QUOTIENT-FIRST DECODING: SHARP TRUNCATION GAPS WITHOUT SUPPORT LOSS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a truncation decoder change its output law when only the representation of probability mass changes? We analyze exact mass-preserving refinements relative to a supplied outcome quotient. We characterize exactly the outcome laws reachable by nucleus sampling without changing its support. At each fixed parent law, the fine-uniform-refinement limit simultaneously maximizes every convex f-divergence from the reference decoder. The sharp one-step total-variation (TV) supremum is $\min\{1/2, 1-\rho\}$; its horizon-H counterpart is $1 - \max\{\rho, 1/2\}^H$. We identify the mechanism under threshold-prefix compatibility: refinement thins the boundary outcome, preserves the conditional law above it, and can strictly sharpen the distribution. A binary likelihood-ratio factorization explains how this distortion propagates along autoregressive paths. Quotient-first decoding removes it by aggregating before truncation; classical conditional-preserving projection characterizes its unique faithful lift. In state-preserving experiments on three real language models and 120 held-out ARC-Challenge, GSM8K, and MMLU prompt families, Qwen3-8B reaches path TV $D_{64} = 0.185$ (95% confidence interval (CI) $[0.176, 0.195]$). Uniform refinement reaches 0.256 ($[0.245, 0.267]$) with identical path support. An exact, locally canonical byte quotient gives $D_{32} = 0.038$; quotient-first replay agrees with the reference kernel to 4.44×10^{-16}. These results explain representation dependence and establish an exact, conditional-faithful remedy for the supplied outcome quotient.

1 INTRODUCTION

A truncation decoder does more than remove unlikely outcomes: it decides which *representation atoms* count toward a threshold. If several atoms encode the same outcome, changing their granularity can change the decoded outcome law even when their total probability is unchanged. This matters whenever decoding is intended to operate over a supplied many-to-one outcome map rather than over the representation itself. Nucleus sampling, or TOP-P, retains the shortest descending prefix whose mass reaches ρ (Holtzman et al., 2020); TOP-K and MIN-P use rank and maximum-relative thresholds (Nguyen et al., 2025; Noarov et al., 2025). None is automatically invariant to mass-preserving refinement.

The key distinction extends beyond which outcomes survive. **Nucleus sampling can retain exactly the same outcome support and still induce sharply different path laws.** Refinement can make the nucleus stop inside the boundary outcome, reducing its retained fraction while retaining every higher-ranked outcome completely. Renormalization then shifts probability upward. Repeating these boundary decisions can produce nearly disjoint probability laws on an identical set of supported paths.

We isolate this effect from ordinary tokenizer sensitivity. At each shared parent-token history, virtual children partition each real token's probability, but sampling a child emits and feeds only its parent token. Thus the two decoding orders see the same transformer state and untruncated kernel *conditional on the same history*. Their sampled histories need not agree; changing their distribution is precisely the effect being measured. The tokenizer, model weights, and parent-history transition function remain unchanged.

Our theoretical contribution is an exact characterization of this effect. We characterize the entire set of laws reachable under support-preserving refinement and show that the fine-uniform-refinement

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.

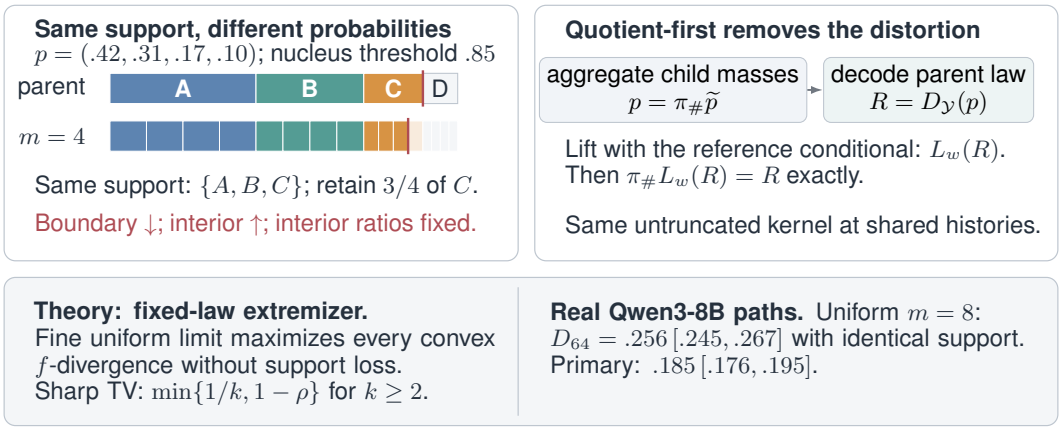

Figure 1: **Support preservation is weaker than distribution preservation.** The mathematical example (left) shows a nucleus stopping inside the same boundary outcome after uniform refinement. Renormalization changes its probability without changing which outcomes are possible. Quotient-first decoding (right) enforces the supplied outcome law. The fine-uniform limit is pointwise extremal over all same-support refinements. The bottom separates theory from real-model estimates; intervals are 95% task-stratified family-bootstrap CIs.

limit simultaneously maximizes every convex f-divergence from the reference decoder. With $k \geq 2$ retained outcomes, the same-support one-step TV supremum is $\min\{1/k, 1-\rho\}$; maximizing over k gives $\min\{1/2, 1-\rho\}$. Its finite-horizon counterpart is also sharp. Uniform refinement realizes a transparent boundary-thinning mechanism, with an exact finite-multiplicity law, majorization, and a binary reduction of both local divergences and path likelihood ratios. These statements distinguish *support preservation*, *probability preservation*, and *path-law preservation*.

The remedy is simple: aggregate to the chosen quotient, decode there, and restore the original conditional distribution within each outcome. We call this *quotient-first decoding*. This builds on aggregate–decode–condition sampling, already used in cluster truncation (Nguyen et al., 2026), and classical conditional-preserving projection (Diaconis & Zabell, 1982; Csiszár, 1975). Our contribution is the exact refinement-reachable set, its universal divergence extremizer, sharp support-constrained limits, and controlled real-model evidence, together with an explicit coherence contract for the remedy. Quotient-blind universal coherence forces full sampling; restricted invariance or a supplied quotient permits nontrivial decoding.

On the real next-token kernels of Qwen3-8B, Qwen2.5-0.5B, and OLMo-3-7B, small initial discrepancies grow into substantial path-law separation. The primary Qwen3 arm reaches $D_{64} = 0.185$; uniform refinement reaches 0.256 without any support change. Common-support reweighting accounts for 82.2% of primary terminal TV. Full-sampling and quotient-first controls close the corresponding gap, whereas an exact byte-quotient study extends noncommutation beyond virtual children.

2 Representation Coherence and the Retention Profile

Let $\mathcal{P}(\mathcal{Y})$ denote the probability laws on finite parent outcomes $\mathcal{Y}$, and let $p \in \mathcal{P}(\mathcal{Y})$. A refinement is a finite child alphabet $\mathcal{C}$, a surjection $\pi : \mathcal{C} \to \mathcal{Y}$, and positive within-fiber weights $w(c \mid y)$ summing to one. Writing $\pi_{\#}$ for pushforward, its lift is

$$L_w(p)(c) = p(\pi(c))w(c \mid \pi(c)), \qquad \pi_{\#} L_w(p) = p. \tag{1}$$

For a deterministic truncation-and-renormalization rule D, define

$$P = D_{\mathcal{Y}}(p), \qquad Q = \pi_{\#} D_{\mathcal{C}}(L_w(p)). \tag{2}$$

We use $\mathrm{TV}(P, Q) = \frac{1}{2}\sum_y |P_y - Q_y|$ and $D_f(q\|p) = \sum_i p_i f(q_i/p_i)$ for lower-semicontinuous convex f with $f(1) = 0$, using the standard zero-mass conventions. A supplied-quotient implementation T is *representation-coherent* if $\pi_{\#} T(L_w(p)) = D_{\mathcal{Y}}(p)$ for every p, and *conditional-faithful* if

its positive-mass fibers retain conditional w. Coherence is a design requirement relative to the chosen outcomes, not a claim that every application should ignore distinctions between tokens.

The *retention profile* exposes the entire one-step discrepancy. Let S be the parent retained set, A the child retained set, and

$$a_y = \sum_{c \in A \cap \pi^{-1}(y)} w(c \mid y), \quad Z = p(S), \quad C = \sum_y p_y a_y. \tag{3}$$

Proposition 2.1 (Exact commutation criterion). *For $Z, C > 0$, $P_y = p_y \mathbf{1}_S(y)/Z$ and $Q_y = p_y a_y / C$. Hence*

$$Q = P \iff a_y = (C/Z)\mathbf{1}_S(y) \quad \textit{for every } y \textit{ with } p_y > 0. \tag{4}$$

If the supports agree, $Q_y/P_y = a_y/\mathbb{E}_P[a]$ on $\operatorname{supp}(P)$.

Thus equal support requires only positive retention on the same fibers; equal laws require *equal retention fractions* across those fibers. Quotient-first decoding selects whole fibers and satisfies this condition automatically. Atom-first truncation generally does not.

3 Sharp gaps, even when every retained outcome survives

We fix stable tie orders, with equal child masses ordered by parent order and then child index. Top-p(ρ) selects the shortest descending prefix with mass at least ρ; Top-k(k) selects the k largest atoms; Min-p(λ) selects atoms with mass at least λ times the maximum. Supremum statements range over finite strictly positive laws and finite refinements, with their sizes unrestricted.

3.1 Exactly which laws can refinement produce?

Work on the positive parent support, sorted as $p_1 \geq \cdots \geq p_n > 0$, let $S_0 = 0$ and $S_j = \sum_{i \leq j} p_i$, and let $S = \{1, \ldots, k\}$ be its nucleus, with $Z = S_k$. The following result characterizes arbitrary refinements, without imposing an ordering on their child blocks.

Theorem 3.1 (Exact reachable laws and a universal extremizer). *For fixed p and $0 < \rho \leq 1$, the parent-outcome laws attainable by child-first nucleus sampling with support S are exactly*

$$\mathcal{R}_\rho(p) = \{q \in \mathcal{P}(S) : \rho q_i \leq p_i \textit{ for all } i \in S\}. \tag{5}$$

Every such q has positive mass on all of S. Define $Q_i^\infty = p_i/\rho$ for $i < k$, $Q_k^\infty = (\rho - S_{k-1})/\rho$, and zero elsewhere. For $P = \text{Top-p}_\rho(p)$ and every convex f-divergence,

$$\max_{q \in \mathcal{R}_\rho(p)} D_f(q \| P) = D_f(Q^\infty \| P). \tag{6}$$

Moreover Q^∞ majorizes every $q \in \mathcal{R}_\rho(p)$, and

$$\max_{q \in \mathcal{R}_\rho(p)} \mathrm{TV}(P, q) = \frac{S_{k-1}(S_k - \rho)}{\rho S_k}. \tag{7}$$

Uniform refinement converges to this extremizer as its multiplicity increases.

Proof idea. Every child nucleus retains mass $C \geq \rho$, so $\rho Q_i \leq p_i$. Conversely, retain one child of mass ρq_i per parent in S, and split all remaining mass into children smaller than the smallest retained child. The nucleus then retains exactly these children, of total mass ρ, and outputs q. The feasible set is a simplex of mass deficits: $\rho q = p_S - (Z - \rho)\lambda$, $\lambda \in \mathcal{P}(S)$. Convexity reduces divergence maximization to its vertices; a two-point Jensen argument selects the smallest parent as the deficit location. Details are in Section A. □

Thus uniform splitting yields a sharp limiting construction. Its limiting law is extremal among *all* support-preserving refinements at that parent distribution. This does not make every finite uniform multiplicity extremal, nor does it rank arbitrary history-dependent path laws.

3.2 IDENTICAL SUPPORT DOES NOT PROTECT THE PATH LAW

Theorem 3.2 (Sharp support-preserving nucleus gaps)**.** *Fix $0 < \rho < 1$. Among pairs in (2) for* TOP-P(ρ) *with identical parent-outcome support of size $k \geq 2$, the sharp one-step supremum is*

$$\Gamma_{\rho,k} = \min\{1/k, 1-\rho\}. \tag{8}$$

For $k = 1$ the gap is zero. Allowing any support size gives $\Gamma_\rho = \min\{1/2, 1-\rho\}$. Among state-preserving autoregressive pairs with identical next-outcome support at every shared history, the sharp horizon-$H \geq 1$ supremum is

$$\sup \mathrm{TV}(P_H, Q_H) = 1 - (1-\Gamma_\rho)^H = 1 - \max\{\rho, 1/2\}^H. \tag{9}$$

If every nonabsorbing history has support size k, replace Γ_ρ by $\Gamma_{\rho,k}$. The suprema can be approached by uniform refinements, and also by refinements with all child masses distinct. At $\rho = 1$ all gaps are zero.

Key argument. Under support equality, the retained child submeasure $s_y = p_y a_y$ is dominated by $r_y = p_y \mathbf{1}_S(y)$, with masses $C \geq \rho$ and $Z \leq 1$. Since $P_y \geq (C/Z)Q_y$,

$$\mathrm{TV}(P, Q) = 1 - \sum_y \min(P_y, Q_y) \leq 1 - C/Z \leq 1 - \rho/Z \leq 1 - \rho.$$

Sorted parent masses satisfy $S_{k-1} < \rho$ and $p_k \leq Z/k$, so $\rho/Z > (k-1)/k$, giving the second bound $\mathrm{TV}(P, Q) < 1/k$. Boundary-thinning constructions approach the smaller constant. Maximal coupling gives the horizon bound; independent repetitions of the same constructions approach it while preserving support at every finite refinement. Full constructions are in Section A. □

The theorem constrains *all* same-support refinements, not only ordered child blocks. Large retained support limits the one-step gap, but no positive local guarantee prevents long-horizon separation. At $\rho = .9$, the same-support local supremum is .1 and the horizon-64 supremum is $1 - .9^{64}$. This is a sharp worst-case statement, not a prediction for a language model.

Write $\Delta_D^* = \sup \mathrm{TV}(P, Q)$ for the unrestricted one-step gap of decoder D.

Theorem 3.3 (Unrestricted sharp gaps)**.** *Without a support constraint, for fixed $k \geq 1$, $0 < \lambda < 1$, and $0 < \rho \leq 1$,*

$$\Delta^*_{\text{TOP-K}(k)} = 1, \quad \Delta^*_{\text{MIN-P}(\lambda)} = 1, \quad \Delta^*_{\text{TOP-P}(\rho)} = \min\{1, (1-\rho)/\rho\}. \tag{10}$$

The corresponding sharp horizon-H supremum is $1 - (1-\Delta^*)^H$. All suprema can be approached without probability ties.

For $\rho > 1/2$, both nucleus orders retain submeasures of the same p with mass at least ρ; their unavoidable overlap yields the bound. TOP-K and MIN-P have no comparable cumulative-mass guarantee. These worst-case orderings need not rank realized effects: our TOP-K(50) control has a much smaller empirical gap than nucleus sampling.

3.3 BOUNDARY THINNING: A SUPPORT-PRESERVING SHARPENING MECHANISM

With the same sorted notation, a refinement is *threshold-prefix compatible* if the selected child prefix contains every child of parents $1, \ldots, k-1$, a nonempty subset of parent k's children, and none from lower-ranked parents. This condition is stronger than support equality. Equal-multiplicity uniform splitting always satisfies it under the declared tie order.

Theorem 3.4 (Boundary thinning and binary reduction)**.** *Under threshold-prefix compatibility, let $\alpha \in (0, 1]$ be the retained fraction of parent k, and $C = S_{k-1} + \alpha p_k$. Both decoded supports are $\{1, \ldots, k\}$. For $k > 1$, define*

$$b = p_k/S_k, \qquad c = \alpha p_k/C, \qquad v_i = p_i/S_{k-1}\ (i < k).$$

Then $c \leq b$, with equality iff $\alpha = 1$, and

$$P = ((1-b)v, b), \quad Q = ((1-c)v, c), \quad \mathrm{TV}(P, Q) = b - c = \frac{S_{k-1}(S_k - C)}{C S_k}. \tag{11}$$

The law Q majorizes P, so every Schur-concave entropy weakly decreases, strictly for strictly Schur-concave entropies when $\alpha < 1$. For every convex f-divergence,

$$D_f(Q\|P) = D_f(\mathrm{Ber}(c)\|\,\mathrm{Ber}(b)). \tag{12}$$

Here Ber *denotes a Bernoulli law. For $k = 1$, both laws are the same point mass.*

Only the choice between the boundary outcome and the interior changes; *conditional on the interior*, the law is identical. Thus the refinement imposes an additional binary selection bias without deleting an outcome. Majorization applies to these local kernels, not automatically to the full history-dependent path laws.

Corollary 3.5 (Uniform refinement and the sharp fixed-law limit)**.** *Let $\beta = (\rho - S_{k-1})/p_k$. Splitting every parent into m equal children gives retained mass*

$$C_m = S_{k-1} + p_k \frac{\lceil m\beta \rceil}{m}, \qquad \rho \le C_m < \rho + p_k/m. \tag{13}$$

The fixed-law maximum in Theorem 3.1, also approachable within the threshold-prefix-compatible class, is

$$\Delta^*_{\mathrm{prefix}}(p, \rho) = \frac{S_{k-1}(S_k - \rho)}{\rho S_k}. \tag{14}$$

Writing $Q_i^\infty = p_i/\rho$ for $i < k$ and $Q_k^\infty = (\rho - S_{k-1})/\rho$, with zeros below k,

$$0 \le \Delta^*_{\mathrm{prefix}} - \mathrm{TV}(P, Q^{(m)}) = \mathrm{TV}(Q^{(m)}, Q^\infty) \le \frac{S_{k-1}p_k}{m\rho^2}. \tag{15}$$

The ceiling creates a sawtooth, not monotonicity at consecutive m. Strictly ordered parent laws admit distinct-child perturbations arbitrarily close to any uniform-refinement law (Corollary A.1); exact ties are not necessary for the mechanism or the sharp suprema.

3.4 How the local distortion propagates

For a parent history $h = y_{<t}$ satisfying the preceding compatibility condition, let b_h, c_h denote its two boundary probabilities and $B_t = \mathbf{1}\{y_t = k(h)\}$. On the common path support,

$$\log \frac{Q_H(y)}{P_H(y)} = \sum_{t=1}^{H} \left[B_t \log \frac{c_h}{b_h} + (1 - B_t) \log \frac{1 - c_h}{1 - b_h} \right]. \tag{16}$$

Histories with singleton support contribute zero. Taking expectations gives the exact chain-rule specialization

$$\mathrm{KL}(P_H\|Q_H) = \mathbb{E}_{P_H} \sum_{t=1}^{H} \mathrm{kl}(b_{Y_{<t}}\|c_{Y_{<t}}), \tag{17}$$

Here KL is Kullback–Leibler divergence and kl its Bernoulli specialization. The identities require no independence or stationarity. They explain how shared-support reweighting accumulates through later boundary decisions; they do not assert additive TV or a universal empirical growth rate. Proofs, a path-affinity identity, and a sufficient condition for large path TV appear in Section A.

4 Quotient-first decoding: exact repair, classical foundation

Given the represented law $\widetilde{p} = L_w(p)$ and the supplied quotient, define

$$\mathrm{QF}_{\pi,w}(\widetilde{p}) = L_w\big(D_{\mathcal{Y}}(\pi_{\#}\widetilde{p})\big). \tag{18}$$

Sample an outcome from $D_{\mathcal{Y}}(p)$, then a child from $w(\cdot \mid y)$ if a represented output is needed. This is exactly coherent for every decoder. For an explicit finite partition, aggregation requires one pass over the represented masses; the decoder then operates on the quotient alphabet. Computing a nonlocal or semantic quotient is a separate problem.

Proposition 4.1 (Conditional-preserving characterization)**.** *Let $R \ll p$. The unique conditional-faithful child law with parent marginal R is $L_w(R)$. For every lower-semicontinuous convex f with $f(1) = 0$ and every $q \ll L_w(p)$ with $\pi_{\#}q = R$,*

$$D_f(q\|L_w(p)) \geq D_f(R\|p) = D_f(L_w(R)\|L_w(p)). \tag{19}$$

The minimizer is unique for strictly convex f when the right side is finite. A common lift is an exact divergence isometry:

$$D_f(L_w(R)\|L_w(S)) = D_f(R\|S) \quad (R \ll S). \tag{20}$$

This is the classical Jeffrey update and convex-divergence projection (Diaconis & Zabell, 1982, Theorem 6.1); the KL projection is also classical (Csiszár, 1975). It identifies the coherent implementation that preserves the within-fiber conditionals. The same argument on the finite path quotient gives path-space canonicality. Exact KL and TV conditional-cost identities, and stability under approximate aggregation, are proved in Section B. In particular, if two aggregated laws retain the same nucleus set, the parent-level implementation error is at most their input TV error divided by ρ.

Why the quotient cannot be omitted universally. If a decoder family commutes with every finite surjection, it must be full sampling: $D_{\mathcal{Y}}(p) = p$. A self-contained proof in Proposition B.3 uses only relabelings, indicator coarse-grainings, and three-to-two merges. This is an identity-only naturality phenomenon, related to classical representation-independence restrictions (Bartels, 2004; Halpern & Koller, 2004). It rules out a nontrivial *universal* quotient-blind repair, not invariance to a restricted class of refinements.

5 Controlled evaluation on real language-model kernels

5.1 Intervention, estimand, and frozen inference

At every shared parent history, virtual refinement partitions the next-token law exactly. Parent-first truncation and child-first truncation then define laws P_H, Q_H on absorbing length-H parent-token paths. We sample paths from both laws and teacher-force each through both decoders at the *same* parent histories. The untruncated model kernels agree there; only decoder order differs.

With $M_H = (P_H + Q_H)/2$ and $g_H(y) = |P_H(y) - Q_H(y)|/(P_H(y) + Q_H(y))$ on the union support,

$$D_H := \mathrm{TV}(P_H, Q_H) = \mathbb{E}_{M_H}[g_H(Y)]. \tag{21}$$

Equal allocation to the two sampling laws yields a bounded unbiased estimate without enumerating paths. Dual likelihoods are scored for the implemented kernels; on common support, $g_H = |\tanh((\log P_H - \log Q_H)/2)|$. We use established likelihood-access methodology (Onak & Sun, 2018; Bhattacharyya et al., 2024; Price et al., 2026) to measure the controlled intervention. The same support decomposition is used by Price et al. (2026, Appendix B.6): for the one-sided-support set U_H,

$$D_H = M_H(U_H) + \mathbb{E}_{M_H}[g_H(Y)\mathbf{1}\{Y \notin U_H\}]. \tag{22}$$

We estimate both terms separately. Direct vocabulary summation gives exact D_1. Parent-path TV upper-bounds TV after any deterministic rendering map; it does *not* identify decoded-string or semantic-output TV.

The frozen design has 120 label-free families, 40 each from the official test splits of ARC-Challenge, GSM8K, and MMLU (Clark et al., 2018; Cobbe et al., 2021; Hendrycks et al., 2021). Selection is deterministic and label-independent; answer labels, correctness, and metric fields are absent from model-facing files. These real benchmarks supply reasoning contexts for the distributional evaluation. Qwen3-8B in non-thinking chat mode is primary; Qwen2.5-0.5B-Instruct and OLMo-3-7B-Instruct provide smaller same-family and non-Qwen checks (Yang et al., 2025; Qwen Team et al., 2024; Team Olmo et al., 2025).

The primary configuration is temperature $T = .7$, Top-p(.9), and a fixed positive Dirichlet(1) split vector of multiplicity eight per token. Controls use uniform $m = 8$, Dirichlet $m \in \{2, 4\}$, quotient-first replay, and full sampling. Each full arm has three split maps, four generation seeds, and two independently namespaced sampling lanes: 24 paths per family, 2,880 per arm, evaluated at

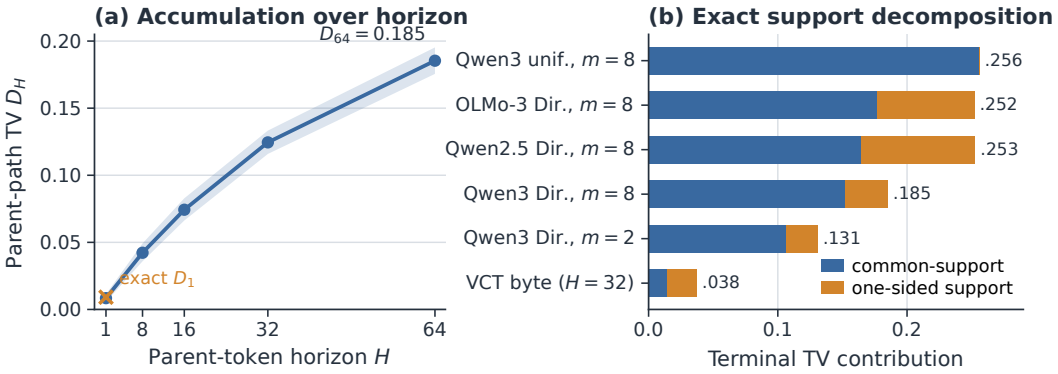

Figure 2: **Small local discrepancies grow; support deletion is not necessary.** (a) Primary Qwen3-8B Dirichlet-$m = 8$ path TV; shading is a 95% task-stratified family-bootstrap CI and the cross is exact D_1. (b) Estimated contributions in the exact decomposition (22). Parent rows use $H = 64$; the valid-covering-tree (VCT) arm uses $H = 32$ bytes. Uniform refinement has zero one-sided-support contribution in exact arithmetic; its recorded contribution is zero.

$H \in \{1, 8, 16, 32, 64\}$. Mass preservation, only real parent IDs entering the model, identical replay histories, pinned revisions, deterministic CUDA, and batch size one are run invariants.

The estimand averages conditional path TVs over families and fixed split maps, rather than taking TV after mixing these conditions. Paths and split maps are averaged within family. Two-sided 95% intervals use 20,000 task-stratified family-bootstrap resamples and equal task weighting (Efron & Tibshirani, 1994). The primary ordered gates require the lower endpoint of D_{64} to exceed .10, then that of $D_{64} - D_1^{\text{exact}}$ to exceed .05. A fixed label-blind 60-family decoder panel is descriptive. All intervals condition on the frozen checkpoints and split maps; families are not a probability sample of language use.

5.2 Results: Support Preservation, Late Divergence, and Calibrated Scope

Primary and cross-model effects. Qwen3-8B reaches $D_{64} = 0.185$ ($[0.176, 0.195]$), from exact $D_1 = 0.009$ ($[0.006, 0.012]$). The contrast is 0.176 ($[0.167, 0.186]$); both ordered gates pass. This is not additive attribution: a common continuation kernel could not increase TV, so the terminal gap requires differences in later transition kernels. Qwen2.5-0.5B and OLMo-3-7B reach 0.253 ($[0.238, 0.267]$) and 0.252 ($[0.240, 0.265]$). Primary task means are .174, .208, and .174; split-map means range from .178 to .198 (Tables 1 and 2).

The support-preserving mechanism is sufficient. The primary arm decomposes into one-sided mass .033 ($[.026, .041]$) and common-support reweighting .152 ($[.146, .159]$), or 82.2% of total TV. This decomposition alone does not imply threshold-prefix compatibility for Dirichlet refinements. In exact arithmetic, uniform $m = 8$ satisfies compatibility at every history and directly instantiates Theorem 3.4 and Equation (16). Its limiting family is pointwise extremal by Theorem 3.1; finite $m = 8$ is not claimed to attain that limit. It reaches $D_{64} = 0.256$ ($[0.245, 0.267]$), the largest Qwen3 endpoint, with *identical path support*. Its task means are .245, .286, and .237; independently namespaced split-cell estimates range from .250 to .261 despite identical uniform weights. Random split imbalance and support deletion are therefore unnecessary for a substantial effect.

Initial agreement does not imply trajectory agreement. The post-hoc diagnostic records 92 uniform-arm families with initial gaps recorded as zero, yet their mean terminal estimate is .248; the corresponding primary-arm figures are 89 families and .175. These are descriptive subgroup summaries, with no additional gate or claimed confidence interval (Table 10). They isolate later-kernel effects up to the numerical zero convention of the recorded initial gaps.

Magnitude depends on the operating point. Qwen3 Dirichlet terminal TV is .131, .169, and .185 at $m = 2, 4, 8$. Paired differences are .038 ($[.031, .045]$) for $m4 - m2$, .016 ($[.009, .024]$) for

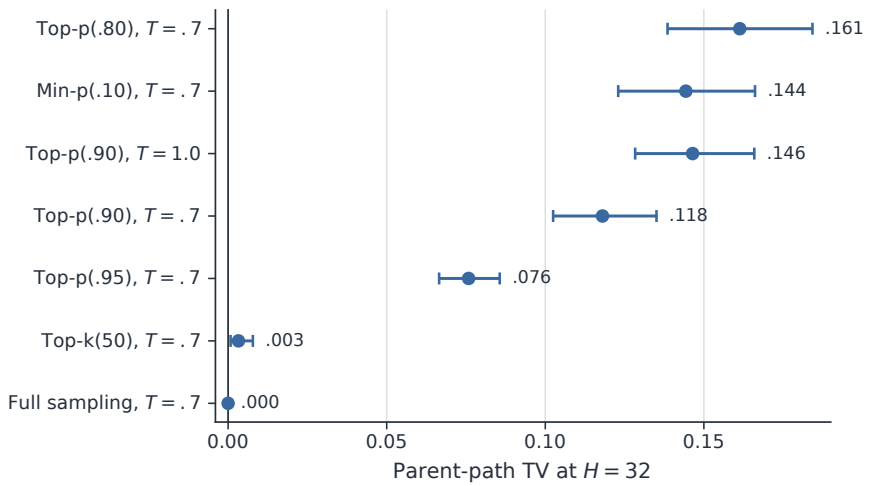

Figure 3: **Realized separation is largest under stronger truncation and vanishes under full sampling.** Points are D_{32}; bars are two-sided 95% task-stratified family-bootstrap intervals on the fixed 60-family Qwen3-8B panel. The panel is descriptive rather than a second confirmatory family.

Table 1: Terminal TV by task stratum. Parent rows use $H = 64$; the VCT row uses $H = 32$. The final column is the equally weighted three-task mean used in the main analysis.

Arm	ARC-C	GSM8K	MMLU	Equal-task
Qwen3 Dirichlet $m = 8$	.174	.208	.174	.185
Qwen3 uniform $m = 8$	.245	.286	.237	.256
Qwen3 Dirichlet $m = 4$	.165	.191	.151	.169
Qwen3 Dirichlet $m = 2$	.120	.146	.127	.131
Qwen2.5 Dirichlet $m = 8$	.287	.265	.207	.253
OLMo-3 Dirichlet $m = 8$	.251	.235	.270	.252
Qwen2.5 VCT byte	.056	.014	.043	.038

$m8 - m4$, and .054 ($[.047, .062]$) for $m8 - m2$. These fixed maps are non-nested, so this is not a universal multiplicity law. The theorem's universal statement is the $O(1/m)$ envelope for compatible refinements. On the fixed decoder panel, D_{32} is .118 at TOP-P(.9), .161 at .8, and .076 at .95. Temperature 1.0 gives .146, MIN-P(.1) gives .144, and TOP-K(50) gives .003. Every arm, including the full-sampling numerical null, is reported in Table 7 and Figure 3.

Repair and an application-defined quotient. All 13 parent/refinement arms pass the declared exact-D_1 calibration check. Quotient-first replay matches the reference parent kernel with maximum absolute residual 4.44×10^{-16}; full sampling has mean terminal TV 4.9×10^{-16}.

The byte study complements virtual refinement. Building on Hayase et al. (2026, Appendix F), which already contrasts token- and byte-level truncation, our Qwen2.5-0.5B study uses their exact valid-covering-tree (VCT) construction: locally canonical token continuations are aggregated into 256 next-byte outcomes and one shared special-stop outcome. At $T = .7$, $\rho = .9$, it compares token-level TOP-P followed by exact byte marginalization against exact byte marginalization followed by byte-level TOP-P. On these declared kernels, $D_{32} = 0.038$ ($[0.027, 0.049]$), and $D_{32} - D_1^{\text{exact}} = 0.024$ ($[0.015, 0.033]$); both ordered .005 gates pass. Task means are .056, .014, and .043. One-sided support contributes .023 and common-support reweighting .015, so the uniform-refinement mechanism is *not* dominant here. Exactness does not extend to unrestricted token sampling or global conditioning on future validity (Phan et al., 2025; Vieira et al., 2025b;a). Byte and parent-token horizons are not directly comparable.

Table 2: Descriptive split-map sensitivity of parent-path D_{64}. Uniform refinement has the same split weights for all seeds, but generation randomness is independently namespaced by split cell, so its three displayed estimates need not be identical.

Arm	Seed 11	Seed 23	Seed 37	Range
Qwen3 Dirichlet $m = 8$	.198	.180	.178	.019
Qwen3 uniform $m = 8$	.261	.257	.250	.011
Qwen3 Dirichlet $m = 4$	.183	.162	.162	.021
Qwen3 Dirichlet $m = 2$	.134	.125	.134	.009
Qwen2.5 Dirichlet $m = 8$	.258	.245	.256	.013
OLMo-3 Dirichlet $m = 8$	.248	.259	.250	.011

6 Related work and scope

Contribution in context. Cluster truncation already performs aggregate–filter–conditional sampling to improve image diversity (Nguyen et al., 2026); coarse-grained speculative speech decoding groups tokens for acceleration (Yanuka et al., 2026). Our analysis adds the exact refinement-reachable law set, its universal divergence extremizer, and sharp support-constrained path limits. Conditional-preserving updates and representation-independence restrictions are classical (Csiszár, 1975; Diaconis & Zabell, 1982; Halpern & Koller, 2004; Bartels, 2004). These classical foundations support our explicit decoding contract.

Relation to tokenization and truncation analyses. Marginalization over tokenizations and character-, byte-, or transducer-level language models address output representations (Cao & Rimell, 2021; Chirkova et al., 2023; Phan et al., 2024; Geh et al., 2024; Chatzi et al., 2025a; Gu et al., 2024; Snæbjarnarson et al., 2026). Noncanonical tokenization and semantics-preserving rewrites can change model behavior (Zheng et al., 2025; Jain et al., 2026; Li et al., 2026); our intervention instead fixes the conditional model kernel. Desmoothing, nontermination, local/global normalization, and simplex or geometric views analyze other aspects of truncation (Hewitt et al., 2022; Finlayson et al., 2024; Welleck et al., 2020; Gareev et al., 2024; Kempton & Burrell, 2025; Ji et al., 2026; Gholami Davoodi et al., 2026). Temperature scaling also admits local majorization while autoregressive path entropy need not be monotone (Mattei & Loureiro, 2026); our distortion instead arises from representation refinement at fixed temperature. Coupled generation supports counterfactuals and evaluation (Chatzi et al., 2025b; Corvelo Benz et al., 2026).

Limitations. The quotient is supplied; constructing semantic quotients may be difficult or nonlocal. Virtual refinement isolates the decoder-order effect from retokenization. Evidence covers three checkpoints, three question-answering (QA) strata, fixed split maps, and finite horizons; VCT uses one model. Same-support bounds are worst-case; local entropy reduction need not extend to path entropy. Parent-path TV does not identify decoded-string TV or downstream utility. See Sections E and F for further scope and protocol details.

7 Conclusion

We characterize boundary thinning, establish sharp same-support gaps, and demonstrate substantial real-model path-law separation under mass-preserving refinement. Quotient-first decoding enforces invariance relative to supplied outcomes while preserving their within-fiber conditional. The resulting coherence principle preserves both *supported outcomes* and *the sampled law*.

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.

REPRODUCIBILITY STATEMENT

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

REFERENCES

Falk Bartels. *On Generalised Coinduction and Probabilistic Specification Formats: Distributive Laws in Coalgebraic Modelling*. PhD thesis, Vrije Universiteit Amsterdam, 2004. URL https://ir.cwi.nl/pub/17104.

Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, Dimitrios Myrisiotis, A. Pavan, and N. V. Vinodchandran. Total variation distance meets probabilistic inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3776–3794. PMLR, 2024. URL https://proceedings.mlr.press/v235/bhattacharyya24a.html.

Kris Cao and Laura Rimell. You should evaluate your language model on marginal likelihood over tokenisations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2104–2114, 2021. doi: 10.18653/v1/2021.emnlp-main.161. URL https://aclanthology.org/2021.emnlp-main.161/.

Ivi Chatzi, Nina Corvelo Benz, Stratis Tsirtsis, and Manuel Gomez-Rodriguez. Tokenization multiplicity leads to arbitrary price variation in LLM-as-a-service. *arXiv preprint arXiv:2506.06446*, 2025a. URL https://arxiv.org/abs/2506.06446.

Ivi Chatzi, Nina L. Corvelo Benz, Eleni Straitouri, Stratis Tsirtsis, and Manuel Gomez-Rodriguez. Counterfactual token generation in large language models. In *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pp. 1291–1315. PMLR, 2025b. URL https://proceedings.mlr.press/v275/chatzi25a.html.

Nadezhda Chirkova, Germán Kruszewski, Jos Rozen, and Marc Dymetman. Should you marginalize over possible tokenizations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–12, 2023. doi: 10.18653/v1/2023.acl-short.1. URL https://aclanthology.org/2023.acl-short.1/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

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. URL https://proceedings.mlr.press/v300/benz26a.html.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.

Persi Diaconis and Sandy L. Zabell. Updating subjective probability. *Journal of the American Statistical Association*, 77(380):822–830, 1982. doi: 10.1080/01621459.1982.10477893.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994. doi: 10.1201/9780429246593.

Matthew Finlayson, John Hewitt, Alexander Koller, Swabha Swayamdipta, and Ashish Sabharwal. Closing the curious case of neural text degeneration. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dONpC9GL1o.

Daniel Gareev, Thomas Hofmann, Ezhilmathi Krishnasamy, and Tiago Pimentel. Local and global decoding in text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14577–14597, 2024. doi: 10.18653/v1/2024.findings-emnlp.854. URL https://aclanthology.org/2024.findings-emnlp.854/.

Renato Geh, Honghua Zhang, Kareem Ahmed, Benjie Wang, and Guy Van den Broeck. Where is the signal in tokenization space? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3966–3979, 2024. doi: 10.18653/v1/2024.emnlp-main.230. URL https://aclanthology.org/2024.emnlp-main.230/.

Arash Gholami Davoodi, Navid Rezazadeh, Seyed Pouyan Mousavi Davoudi, and Pouya Pezeshkpour. Geometry-aware decoding with Wasserstein-regularized truncation and mass penalties for large language models. *arXiv preprint arXiv:2602.10346*, 2026. URL https://arxiv.org/abs/2602.10346.

Kevin Gu, Eva Tuecke, Dmitriy Katz, Raya Horesh, David Alvarez-Melis, and Mikhail Yurochkin. CharED: Character-wise ensemble decoding for large language models. *arXiv preprint arXiv:2407.11009*, 2024. URL https://arxiv.org/abs/2407.11009.

Joseph Y. Halpern and Daphne Koller. Representation dependence in probabilistic inference. *Journal of Artificial Intelligence Research*, 21:319–356, 2004. doi: 10.1613/JAIR.1292.

Jonathan Hayase, Alisa Liu, Noah A. Smith, and Sewoong Oh. Sampling from your language model one byte at a time. In *Forty-third International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2506.14123.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

John Hewitt, Christopher D. Manning, and Percy Liang. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, 2022. doi: 10.18653/v1/2022.findings-emnlp.249. URL https://aclanthology.org/2022.findings-emnlp.249/.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Kanishk Jain, Matthew Day, and Tankut Can. Emergent retokenization symmetry in large language models: Phenomenology and applications. *arXiv preprint arXiv:2606.15521*, 2026. URL https://arxiv.org/abs/2606.15521.

Xiaotong Ji, Rasul Tutunov, Matthieu Zimmer, and Haitham Bou-Ammar. Decoding as optimisation on the probability simplex: From top-k to top-p (nucleus) to best-of-k samplers. *arXiv preprint arXiv:2602.18292*, 2026. URL https://arxiv.org/abs/2602.18292.

Tom Kempton and Stuart Burrell. Local normalization distortion and the thermodynamic formalism of decoding strategies for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22216–22231, 2025. doi: 10.18653/v1/2025.findings-emnlp.1210. URL https://aclanthology.org/2025.findings-emnlp.1210/.

Yinxi Li, Yuntian Deng, and Pengyu Nie. TokDrift: When LLM speaks in subwords but code speaks in grammar. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47605–47629, 2026. doi: 10.18653/v1/2026.acl-long.2199. URL https://aclanthology.org/2026.acl-long.2199/.

Pierre-Alexandre Mattei and Bruno Loureiro. The well-tempered classifier: Some elementary properties of temperature scaling. *arXiv preprint arXiv:2602.14862*, 2026. URL https://arxiv.org/abs/2602.14862.

Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/afa5f124e36bed5cc2125067005d43f5-Abstract-Conference.html.

Trang Nguyen, Shuang Wu, Runyan Tan, and Phillip Howard. Improving sample diversity in autoregressive text-to-image generation via cluster truncation. *arXiv preprint arXiv:2607.10535*, 2026. URL https://arxiv.org/abs/2607.10535.

Georgy Noarov, Soham Mallick, Tao Wang, Sunay Joshi, Yan Sun, Yangxinyu Xie, Mengxin Yu, and Edgar Dobriban. Foundations of top-k decoding for language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3890. URL https://papers.neurips.cc/paper_files/paper/2025/hash/a96d4fda3017f1773b261a52a3efc8dd-Abstract-Conference.html.

Krzysztof Onak and Xiaorui Sun. Probability–revealing samples. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 2018–2026. PMLR, 2018. URL https://proceedings.mlr.press/v84/onak18a.html.

Buu Phan, Marton Havasi, Matthew Muckley, and Karen Ullrich. Understanding and mitigating tokenization bias in language models. *arXiv preprint arXiv:2406.16829*, 2024. URL https://arxiv.org/abs/2406.16829.

Buu Phan, Brandon Amos, Itai Gat, Marton Havasi, Matthew J. Muckley, and Karen Ullrich. Exact byte-level probabilities from tokenized language models for FIM-tasks and model ensembles. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=zGej22CBnS.

Eric Price, Kevin Tian, Zhiyang Xun, and Yusong Zhu. Total variation distance estimation in autoregressive models. *arXiv preprint arXiv:2607.19510*, 2026. URL https://arxiv.org/abs/2607.19510.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Vésteinn Snæbjarnarson, Samuel Kiegeland, Tianyu Liu, Reda Boumasmoud, Ryan Cotterell, and Tim Vieira. Transducing language models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.05193.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025. URL https://arxiv.org/abs/2512.13961.

Tim Vieira, Benjamin LeBrun, Mario Giulianelli, Juan Luis Gastaldi, Brian DuSell, John Terilla, Timothy J. O'Donnell, and Ryan Cotterell. From language models over tokens to language models over characters. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 61391–61412, 2025a. URL https://proceedings.mlr.press/v267/vieira25a.html.

Tim Vieira, Tianyu Liu, Clemente Pasti, Yahya Emara, Brian DuSell, Benjamin LeBrun, Mario Giulianelli, Juan Luis Gastaldi, Timothy J. O'Donnell, and Ryan Cotterell. Language models over canonical byte-pair encodings. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 61413–61443, 2025b. URL https://proceedings.mlr.press/v267/vieira25b.html.

Sean Welleck, Ilia Kulikov, Jaedeok Kim, Richard Yuanzhe Pang, and Kyunghyun Cho. Consistency of a recurrent language model with respect to incomplete decoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 5553–5568, 2020. doi: 10.18653/v1/2020.emnlp-main.448. URL https://aclanthology.org/2020.emnlp-main.448/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Moran Yanuka, Paul Dixon, Eyal Finkelshtein, Daniel Rotman, and Raja Giryes. Principled coarse-grained acceptance for speculative decoding in speech. *arXiv preprint arXiv:2511.13732*, 2026. URL https://arxiv.org/abs/2511.13732. Accepted to ICASSP 2026.

Brian Siyuan Zheng, Alisa Liu, Orevaoghene Ahia, Jonathan Hayase, Yejin Choi, and Noah A. Smith. Broken tokens? your language model can secretly handle non-canonical tokenizations. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1017. URL https://papers.neurips.cc/paper_files/paper/2025/hash/2b5c5689fae6fa9a4883e73e511d52c8-Abstract-Conference.html.

A PROOFS OF THE REFINEMENT AND PATH RESULTS

This supplement gives complete proofs, including sharpness constructions. All refinements used below preserve each parent mass exactly. Supremum constructions are mathematical examples, not additional language-model experiments.

A.1 EXACT REACHABLE LAWS AND UNIVERSAL DIVERGENCE EXTREMALITY

The capacity constraint $\rho q \le p$ is a bounded likelihood-ratio condition, related to the expanded-space conditioning criterion of Diaconis & Zabell (1982, Theorem 2.1). Here the additional task is to realize it by a finite descending nucleus prefix with positive child masses and to identify a common divergence extremizer under the fixed support constraint.

Proof of Theorem 3.1: reachability. For any same-support child nucleus, its retained mass $C \ge \rho$ and retained parent submeasure $s_i \le p_i$ give $\rho Q_i = \rho s_i / C \le p_i$. Hence its law lies in (5). Conversely, let q satisfy those inequalities. Every coordinate of q on S is positive: if $q_j = 0$ then $1 = \sum_{i \ne j} q_i \le (Z - p_j)/\rho \le S_{k-1}/\rho < 1$, a contradiction. Set $d = \min_{i \in S} \rho q_i > 0$. For every $i \in S$, create one retained child of mass ρq_i. Split each positive remainder $p_i - \rho q_i$, and every outside-parent mass, into finitely many positive children of mass strictly less than d. The selected children precede every other child and total exactly ρ; every proper prefix has smaller mass. Thus the child nucleus retains exactly those children and pushes forward to q. Dividing each child's mass by its parent's probability defines positive row-stochastic conditional weights. This proves exact finite attainability, not merely density of attainable laws.

If $Z = \rho$, the coordinate capacities sum to one and force $q = p_S/\rho = P = Q^\infty$. Otherwise put $s = Z - \rho > 0$. Every feasible law can be written uniquely as

$$q = \frac{p_S - s\lambda}{\rho}, \qquad \lambda \in \mathcal{P}(S).$$

Indeed the deficits $p_i - \rho q_i$ are nonnegative and sum to s. Since $s = p_k - (\rho - S_{k-1}) < p_k \le p_i$, every point of this simplex is positive on S. Its vertices $q^{[j]}$ put the entire deficit s on coordinate j. □

Proof of Theorem 3.1: extremality. For the nontrivial case $Z > \rho$, write $r = Z/\rho > 1$ and $x_j = P_j = p_j/Z$. At vertex $q^{[j]}$, the likelihood ratio $q_i^{[j]}/P_i$ is r for $i \ne j$, and $t_{x_j} = r - (r-1)/x_j$ at j. Therefore

$$D_f(q^{[j]} \| P) = F(x_j), \qquad F(x) = (1-x)f(r) + xf(t_x).$$

For $0 < x \le y$ in the feasible range, $t_y = (x/y)t_x + (1 - x/y)r$. Jensen gives $yf(t_y) \le xf(t_x) + (y-x)f(r)$ and hence $F(y) \le F(x)$. Thus a smallest parent probability maximizes the vertex divergence. Every feasible q is a convex combination of the vertices, and $D_f(q\|P)$ is convex in q, so its value cannot exceed $F(x_k) = D_f(Q^\infty \| P)$. This proves simultaneous maximization for every convex generator, with extended values interpreted in the standard way. It also covers reverse divergences through the conjugate generator $f^*(t) = tf(1/t)$.

For majorization, any set $B \subset S$ with $|B| = j < k$ has $q(B) \le p(B)/\rho \le S_j/\rho$. The largest j coordinates of Q^∞ sum to exactly S_j/ρ: its boundary coordinate is no larger than the other retained coordinates, since $\rho - S_{k-1} \le p_k$. At $j = k$ both sums are one. These inequalities prove that Q^∞ majorizes all feasible laws. In particular it minimizes all Schur-concave entropies on the feasible set.

Finally, all interior coordinates of Q^∞ exceed or equal those of P, and its boundary coordinate decreases. Their TV distance is

$$\frac{p_k}{Z} - \frac{\rho - S_{k-1}}{\rho} = \frac{S_{k-1}(Z - \rho)}{\rho Z}.$$

It is maximal either by the f-divergence result with $f(t) = |t-1|/2$, or directly: the positive-difference event $B = \{i : q_i > P_i\}$ is a proper subset of S, so $p(B) \le S_{k-1}$ and $q(B) - P(B) \le p(B)(1/\rho - 1/Z)$. The uniform-limit conclusion follows from the explicit nucleus stopping calculation in Corollary 3.5. Along that limit the likelihood ratios lie on the segment between 1 and their extremal values, so lower-semicontinuity and convexity also give convergence of the divergences, including infinite limiting values. □

A.2 Retention profiles and boundary thinning

Proof of Proposition 2.1. Truncation and renormalization on the child alphabet gives mass $p_y w(c \mid y)/C$ to each retained child. Summing within a fiber gives $Q_y = p_y a_y/C$. Comparing with $P_y = p_y \mathbf{1}_S(y)/Z$ proves (4), after cancelling $p_y > 0$. Under support equality, $a_y = 0$ off S for every positive-mass parent, so $\mathbb{E}_P[a] = C/Z$ and $Q_y/P_y = a_y/\mathbb{E}_P[a]$ on $\mathrm{supp}(P)$. In particular,

$$\mathrm{TV}(P, Q) = \frac{1}{2} \sum_y p_y \left| \frac{a_y}{C} - \frac{\mathbf{1}_S(y)}{Z} \right|.$$

Zero-mass fibers have no effect on any of these identities. □

Proof of Theorem 3.4. Compatibility gives retained parent submeasure $(p_1, \ldots, p_{k-1}, \alpha p_k, 0, \ldots, 0)$, with $C = S_{k-1} + \alpha p_k \geq \rho$. Parent-first retains $(p_1, \ldots, p_k, 0, \ldots, 0)$, of mass S_k. The two laws have the same support. For $k > 1$, their shared conditional law on $I = \{1, \ldots, k-1\}$ is $v_i = p_i/S_{k-1}$. Their boundary probabilities are $b = p_k/S_k$ and $c = \alpha p_k/C$; $c \leq b$ follows from $\alpha S_{k-1} \leq S_{k-1}$. Equality holds iff $\alpha = 1$. This proves the binary representation.

All interior coordinates increase when $\alpha < 1$, and the boundary coordinate decreases. Thus

$$\mathrm{TV}(P, Q) = b - c = \frac{p_k(C - \alpha S_k)}{C S_k} = \frac{S_{k-1}(S_k - C)}{C S_k}.$$

For every $j < k$, $\sum_{i \leq j} Q_i = S_j/C \geq S_j/S_k = \sum_{i \leq j} P_i$, and equality holds at $j = k$. Both vectors are sorted because $\alpha p_k \leq p_k \leq p_{k-1}$. These inequalities establish majorization and the entropy statements. The likelihood ratio is $(1-c)/(1-b)$ throughout I and c/b at the boundary, so

$$D_f(Q \| P) = (1-b) f\left(\frac{1-c}{1-b}\right) + b f\left(\frac{c}{b}\right),$$

which is exactly the Bernoulli divergence. For $k = 1$ both normalized laws are point masses regardless of α. □

Proof of Corollary 3.5. Every compatible refinement has $C \geq \rho$. Its TV gap is $S_{k-1}(1/C - 1/S_k)$, a decreasing function of C, and therefore is bounded by (14). Uniform refinement orders all children of parent i before those of parent $i+1$, with the fixed parent-major tie rule when needed. It retains all children of parents $i < k$ and exactly $r = \lceil m\beta \rceil$ children of parent k, proving (13). Because $0 < \beta \leq 1$, $1 \leq r \leq m$; in particular the boundary parent remains supported. The ceiling inequality gives $\rho \leq C_m < \rho + p_k/m$, so $C_m \to \rho$ and the displayed Q^∞ is the limit. This proves approach to the fixed-law maximum by uniform refinement, including the zero-gap case $S_k = \rho$; exact attainability by a general finite refinement was proved in Theorem 3.1.

The laws $P, Q^{(m)}, Q^\infty$ differ only in mass transferred from the boundary to the same interior conditional v. Their distances therefore add on that one-dimensional segment:

$$\mathrm{TV}(Q^{(m)}, Q^\infty) = \Delta^*_{\mathrm{prefix}} - \mathrm{TV}(P, Q^{(m)}) = \frac{S_{k-1}(C_m - \rho)}{\rho C_m} \leq \frac{S_{k-1} p_k}{m \rho^2}.$$

This includes $k = 1$, for which all terms vanish. □

Corollary A.1 (No-tie robustness)**.** *Suppose the positive parent masses are strictly ordered. For every uniform multiplicity m and every $\delta > 0$, there exists a positive refinement with the same multiplicity, all child masses distinct, a threshold-prefix-compatible retained prefix, and aggregated law within δ TV of the uniform-refinement law.*

Proof. For $m = 1$ the original law already has distinct masses. For $m > 1$, the uniform child law has strict inequalities between different parent blocks. Sufficiently small zero-sum perturbations inside each fiber preserve these inequalities and positivity. Let r be the number of retained boundary children. If their cumulative retained mass is strictly above ρ, while the preceding child prefix is strictly below it, a sufficiently small perturbation keeps both conditions. The retained boundary count stays r, and the aggregated law converges to the uniform one.

If the selected cumulative mass equals ρ and $r < m$, choose distinct within-boundary perturbations whose r largest entries have positive sum; they can be made arbitrarily small and have zero total sum. The selected cumulative mass is then above ρ while its predecessor stays below. If instead $r = m$, the entire boundary fiber is retained, so its mass and the selected cumulative mass remain exactly unchanged under every zero-sum perturbation; the preceding prefix remains below ρ. Perturb the other fibers generically as well. Since distinctness excludes only finitely many equality hyperplanes, all child masses can be made distinct. In each case the aggregated law tends to the uniform law, establishing the claim. A locally unchanged literal child prefix is asserted only away from the relevant rank and threshold boundaries; at an exact threshold the conclusion is convergence along the specified perturbations. □

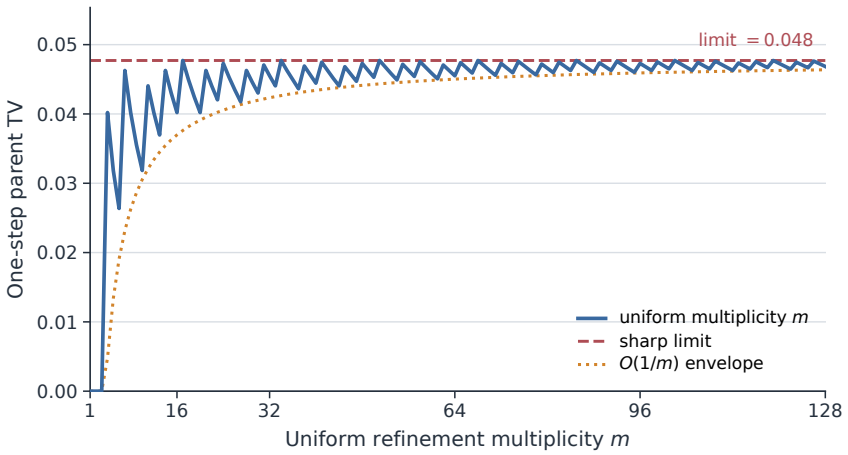

Figure 4: **Analytic illustration of Corollary 3.5, not language-model data.** For $p = (.42, .31, .17, .10)$ and $\rho = .85$, uniform multiplicity exhibits the ceiling-induced sawtooth while approaching the pointwise same-support TV maximum with an $O(1/m)$ envelope.

A.3 Sharp identical-support bounds and their constructions

Proof of Theorem 3.2: upper bounds. The main-text proof gives $\mathrm{TV}(P, Q) \le 1 - \rho/Z \le 1 - \rho$ and $\mathrm{TV}(P, Q) < 1/k$ for $k \ge 2$. These inequalities need only domination of the retained submeasures and the parent nucleus stopping rule; no child block-order assumption is used. If $k = 1$, support equality forces both laws to be the same point mass. Taking the largest bound over $k \ge 2$ gives $\Gamma_\rho = \min\{1/2, 1 - \rho\}$.

For the path upper bound, couple the next-outcome laws maximally whenever the histories still agree. Conditional agreement at each transition is at least $1 - \Gamma_\rho$, including absorbing steps where the discrepancy is zero. Induction gives agreement through H with probability at least $(1 - \Gamma_\rho)^H$. The coupling inequality proves (9). If every nonabsorbing history has nucleus size k, use $\Gamma_{\rho,k}$ instead. Equal next-outcome support at every shared history also implies equal path support, by induction on the prefix length. □

Proof of Theorem 3.2: sharpness. We first construct strictly ordered positive parent laws with nucleus size exactly k. Write $a = S_{k-1}$, $b = p_k$, $Z = a + b$. Uniform refinement with $m \to \infty$ gives boundary probability $(\rho - a)/\rho$ and interior law proportional to the first $k - 1$ parent masses, by Corollary 3.5.

Case 1: $\rho > (k - 1)/k$. Set $a = \rho - \varepsilon$ and $b = 1 - \rho + \varepsilon$, so $Z = 1$. For sufficiently small $\varepsilon > 0$, $a/(k - 1) > b$. Distribute a over $k - 1$ strictly decreasing masses all greater than b (for $k = 2$ there is only one such mass). There are no tail atoms. Since $a < \rho < Z$, the parent nucleus size is k. Its boundary mass tends to $1 - \rho$, while the uniformly refined limiting boundary mass ε/ρ tends to zero. The TV gap tends to $1 - \rho$.

Case 2: $0 < \rho \le (k-1)/k$. Again set $a = \rho - \varepsilon$, but let $b = a/(k-1) - \varepsilon$. For sufficiently small ε, $b > 0$, $Z = a + b > \rho$, and $Z < 1$. Choose the first $k-1$ masses strictly decreasing with sum a and minimum greater than b. Partition the remaining mass $1 - Z$ into sufficiently many distinct positive atoms, all smaller than b. This preserves strict parent ordering and nucleus size k. As $\varepsilon \downarrow 0$, $b/Z \to 1/k$, while the uniformly refined limiting boundary probability $\varepsilon/\rho \to 0$. Hence the TV gap tends to $1/k$. This construction includes the equality case $\rho = (k-1)/k$.

In both cases, first take m sufficiently large for each ε, then send ε to zero; this yields a sequence of *finite* refinements approaching $\Gamma_{\rho,k}$. Every member of the sequence has strictly positive probabilities on the same k retained outcomes. Distinct-child versions follow from Corollary A.1 with perturbation errors tending to zero.

For path sharpness, repeat a constructed pair independently at every step. Let $u_\varepsilon = a/Z$ be the parent-first interior probability and $u'_{\varepsilon,m}$ the child-first interior probability. Both laws have exactly the same conditional distribution on the interior. Along the construction, $u_\varepsilon \to 1 - \Gamma_{\rho,k}$ and $u'_{\varepsilon,m} \to 1$. For the event E_H that all H outcomes are interior,

$$Q_H(E_H) - P_H(E_H) = (u'_{\varepsilon,m})^H - u_\varepsilon^H \longrightarrow 1 - (1 - \Gamma_{\rho,k})^H.$$

TV is at least this event difference, matching the upper bound. Taking $k = 2$ establishes the unrestricted-support-size sharp constant. Each finite construction still has identical path support. Distinct-child perturbations preserve the limiting path gap because finite products are continuous in the transition probabilities. Finally $\rho = 1$ retains every atom in either representation and gives zero exactly. □

A.4 Unrestricted refinements

Lemma A.2 (Normalized submeasures)**.** *If $r, s \le p$ have masses $z_r, z_s \ge \rho > 1/2$, then* $\mathrm{TV}(r/z_r, s/z_s) \le (1-\rho)/\rho$.

Proof. Assume $z_r \le z_s$. Coordinatewise, $\min(r_y/z_r, s_y/z_s) \ge \min(r_y, s_y)/z_s$. Also $\sum_y \min(r_y, s_y) = z_r + z_s - \sum_y \max(r_y, s_y) \ge z_r + z_s - 1$. Thus

$$\mathrm{TV}(r/z_r, s/z_s) \le 1 - \frac{z_r + z_s - 1}{z_s} = \frac{1 - z_r}{z_s} \le \frac{1-\rho}{\rho}.$$

□

Proof of Theorem 3.3. For nucleus sampling, the parent masses retained by either order are submeasures of p, each of mass at least ρ. Apply Lemma A.2 for $\rho > 1/2$; TV ≤ 1 covers smaller thresholds. At $\rho = 1$, both orders are full sampling.

For sharpness when $1/2 < \rho < 1$, take parent masses $(\rho, 1-\rho)$. Parent-first keeps only the first parent. Split that parent into unequal children smaller than the second parent's mass, with maximum child mass tending to zero. Child-first retains the second parent followed by enough first-parent children to reach ρ. Its total retained mass C tends to ρ, so its second-parent probability and TV gap tend to $(1-\rho)/\rho$. For $\rho < 1/2$, choose $p_A > p_B \ge \rho$ with $p_A + p_B \le 1$, distribute any remaining mass over smaller parents, and split every parent except B into children below p_B. The two nuclei are the disjoint point masses at A and B. For $\rho = 1/2$, take $p_A = 1/2 + \varepsilon$, $p_B = 1/2 - \varepsilon$ and split A into children with maximum mass $o(\varepsilon)$. The gap tends to one.

For fixed Top-k(k), take two disjoint sets of k parents with all first-set masses larger than all second-set masses. Split each first-set parent into sufficiently many unequal children below every second-set mass. The decoded parent supports are disjoint, giving gap one. For Min-p(λ), take two parents with $p_B > p_A$ and $p_A < \lambda p_B$. Split B so finely that all its children are below λp_A. Parent-first selects B and child-first selects A, again giving gap one. Small generic perturbations remove ties while preserving the required strict inequalities.

The maximal-coupling argument above gives the horizon upper bound $1 - (1-\Delta^*)^H$. When $\Delta^* = 1$, disjoint one-step supports already give path TV one; the limiting $\rho = 1/2$ construction approaches it. When $0 < \Delta^* < 1$, the nucleus construction pairs a point mass with a law whose mass on that point tends to $1 - \Delta^*$. Independent repetition has path overlap tending to $(1-\Delta^*)^H$. This proves sharpness; the disjoint-support construction handles Top-k(k) without any point-mass requirement. □

A.5 PATH LIKELIHOODS, AFFINITY, AND ESTIMATION

Proof of (16) *and* (17). At a shared history with at least two retained outcomes, Theorem 3.4 gives likelihood ratio c_h/b_h at the boundary and $(1-c_h)/(1-b_h)$ throughout the interior. Multiplying the transition ratios and taking logs gives (16). The reverse log ratio has conditional expectation under P equal to $\mathrm{kl}(b_h\|c_h)$. Summation and the tower property give (17). Both kernels are strictly positive on their common support, so the finite-alphabet, finite-horizon expressions are finite. Singleton and absorbing histories contribute zero. □

Proposition A.3 (Affinity representation and a sufficient separation condition)**.** *Under threshold-prefix compatibility at every shared history, let*

$$A_h = \sqrt{b_h c_h} + \sqrt{(1-b_h)(1-c_h)}, \qquad K_h(y) = \frac{\sqrt{P_h(y)Q_h(y)}}{A_h},$$

with $A_h = 1$ and $K_h = P_h = Q_h$ at singleton histories. Let K_H be the path law generated by these kernels. Then

$$\sum_y \sqrt{P_H(y)Q_H(y)} = \mathbb{E}_{K_H} \prod_{t=1}^{H} A_{Y_{<t}}. \tag{23}$$

If $\sum_{t=1}^{H}(1-A_{y_{<t}}) \geq \kappa$ on every supported K_H path, then $\mathrm{TV}(P_H, Q_H) \geq 1 - e^{-\kappa}$.

Proof. The shared interior conditional implies $\sum_y \sqrt{P_h(y)Q_h(y)} = A_h$, so K_h is a probability kernel. Multiplying $\sqrt{P_h Q_h} = A_h K_h$ over time and summing paths proves (23). Since $A_h \leq e^{-(1-A_h)}$, the assumed lower bound on cumulative affinity deficit bounds the right side by $e^{-\kappa}$. Finally $\min(P_H(y), Q_H(y)) \leq \sqrt{P_H(y)Q_H(y)}$ and $\mathrm{TV} = 1 - \sum \min$ give the result. □

The sufficient condition is not claimed for the measured language-model paths. It clarifies mathematically how recurring nonzero binary distortions can force separation even without support loss. In the history-independent case, the affinity is simply A^H.

Estimator and decomposition identities. On the union support, multiplying g_H by M_H gives $|P_H - Q_H|/2$, proving (21). On U_H exactly one of P_H, Q_H is positive, hence $g_H = 1$. Splitting the expectation proves (22). The identity is exact; the plotted quantities are Monte Carlo estimates of its terms.

With N independent paths from each law, the equal-allocation estimate has variance

$$\mathrm{Var}(\widehat{D}_H) = \frac{\mathrm{Var}_{P_H}(g_H) + \mathrm{Var}_{Q_H}(g_H)}{4N}.$$

Its $2N$ bounded independent summands imply $\Pr(|\widehat{D}_H - D_H| \geq \epsilon) \leq 2e^{-4N\epsilon^2}$. These statements condition on the kernels and sampling design; they are distinct from the family-bootstrap uncertainty used to summarize prompt heterogeneity. A deterministic rendering map contracts TV. A common continuation kernel applied after a differing first step cannot increase it, which justifies interpreting larger terminal gaps as requiring differences in later kernels, not an additive decomposition into time-step contributions. □

B CANONICAL LIFTS AND APPROXIMATE IMPLEMENTATION

The results in this section develop the conditional-preserving characterization used by quotient-first decoding. They are consequences of classical conditional updating, divergence contraction, and finite-space conditioning. The universal convex-divergence projection and strict-convexity uniqueness specialize Diaconis & Zabell (1982, Theorem 6.1, p. 829); the KL projection is also classical (Csiszár, 1975). These classical results precisely characterize the conditional-faithful repair.

B.1 FINITE-ALPHABET PROJECTION AND CONDITIONAL EXCESS

For a lower-semicontinuous convex generator f with $f(1) = 0$, use the standard lower-semicontinuous extension of the finite-alphabet f-divergence at zero masses.

Proof of Proposition 4.1. The target marginal and conditional fidelity force $q(c) = R(y)w(c \mid y)$, proving uniqueness of the faithful lift. If $p(y) = 0$, then $R(y) = 0$ and no admissible q puts mass on that fiber. For $p(y) > 0$,

$$\sum_{c:\pi(c)=y} w(c \mid y)\frac{q(c)}{L_w(p)(c)} = \frac{R(y)}{p(y)}.$$

Jensen's inequality gives

$$\sum_{c:\pi(c)=y} L_w(p)(c) f\left(\frac{q(c)}{L_w(p)(c)}\right) \geq p(y) f\left(\frac{R(y)}{p(y)}\right).$$

Sum over fibers. For $q = L_w(R)$, the likelihood ratio is constant inside each fiber, so equality holds. With strict convexity and a finite minimum, equality in Jensen requires that same constant ratio in every positive-mass fiber, proving uniqueness. For the isometry, the ratio $L_w(R)(c)/L_w(S)(c) = R(y)/S(y)$ is fiber-constant; summing its divergence contribution within a fiber reduces it to the parent contribution. The same conclusion follows at boundary masses by the stated extension. □

Corollary B.1 (Operational uniqueness and exact conditional excess)**.** *On the represented family $\{L_w(p) : p \in \mathcal{P}(\mathcal{Y})\}$, quotient-first decoding is the unique coherent, conditional-faithful implementation. More generally, if $q \ll L_w(p)$ and $R = \pi_\# q$, then*

$$\mathrm{KL}(q \| L_w(p)) = \mathrm{KL}(R\|p) + \sum_y R(y)\,\mathrm{KL}(q(\cdot \mid y)\|w(\cdot \mid y)), \tag{24}$$

$$\mathrm{TV}(q, L_w(R)) = \sum_y R(y)\,\mathrm{TV}(q(\cdot \mid y), w(\cdot \mid y)). \tag{25}$$

Zero-R fibers contribute zero.

Proof. Coherence fixes the target marginal $R = D_{\mathcal{Y}}(p)$, after which conditional fidelity invokes Proposition 4.1. For (24), substitute $q(c) = R(y)q(c \mid y)$ and $L_w(p)(c) = p(y)w(c \mid y)$ inside the log ratio and sum first within each fiber. The two resulting terms are the marginal KL and the weighted conditional KL. For (25), the two laws have the same fiber masses $R(y)$, so factor those masses out of the ℓ_1 difference. These are exact excess-cost identities, not claims that TV decomposes additively for laws with different marginals. □

B.2 PATH-SPACE CANONICALITY

Let $y_{1:H}$ be an absorbing parent path. At history $h = y_{<t}$, let π_h and w_h define the refinement; child sequences are valid when $\pi_{y_{<t}}(c_t) = y_t$. The induced path quotient Π_H maps a valid child path to its parent path, and

$$W_{y_{1:H}}(c_{1:H}) = \prod_{t=1}^{H} w_{y_{<t}}(c_t \mid y_t). \tag{26}$$

Corollary B.2 (Path-space canonicality)**.** *For any finite horizon, parent path law P_H, and target $R_H \ll P_H$, $L_W(R_H)$ is the unique conditional-faithful child-path law with marginal R_H, minimizes every convex f-divergence to $L_W(P_H)$, and satisfies*

$$D_f(L_W(R_H)\|L_W(S_H)) = D_f(R_H\|S_H) \qquad (R_H \ll S_H).$$

Applying quotient-first decoding at each history reproduces the parent-first path law exactly after pushforward.

Proof. The path spaces are finite and Π_H is a surjection. For any fixed parent path, (26) is positive and sums to one over its child-path fiber, by repeated conditioning. Thus this is an ordinary finite refinement, and Proposition 4.1 applies directly. For sequential quotient-first decoding, the pushed-forward transition kernel equals the parent-first kernel at every shared history; induction identifies the full pushed-forward path law with the parent-first law. This is an invariance statement conditional on shared histories, not an assertion that independently sampled paths coincide. □

B.3 Universal quotient-blind coherence

The identity-only conclusion is the $m = 1$ case of the classical distribution-functor naturality characterization (Bartels, 2004, Section A.3, p. 180; Theorem A.3.5, p. 184). Our finite, surjection-only proof makes the decoder assumptions explicit in this classical characterization.

Proposition B.3 (Identity-only finite coarse-graining coherence)**.** *Suppose a family $D_{\mathcal{Y}} : \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{Y})$ is defined on every nonempty finite alphabet and commutes with every finite surjection:*

$$\pi_{\#} D_{\mathcal{C}}(q) = D_{\mathcal{Y}}(\pi_{\#} q) \quad \text{for every } q. \tag{27}$$

Then $D_{\mathcal{Y}}(p) = p$ for all p. Relabelings, indicator coarse-grainings to two points, and three-to-two merges suffice.

Proof. Define $g(t)$ as the first-coordinate output of $D_{\{0,1\}}(t, 1-t)$. Bijections imply relabeling equivariance. For any law p on an alphabet with at least two elements, the indicator map isolating a coordinate y gives $D_{\mathcal{Y}}(p)_y = g(p_y)$. Apply this representation to $(a, b, 1-a-b)$, then merge the first two points. Equation (27) implies $g(a+b) = g(a) + g(b)$ whenever $a, b \geq 0$ and $a + b \leq 1$. Setting $a = b = 0$ gives $g(0) = 0$; normalization at $(1, 0)$ gives $g(1) = 1$. Nonnegativity and additivity make g monotone. Therefore $g(r) = r$ for rational $r \in [0, 1]$, and rational approximation plus monotonicity gives $g(t) = t$ for every t. All output coordinates equal their input coordinates; a singleton alphabet is immediate. □

The proposition concerns all finite coarse-grainings simultaneously. It does not preclude a token-level decoder from being coherent for a specific restricted representation family. Nor does it say that the application must adopt the quotient rather than the token alphabet as its outcome space.

B.4 Approximate aggregation and a stable-prefix certificate

Proposition B.4 (Approximate quotient-first implementation)**.** *Let $\widetilde{p} = L_w(p)$ and $\widehat{\widetilde{p}} = L_{\widehat{w}}(\widehat{p})$ have the same quotient π, with $\mathrm{TV}(\widetilde{p}, \widehat{\widetilde{p}}) \leq \epsilon$. Suppose D obeys $\mathrm{TV}(D(p), D(\widehat{p})) \leq L\,\mathrm{TV}(p, \widehat{p})$ for this pair. If $\mathrm{TV}(\widehat{w}(\cdot \mid y), w(\cdot \mid y)) \leq \eta$ on each fiber with $D(p)(y) > 0$, then*

$$\mathrm{TV}(\pi_{\#} L_{\widehat{w}}(D(\widehat{p})), D(p)) \leq L\epsilon, \tag{28}$$

$$\mathrm{TV}(L_{\widehat{w}}(D(\widehat{p})), L_w(D(p))) \leq L\epsilon + \eta. \tag{29}$$

The conditionals on zero-mass fibers may be defined arbitrarily as normalized laws. For nucleus sampling with a common retained set, one may take $L = 1/\rho$.

Proof. Pushforward contracts TV, so $\mathrm{TV}(p, \widehat{p}) \leq \epsilon$. Since $\pi_{\#} L_{\widehat{w}}(D(\widehat{p})) = D(\widehat{p})$, the first bound follows. For the second, insert $L_{\widehat{w}}(D(p))$. A common conditional lift is an isometry in TV, bounding the first triangle term by $L\epsilon$. The second term equals $\sum_y D(p)(y)\,\mathrm{TV}(\widehat{w}(\cdot \mid y), w(\cdot \mid y)) \leq \eta$.

For the nucleus claim, we prove the sharper general conditioning inequality

$$\mathrm{TV}(p(\cdot \mid S), q(\cdot \mid S)) \leq \frac{\mathrm{TV}(p, q)}{\max\{p(S), q(S)\}}. \tag{30}$$

Assume $z = p(S) \geq z' = q(S) > 0$, and let $B \subseteq S$ be the coordinates where $p_y/z > q_y/z'$. Summing the positive difference gives

$$\mathrm{TV}(p(\cdot \mid S), q(\cdot \mid S)) = \sum_{y \in B} \left(\frac{p_y}{z} - \frac{q_y}{z'} \right) \leq \frac{1}{z} \sum_{y \in B} (p_y - q_y) \leq \frac{\mathrm{TV}(p, q)}{z}.$$

The opposite ordering is symmetric. Both nucleus retained masses are at least ρ, proving $L = 1/\rho$. This factor is sharp for conditioning on a set of fixed mass ρ by transferring mass between two of its coordinates. □

Corollary B.5 (A checkable stable-nucleus margin)**.** *Let $p_1 \geq \cdots \geq p_n$, let its nucleus size be k, and put $\delta = \mathrm{TV}(p, q)$. The law q retains the same nucleus set $S = \{1, \ldots, k\}$ whenever*

$$\delta < \rho - S_{k-1}, \qquad \delta \leq S_k - \rho, \qquad 2\delta < p_k - p_{k+1}, \tag{31}$$

where the last inequality is omitted for $k = n$. Under these conditions its decoded-law error is at most δ/ρ.

Proof. Every coordinate changes by at most δ. The last inequality therefore keeps all members of S above all nonmembers, without requiring their internal order to stay fixed. Every $(k-1)$-element set has p-mass at most S_{k-1} and hence q-mass at most $S_{k-1} + \delta < \rho$. Meanwhile $q(S) \geq S_k - \delta \geq \rho$. Thus q reaches the threshold for the first time at exactly S. Apply (30). □

This is a sufficient certificate, not a global smoothness claim. Near rank or threshold boundaries hard truncation is discontinuous. Full sampling at $\rho = 1$ is globally coherent without a margin certificate.

Corollary B.6 (Finite-horizon approximate coherence)**.** *If two parent transition kernels differ by at most $\delta_t \in [0, 1]$ TV uniformly over shared histories at time t, then*

$$\mathrm{TV}(P_H, \widehat{P}_H) \leq 1 - \prod_{t=1}^{H} (1 - \delta_t) \leq \sum_{t=1}^{H} \delta_t.$$

For approximate quotient-first aggregation one may substitute $\delta_t = \min\{1, L_t \epsilon_t\}$ under Proposition B.4's local assumptions.

Proof. Maximally couple the next-parent kernels while histories agree. The probability of agreement through time H is at least the displayed product. The coupling inequality gives the first bound and the elementary union bound gives the second. □

C Complete numerical results and robustness checks

All intervals below are two-sided 95% task-stratified family-bootstrap intervals. The complete arms, operating points, and descriptive sensitivity summaries are reported; displayed precision follows the supplied numerical tables.

Table 3: **Full parent-path results across all model/refinement arms.** Intervals are two-sided 95% task-stratified family-bootstrap intervals. The final column is the one-sided-support contribution at $H = 64$; the remainder is common-support reweighting.

Arm	Exact D_1	D_{64}	One-sided $H = 64$
Qwen3-8B, Dirichlet $m = 8$	.009 [.006,.012]	.185 [.176,.195]	.033 [.026,.041]
Qwen3-8B, uniform $m = 8$	.014 [.009,.019]	.256 [.245,.267]	.000 [.000,.000]
Qwen3-8B, Dirichlet $m = 4$	.009 [.006,.012]	.169 [.160,.178]	.028 [.022,.035]
Qwen3-8B, Dirichlet $m = 2$	.006 [.004,.009]	.131 [.123,.139]	.025 [.019,.031]
Qwen2.5-0.5B, Dirichlet $m = 8$	.031 [.028,.033]	.253 [.238,.267]	.088 [.077,.100]
OLMo-3-7B, Dirichlet $m = 8$	.013 [.010,.016]	.252 [.240,.265]	.076 [.066,.086]

Table 4: Primary Qwen3-8B Dirichlet-$m = 8$ parent-path trajectory.

H	$\widehat{D}_H$	95% interval
1	.00842	[.00567,.01137]
8	.04226	[.03586,.04921]
16	.07431	[.06627,.08284]
32	.12450	[.11588,.13329]
64	.18536	[.17555,.19528]

Table 5: Paired Qwen3 multiplicity and split-scheme contrasts.

Contrast	D_{32} difference	D_{64} difference
$m = 8 - m = 4$	.010 [.004,.017]	.016 [.009,.024]
$m = 4 - m = 2$	.025 [.019,.030]	.038 [.031,.045]
$m = 8 - m = 2$	.035 [.029,.041]	.054 [.047,.062]
Dirichlet $m = 8$ minus uniform $m = 8$	$-.054$ [$-.062$,$-.046$]	$-.071$ [$-.079$,$-.063$]

Table 6: **Estimated terms of the exact terminal-TV decomposition.** Parent rows use $H = 64$; VCT uses $H = 32$.

Arm	Total TV	One-sided	Common support
Qwen3-8B, Dirichlet $m = 8$	.185 [.176,.195]	.033 [.026,.041]	.152 [.146,.159]
Qwen3-8B, uniform $m = 8$	.256 [.245,.267]	.000 [.000,.000]	.256 [.245,.267]
Qwen3-8B, Dirichlet $m = 2$	.131 [.123,.139]	.025 [.019,.031]	.106 [.102,.111]
Qwen2.5-0.5B, Dirichlet $m = 8$	.253 [.238,.267]	.088 [.077,.100]	.165 [.157,.173]
OLMo-3-7B, Dirichlet $m = 8$	.252 [.240,.265]	.076 [.066,.086]	.177 [.169,.185]
Qwen2.5-0.5B, VCT byte	.038 [.027,.049]	.023 [.014,.033]	.015 [.012,.017]

Table 7: **Decoder and operating-point robustness on the fixed 60-family panel.** Every predeclared arm is shown.

Variant	Exact D_1	D_{32}	One-sided $H = 32$
TOP-P$(.90)$, $T = .7$	.008 [.005,.012]	.118 [.102,.135]	.010 [.002,.019]
TOP-P$(.80)$, $T = .7$	.012 [.007,.018]	.161 [.139,.184]	.044 [.027,.060]
TOP-P$(.95)$, $T = .7$	.007 [.004,.009]	.076 [.067,.086]	.004 [.000,.010]
TOP-P$(.90)$, $T = 1.0$	.013 [.008,.017]	.146 [.128,.166]	.040 [.021,.060]
TOP-K(50), $T = .7$	.000 [.000,.000]	.003 [.001,.008]	.002 [.000,.006]
MIN-P$(.10)$, $T = .7$	.009 [.005,.014]	.144 [.123,.166]	.033 [.017,.052]
Full sampling, $T = .7$	.000 [.000,.000]	.000 [.000,.000]	.000 [.000,.000]

Table 8: Held-out VCT-byte result. Intervals are two-sided 95% task-stratified family-bootstrap intervals.

Quantity	Estimate [95% interval]
Exact D_1	.01381 [.01006,.01772]
$\widehat{D}_1$	.01445 [.00983,.01971]
D_{32}	.03755 [.02727,.04904]
$D_{32} - D_1^{\text{exact}}$	.02374 [.01513,.03345]
Ordered minimum-effect gates	pass/pass

D TASK, SPLIT-MAP, AND FAMILY-LEVEL SENSITIVITY

The task-stratified and split-map displays were frozen before confirmatory outputs were opened. They are descriptive robustness analyses; no task-specific or split-specific gate is applied.

Table 9: Post-hoc descriptive distribution of family-level D_{64} estimates. These quantiles are reported to expose heterogeneity, not as additional confirmatory tests.

Arm	Min	10%	Median	90%	Max
Qwen3 Dirichlet $m = 8$	.049	.111	.179	.261	.324
Qwen3 uniform $m = 8$	.048	.172	.256	.334	.476
Qwen2.5 Dirichlet $m = 8$	.045	.127	.257	.367	.489
OLMo-3 Dirichlet $m = 8$	.101	.168	.248	.351	.464

Table 10: **Later divergence despite recorded initial agreement.** Existing post-hoc diagnostics from the frozen aggregate records. "Zero-D_1 families" counts families with recorded exact initial TV equal to zero; the next column is their stored mean terminal-TV estimate. Parent rows use $H = 64$ and the VCT row uses $H = 32$. r is the Pearson correlation between exact initial and terminal TV. No subgroup confidence interval or additional hypothesis test is claimed.

Arm	Zero-D_1 families	Subgroup TV	Initial–terminal r
Qwen3 Dirichlet $m = 8$	89	.175	.331
Qwen3 uniform $m = 8$	92	.248	.242
Qwen3 Dirichlet $m = 2$	91	.126	.207
Qwen2.5 Dirichlet $m = 8$	12	.223	.137
OLMo-3 Dirichlet $m = 8$	71	.243	.103
Qwen2.5 VCT byte	22	.011	.579

E ADDITIONAL SCOPE AND LIMITATIONS

The virtual-child intervention is intentionally narrower than changing a tokenizer: a real tokenizer can also alter input segmentation, learned probabilities, and hidden states. This controlled design identifies the decoder-order effect. The quotient is supplied rather than learned; discovering, validating, and computing semantic quotients may be difficult or nonlocal. The exact VCT result applies only to the declared locally canonicalized kernels, one shared special-stop outcome, and Qwen2.5-0.5B. The larger virtual-refinement results span three checkpoints but only three QA prompt strata, finite horizons through 64 parent tokens, and three fixed split maps; the families are not a probability sample of language use, and the bootstrap conditions on checkpoints and maps. Uniform splitting contains within-fiber ties, although Corollary A.1 proves that ties are not required mathematically. Parent-path TV upper-bounds but does not identify decoded-string or semantic-output TV. The evaluation measures distributional coherence; assessing generation quality, accuracy, calibration, safety, and human preference requires separate task-level evaluation.

F PROTOCOL AND CALIBRATION

Estimator calibration. For every parent/refinement arm, the Monte Carlo estimate at $H = 1$ is compared against direct vocabulary summation. The absolute errors are below the declared distribution-free tolerance in every arm; representative values are 7.36×10^{-4} for the primary Qwen3 arm, 7.14×10^{-4} for uniform Qwen3, 2.14×10^{-3} for Qwen2.5, and 1.72×10^{-4} for OLMo-3. The VCT path estimate at $H = 1$ is 0.01445 ($[0.00983, 0.01971]$), compared with exact $D_1 = 0.01381$ ($[0.01006, 0.01772]$).

Supplementary material. The supplementary package contains the execution source, frozen configurations, raw path/checkpoint records, family-level analysis outputs, bootstrap implementation

and seed, logs, and the executed byte-sampler patch corresponding to the reported experimental suite. Model weights and caches are not included.

G EXECUTION INVARIANTS AND RECONSTRUCTION

The implementation checks positive normalized split weights, parent-mass preservation, quotient-first agreement, and dual scoring at identical parent histories. Only real parent tokens enter the transformer, and both path laws use a common absorbing end-of-sequence space. The VCT checks 257-outcome normalization, agreement with direct token aggregation at prompt boundaries, shared byte prefixes, and a common absorbing stop. These checks enforce the shared-kernel intervention.

Family bootstraps resample within task, average the three task means, and use matched family differences for paired comparisons. The resampling unit is the source family, with paths and split maps averaged within family. The robustness subset is selected before inference without labels or outcomes. Raw estimates are reported without post-processing.

Experiments used one NVIDIA A100-PCIE-40GB, PyTorch 2.9.0+cu128, CUDA 12.8, bfloat16 arithmetic, and scaled dot-product attention. Environment and resource details are retained with the supplementary code.

Conditioning of the reported estimand. For each full parent arm, let $P_H^{a,f}$ be the reference law for task a and family f, and let $Q_H^{a,f,j}$ be the corresponding refined law for fixed split map j. The equally weighted estimand is

$$\overline{D}_H = \frac{1}{3}\sum_{a=1}^{3}\frac{1}{40}\sum_{f=1}^{40}\frac{1}{3}\sum_{j=1}^{3}\mathrm{TV}(P_H^{a,f}, Q_H^{a,f,j}). \tag{32}$$

The reported D_H abbreviates this conditional average. Each distance is estimated by equally allocated paths from its two conditional laws, before averaging within family and task. It is not generally the TV between laws obtained by marginalizing the task, family, or split map. The byte arm has no virtual split-map average. Paired intervals use matched family differences, not a difference between independently bootstrapped endpoints. The supplementary package includes the family-level records, bootstrap seed, and resampling implementation used for these intervals.

Numerical support and calibration scope. The support and no-tie results are exact statements for the declared finite probability laws and stable tie order. A finite replay residual does not certify rank or threshold margins at every reachable floating-point history; Corollary B.5 supplies a sufficient local certificate when those quantities are available. The uniform arm's recorded one-sided-support mass is zero, consistent with its exact-arithmetic support identity. The post-hoc zero-initial-gap counts use the classification recorded in the supplied summaries, not an independently certified rational equality. The reported distribution-free $H = 1$ half-width is approximately $.02531$ for the full parent arms and $.06199$ for the decoder panel. Passing this coarse calibration check is a consistency diagnostic, not a fine-accuracy or path-uniform numerical certificate.