be placed on an unobserved maximizing state. Limits handle zero radius, constant v, and infinite radius. The code minimizes this one-dimensional dual, so it returns an upper dual objective rather than an unchecked feasible-primal claim. Sixteen independently optimized primal problems and sparse-support edge cases validate the solver. Floating-point agreement is not formal interval arithmetic.

Every method is calibrated at the same total nominal error budget, but the regions have different shapes and finite-sample overheads. The likelihood region is a stronger *comparison design*, not a theorem that it dominates cell boxes. Indeed it is wider in some measured settings. These results do not establish an intrinsic statistical penalty for retaining full outputs independent of the chosen confidence procedure.

H Memory capacity and selective evidence

The pressure environment is a cache-like, bounded-record compactor, not a language summarizer. There are d immutable named facts, resident set S_t of size at most B, and an independently queried key with weights $w_i \ge 0$, $\sum_i w_i = 1$. Only correct resident facts can be recalled; guessing absent values or encoding several facts into one record is outside this interface. A step brings at most one correct proposal, and a candidate store larger than B must discard a record. Let A_H be the keys seen initially or proposed by horizon H.

Proposition 10 (Exact record-interface capacity limit)**.** *For every policy respecting the interface,*

$$R_H \ge 1 - \mathbb{E}\left[\sum_{i \in \mathrm{Top}_B(A_H; w)} w_i \right] \ge 1 - \sum_{j=1}^{B} w_{(j)}.$$

Here $w_{(1)} \ge \cdots \ge w_{(d)}$ and the top-B operation retains all available keys if fewer than B have been seen. With known fixed query weights, the first bound is attained simultaneously at every horizon by retaining the B highest-weight keys seen so far.

Proof. Conditional on the resident set, success probability is $\sum_{i \in S_H} w_i$. Since $S_H \subseteq A_H$ and $|S_H| \le B$, this sum is bounded by the top-B available weights. Taking expectations proves the first inequality, and enlarging A_H to the full universe proves the second. For attainment, induct on

proposal arrivals. A key previously discarded had no greater weight than every retained key at that time; immutable weights imply it cannot later enter the top-B set merely because more keys arrive. Keeping the largest weights among the current store and the new proposal therefore maintains a top-B set of all keys ever seen. This argument also handles ties by a fixed ordering. □

This elementary selection bound is a benchmark-specific capacity statement, not a new general coding theorem for B bits. It rules out a trivial all-facts-fit regime and supplies an exact task floor. It does not predict performance of a learned value-corrupting summarizer.

In the experiment, $d = 8$, $B \in \{3, 4, 5, 6\}$, and all eight facts have positive query weights: $(.25, .25, .25, .05, .05, .05, .05, .05)$. Arrival probabilities are $(.12, .10, .08, .04, .04, .04, .04, .04)$, with no proposal on the remaining half of steps. Keys are retained using fixed priority scores $(3.5, 3.5, 3.5, 0, 0, 0, 0, 0)$: when the candidate store has $B + 1$ distinct keys, the probability of dropping a key is proportional to the exponential of its negative score. The compactor sees keys and these scores, not evaluator-only error labels. The initial store contains the last B keys.

The state space is the *physically realizable* resident subsets, of size $\binom{8}{B}$. This state is sufficient because arrivals and eviction depend only on it. All B slots remain occupied, so the total count of absent records is always $8 - B$: that output statistic cannot identify which important keys are missing. Instead retain $\phi(S)$, the number of missing keys among the first three. Terminal query loss is exactly

$$L(S) = .05(8 - B) + .20\phi(S).$$

Every concrete resident input is kept separate; there is no count pooling. The resulting task-tail certificate uses the same nested-set operator on valid resident states. The total-count-only upper bound is $\max_S L(S)$ and is uninformative even with unlimited samples. Exact risk and the top-B-available oracle are computed by separate finite-state recursions. Fresh proposals are selective, all useful facts do not fit, and the capacity floor is strictly positive in every configuration.

I From exact identification to useful test decisions

The exact-identification theorem distinguishes zero from nonzero ambiguity; it does not distinguish harmless ambiguity from a failed certification target. This section supplies a quantitative, support-aware answer at the first unresolved horizon and a test intervention that removes a specified continuation obstruction. The row optimization and its dual use standard finite linear programming, not a new robust-expectation principle.

I.1 The first unresolved horizon is a stationary linear problem

Let $\mathcal{C}_{M,S}$ contain all nonnegative row-stochastic matrices satisfying $PF = M$ and $P(e, y) = 0$ for $y \notin S_e$. The sets S_e are *declared* allowed supports; an unobserved transition is not automatically a structural zero. Suppose $v_0 = L, \ldots, v_{j-1} \in \mathrm{col}(F)$, with $v_{i+1} = Ma_i$ whenever $v_i = Fa_i$. This recursively defines a common $v_j = P^j L$ for all compatible kernels, whether or not v_j belongs to the tested span. The convention $j = 0$ permits L itself to be unrepresented.

Proposition 11 (Exact stationary extrema at the first unresolved horizon)**.** *Assume $\mathcal{C}_{M,S}$ is nonempty. Its largest and smallest endpoint risks at horizon $j + 1$ equal*

$$\sup_{P \in \mathcal{C}_{M,S}} \mu P^{j+1} L = \sum_e \mu_e \max_{p \in \Delta(S_e):\, pF = M_e} p v_j, \tag{22}$$

with maxima replaced by minima for the lower endpoint. Both are attained by frozen kernels. This statement requires neither full support nor a validated reference transition matrix: v_j is reconstructed from the exact retained moments M and the earlier span representations.

Proof. Induction gives $P^i L = v_i$ for $0 \le i \le j$: if $P^i L = v_i = Fa_i$, then $P^{i+1} L = PFa_i = Ma_i$. Therefore $\mu P^{j+1} L = \sum_e \mu_e p_e v_j$. The row constraints are separate compact polytopes. Optimize each row of this linear expression independently and collect the optimizers into one matrix. The same fixed matrix is used at every step; the induction still applies to that matrix. Compactness gives attainment. At later horizons, $P^h L$ need not be shared, so this argument does not solve general

stationary optimization. Likewise, the result need not hold with unmodeled constraints coupling different rows. □

Proposition 12 (Support-aware ambiguity at the first unresolved horizon)**.** *Suppose exact output-class laws $q_e(c)$ are known for classes C_c, and earlier shared continuations $v_0 = L, \dots, v_{j-1}$ are class-measurable. Then v_j is determined by those laws. With independently declared allowed supports S_e, the width of the compatible frozen-risk interval at $j+1$ is*

$$W_{j+1} = \sum_{e,c} \mu_e q_e(c) \left[\max_{y \in C_c \cap S_e} v_j(y) - \min_{y \in C_c \cap S_e} v_j(y) \right]. \tag{23}$$

Zero-mass classes contribute zero. Both endpoints are attained by stationary kernels. No full-support assumption is needed.

Proof of Proposition 12. For partition features, M_e is the class law $q_e(c)$. Each supported class $C_c \cap S_e$ can independently assign its mass to any member. Its extrema are $q_e(c)$ times the minimum or maximum of v_j on that set. Summing the differences gives the stated exact width. A positive mass on an empty supported class makes the model infeasible, which must be reported rather than silently deleting that mass. Zero-mass classes contribute zero. The attaining kernels can be on the boundary. If the reference and allowed class have full support and one insists on positive compatible kernels, mixing either optimizer with a positive compatible kernel approaches the same extrema; the supremum and infimum are unchanged. □

This gives an exact answer about *how much* ambiguity matters at the first unresolved horizon, including under known structural zeros. It is not a replacement for the generic interior necessity theorem at every later horizon. With uncertain moments, the formula cannot be used as if their plug-in estimates were exact: the uncertainty in both the row constraints and the reconstructed continuation must be propagated. The finite-sample feature experiment below instead uses simultaneous confidence regions directly.

I.2 Controlled distance and support-boundary experiments

Use states $00, 10, 01, 11$, loss $L = (0, 1/2, 1/2, 1)^\top$, initial state 00, and the exact count-class test. For $0 \le \lambda < 1/4$, $0 < q < 1$, and $0 \le b \le 1$, let

$$K_{\lambda,q,b} = \begin{pmatrix} (1-q)/2 & q(1+b)/2 & q(1-b)/2 & (1-q)/2 \\ 1/4+\lambda & 1/4 & 1/4 & 1/4-\lambda \\ 1/4-\lambda & 1/4 & 1/4 & 1/4+\lambda \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}. \tag{24}$$

Here $KL = (1/2, 1/2-\lambda, 1/2+\lambda, 1/2)^\top$. Its Euclidean residual from count-measurable functions is $(0, -\lambda, \lambda, 0)^\top$, with sup norm λ. The unrestricted count-compatible stationary two-step range is exactly

$$[\,1/2 - q\lambda,\ 1/2 + q\lambda\,], \qquad W_2 = 2q\lambda. \tag{25}$$

The nominal risk is $1/2 - qb\lambda$. Equation (25) follows either from Proposition 12 or by moving the first row's count-one mass between the two count-one outputs. Every other row retains its count law, so its one-step expectation of L is unchanged.

The generic approximate bound in Appendix D gives the conservative upper endpoint $\min(1, R_2(K) + 2\lambda)$. It is valid but can be much looser than the exact supported-class calculation. The displayed interior witness in the proof of Theorem 5 is $2\lambda \min_y K(00, y)$, which tends to zero when $b \uparrow 1$. In contrast, the ambiguity width $2q\lambda$ is unchanged if the zero is *not* declared structural. If $b = 1$ and the first row's missing transition to 01 is independently declared impossible, that row can no longer move its count-one mass, and its two-step width is zero. This does not infer structural zeros from finite samples.

We vary λ across eleven values from zero to $.24$, q across six values from $.001$ to $.9$, and b across five values from zero to one, giving 330 exact geometries. Each is checked against its explicit frozen optimizers. The target grid is $.505, .51, .55, .60$. For example, at $(\lambda, q, b) = (.01, .2, 0)$, exact identification fails, but the upper endpoint 5.02×10^{-1} certifies target $.51$; the generic residual upper

bound .52 does not. At $(.1, .01, 0)$, the residual is ten times larger but the exact upper endpoint is only 5.01×10^{-1}. Thus unweighted distance from closure alone does not determine practical significance. At $(.1, .2, .9999)$, the interior perturbation witness is 2.00×10^{-6} while the exact width is .04. At the declared support boundary it collapses instead. These examples distinguish near-closure behavior, rare class occupancy, and an irrelevant small reference probability.

I.3 Adding a binary output statistic without learning the kernel

For a predeclared binary statistic $b(y)$, intersect each existing count-tail region with $\mathbb{E}b(Y) \le u_e$. No count pooling or output independence is required. Assume every nested set $A_c = \{y : \phi(y) \le c\}$ contains a feature-zero state; in the bit experiment the all-zero output suffices. Define

$$A_c(v) = \max_{y \in A_c : b(y)=0} v(y), \qquad B_c(v) = \max_{y \in A_c : b(y)=1} v(y),$$

with an empty maximum in B_c interpreted as $-\infty$. If $\pi_{e,c}$ are the tightened tail differences, the augmented upper expectation is

$$(\mathcal{A}_{U,u} v)(e) = \inf_{s \ge 0} \left\{ s u_e + \sum_c \pi_{e,c} \max\{A_c(v), B_c(v) - s\} \right\}. \tag{26}$$

Proof. Dualize the extra inequality $\mathbb{E}b(Y) \le u_e$ in the finite row LP. Strong LP duality gives $\inf_{s \ge 0}\{s u_e + \sup_{p \in \mathcal{Q}_e(U)} \mathbb{E}_p[v(Y) - s b(Y)]\}$. Apply the existing nested-set upper expectation to the inner support function. Nonemptiness follows from the available feature-zero, level-zero output. The piecewise-linear dual changes slope only at positive differences $B_c(v) - A_c(v)$; after the largest such difference its slope is $u_e \ge 0$. Consequently a minimizer exists at zero or at one of these at most k breakpoints, including the flat-tail case $u_e = 0$. Evaluating all candidates costs $O(Rk^2 + N)$ per step for R input rows and N output states. This is a standard constrained support-function dual with a small specialized candidate set, not a new duality principle. □

Because the extra constraint only shrinks the *same* row region, $\mathcal{A}_{U,u} v \le \mathcal{B}_U v$ and backward induction gives the same ordering over all horizons. A separately calibrated count-only baseline can have slightly narrower count caps: the ordering need not survive that change of confidence region. For every input, the evaluator sees sampled counts and the sampled bit only, never K, its parameters, or an oracle continuation value. This bound treats the statistic as predeclared. The data-selected procedure in Appendix C applies simultaneous correction across candidates and compares with both the native count baseline and the shared-table control.

I.4 What the scaling family does and does not demonstrate

With the influential coordinate relabeled as 1, each next bit has conditional mean $p(e) = a + \beta e_1 + \gamma L(e)$, with $a = .003$, $\beta = .05$, $\gamma = .08$. Conditional independence is used by the synthetic generator but not by the evaluators or the following algebra. Since $KL = Ke_1 = p$, the span of $1, L, e_1$ is invariant. For $s = \beta + \gamma$, induction gives

$$K^H L = a \sum_{j=0}^{H-1} s^j + s^{H-1}\{\beta e_1 + \gamma L\}. \tag{27}$$

At the all-correct initial state this is $a(1 - s^H)/(1 - s)$.

Exact count tails imply $\mathbb{E}L(Y) \le p(e)$, and adding the oracle upper bound $\mathbb{E}e_1(Y) \le p(e)$ bounds every nonnegative linear combination of these two features. The true kernel attains both simultaneously. Because the coefficients in (27) are nonnegative, the augmented upper backup is exact at every horizon, even though it retains only upper inequalities. The same reasoning does not work with e_2 when $k > 2$: e_1 is then not in the span of count-measurable functions and e_2. The selected-feature experiment tests $k = 4, 8, 12$ to avoid the special two-bit identity $e_1 = 2L - e_2$.

Exact affine closure does not require enumerating all 2^k inputs. With a stronger, independently justified shared-affine-mean model, the all-correct and all-error rows alone suffice. Their unknown means are a and $a + \beta + \gamma$. One random queried output bit per reset, obtainable by randomized

thinning of the output count, gives two Bernoulli streams. Allocate $\delta/2$ to each CP upper endpoint u_0, u_1 and set $\bar{u}_1 = \max(u_0, u_1)$. Then

$$r_0^+ = 0, \qquad r_{h+1}^+ = u_0(1 - r_h^+) + \bar{u}_1 r_h^+$$

upper-bounds every horizon. Indeed after one step all output-bit means and the average error agree, so the true mean obeys $r_{h+1} = a(1-r_h)+(a+\beta+\gamma)r_h$. The upper map is nondecreasing and dominates the true map on $[0, 1]$; induction proves the bound. This control uses unknown fitted means, not known numerical parameters, but its *structural assumption is much stronger* than the generic row test's. It also shows that count information alone can suffice after restricting the model class. At total budgets 512, 2,048, 8,192, and 32,768, 200 independent tests each give mean $H = 128$ bounds $0.02, 0.01, 6.05 \times 10^{-3}, 4.56 \times 10^{-3}$, independent of k. Thus the large enumerated budgets are a cost of the deliberately unrestricted input model, not an unavoidable law of this structured family. The two-row control is elementary, not an optimal design claim. It has 11 observed underbounds in 800 tests at the nominal five-percent level; we do not claim empirical zero failure.

I.5 A MECHANISM-DERIVED, DEPLOYMENT-WIDE MISMATCH ALLOWANCE

The general mismatch theorem does not discover a uniform allowance from samples. Here the allowance is enforced by an independently inspectable wrapper, rather than chosen by comparing with the evaluator's transition matrix. The ordinary branch runs the tested compactor using its declared memory input and fresh observation mechanism. A separate branch may reuse a cached snapshot chosen from arbitrary earlier history. Before drawing or inspecting the ordinary output, a fresh independent gate permits this branch with conditional probability $a_t \le g$. On rejection, the ordinary branch must run unchanged. The wrapper receives neither K, the evaluation loss, nor the error state.

Given any history ending at e, the output law is $(1 - a_t)K_e + a_t Q_t$, where Q_t may depend on all hidden cache state. Consequently for every tail event,

$$\mathbb{P}(\phi(E_{t+1}) \ge \ell \mid \mathcal{F}_t) \le (1 - a_t)T_e(\ell) + a_t \le T_e(\ell) + g[1 - T_e(\ell)]. \tag{28}$$

On $T \le U$, propagate $U_e^g(\ell) = U_e(\ell) + g[1 - U_e(\ell)]$. Monotonicity of this transform and backward conditional expectation give the endpoint guarantee. In particular, g is also an admissible allowance for Theorem 9. The multiplicative form is tighter than simply adding g. It requires choosing the gate independently *before* the ordinary output; a rule that selectively replaces successful ordinary outputs can violate it. Such a rule would retain only the additive total-variation bound when its override probability is at most g.

The contract has real scope. The nominal branch's specified behavior still needs justification. Other unguarded sources of state insufficiency remain uncovered; a buggy gate, an incorrectly modeled ordinary branch, or cache access outside the wrapper invalidates the argument. This is an enforceable mechanism example, not a method for discovering unknown abstraction error. A cached B-record snapshot is additional storage: the example charges up to B active plus B auxiliary records, rather than claiming a B-slot system has free hidden memory.

For the finite experiment, the ordinary branch is the disclosed eight-key/six-slot record compactor. Independent nominal resets test all 28 valid resident states. The hidden cache mode has two states, transition matrix $\left(\begin{smallmatrix} .85 & .15 \\ .10 & .90 \end{smallmatrix}\right)$, and acceptance probabilities $g/4$ and g. The two cache snapshots are fixed valid resident subsets of high task loss. They are specified in the simulator before any test, not inferred by the certifier. The evaluator tracks the 56-state joint process exactly; the evaluator retains no cache mode and receives no joint kernel. We test $g \in \{0, .001, .003, .01, .03\}$, 24 nominal test datasets at each $n = 256, 1024, 4096$ resets per row, and $H = 8, 32, 128, 512$. All gate values and horizons reuse a dataset, giving 72 independent test datasets, not 1,440 independent experiments.

At 114,688 calls and $H = 512$, the nominal mean bound 0.12 is invalid at $g = .01$, where actual risk is 0.13. The mechanism-derived bound is 0.17, compared with additive inflation 0.17. For $g = .003$, the guarded mean bound 0.14 remains below target $.15$. These observations verify implementation and informativeness of a bound already justified by the branch contract; they are not how g is obtained.

J BASELINE CERTIFICATE STATEMENTS

Assumption 1 (Sufficient stationary local responses)**.** For known parent sets $\mathrm{pa}(i)$ of size at most k, there are fixed tables f_i such that

$$\mathbb{P}(E_{t+1,i} = 1 \mid \mathcal{F}_t) = f_i(E_{t,\mathrm{pa}(i)}). \tag{29}$$

The same law holds in independently reset tests. Unless explicitly stated otherwise, the test input design is fixed or randomized independently of the outputs.

Proposition 13 (Clean tests do not identify recovery)**.** *Two-state compactors with $f(0) = \epsilon$ and $f(1) = 1 - r$ agree on every clean-input test. Starting clean, their endpoint errors satisfy*

$$p_H^{(r)} = \frac{\epsilon}{\epsilon + r}\{1 - (1 - \epsilon - r)^H\}, \qquad p_H^{(0)} = 1 - (1 - \epsilon)^H, \tag{30}$$

where $\epsilon + r > 0$. Both have first-error probability $1 - (1 - \epsilon)^H$.

J.1 MARGINAL TABLES: A BASELINE WITHOUT OUTPUT INDEPENDENCE

Given a local response f and upper input-error marginals u, keep every joint input law consistent with those caps:

$$\mathcal{T}_f(u) = \max_{\mu \in \Delta_{2^k} : \sum_z \mu_z z \le u} \sum_z \mu_z f(z). \tag{31}$$

Theorem 14 (Optimal local cap)**.** *The maximum is attained, is monotone and concave in u, and equals*

$$\min_{b \in \mathbb{R},\, a \ge 0} \{b + a^\top u :\ b + a^\top z \ge f(z)\ \forall z\}. \tag{32}$$

It is the smallest valid scalar bound given only $\mathbb{E}Z \le u$ and f. Starting with $u_0 \ge p_0$, the update

$$u_{t+1,i} = \mathcal{T}_{f_i}(u_{t,\mathrm{pa}(i)}) \tag{33}$$

bounds p_t for every t, and is no larger than composing any fixed nonnegative affine majorants, clipped at one.

This uses standard LP duality (Boyd & Vandenberghe, 2004) and dependence-robust expectation optimization (Agrawal et al., 2012). Its sharpness is *one-step*: separate optimizers can be incompatible across records and time. For an AND response, the cap is $\min(u_1, u_2)$, not the product $u_1 u_2$ that an unjustified independence assumption would suggest.

Let $m = \sum_i 2^{|\mathrm{pa}(i)|}$. Use one-sided Clopper–Pearson (CP) upper bounds $U_i(z)$ with failure allowance δ/m per cell and assign one to an unseen cell (Clopper & Pearson, 1934). Under a fixed input design,

$$\mathbb{P}_{\mathcal{D}}\{f_i(z) \le U_i(z) \text{ for every } i, z\} \ge 1 - \delta. \tag{34}$$

Theorem 15 (One test, every deployment horizon)**.** *Replacing f_i by U_i in* (33) *gives $p_t \le u_t$ for all t on the event* (34)*. Thus endpoint and average bounds hold simultaneously for every H and $w \in \Delta_d$.*

No union over deployment horizons is needed. A perturbation bound controls statistical excess above the *oracle cap*, not all slack above true risk (Theorem 17). The local LP has exponentially many patterns; our cached implementation supports at most three parents.

J.2 OUTPUT COUNTS: AN EXACT BOUND WHEN POOLING IS JUSTIFIED

In a closed block of k records, let $S_t = |E_t|$. Suppose its next-count law depends only on S_t, remains fixed between test and deployment, and has tails $F(r, \ell) = \mathbb{P}(S_{t+1} \ge \ell \mid S_t = r)$ nondecreasing in r. In plain terms, equal error counts have equal next-count laws, and starting with more errors cannot improve that distribution. These are additional assumptions, not conclusions drawn from a small test.

For each count row and threshold, estimate a simultaneous upper tail $U(r, \ell)$. This retains $k(k+1)$ numbers instead of 4^k full transition entries. Calls and precision, not the parameter count alone, determine sample cost.

Theorem 16 (Horizon-sharp ordered certificate)**.** *Define*

$$F^+(r,\ell) = \min_{j\ge r,\, m\le \ell} U(j,m), \qquad P^+(r,j) = F^+(r,j) - F^+(r,j+1), \tag{35}$$

with boundary tails one and zero. Then P^+ is the greatest stochastically monotone kernel under the raw tail bounds. For every increasing loss g,

$$\sup_{P:\ F_P\le U,\ P \text{ monotone}} (P^H g)(r) = ((P^+)^H g)(r) \tag{36}$$

for every H, r. The same kernel maximizes every bounded increasing finite-path loss. A simultaneous tail-confidence event makes these bounds valid for the true process at all horizons.

The envelope is explicit; stochastic-order comparison is classical (Massey, 1987). This is global sharpness *within the abstract monotone-kernel class*, not every physically realizable record compactor. For a uniform record query, $g(r) = r/k$. If queries are independent given the state path and do not affect compaction, the any-failure bound is $1 - e_0^\top (P^+ D)^H \mathbf{1}$, where $D = \mathrm{diag}(1-g)$. We do not multiply survival probabilities of dependent blocks.

J.3 How statistical uncertainty propagates

Define the nonnegative influence matrix

$$C_{ij} = \max_z |f_i(z) - f_i(z \oplus e_j)| \quad (j \in \mathrm{pa}(i)), \tag{37}$$

with zero entries outside the parent sets and local bit indexing understood.

Theorem 17 (Statistical excess above the oracle cap)**.** *Suppose $0 \le U_i(z) - f_i(z) \le \eta_i$ for all cells. Let u^U, u^f denote* (33) *from the same initial caps. Then*

$$0 \le u_H^U - u_H^f \le \sum_{s=0}^{H-1} C^s \eta. \tag{38}$$

If $Cv \le \alpha v$ for $v > 0$ and $\alpha < 1$, the weighted sup-norm excess is at most $\|\eta\|_v (1-\alpha^H)/(1-\alpha)$, where $\|x\|_v = \max_i |x_i|/v_i$.

A coupling that thins selected one-bits proves the required Lipschitz inequality for $\mathcal{T}_f$ (Appendix J.10). This is a perturbation bound for the nonlinear cap operator. It measures *estimation error above the oracle cap*, not the oracle cap's slack above actual risk.

For example, a two-sided Hoeffding construction with n tests per local cell and row sums of C at most $\alpha < 1$ gives horizon-uniform statistical excess at most ε once

$$n \ge \frac{2\log(2m/\delta)}{\varepsilon^2 (1-\alpha)^2}. \tag{39}$$

This is a sufficient, potentially loose rate; it is not a minimax result for the general local model. The experiment uses binomial rather than Hoeffding endpoints. A contraction condition is also not a low-risk guarantee: error injected at clean inputs can sustain a large error floor. Failure of this sufficient contraction test does not prove instability.

J.4 Even complete marginal tables can be insufficient

Complete input coverage does not solve an output-dependence problem. The following family quantifies it exactly.

The main-text Theorem 1 gives the exact stationary dependence ambiguity; its proof is in Appendix J.11.

Common shocks produce the upper endpoint. A joint law avoiding the all-incorrect state while preserving each marginal ϵ produces the lower endpoint. Thus marginal-response tables cannot determine long-run risk even with unlimited samples. *Full-vector tests can distinguish these kernels*: retaining joint outputs, rather than only their marginal counts, can resolve this ambiguity. The theorem is a limitation of the retained statistics, not of all short interventions.

J.5 NOTATION AND GUARANTEES

Vectors are ordered componentwise. Δ_n denotes the simplex on n outcomes. A local bit pattern is indexed by an integer with the first parent as its least-significant bit. A query weighting $w \in \Delta_d$ selects a record independently of the current deployment error; it is not an observation-dependent adversarial query. A deterministic compound query can instead use an appropriate union bound on its required records, but its exact loss is not identified by marginal error probabilities.

Three probabilities must not be conflated. R_H integrates randomness in a future deployment at a fixed horizon. The confidence probability $1 - \delta$ integrates randomness in the test dataset used to bound R_H. The probability of any failure in a single deployment is a path event. Simultaneous confidence across horizons is not a claim that a deployment has no errors at any horizon.

The baseline confidence results in this appendix condition on a frozen compactor, known parent sets, a sufficient error representation, and a fixed observation-generating mechanism. The bounded-mismatch result in Theorem 9 instead permits the stated history-dependent deviations. Independent reset calls are a generative-model assumption. Without such an allowance, hidden state or an observation change can invalidate the baseline transfer from test to deployment. No theorem in this paper infers locality itself from finite tests of selected representatives.

J.6 CLEAN-INPUT NON-IDENTIFIABILITY

Proof of Proposition 13. For a one-bit memory, conditional expectation gives

$$p_{t+1} = \epsilon(1 - p_t) + (1 - r)p_t = \epsilon + (1 - r - \epsilon)p_t.$$

Starting at zero, summing the finite geometric series yields (30) whenever $\epsilon + r > 0$. Setting $r = 0$ yields $1 - (1 - \epsilon)^H$. If $r > 0$ is fixed, $\epsilon/(\epsilon + r) = O(\epsilon)$ and the magnitude of the bracket is at most two; for sufficiently small ϵ and $r < 1$, it is at most one. With $H = \lceil c/\epsilon \rceil$, the persistent chain's risk tends to $1 - e^{-c}$. Every independently reset clean-input output is Bernoulli (ϵ) in both models, so their entire clean-test transcript laws agree. Before the first error, the process remains at state zero and incurs error probability ϵ on each update, hence first-error survival through H is $(1 - \epsilon)^H$ regardless of r. □

For the numerical illustration, $1 - .99^{100} \approx 0.63$, whereas the recovering chain with $r = .5$ has error $.01/.51\{1 - .49^{100}\} \approx 0.02$. These are exact model calculations, not estimates from a favorable random seed.

J.7 THE LOCAL MARGINAL OPTIMIZATION

Proof of Theorem 14. The primal feasible set in (31) is nonempty because the point mass at the zero vector is feasible for every $u \geq 0$. It is a closed subset of a finite simplex and is therefore compact. Its linear objective attains its maximum. Associate nonnegative multipliers a with the inequalities $\sum_z \mu_z z \leq u$ and a free multiplier b with $\sum_z \mu_z = 1$. Finite LP duality gives (32). The constraint at $z = 0$ implies $b \geq f(0) \geq 0$, although b is formally a free variable.

Every law with the specified caps is feasible in the primal, so its expectation is at most $\mathcal{T}_f(u)$. Conversely, an attaining law is a witness that no smaller scalar bound is valid on this information alone. Monotonicity follows from inclusion of feasible sets as caps increase. Concavity follows from the dual minimum of affine functions (or by mixing primal laws). Since f takes values in $[0, 1]$, the value also lies in $[0, 1]$.

Suppose $p_t \leq u_t$. The actual law of $E_{t,\mathrm{pa}(i)}$ is feasible for the local primal at caps $u_{t,\mathrm{pa}(i)}$. Using Assumption 1,

$$p_{t+1,i} = \mathbb{E} f_i(E_{t,\mathrm{pa}(i)}) \leq \mathcal{T}_{f_i}(u_{t,\mathrm{pa}(i)}) = u_{t+1,i}.$$

Induction from $p_0 \leq u_0$ proves composition. Let (b_i, a_i) be any fixed feasible dual majorants and define $v_{t+1,i} = \min(1, b_i + a_i^\top v_{t,\mathrm{pa}(i)})$, with $v_0 = u_0$. If $u_t \leq v_t$, then optimality of the dual and $a_i \geq 0$ give

$$u_{t+1,i} \leq \min(1, b_i + a_i^\top u_{t,\mathrm{pa}(i)}) \leq v_{t+1,i}.$$

Another induction establishes domination. This argument does not assert that locally attaining laws across different i, t belong to a single compatible process. □

For $f(z) = z_1 z_2$, $\mathbb{E}f(Z) = \mathbb{P}(Z_1 = Z_2 = 1) \le \min(u_1, u_2)$. Putting mass $\min(u_1, u_2)$ at $(1, 1)$ and the rest at $(0, 0)$ attains the value. Both dual planes u_1 and u_2 are feasible, and their pointwise minimum is required. The interpretation as an upper-marginal version of a concave closure connects the construction to the established worst-dependence optimization literature; the LP itself is not a new duality principle.

J.8 Numerical implementation

For $k \le 3$, the dual feasible set is given by 2^k majorant inequalities and k nonnegative slope inequalities in $k + 1$ variables. We enumerate full-rank active sets of $k + 1$ constraints, solve for their vertices, and retain feasible vertices. The polyhedron is pointed and has a finite optimum for the nonnegative cap objectives, so an optimal vertex exists. A constant $\max_z f(z)$ plane supplies a feasible fallback.

The implementation clips very small negative slopes upward and raises each candidate intercept to satisfy all majorant constraints after that change. It independently compares the resulting values with a SciPy primal LP and checks the returned primal witnesses. These are practical numerical checks. They are not formally verified rounding bounds, and claims at less than numerical tolerance are not interpreted as meaningful violations.

J.9 Confidence from independently reset tests

For s erroneous outputs out of $n > 0$ independent Bernoulli trials, write

$$U_{\mathrm{CP}}(s, n; \alpha) = \begin{cases} 1, & s = n, \\ \mathrm{Beta}^{-1}(1 - \alpha; s + 1, n - s), & s < n. \end{cases}$$

For $n = 0$ define the endpoint to be one. The upper endpoint inverts a binomial lower-tail test: for every success probability q, $\mathbb{P}_q(q > U_{\mathrm{CP}}) \le \alpha$. This is the one-sided Clopper–Pearson construction (Clopper & Pearson, 1934), not a posterior credible interval.

Proof of Theorem 15. Condition on the input design. For a particular cell (i, z), the observations selected by that design are independent Bernoulli variables with common mean $f_i(z)$ under the reset and locality assumptions. The cell count can depend on the randomized input design but not on the outcomes; conditional on the design it is fixed. Thus the upper endpoint fails with probability at most δ/m. Union over the m cells establishes (34) conditional on the design and hence unconditionally. This union does not require independence among different cells or coordinates.

On the event $f \le U$, monotonicity in the objective table gives $\mathcal{T}_{f_i}(v) \le \mathcal{T}_{U_i}(v)$ for every cap vector v. The induction in Theorem 14 therefore applies at every integer time on that same event. Taking a nonnegative weighted sum and then, if desired, averaging over times preserves the inequalities. Because the event constrains the tables rather than a chosen horizon or weighting, the conclusion holds for every H and w simultaneously. No additional confidence expenditure is needed to choose H or w after the test. □

If δ is reused for several separately evaluated compactors, each certificate is individually valid at confidence $1 - \delta$; their simultaneous familywise confidence is not automatically $1 - \delta$. Our experiments report each specified method separately. Selecting among compactors on the same data requires a confidence allocation across those candidates or another valid selection argument, which we do not claim here.

J.10 Perturbation of the nonlinear cap

Lemma 18 (Cap Lipschitz property)**.** *Let* $c_j = \max_z |f(z) - f(z \oplus e_j)|$. *For any* $u, v \in [0, 1]^k$,

$$|\mathcal{T}_f(u) - \mathcal{T}_f(v)| \le \sum_j c_j |u_j - v_j|.$$

Proof. First suppose $v \le u$. Take a maximizing law for the cap u, and let Z have that law with marginal means $m_j \le u_j$. When $m_j > v_j$, thin the one-bits in coordinate j with probability

$1 - v_j/m_j$; leave the other coordinates unchanged. The resulting vector Z' has marginal means at most v and expected numbers of changed bits $(m_j - v_j)_+ \le u_j - v_j$. A path that flips those bits one by one gives

$$|f(Z) - f(Z')| \le \sum_j c_j \mathbf{1}\{Z_j \ne Z'_j\}.$$

Consequently, $\mathcal{T}_f(u) = \mathbb{E}f(Z) \le \mathbb{E}f(Z') + \sum_j c_j(u_j - v_j) \le \mathcal{T}_f(v) + \sum_j c_j(u_j - v_j)$. The opposite inequality in the ordered case follows from monotonicity. For general caps, increase coordinates from v to $u \vee v$ to obtain $\mathcal{T}_f(u) - \mathcal{T}_f(v) \le \sum_j c_j(u_j - v_j)_+$. Swap u and v for the other direction. □

Proof of Theorem 17. For any cap v, the assumed uniform table error gives $0 \le \mathcal{T}_{U_i}(v) - \mathcal{T}_{f_i}(v) \le \eta_i$, since every feasible law averages errors at most η_i. Monotonicity implies $u_t^U \ge u_t^f$. Applying Lemma 18 to the true tables yields

$$u_{t+1}^U - u_{t+1}^f \le \eta + C(u_t^U - u_t^f).$$

Unroll from zero initial difference to obtain (38). For $Cv \le \alpha v$, the nonnegative matrix satisfies $\|Cx\|_v \le \alpha \|x\|_v$, so summing the geometric series gives the weighted statement. □

For the Hoeffding corollary, with n samples per cell let $a = \sqrt{\log(2m/\delta)/(2n)}$ and $U_i(z) = \min(1, \widehat{f}_i(z) + a)$. A union of two-sided Hoeffding events gives $0 \le U - f \le 2a$ with probability at least $1 - \delta$. With $v = \mathbf{1}$ and row sums bounded by α, statistical excess is at most $2a/(1-\alpha)$. Solving for n gives (39). This uses the true influence matrix as an analysis condition. A data-dependent claim about contraction would need confidence bounds on the influences as well; it cannot simply substitute an unqualified plug-in estimate.

The bound is attained, away from clipping, in a one-bit model $f(0) = \beta$, $f(1) = \beta + \alpha$ when both table entries are increased by η: the marginal cap is affine on $[0, 1]$ and its difference after H steps equals $\eta \sum_{s=0}^{H-1} \alpha^s$. The sensitivity experiment checks this equality for 60 parameter/horizon combinations. This equality does not imply that the Hoeffding sample bound is minimax, because it does not account for the variance of the observations.

J.11 Dependence ambiguity with complete marginal responses

Proof of Theorem 1. For every process with the specified tables, $f_i(e) \ge \epsilon$, so $p_{H,i} \ge \epsilon$ for every $H \ge 1$. For the local upper optimization,

$$\mathcal{T}_{f_i}(u) = \epsilon + (b - \epsilon) \max_{\mathbb{E}Z \le u} \mathbb{P}(Z = \mathbf{1}) = \epsilon + (b - \epsilon) \min_j u_j.$$

The upper probability is attained by mass $\min_j u_j$ on $\mathbf{1}$ and the rest on zero. Starting at zero, all caps remain equal and obey $r_{t+1} = \epsilon + (b - \epsilon) r_t$, giving (4). Theorem 14 proves that this bounds every process in the class.

For global upper attainment, choose a stationary common-shock kernel: given current state e, output all ones with probability $f_i(e)$ and all zeros otherwise. Its conditional marginal tables are exactly (3), and from clean initialization it remains in $\{0, \mathbf{1}\}$. Its probability of being all incorrect obeys the same scalar recursion, so uniform-query risk equals r_H.

For lower attainment, construct a distribution Q that places probability $d\epsilon/(d-1)$ on a uniformly selected vector with exactly one zero, and places the remaining probability on the zero vector. The restriction $\epsilon \le (d-1)/d$ makes this a probability law. It has every marginal equal to ϵ and never outputs $\mathbf{1}$. Define a stationary kernel that uses Q at every state other than $\mathbf{1}$; at $\mathbf{1}$ use any law with all marginals b, for example the common-shock law. It has the same conditional tables. Starting clean, it never visits $\mathbf{1}$, and every subsequent marginal equals ϵ. This attains the lower bound at every horizon. □

The conclusion is stronger than a failure of clean testing: the two kernels agree on each coordinate's conditional mean at *every* input. It is also narrower than non-identifiability from full-vector interventions: those interventions reveal whether outputs co-occur. Our certificate deliberately discards most of that joint information, and the theorem quantifies the resulting possible loss. This is distinct from the gap created by locally incompatible cap optimizers; the upper bound happens to be jointly attainable in this family but need not be so for arbitrary tables.

J.12 JOINT-BLOCK AND DIRECT-EVALUATION BASELINES

J.13 INTERVAL MARKOV CERTIFICATES ON KNOWN CLOSED BLOCKS

A block G is closed if its next-state distribution conditioned on full history depends only on its current block state, through a stationary kernel K_G. This is stronger than knowing the individual parent sets when joint output dependence might vary with external state. In the symbolic environment, groups are closed by construction.

For each closed block, count transitions from each input state to each output state using the same full-memory test calls. If there are $m_B = \sum_G 2^{2|G|}$ transition cells, assign tail probability $\delta/(2m_B)$ to each one-sided upper and lower CP endpoint. The lower endpoint can be computed as one minus the upper endpoint on complement counts. A union bound yields confidence at least $1 - \delta$ that every row of K_G lies in its interval box intersected with the simplex. If a computed interval box is infeasible, the implementation enlarges it to the whole simplex rather than claiming a bound from an empty uncertainty set.

Let $g_G(s) = |G|^{-1} \sum_{j \in G} s_j$. Starting with $V_{G,0} = g_G$, apply

$$V_{G,h+1}(s) = \max_{q \in \Delta:\, L_G(s,\cdot) \le q \le U_G(s,\cdot)} q^\top V_{G,h}.$$

A feasible true row and monotonicity prove by induction that $V_{G,h}(0)$ upper bounds the expected block loss at horizon h. Weighting blocks by $|G|/d$ gives a uniform-query bound without requiring independence between blocks. It is simultaneously valid for all horizons on the same row-confidence event. The maximization is solved by starting at the lower endpoints and allocating residual mass in decreasing order of V until upper endpoints bind. An independent LP validates this greedy row optimizer in the tests. This is a standard rectangular-uncertainty model-checking baseline, not a new algorithm.

J.14 DIRECT RANDOM-QUERY AND FULL-ENDPOINT BOUNDS

With budget A and horizon H, the direct methods run $n = \lfloor A/H \rfloor$ independent full trajectories. The random-query method samples one coordinate uniformly at each endpoint and applies a CP interval to its binary error. The full-endpoint method uses $Y_\ell = d^{-1} \sum_i E^{(\ell)}_{H,i} \in [0,1]$ for each rollout. The Y_ℓ are independent across rollouts, but coordinates within a rollout need not be independent.

For a bounded variable with mean p, convexity gives $\mathbb{E}e^{-\lambda Y} \le 1 - p + pe^{-\lambda}$. Chernoff optimization therefore yields $\mathbb{P}(\overline{Y} \le x) \le \exp[-n\,\mathrm{kl}(x,p)]$ for $x < p$. The full-endpoint upper bound is the largest $q \in [\overline{Y}, 1]$ with $n\,\mathrm{kl}(\overline{Y}, q) \le \log(1/\delta)$, taking value one when $n = 0$ or $\overline{Y} = 1$. It is a valid one-sided confidence bound for R_H and does not treat fractional losses as binomial counts. The implementation uses bisection. It is not guaranteed to dominate the random-query CP interval on every finite sample; the latter can exploit discreteness when few failures occur. Both are retained rather than selectively choosing their minimum at unchanged confidence.

K SUPPLEMENTARY COUNT AND SEQUENTIAL PROOFS

K.1 RETAINED OUTPUT ORDER: AN EXACT OBSTRUCTION

This section separates *input coverage* from *the order of the output statistics retained*. Complete input coverage does not compensate for discarding the relevant output event. The construction uses standard parity/limited-independence structure, not a new construction of independent random variables; see, for example, the broader limited-independence literature (Naor & Naor, 1993).

Theorem 19 (Exact risk range under every proper output marginal)**.** *Let $k \ge 2$ and $a = \mathbf{1} \in \{0,1\}^k$. State a is absorbing. At every other input, require that the output on every proper subset of the k coordinates be uniform on that subset. Starting at zero, the smallest and largest uniform-query endpoint risks at any $H \ge 1$ are*

$$\frac{1}{2} \quad \textit{and} \quad 1 - \frac{1}{2}\big(1 - 2^{-(k-1)}\big)^{H-1}.$$

Both are attained for all horizons by stationary kernels with identical conditional output distributions on every proper subset at every input.

Proof. For a nonabsorbing row, let Y have the output law and write the Walsh characters as $\chi_S(y) = (-1)^{\sum_{i\in S} y_i}$. Uniformity on all proper subsets implies $\mathbb{E}\chi_S(Y) = 0$ for all nonempty proper S. Fourier inversion on the Boolean cube therefore gives

$$\mathbb{P}(Y = y) = 2^{-k}\{1 + \theta\chi_{[k]}(y)\}, \qquad |\theta| \le 1.$$

The bound on θ is equivalent to nonnegativity, since both parities occur. Thus the probability of a is between zero and $q_* = 2^{-(k-1)}$. Let $A_t = \mathbb{P}(E_t = a)$. Absorption gives $A_t \ge A_{t-1}$, and the rowwise trigger bound gives

$$A_t \le A_{t-1} + (1 - A_{t-1})q_*, \qquad A_0 = 0.$$

Consequently $0 \le A_t \le 1 - (1 - q_*)^t$. A uniform record query after update H has conditional risk one when $E_{H-1} = a$ and $1/2$ otherwise, so

$$R_H = \tfrac{1}{2} + \tfrac{1}{2}A_{H-1}.$$

This yields the stated extrema. On every nonabsorbing row, refresh uniformly on the parity class excluding a to attain $A_t = 0$; refresh on the parity class including a to attain the upper recurrence. Both choices use a fixed row law away from a, are stationary, and agree on all proper output marginals, including at a where both output a deterministically. □

The result applies even with exact population tables of all output orders below k. More narrowly, a test interface exposing only a chosen proper subset on each independent call gives identical transcript laws for the two endpoint kernels, even if inputs and subsets are selected adaptively from earlier transcripts. Conditional on that transcript, the next exposed proper-subset law is identical, so induction proves the assertion. This interface is *not* the full-vector interface used in our experiments. Full-vector outputs distinguish parity immediately; an implementation need not discard this information. We do not claim that arbitrary joint distributions of empirical moment estimates remain indistinguishable after full-vector observations. The example's purpose is to identify which population information a test summary loses.

For the refresh family, retaining the full-order event $\{E_{t+1} = a\}$ identifies its trigger probability q and hence $R_H = 1 - (1 - q)^{H-1}/2$. General processes need not be characterized by a single such event. In particular, the count method below retains a set of nested events, rather than assuming that one trigger always suffices.

K.2 Ordered-statistic certificates

K.3 Assumptions and confidence

For a block of k records, let $S_t \in \{0, \dots, k\}$ be its number of errors. A pooled count test makes two additional assumptions beyond a generic marginal-error model:

Assumption 2 (Sufficient monotone count dynamics)**.** There is a fixed stochastic kernel P on $\{0, \dots, k\}$ such that $\mathbb{P}(S_{t+1} = j \mid \mathcal{F}_t) = P(S_t, j)$, and the same law applies to independent reset tests. Its tail rows $F(r, \ell) = \sum_{j\ge\ell} P(r, j)$ are nondecreasing in r for every ℓ.

Sufficiency requires that different concrete memories with the same count have the same next-count law (count-lumpability), not merely the same *mean* next count. Neither sufficiency nor stochastic monotonicity is inferred from the calibration samples. A chosen ordered statistic other than the count can be used with its own corresponding assumptions and a nondecreasing query loss.

For a predetermined input design, collect n_r outputs at input count r. For each $\ell = 1, \dots, k$, the number with output count at least ℓ is binomial with parameter $F(r, \ell)$. Use a one-sided CP upper endpoint $U(r, \ell)$, assigning one to an unseen row. Spending $\delta/[k(k + 1)]$ per row/threshold gives

$$\mathbb{P}_{\mathcal{D}}\{F(r, \ell) \le U(r, \ell) \text{ for all } r, \ell\} \ge 1 - \delta. \tag{40}$$

For several blocks, replace the denominator by $\sum_g k_g(k_g + 1)$. Threshold indicators on the same call, and outputs across blocks, need not be independent. Identical conditional count laws within each pooled row and independence across reset calls justify the binomial statement conditional on the input design.

K.4 A MAXIMAL COMPATIBLE KERNEL

Proof of Theorem 16. Write $F^+(r,\ell) = \min_{j\ge r,\, m\le \ell} U(j,m)$. Increasing r shrinks the minimization set, so F^+ is nondecreasing in r. Increasing ℓ enlarges that set, so it is nonincreasing in ℓ. Every entry is in $[0,1]$, and the choice $(j,m) = (r,\ell)$ gives $F^+ \le U$. Appending tails $F^+(r,0) = 1$ and $F^+(r,k+1) = 0$ therefore defines a valid stochastically monotone kernel by adjacent differences.

If a monotone tail matrix F satisfies $F \le U$, then for $j \ge r$ and $m \le \ell$,

$$F(r,\ell) \le F(j,\ell) \le F(j,m) \le U(j,m).$$

Taking the minimum gives $F \le F^+$. Thus F^+ is the *greatest* monotone tail matrix under the raw upper bounds, not merely a feasible smoothing or an isotonic least-squares fit. In particular F^+ itself belongs to the compatible abstract class, which is always nonempty (the all-zero tail matrix is feasible).

Let $G_r(v)$ and $G_r^+(v)$ be inverse cumulative distribution function (CDF) draws from $P(r,\cdot)$ and $P^+(r,\cdot)$ at a common $v \in (0,1)$. Tail domination implies $G_r(v) \le G_r^+(v)$; monotonicity gives $G_r^+(v) \le G_s^+(v)$ when $r \le s$. With independent uniform draws V_t, define the two chains using these inverse CDFs. Induction yields $S_t \le S_t^+$ at every time when $S_0 \le S_0^+$. Hence every bounded functional nondecreasing in each state along a finite path has no larger expectation for S than for S^+. The coupling is the standard stochastic-order comparison argument (Massey, 1987); here the closed-form test envelope supplies the dominating kernel.

Since P^+ is itself compatible, it attains all these suprema simultaneously over the abstract monotone kernel class. In particular, for any nondecreasing terminal loss g, $\sup_P (P^H g)(r) = ((P^+)^H g)(r)$ for every r, H. On the single event (40), the true kernel is compatible, so all of the resulting bounds hold simultaneously over horizons and losses. □

The sharpness class is important. A tail-bounded monotone kernel need not be realizable by a particular binary record compactor with additional constraints. The theorem is globally sharp *within the stated ordered abstraction*; it is not a claim that the finite-sample upper kernel is the original system's worst physically realizable perturbation. Known transition constraints or lower confidence bounds could tighten a different uncertainty class. Our use of upper tails is deliberate: it yields an explicit single dominating kernel rather than a horizon-dependent row optimizer.

Computing F^+ uses a prefix minimum over thresholds followed by a suffix minimum over rows, costing $O(k^2)$. Successive matrix-vector products compute all endpoint bounds through H in $O(Hk^2)$ work. The count statistic retains joint output information, including the all-error event, with $k(k+1)$ row/threshold probabilities. This is not the number of required calls: each full-vector call contributes to every threshold in its row, and estimation precision still determines sample size.

K.5 PATH RISK VERSUS ENDPOINT RISK

Let queries be independent conditional on the state trajectory, have nondecreasing failure probability $g(S_t)$, occur after each update, and not affect future compaction. For a uniform record query confined to the block, $g(r) = r/k$. Conditional survival through H is $\prod_{t=1}^{H}[1 - g(S_t)]$, so the failure functional is increasing in the entire count path. The theorem therefore gives

$$F_H \le 1 - e_{s_0}^\top (P^+ D)^H \mathbf{1}, \qquad D = \mathrm{diag}(1-g(0),\dots,1-g(k)). \tag{41}$$

This can be much tighter than summing marginal bounds, but need not remain small as H grows. For several possibly dependent blocks, weighted sums of their endpoint bounds are valid for an independent random-record query. Multiplying block survival probabilities would require additional independence and is *not* done. Our path experiments query one specified block only. An observation-dependent query that preferentially selects erroneous records is not covered by a uniform-query formula.

K.6 THE UNPOOLED SAFEGUARD

We can dispense with count-lumpability and true count monotonicity, at a larger statistical cost. Assume instead that every concrete block pattern $e \in \{0,1\}^k$ has a fixed conditional next-count

tail $h_e(\ell)$ given the full history, and that resets share that law. This closed-block count response is stronger than knowledge of individual next-bit means, but weaker than requiring a Markov count process.

Test every concrete row with pre-output counts and obtain upper endpoints $U_e(\ell)$ simultaneously at confidence $1-\delta$, assigning one to untested rows. There are $k2^k$ row/threshold probabilities per block. Define

$$V(r,\ell) = \max_{\substack{|e|\le r \\ m\ge \ell}} U_e(m), \qquad P^{\mathrm{safe}}(r,j) = V(r,j) - V(r,j+1), \tag{42}$$

with endpoint tails one and zero as before.

Proposition 20 (Closed-block safeguard)**.** *On the simultaneous concrete-row event, a chain with kernel P^{safe} dominates the actual count path from the same initial count. Thus the endpoint and path formulas remain valid without count-lumpability or stochastic monotonicity of the actual process.*

Proof. The definition makes V nondecreasing in r and nonincreasing in ℓ, and it dominates each $U_e(\ell)$ when $|e| = r$. It is the smallest monotone tail matrix with this domination property: any such matrix must dominate $U_e(m)$ for every $|e| \le r, m \ge \ell$. On the confidence event, the conditional law of the actual next count, given the whole current history, is dominated by row r of P^{safe} at $r = |e|$. If the dominating chain is currently at $s \ge r$, monotonicity further dominates it by row s. A conditional inverse-CDF coupling at each step therefore preserves the path order by induction, even when the count process is not Markov. □

In the implementation the per-row CP tails already decrease with threshold, so the extra maximum over m is redundant, but the stated formula remains valid for arbitrary raw upper bounds. Compare the operations carefully: justified pooling takes a *minimum* across more-corrupted input levels to exploit known monotonicity; the unpooled safeguard takes a *maximum* to cover heterogeneous concrete states. Replacing one by the other is not a harmless implementation choice. Unseen concrete rows keep their worst-case tail one.

The safeguard is not asserted globally sharp for the original concrete system. Several different concrete rows may maximize different thresholds and times, creating slack. Its role is to retain useful output dependence without adding an unjustified state aggregation. Closed-block abstraction and interval model checking have extensive prior literature (Zacchia Lun et al., 2018); this result identifies the specific sufficient response information and monotone envelope used by the compaction test.

Selecting a certificate after seeing the data. Suppose a marginal method and an ordered method are each valid for the given frozen compactor, at error levels δ_1, δ_2 with $\delta_1 + \delta_2 = \delta$. On the intersection of their confidence events, the pointwise minimum of their endpoint bounds remains valid simultaneously at all horizons. The union bound proves probability at least $1-\delta$ without assuming independent data. Our hybrid methods use $\delta_1 = \delta_2 = .025$. Taking the minimum of two separate .05 bounds is not claimed to have .95 coverage.

K.7 Why the symbolic count assumptions hold

Within each three-record group of the rewrite, retain, and merge processes, the next error in slot i is

$$Y_i = (B_i Z_i + (1-B_i)e_i) \ \vee \ (G_i \wedge e_j \wedge e_\ell), \qquad \{j,\ell\} = [3] \setminus \{i\},$$

where $B_i \sim \mathrm{Bernoulli}(q)$ chooses a fresh proposal, $Z_i \sim \mathrm{Bernoulli}(\eta)$ is its error, and $G_i \sim \mathrm{Bernoulli}(\gamma)$ enables an invalid merge. These draws are independent within this synthetic implementation; for rewrite/retain take $\gamma = 0$, and rewrite additionally has $q = 1$.

Proposition 21 (Exact monotone count closure)**.** *For all $q, \eta, \gamma \in [0,1]$, this group's error count has a sufficient stationary stochastically monotone four-state kernel.*

Proof. The transition law is invariant under jointly permuting the input slots, output slots, and identically distributed random draws. Any two three-bit input vectors with equal count are permutations, so their output counts have the same law. This gives count-lumpability. For nested input vectors $e \le e'$, use identical B, Z, G draws; both the retained-or-proposed error and the invalid-merge event

are nondecreasing in the input bits. Thus $Y(e) \le Y(e')$ coordinatewise. Choose nested representatives for input counts $r \le s$ and invoke count-lumpability to obtain stochastic monotonicity. Fresh independent draws establish the stationary conditional law required in Assumption 2. □

The exact eight-state transition is aggregated and checked against this four-state kernel in the code. This is a check on a known model, not an inference of sufficiency from finite samples. Source/alternative labels in these processes deliberately expose an artificial conflict marker. This modeling assumption is not inferred from tests.

K.8 Input-specific count tails: proof and implementation

This section retains every concrete input error pattern but compresses the *output* information to error-count tails. It neither pools equal-count inputs nor assumes monotonicity of their actual responses. The price is an exponential number of input patterns per block. Robust Bellman recursion under rowwise uncertainty is classical (Givan et al., 2000; Iyengar, 2005; Nilim & El Ghaoui, 2005); the nested-set expectation is the standard random-set construction of Utkin & Destercke (2009). We include the specialization and its attaining law for reproducibility, not as a new upper-expectation principle.

Fix a block of $k \ge 1$ binary records, state space $\mathcal{E} = \{0,1\}^k$, and a table $U_e(\ell) \in [0,1]$. The uncertainty set for row e is

$$\mathcal{Q}_e(U) = \left\{ q \in \Delta_{2^k} : \sum_{y:|y|\ge\ell} q(y) \le U_e(\ell), \quad 1 \le \ell \le k \right\}. \tag{43}$$

It contains the point mass at zero, so it is nonempty. It is compact. For a continuation value $v : \mathcal{E} \to [0,1]$, write

$$w_c = \max_{|y|=c} v(y), \quad a_c = \max_{j\le c} w_j, \quad \bar{U}_e(\ell) = \min_{m\le\ell} U_e(m).$$

Here a_c is the largest continuation loss achievable with *at most* c errors. This running maximum is essential: v need not increase with error count after earlier robust updates.

Proof of Proposition 3. Any feasible output count $S = |Y|$ has tail bounded by $\bar{U}_e$, because $\mathbb{P}(S \ge \ell) \le \mathbb{P}(S \ge m) \le U_e(m)$ for $m \le \ell$. The sequence $\bar{U}_e(\ell)$ is nonincreasing in ℓ. Append the boundary values one and zero to obtain a count variable T_e with precisely these tails. Stochastic dominance of counts and monotonicity of a give

$$\mathbb{E}_q v(Y) \le \mathbb{E}_q a_{|Y|} \le \mathbb{E} a_{T_e} = a_0 + \sum_{\ell=1}^{k} (a_\ell - a_{\ell-1}) \bar{U}_e(\ell).$$

The second inequality can be proved directly by summation by parts; it does not require an unproved coupling theorem. To attain equality, for every c choose a fixed state y_c with $|y_c| \le c$ and $v(y_c) = a_c$. Such a maximizer exists in the finite state space. Draw T_e and output $Y = y_{T_e}$. Its count is at most T_e, so its tail is bounded by $\bar{U}_e \le U_e$, making its law feasible in (43). Its expected value is exactly $\mathbb{E} a_{T_e}$. This proves the closed-form robust update, including attainment for arbitrary $v \in [0,1]^{2^k}$.

For composition, let the conditional output law of the actual process, given its entire past ending at e, belong to $\mathcal{Q}_e(U)$. This law may depend on the past and on time. Define $v_0 = L$ and $v_{h+1} = \mathcal{B}_U v_h$. Conditional expectation and backward induction give $\mathbb{E}[L(E_H) \mid E_0 = e] \le v_H(e)$. Conversely, at every remaining horizon and every state choose the attaining distribution constructed above. These choices form a time-inhomogeneous Markov process with rows in the allowed sets and attain $v_H(e)$. Thus the recursion is horizon-sharp over the *rectangular class allowing a different compatible row law at different times*. This does not prove attainment by one stationary kernel, or by a particular record compactor. That distinction separates this theorem from Theorem 16.

For a fixed test design, collect independent reset outputs at each concrete input. A tail indicator is Bernoulli with its fixed conditional probability, even if different indicators on one call are dependent. One-sided CP endpoints with total expenditure δ across the $k2^k$ row/threshold pairs cover all true tails simultaneously. Missing rows get one. On that event, the above deterministic induction

applies to every H, and to every loss $L \in [0,1]^{2^k}$ with no additional union bound. For multiple closed blocks, allocate confidence over their total row/threshold count and average each block's uniform-query bound with its fraction of records. Cross-block independence is not required for this expectation. □

Proposition 22 (Dominance over the count-only unpooled safeguard)**.** *Let* $V(r,\ell) = \max_{|e|\le r,\, m\ge \ell} U_e(m)$ *and let* P_V *be its count kernel. For any nondecreasing* $g : \{0,\dots,k\} \to [0,1]$ *and* $L(e) = g(|e|)$*,*

$$(\mathcal{B}_U^H L)(e) \le (P_V^H g)(|e|) \qquad \textit{for all } e, H.$$

This compares bounds formed from the same *upper-tail table and confidence event.*

Proof. The envelope V is nondecreasing in r and nonincreasing in ℓ, so P_V preserves increasing functions. Suppose $v_h(y) \le g_h(|y|)$, where $g_h = P_V^h g$ is increasing. Every feasible row in $\mathcal{Q}_e(U)$ has output count stochastically bounded by $P_V(|e|,\cdot)$, since $U_e(\ell) \le V(|e|,\ell)$. Taking its expectation of g_h bounds the robust maximum of v_h by $g_{h+1}(|e|)$. Begin at equality when $h = 0$ and induct. □

The inequality can be strict. For $k = 2$, list inputs in the order $00, 10, 01, 11$ and choose raw tail rows $(.2, .05), (.9, .05), (.4, .4), (.6, .1)$. At $H = 8$, starting clean with uniform record loss, the input-specific recursion gives 0.34, whereas the count-only safeguard gives 0.45. The latter separately maximizes the first and second tail over different one-error inputs, allowing a combination that need not belong to any single input row. These values are a deterministic illustration, not a calibration experiment.

Complexity and numerical checks. Compute the prefix minima once in $O(k2^k)$ time. Each backup forms group maxima and their running maximum, then applies one tail matrix to the k increments, costing $O(k2^k)$ time and $O(k2^k)$ storage. Through H updates, time is $O(Hk2^k)$. A full joint row table has 4^k entries. These are storage and arithmetic counts, *not* a sample-complexity optimality theorem. The implementation validates probabilities and input shapes, handles unseen cells conservatively, and compares the formula with independent primal linear programs. Bounds are evaluated in ordinary floating-point arithmetic, not certified interval arithmetic.

K.9 Tests that choose inputs and stopping times adaptively

The deployment horizon and the number of test calls are different clocks. A fixed-design table confidence event already supports every deployment horizon. It does not, by itself, remain valid if the evaluator repeatedly inspects fixed-sample intervals and stops when they look favorable. We use a standard mixture confidence sequence (Howard et al., 2021; Waudby-Smith & Ramdas, 2024); no new confidence-sequence construction is claimed.

Predeclare m row/threshold streams. The next reset row is selected using the test history before its next output is seen. For stream j, every new observed indicator has conditional Bernoulli probability p_j given the entire preceding test history and selection. A single call can reveal several dependent indicators. A stream not observed on a call remains unchanged. These conditions allow adaptive row allocation but exclude nonstationary per-row laws, selecting a sample after seeing its outcome, or changing the compactor while reusing its old calibration.

Let $N_j(n), S_j(n)$ denote the trial and success counts after global test call n. For candidate $p \in (0,1)$ define

$$Z_{j,n}(p) = \frac{\mathrm{B}(S_j(n) + 1/2, N_j(n) - S_j(n) + 1/2)}{\mathrm{B}(1/2,1/2)\, p^{S_j(n)}(1-p)^{N_j(n)-S_j(n)}}. \tag{44}$$

This is a beta $(1/2, 1/2)$ mixture of Bernoulli likelihood ratios, not a Bayesian credible interval. The displayed confidence set is obtained by a frequentist crossing inequality.

Theorem 23 (Two-clock certificate)**.** *Choose* $\alpha_j > 0$ *with* $\sum_j \alpha_j \le \delta$*. Let* $U_{j,n}$ *be the upper endpoint of* $\{p : Z_{j,n}(p) \le 1/\alpha_j\}$*, with the continuous boundary convention and* $U = 1$ *for an unobserved stream. Then*

$$\mathbb{P}\{p_j \le U_{j,n} \textit{ for every } j, n\} \ge 1 - \delta. \tag{45}$$

On this same event, any of the deterministic certificate maps in Theorems 15 and 16, and Proposition 3 is valid for every test time n and deployment horizon H, when its respective structural assumptions hold. The same is true at arbitrary output-dependent stopping times and after replacing each endpoint by its running minimum over test times.

Proof. Fix a stream and a candidate alternative $q \in (0, 1)$. Under its true probability $p_j \in (0, 1)$, the likelihood ratio is unchanged when the stream is not sampled. When it is sampled, its multiplicative factor is q/p_j on success and $(1-q)/(1-p_j)$ otherwise; conditional expectation is one. Predictable selection is exactly what permits this conditional calculation. Thus the likelihood-ratio process is a nonnegative martingale in the *global* test filtration. Integrating with respect to a beta $(1/2, 1/2)$ distribution and applying conditional Tonelli gives (44), another nonnegative martingale starting at one. At $p_j = 0$ or 1, define the ratio on the true support by continuity; it is a nonnegative supermartingale, which is sufficient.

For completeness, the maximal inequality follows by stopping this nonnegative supermartingale at its first crossing of $1/\alpha_j$, truncated at N. Its expectation is at most one, so the probability of a crossing by N is at most α_j. Letting N increase gives the same bound over all test times. Outside that crossing event, the true p_j lies in the confidence set and hence below its upper endpoint at every time. A union bound over the predeclared streams proves (45), without requiring stream independence.

Every deterministic certificate proof conditions only on its table of upper bounds covering the true conditional quantities. That property holds simultaneously for all n on (45). Apply the appropriate proof at each n; its conclusion already holds for all H. Taking a running minimum retains upper coverage at all times. Therefore no additional union over test times, deployment horizons, or stopped versions of the same table is needed. Combining different certificate families still requires that their input statistics be jointly covered; reusing δ independently for each family does not establish a joint $1 - \delta$ claim. □

The likelihood-ratio logarithm is convex in p, and its minimum is at S/N. At that point it is at most zero because a mixture likelihood is no larger than the maximum likelihood. Thus its sublevel set at $\log(1/\alpha)$ is an interval. For $S < N$ the upper endpoint is the root above S/N; for $S = N$ it is one. The implementation uses 55 bisection steps and returns the upper bracket. Unit tests check the defining equation, binomial expectation, symmetry, boundaries, and exact crossing probabilities on all length-ten paths.

Allocation is a heuristic, validity is not. After visiting every count row once, the adaptive scheduler estimates how much a batch would reduce the current endpoint certificate. For row r, it moves the current upper tail toward the empirical tail by fraction $b/[2(n_r + b)]$, recomputes the endpoint bound, and chooses the largest hypothetical decrease; ties use the least sampled row. The hypothetical table is used *only* to select the next row, never to issue a certificate. Actual bounds use the mixture endpoints and their running minima. The theorem covers this predictable rule, but does not make it cost-optimal. In particular, our merge experiment reaches its target later under adaptive scheduling than under balanced scheduling. The respective mean stopping costs are 1,540.3 and 1,390.9 calls. At the common final budget, adaptive mean row allocations are $(1603.2, 1141.3, 677.3, 674.1)$ versus $(1024, 1024, 1024, 1024)$ for balanced sampling. The one-batch proxy favors current sensitivity and can neglect a row that becomes limiting later; the allocation pattern is consistent with this explanation but is not a causal ablation or a proof that these final counts caused the delay.

K.10 A LEARNED JOINT BOUNDED-RECORD COMPACTOR

The learned benchmark creates dependent retained outputs by choosing a complete record set jointly from observations of two independent latent facts.

Draw independent bits a, b and define the true record triple $x = (a, b, a \oplus b)$. The third record is a consistency check, not an additional independent fact. An update receives three old records, three new proposals with observed binary quality tags, and three irrelevant-key records; it returns exactly three task-key records. Distractor values cannot affect the schema-constrained scorer, so the batched simulator analytically omits drawing their values. The explicit record interface accepts these

distractors and discards their keys. This is a nine-to-three record-selection task, not natural-language compression or a learned retrieval system.

For source i, a high-quality tag has probability $(.70, .45, .60)_i$. Given low/high tag, the proposal error probabilities are respectively $(.16, .01)$, $(.24, .035)$, and $(.12, .02)$. Independent eviction flags with probabilities $(.10, .18, .25)$ prevent an unavailable old record from being retained. Eviction is an explicit part of the generator, ensuring that copying old memory forever is not a trivial zero-error policy. There is no world drift: latent facts remain fixed within a deployment.

A candidate is a three-bit mask choosing an old record or proposal for each key. Masks violating eviction constraints are unavailable. The six scorer features are three source-specific proposal indicators, mean selected quality tag, whether the candidate triple has even parity, and the fraction of values changed relative to the old triple. The policy samples one whole candidate from a softmax of a linear feature score. It never receives latent truth, the current error mask, or a correctness indicator. Quality tags are noisy, not authoritative evidence.

For training only, the teacher selects the available candidate with the fewest errors against known synthetic truth. Ties prefer fewer proposals, then the fixed mask order. Each of 18 separately fitted scorers uses 6,000 independently generated training examples, cross-entropy loss, and $0.01\|\theta\|_2^2/2$ regularization. Six training seeds are used for each old-record corruption probability $.02, .35, .65$. Optimization uses deterministic bound-constrained L-BFGS (L-BFGS-B) with analytic gradients and checked convergence. Each scorer is frozen before five independent test datasets are generated. Test queries do not affect the memory process.

Proposition 24 (Sufficient error state and exact joint evaluation)**.** *The joint scorer's conditional error-output law is independent of the latent codeword x. Therefore its exact eight-state error transition kernel can be computed with canonical truth 000. Its joint law need not factor into its output marginals and need not be count-lumpable.*

Proof. Changing truth from one valid codeword to another translates by bitwise exclusive OR (XOR) old and proposed triples by a valid codeword. It leaves masks, source tags, eviction constraints, and changed-value indicators unchanged. Both valid codewords have even parity, so their XOR does as well; candidate parity is invariant under this translation. Every candidate's features and hence softmax probability are unchanged. Outputting the same sampled mask translates its values by the same codeword and leaves the resulting error triple unchanged. The source noise law, expressed as XOR errors conditional on tags, also remains unchanged. This proves sufficiency and justifies canonical truth evaluation.

For each of eight current error patterns, enumerate all 8 tag patterns, 8 proposal-error patterns, and 8 eviction patterns. Multiply the explicitly independent source/eviction probabilities and sum each mask's softmax-weighted mass into the corresponding next error pattern. The resulting row is the actual joint law. Because one sampled mask selects an entire triple using its parity feature, there is no conditional-output-independence argument. Asymmetric source parameters likewise give no count-lumpability argument. The fitted models exhibit departures from both properties; certification assumes neither property. □

Exact endpoint risks use powers of this joint transition, never powers of the product of its marginal tables. The same 18 fitted scorers are evaluated at budgets $A = 512, 2048, 8192, 32768$, with five independent test datasets per scorer and budget, balanced over all eight concrete input patterns. Comparisons across budgets reuse the fitted scorers and are not independent model replications. All local methods consume the same full-vector outputs. Marginal and full joint-box baselines use their respective simultaneous confidence events; row-aware and count-only unpooled tail methods share precisely the same tail event. Direct rollout methods use $\lfloor A/H \rfloor$ independent length-H trajectories from the genuine joint kernel. A uniformly sampled endpoint record gives an exact-binomial baseline; averaging endpoint errors gives a bounded-mean baseline. The 6,000 training examples per scorer are common to its tests and are excluded from the reset-call budget.

We average the five test replicates within each trained scorer before computing standard errors over the six independently trained scorers in a regime. We do not treat 30 test outcomes from six models as 30 independent trained models. None of these models is a language model, and none is used as evidence that the structural assumptions hold for natural-language summaries.

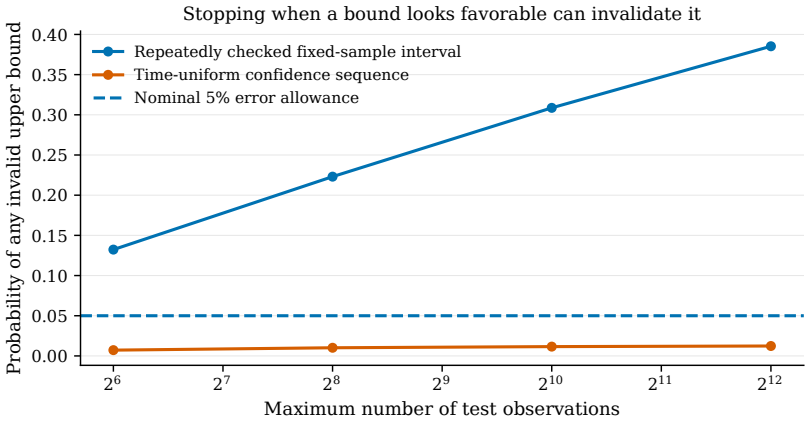

Figure 3: Exact probability of any upper-bound failure under repeated inspection, for Bernoulli error probability $p = .20$. Inspection occurs after every observation, not only the four plotted horizons. Fixed-sample coverage at each predetermined time is not coverage at an output-dependent stopping time.

K.11 Validation protocol and exact stopping calculation

Independent geometry check. For 160 indexed random cases with one to four bits, we compare every row of the closed-form robust update against an independent SciPy linear program with the same raw tail constraints and an arbitrary bounded continuation loss. We separately draw a full joint kernel, add nonnegative tail uncertainty, and compare exact risk, the input-specific certificate, and the count-only safeguard through 32 updates for all initial states. This checks implementation equivalence, not a new statistical confidence theorem.

Peeking is calculated exactly. For each $p \in \{.01, .05, .2, .5\}$, let $a_n(s)$ be the probability of having s successes after n trials without an earlier upper-confidence violation. Before removing violations, update

$$\tilde{a}_{n+1}(s) = (1-p)a_n(s) + p\, a_n(s-1).$$

Set entries to zero if the upper endpoint is strictly below p. For fixed-sample CP endpoints this means $\mathbb{P}\{\mathrm{Bin}(n,p) \le s\} < .05$. For the mixture upper endpoint it means $s/n < p$ and $Z_n(p) > 20$. Start with $a_0(0) = 1$. The exact probability of *any* violation through N is $1 - \sum_s a_N(s)$. We check at every $n \le 4096$, retaining summaries at 64, 256, 1024, and 4096. These are exact finite recursions with floating-point arithmetic, not Monte Carlo trials; different reported horizons share the same crossing event history. Each p is a separate nominal five-percent problem.

Adaptive design. For each of rewrite, retain, and merge, 60 independently generated row-specific output banks are each used for a paired balanced/adaptive comparison. The n th observation from a given row is the same within a pair, but distinct pairs are independent. The test budget is 4,096 calls in batches of 64, with deployment horizon 256. Target bounds are $.08, .12, .35$ respectively. We record the first successful certification time and continue to the budget for comparable final bounds. If no target is met, the recorded cost is right-censored at the budget; certification status is separate. All 180 pairs meet their prescribed targets under both schedules in this run. Paired standard errors use 60 independent banks per policy, not 120 independent schedules. The fixed-design CP baseline is computed only for the predetermined balanced schedule at its final budget; adaptive row counts do not justify that fixed-design calculation, even at a fixed total budget.

L Selected diagnostic curves and tables

The plots and tables here illustrate contrasts across budgets, horizons, and confidence-region constructions. Learned-regime means use six fitted models and five tests per model; feature-

augmentation intervals instead use 24 independent test datasets for a fixed controlled process. Neither kind of descriptive interval is a simultaneous confidence band.

L.1 Likelihood geometry and modest propagation gains

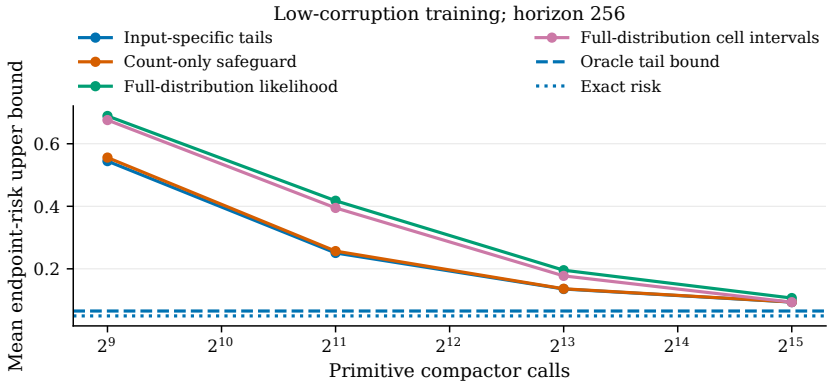

Figure 4: Learned low-corruption regime, $H = 256$. The gain from row-specific propagation is modest. Most small-budget slack is statistical; count-only and full-distribution confidence-region rankings change with budget.

Table 3: Count and full-joint likelihood regions use the same calibration principle on different alphabets. The final row-tail column changes both the retained information and confidence geometry.

Regime	Calls	Count likelihood	Joint likelihood	CP row tails
Low	512	0.62	0.69	0.54
Low	2,048	0.35	0.42	0.25
Low	8,192	0.18	0.20	0.14
Low	32,768	0.11	0.11	0.09
Medium	512	0.33	0.42	0.26
Medium	2,048	0.15	0.20	0.12
Medium	8,192	0.09	0.11	0.08
Medium	32,768	0.07	0.07	0.06
High	512	0.28	0.35	0.22
High	2,048	0.14	0.17	0.12
High	8,192	0.10	0.11	0.09
High	32,768	0.08	0.08	0.07

L.2 Capacity, enumeration, and a guarded hidden cache

Table 4: Enumeration costs for 2^k concrete inputs at 512 resets per input: $512\,2^k$ calls, $k2^k$ tail entries, and 4^k joint entries. This is an implementation/model-class choice, not a lower bound for the shared-affine family; Appendix I.4 gives a two-input stronger-assumption control.

k	Inputs	Calls	Tail entries	Joint entries
2	4	2,048	8	16
3	8	4,096	24	64
4	16	8,192	64	256
6	64	32,768	384	4,096
8	256	131,072	2,048	65,536
10	1,024	524,288	10,240	1,048,576
12	4,096	2,097,152	49,152	16,777,216

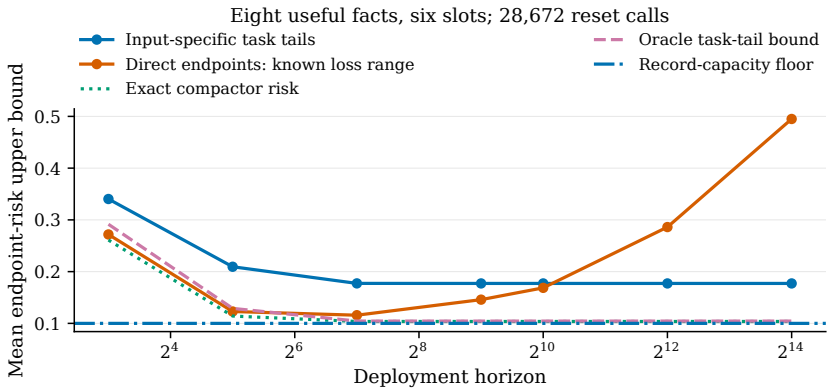

Figure 5: Six slots, eight useful facts, 28,672 calls. Range-aware direct testing remains tighter at 1,024 steps, while reusing the local test is tighter at 4,096. The capacity floor, policy error, and certificate slack are distinct.

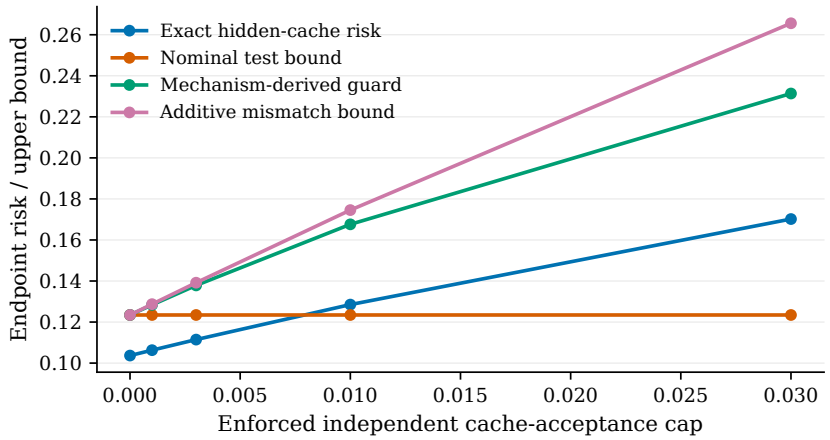

Figure 6: Hidden-cache deployment at $H = 512$, 114,688 nominal test calls. The allowance comes from a fresh independent gate enforced before the ordinary output, not from the exact evaluator. The ordinary-branch contract and extra snapshot storage remain explicit requirements.