PREFIX-BALANCED DECODING: OPTIMAL OCCUPANCY UNDER CAUSAL COUPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Correlated autoregressive samples can improve coverage without changing each completed sample's marginal. That guarantee need not survive inspection of the partial batch: other lanes may reveal shared randomness and distort a selected lane's next-token law. We study couplings that preserve every named lane's conditional given the full public history. Prefix-balanced decoding (PBD) meets this contract by refreshing a systematic categorical coupling within exact-prefix groups. We prove that one target-free process minimizes every individual prefix count in convex order, over all history-adapted admissible couplings. More strongly, local floor/ceiling counts characterize all such universal optimizers; minimizing one positive weighted sum of prefix-count variances is sufficient to identify this entire class. A Bellman gap identity, a conditional restart theorem, and two ordering results explain the mechanism: more balancing improves occupancy on a fixed tree, whereas exposing a finer factorization can worsen the causal optimum. On finite intent decoders induced by three real language models, PBD improves pass@8 by 1.116 percentage points, recovers 99.5% of the endpoint-aligned opportunity, and retains 98.7% of the gain with four balanced levels. Under a shared adaptive budget, full PBD also improves target-free pass by 1.09 percentage points. These results connect an exact characterization of sampling dependence to effective shallow coordination under an inspection-stable contract.

1 INTRODUCTION

When a decoder generates K candidates, correctness of each sample does not require independence of the batch. Arithmetic Sampling and quasi-random multi-sample inference exploit this freedom to reduce redundant completions while preserving their marginal distribution (Vilnis et al., 2023; Amballa et al., 2025; Li et al., 2026). A different requirement arises when the batch is consumed *before* completion. A controller may inspect every prefix, retain selected lanes, and spend the remaining budget on their continuations. Information carried by other lanes can then change a retained lane's conditional distribution even though its original endpoint marginal was correct.

We ask a precise question: *among all couplings preserving each named lane's next-token law after the complete public prefix history is observed, can one target-free sampler optimize every prefix occupancy simultaneously?* This is a dependence-design problem, not a change to the underlying decoder or a search for a better temperature. Its answer separates the useful freedom to correlate samples from the information that an adaptive consumer is allowed to observe.

Our construction, *prefix-balanced decoding* (PBD), applies a fresh shifted systematic grid within each exact-prefix group. A child of probability p receives either $\lfloor np \rfloor$ or $\lceil np \rceil$ of its parent's n lanes. This one-step rounding rule is classical. The contribution is its *dynamic characterization*: it solves every convex prefix-count objective against every history-dependent admissible coupling, and every universal optimizer must have these local adjacent counts. In particular, a single strictly positive aggregate variance objective identifies the same optimal class. Sibling correlations remain nonunique; arbitrary semantic-set utilities are outside the guarantee.

The characterization has operational consequences. A Bellman identity localizes any excess convex cost to nonnegative conditional rounding gaps. Balancing more transitions improves all nodewise objectives on a fixed tree; revealing extra intermediate prefixes can instead increase the price of the

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.

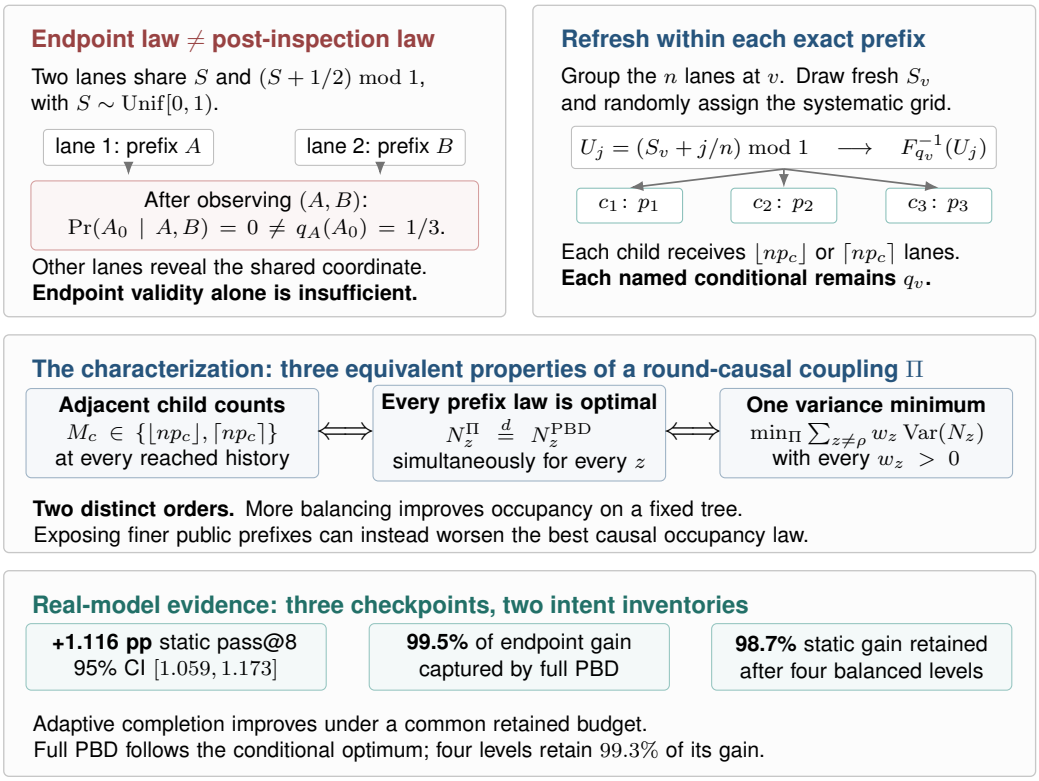

Figure 1: **Local balance characterizes global occupancy optimality.** Shared endpoint randomness can become informative after the prefix batch is observed. Fresh within-prefix coupling instead preserves every named conditional and adjacently rounds child counts. These local rules characterize all universal prefix optimizers and all minimizers of a positive all-prefix variance sum. Real-model experiments show consistent gains across six cells and near-complete recovery of the endpoint-aligned opportunity. CI denotes confidence interval; pp denotes percentage points.

causal contract. At a common inspected history, fresh PBD is optimal for completing any once-selected subforest. This last statement connects the theory to budgeted consumption and supplies a strong implementation check.

We evaluate the theory on Banking77 and CLINC150 using finite constrained decoders induced by three frozen language models. Exact occupancy analysis shows positive pass@8 gains in all six cells and recovers 99.5% of the endpoint-aligned opportunity. Four balanced levels retain 98.7% of the static gain, demonstrating the value of shallow coordination. Under a common adaptive budget, full PBD also improves completion utility and satisfies the conditional ordering predicted by the restart theorem (Section 5.1).

2 The Contract and the Balancing Construction

Decoder and public information. Fix a prompt and a finite rooted prefix tree $\mathcal{T}$ with child kernel q_v at node v. Completed leaves are absorbing. There are K named lanes; V_t^r is lane r's prefix after synchronized round t. The public sigma-field $\mathcal{F}_t$ contains the full prefix vector, previous controller actions, and any already revealed randomness, but not private randomness for the next draw. The decoder kernel is fixed; a controller can select lanes without silently changing q_v.

Definition 1 (Public-history causal validity)**.** *A coupling Π is* round-causal *if, almost surely, every active named lane r satisfies*

$$\mathbb{P}_{\Pi}(V_{t+1}^r = c \mid \mathcal{F}_t) = q_{V_t^r}(c). \tag{1}$$

The next joint transition may otherwise depend on the whole history and correlate all lanes and prefix groups.

The main occupancy results concern K lanes completed without pruning. Equation (1) then implies each endpoint Y^r has the base law P. For an $\mathcal{F}_t$-measurable retained lane R, it also gives $\mathbb{P}(V_{t+1}^R = c \mid \mathcal{F}_t) = q_{V_t^R}(c)$. This is a conditional continuation guarantee, *not* a claim that a selection-biased returned batch has endpoint law P. Decisions made inside a partly revealed round are not covered.

Why endpoint validity is weaker. Let $K = 2$ and use arithmetic coordinates S and $(S+1/2) \bmod 1$, with $S \sim \mathrm{Unif}[0,1)$. Let root intervals be $A = [0, 3/4)$ and $B = [3/4, 1)$; within A, let $A_0 = [0, 1/4)$ and $A_1 = [1/4, 3/4)$. Both named endpoints have the prescribed marginal. But the positive-probability public event $(V_1^1, V_1^2) = (A, B)$ forces $S \in [1/4, 1/2)$, hence $\mathbb{P}(V_2^1 = A_0 \mid \mathcal{F}_1) = 0 \neq q_A(A_0) = 1/3$. The endpoint sampler is not invalid for its intended contract; the stronger conditioning changes the feasible set.

Prefix-balanced decoding. We build on systematic categorical resampling (Gerber et al., 2019). After the public history and any selection action are fixed, group the active lanes by exact prefix. For each occupied v with n lanes, draw fresh $S_v \sim \mathrm{Unif}[0,1)$ and a uniform random permutation σ_v. Map

$$U_j = (S_v + j/n) \bmod 1, \quad j = 0, \dots, n-1, \qquad V_{t+1}^{\sigma_v(j)} = F_{q_v}^{-1}(U_j) \tag{2}$$

through the fixed child cumulative distribution function (CDF) F_{q_v}. Repeat with independent fresh draws at the next public boundary. Each named coordinate is conditionally uniform, so Eq. (1) holds. Simultaneously, each child count M_c obeys

$$M_c \mid \mathcal{F}_t \sim \mathrm{Adj}(nq_v(c)), \qquad \mathrm{Adj}(x) = \lfloor x \rfloor + \mathrm{Bernoulli}(x - \lfloor x \rfloor). \tag{3}$$

Permutation makes the construction exchangeable; it is not needed for the marginal uniformity of an already named shifted grid. The method uses the same next-token kernels as the base decoder, but requires synchronization and prefix grouping.

3 A CHARACTERIZATION OF OPTIMAL PREFIX OCCUPANCY

For a prefix z, let $s_z = P(z \preceq Y)$ and $N_z = \sum_{r=1}^K \mathbf{1}\{z \preceq Y^r\}$. Every unpruned round-causal coupling has $\mathbb{E}N_z = Ks_z$. Write $X \preceq_{\mathrm{cx}} Y$ when $\mathbb{E}\phi(X) \leq \mathbb{E}\phi(Y)$ for every convex ϕ; here this order compares integer-valued laws of equal mean (Shaked & Shanthikumar, 2007).

Lemma 1 (The complete scalar one-step feasible set)**.** *Given any public history with n lanes at v, the feasible laws of the count sent to a designated child of probability p are exactly the distributions on $\{0, \dots, n\}$ with mean np.*

Necessity follows by summing named-lane marginals. Conversely, draw a count with this mean, choose that many lanes uniformly for the designated child, and distribute the remaining lanes according to the renormalized other-child kernel. This describes a *scalar* feasible set, not arbitrary simultaneous sibling-count laws.

Theorem 1 (Universal occupancy optimum)**.** *For every finite decoder tree, $K \geq 1$, prefix z, and round-causal coupling Π,*

$$N_z^{\mathrm{PBD}} \preceq_{\mathrm{cx}} N_z^{\Pi}. \tag{4}$$

One target-free PBD process attains these optima simultaneously for all nodes and all convex costs, including against couplings using arbitrary off-path history and side information.

Why local balance remains globally optimal. On the path to z, let $p_1, \dots, p_d$ be its edge probabilities. For a discretely convex function f on $\{0, \dots, K\}$, let $\bar{f}$ be its piecewise-linear interpolation and set $(A_p f)(n) = \bar{f}(np)$. Define $V_d = \phi$ and $V_{t-1} = A_{p_t} V_t$. Each V_t remains discretely convex. At every complete history, Lemma 1 and conditional Jensen give $\mathbb{E}[V_t(N_t) \mid \mathcal{F}_{t-1}] \geq \bar{V}_t(p_t N_{t-1})$. Backward induction proves the lower bound $V_0(K)$; adjacent rounding attains it at every step. A systematic categorical grid attains this for every sibling at once. This is why no target-specific optimization or off-path information can improve any individual prefix-count law.

The result characterizes more than one particular sampler.

Theorem 2 (All universal optimizers are locally balanced)**.** *Fix a finite tree and positive weights $w_z > 0$ for every nonroot node. For a round-causal coupling Π, the following are equivalent:*

(i) every N_z^{Π} has the same law as N_z^{PBD};

(ii) Π minimizes $\sum_{z \neq \mathrm{root}} w_z \operatorname{Var}(N_z)$ over all round-causal couplings;

(iii) at every reached public history, every child count is almost surely in $\{\lfloor N_v q_v(c) \rfloor, \lceil N_v q_v(c) \rceil\}$.

Thus one positive aggregate variance objective characterizes the entire class of simultaneous convex-order optimizers. It does not identify a unique joint law across siblings.

Proof idea and an exact optimality certificate. All individual variance gaps are nonnegative by Theorem 1. Equality of their positive weighted sum forces equality at every node. For integer laws on $\{0, \dots, K\}$, convex order plus equal second moments forces equality of laws: express n^2 as $n + 2\sum_{a=1}^{K-1}(n-a)_+$. At a child with $p = q_v(c)$ and $a = \lfloor pN_v \rfloor$, the conditional quadratic excess is

$$\mathbb{E}[(N_c - a)(N_c - a - 1) \mid \mathcal{F}_t] \geq 0. \tag{5}$$

It vanishes precisely for adjacent counts, with the conditional mean forcing a deterministic count when pN_v is integral. Equal parent laws and child second moments force the expected excess to vanish. Conversely, adjacent conditional counts determine every root-to-node scalar transition law. Appendix A supplies all details.

More generally, the Bellman construction gives the *exact* gap identity

$$\mathbb{E}_{\Pi}\phi(N_d) - V_0(K) = \sum_{t=1}^{d} \mathbb{E}_{\Pi}\left[\mathbb{E}_{\Pi}[V_t(N_t) \mid \mathcal{F}_{t-1}] - \bar{V}_t(p_t N_{t-1})\right]. \tag{6}$$

Every summand is nonnegative. The identity localizes avoidable occupancy cost to conditional rounding decisions rather than attributing it to an unspecified global diversity effect.

Exactly which utilities are optimized. For any fixed finite node set and nonnegative weights, PBD maximizes $\mathbb{E}\sum_z w_z u_z(N_z)$ for concave u_z, and minimizes it for convex u_z. This includes node coverage, capped multiplicity, empirical entropy on a leaf partition, pair collisions, and count-frequency mean-square error. The guarantee is nodewise and separable, not a multivariate convex order. For example, with four equiprobable root leaves a, b, c, d and $K = 2$, the balanced pairings $\{a, c\}/\{b, d\}$ and $\{a, b\}/\{c, d\}$ both optimize every leaf count. Yet they hit the semantic union $\{a, c\}$ with probabilities $1/2$ and 1, respectively, versus $3/4$ for independent and identically distributed (IID) sampling. All named marginals are uniform. This sharp boundary also explains the nonuniqueness in Theorem 2.

4 Balancing, revelation, and selective continuation

More balancing on a fixed tree helps. Let PBD_L balance the first L transitions and use conditionally IID continuations afterwards; $\mathrm{PBD}_0 = \mathrm{IID}$.

Theorem 3 (Depth hierarchy)**.** *For $0 \leq L \leq L'$ and every prefix z,*

$$N_z^{\mathrm{PBD}} \preceq_{\mathrm{cx}} N_z^{\mathrm{PBD}_{L'}} \preceq_{\mathrm{cx}} N_z^{\mathrm{PBD}_L} \preceq_{\mathrm{cx}} N_z^{\mathrm{IID}}. \tag{7}$$

The same order holds when any deterministic set of balanced transitions is enlarged. All fixed separable concave occupancy utilities improve monotonically.

The key is that the IID operator $(B_p f)(n) = \mathbb{E}f(\mathrm{Bin}(n, p))$ preserves discrete convexity and satisfies $A_p f \leq B_p f$. Earlier expectation operators preserve this inequality. This justifies a depth ablation mathematically; it is not a general rule for arbitrary diversity scores.

The role of public factorization. For endpoint marginals alone, a prefix count may attain $\mathrm{Adj}(Ks_z)$, with hit probability $\min\{1, Ks_z\}$, the marginal-preserving union-bound ceiling also used by Li et al. (2026). Define the causal coverage price

$$\Delta_z = \min\{1, Ks_z\} - \mathbb{P}(N_z^{\mathrm{PBD}} > 0) \geq 0. \tag{8}$$

Theorem 4 (Factorization-refinement order)**.** *Replace a path transition of probability p by publicly exposed transitions $p_1, \dots, p_m$ with product p, leaving all other path probabilities unchanged. For the corresponding coarse and refined PBD occupancies,*

$$\mathrm{Adj}(Ks_z) \preceq_{\mathrm{cx}} N_z^{\mathrm{coarse}} \preceq_{\mathrm{cx}} N_z^{\mathrm{refined}} \preceq_{\mathrm{cx}} \mathrm{Bin}(K, s_z). \tag{9}$$

Consequently refinement cannot reduce the causal coverage price.

Indeed $A_{p_1 \cdots p_m} f \le A_{p_1} \cdots A_{p_m} f$ for convex f, by Jensen applied to the final integer count with the same mean. For $K = 2$ and the refined path $(3/4, 3/4)$, PBD hits with probability $7/8$, strictly between IID's $207/256$ and the endpoint ceiling 1. The two theorems concern different interventions: adding *balancing* on a fixed representation helps, while adding *public constraints* to the representation may hurt. A known laminar decision hierarchy can instead be emitted before surface realization while preserving P; its top-level classes then attain the endpoint ceiling (Appendix A.10). This requires access to the class masses, not merely relabeling an existing language-model (LM) prompt.

Optimal continuation after a common selection. The public-history contract is stable under inspection, but arbitrary controller performance is not automatically optimized. The precise useful extension is the following.

Theorem 5 (Conditional restart on a selected subforest)**.** *After an unpruned synchronous run to an almost-surely finite public stopping boundary τ, select any $\mathcal{F}_\tau$-measurable subset of lanes once, and then complete these lanes without further pruning. Conditional on $\mathcal{F}_\tau$, fresh PBD within the retained prefix groups minimizes every descendant occupancy count in convex order among all round-causal continuations from that same selected state. It optimizes every conditionally fixed separable occupancy objective, including target-leaf hit and distinct completed leaves.*

Each occupied boundary prefix roots a disjoint subtree with a known retained count, so the history-uniform proof restarts there. In particular, when $L \ge \tau$ and selection uses the same observed histories, full PBD cannot be worse than PBD_L on these completion utilities. This does not compare controllers with different observation laws, repeated future pruning, or nonseparable semantic rewards.

Finite-budget error and sure coverage. On a depth-d path, let $a_j = \prod_{\ell=j+1}^{d} p_\ell$ and $\rho_j = \{p_j N_{j-1}\}$. Under any locally balanced optimizer,

$$\mathrm{Var}(N_z) = \sum_{j=1}^{d} a_j^2 \mathbb{E}[\rho_j(1-\rho_j)], \qquad \mathbb{E}(N_z/K - s_z)^2 \le \min\left\{ \frac{d}{4K^2}, \frac{s_z(1-s_z)}{K} \right\}. \tag{10}$$

The $O(K^{-2})$ statement is at fixed depth, not uniform over growing-length outputs. For $p_j > 0$, the recursion $R_{d+1} = 1$, $R_j = \lceil R_{j+1}/p_j \rceil$ gives the exact minimum batch size R_1 for sure coverage under *any* round-causal coupling. Appendix A.9 proves these statements and a pathwise discrepancy bound; finite-horizon results extend to locally finite infinite trees.

5 Evidence from real finite language-model decoders

Controlled evaluation of sampling dependence. We use Banking77 (Casanueva et al., 2020), CLINC150 (Larson et al., 2019), Qwen2.5-3B/7B-Instruct (Qwen, 2025), and Phi-3.5-mini-instruct (Abdin et al., 2024). For each real test query, the model scores every canonical label plus the end-of-sequence (EOS) token. With token logits a_j,

$$\ell_T(y \mid x) = \sum_j \log \mathrm{softmax}(a_j(x, y_{<j})/T)_{y_j}, \qquad \pi_T(y \mid x) = \frac{e^{\ell_T(y|x)}}{\sum_{y'} e^{\ell_T(y'|x)}}. \tag{11}$$

Subtree-mass ratios define an autoregressive trie with leaf law π_T. This is the LM law *conditioned on the finite label inventory*, not unrestricted generation or ordinary locally masked decoding. Every method receives exactly the same posterior. Static count expectations are computed by exact recursion, conditional on the stored floating-point graphics processing unit (GPU) scores; pass@K is the actual coupled target-hit probability, not an IID estimator applied to correlated samples.

Table 1: **Consistent static gains across all six model–dataset cells.** Pass@8 is in percent; gains and paired 95% intervals are in percentage points (pp). Recovery is the ratio of summed PBD–IID gain to summed endpoint-ceiling opportunity within the cell, not an average of unstable per-query ratios. The six-cell summary is descriptive; the frozen primary estimand is the four Qwen cells. B77 denotes Banking77.

Task	Model	N	IID	PBD	Gain [95% CI]	Recovery (%)
B77	Φ-3.5	3,080	50.90	52.21	+1.309 [+1.154, +1.469]	99.3
B77	Qwen-3B	3,080	59.03	59.96	+0.936 [+0.805, +1.073]	99.3
B77	Qwen-7B	3,080	69.34	70.31	+0.966 [+0.825, +1.116]	99.6
CLINC	Φ-3.5	4,530	68.44	69.93	+1.492 [+1.350, +1.639]	99.7
CLINC	Qwen-3B	4,530	72.58	73.82	+1.246 [+1.116, +1.380]	99.5
CLINC	Qwen-7B	4,530	81.49	82.24	+0.744 [+0.646, +0.844]	99.7

Exact finite-decoder evaluation. For a prefix-free inventory with leaf probabilities π_j, the conditional on an edge is child subtree mass divided by parent subtree mass. Full-PBD node laws are propagated by a sparse adjacent-thinning matrix: each incoming count has at most two successors. Once subtree masses are available, all full-PBD node laws and any fixed collection of separable functionals cost $O(K|\mathcal{T}|)$ arithmetic operations. IID and endpoint-aligned hit probabilities are $1-(1-\pi_j)^K$ and $\min\{1, K\pi_j\}$.

For depth-limited PBD followed by IID, a literal binomial count-law transition is dense and can cost $O(K^2)$ per edge, so a direct full-distribution implementation is bounded by $O(K^2|\mathcal{T}|)$ rather than universally $O(K|\mathcal{T}|)$. Hit probabilities admit shortcuts after the balanced frontier: if the remaining path mass is r, then conditional on n lanes its failure probability is $(1-r)^n$. None of these arithmetic-operation statements includes language-model evaluation, hardware synchronization, or vocabulary CDF construction.

The primary operating point is $T = 1, K = 8$. Banking77 uses all 3,080 test queries; CLINC uses 4,530 class-balanced primary queries, with all 5,500 as robustness. Intervals use 10,000 intent-stratified prompt bootstrap replicates, paired across models within each dataset. These are fixed checkpoints, not training-seed replicates. Appendix B gives the scoring protocol, out-of-scope (OOS) subset construction, baselines, and complete results; the supplement records exact asset revisions and execution settings.

Consistent gains and opportunity recovery. All six cells improve (Table 1). Their equal-cell descriptive mean is $+1.116$ pp, 95% CI $[1.059, 1.173]$; opportunity recovery is 99.5%, with mean causal price 0.0053 pp. The original Qwen estimand is $+0.973$ pp $[0.909, 1.039]$, and the separate Phi replication is $+1.401$ pp $[1.294, 1.509]$. The independent replication and original practical criterion are reported in Appendix B. Together, the six cells show that PBD captures nearly all of the available marginal-preserving coverage improvement on these decoders.

Mechanism and alternative objectives. One, two, and four balanced levels retain 75.7%, 92.1%, and 98.7% of the full static gain, respectively (Figure 2). Target-free quantities also move in the predicted direction in all cells: equal-cell changes are $+0.0659$ distinct labels, $+0.02236$ empirical entropy, 0.36596 fewer collision pairs, and 0.73192 lower summed count variance. The unchanged-posterior endpoint lattice is slightly better at coverage, as permitted by its weaker contract. Gumbel and deterministic top-K achieve much larger target-hit probabilities but do not preserve the named sampling marginal; their full values remain in Table 11.

The target-free gains in Table 2 complement target-hit coverage: PBD increases the number and entropy of completed labels while reducing repeated pairs and count variability. Figure 3 shows positive pass gains throughout the prespecified temperature and batch-size grid, using the same fixed model outputs.

Table 3 reports the complete CLINC test results. This complete-test analysis includes all 1,000 out-of-scope examples and removes dependence on the reconstructed 30-example primary OOS subset.

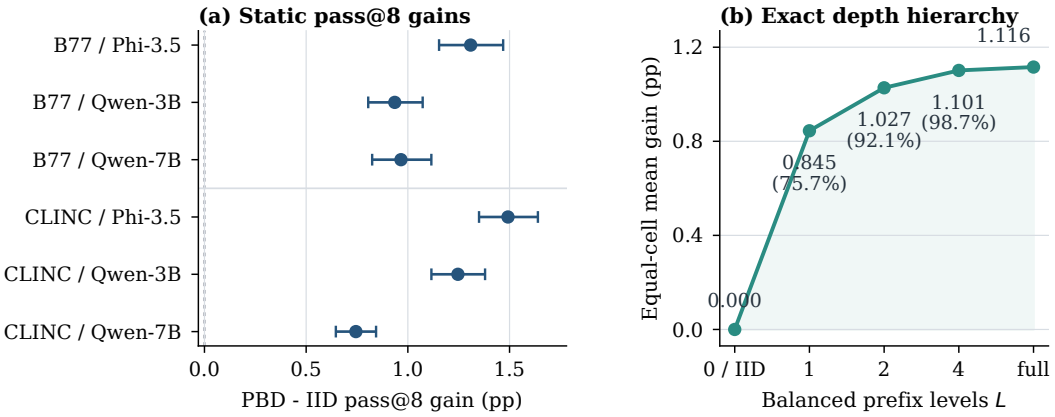

Figure 2: **Cell-wise gains and the predicted depth hierarchy.** Left: paired 95% prompt intervals for full PBD versus IID. Right: exact six-cell mean gain as more prefix levels are balanced; percentages measure retention relative to full PBD, not raw accuracy. Four levels retain 98.7% of the static gain.

Table 2: **PBD improves every prespecified target-free occupancy statistic in every cell.** Exact changes at $T = 1, K = 8$. Positive values indicate improvement; for collisions and count variance, the table reports IID minus PBD.

Task	Model	Distinct outputs	Entropy	Pair collisions	Count variance
B77	Φ-3.5	+0.091	+0.0303	+0.490	+0.980
B77	Qwen-3B	+0.071	+0.0244	+0.398	+0.795
B77	Qwen-7B	+0.045	+0.0153	+0.254	+0.508
CLINC	Φ-3.5	+0.078	+0.0266	+0.440	+0.879
CLINC	Qwen-3B	+0.072	+0.0247	+0.406	+0.812
CLINC	Qwen-7B	+0.039	+0.0129	+0.208	+0.416

Table 3: **The positive gain persists on the complete 5,500-item CLINC test set.** Results use $T = 1, K = 8$; the 4,530-item rows are the primary subset shown for comparison. IID and PBD columns are percentages.

Model	N	IID	PBD	Δ pp
Φ-3.5	4,530	68.44	69.93	+1.492
Qwen-3B	4,530	72.58	73.82	+1.246
Qwen-7B	4,530	81.49	82.24	+0.744
Φ-3.5	5,500	57.02	58.29	+1.269
Qwen-3B	5,500	62.12	63.26	+1.145
Qwen-7B	5,500	78.15	78.93	+0.781

5.1 Adaptive consumption under a shared budget

The frozen adaptive protocol generates eight lanes, observes two token rounds, retains four lanes once, and completes them. A target-free controller prioritizes distinct prefixes by their frozen subtree entropy, then mass and token order, and fills remaining slots round-robin. A target-aware oracle is a diagnostic, not a deployable method. Boundary occupancies use the original counter-based Monte Carlo draws; continuation expectations and the oracle are exact. Global endpoint lattices are integrated exactly on a predeclared 1,368-row model–query subset. Appendix D gives all rules, uncertainty estimates, robustness cases, and precision diagnostics.

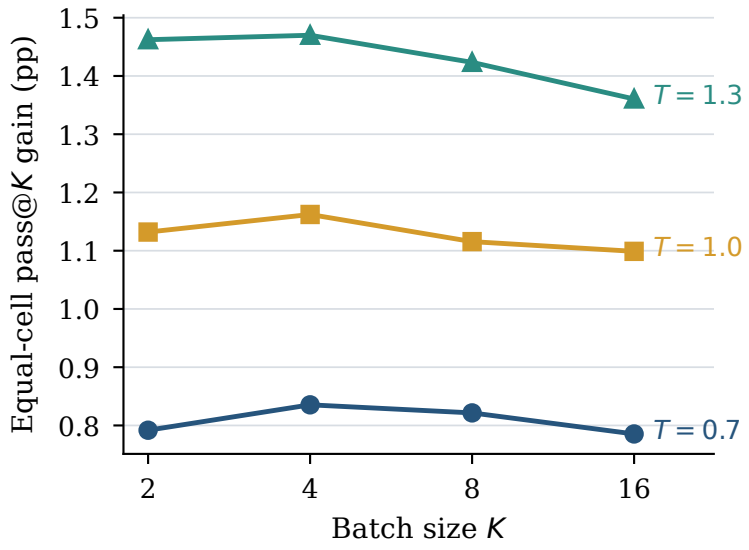

Figure 3: **Static gains are stable across batch size within each temperature.** The plotted values are the exact entries of Table 10.

Table 4: **Positive completion gains under a shared adaptive budget.** Differences from IID under the same target-free controller; oracle columns use the separately labeled target-aware controller. Pass changes and paired 95% intervals are in pp. All methods use the same observation depth and retained-lane budget.

Method	Target-free pass	Oracle pass	Distinct outputs
PBD_1	+0.82 [0.77, 0.87]	+0.85 [0.80, 0.89]	+0.0457 [0.0443, 0.0471]
PBD_2	+1.00 [0.94, 1.05]	+1.03 [0.98, 1.08]	+0.0571 [0.0555, 0.0587]
PBD_4	+1.08 [1.03, 1.14]	+1.11 [1.05, 1.17]	+0.0609 [0.0593, 0.0626]
Full PBD	+1.09 [1.04, 1.15]	+1.12 [1.06, 1.18]	+0.0614 [0.0598, 0.0631]

Full PBD gives +1.09 [1.04, 1.15] pp target-free pass and +0.0614 [0.0598, 0.0631] distinct outputs over IID. The four-level variant gives $+1.08$ pp $[1.03, 1.14]$ and $+0.0609$ distinct outputs $[0.0593, 0.0626]$. Its target-free gain retention is 99.3% $[99.0, 99.6]$. The pointwise full-versus-four-level ordering agrees with Theorem 5; Appendix D.2 records the continuation validation.

Table 5 resolves the cell-wise results; Table 6 reports the remaining separable utilities.

Table 5: **Full-PBD gains are positive in every model–dataset cell.** Target-free pass differences and their paired 95% intervals are in percentage points. Distinct-output changes are expected counts.

		Target-free pass Δ pp [95% CI]		Distinct Δ	
Task	Model	PBD_4	Full	PBD_4	Full
B77	Φ-3.5	+1.22 [1.06, 1.38]	+1.26 [1.10, 1.42]	+0.0835	+0.0859
B77	Qwen-3B	+0.94 [0.81, 1.08]	+0.95 [0.82, 1.08]	+0.0650	+0.0651
B77	Qwen-7B	+0.96 [0.82, 1.11]	+0.96 [0.82, 1.11]	+0.0419	+0.0422
CLINC	Φ-3.5	+1.45 [1.30, 1.59]	+1.45 [1.30, 1.59]	+0.0736	+0.0736
CLINC	Qwen-3B	+1.19 [1.06, 1.33]	+1.19 [1.06, 1.33]	+0.0656	+0.0656
CLINC	Qwen-7B	+0.74 [0.64, 0.85]	+0.74 [0.64, 0.85]	+0.0360	+0.0360

Table 6: **The remaining separable utilities improve in the predicted direction.** Primary equal-cell changes relative to IID with paired 95% intervals. Positive values indicate improvement; collisions and conditional count variance are reported as IID minus method. The latter conditions on inspection and selection.

Quantity	PBD_4	Full PBD
Entropy	+0.02968 [0.02896, 0.03041]	+0.02992 [0.02919, 0.03064]
Fewer collisions	+0.1386 [0.1355, 0.1418]	+0.1398 [0.1365, 0.1429]
Lower conditional variance	+0.0179 [0.0161, 0.0198]	+0.0201 [0.0181, 0.0221]

Conditional fidelity after inspection. On the fixed integration subset, token-lexicographic endpoint coupling gives $+0.94$ pp target-free pass and $+0.0548$ distinct outputs over IID. Its mean post-inspection total-variation (TV) discrepancy is 0.00236 $[0.00153, 0.00319]$; oracle-selected target calibration gap is 0.00143 $[0.00068, 0.00218]$. Alphabetical ordering gives a similar TV gap. These measured discrepancies are below the predeclared 0.01 materiality threshold. PBD preserves the conditional law exactly; this experiment quantifies the contract difference in the tested finite-decoder regime.

6 Related work and scope

Relation to prior coupling methods. Arithmetic Sampling and later quasi-random decoding preserve endpoint marginals and exploit parallel structured randomness (Vilnis et al., 2023; Amballa et al., 2025; Li et al., 2026). Our finite endpoint lattice is a contract reference, not a reproduction of all variants or reasoning benchmarks in those papers. Systematic resampling, array randomized quasi-Monte Carlo (Array-RQMC), and dependent rounding study local balance and stagewise variance reduction (L'Ecuyer et al., 2008; Gerber et al., 2019; Gandhi et al., 2006). Our contribution is an exact history-uniform occupancy optimum and a necessary-and-sufficient characterization of *all* optimizers.

Causal/adapted transport supplies a general language and dynamic programming for filtered couplings (Backhoff et al., 2017; Eckstein & Pammer, 2024; Acciaio et al., 2025); coupled Markov samplers optimize other criteria such as asymptotic variance (Nüsken & Pavliotis, 2019). Convexity propagation by backward induction is also classical (Pagès, 2014). Our solution identifies the complete optimizer class for laminar prefix counts, not for general multicausal costs. Coupled token generation compares different models (Corvelo Benz et al., 2026); PBD coordinates replicas of one model. Diverse beam, without-replacement, and determinantal search, as well as semantic-guided diversity decoding, instead relax the named-marginal constraint (Vijayakumar et al., 2018; Kool et al., 2019; Meister et al., 2021; Shi et al., 2025). Appendix E gives claim-level comparisons.

Scope and use. The theory covers individual prefix counts and fixed separable sums under a synchronized public-history contract. Arbitrary semantic unions and repeated adaptive pruning require additional analysis. The empirical results concern short, globally conditioned intent inventories; long-form generation and production-serving throughput remain to be evaluated. Prefix synchronization has a systems cost, and decision-first factorization requires computable conditional class masses. PBD is particularly suited to consumers requiring post-inspection conditional fidelity.

7 Conclusion

Under public-history causal validity, locally adjacent child counts are exactly the universal prefix-occupancy optimizers. PBD gives a simple, target-free realization, with a Bellman certificate, a factorization order, and an optimal conditional restart rule. Across six real-model finite decoders, it improves coverage consistently and recovers 99.5% of the endpoint-aligned opportunity. Four balanced levels retain 98.7% of the static gain and 99.3% of the adaptive target-free gain. These results combine exact full-depth optimality with effective shallow coordination under an inspection-stable sampling contract.

REPRODUCIBILITY STATEMENT

We provide anonymized source code for the results reported in this paper in the supplementary material.

ARTIFICIAL INTELLIGENCE (AI) USE STATEMENT

The author developed the research idea and proofs. AI tools provided coding, debugging, writing, and editing support, including polishing the exposition and presentation of the proofs. The author reviewed and verified the resulting material and retains responsibility for the research questions, methods, proofs, experiments, interpretation, and final content.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Beatrice Acciaio, Daniel Kršek, and Gudmund Pammer. Multicausal transport: Barycenters and dynamic matching. *SIAM Journal on Financial Mathematics*, 16(3):1104–1138, 2025. doi: 10.1137/24M1634096.

Avinash Amballa, Aditya Parashar, Aditya Vikram Singh, Jinlin Lai, and Benjamin Rozonoyer. Quasi-random multi-sample inference for large language models. In *ICLR Workshop on Frontiers in Probabilistic Inference: Sampling Meets Learning*, 2025. URL https://openreview.net/forum?id=aUAExbvozj.

Julio Backhoff, Mathias Beiglböck, Yiqing Lin, and Anastasiia Zalashko. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562, 2017. doi: 10.1137/16M1080197.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.nlp4convai-1.5.

Nina L. Corvelo Benz, Stratis Tsirtsis, Eleni Straitouri, Ivi Chatzi, Ander Artola Velasco, Suhas Thejaswi, and Manuel Gomez Rodriguez. Evaluation of large language models via coupled token generation. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 2215–2223. PMLR, 2026.

Stephan Eckstein and Gudmund Pammer. Computational methods for adapted optimal transport. *The Annals of Applied Probability*, 34(1A):675–713, 2024. doi: 10.1214/23-AAP1975.

Rajiv Gandhi, Samir Khuller, Srinivasan Parthasarathy, and Aravind Srinivasan. Dependent rounding and its applications to approximation algorithms. *Journal of the ACM*, 53(3):324–360, 2006. doi: 10.1145/1147954.1147956.

Mathieu Gerber, Nicolas Chopin, and Nick Whiteley. Negative association, ordering and convergence of resampling methods. *The Annals of Statistics*, 47(4):2236–2260, 2019. doi: 10.1214/18-AOS1746.

Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The Gumbel-top-k trick for sampling sequences without replacement. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3499–3508. PMLR, 2019.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the*

2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 1311–1316. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1131.

Pierre L'Ecuyer, Christian Lécot, and Bruno Tuffin. A randomized quasi-Monte Carlo simulation method for Markov chains. *Operations Research*, 56(4):958–975, 2008. doi: 10.1287/opre.1080.0556.

Michael Y. Li, Anthony Zhan, Kanishk Gandhi, Noah D. Goodman, and Emily B. Fox. QuasiMoTTo: Quasi-monte carlo test-time scaling. *arXiv preprint arXiv:2607.01179*, 2026.

Clara Meister, Martina Forster, and Ryan Cotterell. Determinantal beam search. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6551–6562. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.512.

N. Nüsken and G. A. Pavliotis. Constructing sampling schemes via coupling: Markov semigroups and optimal transport. *SIAM/ASA Journal on Uncertainty Quantification*, 7(1):324–382, 2019. doi: 10.1137/18M119896X.

Gilles Pagès. Convex order for path-dependent derivatives: a dynamic programming approach. *arXiv preprint arXiv:1407.6348*, 2014. URL `https://arxiv.org/abs/1407.6348`.

Qwen. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Moshe Shaked and J. George Shanthikumar. *Stochastic Orders*. Springer Series in Statistics. Springer, New York, 2007. doi: 10.1007/978-0-387-34675-5.

Weijie Shi, Yue Cui, Yaguang Wu, Jingzhi Fang, Shibo Zhang, Mengze Li, Sirui Han, Jia Zhu, Jiajie Xu, and Xiaofang Zhou. Semantic-guided diverse decoding for large language model. *arXiv preprint arXiv:2506.23601*, 2025.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.12340.

Luke Vilnis, Yury Zemlyanskiy, Patrick Murray, Alexandre Tachard Passos, and Sumit Sanghai. Arithmetic sampling: Parallel diverse decoding for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35120–35136. PMLR, 2023.

A COMPLETE PROOFS AND THE SCOPE OF THE CHARACTERIZATION

The proof dependencies are as follows. The systematic-grid lemma establishes feasibility; the scalar feasible-set lemma and convexity-preserving adjacent operator give Theorem 1. Equality of convex-order laws at fixed second moment then gives Theorem 2. The depth and refinement orders use different operator comparisons. The restart result conditions the universal proof on a selected boundary state. No empirical observation is used in a proof.

A.1 PROBABILITY SPACE, OCCUPANCIES, AND DISCRETE CONVEXITY

Throughout the unpruned results, K named lanes start at the root of a fixed finite tree and are completed; q_v is the same kernel in every lane. The sigma-field $\mathcal{F}_t$ includes their full public past through synchronized round t. Conditional equalities are almost-sure equalities of regular conditional probabilities; we do not condition only on histories with positive point mass, since public random variables may be continuous. The only restriction on a competitor's joint transition is Eq. (1). Fresh independent randomness is a sufficient PBD implementation, not an extra assumption imposed on competitors.

For a node $z = v_d$, write $v_0, \dots, v_d$ for its root path, $p_t = q_{v_{t-1}}(v_t)$, and $s_t = \prod_{j=1}^{t} p_j$. Define $N_t = \sum_r \mathbf{1}\{V_t^r = v_t\}$ for $0 \le t \le d$, with $N_0 = K$. No off-path lane can later enter this prefix path. Therefore

$$\mathbb{E}[N_t \mid \mathcal{F}_{t-1}] = p_t N_{t-1}, \qquad \mathbb{E}N_t = K s_t. \tag{12}$$

All results allow zero and unit probabilities. A zero-probability edge makes its descendants unreachable. Empty prefix groups need no random draw. Absorbing leaves can be extended by deterministic self-transitions when a common finite horizon is convenient.

A function $f : \{0, \dots, K\} \to \mathbb{R}$ is discretely convex if $\Delta^2 f(n) = f(n+2) - 2f(n+1) + f(n) \ge 0$. Its piecewise-linear interpolation $\bar{f}$ on $[0, K]$ is convex exactly when these inequalities hold. Every such f admits the representation

$$f(n) = f(0) + n\Delta f(0) + \sum_{a=1}^{K-1} \Delta^2 f(a-1)(n-a)_+. \tag{13}$$

The empty sum convention handles $K = 1$. This identity follows by matching the initial value and all successive first differences. Testing discretely convex costs is equivalent to ordinary convex order on this integer support.

A.2 SYSTEMATIC COUNTS AND SCALAR FEASIBILITY

Lemma 2 (Systematic categorical counts)**.** *A uniformly shifted regular n-point grid puts either $\lfloor np \rfloor$ or $\lceil np \rceil$ points in every half-open circular interval of length p, with mean np. Thus all categorical child counts are adjacent simultaneously. Each named point is uniform; a random permutation additionally makes the joint assignment exchangeable.*

Proof. After scaling the circle by n, an interval has length np and the grid is a translated integer lattice. The number of integers in a half-open interval of length np is one of its floor and ceiling; wrapping around the circle does not change this conclusion. Each of the n shifted points has a uniform marginal, so the expected count is np. If this mean is nonintegral, the two possible counts and their mean uniquely determine their probabilities; if it is integral, the count is deterministic. These statements hold for all child CDF intervals in the same grid realization.

At a fixed public history, q_v, the prefix groups, and all selection actions are fixed. PBD draws a fresh uniform shift and permutation. Each assigned coordinate is uniform conditional on that history, so inverse-CDF mapping has conditional law q_v for every named lane. Applying this at each group and next synchronized boundary proves Eq. (1). Nothing here asserts conditional validity after part of the *same* joint draw has already been exposed. □

Proof of Lemma 1. At a public history let G_v contain n named lanes and let the designated child be c with probability p. Necessity follows from

$$\mathbb{E}[M_c \mid \mathcal{F}_t] = \sum_{r \in G_v} \mathbb{P}(V_{t+1}^r = c \mid \mathcal{F}_t) = np.$$

For sufficiency, let μ be any law on $\{0, \dots, n\}$ with mean np. Draw $M \sim \mu$, select a uniform M-subset of G_v, and send that subset to c. When $0 < p < 1$, send each other lane to a child $a \neq c$ with probability $q_v(a)/(1-p)$. For every named r, its inclusion probability is $\mathbb{E}M/n = p$, and its probability of $a \neq c$ is $(1-p)q_v(a)/(1-p) = q_v(a)$. Other prefix groups may use any valid kernel. The cases $n = 0$ or $p \in \{0, 1\}$ are forced and immediate. Thus the specified scalar count law is feasible. The construction need not prescribe arbitrary laws for the other sibling counts simultaneously; that stronger claim is neither used nor true in general. □

A.3 Universal optimum and the Bellman gap identity

For $p \in [0, 1]$, define the adjacent expectation operator

$$(A_p f)(n) = \bar{f}(np) = (1 - \{np\})f(\lfloor np \rfloor) + \{np\}f(\lceil np \rceil). \tag{14}$$

The definition at integral np simply returns $f(np)$. The map is positive and linear in f, and it preserves discrete convexity: it is the integer restriction of the convex function $x \mapsto \bar{f}(px)$.

Proof of Theorem 1. Fix z, its path probabilities, and a terminal discretely convex cost ϕ. Set $V_d = \phi$ and $V_{t-1} = A_{p_t} V_t$. All V_t are discretely convex. We prove a history-uniform assertion: at any public history after round t, every admissible continuation has conditional expected terminal cost at least $V_t(N_t)$.

At $t = d$ this is equality. If the assertion holds at t, then at every history one round earlier,

$$\begin{aligned}
\mathbb{E}[\phi(N_d) \mid \mathcal{F}_{t-1}] &\geq \mathbb{E}[V_t(N_t) \mid \mathcal{F}_{t-1}] \\
&= \mathbb{E}[\bar{V}_t(N_t) \mid \mathcal{F}_{t-1}] \\
&\geq \bar{V}_t\big(\mathbb{E}[N_t \mid \mathcal{F}_{t-1}]\big) \\
&= \bar{V}_t(p_t N_{t-1}) = V_{t-1}(N_{t-1}).
\end{aligned}$$

The first inequality holds separately after every realized next public history. It is thus unaffected by lane identities, off-path configurations, or extra public side information. The second uses conditional Jensen, and the last uses Eq. (12). At the root the lower bound is $V_0(K)$.

An adjacent child-count law attains equality in the interpolation/Jensen step. Lemma 2 realizes this law at every child at once. Using the same construction recursively attains every remaining value V_t, so $\mathbb{E}\phi(N_z^{\mathrm{PBD}}) = V_0(K)$. All competing occupancies have mean Ks_z; the inequality for every convex ϕ is therefore exactly Eq. (4). Neither the PBD construction nor its child ordering needs to know z or ϕ. One run is consequently optimal for every node individually, even though their joint correlations are not uniquely specified. □

Proposition 1 (Exact excess-cost decomposition)**.** *For the above values and any admissible unpruned coupling, the quantity*

$$\delta_t = \mathbb{E}[V_t(N_t) \mid \mathcal{F}_{t-1}] - V_{t-1}(N_{t-1})$$

is nonnegative almost surely, and $\mathbb{E}\phi(N_d) - V_0(K) = \sum_{t=1}^{d} \mathbb{E}\delta_t$.

Proof. Nonnegativity is precisely conditional Jensen with the prescribed count mean. The tower property makes the sum telescope in expectation. This proves Eq. (6). For a single cost, zero gap can occur without adjacent rounding on an interval where the relevant value function is affine. The all-node variance characterization below is what makes local adjacency necessary, not equality for one possibly affine Bellman cost. □

A.4 NECESSITY: ONE AGGREGATE VARIANCE OBJECTIVE IDENTIFIES ALL OPTIMIZERS

Lemma 3 (Equality of second moments in integer convex order)**.** *If $X \preceq_{\mathrm{cx}} Y$ take values in $\{0, \dots, K\}$ and $\mathbb{E}X^2 = \mathbb{E}Y^2$, then X and Y have the same law.*

Proof. The means are equal. In Eq. (13) for $f(n) = n^2$, every stop-loss coefficient equals 2. Convex order makes every difference $\mathbb{E}(Y-a)_+ - \mathbb{E}(X-a)_+$ nonnegative, while second-moment equality forces their sum to be zero. They therefore vanish individually for $a = 1, \dots, K-1$. The thresholds $a = 0, K$ agree by the common mean and bounded support. The identity $\mathbb{E}(X-a)_+ - \mathbb{E}(X-a-1)_+ = \mathbb{P}(X \geq a+1)$ recovers all tail probabilities, and hence the distribution. □

Proof of Theorem 2. By Theorem 1, each variance is at least the corresponding PBD variance, and PBD attains all of them. Strict positivity of every w_z therefore makes (ii) equivalent to equality of the variances at every nonroot node. Since all means agree, Lemma 3 makes this equivalent to (i).

Assume (i). Consider a parent v at depth t and a child c of probability p. Conditional on $\mathcal{F}_t$, let $n = N_v$, $m = N_c$, and $a = \lfloor np \rfloor$. For every integer m, $(m-a)(m-a-1) \geq 0$. Moreover,

$$\mathbb{E}[(m-a)(m-a-1) \mid \mathcal{F}_t] = \mathbb{E}[m^2 \mid \mathcal{F}_t] - G(np), \qquad G(x) = (2\lfloor x \rfloor + 1)x - \lfloor x \rfloor (\lfloor x \rfloor + 1). \tag{15}$$

Here $G(x)$ is the second moment of $\mathrm{Adj}(x)$. Under PBD, conditional adjacency gives $\mathbb{E}[(N_c^{\mathrm{PBD}})^2] = \mathbb{E}G(pN_v^{\mathrm{PBD}})$. Under (i), the parent law and the child's second moment are the same, so the expectation of the nonnegative left side of Eq. (15) is zero. Thus its conditional expectation is zero almost surely, and $m \in \{a, a+1\}$ almost surely. If np is integral, the additional conditional mean $\mathbb{E}m = np = a$ forces $m = a$; otherwise this is exactly the adjacent pair $\{\lfloor np \rfloor, \lceil np \rceil\}$. The statement holds at every reached parent and child, proving (iii).

Finally, assume (iii). Conditional on any full history, support on the adjacent pair and the required mean determine the child's scalar law uniquely as $\mathrm{Adj}(pN_v)$. Inducting down a root-to-node path therefore yields exactly the same count law as PBD, regardless of how siblings or distinct prefix groups are coupled. This proves (i) and completes the equivalence. □

Scope of the necessity statement. Every nonroot node has positive weight, not merely the leaves. Optimizing only some counts need not force adjacent rounding elsewhere. The theorem characterizes conditional *counts*, not the particular shifted-grid random seed, lane permutation, child order, or inter-group copula. Such choices can change nonseparable objectives without changing any individual prefix-count law.

A.5 SEPARABLE OBJECTIVES AND A SHARP NONSEPARABLE BOUNDARY

Corollary 1 (Separable occupancy objectives)**.** *For fixed finite $\mathcal{Z} \subseteq \mathcal{T}$, $w_z \geq 0$, and concave u_z, any locally balanced optimizer maximizes $\mathbb{E}\sum_{z\in\mathcal{Z}} w_z u_z(N_z)$. Convex u_z give minimization.*

Proof. Apply Theorem 1 to $-u_z$ or u_z, multiply by w_z, and sum. The same process attains every term. No independence between counts is required. □

Coverage $u(n) = \mathbf{1}\{n > 0\}$ and capped coverage $u(n) = \min\{n, b\}$ are discretely concave. On a fixed leaf partition, empirical entropy is the sum of $u(n) = -(n/K)\log(n/K)$, with $0 \log 0 = 0$. The collision cost $n(n-1)/2$ and centered square $(n/K - s_z)^2$ are convex. For arbitrary weights on a *linear combination* of counts, variance includes cross-covariances and is not covered by the separable-variance statement.

Proposition 2 (Optimal node counts do not optimize arbitrary semantic unions)**.** *There exist two round-causal universally occupancy-optimal samplers with semantic-union hit probabilities $1/2$ and 1 for the same event, whereas IID has hit probability $3/4$.*

Proof. Use a one-step tree with four equally likely leaves a, b, c, d and $K = 2$. Sampler I chooses the unordered pair $\{a, c\}$ or $\{b, d\}$ equiprobably; sampler II chooses $\{a, b\}$ or $\{c, d\}$ equiprobably. Each independently permutes the pair across the named lanes. Each lane is uniform, and every leaf has an adjacent Bernoulli count with mean $1/2$. By Theorem 2, both optimize every individual

count. For $C = \{a, c\}$, sampler I hits only in its first case, whereas sampler II always hits. IID hits with probability $1 - (1 - 1/2)^2 = 3/4$. Sampler I is exactly a shifted systematic grid with child order a, b, c, d. This is an analytical counterexample delimiting the theorem, not a real-model experiment. □

A.6 Depth hierarchy on a fixed factorization

Define the IID thinning operator

$$(B_p f)(n) = \mathbb{E} f(X_n), \qquad X_n \sim \mathrm{Bin}(n, p). \tag{16}$$

Lemma 4 (Positive convexity-preserving operators)**.** *Both A_p and B_p are positive linear operators preserving discrete convexity, and $A_p f(n) \le B_p f(n)$ for every discretely convex f.*

Proof. The properties of A_p follow from interpolation. For B_p, couple $X_{n+1} = X_n + Z$ with independent $Z \sim \mathrm{Bernoulli}(p)$. Then

$$\Delta B_p f(n) = p\mathbb{E}\Delta f(X_n), \qquad \Delta^2 B_p f(n) = p^2 \mathbb{E}\Delta^2 f(X_n) \ge 0.$$

Finally Jensen gives $B_p f(n) = \mathbb{E}\bar{f}(X_n) \ge \bar{f}(np) = A_p f(n)$. Positivity and linearity are immediate from both expectation representations. □

Proof of Theorem 3. For a fixed path and a deterministic choice of balanced versus IID transitions, its terminal cost is the corresponding composition of A_{p_t} and B_{p_t} applied backwards to ϕ. Each suffix composition is discretely convex. Replacing one B_p with A_p decreases this suffix pointwise by Lemma 4; all earlier positive operators preserve the inequality. Repeated replacement proves the entire chain of expectations. Their common mean is Ks_z, so this is Eq. (7). The same argument applies to any deterministic nested set of balanced transitions, including node-specific choices. The construction is target-free, and all fixed separable consequences follow by summation. □

A.7 Factorization refinement and the price of public revelation

Proof of Theorem 4. Fix n and convex f. Start with n lanes and perform adjacent thinning through $p_1, \ldots, p_m$. Its final count M is integer-valued and has mean $n \prod_j p_j = np$, by iterated conditional expectation. Hence

$$(A_{p_1} \cdots A_{p_m} f)(n) = \mathbb{E}\bar{f}(M) \ge \bar{f}(np) = (A_p f)(n). \tag{17}$$

In a larger path, apply this inequality to the convex continuation value after the replaced block. Earlier positive operators preserve the inequality. Both endpoint means agree, giving the coarse-to-refined convex order. For every integer count with mean Ks_z, adjacent rounding at that mean is convex-order minimal by the same Jensen argument, proving the first order. The refined tree is still a fixed decoder, so Theorem 1 compares its PBD count with the IID $\mathrm{Bin}(K, s_z)$ count. Applying the result to $\mathbf{1}\{n = 0\}$ gives the coverage-price inequality. □

This result compares factorizations with exactly the same terminal-event mass. It does not compare arbitrary tokenizers whose endpoint distributions differ, and it does not say every deeper model is worse. A deterministic probability-one refinement has no effect. The interpretation is that requiring validity after more intermediate public revelations can impose additional constraints even when the final marginal is unchanged.

Strict example. For $K = 2$ and path $(3/4, 3/4)$, the first balanced count is 1 or 2 with equal probability. Their second-step hit probabilities are $3/4$ and 1, so the overall hit probability is $7/8$. IID gives $1 - (1 - 9/16)^2 = 207/256$. Collapsing the path to one edge of probability $9/16$ gives $\mathrm{Adj}(9/8)$ and hence certain coverage. This proves that all three bounds can be distinct.

A.8 CONDITIONAL OPTIMALITY AFTER A COMMON INSPECTED SELECTION

Proof of Theorem 5. Condition on $\mathcal{F}_\tau$ and the selected set. At a synchronized boundary, nonterminal selected prefixes have the same depth, and already completed leaves are absorbing. Distinct occupied boundary nodes therefore form a cut: none is a strict ancestor of another. Let n_v be the known retained count at boundary node v.

For any strict descendant z of v, only these n_v lanes can reach z. Their future conditional transitions satisfy Eq. (1), even if coupled to other retained groups and any permitted off-path information. The history-uniform induction in Theorem 1 applies with initial count n_v on this subtree. Fresh PBD simultaneously attains every such conditional bound. For a boundary or ancestral node the selected count is already fixed, and completed retained leaves contribute constants. Thus every relevant future leaf or descendant count has the asserted conditional convex-order minimum.

Conditionally fixed nonnegative weights and concave utilities may now depend on $\mathcal{F}_\tau$, because they are fixed for the conditional problem. In particular, the retained budget B is fixed after conditioning, so empirical entropy can use normalization by B when $B > 0$. Sum the conditional inequalities to obtain target-hit, distinct-leaf, entropy, and collision consequences. If two methods have exactly the same preselection law and rule, integrating against this common law preserves the inequality. The depth-operator argument also applies on each restarted path: additional balanced suffix transitions cannot hurt its separable utilities. □

Common-state comparison. This is not a theorem about the best initial coupling for an arbitrary controller. Different preselection histories can induce different retained states and even different conditional means. Future repeated pruning may also make the objective nonseparable. Merely knowing that two methods are round-causal is insufficient to order such policies. The adaptive comparison in the paper does satisfy the common-history condition for full PBD versus PBD_4: both balance through the observed depth, use the same frozen controller, and perform no subsequent selection.

A.9 PATHWISE DISCREPANCY, EXACT VARIANCE, AND SURE COVERAGE

Under local balance, define $\xi_t = N_t - p_t N_{t-1}$ and $\rho_t = \{p_t N_{t-1}\}$. Then $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0$, $|\xi_t| < 1$, and $\mathbb{E}[\xi_t^2 \mid \mathcal{F}_{t-1}] = \rho_t(1-\rho_t)$. Iteration yields

$$N_d - K s_d = \sum_{j=1}^{d} a_j \xi_j, \qquad a_j = \prod_{\ell=j+1}^{d} p_\ell, \qquad |N_d - K s_d| \le \sum_{j=1}^{d} a_j \le d. \tag{18}$$

The martingale differences are orthogonal: for $i < j$, $\mathbb{E}\xi_i\xi_j = \mathbb{E}[\xi_i \mathbb{E}(\xi_j \mid \mathcal{F}_{j-1})] = 0$. Hence

$$\mathrm{Var}(N_d) = \sum_{j=1}^{d} a_j^2 \mathbb{E}[\rho_j(1-\rho_j)] \le \frac{1}{4}\sum_{j=1}^{d} a_j^2 \le \frac{d}{4}. \tag{19}$$

Dividing by K^2 and using the common mean gives the first bound in Eq. (10); convex order against IID gives the second. The exact identity is stronger than the depth-only bound, particularly when downstream probabilities attenuate early rounding. It holds for every optimizer in Theorem 2, without independence of different groups.

Proposition 3 (Sharp sure-coverage threshold)**.** *For a positive-probability path, define*

$$R_{d+1} = 1, \qquad R_t = \lceil R_{t+1}/p_t \rceil, \quad t = d, \dots, 1. \tag{20}$$

The minimum initial batch size for guaranteed hit under any round-causal coupling is exactly R_1; PBD attains it. If some $p_t = 0$, no finite batch can guarantee a hit.

Proof. With n incoming lanes, the minimum supported adjacent child count is $\lfloor np_t \rfloor$, occurring with positive probability unless it is the only count. The condition $\lfloor np_t \rfloor \ge R_{t+1}$ is equivalent to $n \ge \lceil R_{t+1}/p_t \rceil$. Backward induction proves that R_t is the smallest incoming count from which PBD succeeds surely from stage t. Below it, a supported deficient child count has a positive conditional failure probability downstream, so failure has positive probability. Finally, $\mathbf{1}\{n = 0\}$ is discretely convex, and PBD minimizes its expectation by Theorem 1. If PBD can fail at that batch size, no admissible coupling can have zero failure probability. □

A.10 LAMINAR DECISION-FIRST FACTORIZATION

Theorem 6 (Hierarchy alignment)**.** *Let P be a finite output law and let $\mathcal{H}$ be a finite laminar family of output subsets (two members are nested or disjoint). There is a decoder that first traverses the corresponding decision hierarchy and then realizes a surface output, with exactly the same terminal law P. If every named hierarchy node occurs by depth h, then PBD_h has the full optimum for all hierarchy-node occupancies under this refactored public history. Every top-level class C attains*

$$\mathbb{P}(N_C > 0) = \min\{1, KP(C)\}. \tag{21}$$

Proof. Add the universe as a root. At each hierarchy node A, its maximal strict members in $\mathcal{H}$ are disjoint. Together with their residual complement in A, they partition A. Omit empty or zero-mass children and assign a child B probability $P(B)/P(A)$. At a final hierarchy atom, draw a surface output according to $P(\cdot \mid A)$. Multiplication of the hierarchy probabilities telescopes, giving exactly P for each terminal output.

Every named hierarchy node is reached within h transitions, so PBD_h already has the full-PBD count law there. Theorem 1 applies to the refactored tree. Top-level C is one categorical transition of probability $P(C)$, hence $N_C \sim \mathrm{Adj}(KP(C))$, yielding the claimed hit probability simultaneously for all top-level classes. Computing these hierarchy probabilities is an assumption of the construction; changing a language model's textual prompt alone need not implement it. □

A.11 LOCALLY FINITE INFINITE TREES

Corollary 2 (Finite-depth extension)**.** *On a locally finite, possibly infinite decoder tree, define a node's occupancy by the number of lanes visiting it, without requiring eventual termination. Every fixed finite-depth node has the same universal PBD occupancy optimum. The results also hold for any fixed finite set of such nodes and fixed separable sums.*

Proof. Truncate at a finite depth D and make the boundary absorbing. Local finiteness makes the truncated tree finite. Restricting any original round-causal process to these rounds preserves validity; restricting PBD gives exactly the truncated PBD process. Apply the finite-tree theorem and choose D at least as large as the deepest queried node. This makes no assertion about unbounded terminal costs or infinite sums without additional integrability assumptions. □

B EXPERIMENTAL PROTOCOL AND COMPLETE RESULTS

B.1 SCORING AND EVALUATION PROTOCOL

The three frozen checkpoints are evaluated in bfloat16 on one NVIDIA A100. The supplementary protocol document records exact model and dataset revisions, software versions, prompts, scoring timings, and numerical tolerances.

Posterior construction and temperature. Each prompt lists the complete canonical intent inventory and requests exactly one label. For each candidate label plus EOS, the scorer divides the teacher-forced token logits by T before softmax, sums token log-probabilities, and normalizes the candidate weights over the inventory, as in Eq. (11). Every coupling and baseline receives the same finite-decoder posterior. Temperature-specific scores are stored for $T \in \{0.7, 1, 1.3\}$; the prespecified robustness grid uses $K \in \{2, 4, 8, 16\}$.

The primary CLINC set contains all 30 examples from each in-scope class and 30 OOS examples. The independent replication defines its OOS subset by a predeclared SHA-256 hash-ranking rule on the 1,000 official OOS queries and also evaluates the complete 5,500-item test set. Banking77 uses its complete official test split. No held-out result was used to choose prompts, models, temperatures, batch sizes, or method variants.

Scorer and numerical validation. For every candidate label at $T = 1$ on one held-out prompt per cell, the packed trie scorer is checked against an independently masked single branch in a shape-aligned padded GPU tensor. This checks branch isolation and position handling at the same tensor

Table 7: **The independent A100 rerun closely reproduces both frozen model-family aggregates.** Values are PBD–IID percentage points with paired 95% intervals; the final column is rerun minus frozen.

Family	Frozen	A100 rerun	Difference
Qwen (4 cells)	+0.986 [+0.921, +1.052]	+0.973 [+0.909, +1.039]	-0.013
Φ-3.5 (2 cells)	+1.411 [+1.301, +1.521]	+1.401 [+1.294, +1.509]	-0.010

Table 8: **Most of the static gain is captured by the first few balanced levels.** Pass@8 gain over IID, in percentage points. Full PBD and $L = 4$ agree at the displayed precision in the three CLINC cells; display equality is not a claim that every token path has ended.

		Pass@8 gain over IID (pp)			
Task	Model	$L = 1$	$L = 2$	$L = 4$	Full
B77	Φ-3.5	+0.929	+1.203	+1.226	+1.309
B77	Qwen-3B	+0.674	+0.828	+0.933	+0.936
B77	Qwen-7B	+0.685	+0.834	+0.966	+0.966
CLINC	Φ-3.5	+1.229	+1.384	+1.492	+1.492
CLINC	Qwen-3B	+1.023	+1.214	+1.246	+1.246
CLINC	Qwen-7B	+0.529	+0.702	+0.744	+0.744

shape. The maximum absolute log-probability discrepancy is 2.29×10^{-5}, below the 2×10^{-3} tolerance. Exact finite-decoder evaluation means analytic count recursion or integration rather than simulation of completed batches; the implementation uses float64 arithmetic. Separate exact-rational checks verify the algebra independently of floating-point tolerances.

B.2 Frozen evidence and independent replication

The original Qwen analysis prespecified a $+1.000$-point practical threshold. Its $+0.986$-point estimate was 0.014 points below that threshold; the independent replication and separate Phi evaluation retain this original criterion rather than redefining it.

B.3 Depth-limited results and mechanism diagnostics

The pooled prompt-level diagnostics relate the residual factorization price to target mass, path length, branching depth, and the minimum conditional probability on the target path. Their Spearman correlations are respectively -0.126, $+0.055$, $+0.105$, and -0.125. These correlations are descriptive mechanism probes, not selection criteria or multiplicity-adjusted discoveries.

B.4 Target-free occupancy objectives

Table 2 in the main text reports the complete cell-wise changes; Table 9 resolves the prefix-depth pattern.

The equal-cell mean changes are $+0.0659$ expected distinct outputs, $+0.02236$ empirical entropy, 0.36596 fewer pair collisions, and 0.73192 lower summed count variance. The endpoint-aligned coupling is slightly better on these quantities but is not round-causally valid in the original trie factorization.

B.5 Temperature, batch size, and complete-test robustness

Table 3 in the main text reports the complete CLINC test results.

Table 9: **Prefix diversity improves most at shallow depths and remains nonnegative throughout.** Expected distinct positive-mass prefixes by token depth, averaged equally across the primary cells with nodes at that depth, at $T = 1, K = 8$.

Depth	IID	PBD	Δ
1	1.279	1.329	+0.050
2	1.344	1.406	+0.061
3	1.214	1.271	+0.057
4	0.688	0.723	+0.035
5	0.327	0.345	+0.018
6	0.145	0.154	+0.009
7	0.055	0.058	+0.003
8	0.013	0.014	+0.001
9	0.014	0.015	+0.000
10	0.011	0.011	+0.000
11	0.011	0.011	+0.000
12	0.011	0.011	+0.000

Table 10: **Positive pass gains across every prespecified temperature–batch-size setting.** Equal-cell mean PBD–IID pass@K gain in percentage points, computed from the same frozen model outputs using Eq. (11).

$T\backslash K$	2	4	8	16
0.7	+0.792	+0.835	+0.822	+0.785
1.0	+1.132	+1.162	+1.116	+1.099
1.3	+1.462	+1.470	+1.423	+1.360

B.6 Contract boundaries and utility references

IID, PBD, and PBD_L satisfy the round-causal named-lane contract. The global endpoint lattice applies one randomized systematic categorical grid directly to the terminal inventory and is an endpoint-valid ceiling reference. A sample assigned to a fixed complete-path stratum is not individually drawn from P; uniformly randomizing stratum assignment can restore endpoint marginals, but not the public-history conditional contract in general. Gumbel top-K and deterministic top-K also change the named marginal. These utility references complement the round-causal comparisons rather than sharing their feasible set. Every reference uses the same frozen posterior.

Table 11: **Utility references answer different validity contracts.** Static pass@8 probabilities at the primary operating point. "Aligned" is the endpoint-marginal ceiling, "strata" is the complete-path stratification reference, and the two top-K methods do not preserve each named sample's decoder marginal.

Task	Model	Aligned	Strata	Gumbel top-K	Top-K
B77	Φ-3.5	0.522	0.521	0.764	0.767
B77	Qwen-3B	0.600	0.599	0.827	0.831
B77	Qwen-7B	0.703	0.703	0.868	0.869
CLINC	Φ-3.5	0.699	0.698	0.873	0.878
CLINC	Qwen-3B	0.738	0.737	0.911	0.915
CLINC	Qwen-7B	0.822	0.822	0.949	0.952

C Exact computational studies

C.1 Bellman and depth-hierarchy audits

The universal-optimality audit uses two independent recursions. PBD applies adjacent interpolation. The unrestricted comparator uses Lemma 1: at each state, it enumerates every two-point count law that brackets the required mean. This suffices because the finite-dimensional feasible polytope is defined by normalization and one mean constraint. Equality for all stop-loss costs $(n-a)_+$ certifies convex order on a finite grid. The audit covers probability grid $\{0, 1/8, \ldots, 1\}$, path depths 1–3, $K = 1, \ldots, 8$, and 3,000 additional random convex costs generated from nonnegative second differences.

The independent depth audit directly composes adjacent and binomial count-law operators. It checks every adjacent-L pair on the same grid and a second set of 3,000 random convex cases with depth up to 12 and K up to 16.

These are implementation audits, not substitutes for the proofs.

C.2 Random trees

The random-tree study fixes 100 independent four-ary depth-four trees with seed 20260821 and branch laws drawn from $\mathrm{Dirichlet}(0.3, 0.3, 0.3, 0.3)$. Every entry in Table 13 is an average of exact conditional expectations.

C.3 Factorization stress tests

To isolate factorization depth, a target event of terminal mass 0.1 is split across paths with the same endpoint probability but increasing numbers of uncertain transitions. At $K = 8$, the endpoint-aligned ceiling remains 0.8 and IID remains 0.5695, while fixed-factorization PBD pays a growing causal price.

At $K = 64$, the PBD/IID prefix-frequency mean squared error (MSE) ratios for depths 1, 2, 4, and 8 are $0.0417, 0.0417, 0.0418, 0.0422$, approximately a 24-fold reduction and consistent with Eq. (10).

Table 12: **Independent finite-state audits agree with the claimed Bellman and depth identities to numerical precision.** Values are maximum absolute discrepancies or order violations.

Audit	Comparisons	Maximum
Universal stop-loss grid	232,596	8.88×10^{-16}
Universal random convex	3,000	1.78×10^{-15}
Depth stop-loss grid	103,752	8.88×10^{-16}
Depth random convex	3,000	4.44×10^{-16}

Table 13: **PBD increases exact completed-leaf diversity on every random-tree batch size.**

	Distinct		Entropy		Collisions	
K	IID	PBD	IID	PBD	IID	PBD
4	3.373	3.730	1.153	1.289	0.753	0.301
8	5.764	6.644	1.621	1.808	3.514	1.982
16	9.303	10.871	1.998	2.191	15.061	10.895
32	14.163	16.606	2.283	2.455	62.253	52.051

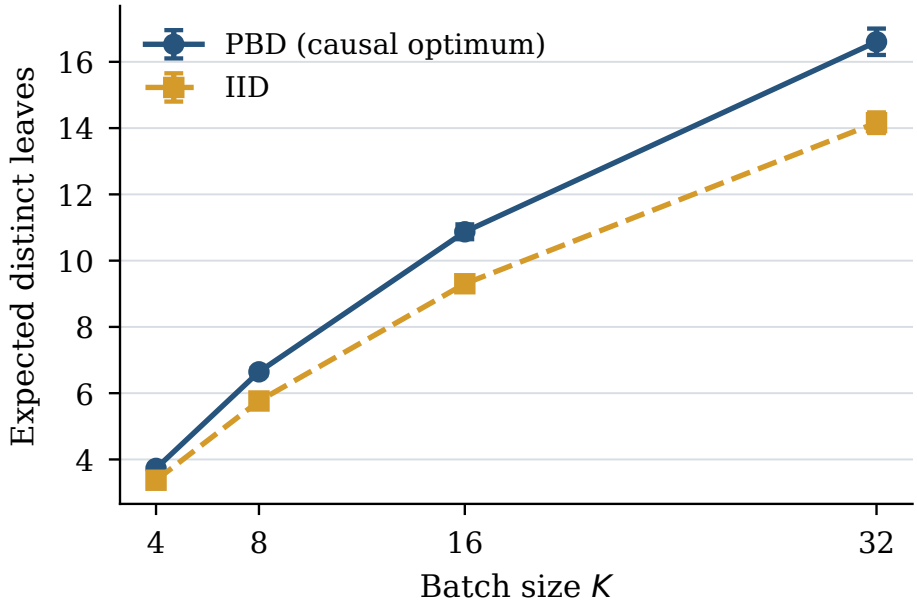

Figure 4: **The random-tree diversity advantage grows with batch size.** Error bars are standard errors across the frozen 100 trees; each tree-level value is computed exactly.

Table 14: Exact factorization-depth stress test at target mass 0.1, $K = 8$. "Price" is aligned minus PBD in percentage points; K_{sure} is the threshold from Eq. (20).

Depth	IID	PBD	Aligned	Price pp	K_{sure}
1	0.570	0.800	0.800	0.000	10
2	0.570	0.800	0.800	0.000	13
4	0.570	0.747	0.800	5.269	15
8	0.570	0.674	0.800	12.646	26

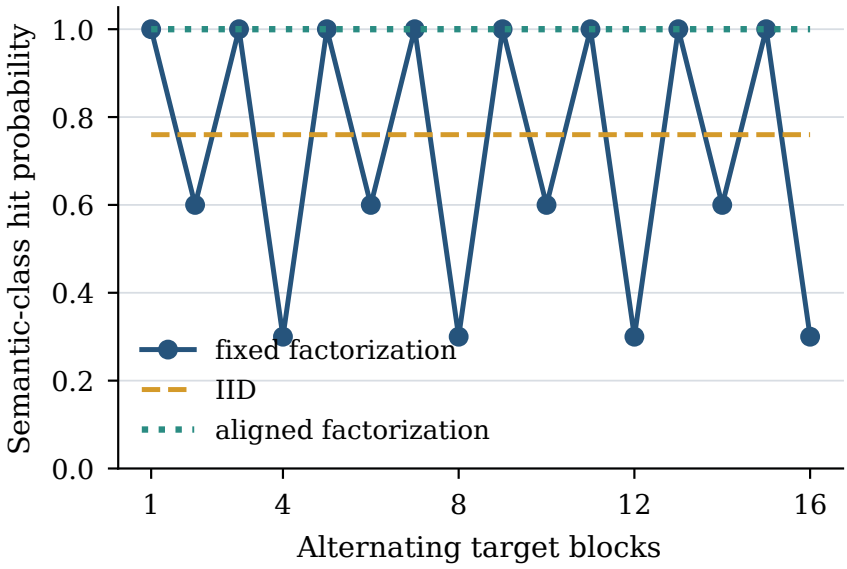

Figure 5: **Decision-first alignment removes sensitivity to surface fragmentation.** The fixed global systematic reference is endpoint-valid; its coverage of a noncontiguous semantic union can be worse than IID. This is a constructed mathematical boundary case, not an additional real-model experiment.

A separate semantic-fragmentation construction alternates target and non-target intervals while holding target mass at 0.3. With $K = 4$, IID class coverage is 0.7599; a fixed global systematic order can fall to 0.3 under fragmentation, whereas decision-first alignment reaches the endpoint-marginal ceiling 1.0 for every fragmentation pattern.

C.4 Numerical integrity

Across all A100 posterior records, the maximum normalization error is 1.11×10^{-15}, the maximum target path-product error is 1.11×10^{-15}, the maximum violation of IID $\leq$ PBD $\leq$ aligned hit probability is 1.42×10^{-14}, and the maximum discrepancy between full-depth PBD_L and full PBD is zero. A Gumbel-quadrature resolution audit over 207 records differs by at most 3.11×10^{-15}.

D Adaptive public-history consumption

This appendix specifies the common adaptive controller, uncertainty estimates, continuation validation, and complete finite-decoder results.

D.1 Design, controllers, and estimands

The primary design uses $K = 8$ named lanes at $T = 1$, observes $L_{\mathrm{obs}} = 2$ synchronized token transitions, selects exactly $B = 4$ lanes once, and completes them without further pruning. Child CDF order is ascending token order. All methods receive the same posterior, queries, initial exposure, budget, and deterministic tie rules.

The *target-free controller* prioritizes occupied exact-prefix groups by frozen subtree entropy, then subtree mass, then token-prefix order. It allocates one lane per group in this order and fills remaining slots round-robin. For the named endpoint lattice, lanes within a chosen group are retained in increasing lane-index order. Absorbing completed lanes can occupy retained slots, so B counts retained outputs, not necessarily B nonterminal continuations. The *oracle mechanism controller* allocates as many slots as possible to the prefix containing the ground-truth label. It is a target-aware

mechanism diagnostic, not a deployable verifier. The IID random-budget reference instead retains B randomly chosen lanes; it is not the primary IID controller baseline.

There are 3,080 Banking77 and 4,530 primary CLINC queries per model, or 22,830 primary model–query rows. Exact endpoint-lattice integration and six robustness cases use the predeclared SHA256-ranked subset of two queries per intent per dataset, evaluated in all three models: 1,368 model–query rows. Banking77 has 77 classes; CLINC uses 151 including out-of-scope. The subset is not selected for large gaps or successful outcomes.

Boundary occupancies use counter-based Monte Carlo with 1,024 replicates for the primary case and 256 for robustness, in eight batches. Seeds and deterministic stream construction are documented in the supplement. Continuation utilities conditional on these boundary counts and the oracle expectation are integrated exactly on the finite trie. Thus only boundary simulation, not model sampling or continuation Monte Carlo, contributes simulation noise.

Adaptive pass is the probability that at least one retained lane reaches the correct label. Distinct outputs, empirical entropy, and pair collisions are computed on the retained B completed labels. The reported count-variance quantity is the expected sum of *conditional continuation* count variances after the public history and selection. It is not the unconditional full-batch count variance of the static experiment.

Prompt and simulation uncertainty. All intervals use 10,000 paired intent-stratified bootstrap replicates with the fixed seed schedule recorded in the supplement. Within each intent, a resampled query index is shared across all three models. Dataset resamples are independent; model–dataset means receive equal weight. The same resamples are used for the numerator and denominator of each gain retention ratio. No denominator draw is nonpositive. These intervals describe query uncertainty for fixed checkpoints, not a population of trained models. The Monte Carlo diagnostics in Table 18 are separate from the prompt bootstrap. The standard-error bound of 7×10^{-5} holds for primary pass and distinct counts, not for every collision or selected-prefix metric.

D.2 Continuation validation

The full-depth analysis corrects an adaptive continuation-dispatch error: the original implementation balanced through inspection but then used IID completion. Static full-PBD results and all nonnegative-depth variants are unaffected. The correction uses identical stored posteriors, controllers, queries, and boundary draws. The supplement retains the observation-only results separately and provides the implementation details and source-to-result mapping.

Full PBD and PBD_4 share the same preselection law at the primary observation depth. Theorem 5 therefore gives an independent check on their completion ordering. Validation covers 279,342 scalar comparisons over 31,038 query–cases; replay of the observation-only calculation differs by at most 1.11×10^{-16}. The maximum violation of the full-versus-four-level ordering is 1.62×10^{-17}.

D.3 Adaptive completion and shallow-depth retention

Tables 4 and 5 in the main text give the primary aggregate and cell-wise comparisons. Full PBD improves target-free pass by +1.09 [1.04, 1.15] pp and distinct outputs by +0.0614 [0.0598, 0.0631] relative to the same IID controller. The PBD_4 gain is $+1.08$ pp $[1.03, 1.14]$ and $+0.0609$ distinct outputs $[0.0593, 0.0626]$. At a common boundary state, extra balanced completion cannot hurt these objectives, and the results satisfy this order on every saved row up to arithmetic roundoff.

D.4 Endpoint-lattice conditional fidelity

The global lattice draws one shift before generation and maps its grid points directly through a fixed terminal-label ordering. For each observed prefix vector, exact integration partitions the shift interval at relevant terminal CDF and prefix boundaries. Conditional next-token laws and target-probability gaps are then integrated over all consistent shifts. These are exact finite-posterior conditional quantities, not nested Monte Carlo estimates.

The token-lexicographic lattice gains 0.94 pp target-free pass and 0.0548 distinct outputs. Its mean all-lane TV is 0.00236 $[0.00153, 0.00319]$; TV for diversity-selected and oracle-selected lanes is

Table 15: **Four balanced levels retain almost all of full PBD's adaptive gain.** Ratios compare gains over the same IID controller; the common prompt bootstrap preserves numerator–denominator pairing. No ratio draw has a nonpositive denominator.

Quantity	Retention (%)	Paired 95% CI
Target-free pass	99.3	[99.0, 99.6]
Oracle pass	99.0	[98.4, 99.5]
Distinct outputs	99.2	[99.0, 99.4]

Table 16: **Post-inspection fidelity and utility of two endpoint-lattice orders.** The predeclared integration subset is shared by the two terminal orders. Pass differences are pp relative to the same adaptive IID controller. TV averages next-token total variation over all K named lanes, with zero contribution from absorbing terminal lanes; calibration is the oracle-selected target-probability absolute gap.

Terminal order	Target-free pass Δ pp	Oracle pass Δ pp	Distinct Δ	Mean TV	Calibration
Token-lexicographic	+0.94	+1.02	+0.0548	0.00236	0.00143
Alphabetical-label	+0.92	+1.00	+0.0532	0.00235	0.00140

0.00245 and 0.00276, respectively. Selected-lane TV averages over the B retained lanes, again including absorbing terminals with zero discrepancy. The oracle-selected target-calibration gap is 0.00143 $[0.00068, 0.00218]$. Alphabetical label order gives mean TV 0.00235. Both are below the prespecified 0.01 materiality threshold. Full-primary and subset utility means refer to different query populations and are not a paired head-to-head comparison.

D.5 Frozen robustness cases and simulation precision

The cases vary observation depth to one or four, retained budget to two or six, and temperature to 0.7 or 1.3. The largest PBD_4 oracle gain is 1.43 pp at $T = 1.3$. Lattice TV is largest after one observed round (0.00591) and smallest after four (0.00075). Cases and subset were frozen before execution and are not used to choose the primary design.

D.6 Reproducibility

The supplement includes the fixed model probabilities, adaptive records, source implementations, validation checks, and table-generation scripts. The full analysis can be regenerated offline from the supplied finite-decoder probabilities. The resulting comparisons support positive causal-coupling gains and approximately 99% adaptive gain retention with four balanced levels under the common one-shot controller.

Table 17: **Shallow and full PBD retain positive gains throughout the robustness cases.** Target-free pass differences are pp. The last column is the lexicographic lattice's all-lane TV. The same scores and boundary draws are used across depth variants.

Case	PBD_4 pass Δ	Full pass Δ	Full distinct Δ	Lattice TV
$L_{\mathrm{obs}} = 1$	+1.07	+1.07	+0.0565	0.00591
$L_{\mathrm{obs}} = 4$	+1.03	+1.04	+0.0577	0.00075
$B = 2$	+0.91	+0.91	+0.0305	0.00236
$B = 6$	+1.10	+1.11	+0.0629	0.00236
$T = 0.7$	+0.84	+0.84	+0.0426	0.00156
$T = 1.3$	+1.20	+1.20	+0.0768	0.00354

Table 18: **Boundary Monte Carlo error is reported separately from query uncertainty.** Equal-cell mean Monte Carlo standard-error diagnostics from eight batches; the observation-only row balances through inspection and then uses IID.

Method	Pass	Distinct	Entropy	Collisions	Selected prefixes
IID	2.73e-05	6.98e-05	3.57e-05	1.79e-04	7.01e-05
PBD_1	2.36e-05	5.59e-05	2.82e-05	1.41e-04	5.62e-05
PBD_2	2.28e-05	5.34e-05	2.65e-05	1.31e-04	5.36e-05
PBD_4	2.28e-05	5.33e-05	2.65e-05	1.31e-04	5.36e-05
$\mathrm{PBD}_{\mathrm{obs}} \to \mathrm{IID}$	2.28e-05	5.34e-05	2.65e-05	1.31e-04	5.36e-05
Full PBD	2.28e-05	5.33e-05	2.65e-05	1.31e-04	5.36e-05

E FOUNDATIONS AND DYNAMIC CHARACTERIZATION

Classical foundations. For an integer-valued variable with a fixed mean, the least-dispersed law is supported on the adjacent floor and ceiling values; a uniformly shifted systematic grid realizes those counts for categorical intervals (Shaked & Shanthikumar, 2007; Gerber et al., 2019). These local facts form the basis of the dynamic characterization.

Dynamic characterization. Lemma 1 records the elementary scalar feasible set: after conditioning on an arbitrary public history, a designated-child count law is constrained only by its mean. Theorem 1 then proves that the adjacent law remains optimal under backward composition through every future decoder transition, uniformly over off-path state and history-dependent side information. Finally, one systematic categorical draw balances all siblings at once, so a single target-free process attains the optimum for every node rather than for one preselected target. Theorem 3 adds an exact nested order from IID to the full optimum. Theorem 2 strengthens sufficiency to necessity and identifies the entire optimizer class through one aggregate variance objective. Theorem 5 turns the history-uniform proof into a conditional completion guarantee, while Theorem 4 orders different public factorizations with unchanged endpoint mass.