BANKING EXPIRING COMPUTE: HOW REUSE SHAPES THE VALUE OF PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowing the expected work of every feasible preparation need not reveal which one is most likely to finish within a fixed budget. We study banking expiring compute: using a temporary allowance to retain intermediate results, reusable while their inputs remain unchanged, under limited storage. For ordered associative computations, both the learner and an informed comparator may choose the evaluation tree and stored results. Under ideal reuse, individual interval-validity probabilities exactly characterize all feasible mean work; joint validity of collections that fit preparation and storage budgets characterizes full work distributions. We construct laws agreeing on every feasible mean, yet even randomized mean-only decisions have exponentially smaller worst-case completion probability than informed decisions with the same resources. On this family, increasing only the execution budget makes a history-free randomized policy exponentially close to optimal in a specified slack range. Logarithmically many complete i.i.d. histories from the same update law recover an optimal preparation with high probability. More generally, physical reuse capacity bounds the sample complexity of near-optimal preparation, with matching worst-case bounds in identified ideal-interface regimes. The upper bound has no additional input-count dependence beyond resource parameters and survives fixed nonnegative checking and restoration costs. Together, these results tie the value of predictive information and the statistical cost of learning a preparation to feasible reuse and execution budgets.

1 INTRODUCTION

Compute available now may be unavailable when a future request arrives. Continual computation and sleep-time compute exploit this opportunity to prepare ahead of demand (Horvitz, 2001; Lin et al., 2025). But storage is limited, and saved results can become invalid as inputs change. With a hard budget for future work, the question goes beyond how much computation to save: *what must be predicted to choose a preparation that will finish?*

An offloaded mixture-of-experts model illustrates the information question: experts must be chosen for residency before requests arrive (Eliseev & Mazur, 2023). Individual invocation frequencies need not reveal which requests fit a transfer budget; which experts are needed together can matter.

Now suppose we know the exact expected remaining work of *every* feasible preparation. Even this information can leave the best completion decision unresolved. We construct update distributions that agree on all these means yet require different preparations. Neither more accurate means nor a different decision rule can resolve this ambiguity. Figure 1 illustrates the missing joint information.

We establish the exact theory for ordered associative computations, including interval aggregates and compact affine-transition composition (Tangwongsan et al., 2017; Smith et al., 2023). A preparation chooses an evaluation tree and a bank of saved results, fixed before inputs change. Shared preparation is charged once; nested records let a valid child survive an invalid parent. All comparisons use the same preparation and storage budgets.

Execution structure identifies what information is sufficient. Under cost-free checking and restoration, individual interval-validity probabilities exactly characterize all feasible mean work, whereas joint validity of collections that fit the budgets characterizes all feasible work distributions. This identifies information sufficient for every deadline without specifying the entire update law; it does

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.

not claim minimality for one deadline. The relevant dependence is therefore tied to what can actually be prepared and reused.

The deadline determines how much missing information matters. In a constructed family, even randomized decisions based on all feasible means have exponentially smaller worst-case completion probability than law-aware decisions with the same resources. This holds over every legal tree and nested bank. Increasing only the execution budget makes a history-free randomized policy exponentially close to optimal in a specified range. On the same family, logarithmically many complete, independent histories recover an optimal bank with high probability.

Learning a preparation need not mean learning the environment. Physical reuse capacity bounds the histories needed for near-optimal completion in terms of storage, preparation shortfall and execution slack. There is no additional input-count term, though these resources may grow with input size. Worst-case bounds match in identified regimes; the upper bound survives fixed nonnegative checking and restoration costs. Histories must be complete, independent and drawn from the future's law; statistical learnability does not guarantee efficient planning. Together, these results make prediction an execution-dependent design choice: resource constraints shape both the relevant distinctions between futures and the histories needed to learn a preparation.

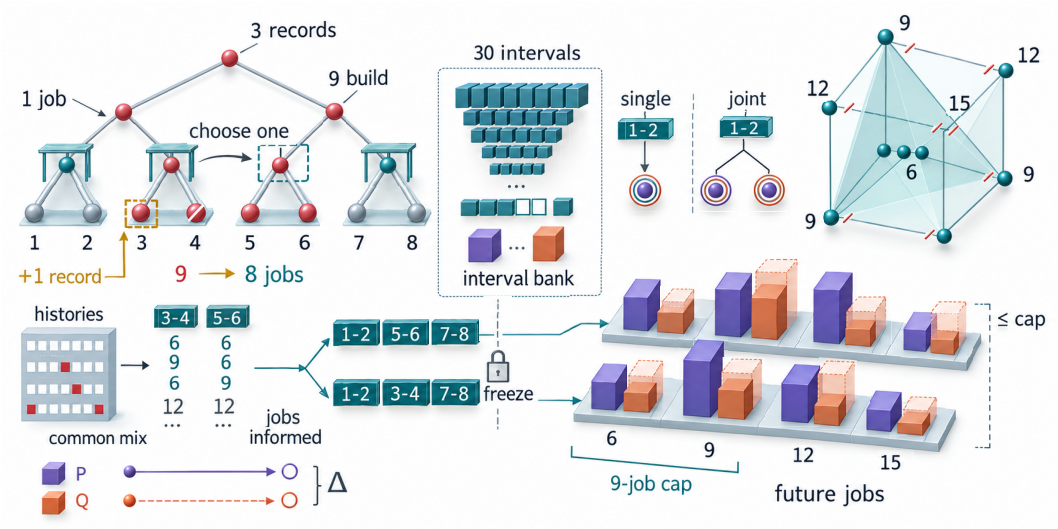

Figure 1: **Equal means, different completion decisions.** Both banks use nine Build jobs and three records; gold indicates a separate fourth-record variant. The full-support laws P (purple) and Q (orange) match every feasible mean but differ in joint validity and completion. Cube labels give work; the nine-job cap separates success from failure. Slashes mark invalidity. Heights and perspective are schematic, not empirical estimates; exact probabilities appear in Appendix H.2.

2 Related Work

Accurate predictions and useful information. Decision-focused learning aligns prediction with downstream optimization (Donti et al., 2017; Elmachtoub & Grigas, 2022); Cameron et al. (2022) show that exact target means can still yield poor decisions under nonlinear objective parameterization. We instead reveal the exact expected work of every feasible plan and allow any randomized decision rule. The limitation is in this entire observation, not a particular objective substitution. Moment optimization and correlation-gap analysis already show why dependence beyond given summaries matters (Bertsimas & Popescu, 2005; Agrawal et al., 2012). We identify which summaries legal reuse distinguishes and how changing only the execution budget changes the decision value of the missing information.

Physically feasible reuse. Sparse prefix caching minimizes expected recomputation from the deepest reusable checkpoint, with exact optimization and learning guarantees from overlap-depth

observations (Shirokikh & Nikolenko, 2026). Tail-optimized caching studies threshold-referenced excess latency under online cache decisions in a birth–death conversation model (Zhang et al., 2025). Our bank is fixed before changes; success means finishing within a work budget. The information comparison covers every feasible ordered tree and nested bank under common resources, rather than a selected set of policies. This physical restriction also determines which joint-validity events can be used.

Learning a decision rather than a distribution. Data-driven algorithm selection bounds observations needed to choose an effective algorithm (Gupta & Roughgarden, 2017). We use this framework; the new input is a bound on decision-class complexity from physical reuse capacity. Bounded-bag analysis constrains individual observation size (Sabato & Tishby, 2012). Our complete histories need not be sparse and may contain arbitrary within-history dependence. Instead, we bound how many input locations can each change while one preparation still finishes. Thus execution resources, rather than an assumed bound on changes per history, control the learning guarantee.

3 Which information does legal execution distinguish?

Execution and budgets. There are m ordered, versioned inputs and a supplied deterministic, exactly associative operation. A legal tree T is full binary and order-preserving. Each leaf and combination costs one; cold work is $n = 2m-1$. A plan $\pi = (T, A)$ retains at most M completed subtree results. For operations $D_T(v)$ and input interval I_v below v, Build from an unchanged snapshot evaluates shared dependencies once:

$$C_{\mathrm{build}}(\pi) = \left| \bigcup_{v \in A} D_T(v) \right| \le b = n - d. \tag{1}$$

Budgets are nonnegative integers, $0 \le d \le n$. Records contain exact values, endpoints and version descriptors with $O(\log m)$ metadata; planning, history reading and transient workspace are separate resources. After expiry, the change mask $J \subseteq [m]$ identifies inputs whose versions change. A record skips its subtree exactly when $J \cap I_v = \varnothing$; otherwise execution recurses below it. Access is free until Section 5. Repair, inverse updates, arbitrary snapshot encodings, new saved state and remote computation are excluded. Changed versions invalidate numerically coincident values too. Exact associativity preserves cold output, not necessarily floating-point bitwise identity.

Let $C_\pi(J)$ be the work after changes J. For a law P and cap L, mean work, completion and slack are

$$W_P(\pi) = \mathbb{E}_P C_\pi(J), \qquad F_P(\pi) = \mathbb{P}_P(C_\pi(J) \le L), \qquad \Delta = L - d. \tag{2}$$

Let $\mathrm{OPT}_P = \max_{\pi \text{ feasible}} F_P(\pi)$. The informed comparator knows P, not the realized J. All decision coins are independent of future changes.

One execution identity. Assign each operation to its nearest retained ancestor, or a virtual component if none exists. Real sizes are $w_v > 0$ and the virtual size is a. If r maximal retained roots cover $m - u$ inputs, and v_i are the maximal retained descendants of v, then

$$C_\pi(J) = a + \sum_{v \in A} w_v \mathbf{1}\{J \cap I_v \neq \varnothing\}, \quad a = n - C_{\mathrm{build}}(\pi) = 2u + r - 1 \ge d, \tag{3}$$

$$w_v = 2|I_v| - 1 - \sum_i (2|I_{v_i}| - 1), \qquad a + \sum_v w_v = n. \tag{4}$$

An invalid record has only invalid retained ancestors, so exactly its assigned operations execute. Interval descriptors determine costs; conversely, every laminar interval family admits an ordered binary refinement, without privately encoding unsaved parts.

For example, affine transitions compose as $(A_X, b_X) \star (A_Y, b_Y) = (A_Y A_X, A_Y b_X + b_Y)$. Reuse requires compact closure and endpoint or sparse-output queries; a general multilayer model need not satisfy this interface. The unit-cost analysis counts operations, not bytes or latency. In Figure 1, storing $[1, 2], [3, 4], [7, 8]$ costs nine Build jobs, with $a = 6$ and three weights of size three. Adding the nested record $[3, 3]$ costs no extra Build but a fourth slot: when inputs 4 and 5 change, work falls from nine to eight because the valid child survives its invalid parent.

Let $\nu_P(U) = \mathbb{P}_P(J \cap U = \varnothing)$, and let $\mathcal{I}_b$ contain the nonempty intervals with $2|I| - 1 \le b$. Write

$$\mathcal{U}_{b,M} = \{U \subseteq [m] : \mathrm{runs}(U) \le M,\ 2|U| - \min(M, |U|) \le b\}. \tag{5}$$

These are exactly the unions of at most M jointly Build-feasible saved intervals, including $\varnothing$.

Theorem 1 (Information distinguished by legal execution)**.** *For integers $M \ge 1$, $0 \le b \le n$, and any update laws P, Q,*

$$W_P(\pi) = W_Q(\pi)\ \forall \pi \textit{ feasible} \iff \nu_P(I) = \nu_Q(I)\ \forall I \in \mathcal{I}_b, \tag{6}$$

$$\mathcal{L}_P(C_\pi) = \mathcal{L}_Q(C_\pi)\ \forall \pi \textit{ feasible} \iff \nu_P(U) = \nu_Q(U)\ \forall U \in \mathcal{U}_{b,M}. \tag{7}$$

The second equivalence concerns complete work distributions, not necessity at one specified cap.

Proof. Equation (3) gives $W_P(\pi) = n - \sum_v w_v \nu_P(I_v)$; single-record plans probe every $I \in \mathcal{I}_b$, proving (6). Validity intersections are union avoidance. A selected collection with r maximal roots has Build $2|U| - r \le b$ and $r \le \min(M, |U|)$, placing U in $\mathcal{U}_{b,M}$. Inclusion–exclusion recovers the validity vector's law, hence its work law. Conversely, split the runs of nonempty $U \in \mathcal{U}_{b,M}$ into $\min(M, |U|)$ disjoint intervals. Their roots cost at most b; positive weights make minimum work equivalent to avoiding U, recovering $\nu_P(U)$. Empty-set avoidance is one. □

One record has a two-point work law, determined by its mean.

Theorem 2 (Complete mean information across a deadline transition)**.** *For each integer $g \ge 2$, there are laws P_H, indexed by $H \in \binom{[2g]}{g}$, at budgets $m = 8g,\ M = g + 1,\ b = 9g - 1,\ d = 7g$. Every feasible plan has the same mean across H. Put $F_{H,e}(\pi) = P_H(C_\pi \le d + e)$ and define*

$$V_g(e) = \sup_{\mathcal{A}\ \text{mean-only}} \inf_H \mathbb{E}_{\mathcal{A}} P_H(C_{\widehat{\pi}_{\mathcal{A}}} \le d + e). \tag{8}$$

The procedure knows the family and all feasible means, but no other H-dependent information. For integer $0 \le e < g/2$, the informed optimum is $1/2$. One history-free policy attains $V_g(e)$ for every H and cap, averaging over its coins. With $D(x\|p) = x\ln(x/p) + (1-x)\ln((1-x)/(1-p))$ and $0 \ln 0 = 0$,

$$V_g(e) \le \exp[-gD(e/g\|1/4)], \qquad 0 \le e < g/4, \tag{9}$$

$$0 \le \tfrac{1}{2} - V_g(e) \le \tfrac{1}{2}\exp[-gD(e/g\|1/4)], \qquad g/4 < e < g/2. \tag{10}$$

For $0 < \delta < 1$, $s \ge \lceil \log_2(g(g+1)/(2\delta)) \rceil$ complete i.i.d. histories suffice to recover one bank optimal for every integer $0 \le e < g/2$, with probability at least $1 - \delta$.

Why no other tree escapes the comparison. Interleave $2g$ candidate inputs with always-changing separators and append $4g$ stable inputs. Candidates in H share a fair change bit; outsiders have independent fair bits. Every interval's validity is zero, one or $1/2$, independently of H. Delete always-invalid records. If a suffix record remains, at most g candidate leaves remain: replace the suffix records by its whole root and fill to g candidates, preserving or increasing skipped work. Otherwise at most $g + 1$ candidate operations are skipped, dominated by any full-suffix/g-candidate bank. Every replacement admits an ordered tree within b, M (Lemma A.1). Thus normalization covers all trees and nested banks; actual usage can increase within the allowances, and a positive-slack successful bank need not originally be normalized.

A normalized bank selecting R with $h = |R \setminus H|$ outsiders has work $d + B_h + (g - h)Z$, where $B_h \sim \mathrm{Bin}(h, 1/2)$ and $Z \sim \mathrm{Bernoulli}(1/2)$ are independent. Complementing the fair bits makes excess work symmetric around $g/2$, so completion below that threshold is at most $1/2$, attained by $R = H$. Write $\beta_h(k) = \mathbb{P}(\mathrm{Bin}(h, 1/2) \le k)$, extended by zero for $k < 0$ and one for $k \ge h$. Then

$$V_g(e) = \frac{1}{2\binom{2g}{g}} \sum_{h=0}^{g} \binom{g}{h}^2 [\beta_h(e) + \beta_h(e - g + h)]. \tag{11}$$

Every fixed R has this average under uniform H, bounding any randomized procedure's worst-law value. Conversely, uniform R attains it for every fixed H; the adversary chooses H before the coins. Appendix A proves the tail and critical-window bounds. Identical-column recovery fails with probability at most $g(g+1)2^{-s}/2$ and reads $\Theta(gs)$ bits.

Figure 2 (left) shows the family-specific transition. Pairwise covariances identify H, so the theorem limits procedures restricted to means; it is not a hierarchy of information orders.

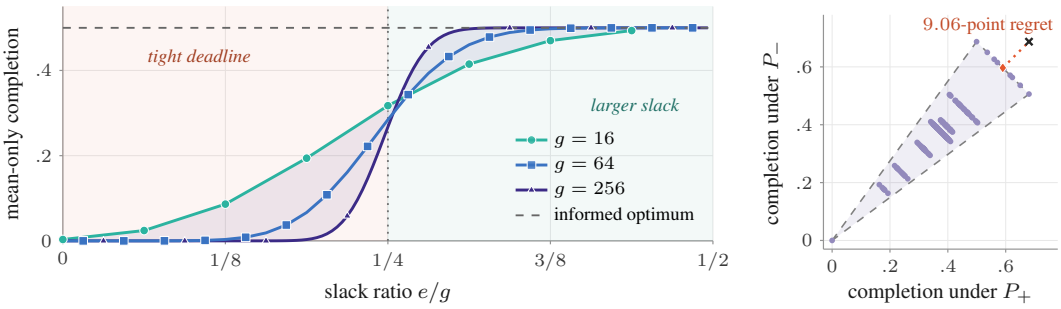

Figure 2: Left: exact mean-only completion in Theorem 2 against the slack ratio e/g, for three group sizes; throughout $e < g/2$ informed completion is $1/2$ (dashed), the dotted line is the limiting critical ratio $1/4$, and points evaluate the finite formula, not sampled trials. Right: the matched pair of Section 6, identical statistics with incompatible optimal preparations. Points are feasible banks; the dashed hull allows common randomization; the cross combines separately informed optima, not one plan; the diamond attains the 9.06-point worst-law regret, the dotted segment to the cross. Exact finite values, not prevalence.

No deterministic input is needed. Mix P_H equally with independent $3/4$-rate changes: every mask has positive probability and every feasible mean remains common. Proposition A.3 bounds mixed minimax completion by $V_g(e)/2 \le \bar{V}_g(e) \le V_g(e)/2 + r_g/2$ and informed completion by $1/4 \le \overline{\mathrm{OPT}}_{H,e} \le 1/4 + r_g/2$, where $r_g = (3/4)^{4g}$. Indeed, any legal plan needs $|J| \le 4g$ to finish below $8g$, an event of probability at most r_g in the independent component. Both sides of the transition persist. With probability $1-\delta$, $O(\log(g/\delta))$ histories recover a bank with regret at most $r_g/2$ throughout the band, optimal at zero slack; Appendix A.1 gives the explicit sample bound. This does not cover arbitrary perturbations or paid reuse.

4 Learning a Preparation, Not the Entire Update Law

The preceding results concern which information supports a decision. We now ask how many complete histories suffice to learn that decision directly. Under the same execution model, complete i.i.d. masks permit arbitrary within-mask correlation and counterfactual evaluation of every feasible plan. Physical capacity limits the complexity of these decisions.

Physical capacity. Let $S_\pi = \{j : C_\pi(\{j\}) \le L\}$. A positive interval sum changes from failure to success only initially or after an interval ends, so S_π has at most $M+1$ runs. If $a > L$, then S_π is empty. Otherwise at most $u \le (a+1)/2$ uncovered inputs and $L-a$ owned inputs per saved component can be timely. Hence $|S_\pi| \le (a+1)/2 + M(L-a) \le (d+1)/2 + M\Delta$, giving

$$|S_\pi| \le G := \min\{m, \lfloor (d+1)/2 \rfloor + M\Delta\}, \quad 0 \le d \le L < n, \tag{12}$$

$$\tfrac{1}{12}\min\{m, d + M\Delta\} \le \max_\pi |S_\pi| \le G. \tag{13}$$

The lower bound in (13) additionally requires $2 \le M \le \lfloor m/2 \rfloor$ and $M \le d \le m$. Uncovered inputs supply the d term; small recomputable blocks supply $M\Delta$. Appendix C proves both directions. This constrains timely inputs, not the number of trees: $M = 3, d = 4, \Delta = 1$ gives $G \le 5$ regardless of input count.

Given s histories, we evaluate an independent future from the same unknown P. Let $\mathcal{H} = \{\mathbf{1}\{C_\pi(J) \le L\} : \pi \text{ feasible}\}$.

Theorem 3 (Physical capacity controls history complexity)**.** *For $M \ge 1$ and $0 \le d \le L < n$, an absolute constant c_0 satisfies*

$$\mathrm{VCdim}(\mathcal{H}) \le c_0 M \log_2(2 + d/M + \Delta). \tag{14}$$

For $\epsilon, \delta \in (0,1)$, empirical maximization gives $F_P(\widehat{\pi}) \ge \mathrm{OPT}_P - \epsilon$ with probability at least $1-\delta$ whenever

$$s \ge c_0\,[M \log_2(2 + d/M + \Delta) + \log(1/\delta)]/\epsilon^2. \tag{15}$$

Expected regret is $O(\min\{1, \sqrt{M \log(2 + d/M + \Delta)/s}\})$. An additive τ empirical approximation adds at most τ to either bound.

The capacity-to-learning step. If $J_1, \ldots, J_v$ are shattered, the all-positive labeling has a feasible witness π_+ completing all of them. For each j in $Z = \bigcup_i J_i$, monotonicity gives $C_{\pi_+}(\{j\}) \le L$. Hence $|Z| \le G$ by (12). Project every other bank onto this ordered union. At most M laminar intervals remain; projected duplicates are allowed. There are at most $\sum_{r=0}^{M} 4^r \binom{G+2r}{2r}$ such interval structures, and each acts through an affine threshold of at most M hit features. The halfspace growth bound (Shalev-Shwartz & Ben-David, 2014) yields, for $v \ge M + 1$,

$$2^v \le \left[\sum_{r=0}^{M} 4^r \binom{G+2r}{2r}\right] \left(\frac{ev}{M+1}\right)^{M+1},$$

so $v = O(M \log(2 + G/M))$. Since $G/M \le (d+1)/(2M) + \Delta$, this proves (14). The witness is a shattering argument, not an assumption that actual training masks are jointly completable. Standard statistical transfer gives (15) (Gupta & Roughgarden, 2017).

Comparison with direct counting. Direct counting gives $O(M \log(2 + m/M))$ for sampled chance constraints (Luedtke & Ahmed, 2008). Our G-bound replaces m/M by $d/M + \Delta$: bounded ratios remove input-count dependence; at order m/M, direct counting is as good. As $d = n - b$, the refinement needs near-full rather than fixed-fraction Build. Theorem 4 matches the stated regimes.

Theorem 4 (Matching worst-instance envelope)**.** *Fix $M \ge 2$, $\bar{d} \ge M$, $\bar{\Delta} \ge 0$, $s \ge 1$. Over known instances with $M \le d \le \bar{d}$, $0 \le \Delta \le \bar{\Delta}$, $d + \Delta < n$, requiring a feasible output, worst-case minimax expected regret is*

$$\Theta\left(\min\left\{1, \sqrt{\frac{M \log_2(2 + \bar{d}/M + \bar{\Delta})}{s}}\right\}\right) \tag{16}$$

if $\bar{\Delta} = 0$, or if $M \ge 4$ and $\bar{d} \ge \lceil 3M/2 \rceil$. The upper bound holds throughout.

Why execution slack creates a learning cost. The zero-slack construction uses Lemma A.1 to force choices among g groups of K candidates; paired singletons cover bounded deficit ratios. For positive slack, take $g \ge 2$, $K \ge 2$, $\zeta \in \{0, 1\}$ and $r = g + \zeta$. Each variable region contains K alternating candidate/separator pairs; the optional last region has $2K$ stable inputs. Always-changing inputs separate regions. Set

$$m = r(2K + 1), \quad M = 2g + \zeta, \quad b = r(4K - 1), \quad d = 3r - 1, \quad \Delta = 4K - 2.$$

One uniformly chosen variable region activates. Its internal separators change, and candidates change independently with probability $p = 1/16$, except a preferred candidate with probability $p - \eta$, $0 < \eta \le p/2$; other regions stay stable.

Enlarging usable records to candidate leaves and region roots dominates every original plan pointwise within the budgets. Two missing roots or a missing stable root force failure. With one missing variable root, success requires its activation and has probability at most $1/g$. All roots and one fixed candidate per region achieve at least $15/16$. With all roots present, g candidate slots remain. Moving a second candidate to an empty region loses at most p/g but gains at least $(1 - p)/g$, independently of the preferences. This last exchange is distribution-wise, not pointwise. Thus every legal output reduces to one candidate per region. Missing h preferences costs exactly $\eta h/g$ regret, while one history has divergence at most $128\eta^2 \operatorname{Ham}(\theta, \theta')/g$ between preference vectors. Testing gives $\Omega(\min\{1, \sqrt{g \log K/s}\})$. Choosing $g = \lfloor M/2 \rfloor$ and $K = \Theta(\bar{\Delta})$ supplies the slack term; Appendix E gives the integer choices and combines both lower bounds. This worst-instance envelope does not imply monotone hardness in one system's cap. Narrow-deficit and fully independent positive-slack tightness remain open.

Computation is a separate resource. For singleton-or-no-change laws, the run bound gives sample complexity $O((M + 1 + \log(1/\delta))/\epsilon^2)$ and matching worst-instance regret $\Theta(\min\{1, \sqrt{M/s}\})$ for $M \ge 2$. Exact compilers use, respectively,

$$O\big(m^3 (M+1)^2 (L+2)^3\big) \qquad \text{(singleton laws)}, \tag{17}$$

$$O\big(m^3 s (M+1)^2 (L+2)^2 (\Delta+2)^{2s}\big) \qquad \text{(general masks)}. \tag{18}$$

The first has polynomial bit complexity; the second stores actual plans indexed by clipped residual-cost vectors. At fixed M, scoring all feasible laminar interval families gives $O(M(M+s)m^{2M})$ operations for $m \geq 2$, using (3)–(4) and ordered realizability. Appendix F proves reconstruction and NP-hardness when M and the scenario count are inputs, already with unit costs, $L = d$, at most two changes per scenario and optimum at least $7/8$; a deterministic FPTAS there implies $\mathrm{P} = \mathrm{NP}$. History reading, planning and Build are not priced by the sample bound.

5 Fixed reuse fees preserve the learning upper bound

The same capacity argument controls learning when reuse itself is charged. Fix public nonnegative real fees κ_I, ρ_I before sampling, independent of values, futures and the update law. Every reached retained interval pays κ_I; a valid hit also pays ρ_I. Build still counts primitive preparation jobs. Paid work C^f_π is at least ideal work C^0_π, but need not be monotone: a change can avoid an expensive restoration.

Proposition 5 (Learning with fixed reuse fees)**.** *For $M \geq 1$ and integer $0 \leq d \leq L < n$, the entire paid completion class satisfies the bound* (14)*, independently of fee magnitude or precision. Theorem 3's sample and regret bounds hold against the paid optimum. A future-independent deletion of records produces a subbank on the same tree with no greater Build, storage or paid cost on any mask, and positive interval-exposure coefficients.*

Let $Z_v = \mathbf{1}\{J \cap I_v \neq \varnothing\}$, $\mathcal{R}$ be the retained roots and $\mathrm{ch}(v)$ the retained-forest children of v. A nonroot u with parent v incurs $\kappa_u Z_v + \rho_u(Z_v - Z_u)$, whereas a root incurs $\kappa_u + \rho_u(1 - Z_u)$. Adding fees to (3) gives

$$C^f_\pi = a + \sum_{v \in \mathcal{R}} (\kappa_v + \rho_v) + \sum_{v \in A} \left[w_v - \rho_v + \sum_{u \in \mathrm{ch}(v)} (\kappa_u + \rho_u) \right] Z_v.$$

An all-positive witness satisfies $C^0_{\pi_+}(\{j\}) \leq C^0_{\pi_+}(J_i) \leq C^f_{\pi_+}(J_i) \leq L$, preserving G and affine-threshold counting even with signed coefficients. Delete a record when checking plus restoration costs at least its unchanged fallback (Appendix G). Ideal lower bounds and planning guarantees need not extend to nonzero fees; shared contention and future-dependent charges are excluded.

6 Evidence for information choice and preparation learning

Constructed laws test information separation and learning; matched-objective comparisons on MoE requests and revision traces test the value of richer statistics. Execution studies separately assess runtime gains.

An exact full-support separation. At $(m, M, b, L) = (8, 3, 10, 9)$, P_+, P_- (the P, Q of Figure 1) match all 36 input first/second moments and all 30 feasible interval validities, hence every legal mean. All 429 ordered trees yield the same 2,236 banks as laminar enumeration, evaluated on all 256 masks. Selected by moment-constrained search (Bertsimas & Popescu, 2005, Sec. 2), the laws demonstrate existence, not prevalence. Their midpoint is independent with change rate $1/4$. Every decision receiving the common population statistics has worst-law regret at least $7419/81920$, or 9.06 points (Figure 2, right). A mixture of the two banks in Figure 1 attains it. Appendix H.2 specifies the laws, attaining weights and fixed-fee cases where the gap vanishes despite positive completion.

Population information and finite-history learning. For the fixed pair P_+, P_-, form $\widehat{p}(x) = (n_x + 1/256)/(s+1)$ from s masks. Full-history training maximizes completion; interval and pairwise baselines fit maximum-entropy laws to their moments, then maximize completion. Their regrets are 3.37/8.88/7.87 points at 32 histories and 0.02/9.23/8.94 at 512 (64 independent trials per world, exact evaluation). Empirical-moment fluctuations can reveal information absent from population moments; the population bound therefore does not directly lower-bound these learners. At $(g, e) = (64, 8)$, grouping and covariance learners find optimal banks in all 256 eight-history noiseless trials despite about 76% exact group recovery: optimal decisions can precede exact identification in this family. Noise disrupts grouping sooner than covariance selection but can also lower informed completion (Appendix H.1).

Selecting information for feasible decisions. For $P_{\lambda,\pm} = (1-\lambda)P_0 + \lambda P_\pm$, where P_0 is the independent midpoint, 32 pilot masks select six joint-avoidance events separating the eight best empirical completion functions. Frequencies and interval validities are estimated on 1,024 fresh masks; controls receive all 1,056. Table 1 evaluates 32 trials per law against the exact population.

Table 1: Six selected joint events on the matched family: completion differences in points. All paired 95% trial-bootstrap intervals exclude zero. Selection, estimation and controls: Appendix H.9.

Comparison	$\lambda = .25$	.5	1
Gain over pairwise moments	1.03	3.28	7.70
Gain over six random events	0.66	1.44	3.34
Shortfall from full history	0.79	0.84	1.31

Decision-selected events outperform random counters and pairwise moments on this family, while retaining a shortfall from full history. Gains are small or zero on independent and burst laws. The benefit is in information choice; counter counts exclude identifiers and processing and therefore do not establish storage savings.

When richer statistics help in applications. Under the path-request, ancestor-closed prefix interface, node visits determine endpoint probabilities and hence every work distribution. For set-valued requests, frequencies can omit which items are needed together (Proposition I.1). Expert residency tests this information question, not the ordered-tree learning bounds.

Expert residency at matched objectives and budgets. Four MoE models are teacher-forced on WildChat response tokens (Zhao et al., 2024). The first 70% of conversations fit static resident sets; later conversations test per-token copy-budget completion. The same optimizer uses joint requests or synthetic requests matching per-expert inclusion rates independently across layers. A separate control shuffles each layer's intact routed sets across training tokens. Capacities span $1/8$–$3/4$ of layer–expert pairs; caps are training miss quantiles. The protocol is unchanged across models (Table 29).

Table 2: Static expert residency: completion gains (points; synthetic near-ceiling ties excluded). Brackets: 95% bootstrap intervals over seed-averaged size–budget configurations, conditional on the held-out corpus.

Model	Joint − marginal	Joint − shuffled
OLMoE-1B-7B	9.3 [7.0, 11.9]	6.8 [5.2, 8.6]
Granite-3.0-3B	11.3 [8.5, 14.4]	10.6 [7.7, 13.8]
Qwen1.5-MoE	11.0 [7.0, 15.3]	9.0 [5.9, 12.3]
DeepSeek-V2-Lite	13.7 [10.4, 17.2]	12.1 [9.3, 15.1]

Including all configurations without the near-ceiling exclusion, joint-minus-marginal gains are 9.2, 11.3, 11.0 and 12.3 percentage points in Table 2's model order. The shuffle contrast supports a contribution from cross-layer dependence. Exclusions require all five plans (global-frequency, per-layer-frequency, marginal, joint and layer-shuffled; Appendix I) to reach 99.9% completion on synthetic requests, irrespective of test gains. Joint plans are within 0.4 points of diagnostic cross-fitted, same-optimizer references, not certified optima. Static plans lose or nearly tie LRU on single conversations but lead by 19.9–43.3 points with 8/32 conversations in rotation. A pinned share chosen on earlier conversations matches this lead without observed losses to LRU. These single-corpus routing replays model neither batching nor queueing (Appendix I).

Objective, information and plan class on natural updates. On 24 Python projects, 32 ML systems repositories and 24 Wikipedia articles, eight fixed path or section regions define changes, not program dependencies. All methods train on the first 256 episodes and test chronologically at five

caps. Table 3 separates three effects, pooling one- and four-update episodes. The full baseline matrix, including ordinary laws and tail objectives, is in Table 6.

Table 3: Full-history completion gains (points; 95% bootstrap intervals over repository/article histories), at $(m, M, b) = (8, 3, 10)$ and caps $5, 7, 9, 11, 13$. Each contrast changes only the named objective, information or plan-class restriction.

Comparator	Python	ML systems	Wikipedia
Mean-work objective	$8.4\ [6.8, 10.1]$	$8.0\ [6.7, 9.3]$	$7.4\ [5.7, 9.1]$
Pairwise information	$0.2\ [-0.1, 0.6]$	$-0.4\ [-0.9, -0.03]$	$-0.1\ [-0.2, 0.1]$
Balanced tree only	$7.7\ [4.4, 11.3]$	$4.5\ [2.7, 6.5]$	$3.9\ [1.7, 6.3]$

Here objective and tree choice dominate: interval, pairwise and marginal models are within two points of full history on every natural row; pairwise is significantly better on ML traces. Unlike the MoE comparisons, these traces offer little added value from full joint histories. The balanced-tree loss concerns plan choice, not missing information. Expected overshoot and buffered exceedance trail completion by 2.9–4.1 and 5.7–6.8 points. Dependent traces do not test the i.i.d. guarantees; mean-class exposure measures possible loss, not prevalence (Appendices H.11 and H.12).

Transfer across reuse workloads. Mooncake prefix staging improves completion through objective choice, without omitted joint information; its timing excludes pool transfer. BEIR caches benefit under imposed recurrence, but production RAG shows no held-out gain over frequency caching after optimizer repair: a generalization boundary distinct from runtime overhead (Appendix I).

From completion decisions to runtime. Across 16–256 regions, objective gains persist with scarce records and known unequal costs, but misspecified caps or unit-cost training can reverse them. Averaging completion over a cap band retains 68–88% of nominal gain and mitigates the tested cap errors (Appendix H.11).

The associative-model timing study did not establish net GPU savings under the primary protocol, with validation, planning and data-movement costs (Appendices H.4–H.5, H.8 and H.10). The MoE results measure routing-replay completion under a copy budget, not end-to-end serving speedup. Cost-model error can reverse measured gains; later successful, adaptively chosen safety margins remain exploratory (Appendix I).

7 Discussion and limitations

Under ideal reuse, even exact means for every feasible preparation can leave completion decisions unresolved: budget-feasible joint validity determines work distributions. Physical reuse capacity bounds the histories needed to learn a near-optimal bank without reconstructing the update law.

Recovering work distributions suffices for every cap, but a particular cap may require less information. In the separation family, increasing only the cap takes a mean-only decision from exponentially small completion to nearly informed-optimal performance, without changing the feasible banks. This transition distinguishes missing statistical information from its decision value: the same information gap can matter greatly at one deadline and little at another.

The analysis assumes ordered associative computation, exact version-based reuse and a bank fixed before updates. Separations describe worst cases, not typical gains. Learning needs complete, independent histories from the future's law; partial observations, drift and adaptive repair are outside the guarantees. Fixed nonnegative checking and restoration costs preserve the learning upper bound; efficient planning remains a separate problem.

MoE routing replays show joint-information gains at matched objectives; on natural revision traces, simpler statistics approach full-history performance. The evidence supports workload-dependent information choice, not universal superiority of full history. Net speedup also depends on planning, preparation, data movement and cost-model accuracy. Predictive information should be chosen for the execution decisions it improves.

AI USE STATEMENT

We used generative AI to develop the framework and hypotheses, formulate mathematical claims, provide critical proof ingredients and assist proof writing, design experiments and suggest parameters, implement methods, generate synthetic update laws and edits, clean and reformat revision data, and interpret results. AI also assisted literature search and synthesis, coding, figures, and manuscript drafting and editing. We did not use AI for translation; qualitative or thematic analysis was not applicable. We reviewed AI-assisted work, checking mathematical claims by derivation and exhaustive finite exact computation, not formal verification. The evaluated language models are research subjects, not assistants. The authors take responsibility for all content.

ETHICS STATEMENT

This work uses synthetic update laws and existing public datasets, including revision histories and conversational traces. Results are reported as aggregate performance measurements. No human participants were recruited, and no new human annotations were commissioned.

REPRODUCIBILITY STATEMENT

The main text and Appendices A–G state the assumptions and provide proofs of the theoretical results. Appendices H and I describe the data sources, model configurations, evaluation protocols, resource budgets and uncertainty estimates. We will publicly release the code upon acceptance of the paper.

REFERENCES

Shipra Agrawal, Yichuan Ding, Amin Saberi, and Yinyu Ye. Price of correlations in stochastic optimization. *Operations Research*, 60(1):150–162, 2012.

Ganesh Ananthanarayanan, Ali Ghodsi, Andrew Wang, Dhruba Borthakur, Srikanth Kandula, Scott Shenker, and Ion Stoica. PACMan: Coordinated memory caching for parallel jobs. In *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, pp. 267–280, 2012.

Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015.

Dimitris Bertsimas and Ioana Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.

Pramod Bhatotia, Umut A. Acar, Flavio P. Junqueira, and Rodrigo Rodrigues. Slider: Incremental sliding window analytics. In *Proceedings of the 15th International Middleware Conference*, pp. 61–72, 2014.

Chris Cameron, Jason Hartford, Taylor Lundy, and Kevin Leyton-Brown. The perils of learning before optimizing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3708–3715, 2022.

DeepSeek-AI. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. MiniRocket: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 248–257, 2021.

Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5484–5494, 2017.

Artyom Eliseev and Denis Mazur. Fast inference of mixture-of-experts language models with offloading. *arXiv preprint arXiv:2312.17238*, 2023.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, 2022.

IBM Granite Team. Granite 3.0 language models. Technical report, IBM, 2024. URL https://github.com/ibm-granite/granite-3.0-language-models/.

Rishi Gupta and Tim Roughgarden. A PAC approach to application-specific algorithm selection. *SIAM Journal on Computing*, 46(3):992–1017, 2017.

Eric Horvitz. Principles and applications of continual computation. *Artificial Intelligence*, 126(1–2): 159–196, 2001.

Junhao Hu, Wenrui Huang, Weidong Wang, Haoyi Wang, Tiancheng Hu, Zhang Qin, Hao Feng, Xusheng Chen, Yizhou Shan, and Tao Xie. EPIC: Efficient position-independent caching for serving large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24391–24402, 2025.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.

Misha Khodak, Maria-Florina Balcan, Ameet Talwalkar, and Sergei Vassilvitskii. Learning predictions for algorithms with predictions. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3542–3555, 2022.

Xinran Li and Peng Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769, 2017.

Kevin Lin, Charlie Snell, Yu Wang, Charles Packer, Sarah Wooders, Ion Stoica, and Joseph E. Gonzalez. Sleep-time compute: Beyond inference scaling at test-time. *arXiv preprint arXiv:2504.13171*, 2025.

James Luedtke and Shabbir Ahmed. A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, 19(2):674–699, 2008.

Thodoris Lykouris and Sergei Vassilvitskii. Competitive caching with machine learned advice. *Journal of the ACM*, 68(4):24:1–24:25, 2021.

Alexander Mafusalov and Stan Uryasev. Buffered probability of exceedance: Mathematical properties and optimization. *SIAM Journal on Optimization*, 28(2):1077–1103, 2018.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open mixture-of-experts language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2007.

Ruoyu Qin, Zheming Li, Weiran He, Jialei Cui, Feng Ren, Mingxing Zhang, Yongwei Wu, Weimin Zheng, and Xinran Xu. Mooncake: Trading more storage for less computation — a KVCache-centric architecture for serving LLM chatbot. In *23rd USENIX Conference on File and Storage Technologies (FAST 25)*, pp. 155–170, 2025.

Qwen Team. Qwen1.5-MoE: Matching 7B model performance with 1/3 activated parameters. Qwen blog, 2024. URL https://qwenlm.github.io/blog/qwen-moe/.

Johannes Rauh, Pradeep Kr. Banerjee, Eckehard Olbrich, Jürgen Jost, Nils Bertschinger, and David Wolpert. Coarse-graining and the Blackwell order. *Entropy*, 19(10):527, 2017.

Sivan Sabato and Naftali Tishby. Multi-instance learning with any hypothesis class. *Journal of Machine Learning Research*, 13(97):2999–3039, 2012.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Mikhail Shirokikh and Sergey Nikolenko. Sparse prefix caching for hybrid and recurrent LLM serving. *arXiv preprint arXiv:2605.05219*, 2026.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Kanat Tangwongsan, Martin Hirzel, Scott Schneider, and Kun-Lung Wu. General incremental sliding-window aggregation. *Proceedings of the VLDB Endowment*, 8(7):702–713, 2015.

Kanat Tangwongsan, Martin Hirzel, and Scott Schneider. Low-latency sliding-window aggregation in worst-case constant time. In *Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems*, pp. 66–77, 2017.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Zhengchao Wang, Yitao Hu, Jianing Ye, Zhuxuan Chang, Jiazheng Yu, Youpeng Deng, and Keqiu Li. RAGPulse: An open-source RAG workload trace to optimize RAG serving systems. *arXiv preprint arXiv:2511.12979*, 2025.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, 2024.

Jiayi Yao, Hanchen Li, Yuhan Liu, Siddhant Ray, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, and Junchen Jiang. CacheBlend: Fast large language model serving for RAG with cached knowledge fusion. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 94–109, 2025.

Wenxin Zhang, Yueying Li, Ciamac C. Moallemi, and Tianyi Peng. Tail-optimized caching for LLM inference. In *Advances in Neural Information Processing Systems*, volume 38, pp. 164419–164447, 2025.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. SGLang: Efficient execution of structured language model programs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 62557–62583, 2024.

A SELECTION REDUCTIONS AND FULL-SUPPORT SEPARATION

The reductions preserve arbitrary ordered trees, nested banks, unit costs and resource allowances, supporting the separation here and the learning and planning proofs in Appendices E–F. Capacity and fixed fees are treated in Appendices C and G; complementary mean-work and resource results in Appendices B and D; experimental evidence in Appendices H–I.

Lemma A.1 (Deadline-independent normalization and zero-slack forcing)**.** *Interleave $N \geq g \geq 1$ candidate inputs c_i with always-changing inputs z_i, then append $2N$ stable inputs. Set*

$$m = 4N, \quad M = g + 1, \quad d = 4N - g, \quad b = 4N + g - 1. \tag{19}$$

Every feasible ordered tree and nested bank has a replacement, independent of the update law and deadline, retaining the complete stable-suffix root and exactly g candidate leaves. It uses M records and b Build jobs, and has no greater work on every mask in which all z_i change and the suffix is stable. Actual usage may increase within the original allowances. Every such replacement is feasible and has work $d + |R \cap J_c|$, where R is its candidate set and J_c the changed candidates. At $L = d$, any original bank completing even one supported scenario must itself have this form, and succeeds exactly when no selected candidate changes.

Proof. A saved interval meeting a z_i is always invalid. Deleting its record leaves ideal work unchanged and cannot increase Build or storage. Every remaining core record is a candidate leaf; all others lie in the suffix, which the final z_N separates from the core.

If a suffix record remains, at most g candidate records remain. Replace all suffix records by the complete suffix root and extend the candidate set to size g by a fixed public rule. For each mask, the old valid suffix roots skipped at most $4N - 1$ operations; the new root skips exactly that many. Retained candidates are preserved and added candidates save one operation when valid. The disjoint replacement intervals admit an ordered binary refinement and use $(4N - 1) + g = b$ Build jobs and $g + 1 = M$ records. The count of skipped operations proves domination even if the original tree did not have a full-suffix node; unsaved refinements do not change interval costs.

If no suffix record remains, at most $g + 1$ candidate leaf operations can be skipped, so original work is at least $8N - g - 2$. Any full-suffix/g-candidate bank has work at most $4N \leq 8N - g - 2$, using $N \geq g \geq 1$. It therefore dominates in this case too. The replacement's baseline is $n - b = d$, and each changed selected candidate adds one operation. Its disjoint maximal subtrees admit a child-first schedule of exactly b Build jobs.

For zero-slack forcing, a successful original plan has $a = d$ by (3). No always-invalid retained interval can occur, since its positive component would exceed $L = d$. If t candidates are saved and r maximal suffix roots leave u suffix inputs uncovered, then

$$4N - t + 2u + r - 1 = 4N - g, \qquad t + r \leq g + 1.$$

Thus $t = g + 2u + r - 1$ and $u + r \leq 1$. If $r = 0$, then $u = 2N$, impossible; hence $r = 1, u = 0, t = g$. These maximal roots exhaust the record budget, leaving no nested exception. This retains the forcing assertion used by the later zero-slack learning converse. □

Proof of Theorem 2. **Law and informed optimum.** Take $N = 2g$ in Lemma A.1, with the snapshot and initial information fixed independently of H. In each independent history and future, candidates in H share a fair change bit Z; the other g candidates have independent fair bits, independent of Z. Every nonempty interval either contains a separator, lies in the stable suffix, or is one candidate leaf. Its validity probability is respectively 0, 1, or $1/2$, independently of H. Theorem 1 makes the *entire* feasible mean-work response identical across H.

For a normalized g-set R, put $h = |R \setminus H|$. Its exact work is

$$C_R = d + S_h, \qquad S_h = B_h + (g - h)Z, \qquad B_h \sim \mathrm{Bin}(h, 1/2), \quad B_h \perp Z. \tag{20}$$

Complementing all fair bits gives $S_h \overset{d}{=} g - S_h$. For $0 \leq e < g/2$, the events $S_h \leq e$ and $S_h \geq g - e$ are disjoint and equally probable. Completion is therefore at most $1/2$, attained by $R = H$. Pointwise domination extends this optimum to all feasible trees and banks. The endpoint $e = g/2$ is excluded. At zero slack, (20) recovers the formula used in the finite evaluations:

$$F_{H,0}(R) = 2^{-|R \setminus H| - \mathbf{1}\{R \cap H \neq \varnothing\}}. \tag{21}$$

Exact randomized minimax. Define $\beta_h(k) = \mathbb{P}\{\mathrm{Bin}(h, 1/2) \le k\}$, with value zero for $k < 0$ and one for $k \ge h$. Then

$$F_{H,e}(R) = \tfrac{1}{2}\beta_h(e) + \tfrac{1}{2}\beta_h(e - g + h).$$

For fixed R and uniform H, the overlap deficit h has mass $\binom{g}{h}^2/\binom{2g}{g}$, giving the value in (11). Indeed, normalize any mean-only procedure's output conditional on its independent coins. Its selected R is independent of H. Each fixed R has the displayed uniform-H average, so the original procedure's average, and hence worst-H value, is at most this number. Conversely, a uniform random R attains this number for every fixed H. This proves equality without interchanging extrema. Every selection admits a same-budget ordered realization; relabeling candidates changes retained intervals, not operand order. One randomized policy is thus minimax for every integer cap, in expectation over its coins. The adversary fixes H before those coins are drawn.

Let $\mathbb{P}(Q = h) = \binom{g}{h}^2/\binom{2g}{g}$ and $Y \mid Q \sim \mathrm{Bin}(Q, 1/2)$. Binomial symmetry rewrites the exact value, for every integer e, as

$$V_g(e) = \tfrac{1}{2}\mathbb{P}(Y \le e) + \tfrac{1}{2}\mathbb{P}(Y \ge g - e). \tag{22}$$

Two-sided bounds and the deadline band. For $q \ge 1$, expand over subsets of a fixed g-set R. With $(x)_a = x(x-1)\cdots(x-a+1)$,

$$\mathbb{E}q^Q = \sum_{A \subseteq R} (q-1)^{|A|}\mathbb{P}(A \subseteq H^c) \le \sum_{A \subseteq R} \left(\frac{q-1}{2}\right)^{|A|} = \left(\frac{1+q}{2}\right)^g,$$

since $(g)_a/(2g)_a \le 2^{-a}$. For $0 < q < 1$, use $Q \overset{d}{=} g - Q$ and the preceding bound at q^{-1}. Setting $q = (1 + e^\theta)/2$ gives, for every real θ,

$$\mathbb{E}e^{\theta Y} \le \left(\frac{3 + e^\theta}{4}\right)^g. \tag{23}$$

This is the standard sampling-without-replacement convex-order/Chernoff mechanism (Bardenet & Maillard, 2015, Lemma 1.1), here derived for both signs. It is an mgf bound, not a claim that Y is binomial or stochastically dominated by a binomial variable.

For $0 < e/g < 1/4$, Markov's inequality with $\theta < 0$ and minimizer $e^\theta = 3(e/g)/(1 - e/g)$ yields

$$\mathbb{P}(Y \le e) \le \inf_{\theta < 0} e^{-\theta e}\left(\frac{3 + e^\theta}{4}\right)^g = e^{-gD(e/g\|1/4)}.$$

At $e = 0$, take $\theta \to -\infty$. Since $S_h \ge B_h$, averaging gives $V_g(e) \le \mathbb{P}(Y \le e)$ and proves (9), including $(3/4)^g$ at zero slack. For $g/4 < e < g/2$, the positive-θ bound gives $\mathbb{P}(Y > e) \le e^{-gD(e/g\|1/4)}$. Combining (22) with $V_g(e) \le 1/2$ proves (10). These KL exponents are bounds, not claimed exact large-deviation rates.

For fixed $0 < \chi < 1/4$, $D(x\|1/4)$ decreases on $[0, \chi]$, giving the bound $e^{-gD(\chi\|1/4)}$ throughout $E_\chi = \{0, \dots, \lfloor \chi g \rfloor\}$. Moreover, for any prescribed collection of cap-dependent mean-only procedures $\{\mathcal{A}_e : e \in E_\chi\}$, some single H satisfies

$$\max_{e \in E_\chi} \mathbb{E}F_{H,e}(\widehat{\pi}_{\mathcal{A}_e}) \le |E_\chi| e^{-gD(\chi\|1/4)}. \tag{24}$$

Average the sum of their nonnegative completion probabilities over uniform H and select an H no worse than that average. The $O(g)$ factor preserves exponential separation. This H can depend on the prescribed procedures; it is not one law defeating every possible algorithm. For fixed $1/4 < \chi < 1/2$, (10) instead makes $1/2 - V_g(\lfloor \chi g \rfloor)$ exponentially small for all sufficiently large g.

The critical window. The hypergeometric variable Q has mean $g/2$ and variance $g^2/[4(2g-1)]$. The finite-population CLT applies to sampling g entries from g zeros and g ones: the maximum squared centered entry divided by population variance and by the smaller sample/complement size is $O(1/g)$ (Li & Ding, 2017, Theorem 1 and Example 1). Hence

$$(Q - g/2)/\sqrt{g} \Rightarrow N(0, 1/8), \qquad Q/g \to 1/2 \quad \text{in probability.}$$

For fixed real u, conditioning on Q gives

$$\mathbb{E}e^{iu(Y-g/4)/\sqrt{g}} = \mathbb{E}\left[e^{iu(Q-g/2)/(2\sqrt{g})}\cos\left(\frac{u}{2\sqrt{g}}\right)^Q\right]$$
$$\longrightarrow e^{-u^2/64}e^{-u^2/16} = e^{-5u^2/64}.$$

Here $\log\cos(u/(2\sqrt{g})) = -u^2/(8g) + O(g^{-2})$, so the cosine factor converges in probability and, by boundedness, in L^1 to $e^{-u^2/16}$. Replacing this factor by its limit is therefore valid without an independence assumption. In particular,

$$\frac{Y - g/4}{\sqrt{5g/32}} \Rightarrow N(0,1), \qquad \mathrm{Var}(Y) = \frac{g}{8} + \frac{g^2}{16(2g-1)}.$$

For $e_g = \left\lfloor g/4 + t\sqrt{5g/32}\right\rfloor$, the first probability in (22) tends to $\Phi(t)$. Eventually $g - e_g \geq g/2$, so (23) bounds the second by $e^{-gD(1/2\|1/4)}$. Thus, for every fixed $t \in \mathbb{R}$,

$$\lim_{g\to\infty} V_g\Big(\Big\lfloor g/4 + t\sqrt{5g/32}\Big\rfloor\Big) = \tfrac{1}{2}\Phi(t). \tag{25}$$

This includes $V_g(\lfloor g/4\rfloor) \to 1/4$, not a finite-$g$ step. No quantitative normal-approximation error is asserted.

History recovery and resource accounting. Across s complete histories, members of H have one common binary word. The other g words are independent uniform elements of $\{0,1\}^s$. Except on an event of probability at most

$$\left(g + \binom{g}{2}\right)2^{-s} = g(g+1)2^{-s}/2,$$

none matches the shared word or another outside word. Since $g \geq 2$, the unique repeated group of size g then identifies H. Return its bank, or a fixed legal g-candidate bank on ambiguous samples. The stated s makes one event and one bank optimal simultaneously for all integer $0 \leq e < g/2$. Reading the masks and grouping their words in a binary trie takes $O(gs)$ operations and $O(gs)$ transient words; tree and schedule output add $O(m+b)$. Build is exactly b and deployed storage M. Logarithmically many complete histories are not logarithmically many observed bits: their representation has $\Theta(gs)$ bits. □

Small failure witnesses. For integer $\Delta \geq 0$, any failing change set has a failing subset of size at most $\min(M, \Delta+1)$. If $a > L$, the empty set suffices. Otherwise, while work is at most L, choose an input of the failing set that hits a previously inactive retained component. Such an input exists until failure; each choice increases work by at least one. At most M choices activate all components, and at most $\Delta+1$ increase work beyond L, since $a \geq d$. Thus the downward co-change support up to this order determines whether a supplied family has a common feasible bank, not its completion probabilities.

Lemma A.2 (Uncovered candidates)**.** *Place p candidates at positions $2, 4, \ldots, 2p$ among $m = 2p+1$ inputs; all other inputs are spacers. If t candidates are uncovered by a bank, its baseline obeys $a \geq 3t$. At $d = L = 3g$, $1 \leq g \leq p$, and $M = g+1$, every g-subset of candidates is a feasible timely set. Under candidate-only changes, every positive-value bank is dominated by one of these choices.*

Proof. Let the uncovered inputs have total size u, c runs, and $e \in \{0,1,2\}$ sequence endpoints. Their complement has $c+1-e$ runs, each requiring a maximal saved root, so $a = 2u + r - 1 \geq 2u + c - e$. A run with $t' > 0$ candidates and e' endpoints has at least $2t' - 1 + e'$ inputs; its contribution is at least $4t' - 1 + e' \geq 3t'$. Candidate-free runs contribute nonnegatively. Summing gives $a \geq 3t$. At zero slack, a positive-value plan has $a = d$, and a changed candidate is timely exactly when uncovered. Thus it protects at most g candidates. Leaving any chosen g candidates uncovered and saving its $g+1$ complementary runs gives $a = 2g + (g+1) - 1 = 3g$ and exact Build $n - 3g$. Extend a smaller timely set to g candidates to dominate all candidate-only scenarios. □

A.1 FULL SUPPORT WITHOUT DETERMINISTIC INPUTS

Keep Theorem 2's laws P_H and budgets $(m, M, b, d) = (8g, g+1, 9g-1, 7g)$, $g \geq 2$. Let R_g independently change each input with probability $3/4$, and define

$$\bar{P}_H = \tfrac{1}{2}P_H + \tfrac{1}{2}R_g, \qquad r_g = (3/4)^{4g}. \tag{26}$$

Each complete history and future independently draws a fresh mixture coin, which is unobserved. Write $\bar{F}_{H,e}(\pi) = \bar{P}_H(C_\pi \leq d+e)$, $\overline{\mathrm{OPT}}_{H,e} = \max_\pi \bar{F}_{H,e}(\pi)$, and let $\bar{V}_g(e)$ be the mean-only randomized minimax value for this known family, with no other H-dependent information. The bank π_H saves the full suffix and the g candidates in H; π_{unif} instead selects a uniform g-subset. Both use the original budgets, before the future is known.

Proposition A.3 (Full-support separation and learning)**.** *Every mask has positive $\bar{P}_H$ probability. Candidate, separator and suffix change probabilities are $5/8$, $7/8$ and $3/8$, respectively, and every feasible plan's mean is independent of H. For every integer $0 \leq e < g/2$,*

$$\tfrac{1}{4} \leq \overline{\mathrm{OPT}}_{H,e} \leq \tfrac{1}{4} + \tfrac{1}{2}r_g, \qquad 0 \leq \bar{V}_g(e) - \tfrac{1}{2}V_g(e) \leq \tfrac{1}{2}r_g, \tag{27}$$

$$\overline{\mathrm{OPT}}_{H,e} - \bar{F}_{H,e}(\pi_{\mathrm{unif}}) \leq \tfrac{1}{2}\big(\tfrac{1}{2} - V_g(e)\big) + \tfrac{1}{2}r_g, \qquad \overline{\mathrm{OPT}}_{H,e} - \bar{F}_{H,e}(\pi_H) \leq \tfrac{1}{2}r_g. \tag{28}$$

The first policy's value averages its independent coins. For each fixed real t,

$$\bar{V}_g\Big(\Big\lfloor g/4 + t\sqrt{5g/32}\Big\rfloor\Big) \longrightarrow \tfrac{1}{4}\Phi(t). \tag{29}$$

For $\delta \in (0,1)$, an explicit learner using

$$s \geq \left\lceil 128 \ln\left(2\binom{2g}{2}/\delta\right)\right\rceil \tag{30}$$

complete i.i.d. histories returns, with probability at least $1-\delta$, one bank with regret at most $r_g/2$ simultaneously throughout this cap band. At zero slack it is exactly optimal on that event, and

$$\overline{\mathrm{OPT}}_{H,0} = \tfrac{1}{4} + \tfrac{1}{2}\,4^{-5g}, \qquad \bar{V}_g(0) = \tfrac{1}{2}V_g(0) + \tfrac{1}{2}\,4^{-5g}. \tag{31}$$

Proof. **Common means and a uniform completion bound.** For each $J \subseteq [8g]$, $R_g(J) = 3^{|J|}/4^{8g} > 0$. Mixing the original marginals with $3/4$ gives the stated change probabilities. Also $W_{\bar{P}_H}(\pi) = W_{P_H}(\pi)/2 + W_{R_g}(\pi)/2$ is independent of H.

For every legal tree and bank,

$$C_\pi(J) \geq 2|J| - 1. \tag{32}$$

Indeed, no valid saved subtree contains a changed input. The $|J|$ changed leaf operations execute, as do the at least $|J|-1$ branching nodes connecting them; intervening unary paths only add work. The empty case is immediate. In the stated band, $d+e \leq 8g-1$, so success requires $|J| \leq 4g$. For $K \sim \mathrm{Bin}(8g, 3/4)$,

$$\sup_\pi R_g(C_\pi \leq d+e) \leq \mathbb{P}(K \leq 4g) \leq 3^{4g}\mathbb{E}3^{-K} = 3^{4g}(1/2)^{8g} = r_g. \tag{33}$$

This bound covers every tree and bank on the added component; the structured-mask normalization is not asserted on its full support.

Values and the transition. On P_H, all plans have completion at most $1/2$, attained by π_H. Combining this with (33) proves the optimum bounds and the regret bound for π_H. The uniform policy has structured completion $V_g(e)$ for every H, giving $\bar{V}_g(e) \geq V_g(e)/2$ and its claimed regret bound. For any common randomized preparation μ, the normalization and averaging argument proving (11) gives $\mathbb{E}_H F_{H,e}(\mu) \leq V_g(e)$ for uniform H. Thus

$$\inf_H \bar{F}_{H,e}(\mu) \leq \mathbb{E}_H \bar{F}_{H,e}(\mu) \leq \tfrac{1}{2}V_g(e) + \tfrac{1}{2}r_g,$$

proving (27). In particular,

$$\bar{V}_g(e) \leq \tfrac{1}{2}e^{-gD(e/g\|1/4)} + \tfrac{1}{2}r_g \qquad (e < g/4),$$

$$\overline{\mathrm{OPT}}_{H,e} - \bar{F}_{H,e}(\pi_{\mathrm{unif}}) \leq \tfrac{1}{4}e^{-gD(e/g\|1/4)} + \tfrac{1}{2}r_g \qquad (g/4 < e < g/2).$$

These preserve exponential separation and near-optimality at fixed ratios on either side of $1/4$. Equation (25) and the vanishing sandwich error give (29). A generic total-variation bound for a half-weight mixture would leave a constant error; the uniform rare-completion bound is essential.

Learning without observing the mixture component. For distinct candidate indices i, j, their joint change probability is $17/32$ if both are in H, and $13/32$ otherwise. Join i, j when their empirical joint change frequency exceeds $15/32$. Each incorrect classification requires deviation at least $1/16$. For the centered Bernoulli pair indicator, the bounded-variable mgf bound (Bardenet & Maillard, 2015, Lemma 1.3) is $\log \mathbb{E}e^{\lambda(X-\mathbb{E}X)} \le \lambda^2/8$. Independence across histories and Chernoff's argument therefore give $2e^{-2s(1/16)^2} = 2e^{-s/128}$ per pair. Union bounding over $\binom{2g}{2}$ pairs proves simultaneous correctness under (30). The nonisolated vertices are then precisely H, since $g \ge 2$. Return their bank if there are exactly g such vertices, or a fixed legal bank otherwise. On the successful event the same output is π_H for all caps, proving the regret claim. A streaming implementation takes $O(g^2 s)$ arithmetic operations and $O(g^2)$ transient counters; reading costs $O(gs)$ and tree/schedule output adds $O(m + b)$. The complete histories contain $\Theta(gs)$ bits, not logarithmically many bits. These costs are separate from the deployed $g + 1$ records and $9g - 1$ Build jobs.

Exactness at zero slack. By Lemma A.1, a plan with positive completion on any structured mask must itself save the full suffix and g candidate leaves. Each such normalized plan has R_g completion $q_g = 4^{-5g}$: all $4g$ suffix inputs and all g selected candidates must remain unchanged. Its structured completion in every world is at least 2^{-g} by (21). A nonnormalized plan has zero structured completion and mixed completion at most $r_g/2$. Since $r_g = (81/256)^g < 2^{-g}$, every such plan is dominated in every world by any normalized plan. Both optimizations thus reduce to the original normalized game plus the common term $q_g/2$, proving (31). □

Scope of the perturbation. The mixture removes deterministic coordinates with a weight independent of g; it does not establish arbitrary-noise, paid-interface or natural-workload robustness. For example, independently flipping every original change bit with probability $\eta \in (0, 1/2)$ reveals the hidden group through an exact mean. Compare a group containing both c_1, c_2 with one containing neither. The interval (c_1, z_1, c_2) has an avoidance-probability difference

$$\eta\left(\frac{(1-\eta)^2 + \eta^2}{2} - \frac{1}{4}\right) = \eta(1/2 - \eta)^2 > 0.$$

Its single-record plan has Build five and absolute mean-work difference $5\eta(1/2-\eta)^2$. Distribution-level mixing and coordinate-wise bit flips therefore preserve different information interfaces.

B MEAN WORK AND THE PERFECT-TREE BARRIER

The following workload-specific mean-work result complements the separation of Section 3, rather than adding assumptions to it. Write $\mathrm{OPT}_P^W = \min_{\pi \text{ feasible}} W_P(\pi)$.

Theorem B.1 (Reassociation below full preparation)**.** *For $1 \le M \le \min\{d, \lfloor m/2 \rfloor\}$ and $d \le m$, an explicit history-free plan π_0 uses M records and at most $n - d$ Build. For every exchangeable update law (positions uniform given $K = |J|$),*

$$\Lambda_P/7 \le \mathrm{OPT}_P^W \le W_P(\pi_0) \le 6\Lambda_P, \qquad \Lambda_P = d + n\mathbb{E}_P \min\{1, K/(M+1)\}. \tag{34}$$

At $m = 2^H, M = d = H, H \ge 16$, the perfect-tree zero-history minimax mean ratio is $\Theta(H/\log_2(H+1))$. Allowing depth $H + 1$ instead suffices for a constant ratio against the unrestricted informed comparator; depth H forces the perfect tree.

The construction leaves $\lceil (d - M + 1)/2 \rceil$ prefix inputs unprepared and saves M nearly equal suffix blocks. Uniform changes expose component work, and Cauchy–Schwarz gives the lower bound, proved below. This is a mean-work guarantee under exchangeability, not a completion guarantee for arbitrary updates.

The all-tree lower bound. For a uniform k-set J_k, $k \ge 1$, a real component of weight w lies inside an interval of length $\ell \ge w/2$. Hence

$$\mathbb{P}(J_k \cap I \ne \varnothing) = 1 - \frac{\binom{m-\ell}{k}}{\binom{m}{k}} \ge 1 - e^{-k\ell/m} \ge \tfrac{1}{2}\min\{kw/(2m), 1\}.$$

Binomial coefficients with too small an upper argument are zero. The always-active virtual component satisfies the same bound. For $x = k/(2m)$ and at most $r_A \le M + 1$ nonempty components, $\min(z, 1) \ge z/(1 + z)$ and Cauchy–Schwarz give

$$\mathbb{E}C_\pi(J_k) \ge \frac{1}{2}\sum_v \frac{xw_v^2}{1 + xw_v} \ge \frac{xn^2}{2(r_A + xn)} \ge \frac{nk}{2k + 4(M + 1)} \ge \frac{n}{6}\min\{1, k/(M + 1)\}. \quad (35)$$

Here the sum includes the virtual weight. Combining $C \ge d$ with $\max(d, z/6) \ge (d + z)/7$ proves the lower side of (34); $k = 0$ uses only $a \ge d$. Nothing assumes independent component activations. Averaging over k preserves the bound for every exchangeable law and every informed or randomized choice of tree.

The universal construction. Set $u = \lceil (d - M + 1)/2 \rceil$, so $u \ge 1$ and $m - u \ge M$ in Theorem B.1's range. Save M disjoint suffix roots with lengths differing by at most one, and refine inside each block and over the ordered prefix/block units. Then

$$C_{\mathrm{build}} = 2(m - u) - M \le n - d, \qquad a = 2u + M - 1 \in \{d, d + 1\}.$$

For every k-set, regardless of its positions,

$$\begin{aligned} C_{\pi_0}(J) &\le \min\{n, a + (2\lceil (m - u)/M \rceil - 1)\min(k, M)\} \\ &\le 2d + 3m\min(1, k/M) \le 6[d + n\min\{1, k/(M + 1)\}]. \end{aligned} \quad (36)$$

We used $a \le 2d$, $2\lceil (m - u)/M \rceil - 1 \le 3m/M$, and $(M + 1)/M \le 2$. A postorder executes the exact Build union; constructing the tree and bank takes $O(m)$ operations. This proves the upper side and factor 42. Full coverage with at most M maximal roots costs at least $2m - M > n - d$.

The critical perfect tree. Fix $m = 2^H, M = d = H, H \ge 16$. The fixed-tree zero-history ratio is the infimum over distributions on feasible banks, independent of P, of the supremum over exchangeable P of mean work divided by its informed *fixed-tree* optimum. Any feasible bank leaves an uncovered input. Its H off-path siblings have sizes $1, 2, \ldots, m/2$. Every sibling larger than the total uncovered count u needs a record, so the largest sibling contains at most $\lfloor \log_2 u \rfloor + 2$ records, including nested records. Component Cauchy–Schwarz inside that half-tree yields

$$\mathbb{E}C_\pi(J_1) \ge \frac{m}{8(3 + \log_2 u)} \ge \frac{m}{8(3 + \log_2(a + 1))}. \quad (37)$$

For example, its $m - 1$ operations form at most $\lfloor \log_2 u \rfloor + 3$ components; a real component of weight w activates with whole-tree singleton probability at least $w/(2m)$, and the virtual component with probability $1/2$. This gives $(m - 1)^2/[2m(\lfloor \log_2 u \rfloor + 3)]$, which implies (37). The function $(3 + \log_2(a + 1))^{-1}$ is convex, so the same lower bound holds for randomized banks with a replaced by $\bar{a} = \mathbb{E}a$.

For an upper construction put $s_0 = 2^{\lceil \log_2 H \rceil}$, $h_0 = \log_2 s_0$, $t = \lfloor \log_2 h_0 \rfloor$, $R = 2^t$, and $q = H - h_0$. Leave the rightmost depth-q subtree unprepared. Retain the other $R - 1$ depth-t roots and the siblings along the rightmost path at depths $t + 1, \ldots, q$. This uses $R - 1 + H - h_0 - t \le H$ records, has $H \le a \le 5H$, and every retained block has at most $2m/h_0$ inputs. Consequently

$$C_U(J) \le \min\{n, 5H + 4m|J|/\log_2 H\}. \quad (38)$$

Comparing each count to (35) and $C \ge H$ gives ratio $O(H/\log_2(H + 1))$ for every exchangeable count mixture.

For the matching lower bound only two independent per-input update rates are needed: m^{-2} and m^{-1}. At the slower rate, (38) and the probability of any change at most $1/m$ give optimum at most $6H$. At the faster rate, form a largest-first dyadic frontier of H blocks and delete one largest block of size ℓ. Saving the other $H - 1$ roots is feasible, has baseline $H + 2\ell - 2$, and all block sizes at most $2m/H$; since $\mathbb{E}|J| = 1$ and $H^2 \le m$, its mean is at most $9m/H$. If a randomized bank has ratio ρ on both laws, then $\bar{a} \le 6\rho H$. The faster law has exactly one uniform change with probability at least $1/4$, so (37) gives

$$9\rho m/H \ge \frac{m}{32[3 + \log_2(6\rho H + 1)]}.$$

For $\rho < H$, the denominator's bracket is at most $4\log_2 H$, giving $\rho \ge H/(1152\log_2 H)$; otherwise the desired bound is immediate. This covers all randomized nested banks.

One additional level. A binary tree of depth at most H with 2^H leaves must be perfect, by equality in the recursive leaf bound. For depth $H+1$, put one unprepared input beside a recursively balanced tree on the other $m-1$ inputs. Split a largest frontier subtree until there are H saved blocks. If the largest final block has t leaves, each final block has at least $(t-1)/2$, so $t \le 2(m-1)/H + 1 \le 4m/H$. Build is $n-H-1$, baseline $H+1$, and

$$C(J) \le \min\{n, H+1+(8m/H)\min(|J|, H)\}.$$

Equation (35) proves a constant ratio even against the all-tree comparator. Expression depth is not measured latency or neural depth.

C Capacity and forecast geometry

Finite capacity constants. If $a > L$, S_π is empty. Otherwise the argument preceding (12) gives

$$|S_\pi| \le u + M(L-a) \le \lfloor (a+1)/2 \rfloor + M(L-a) \le \lfloor (d+1)/2 \rfloor + M(L-d),$$

since the last expression is nonincreasing in integer a for $M \ge 1$. This upper bound needs none of the extra assumptions for a matching construction. For $M \le d \le m$, $M \ge 2$, leave an uncovered prefix of size $u = d/2$ with one complementary saved root when d is even, or $u = (d-1)/2$ with two when d is odd. The remaining inputs admit those nonempty blocks, $a = d$, and $u \ge d/3$ positions are timely.

For $\Delta \ge 2$, set $u = \lceil (d-M+1)/2 \rceil$, $t = m-u$, $f = \lfloor \Delta/2 \rfloor$, and $S = \min\{(M-1)f, t-1\}$. Split S suffix inputs as evenly as possible into $M-1$ nonempty blocks, put the other suffix inputs in one last block, and save all M roots. Here $t \ge M$, $a \in \{d, d+1\}$, and every small-block singleton costs at most $a + 2f - 1 \le L$. Since $M - 1 \ge M/2$, $f \ge \Delta/3$, and $t - 1 \ge m/4$,

$$S \ge \min\{M\Delta/6, m/4\} \ge \tfrac{1}{6}\min\{m, M\Delta\}.$$

The better of this plan and the uncovered-prefix plan proves (13); for $\Delta = 0, 1$, the prefix alone suffices since $M \le d$. Construction is linear and Build has the postorder schedule. For $L \ge n$ the empty bank serves all positions; for $L < d$ none can succeed. On a perfect tree at cap $H+1$, a changed leaf and its H ancestors exhaust the cap. Every off-path sibling must be skipped at its root, using all H records. Any second changed location invalidates one such sibling with no retained fallback. Conversely those H roots cost $n - H - 1$ Build and serve the chosen location.

Equal posterior values can occupy different geometry. For a uniform latent hotspot θ, let a history X and future Y be conditionally i.i.d. from $\frac{1}{2}\,\mathrm{Unif}[m] + \frac{1}{2}\delta_\theta$. Set $m = KR$, $d = 2K+1$, $L = d + \Delta$, with $K \ge 8$, $2 \le M \le K/4$ and a power of two $R > K + 1 + M\Delta$. Number positions from zero. The signals $\lfloor X/K \rfloor$ and $X \bmod R$ have consecutive and striped cells C_s respectively, but match message length, mutual information, conditional entropy, best-point prediction accuracy, and posterior probability multisets:

$$\mathbb{P}(Y = y \mid S = s) = \frac{3}{4m} + \frac{\mathbf{1}\{y \in C_s\}}{4K}. \tag{39}$$

Proposition C.1 (Unresolved locations must share a preparation)**.** *The optimal prior-averaged values satisfy $V_{\mathrm{int}} = \Theta(1)$ and $V_{\mathrm{str}} = \Theta(M/K)$. Refining the striped signal by q bits about the same X gives*

$$V_{\mathrm{str}}(q) = \Theta(\min\{1, 2^q M/K\}), \qquad q_{1/8} = \log_2(K/M) + O(1). \tag{40}$$

The converse permits randomized encoders, arbitrary ordered trees and nested banks at the same budgets.

A consecutive cell can be left uncovered while saving its complement. A timely run cannot contain two striped candidates, whose separation exceeds G. Thus a bank covers at most $M+1$ such candidates. The matching construction is in the proof below. This compares supplied signals, not labeled posterior equality or Blackwell equivalence (Khodak et al., 2022; Rauh et al., 2017).

Proof of Proposition C.1. Let Ω be a finite support with $N = |\Omega|$; here $\Omega = [m]$. For a uniform hotspot prior and conditional iid $X, Y \sim (1-\alpha)\,\mathrm{Unif}(\Omega) + \alpha\delta_\theta$, direct summation gives

$$\mathbb{P}(X = x) = 1/N, \qquad \mathbb{P}(Y = y \mid X = x) = (1-\alpha^2)/N + \alpha^2 \mathbf{1}\{y = x\}. \tag{41}$$

Averaging at $\alpha = 1/2$ over either signal cell proves (39). Equal posterior multisets and uniform marginals prove the asserted information equalities. Leave a consecutive cell's K inputs uncovered and split its complement into exactly two nonempty saved intervals. Then Build $= 2(m-K)-2 = b$, baseline d, and $V_{\mathrm{int}} \geq 1/4 + 3K/(4m)$.

For a striped cell, each timely run has at most one candidate because its length is at most $G < R$. Hence a bank serves at most $M+1$ candidates. Conditional on any independent random tapes, $Q = 2^q$ messages select Q fixed feasible banks; (39) bounds their background contribution by $3G/(4m)$ and their selected-history contribution by $\frac{1}{4}\min\{1, Q(M+1)/K\}$. Averaging the tapes covers randomized encoders and builders.

For the lower bound put $r = 2\lfloor M/2 \rfloor$, $t = r - 1 \geq M/3$, and $u = K + 1 - r/2$. Any nonempty set of at most t candidates can be enlarged by adjacent positions to an uncovered set of size u, without increasing its number of runs. Its complement has at most r runs and at least r inputs; split it into exactly r saved intervals. The resulting bank has

$$|A| = r \leq M, \quad C_{\mathrm{build}} = 2(m-u) - r = b, \quad a = 2u + r - 1 = d.$$

Partition $\min(K, Qt)$ candidates into at most Q groups of size at most t, and send the group containing X, or a default for an unselected X. Thus

$$\frac{3u}{4m} + \frac{\min(K, Qt)}{4K} \leq V_{\mathrm{str}}(q) \leq \frac{3G}{4m} + \frac{1}{4}\min\{1, Q(M+1)/K\}. \tag{42}$$

Because $R > G$, $t \geq M/3$, and $M + 1 \leq 3M/2$, these bounds give (40). For completion $1/8$, $Q \geq K/(2t)$ suffices. If $K \geq 12$, necessity gives $Q(M+1) \geq K/2 - 3 \geq K/4$ and hence $Q \geq K/(6M)$; smaller K has bounded K/M. Every construction uses the original operand order and admits an $O(m)$ construction and a serial schedule.

D THE LOCAL REDUCTION AND EXACT RESOURCE CONSEQUENCES

Consider ordered groups (A_i, x_i, z_i), $i \in [g]$, with $\ell_i \geq 3$ stable inputs in A_i, a possibly changing x_i, and an always-changing separator z_i. A scenario $K \in \binom{[g]}{k}$, $1 \leq k < g$, changes its x_i and all separators. Put $R_i = A_i x_i$, $\nu_i = 2\ell_i + 1$, $b_0 = \sum_i \nu_i$, $d_0 = 3g - 1$, $L_* = d_0 + 2k$, and $L = L_* + e$. The actual deficit is $d = d_0 + \delta_b$, where $\delta_b = b_0 - b$; $\Delta = 2k + e - \delta_b$ is not e.

Theorem D.1 (Pointwise local reduction)**.** *Every legal tree and bank is dominated on every trigger mask by a bank with at most two records per region, without increasing Build or storage. Every original plan has work at least $d_0 + 2|K|$. The profiles in Table 4 suffice and have legal ordered realizations.*

Table 4: Local preparation profiles; c_0 and c_1 denote work without and with the trigger.

Retained results	Build u	Records	(c_0, c_1)
None	0	0	(ν, ν)
One stable interval	$1, 3, \ldots, \nu - 2$	1	$(\nu - u, \nu - u)$
Two stable intervals	$2, 4, \ldots, \nu - 3$	2	$(\nu - u, \nu - u)$
A and x (flat)	$\nu - 1$	2	$(1, 2)$
R	ν	1	$(0, \nu)$
R, A (nested)	ν	2	$(0, 2)$

Global work is $d_0 + \sum_i c_{i,0} + \sum_{i \in K}(c_{i,1} - c_{i,0})$. Deleting separator-crossing records localizes execution. At Build u, no-trigger work is $\nu - u$ and trigger work is at least $\max\{\nu - u, 2\}$. Stable intervals attain it until the last two Build levels, which require the flat and nested modes. The proofs below supply reconstruction and exact capacity $C_e(b, M)$, the largest number of timely k-masks. For $0 \leq b \leq b_0$ and $0 \leq e < \min_i(\nu_i - 2)$ it is computable in $O(g^5)$ integer operations; equal lengths allow every cap in $O(g^8)$. This counts the uniform k-layer, not arbitrary probability weights; tree and schedule output incur their input/output cost.

Corollary D.2 (Joint Build–storage frontier)**.** *For $g \le M \le 2g$, $0 \le \delta_b \le 2g$ and the preceding range of e, let $r = g - k$, $a_* = \max\{2g - M, \delta_b - g, 0\}$ and $f_* = (\delta_b - 2a_*)_+$. One bank serves every k-mask iff*

$$e \ge e_{\mathrm{all}} := 2\min\{a_*, r\} + \min\{f_*, (r - a_*)_+\}. \tag{43}$$

If $e_{\mathrm{all}} < \min_i(\nu_i - 2)$, it is the globally minimum extra work at these budgets.

For $g = 4, k = 2, \ell_i = 4$, cap 17 permits all six masks at both $(b, M) = (34, 7)$ and $(32, 8)$, but at most five at $(32, 7)$. At $(b, M, e) = (36, 8, 0)$, nesting serves all six whereas any flat bank serves at most one. Thus saved internal states can replace distinctions a predictor would otherwise need to resolve; Build and storage minima are not separately attainable.

Compressing *one* noisy observation differs from learning from many histories. Put $N = \binom{g}{k}$, $\rho = C_e/N$. For X, Y conditionally i.i.d. from $(1-\alpha)\,\mathrm{Unif}(\binom{[g]}{k}) + \alpha\delta_\theta$, with $0 \le \alpha \le 1$, an integer $q \ge 0$-bit message about X, with public coins independent of θ, has optimal worst-θ value bounded by

$$(1-\alpha^2)\rho + \alpha^2[1 - (1-\rho)^{2^q}] \le V(q) \le (1-\alpha^2)\rho + \alpha^2 \min\{1, 2^q\rho\}. \tag{44}$$

If $C_e > 0$ and $\alpha = 1/2$, attaining completion $1/8$ needs $\log_2(N/C_e) + O(1)$ bits; if $C_e = 0$, no message helps. The same-budget covering construction is in Appendix D.1 below; only the selected bank is built, but generating the 2^q descriptions can cost $O(2^q g)$. Separators are essential for this local resource reduction, unlike the full-support example of Section 6.

Proof of Theorem D.1. Delete intervals meeting always-changing separators, then use the canonical interval refinement to make each R_i a subtree. This preserves costs and can only reduce Build and records. The separators and connecting skeleton contribute $g + (2g-1) = d_0$. In one region suppress i, and let u be its original Build and r its record count. On no trigger all its saved intervals are valid, so $c_0 = \nu - u$ and monotonicity gives $c_1 \ge c_0$.

For odd $1 \le u \le \nu - 2$, one stable interval with $(u+1)/2 \le \ell$ inputs realizes $(\nu - u, \nu - u)$ with the same Build and one record. For even $2 \le u \le \nu - 3$, at least two original records are necessary; two disjoint stable intervals covering $(u+2)/2 \le \ell$ inputs realize the same pair using two records. Under a trigger, disjoint valid hits cover at most ℓ stable inputs and skip at most $2\ell - 1 = \nu - 2$ operations. Hence $c_1 \ge 2$. At $u = \nu - 1$, at least two records are necessary and the flat children attain $(1, 2)$. At $u = \nu$, one record must be the full root, giving $(0, \nu)$; with at least two, the nested pair attains $(0, 2)$ and dominates. At $u = 0$ use no records. All prescribed intervals have the stated binary refinement and a postorder Build schedule. Summing the local costs proves the global formula and lower bound. □

A common family without a saturated cap. Assume $0 \le b \le b_0$ and $0 \le e < \min_i(\nu_i - 2)$. A positive-success bank cannot use an empty local profile, and an outer-only region cannot trigger in a successful mask: either contributes at least $\nu_i - 2 > e$ above the scenario-wise minimum. Partition regions into S, F, H, O for stable-only, flat children, nested pair, and outer-only. Write $a = |S|$, $f = |F|$, $z = |H|$. If v is total stable-only work beyond saving all the complete A_i in S, direct substitution gives

$$C(K) = L_* + v + 2|S \setminus K| + |F \setminus K| + \sum_{i \in O \cap K} (\nu_i - 2). \tag{45}$$

One-stable-interval profiles have even extra work and two-interval profiles odd extra work. Since success implies $v \le e$, all v can be placed in any one stable region: when even, Build $\nu_i - 2 - v \ge 1$ is a legal odd amount; when odd, it is a legal even amount at least two. This uses only $\eta = v \bmod 2$ extra records and preserves every cost and total Build, also for unequal lengths.

Consequently a nonempty supplied $\mathcal{F} \subseteq \binom{[g]}{k}$ has a common preparation if and only if a partition and integers $v \ge 0$, $\eta \in \{0, 1\}$ satisfy

$$\begin{aligned} &v \equiv \eta \pmod 2, \quad v \ge \eta, \quad \eta \le a, \quad a = 0 \Rightarrow v = \eta = 0, \\ &g + z + f + \eta \le M, \qquad 2a + f + v \ge \delta_b, \\ &K \cap O = \varnothing, \quad 2|S \setminus K| + |F \setminus K| + v \le e \quad (K \in \mathcal{F}). \end{aligned} \tag{46}$$

Necessity follows by the reduction; sufficiency follows by its profiles, using exactly $b_0 - 2a - f - v$ Build and $g + z + f + \eta$ records. Thus the condition concerns one compatible bank, not separate preparations for individual scenarios.

Exact capacity. Enumerate nonnegative a, z, f with $a + z + f \le g$, and $\eta \in \{0, 1\}$ with $\eta \le a$ and $g + z + f + \eta \le M$. If $a = 0$, require $\eta = 0$, $f \ge \delta_b$, and set $v_\eta = 0$; otherwise put

$$v_\eta = \eta + 2\max\{0, \lceil(\delta_b - 2a - f - \eta)/2\rceil\}.$$

Discard $v_\eta > e$. Since the smallest feasible v enlarges every success set, (46) gives exactly

$$C_e(b, M) = \max_{a,z,f,\eta} \sum_{j=0}^{a} \sum_{t=0}^{f} \binom{a}{j}\binom{f}{t}\binom{z}{k-j-t} \mathbf{1}\{2(a-j) + f - t + v_\eta \le e\}. \tag{47}$$

Out-of-range binomial coefficients and an empty maximum are zero. There are $O(g^3)$ states and $O(g^2)$ summands per state. Coefficients have $O(g)$ bits; budget arithmetic uses its input bit length. Realizing the maximizing state takes $O(m)$ tree construction and the exact Build schedule. Equation (47) does not optimize arbitrary probability weights.

Proof of Corollary D.2. For all-mask success, $O = \varnothing$. Restore complete stable prefixes, removing v work everywhere but losing v units of Build deficit. Recover any necessary deficit by transitions $H \to F \to S$. Each adds one unit of deficit, never increases records, and increases work on each mask by at most one. At most v transitions suffice, since the available deficit reaches $2g \ge \delta_b$. Thus $v = \eta = 0$ is no worse at the allowed budgets.

The remaining constraints are $a \ge 2g - M$, $2a + f \ge \delta_b$, and $a + f \le g$. For $r = g - k$, worst extra work is the sum of the r largest entries of a list of a twos, f ones, and zeros:

$$T_r(a, f) = 2\min(a, r) + \min\{f, (r - a)_+\}.$$

For fixed a minimize f at $(\delta_b - 2a)_+$. Feasibility is exactly $a \ge a_*$. Increasing a cannot decrease T_r: two ones are replaced by a two and a zero until the deficit is met, a possible parity step replaces a one by a two, and subsequently only twos are added. Hence a_*, f_* attain the minimum. If it lies in the non-saturated range, every smaller extra-work value also lies there and has been ruled out, proving global minimality. For the stated numerical example, one S and three H profiles give $(34, 7)$, four F give $(32, 8)$, and (47) gives five successes at $(32, 7), e = 2$. □

At $e = 0$, (46) forces $v = 0$. Every F region triggers in every successful mask and can be replaced by full S, saving both resources. Thus successful masks satisfy $S \subseteq K \subseteq S \cup H$. For a prescribed nonempty family, take $S = I = \bigcap \mathcal{F}$ and $H = U \setminus I$, where $U = \bigcup \mathcal{F}$. The simultaneous minima are

$$M \ge g + |U \setminus I|, \qquad b \ge b_0 - 2|I|. \tag{48}$$

Saving R_i for $i \notin I$ and A_i for $i \in U$ attains them. Equivalently,

$$C_0(b, M) = \max\left\{\binom{v}{k-f} : \ 0 \le f \le k,\ k - f \le v \le g - f,\ g + v \le M,\ b_0 - 2f \le b\right\}. \tag{49}$$

The realizing bank has cost $L_* + 2|I \setminus K| + \sum_{i \in K \setminus U}(\nu_i - 2)$, so its success set is precisely the stated interval of the subset lattice.

All caps at equal lengths. If every $\nu_i = \nu$, count outer, nested, flat-children, one-stable, two-stable, and empty profiles by (o, z, f, s_1, s_2, t), summing to g. Require $o + 2z + 2f + s_1 + 2s_2 \le M$ and set $b_f = \nu(o + z) + (\nu - 1)f$. Stable-only Build fills the step-two progression from $s_1 + 2s_2$ to $s_1(\nu - 2) + s_2(\nu - 3)$; choose its largest member $u \le b - b_f$, discarding infeasible states. For $B' = b_f + u$, exact capacity at any $0 \le L \le n$ is

$$\max_{o,z,f,s_1,s_2,t} \sum_{p=0}^{o} \sum_{q=0}^{z} \sum_{r=0}^{f} \binom{o}{p}\binom{z}{q}\binom{f}{r}\binom{g-o-z-f}{k-p-q-r} \mathbf{1}\{n - B' + \nu p + 2q + r \le L\}. \tag{50}$$

The largest stable Build lowers only the common baseline. Triggers in the three other types add $\nu, 2, 1$, respectively, proving the formula from the pointwise reduction. The progression is attainable by distributing increments of two among stable profiles. There are $O(g^5)$ states and $O(g^8)$ arithmetic operations by direct summation. At unequal lengths the pointwise reduction still holds, but this symmetric count elimination generally does not.

Flatness, mean work, and the separator boundary. At $e = 0$, a region varying between active and inactive among successful masks forces both A_i and R_i: equality in the triggered skipped-work bound needs the complete A_i root, and inactive zero local work needs R_i. Thus a flat bank can serve at most one k-mask; saving A_i on that mask and R_i elsewhere realizes it. At $b = b_0, M = 2g$, the nested bank saves both roots everywhere and serves all masks. With one noisy history the flat value is exactly

$$(1-\alpha^2)/\binom{g}{k} + \alpha^2 \min\{1, 2^q/\binom{g}{k}\},$$

by the covering argument and a uniform set of distinct singleton families. At $\alpha = 1/2$ it needs $\log_2 \binom{g}{k} + O(1)$ bits for value $1/8$, versus zero bits for the nested bank.

This separation survives extra slack. At $b = b_0, M = 2g, L = d_0 + \Delta$, and $2k \le \Delta < \min_i \nu_i$, the exact flat capacity is $\binom{\min(g, \Delta-k)}{k}$. To see this, let D contain regions without a complete retained R_i. A triggered retained R_i has no nested fallback and loses $\nu_i > \Delta$ work, so a successful K lies in D. Inactive regions in D cost at least one and active ones at least two. Therefore $C(K) \ge d_0 + |D| + k$, forcing $|D| \le \Delta - k$. Retain flat children on $\min(g, \Delta - k)$ regions and outer roots elsewhere to attain the bound, using $g + |D| \le 2g$ records and $b_0 - |D|$ Build. At $\Delta = 3k, g \ge 4k$, the resulting fraction is at most $(2k/g)^k$, while nested completion is one.

For mean work, saving just every A_i costs g records and $b_0 - 2g \le b_0 - 2k$ Build, and gives constant work $5g - 1$. Every plan needs at least $3g - 1 + 2k$, yielding ratio less than $5/3$. Finally, without separators take A_1, x_1, A_2, x_2 with $|A_i| = 3$. A saved seven-input prefix uses one record and 13 Build jobs; a change of x_2 costs two operations. Every region-local plan needs a connecting operation and at least two operations in the triggered region. Cross-region reuse is thus a genuine omitted freedom.

D.1 One noisy history and physically attainable covers

Let $\mathcal{S}$ be feasible success sets on a finite support Ω, $N = |\Omega|$, with maximum size C. Suppose the images of a maximum set under a transitive group on Ω all have same-budget realizations. The posterior (41) immediately gives (44): Q messages select at most Q sets, with background at most $(1-\alpha^2)C/N$ and union size at most $\min(N, QC)$. A uniform analysis prior upper-bounds worst-hotspot value; conditioning and then averaging independent coins permits randomized procedures.

For achievability draw Q independent uniform group images of a maximum set, and send the first containing X, or a default. Every fixed X is covered with probability $1 - (1 - C/N)^Q$, while every selected set has size C. The rule is equivariant, so its prior-averaged value is also its value for each fixed hotspot. For $\rho = C/N$, $1 - (1-\rho)^Q \ge (1 - e^{-1}) \min(1, Q\rho)$. At $\alpha = 1/2$, $Q \ge (\ln 2)/\rho$ suffices for value $1/8$; when $\rho \le 1/12$ the converse requires $Q\rho \ge 1/4$, and otherwise $\log(1/\rho)$ is bounded. This proves the bit threshold. The zero-bit value is ρ, and observing all of X yields $\alpha^2 + (1-\alpha^2)\rho$. $C = 0$ is immediate. With full X, a uniform maximum set containing it exists by transitivity and attains the endpoint. In particular, for $\alpha < 1$, probability-one completion needs $C = N$, however accurately that one history is described.

The capacity states of (47) depend on partition counts and a stable-work deficit that fits any stable region in the non-saturated range. Permuting the partition therefore preserves total Build, records, and the success set's cardinality even at unequal lengths. At arbitrary caps equal lengths provide this property instead. These are different interval placements, not operand permutations. The menu contains descriptions, not Q prepared banks; generation and search may take $O(Qg)$, and only the chosen tree is built.

At $m = 2^H, M = d = H, L = H + 1$, $H \ge 16$, the singleton hotspot value is

$$V(q) = \Theta\left((1-\alpha^2)\frac{H}{m} + \alpha^2 \min\{1, 2^q H/m\}\right). \tag{51}$$

To realize it, take $h = \lfloor (H-1)/2 \rfloor$. A cyclic core of h uncovered positions has a complement with at most two runs. Split it into exactly $r = H + 1 - 2h \in \{2, 3\}$ saved intervals, giving $a = H$ and Build $n - H$. Choose $Q' = \min\{Q, \lfloor m/h \rfloor\}$ disjoint cores and translate them by one uniform cyclic offset. A history in their union selects its core; otherwise use a default. The guaranteed completion is

$$(1-\alpha^2)h/m + \alpha^2 Q' h/m.$$

Cyclic equivariance makes this a worst-hotspot guarantee. The capacity upper bound $|S_\pi| \le \lfloor (H+1)/2 \rfloor + H$ and the same message-union argument give the matching order. The pointwise bound (36) costs at most 6λ on any singleton, where $\lambda = H + n/(H+1)$. Every comparator pays at least $(1-\alpha)\lambda/7$ from the uniform background, proving mean ratio $42/(1-\alpha)$ for $\alpha < 1$ on this identical workload. At the striped-signal budgets the identical argument gives factor 84 for the half-uniform law; the signal comparison remains a deadline result.

E Statistical complexity and proper-decision lower bounds

E.1 Dimension and statistical transfer

The signed-threshold argument in Appendix G proves (14) for the entire paid class, including the ideal class at zero fees. It uses the capacity bound (12), not this statistical transfer. Write v for the resulting VC dimension.

For empirical success $\widehat{F}$ let $D_s = \sup_\pi |F_P(\pi) - \widehat{F}(\pi)|$. For $v \ge 1$, bounded VC uniform convergence (Gupta & Roughgarden, 2017, Theorem 3.2) gives $D_s \le C\sqrt{(v + \log(1/\delta))/s}$ with probability $1-\delta$, and $\mathbb{E}D_s \le C'\sqrt{v/s}$. An empirical τ-maximizer satisfies

$$\mathrm{OPT}_P - F_P(\widehat{\pi}) \le 2D_s + \tau.$$

Regret is at most one, and when $v = 0$ all classifiers agree so decision regret is zero. This proves Theorem 3. For singleton-or-no-change laws, the $M+1$-run bound instead gives VC dimension at most $2(M+1)+1$: an alternating labeling on $2(M+1)+1$ ordered positions needs too many positive runs, and the empty-mask atom adds at most one.

E.2 One testing argument for the three lower bounds

The following elementary preference-testing statement isolates the standard statistical step from the physical reductions. Consider $\theta \in [K]^g$ with preferred set $S_\theta = \{(i, \theta_i) : i \in [g]\}$. Suppose outputs reduce, independently of θ, to sets T of g candidates and, for a scale $0 < \gamma \le \gamma_0$ with fixed positive γ_0, obey

$$\mathrm{OPT}_\theta - F_\theta(T) \ge c_1 \gamma h/g, \quad h = g - |T \cap S_\theta|, \qquad \mathrm{KL}(P_\theta \| P_{\theta'}) \le c_2 \gamma^2 \,\mathrm{Ham}(\theta, \theta')/g. \quad (52)$$

For constants $c_1, c_2 > 0$, some γ forces expected regret $\Omega(\min\{1, \sqrt{g \log K/s}\})$ from s iid histories.

For $K \ge 16$, greedy Hamming packing at distance greater than $g/4$ has size at least $K^{3g/4}/2^g \ge K^{g/2}$: a radius-$\lfloor g/4 \rfloor$ ball contains at most $2^g K^{g/4}$ words. Choose γ a sufficiently small constant times $\min\{1, \sqrt{g \log K/s}\}$. The information in s histories under the uniform code prior is at most the maximum pairwise divergence, a small fraction of the log code size. Fano's inequality gives a constant decoding error. If $h < g/8$, then $|T \triangle S_\theta| = 2h < g/4$, so nearest-code-set decoding succeeds. Thus $\mathbb{E}h = \Omega(g)$, proving the bound.

For $2 \le K < 16$, restrict preferences to two alternatives per group. Decode a unique selected potential alternative as that group's sign, and otherwise toss a fair coin. If E, D, W count groups with zero, two, and one wrong selected potential alternatives, respectively, then $D \le E$ because at most g potential alternatives were selected. Neutral candidates need not be moved: $h = E + W$, whereas conditional expected sign errors are $W + (E+D)/2 \le h$. Taking γ a small constant times $\min\{1, \sqrt{g/s}\}$, adjacent parameters' s-history total variation is at most $1/4$ by Pinsker's inequality. Pairing adjacent signs under a uniform prior gives expected sign errors at least $3g/8$, hence $\mathbb{E}h \ge 3g/8$. This proves the bounded-K case and covers arbitrary allocations and randomized outputs.

E.3 Singletons and the zero-slack envelope

For the singleton lower bound take $p = 2g$ in Lemma A.2, so $m = 4g+1, M = g+1, d = L = 3g$. Give the empty mask probability $3/4$ and the two candidates in pair i probabilities $(1 \pm \beta\theta_i)/(8g)$, $\theta_i \in \{-1, 1\}$, $0 < \beta \le 1/2$. A feasible plan can be dominated by a timely g-set; its optimum is

$7/8+\beta/8$. Missing h preferred candidates loses $\beta h/(4g)$, even for allocations not selecting one per pair. Adjacent parameters have

$$\mathrm{KL}(P_\theta \| P_{\theta'}) = \frac{\beta}{4g} \log \frac{1+\beta}{1-\beta} \le \beta^2/g.$$

The binary testing argument with $\gamma = \beta$ gives $\Omega(\min\{1, \sqrt{g/s}\})$. With $g = M-1$ this proves the singleton rate; cold completion is zero and optimum exceeds $7/8$.

For the independent-update zero-slack family use Lemma A.1 with $N = gK$. Partition candidates into g groups of K, one preferred per group. Independently across candidates and histories, preferred candidates change with probability $p-\eta$ and all others with p, where $p = 1/(16g)$ and $0 < \eta \le p/2$. Deterministic separators and suffix preserve independence of input indicators. For a selected g-set missing h preferred candidates,

$$F_\theta(T) = (1-p+\eta)^{g-h}(1-p)^h, \quad \mathrm{OPT}_\theta = (1-p+\eta)^g > 15/16, \quad \mathrm{OPT}_\theta - F_\theta(T) \ge (15/16)\eta h. \tag{53}$$

The last bound telescopes h factors and uses $(1-p)^{g-1} \ge 15/16$. The Bernoulli inequality $\mathrm{KL}(\mathrm{Ber}(u)\|\mathrm{Ber}(v)) \le (u-v)^2/[v(1-v)]$ gives complete-history divergence at most $8\eta^2 \,\mathrm{Ham}(\theta,\theta')/p$. With $\gamma = g\eta \le 1/32$, this is at most $128\gamma^2 \,\mathrm{Ham}\,/g$, while regret is at least $(15/16)\gamma h/g$. Lemma A.1 covers all positive-value plans; zero-value outputs are dominated by any selected g-set. Therefore (52) applies and gives $\Omega(\min\{1, \sqrt{g \log K/s}\})$.

To cover all zero-slack envelope parameters, if $\bar{d}/M \ge 8$, choose $g = M-1$, $K = \lfloor \bar{d}/(4g) \rfloor \ge 2$. Then $M \le d = 4gK - g \le \bar{d}$ and $g \log K = \Theta(M \log(2 + \bar{d}/M))$. If $M \ge 3$ and $\bar{d}/M < 8$, instead put $g = \lfloor M/3 \rfloor$, $m = 4g+1$, $d = L = M$ and use the paired singleton law. Lemma A.2's inequality $a \ge 3t$ still permits at most g timely candidates. Every g-set is attainable: split its $g+1$ complementary runs, with $3g+1$ total inputs, into $r = M - 2g + 1$ roots. Writing $M = 3g + j$, $0 \le j \le 2$, gives $g+1 \le r = g+j+1 \le 3g+1$ and $r \le M$, with baseline $2g + r - 1 = M$. Since $g \ge M/5$, the singleton bound has the required order. For $M = 2$ in the remaining bounded-ratio range, $m = 2, d = L = 2, b = 1$ forces saving one leaf and protecting the other; the one-pair law gives the order $s^{-1/2}$. Only the explicit family (19) with these independent Bernoulli coordinates claims independent input changes and optimum above $15/16$.

E.4 Positive slack: learning fallback summaries

Let $g \ge 2, K \ge 2, t \ge 1$, $\zeta \in \{0,1\}$, and $r = g + \zeta$. Each of r regions is followed by an always-changing separator input. A variable region has order $R_i = B_{i1} Z_{i1} \cdots B_{iK} Z_{iK}$, with t inputs in every candidate block B_{ij} and internal separator block Z_{ij}. If $\zeta = 1$, the last region has $2Kt$ stable inputs. Set

$$\begin{gathered} m = r(2Kt+1), \quad M = 2g + \zeta, \quad \nu = 4Kt - 1, \quad w = 2t - 1, \\ b_0 = r\nu, \quad d = 3r - 1, \quad L_e = d + \nu - w + e, \quad 0 \le e < w. \end{gathered} \tag{54}$$

A scenario activates one variable region $I \sim \mathrm{Unif}[g]$: all its internal separators change; each candidate block changes wholly or not at all. Other regions are stable, and global separators always change. The blocks are charged inputs, not free macro-operations, and $\Delta_e = (4K-2)t + e$.

All-tree normalization. Delete intervals meeting global separators. Inside a candidate block replace all retained intervals by its root. Every other interval in a variable region meets an internal separator; if any is retained, replace all such intervals by the region root. In the stable region use its root or nothing. Records do not increase. When active, the replaced non-candidate intervals were invalid and the enlarged candidate roots skip at least as much work; when inactive, the region root skips everything. Thus the new bank dominates pointwise. Its Build may increase, but is at most ν per region and therefore within b_0. Canonical refinement supplies one compatible tree and schedule.

If T_i are selected candidate roots and H_i stay stable when region i activates, active work is $\nu - wH_i$. Inactive work is zero with the region root, and $\nu - w|T_i|$ without it. A rootless region therefore costs at least $\nu - Kw$, and

$$2(\nu - Kw) = 4Kt + 2K - 2 > \nu - 1 \ge L_e - d. \tag{55}$$

Two missing variable roots always fail. One missing variable root permits success only when that region activates; a missing stable root contributes $\nu > L_e - d$ and always fails.

The law and arbitrary allocations. Fix $p = 1/16$, $0 < \eta \le p/2$, and $\theta \in [K]^g$. Conditional on activation of i, its candidate blocks change independently with $p_{ij} = p - \eta \mathbf{1}\{j = \theta_i\}$. Complete masks reveal I and every active candidate bit; inactive coordinates are zero, not censored. A missing-root plan has value at most $1/g \le 1/2$. All region roots and one fixed candidate per region have value at least $1 - p$, so they give a parameter-independent dominating replacement at the allowed budgets. With every region root present, fill the remaining slots to exactly g candidate roots without extra Build. Completion throughout the cap band is exactly $H_I \ge 1$, and hence

$$F_\theta(T) = \frac{1}{g} \sum_{i=1}^{g} \left(1 - \prod_{j:(i,j)\in T} p_{ij} \right), \qquad \mathrm{OPT}_\theta = 1 - p + \eta > 15/16. \tag{56}$$

There is a stronger normalization for this family. If one region has at least two candidates, another is empty. Move any candidate from the former to any fixed candidate of the latter. Removing it loses at most p/g completion, since another selected candidate remains; adding it gains at least $(1-p)/g$. The exchange is independent of θ, stays within the same Build and records, and strictly improves completion for every θ. Iteration leaves one candidate per region. This is distribution-wise domination, not pointwise domination of masks. For the resulting set, missing h preferred candidates gives exactly

$$\mathrm{OPT}_\theta - F_\theta(T) = \eta h / g,$$

proving the optimum in (56) and reducing every proper output to one choice per region.

The observed activation mixture has

$$\mathrm{KL}(P_\theta \| P_{\theta'}) \le \frac{8\eta^2}{gp} \mathrm{Ham}(\theta, \theta') = 128\eta^2 \mathrm{Ham}(\theta, \theta')/g.$$

Repeated bits within a block are deterministic copies and add no information. Equation (52) with $\gamma = \eta$ now proves the lower order $\Omega(\min\{1, \sqrt{g \log K/s}\})$. The same law and η work for all w integer caps in (54). At $t = 1$, $d = \Theta(M)$ and $\Delta = 4K - 2$; at $t = K$, $\Delta = \Theta(K^2)$ and the cap band has width $2K - 1 = \Theta(\sqrt{\Delta})$, matching the upper logarithm in both cases.

Mean work on this family. The pointwise normalization also preserves mean comparisons. Suppose it omits h variable roots and $\ell \in \{0, 1\}$ stable roots. There are at most $g + h + \ell$ selected candidates. With $B = \sum_{i:R_i \text{ omitted}} (\nu - w|T_i|)$, its mean is

$$d + \nu - \frac{w}{g} \sum_{(i,j)\in T} (1 - p_{ij}) + \frac{g-1}{g} B + \ell\nu.$$

Selected stability is at most $g(1-p+\eta) + (h+\ell)(1-p)$, and $B \ge h(\nu - Kw)$. Since $(g-1)(\nu - Kw) > w(1-p)$ and $g\nu > w(1-p)$, each omitted root costs more than its extra candidate can save. All roots with the preferred candidates attain minimum mean $d + \nu - w(1-p+\eta)$. Root-only mean $d+\nu$ therefore has ratio at most $1 + w/(d+\nu-w) < 1 + 1/(2K-1)$, but no completion anywhere in the band. By contrast, fixed candidates attain at least $15/16$ completion without history.

Completing the envelope and its limits. Formally the quantity in Theorem 4 is

$$\sup_{\substack{m,\ M \le d \le \bar{d},\ 0 \le \Delta \le \bar{\Delta} \\ d + \Delta < 2m - 1}} \inf_{\mathcal{A}} \sup_{P} \mathbb{E}[\mathrm{OPT}_P - F_P(\mathcal{A}(J_1, \ldots, J_s))],$$

where the learner knows the instance and returns a feasible plan. The VC bound proves the upper order. For $M \ge 4$, set $g = \lfloor M/2 \rfloor$, $\zeta = M - 2g$ in (54); then $M \le d = 3(g + \zeta) - 1 \le \lceil 3M/2 \rceil$. If $\bar{\Delta} \ge 6$, choose $t = 1, K = \lfloor (\bar{\Delta} + 2)/4 \rfloor \ge 2$. This supplies the $\log(2 + \bar{\Delta})$ term. For $\bar{\Delta} < 6$ it is constant and the zero-slack construction suffices. Finally the maximum of $\log(2 + \bar{d}/M)$ and $\log(2 + \bar{\Delta})$ is of order $\log(2 + \bar{d}/M + \bar{\Delta})$; the two lower bounds need not use the same instance or dependence structure.

The condition $g \ge 2$ is substantive: with $g = 1, K = 2, t = 1, p = 1/16, \eta = 1/32$, two candidate leaves without a region root have value $511/512$, greater than root-plus-preferred value $31/32$. This argument does not settle positive slack for $M = 2, 3$, all $M \le \bar{d} < \lceil 3M/2 \rceil$, or fully independent changes. The growing cap band is not a fixed-relative-width band.

F EXACT COMPILERS AND THE COMPUTATIONAL BOUNDARY

F.1 SINGLETON-OR-NO-CHANGE LAWS

Take rational weights $p_0, p_1, \ldots, p_m$, which may be population probabilities or empirical frequencies. Cap M at n. If $L \geq n$, the empty bank succeeds everywhere; if $d > L$, all plans fail. Otherwise $0 \leq d \leq L < n$. For an interval I, write $a = C_I(\varnothing)$ and $\bar{a} = \min(a, L+1)$. Store

$$F_I(r, \bar{a}, \ell) = \max \sum_{j \in I} p_j \mathbf{1}\{C_I(\{j\}) \leq \ell\}, \qquad \ell \in \{-1, 0, \ldots, L\},$$

over actual interval plans with exactly r records and clipped baseline $\bar{a}$, or $-\infty$ if infeasible. A feasible state at $\ell = -1$ has value zero and must not be discarded: an ancestor may skip its unchanged execution. At leaf j, choices $(r, \bar{a}) = (0, 1), (1, 0)$ both have value $p_j \mathbf{1}\{\ell \geq 1\}$.

For each ordered split $I = XY$, child resource states, and parent retention flag $z \in \{0, 1\}$, set

$$r = r_X + r_Y + z, \qquad \bar{a} = \begin{cases} 0 & z = 1, \\ \min(L+1, 1 + \bar{a}_X + \bar{a}_Y) & z = 0, \end{cases}$$

$$F_I(r, \bar{a}, \ell) \leftarrow \max\big\{F_I(r, \bar{a}, \ell),\ F_X(r_X, \bar{a}_X, \max(-1, \ell - 1 - \bar{a}_Y)) + F_Y(r_Y, \bar{a}_Y, \max(-1, \ell - 1 - \bar{a}_X))\big\}. \tag{57}$$

Discard $r > M$ and retain backpointers. A singleton in X invalidates the parent whether saved or not, leaving one parent operation and unchanged sibling cost a_Y; symmetrically for Y. Once child baselines and record counts are fixed, each objective depends on the other child only through its baseline, so the separately chosen child optima are compatible.

Clipping is exact: if a sibling baseline exceeds L, its other child has a negative remaining cap and contributes zero for every $\ell \leq L$; otherwise it is exact. Nonnegative addition commutes with saturation at $L + 1$. A saved parent can reset an unchanged overflow baseline to zero, explaining why overflow resource states and their $\ell = -1$ entries must be retained. Induction on interval length proves that every legal plan gives an enumerated transition and every finite optimum has one actual witness. At the full root choose

$$\max_{r \leq M,\ d \leq \bar{a} \leq L} \{F_{[m]}(r, \bar{a}, L) + p_0\},$$

with the feasible zero-value empty bank as fallback. Here $\bar{a} = a$ is exact and $a \geq d$ enforces Build $n - a \leq n - d$. Backtracking gives the tree and bank, and postorder its Build schedule. There are $O(m^2(M+1)(L+2)^2)$ states. Across $O(m^3)$ split positions, enumerating two record counts, two baselines, and one cap proves (17). Only sums and comparisons of rational weights occur, with polynomial bit lengths.

F.2 ARBITRARY COMPLETE MASKS

For a concrete interval plan define $\xi_I(J) = C_I(J) - a_I \geq 0$. At split $I = XY$, with $q_I = 1 + a_X + a_Y$ and root flag z,

$$a_I = (1 - z) q_I, \qquad \xi_I(J) = \xi_X(J) + \xi_Y(J) + z q_I \mathbf{1}\{J \cap I \neq \varnothing\}. \tag{58}$$

A positive residual implies an internal change; no valid ancestor can hide it. The final baseline is at least d, so residual exceeding Δ already precludes completion. For every interval, store one actual witness for each state

$$(r, \bar{a}, e_1, \ldots, e_s), \qquad \bar{a} = \min(a_I, L+1), \quad e_i = \min(\xi_I(J_i), \Delta + 1).$$

At a leaf, the unsaved state is $(0, 1, 0, \ldots, 0)$; the saved state is $(1, 0, \mathbf{1}\{j \in J_1\}, \ldots, \mathbf{1}\{j \in J_s\})$. At a split put $q = 1 + \bar{a}_X + \bar{a}_Y$ and enumerate $z = 0, 1$:

$$r = r_X + r_Y + z, \quad \bar{a} = (1 - z)\min(L+1, q), \quad e_i = \min\{\Delta + 1, e_{X,i} + e_{Y,i} + z q \mathbf{1}\{J_i \cap I \neq \varnothing\}\}. \tag{59}$$

If both child baselines are at most L, q is exact. Otherwise both it and the true q_I are at least $L + 2 > \Delta + 1$; their changed shifts saturate alike, and their unchanged shifts vanish alike. Residual

addition commutes with clipping because it is nonnegative. Thus induction proves exact clipped states in both directions. Two witnesses with the same state are interchangeable in every exterior transition, so one representative loses no empirical optimum; no incompatible coordinatewise optima are combined.

At the root accept $r \le M, d \le \bar{a} \le L$ and maximize $|\{i : \bar{a} + e_i \le L\}|$, or return the empty bank if none is accepted. The baseline is exact, and reconstruction gives Build $n - a$. There are at most $(M+1)(L+2)(\Delta+2)^s$ states per interval. Pairs of child states over all splits, with $O(s)$ transition work, give (18), with polynomial integer bit lengths apart from the displayed state count.

F.3 Hardness under the same execution interface

Let a graph have $p \ge k \ge 2$ vertices and a nonempty edge set. Apply Lemma A.2 with $g = k$: $m = 2p + 1, M = k + 1, d = L = 3k, b = 4p + 1 - 3k$. Give the empty scenario probability $7/8$ and each edge's two-candidate change probability $1/(8|E|)$. The lemma reduces every positive-value preparation to an uncovered k-set, and every such set is realizable. Therefore

$$\mathrm{OPT} = \frac{7}{8} + \frac{1}{8|E|} \max_{|S| \le k} |E(G[S])|. \tag{60}$$

Testing whether the induced-edge optimum is $\binom{k}{2}$ decides CLIQUE. Thus exact optimization is NP-hard when M $(= k + 1)$ and the scenario count are both part of the input (for fixed M, scoring all laminar families of at most M intervals is polynomial, and for a fixed scenario count so is the bound (18), since $\Delta < n$), even though every primitive cost is one, all nodes are retainable, Build is binding in realizing banks, cold completion is zero, and optimum is at least $7/8$. A deterministic FPTAS with $\epsilon = 1/(16|E|)$ returns a feasible value within $1/(16|E|)$ of optimum, less than the spacing $1/(8|E|)$, and recovers the exact induced-edge optimum in polynomial time. The same law is empirical: use $7|E|$ empty masks and one mask per edge. This rules out such an FPTAS unless $P = NP$, not every approximation or every subexponential search.

G Fixed checking and restoration fees

This appendix proves Proposition 5 and the dimension bound in Theorem 3 by one argument. The ordered computation and primitive Build costs are unchanged. Public interval fees are fixed before sampling, nonnegative, and independent of values and changes. They charge only post-expiry access. Denote ideal work by C^0_π, and put $Z_v(J) = \mathbf{1}\{J \cap I_v \ne \varnothing\}$.

G.1 A signed interval identity and pointwise deletion

For the retained-node forest, let $\mathrm{ch}(v)$ be the maximal retained descendants of v, and let $\mathcal{R}$ be its maximal roots. The ideal component weights a, w_v are those of (3). Every maximal root is reached. A nonmaximal retained child is reached exactly when its retained parent is invalid. Accordingly,

$$\begin{aligned} C^f_\pi(J) &= A_\pi + \sum_{v \in A} \lambda_v Z_v(J) \ge C^0_\pi(J), \\ A_\pi &= a + \sum_{v \in \mathcal{R}} (\kappa_v + \rho_v), \\ \lambda_v &= w_v - \rho_v + \sum_{u \in \mathrm{ch}(v)} (\kappa_u + \rho_u). \end{aligned} \tag{61}$$

To derive the coefficients directly, each nonmaximal node u with retained parent v contributes $Z_v\kappa_u + (Z_v - Z_u)\rho_u$ in fees; a maximal node contributes $\kappa_u + (1 - Z_u)\rho_u$. Add these terms to the ideal component identity and collect each Z_v. Nonnegative added fees prove the pointwise lower bound, without any sign restriction on λ_v.

Negative coefficients can matter. On four inputs retain $[1, 1]$ and $[2, 4]$, charge two only for restoring $[1, 1]$, and set all other fees to zero. Build is six, but $C^f(\varnothing) = 3 > C^f(\{1\}) = 2$: at cap two an edit changes failure to success. The following deletion addresses such avoidable expense.

Process retained intervals from smallest to largest. Recompute the retained children after each deletion and put

$$f_v = w_v + \sum_{u \in \mathrm{ch}(v)} (\kappa_u + \rho_u).$$

This is the unchanged fallback cost if v is not retained. Delete v when $\kappa_v + \rho_v \geq f_v$. If reached and valid, fallback costs no more than the deleted check and restoration. If reached and invalid, removing v saves its check while leaving the same recursion below it. If a valid ancestor intercepts the request, neither execution visits v. Each deletion therefore weakly lowers cost for every mask and cannot enlarge the Build union or record count. Once descendants have been processed they are unchanged by any subsequent ancestor deletion. Every surviving record consequently satisfies

$$\lambda_v = f_v - \rho_v > \kappa_v \geq 0.$$

Composing the pointwise comparisons yields one future-independent dominating subbank, simultaneously for every law and cap. This does not assert that this subbank dominates all other subbanks. With zero fees, all ideal components have positive integer weights and no record is deleted. In the example, deleting the leaf gives cost two in both cases.

G.2 One dimension argument for both interfaces

Suppose $J_1, \ldots, J_v$ are shattered by paid completion indicators. The all-positive labeling supplies a feasible π_+ with $C^f_{\pi_+}(J_i) \leq L$ for all i. If $j \in J_i$, then

$$C^0_{\pi_+}(\{j\}) \leq C^0_{\pi_+}(J_i) \leq C^f_{\pi_+}(J_i) \leq L.$$

The first inequality uses only ideal monotonicity. Thus the entire ordered union $Z = \bigcup_i J_i$ has size at most $G = \min\{m, \lfloor (d+1)/2 \rfloor + M\Delta\}$ by (12). Neither monotonicity of paid work nor the deletion just proved is required.

Project any other bank onto ordered Z and remove empty projections. At most M laminar intervals remain; distinct original intervals can have the same projection. An r-interval laminar multiset has at most 4^r parenthesis-forest shapes. Its ordered endpoints have at most $\binom{G+2r}{2r}$ weakly increasing placements; allowing impossible placements only enlarges the count. Each fixed structure maps a mask to r binary hit features. By (61), every completion indicator is an affine halfspace on these features, even with arbitrary signed real coefficients. The halfspace dimension and growth bounds (Shalev-Shwartz & Ben-David, 2014, Theorem 9.3 and Lemma 6.10) give, for $v \geq M + 1$,

$$2^v \leq \left[\sum_{r=0}^{M} 4^r \binom{G+2r}{2r} \right] \left(\frac{\mathrm{e}v}{M+1} \right)^{M+1}. \tag{62}$$

The threshold absorbs A_π and includes constant classifiers. For $G \geq 1$, bound the sum by $M+1$ times its last term and use $\binom{G+2M}{2M} \leq [\mathrm{e}(G+2M)/(2M)]^{2M}$. Its logarithm is $O(M \log(2+G/M))$. Writing $x = v/(M+1)$ yields

$$x \leq C \log_2(2 + G/M) + \log_2(\mathrm{e}x),$$

so $x = O(\log(2 + G/M))$; for example, for sufficiently large x the last term is at most $x/2$. Since $G/M \leq (d+1)/(2M) + \Delta$, this proves the asserted dimension. If $v < M + 1$ it is immediate; if $G = 0$, only the empty mask can occur and $v \leq 1$. Setting every fee to zero proves (14) as a special case. No coefficient discretization or fee precision enters.

The uniform-convergence and empirical-optimization argument in Appendix E applies to the paid class verbatim, with its own population and empirical completion values. It proves both regret statements against the paid optimum. These are statistical guarantees for a fixed observation interface and fixed fee schedule; they neither price optimization nor prove an efficient optimizer for general real fees. The ideal lower-bound instances are included when fees vanish, but do not establish a matching lower bound for every prescribed positive fee schedule.

H Supporting experimental evidence

Controlled laws test information requirements and learning; revision histories separate objective, information and plan-class effects; trained readers and measured operators test whether saved work

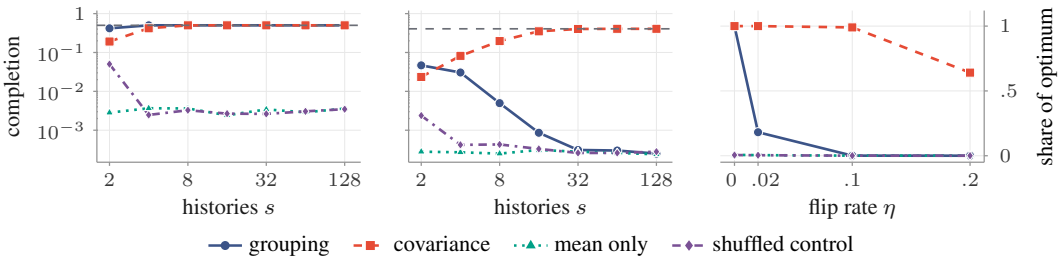

Figure 3: Learning from finite history. Left: noiseless changes. Center: 10% group-internal flips. Both use $g = 64$, $e = g/8$; the dashed rule is the informed optimum. Right: completion at 32 histories as a share of the noise-specific optimum. Under the tested flips, covariance outperforms exact column matching; this is not a uniform noise-robustness guarantee. The shuffled control preserves observed means but destroys synchronous change.

yields runtime benefits. Appendix I compares LLM reuse interfaces, including the expert-residency results in Section 6.

H.1 Finite-history learning and imperfect correlation

We test whether an optimal decision requires identifying the hidden group H in Theorem 2, and how imperfect correlation affects learning.

We use the family at $g \in \{16, 64, 128\}$ and its original (m, M, b, d): members of H share a fair bit followed by independent Bernoulli-η flips, $\eta \in \{0, .02, .1, .2\}$; outsiders are independent. Each of 256 trials per setting draws a uniform H and 128 iid histories (nested prefixes). With $g + 1$ records and $9g - 1$ Build jobs, four arms select g columns: uniformly (mean-only, as exact feasible means are independent of H); a largest class of identical history columns (grouping); the largest sums of each column's $g - 1$ largest off-diagonal sample covariances (covariance); and covariance on time-shuffled columns, which keeps means but destroys synchronous change. Two group columns agree on s histories with probability $[(1-\eta)^2 + \eta^2]^s$, so under noise more history can undermine exact matching without reducing the available information. Each noise level is scored against its own informed optimum, computed exactly by binomial convolution, since selecting H at zero slack completes only with probability $\frac{1}{2}[(1-\eta)^g + \eta^g]$.

At $g = 64$, $\eta = 0$ and eight histories, grouping/covariance identify H in only 196/195 of 256 trials, yet both attain optimal completion at $e = 8$ in all trials. Identification is therefore stricter than decision optimality in these samples. At $\eta = .1$, covariance remains near the optimum while exact matching and shuffled controls deteriorate (Figure 3, Table 5). At $\eta = .2$ it is suboptimal at every size; the $g = 64$ optimum itself is only .04221. Trials, not shared prefixes, caps or arms, are statistical units. These controlled results do not establish a noisy minimax rate, uniform robustness or deployment gain.

H.2 Strong statistics, a matched pair, and revision replay

For the information claim, an exact pair isolates what interval and pairwise statistics miss; ordinary laws and revision histories test whether it matters. All studies use $(m, M, b, d) = (8, 3, 10, 5)$, caps $\mathcal{L} = \{5, 7, 9, 11, 13\}$ fixed before evaluation, and exact evaluation of the 2,236 feasible banks on all 256 masks. Mean training minimizes expected work. Besides the completion learners of Section 6, an independent baseline uses the product of the marginals of $\widehat{p}$ and a balanced baseline restricts full history to a midpoint-split tree. Ties within 10^{-10} go to the first bank by record count, then by lexicographic interval endpoints; maximum-entropy fits minimize the log-partition dual and are accepted only at maximum moment residual $5 \cdot 10^{-7}$.

Ordinary laws. There are three laws of each kind: independent, with eight rates uniform on $[.05, .35]$, or mixtures of three product laws with $\mathrm{Dirichlet}(1.5, 1.5, 1.5)$ weights and rates .65 inside a group and .03 outside, the group being an interval of uniform length in $\{2, 3, 4\}$ (contiguous bursts) or a uniform three-element set (dispersed bursts). All have full support and no forced separator or

Table 5: Completion at 32 histories and $e = g/8$, averaged over 256 trials per size/noise setting. The last column is the noise-specific informed optimum; other columns are returned banks at identical physical budgets.

g	η	Mean-only	Grouping	Covariance	Shuffled	Optimum
16	0	0.0845	0.5000	0.5000	0.0850	0.5000
16	0.02	0.0771	0.2243	0.4982	0.0754	0.4982
16	0.1	0.0430	0.0445	0.3932	0.0445	0.3946
16	0.2	0.0214	0.0214	0.1522	0.0196	0.1759
64	0	0.0034	0.5000	0.5000	0.0026	0.5000
64	0.02	0.0023	0.0907	0.5000	0.0021	0.5000
64	0.1	0.0003	0.0003	0.4026	0.0003	0.4067
64	0.2	$1.5{\cdot}10^{-5}$	$1.3{\cdot}10^{-5}$	0.0271	$1.5{\cdot}10^{-5}$	0.0422
128	0	0.0001	0.5000	0.5000	0.0001	0.5000
128	0.02	$3.8{\cdot}10^{-5}$	0.0311	0.5000	0.0001	0.5000
128	0.1	$5.2{\cdot}10^{-7}$	$7.3{\cdot}10^{-7}$	0.4250	$8.2{\cdot}10^{-7}$	0.4306
128	0.2	$1.5{\cdot}10^{-9}$	$1.5{\cdot}10^{-9}$	0.0040	$2.1{\cdot}10^{-9}$	0.0091

stable suffix. Each law has 32 independent training trials of 512 histories and exact evaluation. Full history is not uniformly preferable (Table 6): the correctly specified independent model is better on independent laws, and interval and pairwise models nearly match full history on bursts.

Table 6: Completion at $(m, M, b) = (8, 3, 10)$, averaged over caps $5, 7, 9, 11, 13$. Top: three laws per kind, 32 training trials of 512 histories, exact evaluation. Bottom: episodes of h updates (commits or revisions), chronological testing after 256 training episodes, averaged over histories. All methods choose among the same 2,236 banks; tail objectives break ties by mean work (Appendices H.2 and H.11).

Update source	h	Mean work	CVaR_{95}	90th pct.	Intervals	Pairwise	Marginals	Full history
Independent		.5431	.2989	.5309	.6378	.6376	.6401	.6364
Contiguous bursts		.6859	.6081	.6694	.7838	.7842	.7410	.7838
Dispersed bursts		.5781	.3641	.5260	.6659	.6663	.6489	.6662
Python projects	1	.6325	.4020	.6396	.7238	.7298	.7148	.7338
	4	.4143	.2800	.3551	.4784	.4798	.4751	.4805
ML systems	1	.6425	.4264	.5867	.7321	.7354	.7250	.7345
	4	.3628	.2301	.3186	.4368	.4386	.4321	.4317
Wikipedia	1	.8122	.6343	.7765	.8715	.8714	.8716	.8700
	4	.6670	.4156	.6340	.7572	.7567	.7598	.7569

An exactly specified full-support pair. For mask index $x = \sum_{i=1}^{8} 2^{i-1}\mathbf{1}\{i \in J\}$, put

$$p_0(x) = \frac{3^{8-|J|}}{4^8}, \qquad p_\pm(x) = p_0(x) \pm \frac{z_x}{1310720}, \tag{63}$$

with the integers z_x listed below. Both laws are normalized, have density ratio in $[1/5, 9/5]$ to p_0, marginal $1/4$ and pair moment $1/16$ for every distinct input pair, and equal Build-feasible interval avoidance probabilities, hence equal feasible plan means. Rational enumeration at cap 9 gives

$$\text{OPT}_+ = \frac{445981}{655360}, \quad \text{OPT}_- = \frac{450323}{655360}, \quad \text{OPT}_0 = \frac{1215}{2048}.$$

Hence the exact-statistic minimax regret is $(\text{OPT}_+ + \text{OPT}_-)/2 - \text{OPT}_0 = 7419/81920$, attained by mixing $\{[1,2],[3,4],[7,8]\}$ and $\{[1,2],[5,6],[7,8]\}$ with weights $70319/149256$ and $78937/149256$. Found by classical moment-constrained search (Bertsimas & Popescu, 2005,

Table 7: Regret on the matched pair. Entries average 64 training trials in each world at cap 9 and free reuse. The optimal exact-population-statistic decision has regret .090564 in each world; it is a different information interface from the empirical baselines.

Histories	Independent	Intervals	Pairwise	Full history
32	.093651	.088838	.078713	.033656
128	.095982	.090318	.085175	.008721
512	.100967	.092257	.089360	.000237
2048	.097059	.091819	.091565	.000000

Sec. 2) over 40 candidates for each of five caps, the pair shows existence, not frequency in ordinary or natural workloads. Both exact-population maximum-entropy closures equal p_0, since $H(X_1, \dots, X_8) \le \sum_i H(X_i)$ with equality only at the independent law, which meets all interval and pairwise constraints.

A continuous family near independence. For public $\lambda \in [0, 1]$, $P_{\lambda,\pm} = (1-\lambda)P_0 + \lambda P_\pm$ keeps the matched statistics of P_0, and

$$1 - \tfrac{4}{5}\lambda \le \frac{p_{\lambda,\pm}(x)}{p_0(x)} \le 1 + \tfrac{4}{5}\lambda, \qquad \inf_{\mathcal{A}} \max_{\sigma \in \{+,-\}} \left[\mathrm{OPT}_{\lambda,\sigma} - \mathbb{E}_{\mathcal{A}} F_{P_{\lambda,\sigma}}(\widehat{\pi})\right] = \lambda \frac{7419}{81920}. \tag{64}$$

Here $\mathcal{A}$ knows the family, λ and the common statistics but not the sign, and may randomize over feasible banks independently of the future; the same mixture is optimal for every λ. At $\lambda = 1/4$ every mask probability is within $[.8, 1.2]$ times its independent reference, yet regret is 2.2641 percentage points; this is one constructed direction, not a neighborhood-wide claim.

Proof. Density and moment statements follow by linearity. The banks $A_+ = \{[1,2],[3,4],[6,7]\}$ and $A_- = \{[2,3],[5,6],[7,8]\}$, with Build nine and three records, satisfy $F_{P_0}(A_\pm) = \mathrm{OPT}_0$ and $F_{P_\pm}(A_\pm) = \mathrm{OPT}_\pm$ exactly. Under a mixed law no bank exceeds the mixture of the two optima, which $A_\pm$ attains, so $\mathrm{OPT}_{\lambda,\pm} = (1-\lambda)\mathrm{OPT}_0 + \lambda \mathrm{OPT}_\pm$. As $(P_{\lambda,+} + P_{\lambda,-})/2 = P_0$, a common randomized decision has average completion at most OPT_0, hence worst-sign regret at least $\lambda[(\mathrm{OPT}_+ + \mathrm{OPT}_-)/2 - \mathrm{OPT}_0] = \lambda\, 7419/81920$. The attaining randomized policy has P_0-completion OPT_0 and expected endpoint regret $7419/81920$ in either world, so by linearity its regret along either segment equals this bound. □

Learning on the pair. Each world uses 64 independent trials with nested history lengths $32, 128, 512, 2048$; learners observe masks, not world labels, and population evaluation is exact. Full-history regret vanishes in all 2048-history trials, whereas moment-model regret persists (Table 7); this is sampled recovery, not a guarantee. The exact-statistic minimax value does not bound empirical-moment methods, whose joint fluctuations may be informative. Reoptimizing with checking/restoration fees $(\kappa, \rho) = (.1, .1)$ reduces minimax regret to $.034148$; at $(.5, 1)$ it vanishes, with both optima still $.237305$.

Coefficients. $z_0, \dots, z_{255}$ in mask-index order:

437, 34992, -4303, -11664, 7754, -11664, -11664, -3888, 34992, -5721, -11664, -3888, -6579, -3888, -1956, -1296, -34992, 5046, 11664, 3888, 11664, -2337, 3888, -13, 782, -3888, 3888, -454, 3888, -1296, -1296, -432, -7798, -11664, 8208, 3888, 1399, 3888, 3888, 1296, -11664, 3888, 3888, 1296, -1713, 1296, 1296, 432, 1993, -3888, -1593, 1296, 794, 1296, -1296, 432, -3888, 1296, -279, 432, 1296, 432, -291, 144, 4885, -11664, -540, 3888, -8208, 3888, 3888, 1296, -11664, 2016, 3888, 1296, 2862, 1296, 1296, 432, 11664, -3888, -3888, -883, -3888, 1296, 20, 432, -3888, 1296, 1296, 432, 136, 432, 432, 144, 11664, -3888, -3888, -1296, -3888, 1296, 41, 432, -3888, 1296, -1296, 432, 1296, 432, 432, 144, 3888, -726, -1296, -432, -1296, 432, -432, 144, 514, 432, -432, 144, -432, 144, -21, 48, -34992, -22, 11664, 3888, 11664, 2614, 3888, 1296, -6712, -3888, 3888, 1296, 3888, 251, 1296, -19, 8651, 3888, -3428, -1296, -3888, -1296, -1296, -432, 3888, -759, -1296, -432, -1296, -432, -432, -144, 11664, -2570, -3888, -1296, -3888, -1296, -1296, -432, 3050, 1296, -1296, -432, -1296, 432, -432, -144, 3888, 1296, -1296, -432, -1296, -432, -432, -144, 1296, 432, -432, -144,

Table 8: Revision-replay details. Independent and balanced columns are completion probabilities; the final three columns are full-history work, Build and record use, averaged over the five fixed caps. Budgets are upper limits, not a requirement to use identical resources.

Trace	Test n	Empty (%)	Indep.	Balanced	Work	Build	Records
Flask, 1	2019	.297	.9913	.7489	7.04	8.6	2.6
Flask, 4	312	.000	.9628	.5878	7.64	8.6	2.6
Requests, 1	2404	.749	.6910	.6980	9.73	7.6	2.8
Requests, 4	409	.244	.4049	.3848	12.26	7.6	2.8
pytest, 1	3840	.000	.6731	.6153	10.13	9.0	2.8
pytest, 4	768	.000	.3768	.3464	12.33	8.4	2.8

Table 9: Fixed-decision replay. Entries are full-history-only successes / baseline-only successes among the same test episodes. All five caps are shown for the positive Requests comparison and the adverse comparisons; counts across caps are paired, not independent trials.

Trace, horizon	Test n	$L = 5$	7	9	11	13
Requests, 1: intervals	2404	$0/0$	$836/589$	$129/74$	$70/31$	$29/14$
Requests, 1: pairwise	2404	$0/0$	$0/0$	$129/74$	$70/31$	$0/0$
Flask, 4: either	312	$0/0$	$0/0$	$0/0$	$0/0$	$0/2$
pytest, 1: intervals	3840	$0/0$	$0/0$	$0/0$	$5/53$	$0/0$
pytest, 4: pairwise	768	$21/0$	$0/0$	$0/0$	$0/0$	$7/44$

-432, 144, -144, -48, 10252, -3888, -3888, 955, -3888, 1296, 1296, 432, -3888, 1296, 351, 432, -621, 432, 432, 144, 3888, -1296, -1296, -432, -1296, 432, -432, -144, -183, 432, -432, -2, -432, 144, -144, 48, 3888, -1296, -1296, -432, -1296, 432, -432, -144, 1256, 432, -432, 144, -432, 144, 95, 48, 1020, -432, -432, -144, -432, 144, -144, -48, -3, 144, -144, -48, -144, 48, -48, -16.

Temporal replay of public revisions. Does full history help on natural traces? We replay at most 4,096 consecutive first-parent transitions of Flask, Requests and pytest in chronological order, with eight regions fixed at lexical quantiles of the earliest snapshot holding at least eight paths; episodes mark every region touched by one or four consecutive, non-overlapping transitions. All six methods train once on the first 256 episodes and are tested on every later complete episode, including empty masks, with banks prepared anew and held immutable; no test outcome selects caps, methods or boundaries (Table 8).

Paired 95% intervals come from 2,000 moving-block resamples of whole test episodes (primary block length 16; caps and methods kept together). For Requests at horizon 1, full history minus intervals is $[1.80, 4.24]$ percentage points and minus pairwise $[.43, 1.15]$, positive also at block lengths 4 and 64. Adverse cases remain: pytest loses .25 points to intervals at horizon 1 ($[-.33, -.18]$) and .42 to pairwise at horizon 4 ($[-.94, .08]$); Flask loses .13 to both at horizon 4 ($[-.32, 0]$). The Requests gain moves which requests cross the deadline rather than lowering work uniformly (Table 9; at cap 9, equal-size full-history and moment-model banks cost 9 versus 12 on 48 episodes and the reverse on 36). At cap 7, 802 of the 836 wins against intervals come from the single-region mask $\{3\}$, so that gain is not by itself evidence that high-order statistics are necessary. These describe fixed plans on dependent traces (horizons share commits) without stationarity guarantee or multiplicity correction; lexical regions are not a dependency graph, and the replay neither measures build acceleration nor establishes the iid theorem, general superiority of full history, or GPU gains.

H.3 Forecast structure and decision learning

We check on exact finite instances that forecasts matched in message length, mutual information and posterior values can differ in decision value (information), and how closely planners fitted to histories approach their known-law optima (learning). We evaluate Proposition C.1's law exactly via (39) on all future positions, over small grids with $K \leq 32$ and every striped refinement q, taking

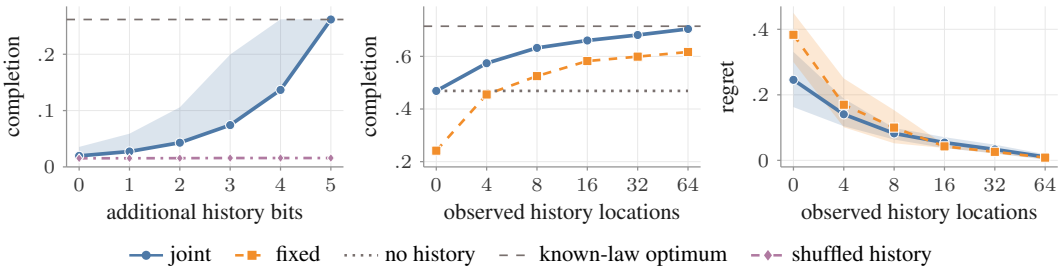

Figure 4: Left: matched forecasts at $K = 32$ (solid: construction; dashed: consecutive; dash-dotted: shuffled; shading: striped-optimum bounds). Center: completion at fixed budgets. Right: regret to each learner's known-law tree-class optimum, with descriptive 95% intervals.

Table 10: Budgeted three-seed fits (not published best accuracies) and independent-edit unit-completion gains of joint over fixed trees, both using history; intervals from 5,000 bootstrap resamples of the 32 TEST snapshots per task.

	Clean TEST accuracy (%)			Unit completion gain
Task	Affine	InceptionTime	MiniRocket	Joint–fixed, pp [95%]
FordA	49.70	95.18	94.70	7.13 [5.76, 8.50]
SCP1	79.86	78.50	91.58	10.55 [9.01, 12.01]
MotorImagery	49.00	53.00	50.00	9.64 [8.08, 11.18]

R the smallest power of two exceeding $K+1+3M$; values are those of the construction, not global optima. At $K = 32, M = 2, \Delta = 0$, consecutive and unrefined striped forecasts complete with probability .2617 and .0195; the construction first reaches $1/8$ at $q = 2, 3, 4$ for $K = 8, 16, 32$. Shuffled histories or independent futures give the null completion $1/R$ at $M = 2, \Delta = 0$.

For learning, $(m, M, d, b, L) = (16, 4, 4, 27, 8)$ and $P_\theta = \frac{1}{2}\delta_\theta + \frac{1}{2}\,\mathrm{Unif}[16]$; thirty independent trials draw a uniform hotspot and nested iid histories of length up to 64, and evaluation enumerates all 16 future positions. With 64 histories, empirical joint and fixed planners reach .7042 and .6167 against known-law optima .7146 and .6250; uniform planners give .4688 and .2417, the mean construction .0646. Joint regret to its own optimum is .0104 (descriptive 95% interval $[.0021, .0188]$, 5,000 whole-trial resamples). Nested prefixes and enumerated futures are not independent trials, and the curves (Figure 4) do not establish a minimax exponent.

H.4 Trained affine readers: quality and charged execution

We distinguish predictive quality from charged execution savings. Token-local affine maps compose as $(A_X, b_X) \star (A_Y, b_Y) = (A_Y A_X, A_Y b_X + b_Y)$ (Smith et al., 2023, Appendix H, Eq. (34)), allowing reuse with a different incoming state. This requires compact closure and endpoint or sparse outputs, not arbitrary multilayer dependencies; floating-point reassociation is not bitwise exact.

Readers. On the official TRAIN/TEST splits of FordA (500×1), SelfRegulationSCP1 (896×6) and MotorImagery (3000×64), with fitting/validation split and normalization taken from TRAIN only, the width-128 reader has one complex diagonal affine layer,

$$a_t = \exp[-\operatorname{softplus}(\lambda) \odot (1 + .3\tanh(x_t W_g)) + i\phi], \qquad b_t = \tanh(x_t W_x + u)/\sqrt{n_{\text{time}}},$$

and a linear final-state readout, frozen before Build. References are one InceptionTime network, not the five-network ensemble (Ismail Fawaz et al., 2020), and MiniRocket with ridge (Dempster et al., 2021); neural models use three seeds under a common epoch, update and time budget. The original linear-readout reader is at chance on FordA and all models are near chance on MotorImagery (Table 10); those timing runs therefore test execution accounting. Edited inputs have no validated labels.

Table 11: FordA clean TEST accuracy (%), averaged over three seeds, with independently fitted readouts. All feature choices are reported; regularization uses TRAIN validation only. These are readout variants, not replacements for the original fits in Table 10.

Frozen state	Linear	Energy	Linear + energy
Fitted	49.55	87.53	86.87
Random initialization	48.84	89.32	88.38

Readout sensitivity. On FordA, logistic readouts use real/imaginary coordinates, energy $|z|^2$, or both from three frozen fitted states and three fresh random states. Features are prespecified, standardization uses fitting data, and TRAIN-validation cross-entropy selects regularization from $10^{-4}, \ldots, 10^{2}$. All variants use the same 1,320 TEST inputs (Table 11). Energy improves both fitted and random states: linear-readout failure does not imply uninformative states, but the control establishes no benefit from learned dynamics. Lower learning rates and removing early stopping do not achieve comparable validation quality within the same update budget.

The fitted-state energy readout preserves cached/cold predictions in 9,216 comparisons (three seeds, eight snapshots, three policies, two edit regimes, 64 requests; maximum logit difference 7.63×10^{-6}). This tests composition, not edited-input accuracy or timing gains; clean quality remains below the stronger references.

Requests and accounting. Banks use $(m, M, d, b, L) = (16, 4, 4, 27, 8)$ and 32 half-hotspot history locations. The seven policies in Table 12 share records, bytes and requests and are chosen before future edits. Independent edits reset inputs; cumulative edits do not. Net savings equal paired cold-minus-bank time over 128 requests, less one-time planning, Build and serialization; checks, restoration and transfers are charged. Cold references (CPU serial, GPU CUDA Graph) and TRAIN-median deadlines are fixed before TEST. Idle-device deterministic timing uses 32 snapshots/three seeds per task on GPU and four snapshots/one seed on CPU. Content fingerprints, rather than version descriptors, determine validity. Across 4,169,880 executions, including controls, no output exceeds tolerance (maximum state discrepancy 1.006×10^{-5}); technical repeats are not independent observations.

Findings. Table 12 shows net losses for all GPU policies and all cumulative-edit CPU policies; joint/history also loses in all five primary GPU request regimes. Independent CPU gains are policy- and task-dependent. SCP1 joint preparation costs .165 s versus about .003 s for fixed/mean banks. Every primary GPU bank has zero completion at the TRAIN-median deadline, which cold also often misses (SCP1 joint/history 1.281 ms versus graph cold .103 ms). Appendix H.5 controls host dispatch. Cumulative SCP1 joint completion falls from 1 to .5625 to 0 after 0, 2 and 16 edits; all tasks reach zero by 16, and reverting inputs restores completion. An eight-block hotspot shift favors joint/uniform on every task. Synthetic edits, unlabeled modified inputs and four CPU snapshots limit deployment interpretation; predictive quality and net joint-planning gains are not jointly established.

H.5 Execution costs, adapted aggregation, and capacity

We ask whether a faster backend reverses the losses above, how searched plans compare with adapted aggregation shapes at equal budgets, and how exact optima compare with the capacity bound (12).

Execution backend. Paired CUDA Graph and host-dispatch measurements hold banks, requests, policies and seeds fixed: RTX 4090 has 288 units/2,016 policy cells; RTX 5090 has 72/504. Graph capture accelerates both bank and cold execution (Table 13). On RTX 4090 their median latencies fall from 3.40 to .273 ms and from .447 to .135 ms, respectively (Figure 5). All 2,520 banks remain slower online and net negative; planning, Build and metadata contribute a median 70.8% of the net loss. Backend acceleration alone does not reverse the $m = 16$ result.

Table 12: All primary bank policies, keeping devices separate. Entries are paired cold-minus-bank execution seconds over 128 requests, minus planning, Build, and serialization once per bank, averaged over timing passes and then banks. Positive means time saved. The TRAIN-selected same-device cold reference has net zero. CPU uses four snapshots and one seed; GPU uses 32 snapshots and three seeds.

	CPU net seconds/bank		GPU net seconds/bank	
Policy	Independent	Cumulative	Independent	Cumulative
FordA				
Fixed/history	.091	−.016	−.143	−.356
Fixed/uniform	.065	−.017	−.195	−.359
Joint/history	−.087	−.182	−.329	−.553
Joint/uniform	−.081	−.184	−.346	−.526
Mean/no history	.102	−.016	−.128	−.358
Shuffled/history	−.098	−.186	−.345	−.525
Greedy	.074	−.022	−.179	−.361
SelfRegulationSCP1				
Fixed/history	.172	−.008	−.160	−.397
Fixed/uniform	.127	−.013	−.211	−.400
Joint/history	.013	−.172	−.347	−.592
Joint/uniform	−.007	−.174	−.368	−.565
Mean/no history	.188	−.009	−.137	−.397
Shuffled/history	−.034	−.180	−.363	−.568
Greedy	.147	−.013	−.193	−.403
MotorImagery				
Fixed/history	.708	−.019	−.208	−.508
Fixed/uniform	.494	−.006	−.262	−.515
Joint/history	.498	−.200	−.415	−.730
Joint/uniform	.419	−.210	−.485	−.705
Mean/no history	.724	−.031	−.199	−.515
Shuffled/history	.267	−.227	−.461	−.701
Greedy	.552	−.031	−.240	−.518

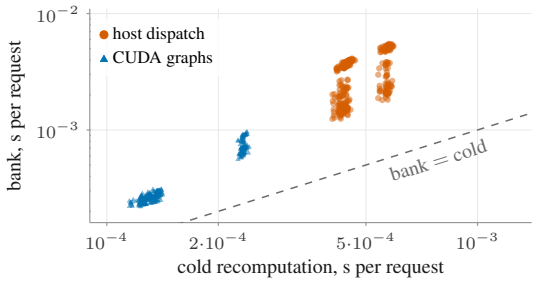

Figure 5: Paired bank and cold latency (orange: host-recursive dispatch; blue: captured CUDA graphs). Each RTX 4090 point compares a bank with its own backend's cold path; every sixth pair is shown. All 2,016 pairs per backend lie above the equality line, including after CUDA Graph capture. This result concerns the tested interface and workloads, not all bank implementations.

Installing records once. An RTX 4090 control retains records in CPU memory after one charged installation, including content scanning/hashing. Its 162 cells cross three datasets, seeds, snapshots and policies (fixed/history, joint/history, mean/no-history) with both edit regimes; each uses 128 requests and four randomized-order timing passes (PyTorch 2.11, CUDA 12.8). Performance-core execution is faster online in 63 cells, but net savings remain nonpositive in all cells, also on efficiency cores. These paired within-run results do not replace Table 12's absolute timings.

Table 13: Paired execution-backend control. Each entry is the median ratio of host-recursive dispatch to captured-graph latency for the same bank on the same requests within one process. Per-cell ratios span 2.07 to 15.60. The final column counts cells whose net sign changes; none does. Two devices are separate replications, not independent samples of one population.

Policy	Speedup: dispatch time / graph-replay time RTX 4090	RTX 5090	Sign changes (either device)
Fixed/history	5.80	8.92	0
Fixed/uniform	10.20	11.11	0
Joint/history	5.12	8.52	0
Joint/uniform	6.02	11.40	0
Mean/no history	6.47	9.38	0
Shuffled/history	7.82	11.83	0
Greedy	9.14	11.04	0
All cells	7.31	11.71	0 of 2,520

Where the time goes. To separate reuse from overhead, the same 162 cells are replayed with 1,024 requests on RTX 4090 and RTX 5090, timing installed records under two validity checks: content fingerprints and edited-block version bits (one bit per block, set when the block is edited; the edit path supplies them and their maintenance is not charged). States match cold execution (maximum error $7.5 \cdot 10^{-6}$), both checks accept the same records for every request, and fingerprint job counts match the exact recursion. Cold execution uses the faster of a sparse frontier and a captured whole-model graph (Table 14). Fingerprinting dominates on long inputs: with it, MotorImagery banks take 3.2–3.8 times the cold latency on both devices, whereas with version bits they are 29% faster than cold execution on RTX 4090 under independent edits, where fixed-tree and history-free banks repay preparation after 133–226 requests (all 18 by 256). Elsewhere banks stay slower even with version bits: per-request overhead around a width-128 reader (file access, frontier staging, kernels) exceeds the computation saved, cumulative MotorImagery edits invalidate most records, and on RTX 5090 whole-model graph replay answers in 0.08–0.16 ms; exact joint planning is repaid in no cell. The net losses of Table 12 thus reflect per-request overhead around a small reader and one-time planning rather than reuse itself; these are controlled replays, not deployment measurements.

Adapted aggregation shapes at fixed budgets. At identical (b, M, L) and requests, we compare at most 900 candidates of Theorem B.1's shape with flat recomputation and budget-feasible balanced and two-stacks prefix/suffix menus. The latter adapt structures from Reactive Aggregator (Tangwongsan et al., 2015), DABA (Tangwongsan et al., 2017) and related associative-window restructuring (Bhatotia et al., 2014); they are not the published online systems, since banks here remain frozen. Four exact associative operators, two noncommutative, give identical structural outcomes, counted by geometry.

The search wins/ties/loses in $2{,}457/353/0$ of 2,810 geometries without violating (12). At 81 representative budgets the empty bank is always a candidate, budgeted two-stacks in 13, and truncated balanced/symmetric two-stacks in none; only the first comparison is therefore guaranteed by menu inclusion. Excluding 71 fully protected ties and 1,180 zero-baseline cases, the protected-count ratio has median 1.047, 90th percentile 5.33 and 23.6% at least $2\times$. This excludes the largest relative margins. Flat recomputation needs a median $14.11\times$ cap to match searched protection (Figure 6). These are lower bounds on full-class headroom under binding storage; unrestricted FlatFAT is preferable. Table 15 separately covers 1,021 configurations.

Uniform-singleton capacity. Large instances use an exact length-indexed program rather than the position-indexed weighted planner of Appendix F. For remaining work t, let $A_t(\ell, k)$ be the largest timely singleton count inside a retained length-ℓ record with at most k additional records, and $B_t(p, \ell, k)$ the analogous maximum over p disjoint retained roots of total length ℓ; invalid states are $-\infty$ and $B_t(0, 0, k) = 0$. Then

$$B_t(p, \ell, k) = \max_{\substack{1 \le x \le \ell \\ 0 \le j \le k}} \{B_t(p-1, \ell - x, k - j) + A_t(x, j)\}.$$

Table 14: Where the time goes: installed banks with content-fingerprint or version-bit validity against the faster cold backend, over 1,024 requests per cell (the first 128 are those of the installation control). Break-even is the median request count at which fixed-tree and history-free banks repay planning, Build and installation (— when they never do); the last column counts those banks with positive net time after 1,024 requests. Exact joint planning is repaid in no cell. Cells: RTX 4090 SCP1 keeps the nine cells per regime timed against the faster backend, which start-up calibration mis-ranked in the others.

			Latency per request (ms)					
Dataset	Edits	Cells	Cold	Fingerprint	Version bits	Bits vs cold	Break-even	Net > 0 at 1,024
RTX 4090								
FordA	indep.	27	0.43	0.49	0.45	+4%	—	0/18
FordA	cumul.	27	0.42	0.55	0.51	+20%	—	0/18
SCP1	indep.	9	0.45	0.56	0.47	+3%	—	0/6
SCP1	cumul.	9	0.45	0.63	0.53	+19%	—	0/6
MotorImagery	indep.	27	0.89	2.93	0.63	−29%	166	18/18
MotorImagery	cumul.	27	0.90	3.36	1.01	+13%	—	0/18
RTX 5090								
FordA	indep.	27	0.08	0.14	0.14	+69%	—	0/18
FordA	cumul.	27	0.08	0.15	0.15	+81%	—	0/18
SCP1	indep.	27	0.10	0.17	0.15	+59%	—	0/18
SCP1	cumul.	27	0.10	0.18	0.17	+75%	—	0/18
MotorImagery	indep.	27	0.15	0.48	0.18	+22%	—	0/18
MotorImagery	cumul.	27	0.16	0.55	0.25	+61%	—	0/18

Table 15: Capacity attainment on the specified $1{,}021$-configuration grid. Entries are median / mean of protected count divided by G, followed by the number with zero protected positions. Menus optimize at each original cap; no column is an exact fixed-tree optimum.

Slack	Count	Searched menu	Balanced menu	Two-stacks menu
0	23	$1.000/1.000/0$	$.000/.402/13$	$1.000/.783/5$
$(0, m/4]$	178	$.563/.562/0$	$.400/.374/38$	$.461/.423/33$
$(m/4, m/2]$	178	$.648/.646/0$	$.267/.431/22$	$.359/.405/0$
$(m/2, m]$	340	$1.000/.867/0$	$1.000/.732/0$	$.547/.594/0$
$> m$	302	$1.000/1.000/0$	$1.000/.923/0$	$.781/.774/0$

A record with p maximal saved children and q uncovered inputs has weight $w = 2q + p - 1$, so $A_t(\ell, k)$ is the maximum of zero, $\ell \mathbf{1}\{2\ell - 1 \le t\}$, and

$$q + B_{t-w}(p, \ell - q, k - p), \qquad 1 \le p \le k, \quad 0 \le q \le \ell - p, \quad 1 \le w \le t.$$

Exclude $(p, q) = (1, 0)$. Child states have strictly smaller t, so the recursion is well founded, and each transition concatenates actual child intervals and uncovered inputs, so it emits a feasible laminar bank. The root maximizes $u + B_{L-a}(r, m - u, M - r)$ over $0 \le r \le M$, $0 \le u \le m$, $d \le a = 2u + r - 1 \le L$ ($r = 0$ requires $u = m$), or is zero if no state is feasible. This gives the exact uniform-singleton optimum in $O((\Delta + 1)m^2(M + 1)^3 + (\Delta + 1)^2 m(M + 1)^2)$ operations; it does not handle position weights or correlated masks.

Table 16 summarizes the exact evaluations, retaining failures outside the sufficient regime. Such failures do not establish necessity of individual assumptions; finite agreement neither proves the bounds nor improves their constants.

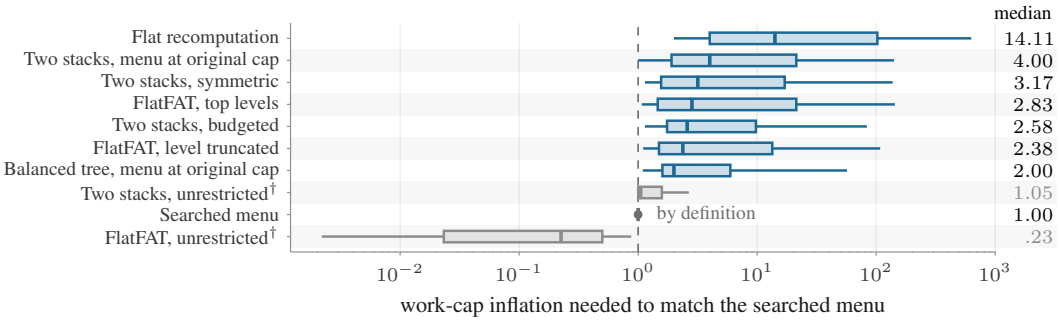

Figure 6: Baseline cap inflation. Ratios to the searched menu span 2,810 parameter geometries. Boxes show the interquartile range, whiskers the 10th–90th percentiles, and heavy rules the medians. Ratios below 1 favor the baseline. Daggered rows exceed the record budget and are not budget-matched. The searched menu's ratio is 1 by definition.

Table 16: Capacity and work-accounting evaluations. Counts describe deterministic checks, not independent statistical observations.

Evaluation	Result
Exact capacity: 227,092 configurations, $m \le 64{,}106$	170,241 in-regime; no bound violation; optimum/G .267–1, median .720
Independent comparisons	10,541 laminar-enumeration and 1,914 recursive-execution agreements
Outside the sufficient regime	30,543 lower-bound failures: mostly $M = 0$, otherwise $M = 1$ or $d < M$
Outside the upper-bound domain	One failure, $M = 0, L = n$
Constructive planner, 815 small instances	706 optimal; worst attainment .400 versus guaranteed $1/12$
Component identity, integer sum/max	1,050,990 executions over 95,991 configurations, $12 \le m \le 65{,}536$; exact agreement

H.6 Learning and execution sensitivity

These studies bound how the learning, deadline and execution results should be read; none measures worst-case sample complexity or net savings.

Learning a localized update law. A hotspot of width $w = \max\{2, \lfloor .1m \rfloor\}$ starts uniformly; inputs change independently at rate $\min\{.9, 2/w\}$ inside and $\min\{.9, .4/(m - w)\}$ outside. A heuristic learner estimates marginal rates from s complete masks and searches for the best plan under their product law, with all M blocks banked, $d = M - 1$ and $L = \lfloor M - 1 + 1.6(2m/M - 1) \rfloor$. Its gap to the known-law search (760 trials per point at $M = 4$; Figure 8) mixes estimation and search effects; it is not exact regret. The history length at which the mean gap first falls to .02 coincides for eight of nine input sizes, but Δ grows with m and (14) depends on M, d, Δ, so this does not show input-size-independent learning; some larger-budget settings miss .01 within 128 histories.

Finite-size deadline window. For $e_g(t) = \lfloor g/4 + t\sqrt{5g/32} \rfloor$, the maximum of the exact $|V_g(e_g(t)) - \Phi(t)/2|$ over the half-step grid $\mathcal{T}_g$ of t from -3 to 3 with $0 \le e_g(t) \le g$ decreases from .158 to .0031 over 21 sizes (Figure 9), with a large-g trend close to $g^{-1/2}$. This is neither a supremum over t nor a proved rate; since $V_g(e) = 1$ for $e \ge g$ while $\Phi(t)/2$ tends to $1/2$, no global uniform approximation of this form holds.

Dense-operator service cost. For ordered products of m dense $D \times D$ float32 matrices, serving from M stored block products while recomputing the middle block (unchanged inputs) falls below cold service time near $D = 128$–192 on an RTX 5090 (medians of three runs; Figure 7). Build,

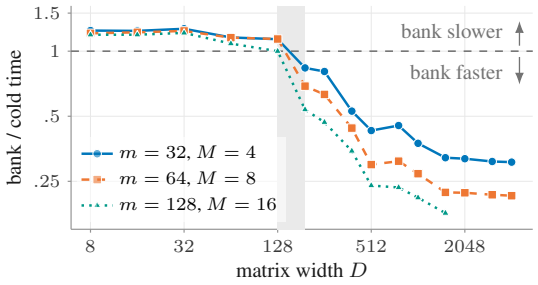

Figure 7: Dense-operator service latency; shading marks the crossover. Ratios are medians over three repeated runs on an RTX 5090, each using 20 synchronized measurements after warmup. The same unchanged matrix input is reused with forced middle-block recomputation. Setup costs are excluded and payload grows as $4MD^2$ bytes; the $m = 128$, $M = 16$ configuration was measured up to $D = 1536$. The service crossover is not an end-to-end deployment gain.

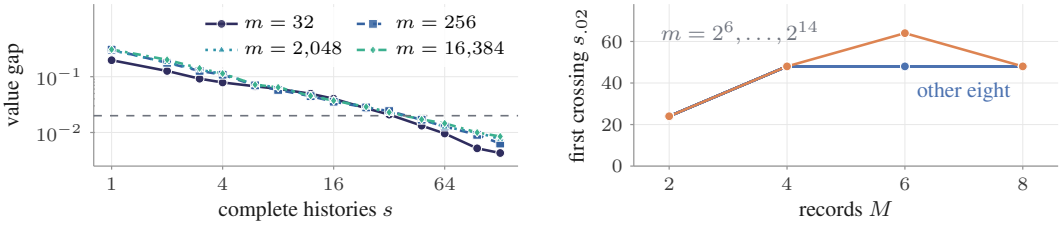

Figure 8: Heuristic learning across input sizes. Left: mean positive population-value gap between known-law and estimated-law searches at $M = 4$, with 760 trials per point; the dashed line is .02. Right: first tested history length with gap at most .02 for nine input sizes, eight of which coincide. Guided candidates depend on the fitted law, and Δ grows with m; this is not a pure statistical-regret comparison.

planning, validity checks and transfers are excluded, and the payload grows as $4MD^2$ bytes. This possible service-level benefit for expensive operators neither reverses the trained-reader net losses nor shows a gain at a fixed byte budget.

H.7 Exact evaluation of the mean-blindness family

We evaluate Theorem 2's family at its original (m, M, b, d) with exact rational probabilities, to quantify how much a mean-only decision loses to the informed optimum $1/2$ and how slack $L = d + e$ changes the loss.

Information premise and blind values. At $g = 2$, exhaustive enumeration gives 68,671 feasible banks; exactly $\binom{4}{2} = 6$ have positive zero-slack value, all of the normalized form. Each bank has the same mean under all six laws, although banks differ (33 distinct plan means): this illustrates the theorem's information premise, not equality of all preparation policies. A fixed selection has worst-group completion 2^{-g}; uniform randomization attains $V_g(0)$. For $2 \le g \le 500$ the exact values satisfy the proved $(3/4)^g$ bound (Table 17); at $g = 500$ the informed-to-blind ratio is $5.79 \cdot 10^{68}$, not the deterministic 2^{499}. This quantifies a proved formula, not independent evidence for its exponent.

Identification and slack. With s noiseless histories and $Q = 2^s$, H is uniquely identified with probability

$$(1 - Q^{-1})^g - (Q - 1)Q^{-g}.$$

Optimal decisions may precede identification (§H.1). At $g = 2$, four of ten positive-value cost functions at $e = 1$ lie outside the normalized shape: Lemma A.1 asserts a dominating replacement, not a classification. With preparation and storage unchanged, slack decides whether the missing information matters: by (11) at $g = 256$ (Table 18), blind completion is $3.72952 \cdot 10^{-8}$, 0.267313 and 0.499999851 for $e/g = 1/8$, $1/4$ and $3/8$, against the informed $1/2$. The transition at $e/g = 1/4$

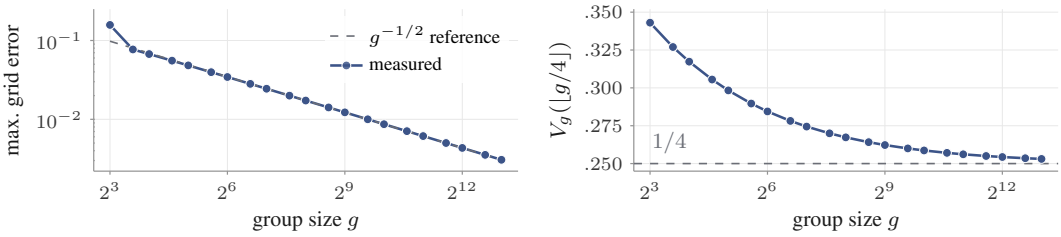

Figure 9: Finite-size deadline window. Left: maximum error on the finite half-step grid $\mathcal{T}_g$ for 21 sizes; the dashed $g^{-1/2}$ line is a visual reference anchored at $g = 256$, not a proved rate. Right: $V_g(\lfloor g/4 \rfloor)$ approaches $1/4$ from above.

Table 17: Exact zero-slack values, rounded for display. Optimum is $1/2$; ratio is randomized completion divided by $(3/4)^g$. The last columns compare unique-identification history requirements with the theorem's sufficient bound at confidence .95.

	Worst-H blind completion		Against the bound			Histories, $\delta = .05$	
g	deterministic	randomized	$(3/4)^g$	ratio	OPT/rand.	unique ID	bound
2	$2.50{\cdot}10^{-1}$	$2.92{\cdot}10^{-1}$	$5.63{\cdot}10^{-1}$	.5185	1.71	6	6
4	$6.25{\cdot}10^{-2}$	$1.44{\cdot}10^{-1}$	$3.16{\cdot}10^{-1}$	.4543	3.48	7	8
8	$3.91{\cdot}10^{-3}$	$4.03{\cdot}10^{-2}$	$1.00{\cdot}10^{-1}$	.4028	$1.24{\cdot}10^{1}$	8	10
16	$1.53{\cdot}10^{-5}$	$3.20{\cdot}10^{-3}$	$1.00{\cdot}10^{-2}$	.3192	$1.56{\cdot}10^{2}$	9	12
64	$5.42{\cdot}10^{-20}$	$8.00{\cdot}10^{-10}$	$1.01{\cdot}10^{-8}$	.0793	$6.25{\cdot}10^{8}$	11	16
256	$8.64{\cdot}10^{-78}$	$3.13{\cdot}10^{-36}$	$1.04{\cdot}10^{-32}$	$3.02{\cdot}10^{-4}$	$1.60{\cdot}10^{35}$	13	20
500	$3.05{\cdot}10^{-151}$	$8.63{\cdot}10^{-70}$	$3.39{\cdot}10^{-63}$	$2.54{\cdot}10^{-7}$	$5.79{\cdot}10^{68}$	14	22

has (25)'s $\sqrt{g}$ window, not a sharp finite-size threshold. These evaluations claim neither noise robustness, nonzero-fee persistence, arbitrary-law generality nor runtime savings.

H.8 Execution on a real transformer, and the two cost models

We compare tree-derived and KV-optimal checkpoint banks with full recomputation and flat prefix caches on GPT-2 Large, GPT-2 Medium and SmolLM2-135M. The FP32 single-GPU grid uses $n = 512, 1024, 2048$, $m = 8, 16, 32$ segments and four seeds: 1,606 configurations, 21,056 edit positions and measured KV-byte budgets. Random token contexts test execution, not language quality; maximum last-token logit error is $7.82 \cdot 10^{-5}$ with no top-1 disagreement.

At equal stored bytes, the checkpoint bank is dominated. In 1,430 exactly byte-matched comparisons, flat caching protects the same positions as KV-optimal banks, recomputes fewer tokens in 1,186 (82.9%) and never more, and is faster in 1,286 (89.9%; median ratio 1.256, IQR $[1.067, 1.405]$). Against 200 byte-matched tree-derived banks it recomputes less in all and is faster in 198. Table 19 reports cold-relative costs; 16.4% of tree-derived banks are slower than cold. A checkpoint at b_k already contains $[0, b_k)$: flat caching resumes anywhere in that prefix, whereas banks resume only at saved boundaries.

Why the tree capacity does not transfer. Fixed-size summaries and byte-priced causal prefixes are different interfaces: their predicted protected counts differ in $125/179$ budget cells, while measured protected counts do not separate the strategies above. Latency tracks recomputed segments (median within-unit Pearson .9455 over 76 units), but this does not validate the tree capacity for KV bytes. Context-2048 prediction coverage is four units from one checkpoint, and one incomplete six-unit batch is excluded, so no architecture or context-length invariance is inferred; nor do the results establish a fixed-fee hardware model or a general deployment speedup.

Table 18: Exact positive-slack values at unchanged preparation/storage budgets. The informed optimum is $1/2$. The middle column approaches $1/4$, rather than equaling it at finite size.

g	$V_g(g/8)$	$V_g(g/4)$	$V_g(3g/8)$
16	0.0862594	0.317309	0.470008
64	0.00331532	0.284428	0.497791
256	$3.72952 \cdot 10^{-8}$	0.267313	0.499999851

Table 19: Prefix-edit recovery over 1,606 configurations. Storage and work are median fractions of full recomputation; latency is the median per-configuration ratio. Slower counts median latencies exceeding cold. Protected-position counts do not distinguish the resumption strategies; byte-matched contrasts are reported in the text.

Strategy	Stored (/full)	Recomputed (/full)	Latency (/full)	Slower (of 1606)	Protected separates?
Full recomputation	0.000	1.000	1.000	—	—
Unbounded flat cache	1.000	0.500	0.659	0	no
Byte-matched flat cache	0.688	0.516	0.690	0	no
Bank, KV-optimal boundaries	0.625	0.750	0.847	129	no
Bank, tree-derived boundaries	0.500	1.000	0.967	263	no

H.9 Which joint events are worth counting

We test whether selecting a few budget-relevant joint events recovers full history's advantage over pairwise counters.

Selection and estimation. Write $h_\pi(J) = \mathbf{1}\{C_\pi(J) \le 9\}$ and $\phi_U(J) = \mathbf{1}\{J \cap U = \varnothing\}$. Subset-lattice inversion gives $h_\pi = \sum_{U \subseteq [8]} c_{\pi,U} \phi_U$ for each of the 2,236 feasible banks. Candidates are the noninterval sets U with $c_{\pi,U} \neq 0$ for some bank; the thirty Build-feasible intervals are always counted. From 32 pilot masks, form $\widehat{p}_p(J) = (n_J + 1/256)/33$, fit the maximum-entropy law q_I matching its interval-validity moments, rank banks by $\mathbb{E}_{\widehat{p}_p} h_\pi$ and keep the eight highest distinct completion functions $\mathcal{P}$, with best member π_0. The six selected events maximize

$$\text{score}(U) = \left|\mathbb{E}_{\widehat{p}_p} \phi_U - \mathbb{E}_{q_I} \phi_U\right| \max_{\pi \in \mathcal{P}} |c_{\pi,U} - c_{\pi_0,U}|.$$

This favors events that interval statistics misestimate and that distinguish promising preparations. The frozen events and the thirty interval moments are estimated on 1,024 independent masks with $(n_J + 1/256)/1025$, and completion is maximized under the fitted maximum-entropy law, with the tolerances of Appendix H.2. In this six-event comparison, controls receive all 1,056 masks: six random events from the same pool, and the pairwise model (eight first and 28 second moments); full history is the reference. The laws are both signs of the matched family of that appendix at $\lambda \in \{.25, .5, 1\}$, the independent midpoint law, and contiguous and dispersed burst mixtures, each with 32 fresh trials (seed 2026092041) and paired 95% whole-trial bootstrap intervals. The separate counter-budget sweep adds $\lambda = .125, .75$, with 32 trials per law. All decisions use 1,024 common estimation masks; selection also uses 32 disjoint pilot masks. Budget k counts joint events beyond the 30 interval counters.

Findings. Selected events outperform random counters and pairwise moments on the matched family but do not recover all full-history value (Figure 10, Table 20). In the counter-budget sweep, four selected events outperform twelve random ones; selection wins in 66 law–budget cells and loses in none. Null and burst gains are small or zero, with both arms within 0.36 points of full history at every counter budget. A same-size static geometric dictionary has lower/higher/equal regret than random selection in $2/5/2$ of nine prespecified cases; the ties are exact for every method.

Table 20: Joint-event selection, in completion points from all 256 future masks. Top: six-event comparison, averaged over both signs; all paired 95% intervals exclude zero. Bottom: separate counter-budget sweep, averaged over both signs at $\lambda \in \{.125, .25, .5, .75, 1\}$, excluding null and burst controls. The panels use different sampling protocols, specified above.

Six selected events, $(m, M, b, L) = (8, 3, 10, 9)$	$\lambda = .25$	.5	1
Gain over pairwise moments	1.03	3.28	7.70
Gain over six random events	0.66	1.44	3.34
Shortfall from full history	0.79	0.84	1.31

Events k	0	1	2	4	6	8	12
Selected: shortfall	4.43	2.64	1.84	1.18	0.67	0.52	0.18
Random: shortfall	4.43	4.06	3.74	2.99	2.44	1.78	1.37

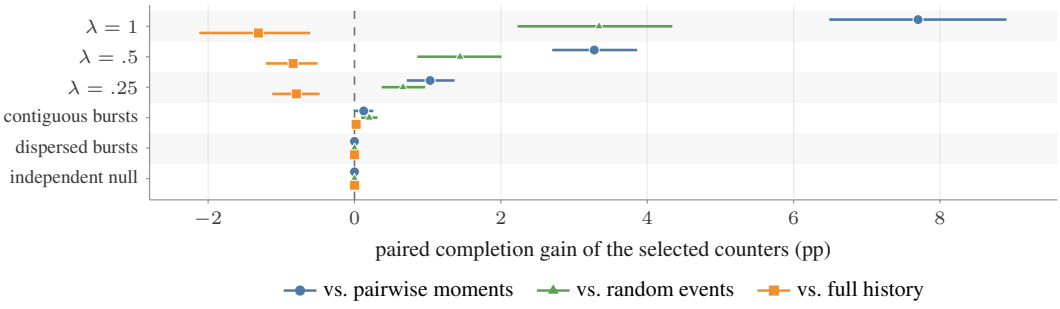

Figure 10: Completion from selected joint events. Paired gains over 32 fresh trials per fixed law, with descriptive 95% whole-trial bootstrap intervals. Thirty interval counters accompany six selected or random events; the pairwise model also uses 36 counters. Selection improves on these two controls in the matched family but remains below full history. Null and burst cases are retained, including zero gains. Counter counts exclude event identifiers and pilot processing.

A diagnostic that does not transfer. Six events selected on 128 training episodes are assessed against an interval closure refitted on the next 128. Their discrepancy summary predicts later full-minus-pairwise gains on the constructed family (rank correlation $+0.72$, versus -0.06 for summed covariances), but not on 24 repositories: lexical/directory correlations are $+0.07/+0.15$ versus $+0.27/+0.37$ for covariances. Correlations use projects because windows overlap; both planners agree in 173/240 windows. Counter selection thus lacks demonstrated predictive value on natural workloads. Counts exclude event identifiers, pilot processing and planning, and do not establish storage compression or scalability.

H.10 The capacity frontier, and what a bank costs to keep

We compare the exact frontier with executed instances, then measure how storage, data movement and natural update structure affect its practical interpretation.

The frontier is a frontier of execution. Four groups of four stable inputs have $\nu_i = 9$, $b_0 = 36$, $d_0 = 11$, $L_* = 15$ and six k-masks. Exact nested/flat-bank enumeration agrees with Corollary D.2 at all 225 points $b = 28, \ldots, 36$, $M = 4, \ldots, 8$, $e = 0, \ldots, 4$. At $e = 2$, nesting serves all six masks at $(b, M) = (34, 7), (32, 8)$, but at most five at $(32, 7)$ (Figure 11). Recursive execution on the actual 24-input, 47-node tree reproduces all eleven optimal local profiles, 3,750 regional decompositions and the 225-point grid. Across 35 families (32 with unequal region sizes), all 18,438 in-regime points agree for $g \le M \le 2g$, $0 \le \delta_b \le 2g$ and $e < \min_i(\nu_i - 2)$. Failures at $\delta_b = 2g + 1$ and $M = g - 1$ preclude unrestricted extension, without proving each assumption necessary. These correlated finite checks (three/four groups, three–six stable inputs per region) do not test large-g complexity.

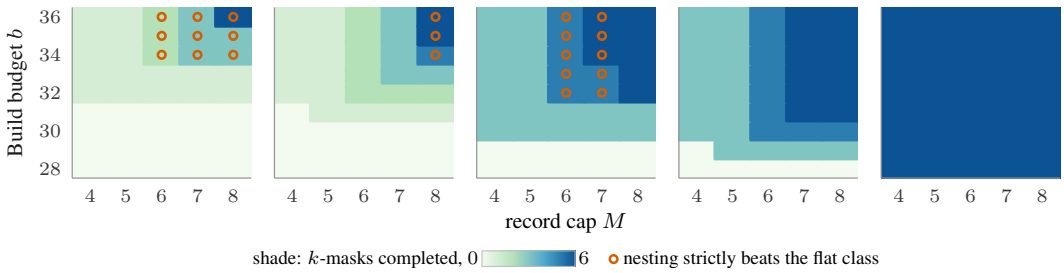

Figure 11: Timely k-masks served by one bank, exact over every legal tree and bank, for extra work $e = 0, \ldots, 4$ (left to right). Circles mark where nesting strictly beats flat storage; the advantage closes by $e = 3$, so extra work and nested storage are exact substitutes.

Table 21: Lifecycle costs. Rebuilding under persistent edits extends, rather than instantiates, the immutable-bank theory. Resident reverted-edit trends use up to 128 rounds.

Execution regime	Measured result
Resident: realized share of product-count gain	70% at $D = 64$–512; 77% at 1024; 97% at 2048
$D = 1024$: speedup at $(m, M) = (64, 4)/(1024, 32)$	$2.63\times$ / $6.36\times$
$D = 2048$: host restoration / computation avoided	2.51 s / .74 s
Single-round reuse	Best ratio $.68\times$
Rebuild when half the records become invalid	3/96 cells above $1\times$; best $1.29\times$

More storage can increase the value of information. For the fixed matched pair of Appendix H.2, at deadline 8 and free reuse, common-statistic minimax regret is

$$0.00,\ 0.00,\ 3.41,\ 4.73,\ 5.80,\ 6.86,\ 7.54,\ 8.61 \text{ points} \quad \text{for} \quad M = 1, \ldots, 8.$$

Increasing storage beyond two records benefits the informed comparator more, even though the laws are unchanged and no feasible bank is lost. In contrast, at eight records, deadlines of at least ten eliminate regret. Values through four records match rational certificates to $4.3 \cdot 10^{-16}$; larger values use uncertified double-precision LPs. The heaviest fee schedule eliminates most of the gap. This is a constructed storage–information interaction, not a general monotonicity claim.

Node counts are exact; time is not. Balanced-tree products use $D = 64$–2048, 64–1024 inputs and 4–32 records over up to 128 reverted/persistent-edit rounds on one idle GPU. Records are resident or round-tripped to host memory; version checks are charged and cold timing is paired within rounds. All 832 configurations match the predicted node count and cold root bitwise. An unpaired concurrent-load measurement is excluded because its apparent ceiling violation does not compare the same device conditions.

Figure 12 and Table 21 separate operation savings from latency. The prediction counts products, not all $2m - 1$ nodes: leaves multiply nothing. At $D = 1024$, speedup rises with record cap and with input count up to 512, but not uniformly beyond. Host transfers erase gains at the three largest dimensions. Repeated-stream net-time ratios have median interquartile spread $0.04\times$. These trends concern one device/kernel/tree shape, not equal hardware cost per node.

Extended revision replay. Twenty-four Python projects follow Appendix H.2's extraction: up to 4,096 recent first-parent transitions, eight frozen lexical regions and one/four-transition episodes. The original three projects are included; evaluation here uses fixed windows. Directory-boundary groups provide a second representation. Windows train on 256 episodes starting at $0, 512, 1024$ and test on the next 256; 41/144 per assignment are too short. Intervals use the widest moving-block estimate over lengths 4, 16 and 64, without multiplicity correction.

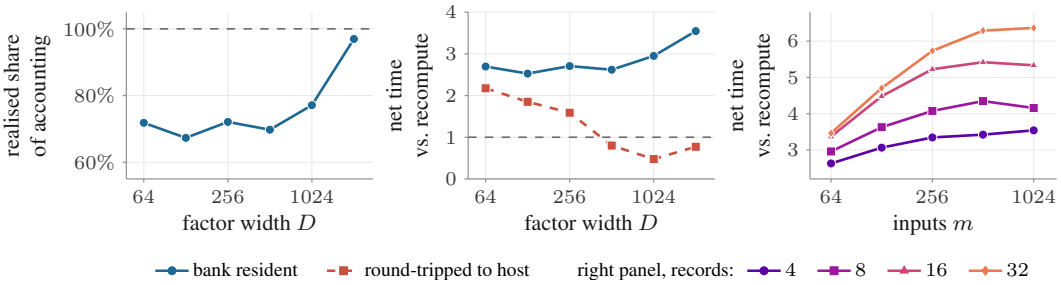

Figure 12: Measured execution over 128 rounds of reverted edits on an idle GPU. Left: realized share of the product-count prediction. Middle: net-time ratio, bank resident or round-tripped to host memory; the round trip inverts the gain at the three largest dimensions. Right: speedup at $D = 1024$ by tree size and record cap.

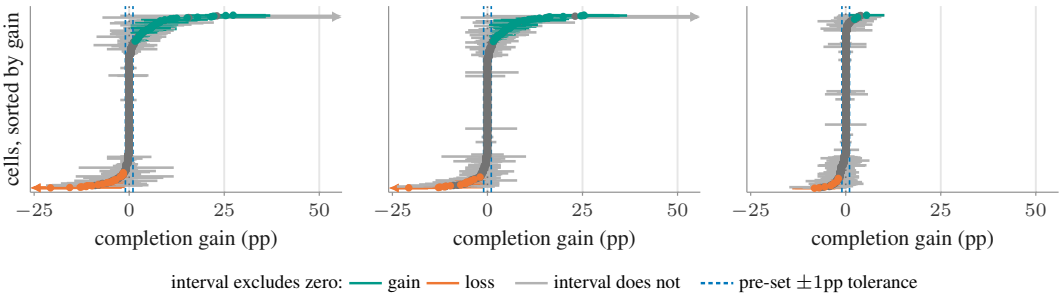

Figure 13: Completion gains on revision traces. Left: directory groups; center: lexical path quantiles; right: an independent-input control. Each panel contains 515 evaluated window–cap cells from the 24-project extension. Gains compare full-history preparation with independence-based completion decisions. Intervals use the widest moving-block estimate over lengths 4, 16 and 64, without multiplicity correction; overlapping windows are not independent evidence. Arrowheads mark intervals cut at the axis limits. Most gains are zero. Neither mapping consistently requires full-history estimation.

Across 1,030 window–cap cells, joint history gains .55 points over independence-based completion (243 wins, 194 losses; 15/24 projects improve). Gains are .66 and .32 for one- and four-transition episodes, and .58 and .52 for lexical and directory regions. Independent redraws at TRAIN rates instead give $-.20$ (46 wins, 97 losses). Most cells tie (Figure 13); intervals exclude zero in 142 real cells (106 positive, 36 negative) and 10/515 control cells (3 positive, 7 negative). The prespecified ± 1-point equivalence criterion is not met. In contrast, the gain over mean-work training is 7.6 points and over pairwise completion only .07: the objective dominates the incremental-statistics effect. Overlapping windows are not independent evidence.

Resolution sensitivity at a fixed history budget. Twelve projects with at least 256 anchor paths use 8/16/32 lexical regions, all active at the finest resolution. Fixed-tree banks have at most three records (63/575/4,991 candidates on 15/31/63 nodes), fit 256 episodes and test on the rest at four caps spanning 0.5–0.8 cold work. Across 96 windows per resolution, joint-minus-independent completion is $+.06, -.01, -.36$ points, versus $-.31$ on the finest independence control. Hindsight regret rises from .80 to 4.23 for joint and .85 to 3.87 for independent fitting; mean-work regret is $6.8, 12.7, 15.6$. Resolution changes geometry, class and cap simultaneously, so it isolates no estimation mechanism. The broader storage/cap grid in Appendix H.11 gives positive repository contrasts from 16 regions onward. Neither slice establishes the absence of higher-order dependence, and decision learning need not estimate all 2^{32} mask probabilities (Theorem 3, under i.i.d. histories).

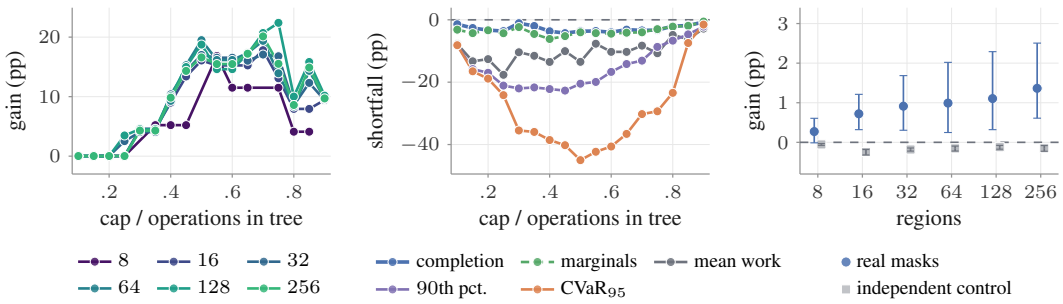

Figure 14: Which objective finishes on time. Left: gain of completion training over exact mean-work training against the cap as a share of the tree, repositories pooled, two records, one line per resolution. Middle: shortfall from the best bank in hindsight of completion, marginal-rate completion, mean work, the 90th percentile and CVaR_{95} at thirty-two regions and four records. Right: joint history against per-region marginals on the real masks and on independence controls, pooled over storage levels and caps between a fifth and four fifths of the tree; bars are history-bootstrap 95% intervals, for the controls at most .08 points on either side.

H.11 Training objectives on natural histories

We separate the effects of the training objective, predictive statistics and tree choice, then test sensitivity to history length, storage, deadlines and operation costs.

Setup. The prespecified collections comprise 24 Python projects (Appendix H.10), 32 ML systems repositories extracted identically, and 24 Wikipedia articles (ten level-one vital and fourteen computing articles, partitioned by section over their most recent 4,097 revisions). Learners fit the first 256 episodes and test on all later episodes. Independence controls redraw regions at their own training rates. Methods follow Table 6, with CVaR_{90} added and tail-objective ties broken by mean work. Its primary protocol searches 2,236 banks over all 429 ordered trees. The width study instead uses a balanced tree, no Build limit and up to M records, with M as large as exhaustive search permits at each resolution from 8 to 256 regions. A history is eligible only when its anchor contains at least one path or section per region. Costs are node counts; all storage levels and prespecified caps from 0.2 to 0.8 of cold work are included, even when every bank has zero held-out completion. Unless stated otherwise, intervals are 95% history-bootstrap intervals.

Objective, information and tree choice. Table 23 separates these effects. Completion training improves on mean-work training even under independence and with one record: means then suffice to evaluate every cap (Theorem 1), but minimizing the mean remains a different decision. Gains persist across five additional resource budgets, directory-based regions and a later training window. They peak near caps of 0.5–0.75 of cold work and vanish near full recomputation (Figure 14). At 32 regions, among cells with positive held-out completion, completion banks use 69% of cold work versus 64% for mean-work banks. About a quarter of these cells contain banks within 2% of the minimum mean whose completion differs by at least ten points.

Cap-aware surrogates. We compare completion with buffered exceedance $\inf_{a\geq 0} \mathbb{E}[(a(C-L)+1)_+]$ (Mafusalov & Uryasev, 2018; Nemirovski & Shapiro, 2007), expected overshoot $\mathbb{E}[(C-L)_+]$, and CVaR with level .5–.95 selected by chronological TRAIN validation. The first two close part of the objective gap; validated CVaR does not (its gains over mean work are $0.0, -1.4, -1.9$ points). Validation rarely distinguishes CVaR levels. Fixed-level tails also underperform mean work (Table 22): among positive-completion cells, CVaR_{95} executes 77–78% of the tree at 32 regions. It targets the most heavily changed episodes, which rarely complete, rather than those near the cap. Fresh training draws from the ordinary laws in Table 6 reproduce this ordering.

Incremental value of joint statistics. Across $8, 16, 32, 64, 128, 256$ regions, joint-minus-marginal completion on the 56 repositories is $0.21, 0.65, 0.80, 0.83, 0.84, 1.06$ points; intervals exclude zero from 16 regions onward. Independence controls instead give -0.05 to -0.23 points. Wikipedia's

Table 22: Held-out completion gain of completion training over each rival, in points: mean over histories of each history's mean over storage levels and caps between a fifth and four fifths of the tree, with history-bootstrap 95% intervals. The last column repeats the contrast against per-region marginals on the independence control. The two tail columns omit their intervals, whose lower bounds are all at least 3.2 points. Zero-completion cells are included.

		completion training minus				
Regions	Hist.	mean work	CVaR_{95}	90th pct.	marginals	marginals, control
Python projects						
8	24	$+2.8$ [2.4, 3.2]	$+3.8$	$+3.6$	-0.05 [-0.21, 0.10]	-0.06 [-0.08, -0.03]
16	20	$+7.4$ [6.5, 8.5]	$+21.6$	$+12.0$	$+0.48$ [0.06, 0.90]	-0.22 [-0.33, -0.13]
32	16	$+8.3$ [7.3, 9.7]	$+26.8$	$+16.4$	$+0.54$ [0.18, 0.92]	-0.20 [-0.25, -0.15]
64	13	$+8.6$ [7.6, 9.9]	$+25.8$	$+17.9$	$+0.49$ [0.12, 0.93]	-0.10 [-0.17, -0.04]
128	12	$+7.7$ [6.6, 9.3]	$+22.8$	$+17.6$	$+0.44$ [-0.08, 0.95]	-0.07 [-0.10, -0.05]
256	12	$+7.9$ [6.6, 9.7]	$+22.8$	$+17.6$	$+0.76$ [0.33, 1.20]	-0.08 [-0.15, -0.03]
ML systems						
8	32	$+3.1$ [2.7, 3.6]	$+4.8$	$+4.1$	$+0.41$ [0.03, 0.83]	-0.04 [-0.08, -0.01]
16	27	$+7.4$ [6.2, 8.9]	$+23.2$	$+12.5$	$+0.78$ [0.24, 1.47]	-0.24 [-0.34, -0.17]
32	24	$+8.5$ [7.0, 10.4]	$+27.7$	$+17.0$	$+0.97$ [0.15, 2.00]	-0.15 [-0.20, -0.11]
64	21	$+9.2$ [7.3, 11.7]	$+27.7$	$+19.8$	$+1.04$ [0.05, 2.33]	-0.15 [-0.23, -0.08]
128	18	$+7.8$ [6.1, 10.2]	$+22.8$	$+17.7$	$+1.12$ [0.26, 2.48]	-0.11 [-0.17, -0.06]
256	18	$+8.0$ [6.2, 10.5]	$+22.7$	$+17.9$	$+1.26$ [0.39, 2.60]	-0.14 [-0.23, -0.07]
Wikipedia						
8	24	$+2.9$ [2.3, 3.4]	$+6.1$	$+5.5$	-0.03 [-0.08, 0.02]	-0.07 [-0.10, -0.04]
16	14	$+6.9$ [5.8, 7.9]	$+25.6$	$+15.0$	$+0.21$ [-0.05, 0.46]	-0.12 [-0.16, -0.09]
32	7	$+7.7$ [6.7, 8.7]	$+30.2$	$+17.1$	$+0.44$ [0.16, 0.71]	-0.11 [-0.23, -0.01]

contrast excludes zero only at 32 regions. These effects are much smaller than the objective and tree-choice effects. Under the primary protocol, interval, pairwise and marginal models are all within two points of full history; pairwise moments significantly outperform it on ML repositories (Table 23). The negative three-record, 0.5–0.8-cap slice in Appendix H.10 is retained: the information gain depends on the resource range.

How much history. With TEST fixed after episode 1,024 on the 56 repositories, 16–64 training episodes bring completion within 1.5–4 points of the hindsight optimum; larger banks benefit most from additional history. On a later TEST window, improvement saturates at 256 episodes and the longest histories slightly degrade, whereas mean-work training remains 5–20 points below the optimum at every length. Rolling 256-episode training and TEST windows over all 104 histories (including replication) yield repository objective gains of 5.0–8.7, 8.9–12.7 and 10.5–12.7 points at 8,

Table 23: Contrasts under the primary protocol of Table 6, averaging its one- and four-update rows. Brackets give 95% history-bootstrap intervals. Intervals for the tight-Build and three cap-aware surrogate contrasts also exclude zero.

Completion gain (points)	Python	ML systems	Wikipedia
Over mean work	$8.4\ [6.8, 10.1]$	$8.0\ [6.7, 9.3]$	$7.4\ [5.7, 9.1]$
Over mean work, $(M, b) = (3, 6)$	2.0	2.4	1.4
Over mean work, independence control	7.1	7.7	7.5
Over CVaR_{95}	26.6	25.5	28.9
Over buffered exceedance	6.1	6.8	5.7
Over expected overshoot	3.7	4.1	2.9
Over validation-selected CVaR	8.4	9.4	9.3
Over pairwise-moment completion	$0.2\ [-0.1, 0.6]$	$-0.4\ [-0.9, -0.03]$	$-0.1\ [-0.2, 0.1]$
Over balanced-tree completion	$7.7\ [4.4, 11.3]$	$4.5\ [2.7, 6.5]$	$3.9\ [1.7, 6.3]$

Table 24: Held-out completion gain of completion training over exact mean-work training under unequal operation costs known to both learners, in points: mean over histories of each history's mean over one to three records, caps at $0.2, 0.3, \ldots, 0.8$ of cold work and ten cost draws, with history-bootstrap 95% intervals. Independent: every operation's cost log-normal with mean one and the stated coefficient of variation; merge: leaf costs so drawn, each combination costing the leaf cost beneath it. Averaged over storage levels, no history favors mean-work training in any cell.

Operation costs	8 regions	16 regions	32 regions
Unit	5.1 [4.6, 5.7]	8.4 [7.5, 9.2]	8.5 [7.6, 9.4]
Combinations only	3.9 [3.4, 4.6]	7.4 [6.8, 8.1]	8.4 [7.5, 9.3]
Independent, CV 0.5	7.2 [6.6, 8.0]	9.7 [8.9, 10.5]	9.6 [8.8, 10.5]
Independent, CV 1.3	7.0 [6.5, 7.5]	10.4 [9.7, 11.1]	10.4 [9.7, 11.1]
Independent, CV 2.9	5.9 [5.5, 6.3]	10.2 [9.6, 10.7]	11.9 [11.1, 12.6]
Merge, CV 0.5	4.5 [4.1, 5.0]	6.6 [6.1, 7.1]	7.1 [6.4, 7.8]
Merge, CV 1.3	4.7 [4.4, 5.2]	7.3 [6.7, 7.9]	8.0 [7.4, 8.7]
Merge, CV 2.9	4.8 [4.5, 5.2]	7.8 [7.3, 8.3]	9.0 [8.3, 9.7]
Histories	80	61	47

16 and 32 regions. Joint information adds 0.1–1.2; all intervals exclude zero except the eight-region Wikipedia joint contrast.

Larger banks. For 64–256 regions, exact mean-work dynamic programming is compared with completion local search (record addition, removal and swaps). Gains at $M = 3, 4, 8, 16$ are respectively 11.5–11.9, 7.3–8.0, 1.5–2.1 and 0.4–0.7 points, all intervals excluding zero. Wikipedia at 16–32 regions gains 8.9–12.3 with three or four records. The gap narrows as both objectives reach 93–95% completion at 16 records. At 64 regions, local search attains the exact training optimum in 79% of 476 three-record cells; for four or eight records it trails exact planning by 0.1–0.4 held-out points in the 193 of 204 cells with a certified mixed-integer optimum.

Misjudged caps and unknown costs. Deadline error is asymmetric. At 32 regions, training one to three operations below the enforced cap retains a 5–11-point gain over mean work; training two or more operations above it reverses the gain. At eight regions, a two-operation error in either direction erases it (one operation is 7% of cold work). Averaging completion over a ± 2-operation cap band retains 68%, 80% and 88% of the nominal gain at 8, 16 and 32 regions (104, 85 and 71 histories) and beats mean work throughout that error range.

Table 24 evaluates synthetic unequal costs on natural masks. Known costs preserve the objective gain, with 2.4–5.9% additional expected cold work. No history favors mean work after averaging

Table 25: Cost misspecification on the same natural histories and cost draws as Table 24. Cap-band training averages completion within approximately 0.1 cold work of the nominal cap. Pooled merge-cost intervals cover the eighty primary histories.

Actual costs; learners assume unit costs	Completion minus mean work (points)
Independent, CV 0.5	$+3.8$, $+4.6$ at 16, 32 regions
Merge; independent CV 2.9	-7.6 to -4.0; -2.9 to -1.8 at all resolutions
Combinations only; independent CV 1.3	-3.2; -1.6 at eight regions
Cap-band objective, merge costs	-1.0 to $+0.2$ pooled; all intervals include zero
Cap-band objective, merge costs by collection	Losses up to 1.8 (ML, 32 regions), 2.2 (Wikipedia), 2.7 (replication)
Cap-band objective, other costs	$+1.1$ to $+6.2$; tie at eight regions, CV 2.9

Table 26: GPU completion gains over mean-work training. Eight-region time and operation gains differ by less than 0.01 points. Under calibration-only rectangular deadlines, only the three-record Wikipedia interval includes zero.

Square products: completion gain (points)	Python	ML systems	Wikipedia
8 regions, 3 records, $D = 2048$, time	4.6	4.9	2.7
32 regions, 4 records, $D = 1024$, time	8.7	6.3	4.7
32 regions, 4 records, $D = 1024$, operations	10.1	8.0	6.0
Rectangular products, 80 histories	2 records	3 records	
Time gain [95%]	7.0 [5.3,8.6]	4.2 [3.1,5.4]	
Histories won / lost	66 / 11	65 / 10	
Calibration-only deadline, time gain	5.7	2.9	

storage levels, although up to five do at an individual level. Incorrectly assuming unit costs can reverse the ranking (Table 25); cost knowledge improves completion by 4.4–20.2 points in every history. A cap band of about ± 0.1 cold work reduces this sensitivity, but sacrifices 1.6–2.6 points when unit costs are correct.

Deadlines in wall-clock time. Held-out episodes are replayed as ordered matrix products on one GPU, with device-resident records and stale-record rebuilding outside the timed request. For eight regions, three records and 2048×2048 products, time is nearly affine in executed products (median $R^2 = 0.996$–0.998). Deadlines of $L + \frac{1}{2}$ product times agree with the five operation caps of Table 6 on at least 99.8% of requests on average. Smaller products at 32 regions weaken this correspondence (Table 26); larger products restore it on the first held-out repository episodes. Four-commit gains are smaller: 0.8 and 1.9 points on the repository collections.

Rectangular products have median operation-time spread $12\times$. Banks use previously measured costs; deadlines use either each history's fitted measured/predicted-time ratio or calibration alone. Both retain positive average gains, but these resident-service comparisons exclude host restoration and do not establish net lifecycle savings (Figure 12).

Replication. The unchanged protocol on 24 additional repositories, mainly in other languages and chosen after the primary results, gives gains of 7.2 points over mean work ($[5.8, 8.4]$), 19.6 over CVaR_{95} and 3.0 from tree choice ($[1.9, 4.4]$); pairwise and marginal models remain within 0.3 of full history. At 8, 16 and 32 regions, objective gains are 2.5, 6.0 and 7.4, and joint-information gains 0.05, 0.44 and 0.60. Larger banks, known unequal costs, rectangular GPU replay, history-length effects, asymmetric cap errors and the reversal under unknown merge costs reproduce the primary patterns. All gain intervals exclude zero except the eight-region joint contrast.

Table 27: Exposure of mean-only decisions around natural laws, in points, averaged over caps 5–13 and histories. The columns give $R_\epsilon(p)$ over the whole mean class and within total variation .10, .05 and .02 of p; the share of histories with at least one point; and, as a contrast, the loss of the interval-closure decision under p itself. h is the number of updates per episode. Every exposure entry is a certified lower bound within 10^{-5} of the exact value; closure losses are exact.

Collection	Laws	h	whole class	≥ 1 pt	TV .10	TV .05	≥ 1 pt	TV .02	closure loss
Python projects	natural	1	5.34	83%	2.17	1.17	67%	0.42	0.49
		4	7.46	92%	2.84	1.29	62%	0.39	0.16
	control	1	6.71	79%	2.50	1.23	62%	0.37	0.02
		4	7.66	92%	2.78	1.30	71%	0.38	0.06
ML systems	natural	1	5.94	84%	2.41	1.30	78%	0.46	0.36
		4	8.31	91%	2.92	1.41	69%	0.40	0.19
	control	1	7.71	94%	3.07	1.51	78%	0.50	0.01
		4	9.30	94%	3.19	1.49	75%	0.45	0.02
Wikipedia	natural	1	1.94	54%	1.31	0.84	46%	0.42	0.10
		4	4.50	79%	2.05	1.07	58%	0.36	0.17
	control	1	3.22	83%	1.79	1.07	58%	0.44	0.01
		4	6.64	83%	2.37	1.12	54%	0.32	0.05
Replication	natural	1	5.94	88%	2.45	1.29	79%	0.48	0.45
		4	8.61	92%	3.13	1.50	71%	0.43	0.26
	control	1	6.46	88%	2.63	1.35	58%	0.47	0.01
		4	9.05	92%	3.13	1.40	79%	0.39	0.06

Scope. These dependent trajectories and path/section regions do not instantiate the i.i.d. theory or a dependency graph. Bootstrap intervals describe the selected histories, not unobserved repositories or articles; collections are neither random nor mutually independent.

H.12 How far exact means leave the choice open around natural laws

Theorem 2 and the pair of Section 6 are constructed. This study asks, for the information claim, how much room exact means leave around the natural laws themselves. For each natural history, episode length and cap $L \in \{5, 7, 9, 11, 13\}$ at $(m, M, b) = (8, 3, 10)$, let p be the empirical law of all episodes over the 256 masks of the eight regions of Table 6. Let $\mathcal{Q}(p)$ contain every law with the same 30 Build-feasible interval validities as p, hence by Theorem 1 the same mean work for every feasible plan, and let $\mathcal{Q}_\epsilon(p)$ also require total-variation distance at most ϵ from p. The exposure

$$R_\epsilon(p) = \min_w \max_{q \in \mathcal{Q}_\epsilon(p)} \big[\mathrm{OPT}_q - F_q(w)\big],$$

with w ranging over randomized banks, is the exact minimax regret of any procedure that receives the feasible means and, for finite ϵ, knows that the law lies within ϵ of p; it is a worst case over a class, not a realized loss under p. Cutting planes against adversary laws in $\mathcal{Q}_\epsilon(p)$, with exact best responses, give certified lower and upper bounds that agree within 10^{-5} on every cell. For scale, the total-variation distance between an empirical law and a multinomial resample of the same size has median .053 over the natural laws. The independent-input controls of Appendix H.11 are evaluated identically.

Findings. Over the 160 primary cells (eighty histories, two episode lengths; Table 27), the whole-class exposure averages 5.7 points (cell-bootstrap interval $[5.1, 6.4]$); it is at least one point in 81% of cells and at least two in 74%. Within total variation .10 it averages 2.3; within .05, about the median sampling distance, 1.2 ($[1.1, 1.3]$), at least one point in 64% of cells but at least two in only 9%; within .02, .4, with no cell reaching one point. The interval-closure decision itself loses only .25 points under the empirical laws (at most 2.6). Independent-input controls are at least as exposed: paired by history and episode length, natural minus control over the whole class is -0.8, -1.4, -1.7 and -0.5 points for the Python, machine-learning, Wikipedia and replication sets, every bootstrap

interval over these pairs excluding zero; within total variation .05 the differences lie between -0.15 and $+0.02$, and only the machine-learning interval ($[-0.25, -0.05]$) excludes zero. The exposure is thus a property of mean-only information at these budgets, not of the dependence in these histories: exact means rarely pin down the preparation, yet within the sampling uncertainty of these laws the room they leave is about one point. The study covers eight regions, one budget and empirical laws as centers; it measures how much room means leave, not how often natural updates realize a loss.

I DEADLINE-AWARE RESIDENT STATE IN LLM INFERENCE

Prefix KV caches (Zheng et al., 2024; Qin et al., 2025), retrieved-chunk caches with partial recomputation (Yao et al., 2025; Hu et al., 2025), and offloaded mixture-of-experts decoders (Eliseev & Mazur, 2023) offer different reuse interfaces. We compare deadline-aware resident-state choices under equal budgets in these three settings. A service-level objective (SLO) specifies a completion deadline; TTFT denotes time to first token.

Proposition I.1 (The reuse interface decides the sufficient statistics)**.** *Let $W(R, S)$ be the prefill work of request R given resident states S, and completion $\mathbb{P}(W(R, S) \le L)$.*

1. Prefix interface. *Requests are root-to-node paths of a rooted tree, resident sets are ancestor-closed, and $W(R, S) = |R| - \max\{\mathrm{depth}(v) : v \in R \cap S\}$. The node-visit probabilities $\pi(v) = \mathbb{P}(v \in R)$ determine the law of $W(R, S)$ for every ancestor-closed S, hence mean work and completion at every cap.*

2. Chunk interface. *Requests are sets of chunks with lengths ℓ_d, and $W(R, S) = q + \sum_{d \in R \setminus S} \ell_d + \rho \sum_{d \in R \cap S} \ell_d$ with $0 \le \rho < 1$. Mean work depends on the law of R only through $\mathbb{P}(d \in R)$, but completion does not: there are laws with identical chunk frequencies under which every frequency-only, possibly randomized, cache has worst-case completion at most $1/4$, while the informed cache completes $1/2$.*

Proof. (1) A path is identified by its endpoint, and $\mathbb{P}(R \text{ ends at } v) = \pi(v) - \sum_{c \text{ child of } v} \pi(c)$, so the π determine the law of R and hence of $W(R, S)$. (2) Linearity gives $\mathbb{E}W = \mathbb{E}q + \sum_d \mathbb{P}(d \in R)\, \ell_d\, [1 - (1 - \rho)\mathbf{1}\{d \in S\}]$. For the separation take four chunks of equal length ℓ, capacity two chunks, constant q, and L with $q + 2\rho\ell \le L < q + (1 + \rho)\ell$, so a two-chunk request completes only if both chunks are resident. Law A draws $\{a, b\}$ or $\{c, d\}$ and law B draws $\{a, c\}$ or $\{b, d\}$, each with probability $1/2$; every chunk has frequency $1/2$ under both. A two-chunk cache completes $1/2$ under at most one law and 0 under the other, so a randomized cache that cannot tell the laws apart has worst-case completion at most $1/4$, attained by mixing $\{a, b\}$ and $\{a, c\}$ equally; knowing the law attains $1/2$. □

Part 1 determines prefix work distributions from all node-visit probabilities. Part 2 gives an analogous chunk-information distinction; with $\rho = 0$ and unit lengths, it describes expert residency. Neither is an instantiation of the ordered-tree learning bound. The experiments compare equal-budget plans with a shared completion optimizer (Figure 15).

Prefix caches: an objective effect. Mooncake conversation and agent traces (Qin et al., 2025) contain 12,031 and 23,608 requests over one hour each. Prefixes use 512-token blocks, with a cap of $B \in \{2, 4, 8, 16\}$ nonresident blocks. Every 60 s, a host pool of $M \in \{1{,}000, 2{,}000, 4{,}000\}$ blocks pins a prefix-closed set of fM blocks, $f \in \{.5, 1\}$; the remainder follows LRU. Hit and SLO staging maximize, respectively, predicted resident blocks and completed requests using the same prior-request-only continuation predictor. Baselines are LRU and T-LRU (Zhang et al., 2025).

Against hit staging, SLO staging raises attainment (Table 28) and lowers block-hit rates in all 24 primary $f = 1$ configurations. A separate 144-cell grid crosses both traces with $M/1{,}000 \in \{1, 2, 4, 8\}$, $B \in \{4, 8, 16\}$, staging intervals $15, 120, 300$ s and lookbacks $600, 3{,}600$ s, at $f = 1$. SLO staging beats hit staging in all cells and LRU in 137; six losses and one tie use 300 s and $M = 8{,}000$. RTX 4090-calibrated TTFT includes prefix attention but excludes pool transfer; 2–5% timing error changes gains by at most 0.3 points. A mean-new-block predictor retains about two thirds of the objective gain for $B \ge 4$, but none at $B = 2$, where it predicts universal failure.

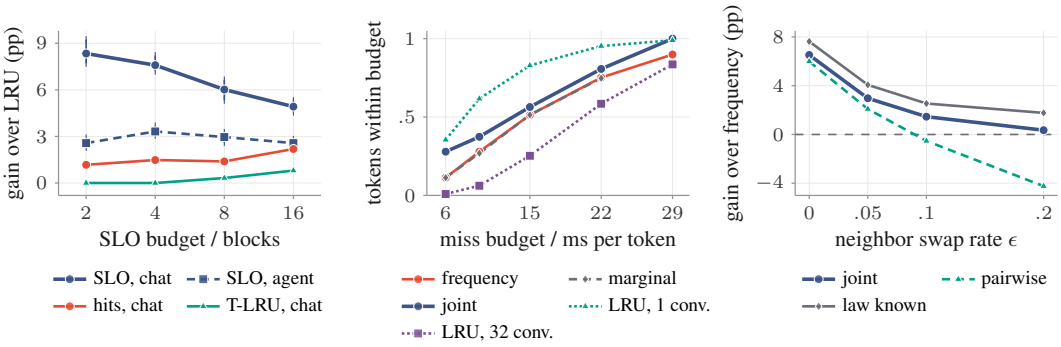

Figure 15: Resident state in LLM inference. Left: prefix KV caches on the Mooncake chat and agent traces ($M = 2{,}000$ blocks, whole pool staged every 60 s), SLO-attainment gain over LRU of SLO staging (with 95% moving-block bootstrap intervals), hit staging and T-LRU. Middle: expert residency on OLMoE-1B-7B, share of later WildChat tokens within a per-token copy budget for static sets of 512 (layer, expert) pairs chosen by frequency, marginal or joint completion (mean of three seeds), and for online LRU of the same size with one or 32 conversations in rotation; the marginal planner's tie at the loosest budget is omitted. Right: BEIR chunk caches ($k = 5$, $L = 600$, $s = 8{,}000$), gain over frequency caching against the probability ϵ of swapping each retrieved document for a near neighbor; "law known" plans on samples from the true law.

Table 28: Prefix-cache controls. Brackets are paired 95% moving-block bootstrap intervals over staging windows, pooled over 12 configurations; they describe within-service temporal variation. Known-request staging uses the actual requests in the window.

SLO-attainment contrast (points), $f = 1$	Conversation	Agent
SLO − hit staging	5.0 [4.6,5.4]	2.1 [2.0,2.3]
SLO staging − LRU	6.2	2.6
T-LRU − LRU	0.4	0.1
SLO − hit staging, calibrated TTFT	4.2 [4.0,4.6]	1.9 [1.8,2.1]
Known-request hit staging − learned SLO	11.6	6.1
Conversation: SLO − LRU under a staging limit		
New blocks/window: 0 / 50 / 200 / 1,000	−0.9 / −0.4 / +0.3 / +6.0	

Prediction and staging capacity both matter: known-request hit staging outperforms learned SLO staging, and the latter moves about 0.7 GB/s, half of hit staging's traffic. Restricting staging can eliminate its benefit. On individual WildChat days (Zhao et al., 2024), all policies lie within 2.6 points and whole-cache staging often loses. Overlaying 4 or 16 days increases half-staged SLO-minus-LRU gains from 0.2 to 0.6 and 1.4 points (50–400 rather than 25–200 cache blocks). Sixteen disjoint later days reproduce gains within 0.4 points; among 48 intervals, 35 favor SLO staging and nine T-LRU, also under TTFT. These are objective effects, not missing joint information (Proposition I.1(1)).

Expert residency: an information gap on a natural workload. OLMoE-1B-7B (Muennighoff et al., 2025) is teacher-forced on 3,991 WildChat conversations, scoring response tokens. The first 70% chronologically train a static resident set; the remaining 30% (872,589 tokens) test it. Sizes $M \in \{128, 256, 384, 512, 768\}$ of 1,024 (layer, expert) pairs are evaluated at the 10th–90th percentiles of frequency caching's TRAIN misses (1.7–51 ms/token on RTX 5090). Frequency caching optimizes hits and mean misses. The same completion optimizer uses either TRAIN tokens (joint) or re-synthesized tokens matching individual (layer, expert) inclusion rates independently across layers (marginal). The five evaluated plans are global-frequency, per-layer-frequency, marginal, joint and layer-shuffled. Three seeds vary re-synthesis and fitting subsamples. Primary policies and caps use TRAIN only; random TEST halves serve only the cross-fitted diagnostic reference below. A cell is a seed–size–budget combination; intervals resample configurations after averaging seeds.

Table 29: Static expert residency across MoE families. Completion gain (points) of joint-history planning over planning from re-synthesized per-layer statistics, on teacher-forced response tokens of later WildChat conversations. Each cell is one seed of a size–budget configuration, with three seeds per configuration. Untied cells exclude cases where all five evaluated plans complete at least 99.9% of re-synthesized tokens; intervals resample configurations after averaging seeds and describe configuration variation conditional on the held-out corpus. "Intermediate regime" additionally excludes cases where all five evaluated plans complete at most 1% of those tokens. "Shuffle removes" is the share of gain lost when each layer's routed experts are permuted across training tokens. The last columns compare the static joint plan with online per-layer LRU of the same size, over the same untied cells, with one or 32 conversations in rotation. Each model is listed with its layers×experts and top-k.

Model	Joint − marginal, untied [95%]	Cells > 0	Intermediate regime	Shuffle removes	Static − LRU 1 conv.	Static − LRU 32 conv.
OLMoE-1B-7B 16×64, top-8	9.3 [7.0, 11.9]	60/60	7.2	73%	−15.9	29.0
Granite-3.0-3B 32×40, top-8	11.3 [8.5, 14.4]	75/75	8.5	93%	−0.7	43.3
Qwen1.5-MoE 24×60, top-4	11.0 [7.0, 15.3]	45/73	8.9	82%	−5.3	26.7
DeepSeek-V2-Lite 26×64, top-6	13.7 [10.4, 17.2]	60/60	11.4	88%	−17.4	31.0

At median/75th-percentile caps, the marginal model favors its own plan by 7.7 points, yet joint planning gains 7.2 on TEST ($[4.8, 10.2]$, all 30 cells), versus an objective-only gain of -0.1. At the two tightest caps, all five evaluated plans complete at most 1% of the marginal model's tokens, but joint planning gains 11.5 points ($[8.0, 15.2]$). The loosest, $\geq 99.9\%$-completion tie is excluded using model tokens, not TEST outcomes. Across 60 untied cells, joint gains 9.3 points and has lower hit rate in 59. A cross-fitted reference using the same optimizer (half of TEST conversations fitted, the other half scored) changes completion by -0.02 points; it is not a certified global optimum. Layer shuffling preserves per-layer sets but leaves only 2.5 points of the gain, implicating cross-layer dependence. Static residency loses to online LRU in all 60 cells for one conversation; with 8/32 conversations in rotation it gains 25.8/29.0 points in 57/60 and 60/60 cells. Rotation approximates concurrency, not full serving.

Replication across MoE families. Table 29 repeats the unchanged protocol on Granite-3.0-3B-A800M (Granite Team, 2024), Qwen1.5-MoE-A2.7B (Qwen Team, 2024) and DeepSeek-V2-Lite (DeepSeek-AI, 2024). Shared experts and DeepSeek's dense first layer always execute and are excluded from residency choices. Sizes remain $1/8$–$3/4$ of routed pairs; cap percentiles and cell-classification rules were fixed on OLMoE. Joint-information gains persist across models, mostly disappear under layer shuffling, and lie within 0.4 points of cross-fitted references. Objective-only gains are $0.0/-0.1$ on Granite/DeepSeek.

Qwen's post hoc breakdown exposes an objective effect: at its two loosest caps, $M/60$ fully resident layers guarantee completion for any routing. Both completion planners find this solution in 28 cells, while frequency caching fails on 7.6–23.7% of tokens (marginal-minus-frequency gain 7.2 points in the intermediate regime). The remaining 45 cells show an 18.4-point joint gain ($[14.8, 22.3]$). Across models, static joint residency gains 19.9–43.3 points over LRU at 8/32 conversations, but loses on average for a single conversation except for the near tie on Granite.

Including all 25 size–budget configurations and three seeds per model gives joint-minus-marginal gains of 9.2 [7.3, 11.3], 11.3 [8.5, 14.4], 11.0 [7.0, 15.3] and 12.3 [9.4, 15.3] points in Table 29's model order, using the same configuration-level bootstrap. The average gains persist without the near-ceiling exclusion.

Choosing between pinning and LRU. Concurrency decides the design, and a deployment can choose it from data. We pin fM expert slots, planned for completion on the first 80% of training

Table 30: Choosing the pinned share on earlier conversations. The pinned fraction $f \in \{0, 1/8, 1/4, 1/2, 3/4, 1\}$ of M is chosen per cell on the last 20% of training conversations and scored on the later test conversations against pure per-layer LRU of the same size; intervals resample size–budget configurations. The last column counts cells where the validated policy is at least LRU at all three concurrency levels.

			Validated policy − LRU (points) at			Cells
Model	Seeds	Cells	1 conv.	8 conv.	32 conv.	$\geq$ LRU
OLMoE-1B-7B	3	75	0.2 [0.0, 0.3]	23.6 [19.8, 27.4]	26.6 [23.4, 29.8]	75/75
Granite-3.0-3B	1	25	6.1 [3.5, 9.1]	41.6 [36.7, 46.8]	43.0 [38.1, 48.1]	25/25
Qwen1.5-MoE	1	25	3.3 [1.7, 5.2]	19.6 [16.2, 23.3]	26.4 [23.9, 29.1]	25/25
DeepSeek-V2-Lite	1	25	1.2 [0.5, 2.0]	24.3 [20.2, 28.6]	28.9 [25.7, 32.4]	25/25

conversations, leave the rest to per-layer LRU, and choose $f \in \{0, 1/8, 1/4, 1/2, 3/4, 1\}$ for each cell and concurrency level on the remaining 20%, in time order. On later conversations the validated policy is never below LRU in any cell of any model (Table 30). For one conversation it matches LRU on OLMoE (0.2 points) and exceeds it on the other three models by 1.2–6.1 points by pinning part or all of the cache (on Qwen1.5-MoE, the whole-layer plan at the loosest caps); with 8 or 32 conversations it matches the static plan's lead. OLMoE uses three seeds, which vary the fitting subsample; the other models use one. With all of M pinned, joint pinning beats pinning the most frequent pairs by 9.4 points on OLMoE ($[7.5, 11.5]$).

Retrieval chunk caches: a joint-information gain under recurrence. Three BEIR collections (Thakur et al., 2021) use bge-small-en-v1.5 retrieval (Xiao et al., 2024), $k \in \{5, 10\}$ documents per query and at most 512 tokens per chunk. Queries follow i.i.d. Zipf popularity: retrieval is real, recurrence semi-synthetic. Work follows Proposition I.1(2), $\rho = .15$. The grid spans SLOs, cache sizes and $s = 500$–$32{,}000$ training requests (3,456 runs, three seeds per cell), with exact population evaluation. Baselines are greedy frequency, frequency per token, and the same completion optimizer fitted to marginal re-synthesis, pairwise co-retrieval re-synthesis or joint history.

At $s = 8{,}000$, joint planning gains 5.6 points over the best frequency-only planner ($[5.1, 6.2]$; 288 configurations bootstrapped with seeds resampled within cells; 97% of 864 runs positive) and 11.1 over LRU, despite lower hit rate (.39 versus .52) and higher mean work. Pairwise counts recover 4.8 points. From $s = 500$ to $32{,}000$, joint planning recovers 29–93% of the known-law gain; marginal completion instead loses 5.6–23.2 points to frequency caching. Near-neighbor substitutions weaken recurrence: at $\epsilon = 0, .05, .1, .2$, joint gains are 6.5, 3.0, 1.5, 0.4 points (Figure 15); pairwise planning loses 4.2 at $\epsilon = .2$.

Baseline and information controls. Variable-length mean-work optimization is a 0–1 knapsack problem, not frequency ordering. A 108-run control solves these knapsacks exactly on the same three collections, with $k \in \{5, 10\}$, Zipf exponents $.6, 1, 1.3$, $M \in \{10{,}000k, 20{,}000k\}$ tokens, $L = 120k$, $s = 8{,}000$ and three seeds. It also replaces weighted sampling without replacement by systematic sampling that preserves document inclusion probabilities, and permutes query lengths to preserve both marginals while removing their association with retrieved sets. Table 31 shows that stronger baselines retain the gap and that query-length association contributes alongside co-retrieval. These subset controls do not re-estimate the full-grid 5.6-point effect.

A production RAG trace: no gain out of sample. RAGPulse (Wang et al., 2025) contains 34,987 university QA requests over 42 days. Daily staging uses earlier days only, with seven warm-up days and 24 budget/SLO/history configurations. Joint completion underperforms frequency caching (Table 32). Greedy optimization loses even on TRAIN in 63 of 72 pooled random-split runs; exact-improvement local search removes this optimization failure but not the generalization gap. Most

Table 31: RAG controls over 108 runs. Both joint–baseline and both query-length contrasts are positive in every run. The marginal-sampling comparison uses median query length.

Contrast on the 108-run control	Gain (points)
Joint completion − exact mean-work knapsack	6.83
Exact knapsack − frequency ordering	0.02
Inclusion-preserving − original marginal sampling	2.15
Joint − best frequency/original/corrected marginal plan	6.84
Joint with permuted query lengths − frequency	4.11
Joint with actual − permuted query-length pairing	2.74

passage sets occur once and only 17% of pooled TEST sets appear in TRAIN. Random splitting therefore does not recover the semi-synthetic recurrence benefit; drift is not its only limitation.

Table 32: Production RAG generalization. Local search accepts only exact TRAIN improvements. Chronological 95% intervals use three-day moving blocks; random-split intervals resample configurations. Dashes denote contrasts not reported here.

RAGPulse contrast (points)	Chronological	Random pooled	Random within-day
Greedy joint − frequency, TEST	-8.1	—	—
Local joint − frequency, TRAIN	—	1.8	4.6
Local joint − frequency, TEST	-0.5 $[-1.0, -0.03]$	0.07 $[-0.2, 0.4]$	-2.1
Marginal − frequency, TEST	-0.8	—	—
Local joint − marginal, TEST	0.25 $[-0.01, 0.57]$	0.7 $[0.4, 0.9]$	1.1 $[0.3, 2.1]$

Wall clock and latency-model error. RTX 4090 replays use 1,500 BEIR requests, resident chunk KV and a Llama-3-8B-shaped decoder (eight of 32 layers timed and scaled by four). Per-replay calibration predicts frequency-cache attainment within three points in 37/42 replays and overpredicts the rest. Joint plans often operate exactly at the deadline; fully cached frequency-policy requests retain slack. Consequently, predicted gains can reverse in measured time (Table 33). The successful 97th-percentile margin was chosen after the earlier failures, not prespecified; eight later configurations were selected after it was fixed. Repeated configurations can change sign under the 90th-percentile rule. In 72 simulated configurations, a 2σ margin mitigates multiplicative timing error. Accurate costs and conservative error estimates remain necessary for the measured gain.

Table 33: RAG latency sensitivity. Simulations average 72 configurations (6.1-point error-free gain). Replays use resident single-GPU KV; later margin choices remain adaptive.

Timing rule	Joint wins / replays	Joint − frequency (points)
Calibrated deadline, no margin	1 / 42	Predicted ahead in all 42
90th-percentile margin	28 / 42	-45 to $+16$
Near-deadline-only margin	10 / 35	—
97th-percentile margin, later replays	31 / 31	$+1.3$ to $+13.7$
Simulated error σ	No margin	Approximate 2σ margin
2%	-3.4	$+5.0$
5%	-6.7	$+3.0$
10%	-9.6	$+1.0$

Scope. Caches stay fixed between staging steps. Traces describe individual services; BEIR popularity is synthetic and ρ fixed. Timing excludes batching and queueing; all four routers use teacher forcing on one corpus and prompt format. These preparation decisions differ from online eviction (Lykouris & Vassilvitskii, 2021); all-or-nothing benefits also motivate coordinated caching (Ananthanarayanan et al., 2012). Prefix gains concern objectives, expert gains joint information; chunk gains also require recurrence and latency calibration.