# BEYOND THE MARGINALS: DEPENDENCE AND PARALLEL DECODING IN MASKED DIFFUSION LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Parallel decoding in masked diffusion language models commits multiple tokens simultaneously from a product of marginals, introducing an error governed by the unobserved conditional total correlation of the revealed set. We establish the fundamental limits of parallel decoding by characterizing the tightest dependence certificates across a hierarchy of available information, defined as the supremum of total correlation over all consistent joint laws. From marginal entropies alone, the excess-bit criterion of Ben-Hamu et al. (2025) is tightest whenever the vocabulary is large relative to the entropies; with full marginal vectors, the tightest certificate is attained by a minimum-entropy coupling and can improve upon the excess bits by an amount on the order of $\log |\mathcal{V}|$ bits. Our main theoretical result characterizes the tightest certificate under truncated conditional queries: a residual coupling constrained by the marginals attains this certificate, and computable brackets decouple approximation slack from missing-information slack. On sources with known joint distributions, this formulation doubles the fraction of certified sets compared to discarding unqueried values, while for larger vocabularies missing query information accounts for most residual slack. Crucially, such query-based certificates presuppose that the denoiser outputs are conditionals of a single coherent joint law; truncation-aware tests on saved LLaDA-8B states reveal that real model outputs violate this precondition. Finally, we characterize the position left unconstrained by the excess-bit criterion, proving that anchoring it to the most uncertain position adds no parallelism and can paradoxically increase dependence.

## 1 INTRODUCTION

Masked diffusion language models denoise many positions in parallel and admit arbitrary generation orders (Austin et al., 2021a; Lou et al., 2024; Sahoo et al., 2024), and recent systems show that the paradigm scales (Nie et al., 2025; Ye et al., 2025; Nie et al., 2026). A denoiser trained with a factorized objective supplies per-position marginals, whereas a parallel step must sample from a *joint distribution*: committing several mutually dependent positions from a product of marginals introduces an error governed by the conditional total correlation of the committed set (Huang et al., 2022; Ben-Hamu et al., 2025), whose accumulation along a trajectory is characterized exactly (Zhang et al., 2026; Wen et al., 2026) and which can degrade generation quality substantially on tasks with strong token dependence (Kang et al., 2026). Existing samplers control how much is revealed per step through confidence (Wu et al., 2026; Cai & Li, 2026; Kasa et al., 2026) or entropy (Ben-Hamu et al., 2025), incorporate dependence through copula models or dependency estimates (Liu et al., 2025; Ringel et al., 2026; Sharma et al., 2026; Kim et al., 2026), test conditional independence with additional forward passes (Azangulov et al., 2026), or use lookahead and estimates of the remaining entropy to guide decoding (Lee et al., 2025; Ye et al., 2026; Saini et al., 2026).

A sampler never observes the total correlation that it must control. It observes per-position entropies, the full marginal probability vectors, and, at the cost of additional forward passes, the outputs of the denoiser queried with chosen positions clamped to chosen values. With exact outputs, a valid certificate must bound the total correlation of every joint law consistent with the observed information, so the tightest such certificate describes a limit on what that information can certify. We ask how much of the dependence of a parallel step can be certified at each of these levels of information, what



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.

Table 1: Upper bounds on the total correlation $\mathrm{TC}(P)$ of a revealed set $S$ at each level of information, valid for exact outputs, with *tightest* as in Definition 4; $\kappa_2 \approx 0.53$ and $\kappa_m \approx 1.22$ for $m \geq 3$.

| Information (extra cost) | Certificate | Status | Results |
|---|---|---|---|
| entropies $\widehat{h}_i$ (none) | $e(S) = \sum_i \widehat{h}_i - \max_i \widehat{h}_i$ | tightest if $\max_i \widehat{h}_i + \lvert S\rvert \leq \log_2 \lvert\mathcal{V}\rvert$; not in general | Thm. 3, Prop. 5 |
| marginal vectors $\widehat{P}_i$ ($\widetilde{O}(\lvert S\rvert^2\lvert\mathcal{V}\rvert)$ arithmetic) | $e_{\mathrm{opt}}(S) = \sum_i \widehat{h}_i - H_{\mathrm{MEC}}$; computable $e_{\mathrm{G}} \leq e_{\mathrm{opt}} \leq e_{\mathrm{MP}} \leq e_{\mathrm{G}} + \kappa_{\lvert S\rvert}$ | tightest here; below $e$ by up to an order of $\log \lvert\mathcal{V}\rvert$ bits | Props. 6, 8, Cor. 7 |
| clamped conditionals of an anchor $a$ (one round of passes) | $e_a^{\mathrm{opt}}(S)$, a conditional MEC; $\beta_R^\star(S)$ if only the values in $R$ are queried | tightest given the queried values; chain-rule $e_T$ valid, not tightest; requires coherent outputs (testable) | Thm. 9, Props. 10, 13–15 |

the highest level requires of the denoiser, and what remains unconstrained once a certificate is fixed (Table 1). Our main contributions are as follows.

- **Limits of entropy and marginal information** (Section 3). The excess bits are the tightest entropy-only certificate whenever $\max_i \widehat{h}_i + |S| \leq \log_2 |\mathcal{V}|$. The tightest marginal-only certificate is given by a minimum-entropy coupling; computable bounds bracket it within $\approx 1.22$ bits, and it can fall below the excess bits by an amount on the order of $\log |\mathcal{V}|$ bits.
- **Certificates from conditional queries: main result** (Section 4.1). When querying only partial anchor values, a residual coupling constrained by marginals yields the tightest certificate, with computable brackets bounding approximation and information slack separately. On a four-symbol source with two queried values, it doubles the certified fraction over discarding unqueried values; for larger vocabularies, missing query information drives most remaining slack.
- **Coherence as a testable precondition** (Section 4.2). Query-based certificates presuppose that the queried outputs are conditionals of a single joint law, a property known to fail for masked language models (Torroba Hennigen & Kim, 2023; Young et al., 2024). We show what truncated queries and entropy summaries can certify about this failure, characterize what they cannot detect, and certify it on saved LLaDA-8B states.
- **Degrees of freedom under a fixed budget** (Section 5). At maximum cardinality, one position of the reveal set remains unconstrained. We give a sufficient condition under which placing it at the most uncertain position commits more information than the entropy-ordered prefix, and we show that this choice can increase dependence.

**Relation to prior work.** The excess-bit bound is due to Ben-Hamu et al. (2025), and minimum-entropy couplings with their meet and profile bounds to Kocaoglu et al. (2017); Cicalese et al. (2019); Compton (2022); Compton et al. (2023). Kasa et al. (2026) also organize guarantees by the information available to the decoder but certify that argmax commits agree with the joint argmax, whereas we bound the total correlation that governs categorical commits. Kim (2026) separates incompatibility from conditional dependence; we ask what truncated queries can certify and detect. Trajectory-level analyses and further related work are discussed in Appendix G.

## 2 Setup and the cost of a parallel step

Let $C$ be a prompt, which may include any side information supplied to the denoiser, and let $X = (X_1, \ldots, X_n)$ be a response over a vocabulary $\mathcal{V}$, with $(C, X) \sim p_{\mathrm{data}}$; information is measured in bits. For a random vector $V = (V_i)_{i \in S}$ and a conditioning variable $U$, the *total correlation* (Watanabe, 1960; Cover & Thomas, 2006) is

$$\mathrm{TC}(V \mid U) = \sum_{i \in S} H(V_i \mid U) - H(V \mid U) = \mathbb{E}_U \Big[ \mathrm{KL}\left( p_{V|U} \,\Big\|\, \bigotimes_{i \in S} p_{V_i|U} \right) \Big] \geq 0. \tag{1}$$

A sampler starts from $Z_0 = (\texttt{[MASK]}, \ldots, \texttt{[MASK]})$; $M(Z_k)$ is the set of masked positions of the state $Z_k$. A *reveal policy* $\sigma$ selects a nonempty $S_k = \sigma(Z_k, C) \subseteq M(Z_k)$, commits those coordinates, and sets $Z_{k+1} = (Z_k, X_{S_k})$; because $\sigma$ depends on the state, reveal indicators are not independent of $X$ (Theorem 26). For a state $z$ and a prompt $c$, $p^\star(\cdot \mid z, c)$ is the conditional law of the masked coordinates, and $\widehat{P}_i = p_\theta(X_i = \cdot \mid z, c)$ is the denoiser marginal at a masked position $i$, with $\widehat{h}_i(z) = -\sum_{v \in \mathcal{V}} \widehat{P}_i(v) \log_2 \widehat{P}_i(v)$ and error $\widehat{e}_i(z) = 1 - \max_{v \in \mathcal{V}} \widehat{P}_i(v)$.

**Assumption 1** (Full support)**.** For every reachable state $z$, every $i \in M(z)$ and every $v \in \mathcal{V}$, we have $p_\theta(X_i = v \mid z, c) > 0$.

**Definition 2** (Excess bits)**.** For a state $z$ and a nonempty $S \subseteq M(z)$, $e(S \mid z) = \sum_{i \in S} \widehat{h}_i(z) - \max_{i \in S} \widehat{h}_i(z)$, the entropy mass of $S$ after discarding its most uncertain position. This is the approximate joint-dependence bound of Ben-Hamu et al. (2025).

Assumption 1 holds for any softmax denoiser. Samplers constrain $|S_k|$, $\max_{i \in S_k} \widehat{e}_i$ (Wu et al., 2026), $\sum_{i \in S_k} \widehat{h}_i$, or $e(S_k \mid Z_k)$ (Ben-Hamu et al., 2025). Fix $z$, $c$ and a nonempty $S \subseteq M(z)$; let $P = p^\star(X_S = \cdot \mid z, c)$ be the target joint with marginals $P_i$, let $Q = \bigotimes_{i \in S} \widehat{P}_i$ be the law from which a parallel step commits, let $e^\star(S) = \sum_{i \in S} H(P_i) - \max_{i \in S} H(P_i)$, and let $\mathrm{CE}(P_i, \widehat{P}_i) = \mathbb{E}_{P_i}[-\log_2 \widehat{P}_i]$ be the cross-entropy of a model marginal under the target marginal.

**Theorem 3** (Commit-error decomposition and the excess-bit bound)**.** *Under Assumption 1:*

*(i)* $\mathrm{KL}(P \,\|\, Q) = \mathrm{TC}(P) + \sum_{i \in S} \mathrm{KL}\left(P_i \,\|\, \widehat{P}_i\right)$;

*(ii)* $0 \le \mathrm{TC}(P) \le e^\star(S)$, *with equality on the right if and only if* $X_S$ *is a deterministic function of* $X_a$ *under* $P$ *for some* $a \in \arg\max_{i \in S} H(P_i)$;

*(iii)* $\mathrm{TV}(Q, P) \le \left(\frac{\ln 2}{2}\left[e^\star(S) + \sum_{i \in S} \mathrm{KL}(P_i \| \widehat{P}_i)\right]\right)^{1/2}$;

*(iv) if* $\widehat{P}_i = P_i$ *for all* $i \in S$, *then* $e(S \mid z) = e^\star(S)$ *and* $\mathrm{TV}(Q, P) \le \sqrt{(\ln 2/2)\, e(S \mid z)}$; *moreover,* $e(S \mid z) = 0$ *if and only if at most one position of* $S$ *has positive entropy, in which case* $Q = P$;

*(v) for every* $a \in S$, $\mathrm{KL}(P \,\|\, Q) \le \sum_{i \in S \setminus \{a\}} \mathrm{CE}(P_i, \widehat{P}_i) + \mathrm{KL}(P_a \,\|\, \widehat{P}_a)$, *with equality if and only if* $X_S$ *is a deterministic function of* $X_a$ *under* $P$.

Parts (i) and (ii) restate results of Ben-Hamu et al. (2025) for a single step; the remaining statements follow directly. Under exact marginals, $e(S \mid z) \le e$ certifies $\mathrm{TC}(P) \le e$ while admitting one position of arbitrarily large entropy. Part (v) is an offline validation bound (Appendix B.6); along a trajectory, the terms in (i) accumulate additively (Ben-Hamu et al., 2025; Zhang et al., 2026; Wen et al., 2026) (Theorem 27).

## 3 What entropies and marginals can certify

We compare levels of information through the best bound that each of them supports.

**Definition 4** (Tightest certificate)**.** Let the information about a reveal set $S$ be a map $\iota$ from joint laws of $X_S$ on $\mathcal{V}^S$ to observations, such as the vector of marginal entropies or the vector of marginal distributions. A *certificate at level* $\iota$ is a function $\beta$ with $\mathrm{TC}(P) \le \beta(\iota(P))$ for every joint law $P$ on $\mathcal{V}^S$. The *tightest* certificate is $\beta^\star_\iota(o) = \sup\{\mathrm{TC}(P) : \iota(P) = o\}$, and every certificate at level $\iota$ satisfies $\beta \ge \beta^\star_\iota$.

**Per-position entropies.** The excess-bit criterion uses only the entropies of the marginals. It is a valid entropy-only certificate by Theorem 3(ii), and on vocabularies that are large relative to the entropies no entropy-only certificate is smaller.

**Proposition 5** (When the excess bits are tightest)**.** *Let* $h_{(1)} \le \cdots \le h_{(m)}$ *be the sorted entries of an entropy vector for a reveal set* $S$ *of size* $m$ *over* $\mathcal{V}$, *and put* $\delta_1 = h_{(1)}$ *and* $\delta_j = h_{(j)} - h_{(j-1)}$ *for* $2 \le j \le m$. *If* $\prod_{j=1}^{m} \lceil 2^{\delta_j} \rceil \le |\mathcal{V}|$, *which holds in particular when* $\max_i h_i + m \le \log_2 |\mathcal{V}|$, *then the tightest certificate given only the entropies equals* $e(S) = \sum_i h_i - \max_i h_i$, *and it is attained by a joint law in which every coordinate is a function of a coordinate of largest entropy.*

The condition cannot be removed (Appendix A.3 gives a binary counterexample); for the 126,464-token vocabulary of Section 6, it holds whenever the largest entropy in a set of size $m$ is at most $16.9 - m$ bits.

**Marginal vectors.** Here $\beta^\star = \sum_i \widehat{h}_i - H_{\mathrm{MEC}}$, where $H_{\mathrm{MEC}}$ is the entropy of a minimum-entropy coupling (MEC) of the marginals (Cicalese et al., 2019; Compton et al., 2023). Computing $H_{\mathrm{MEC}}$ is NP-hard (Kovačević et al., 2015), an additive $\varepsilon$-approximation for $m$ marginals takes time $|\mathcal{V}|^{O(\mathrm{poly}(1/\varepsilon)\exp(m))}$ (Compton, 2025), and the greedy coupling (Kocaoglu et al., 2017), alongside the meet and profile bounds (Marshall & Olkin, 1979; Compton et al., 2023), provide upper and lower bounds on $H_{\mathrm{MEC}}$.

**Proposition 6** (The marginal-only certificate hierarchy)**.** *Fix marginals $\widehat{P}_i$, $i \in S$, with $m = |S| \geq 2$. Let $\bigwedge_{i\in S} \widehat{P}_i$ be the* majorization meet*, whose $k$-th partial sum is $\min_{i\in S} \sum_{l\leq k} \widehat{P}_i^{\downarrow}(l)$ for the decreasing rearrangements $\widehat{P}_i^{\downarrow}$; let $\mathrm{PROF}_S$ and $\mathrm{MPROF}_S$ be the profile and major-profile of Compton et al. (2023); let $G_S$ be the greedy coupling of Kocaoglu et al. (2017); and let $H_{\mathrm{MEC}}$ be the entropy of a minimum-entropy coupling. Put $e_{\mathrm{maj}}(S) = \sum_{i\in S} \widehat{h}_i - H(\bigwedge_{i\in S} \widehat{P}_i)$, and define $e_{\mathrm{prof}}, e_{\mathrm{MP}}, e_{\mathrm{G}}, e_{\mathrm{opt}}$ likewise with $H(\mathrm{PROF}_S)$, $H(\mathrm{MPROF}_S)$, $H(G_S)$ and $H_{\mathrm{MEC}}$. Then, for every joint law $P$ with marginals $\widehat{P}_i$,*

$$\mathrm{TC}(P) \leq e_{\mathrm{opt}}(S) \leq e_{\mathrm{MP}}(S) \leq \min\big(e_{\mathrm{prof}}(S), e_{\mathrm{maj}}(S)\big) \leq e(S \mid z), \qquad e_{\mathrm{G}}(S) \leq e_{\mathrm{opt}}(S).$$

*Moreover, $e_{\mathrm{opt}}(S)$ is the tightest certificate from the marginals, and, with $\kappa_2 = \log_2 e/e \approx 0.53$ and $\kappa_m = (1 + \log_2 e)/2 \approx 1.22$ for $m \geq 3$,*

$$e_{\mathrm{prof}}(S) - e_{\mathrm{G}}(S) \leq \kappa_m, \quad e_{\mathrm{maj}}(S) - e_{\mathrm{G}}(S) \leq \log_2 e, \quad e_{\mathrm{maj}}(S) - e_{\mathrm{opt}}(S) \leq 1 \text{ if } m = 2, \quad (2)$$

*so that $e_{\mathrm{G}}(S) \leq e_{\mathrm{opt}}(S) \leq e_{\mathrm{MP}}(S) \leq e_{\mathrm{G}}(S) + \kappa_m$, where $e_{\mathrm{G}}$, $e_{\mathrm{maj}}$, $e_{\mathrm{prof}}$ and $e_{\mathrm{MP}}$ are computable in $\widetilde{O}(m^2|\mathcal{V}|)$ time.*

The constants in (2) are due to Compton et al. (2023, Thms. 4.1 and 4.2), Compton (2022) and Cicalese et al. (2019). Neither $e_{\mathrm{prof}}$ nor $e_{\mathrm{maj}}$ dominates the other in general (Appendix A.2). The quantity $e_{\mathrm{G}}$ bounds the tightest certificate, not the total correlation: $e_{\mathrm{G}}(S) > \tau$ means that no certificate computed from these marginals can guarantee $\mathrm{TC}(P) \leq \tau$, not that $\mathrm{TC}(P) > \tau$.

**Corollary 7** (Lower bound on marginal-only certificates)**.** *If $\beta(S)$ upper bounds $\mathrm{TC}(P)$ for every joint law $P$ with marginals $\widehat{P}_i$, $i \in S$, then $\beta(S) \geq e_{\mathrm{opt}}(S) \geq e_{\mathrm{G}}(S) \geq \max\{e_{\mathrm{prof}}(S) - \kappa_{|S|},\ e_{\mathrm{maj}}(S) - \log_2 e\}$. The constant $\log_2 e$ cannot be improved uniformly in $|S|$ as a function of the meet alone (Compton, 2022).*

**Proposition 8** (Entropy values versus marginal vectors)**.** *For an integer $s \geq 2$, let $P_s$ be uniform on $2^s$ of $N = 2^{2s} + 1$ symbols, and let $Q_s$ put mass $\frac{1}{2}$ on one symbol and $2^{-(2s+1)}$ on each of the others, so that $H(P_s) = s$ and $H(Q_s) = s + 1$. For a reveal set of two positions on $N$ symbols:*

*(i) given the marginal vectors $(P_s, Q_s)$, $H_{\mathrm{MEC}} = \frac{3s+1}{2}$, so the tightest certificate is $e_{\mathrm{opt}} = \frac{s+1}{2}$;*
*(ii) given only the entropies $(s, s+1)$, the tightest certificate is $s = e(S)$, by Proposition 5.*

*The marginal vectors thus certify $\frac{s-1}{2}$ fewer bits, an amount on the order of $\log N$ that persists under full support; for two positions on $d$ symbols the separation never exceeds $\log_2 d$ (Appendix A.3).*

## 4 Beyond the Marginals: Conditional Queries

### 4.1 Query-based certificates

For a state $z$, a masked position $u$ and a value $v$, write $z[u \leftarrow v]$ for the state in which $u$ is revealed with value $v$. One forward pass at $z[u \leftarrow v]$ returns the *clamped conditionals* $\widehat{P}_{t|u=v} = p_\theta(X_t = \cdot \mid z[u \leftarrow v], c)$ of all remaining masked positions $t$; for an exact denoiser, these are the conditionals of the target joint $P$ given $X_u = v$.

**Theorem 9** (Depth-$T$ certificate)**.** *Let $a_1, \ldots, a_T$ be distinct elements of $S$, $1 \leq T \leq |S|$, and $A_T = \{a_1, \ldots, a_T\}$. Then*

$$\mathrm{TC}(P) \ \leq\ e_T(S) \ :=\ \sum_{i\in S} H(P_i) - \sum_{t=1}^{T} H\big(X_{a_t} \mid X_{a_1}, \ldots, X_{a_{t-1}}\big) = \sum_{i\in S} H(P_i) - H(X_{A_T}), \quad (3)$$

*with gap* $e_T(S) - \mathrm{TC}(P) = H(X_{S\setminus A_T} \mid X_{A_T}) \geq 0$. *Thus* $e_T$ *depends on the anchors only through* $A_T$, *is nonincreasing in* $T$, *and is exact if and only if* $X_{A_T}$ *determines* $X_S$; *for* $|S| = 2$ *and* $T = 2$ *it equals* $\mathrm{TC}(P)$, *and truncating each conditional entropy to the* $k$ *most probable values of the conditioning coordinates preserves validity.*

The certificate $e_T$ requires $T - 1$ rounds of clamped passes, and greedy anchor selection attains a $1 - 1/e$ fraction of the optimal anchor-set entropy (Nemhauser et al., 1978), up to an additive term under truncation (Corollary 20 and Remark 21). Theorem 9 does not imply a general ordering between $e_2$ and $e_{\mathrm{opt}}$: in Remark 22, the anchors of largest marginal entropy give $e_2 = e > e_{\mathrm{opt}}$, whereas greedy anchors make $e_2$ exact.

The chain-rule certificate uses one conditional entropy per value of the anchor, although each pass returns the conditional marginals of all remaining positions. Leveraging all of these conditional marginals yields the tightest certificate at this level.

**Proposition 10** (Tightest certificate from one anchor's queries)**.** *Fix* $S$ *and an anchor* $a \in S$.

*(i) Suppose that the queries at* $a$ *return the exact conditional marginals* $P_{i|a=v}$ *of every* $i \in S \setminus \{a\}$ *for every value* $v$. *Given these and the marginals, the tightest certificate is*

$$e_a^{\mathrm{opt}}(S) = \sum_{i\in S} H(P_i) - H(P_a) - \sum_v P_a(v)\, H_{\mathrm{MEC}}\big(\{P_{i|a=v}\}_{i\neq a}\big),$$

*and* $e_a^{\mathrm{opt}}(S) \leq e_{\mathrm{opt}}(S)$. *Replacing* $H_{\mathrm{MEC}}$ *by the entropy of the majorization meet of the conditional marginals gives* $e_a^{\mathrm{maj}}(S)$, *and replacing it by the largest conditional entropy at each value gives the envelope* $e_a^{\mathrm{env}}(S)$, *with* $\mathrm{TC}(P) \leq e_a^{\mathrm{opt}} \leq e_a^{\mathrm{maj}} \leq e_a^{\mathrm{env}} \leq \min_{b\neq a} e_2(S; a, b)$ *and* $e_a^{\mathrm{maj}} - e_a^{\mathrm{opt}} \leq c_{|S|-1}$, *where* $c_1 = 0$, $c_2 = 1$ *and* $c_r = \log_2 e$ *for* $r \geq 3$.

*(ii) Suppose instead that only the values in a set* $R \subsetneq \mathcal{V}$ *are queried, and let* $r = P_a(\mathcal{V} \setminus R) > 0$. *Define the residual laws* $q_a(v) = P_a(v)/r$ *for* $v \notin R$ *and* $q_i = \big(P_i - \sum_{v\in R} P_a(v) P_{i|a=v}\big)/r$ *for* $i \neq a$. *Given the marginals and the queried conditionals, the tightest certificate is*

$$\begin{aligned}\beta_R^\star(S) = \sum_{i\in S} H(P_i) - H(P_a) - \sum_{v\in R} P_a(v)\, H_{\mathrm{MEC}}\big(\{P_{i|a=v}\}_{i\neq a}\big) \\ - r\Big[H_{\mathrm{MEC}}\big(q_a, \{q_i\}_{i\neq a}\big) - H(q_a)\Big],\end{aligned} \tag{4}$$

*which equals* $e_{\mathrm{opt}}(S)$ *when* $R = \emptyset$ *and satisfies* $e_a^{\mathrm{opt}}(S) \leq \beta_R^\star(S) \leq e_{\mathrm{opt}}(S)$. *Replacing each* $H_{\mathrm{MEC}}$ *in* (4) *by the entropy of the corresponding meet or major-profile gives computable valid certificates.*

In Remark 22, the anchor of largest marginal entropy already gives $e_a^{\mathrm{opt}} = \mathrm{TC}(P)$. Unlike the meet relaxation, the envelope does not admit a constant-gap bound to the optimum. Part (ii) reduces to Proposition 6 for $R = \emptyset$ and to part (i) when a single value is unqueried; discarding the unqueried values instead gives weaker certificates, and a negative entry of a residual law certifies inexact outputs (Appendix B.4, Proposition 13(ii)). Anchor tuples can be accommodated by replacing $X_a$ with $X_A$, and the optimum in (ii) can be bracketed from the same queries.

**Corollary 11** (Computable brackets for truncated queries)**.** *In Proposition 10(ii) with* $m = |S|$, *let* $U_R$ *and* $L_R$ *be obtained from* (4) *by replacing every* $H_{\mathrm{MEC}}$ *with the entropy of the corresponding major-profile and greedy coupling, respectively, and let* $\widetilde{U}_R$ *and* $\widetilde{L}_R$ *be obtained by replacing only the residual* $H_{\mathrm{MEC}}$. *Then* $U_R$ *and* $\widetilde{U}_R$ *are certificates,* $L_R \leq \widetilde{L}_R \leq \beta_R^\star \leq \widetilde{U}_R \leq U_R$, *and, with* $\kappa_1 = 0$,

$$U_R - L_R \leq (1 - r)\,\kappa_{m-1} + r\,\kappa_m, \qquad \widetilde{U}_R - \widetilde{L}_R \leq r\,\kappa_m.$$

*For two positions the two brackets coincide and have width at most* $\kappa_2 r$; *for* $R = \emptyset$ *they equal* $[e_{\mathrm{G}}, e_{\mathrm{MP}}]$. *For* $m \geq 3$, *the width* $U_R - L_R$ *need not vanish as* $r \to 0$ *(Appendix B.4).*

The second bracket requires exact couplings of the $m - 1$ retained conditional marginals, which is immediate for two positions and otherwise feasible only for small vocabularies (Compton et al., 2023).

### 4.2 Coherence of the queried conditionals

The certificates above hold for exact outputs. Exactness cannot be verified without the target distribution, but a necessary condition can be tested: exact outputs are the marginals and conditionals of one

joint law (Arnold & Press, 1989), a condition that fails for masked language models (Torroba Hennigen & Kim, 2023; Young et al., 2024; Kim, 2026). For masked positions $u \neq t$, a set $V_k(u)$ of $k$ retained values of $X_u$ and $S \supseteq \{u, t\}$, write $\widehat{h}_{t|u=v} = H(\widehat{P}_{t|u=v})$ and

$$C_k(u,t) = \sum_{v \in V_k(u)} \widehat{P}_u(v)\, \widehat{h}_{t|u=v}, \quad w_k(u) = \sum_{v \in V_k(u)} \widehat{P}_u(v), \quad d_{2,k}^{u \to t}(S) = \sum_{i \in S} \widehat{h}_i - \widehat{h}_u - C_k(u,t), \tag{5}$$

and let $C_{\text{full}}$ and $d_{2,\text{full}}^{u \to t}$ denote the complete sums, with $V_k(u) = \mathcal{V}$. With exact outputs, $d_{2,\text{full}}^{a \to b}(S)$ is the certificate $e_2(S)$ with anchors $a$ and $b$.

**Definition 12** (Coherent queries)**.** At a state $z$, the direction $u \to t$ is *Bayes-consistent* if $\widehat{P}_t = \sum_{v \in \mathcal{V}} \widehat{P}_u(v)\, \widehat{P}_{t|u=v}$. A pair $(a, b)$ is *jointly coherent* if the order-induced pair laws $Q^{a \to b}(v, w) = \widehat{P}_a(v) \widehat{P}_{b|a=v}(w)$ and $Q^{b \to a}(v, w) = \widehat{P}_b(w) \widehat{P}_{a|b=w}(v)$ coincide, that is, if one law on $\mathcal{V}^2$ has marginals $\widehat{P}_a$ and $\widehat{P}_b$ and both families of conditionals.

Exact outputs are jointly coherent, and joint coherence implies Bayes-consistency in both directions. Joint coherence does not imply exactness: a coherent denoiser can still differ from the target (Section 6.1).

**Proposition 13** (One-direction tests)**.** *Suppose that the direction $u \to t$ is Bayes-consistent at $z$. (i) $C_k(u,t) \leq C_{\text{full}}(u,t) \leq \widehat{h}_t$, and hence $d_{2,k}^{u \to t}(S) \geq \sum_{i \in S \setminus \{u,t\}} \widehat{h}_i \geq 0$ for every $V_k(u)$ and $S \supseteq \{u, t\}$. (ii) $\sum_{v \in R} \widehat{P}_u(v)\, \widehat{P}_{t|u=v}(x) \leq \widehat{P}_t(x)$ for every $R \subseteq \mathcal{V}$ and every $x \in \mathcal{V}$. A negative score in (i) or a violation of (ii) therefore certifies, without reference to the target, that the outputs of the denoiser at $z$ are not exact.*

**Proposition 14** (A two-direction test that accounts for truncation)**.** *For $S \supseteq \{a, b\}$ and $(u, t) \in \{(a, b), (b, a)\}$, the complete score lies in $I_k^{u \to t}(S) = \big[d_{2,k}^{u \to t}(S) - (1 - w_k(u)) \log_2 |\mathcal{V}|,\ d_{2,k}^{u \to t}(S)\big]$. If $(a, b)$ is jointly coherent at $z$, the two complete scores coincide, so the two intervals intersect. Consequently,*

$$\text{IncLB}_k(S) = \max\big\{0,\ \min I_k^{b \to a}(S) - \max I_k^{a \to b}(S),\ \min I_k^{a \to b}(S) - \max I_k^{b \to a}(S)\big\} > 0 \tag{6}$$

*certifies that $(a, b)$ is not jointly coherent at $z$. If the intervals obtained by averaging the endpoints with common nonnegative weights over a family of sets are disjoint, the pair is not jointly coherent at some set of positive weight.*

**Proposition 15** (A probability-level test and limitations of entropy-based tests)**.** *(i) For retained value sets $A_k$ of $X_a$ and $B_k$ of $X_b$, $D_k = \frac{1}{2} \sum_{v \in A_k, w \in B_k} |Q^{a \to b}(v, w) - Q^{b \to a}(v, w)| \leq \text{TV}(Q^{a \to b}, Q^{b \to a})$, and $D_k > 0$ certifies that $(a, b)$ is not jointly coherent. (ii) For uniform binary marginals with forward and reverse agreement probabilities $0.9$ and $0.1$, no entropy test triggers, although $\text{TV}(Q^{a \to b}, Q^{b \to a}) = 0.8$.*

Entropy tests use scalar summaries that can be averaged over states; probability-level tests require the retained vectors but provide higher detection sensitivity. Passing a test establishes neither coherence nor exactness.

## 5 Degrees of Freedom under a Fixed Excess-Bit Budget

The excess-bit certificate bounds how much dependence a reveal set may carry, but not which positions it contains.

**Proposition 16** (Maximum-cardinality reveal sets)**.** *Let $M = M(z)$ with $m = |M| \geq 1$, let $\widehat{h}_{(1)} \leq \cdots \leq \widehat{h}_{(m)}$ be the sorted entropies, with ties broken by index, and for $e \geq 0$ let $J(e) = \max\{j \leq m - 1 : \sum_{l=1}^{j} \widehat{h}_{(l)} \leq e\}$ and $B = \{(1), \ldots, (J(e))\}$. Then:*

- *(a) $\max\{|S| : \emptyset \neq S \subseteq M,\ e(S \mid z) \leq e\} = J(e) + 1$, attained by the* anchored *set $S^\star = B \cup \{(m)\}$, which contains the most uncertain position;*
- *(b) the entropy-ordered prefix $U = B \cup \{(J(e) + 1)\}$ of Ben-Hamu et al. (2025) also attains it, and the entropy masses of $S^\star$ and $U$ differ by $\widehat{h}_{(m)} - \widehat{h}_{(J(e)+1)} \geq 0$;*

Table 2: Two positions, controlled instance ($|\mathcal{V}| = 4$, mean TC $= 0.086$): mean certificate (bits) and percentage of 2,000 sets certified at $\tau = 0.25$ bits ($k = 0$: marginals only). The last column reports sets meeting the budget (93.2%) that cannot be certified.

| $k$ | retained mass | discarding | residual $\beta_R^\star$ | computable $U_R$ | TC $\le \tau <$ $L_R$ |
|---|---|---|---|---|---|
| 0 | 0 | – | 1.090 (1.5%) | 1.138 (1.4%) | 91.7% |
| 1 | 0.444 | 0.838 (4.7%) | 0.586 (14.6%) | 0.606 (14.4%) | 78.4% |
| 2 | 0.724 | 0.474 (24.8%) | 0.260 (50.0%) | 0.271 (48.5%) | 42.7% |
| 3 | 0.901 | 0.227 (58.8%) | 0.086 (93.2%) | 0.086 (93.2%) | 0.0% |

*(c) if the entropies are distinct and either $J(e) = 0$ or $e < \sum_{l \le J(e)} \widehat{h}_{(l)} + \widehat{h}_{(J(e)+1)} - \widehat{h}_{(J(e))}$, the sets of maximum cardinality are exactly $B \cup \{j\}$, $j \in M \setminus B$, and $j$ has the largest entropy in each.*

Anchoring thus adds no parallelism; the criterion leaves unconstrained the position of largest entropy, which we call the *free position*. Selecting a position $j$ commits $H(X_B \mid z) + H(X_j \mid X_B, z)$ bits, and comparing this across choices has a definite sign only under a specific condition.

**Corollary 17** (Information committed at equal certificate)**.** *Under exact marginals, for every $j \in M \setminus B$, $S_j = B \cup \{j\}$ satisfies $H(X_{S_j} \mid z) - H(X_U \mid z) \ge (\widehat{h}_j - \widehat{h}_{(J(e)+1)}) - e$. In particular,*

$$H(X_{S^\star} \mid z) - H(X_U \mid z) \;\ge\; \big(\widehat{h}_{(m)} - \widehat{h}_{(J(e)+1)}\big) - e, \tag{7}$$

*so the anchored set commits strictly more information than the prefix whenever $\widehat{h}_{(m)} - \widehat{h}_{(J(e)+1)} > e$.*

The condition can fail, and anchoring can then increase dependence under exact marginals (Remark 25).

The anchored rule places the free position where marginal uncertainty is largest, as lookahead samplers also do (Ye et al., 2026; Saini et al., 2026); its commit error includes the model error $\mathrm{KL}(P_a \| \widehat{P}_a)$ of Theorem 3(v). Algorithm 1 (Appendix C.4) implements the rule; at $e_k = 0$ it decodes one position per step, and the reveal order influences the output distribution only in the presence of incoherence.

**Proposition 18** (Reveal order under coherent outputs)**.** *Suppose that at every reachable state the denoiser marginals are the conditional marginals of one joint law $Q$ on $\mathcal{V}^n$. Then decoding one position per step with categorical commits samples exactly from $Q$ under every reveal policy that depends only on the current state. Consequently, a dependence of the output law on the reveal order at $e = 0$ certifies that the outputs do not derive from any joint law, and it cannot be explained by calibration error alone.*

The proof applies the chain rule along the realized order; the statement is the sampler-level form of the order consistency of compatible conditionals discussed by Kim (2026) and does not hold for argmax commits.

## 6 Experiments

We compare dependence certificates on tractable sources with known joint laws (Section 6.1) and test the coherence of conditional outputs from LLaDA-8B-Instruct (Nie et al., 2025) (Section 6.2). We then examine reveal-set composition through one-step commits at common states (Section 6.3) and complete generations (Section 6.4).

### 6.1 Exact verification on tractable sources

We use known joint laws: mixtures of products ($n = 7$, $|\mathcal{V}| = 4$, four components), a Markov chain ($n = 8$, $|\mathcal{V}| = 3$), and hidden Markov models with up to $n = 40$ and $|\mathcal{V}| = 8$; unless stated otherwise, the denoiser returns exact conditional marginals (Appendices E.1 and E.2).

**The marginals are structurally insufficient.** On the 12,000 evaluated reveal sets, the inequality chain TC $\le e_{\mathrm{MP}} \le \min(e_{\mathrm{prof}}, e_{\mathrm{maj}}) \le e$ held consistently across all evaluated instances, as

Table 3: Depth-2 scores on saved LLaDA states at $k = 8$ and incompatibility bounds (6) (hierarchical means over 35 prompts, in bits). A positive IncLB certifies incoherence on a state–pair of positive weight (all levels in Table 13).

| | $m = 2$ | | | $m = 3$ | | | $m = 4$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | prefix | anchored | uniform | prefix | anchored | uniform | prefix | anchored | uniform |
| $\bar{d}_1$ | 0.2009 | 0.2009 | 0.2009 | 0.6085 | 0.6085 | 0.6085 | 1.2419 | 1.2419 | 1.2419 |
| $\bar{d}_2^{a\to b}$ | 0.1486 | 0.1198 | 0.1338 | 0.5077 | 0.4642 | 0.4850 | 1.0969 | 1.0400 | 1.0656 |
| $\bar{d}_2^{b\to a}$ | 0.1821 | 0.2407 | 0.2062 | 0.5586 | 0.6467 | 0.5973 | 1.1604 | 1.2717 | 1.2089 |
| $\mathrm{IncLB}_8$ | **0.0045** | **0.0920** | **0.0434** | 0 | **0.0878** | **0.0175** | 0 | **0.0241** | 0 |

did $\mathrm{TC} \le e_{\mathrm{opt}} \le e_{\mathrm{MP}}$ at $|S| = 2$, where the MEC is exact. Even the meet-based lower bound $e_{\mathrm{maj}} - \log_2 e$ of Corollary 7, which is weaker than $e_{\mathrm{G}}$, averages 5.41 and 4.79 times the mean total correlation on the two respective sources. The depth-2 certificate removes 40.5% and 37.7% of the excess bits, but at total variation 0.05 it commits at most 1.29 positions per step, compared to 1.21 for anchoring and 1.13 for the prefix, which does not offset its $1 + k$ forward passes per step (Tables 4 and 5).

**Query-level certificates.** On the controlled instance, the residual certificate of Proposition 10(ii) lowers the discarding certificate by 30% to 62% and raises the fraction of sets certified at 0.25 bits from 5%, 25% and 59% to 15%, 50% and 93% for $k = 1, 2, 3$, and the computable bracket of Corollary 11 retains most of this gain (Table 2); at $k = 3$ the marginals uniquely determine the remaining conditional distribution, a consequence of the four-symbol alphabet. For three positions, the general bracket has mean width 0.05 to 0.08 bits. Since $\mathrm{TC} \le \beta_R^\star$, the slack $U_R - \mathrm{TC}$ splits into $U_R - \beta_R^\star \le U_R - L_R$, which exact couplings could remove, and $\beta_R^\star - \mathrm{TC} \ge L_R - \mathrm{TC}$, which only additional information can remove; for three positions, querying every anchor value still leaves an average gap of 0.86 bits for $e_a^{\mathrm{opt}} - \mathrm{TC}$, rendering 14.2% of the sets, all within the budget, uncertifiable (Table 9). In mixtures with $|\mathcal{V}| \ge 64$ at common retained-mass targets, $U_R - L_R$ averages at most 0.031 bits, $L_R - \mathrm{TC}$ averages 0.13 to 1.29 bits, and at most 0.2% of sets remain undecided at 0.25 bits, so the remaining slack reflects missing query information. This comes at a cost: for two positions, where the full-query optimum equals TC, 44.6% to 67.8% of the sets for $|\mathcal{V}| \ge 16$ meet the budget but provably cannot be certified at a target of 0.8, compared to 3.6% to 4.2% at 0.95, which requires 1.7 to 2.9 times as many queries (Table 10).

**Coherence tests and reveal order.** With the exact or a coherent but miscalibrated denoiser, no test triggers across the 2,000 pairs per denoiser; for incoherent denoisers, the probability-level test detects every pair at $k = 2$ of $|\mathcal{V}| = 4$ values, while the two-direction entropy test detects 2.5% to 32% of the pairs (Table 11). At $e = 0$, the coherent denoiser has the same error, $\mathrm{TV} = 0.482$, under three reveal orders, as Proposition 18 requires, whereas an incoherent one yields TV errors between 0.417 and 0.448 (Table 12).

**Composition and learned denoisers.** The anchored set and the prefix had equal cardinality in all 12,000 comparisons (Proposition 16(b)). With exact marginals, anchoring attains the smallest guaranteed error at matched cost on both reference sources (Table 7) and a lower realized error on the controlled instance, 0.041 compared to 0.057 at $K = 6$ (Table 12). For denoisers fitted by expectation maximization (Dempster et al., 1977), the exact-marginal bound fails at every singleton, whereas Theorem 3(v) always holds (Appendix F.1.3).

### 6.2 Conditional queries to LLaDA: scores and coherence

The target joint is unavailable, so we test necessary conditions for the coherence of LLaDA-8B-Instruct outputs at 6,720 saved pre-commit states from argmax decoding on 35 GSM8K training prompts (Cobbe et al., 2021), for the prefix, the anchored set and the uniform average over feasible size-$m$ sets, $m \in \{2, 3, 4\}$, at a common budget. For the two positions $a, b$ of largest marginal entropy, we clamp one at a time in both directions, retaining the $k \in \{1, 2, 4, 8\}$ most probable values; entropies are over all 126,464 tokens (Appendix E.3).

**Scores and coherence.** In the direction $a \to b$ the depth-2 score lies $11.7\%$ to $40.4\%$ below $d_1$, which is not evidence of low dependence, since this reduction equals $C_k(a, b) \geq 0$ for any denoiser; in the direction $b \to a$ the anchored score exceeds $d_1$ at every $m$ (Table 3). Proposition 14, applied to the hierarchical means with their common weights, gives a positive incompatibility bound in six of nine combinations at $k = 8$, including the anchored set at every $m$, and in seven of 36 over all $k$ (Table 13). Each positive value certifies that some state–pair of positive weight is incompatible with any joint law; this is an existence result, not a prevalence estimate. The largest bound, $0.092$ bits, far exceeds the rounding of the reported aggregates, but the outputs were computed with bf16 weights, and these measurements do not disentangle model incoherence from reduced-precision arithmetic. The mean ratio $d_2/d_1$ is negative in 22 of 72 combinations, so sets violating Proposition 13(i) exist; model-based depth-2 scores serve to diagnose coherence rather than to bound dependence.

### 6.3 Composition at common states

At the same states, every feasible set consists of the $m - 1$ lowest-entropy positions and one free position (Proposition 16(c)). The one-step decrease of the mean model entropy over the next-block positions that no feasible set contains is ordered $D_A > D_{\mathcal{I}} > D_P$ under both commit laws and at every $m$, with $D_A - D_P$ between $0.273$ and $0.505$ bits per position (Appendix E.4 and Table 14): this reflects a change in model entropies rather than evidence of correct commits, consistent with the pivot effect observed by Ye et al. (2026).

### 6.4 Complete generations

We evaluate fixed schedules, the prefix rule, and the anchored rule on all $1{,}319$ GSM8K and $500$ MBPP test problems (Cobbe et al., 2021; Austin et al., 2021b) with one completion per problem and configuration. GSM8K uses argmax commits, which lie outside the categorical guarantees, and full-vocabulary categorical commits at temperature one; MBPP uses argmax commits in one $256$-token block (Appendix E.5 and Table 15).

With argmax commits on GSM8K, the prefix at $e = 1$ matches the $256$-step schedule ($79.00\%$) at $73.78$ mean NFE. At a comparable mean NFE, anchoring at $e = 1$ ($74.75\%$, $63.75$ NFE) and the prefix at $e = 2$ ($74.53\%$, $65.32$ NFE) differ by $0.23$ percentage points in accuracy (anchored minus prefix, $95\%$ interval $[-2.20, 2.65]$) and by $-1.57$ in NFE ($[-2.10, -1.05]$), which does not establish equivalence; both exceed the $64$-step schedule ($63.91\%$). Categorical commits degrade sharply with parallelism ($68.54\%$ at $256$ steps, $25.25\%$ at $128$), and at $e = 1$ the prefix is $9.70$ percentage points more accurate than anchoring at a lower mean NFE. On MBPP, anchoring is more accurate than the prefix at a lower mean NFE for $e \leq 0.5$ and less accurate at larger budgets, and both remain below the $64$-step schedule; we hypothesize that entropy ordering within a single $256$-token block underlies this degradation. Neither rule for the free position is universally superior across these settings, and these measurements do not disentangle dependence, model error, incoherence, and changes in visited states.

## 7 Limitations and conclusion

Our certificates provide theoretical guarantees for exact outputs. For LLaDA, our empirical analysis evaluates aggregated entropy summaries; measuring state-level detection rates, probability-level tests, and precision effects remains for future work (Appendix E.6). Our LLaDA evidence covers one model, with $35$ development prompts for saved-state analyses and $1{,}319$ GSM8K and $500$ MBPP test problems for full-trajectory evaluations under fixed decoding protocols.

A parallel step certifies only as much dependence as its information allows. Whenever the condition of Proposition 5 holds, entropies certify nothing beyond the excess bits; marginal vectors can certify up to an order of $\log |\mathcal{V}|$ fewer bits; and for truncated anchor queries, computable brackets show in the tested sources that the remaining slack stems from missing query information. Interpreting such a certificate as a valid bound for the target requires output coherence, a condition that certain saved LLaDA-8B outputs violate. The choice of the free position alters model entropies at common states, whereas its effects in complete generations depend on the task, the budget, and the commit law.

### Reproducibility statement

All theoretical statements are proved in Appendices A to D. Appendices E.1 and E.2 specify the tractable sources, the controlled experiments and the exact reference calculations, and Appendices E.3 to E.5 specify the LLaDA protocols, including data splits, commit laws, cost accounting and interval methods. Appendix F reports the detailed results. The supplementary document contains the complete numerical tables for the LLaDA experiments and auxiliary tables for the tractable sources, and the supplementary code includes the script that reproduces Tables 2, 9 and 10. The coherence statistics of Section 6.2 are computed from the averaged scores and retained masses in Table 13. No neural network is trained in this work.

### AI use statement

We used large language models as assistive tools in the preparation of this manuscript: to check the statements and proofs in the appendices and propose corrections, to search for and position related work, to restructure the manuscript, and for copy-editing. They were not used to generate the tractable-source distributions of the other experiments, their enumeration code, or the LLaDA measurements. The authors checked all model-generated suggestions and take full responsibility for the statements, proofs and numerical results.

## References


Barry C. Arnold and S. James Press. Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405):152–156, 1989.

Koenraad M. R. Audenaert. A sharp continuity estimate for the von Neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127–8136, 2007.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, 2021a.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021b.

Iskander Azangulov, Teodora Pandeva, Niranjani Prasad, Javier Zazo, and Sushrut Karmalkar. Parallel sampling from masked diffusion models via conditional independence testing. In *International Conference on Learning Representations*, 2026.

Heli Ben-Hamu, Itai Gat, Daniel Severo, Niklas Nolte, and Brian Karrer. Accelerated sampling from masked diffusion models via entropy bounded unmasking. In *Advances in Neural Information Processing Systems*, 2025.

Jean Bretagnolle and Catherine Huber. Estimation des densités: Risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47(2):119–137, 1979.

Changxiao Cai and Gen Li. Confidence-based decoding is provably efficient for diffusion language models. *arXiv preprint arXiv:2603.22248*, 2026.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In *Advances in Neural Information Processing Systems*, 2022.

Clément L. Canonne. A short note on an inequality between KL and TV. *arXiv preprint arXiv:2202.07198*, 2022.

Rodrigo Casado Noguerales, Bernhard Schölkopf, Thomas Hofmann, and Aran Raoufi. What does a discrete diffusion model learn? *arXiv preprint arXiv:2607.05381*, 2026.

Kecheng Chen, Ziru Liu, Xijia Tao, Hui Liu, Xinyu Fu, Suiyun Zhang, Dandan Tu, Lingpeng Kong, Rui Liu, and Haoliang Li. Beyond confidence: Adaptive and coherent decoding for diffusion language models. *arXiv preprint arXiv:2512.02044*, 2025.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Sitan Chen, Kevin Cong, and Jerry Li. Optimal inference schedules for masked diffusion models. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 1279–1311, 2026.

Ferdinando Cicalese, Luisa Gargano, and Ugo Vaccaro. Minimum-entropy couplings and their applications. *IEEE Transactions on Information Theory*, 65(6):3436–3451, 2019.

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.

Spencer Compton. A tighter approximation guarantee for greedy minimum entropy coupling. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 168–173, 2022.

Spencer Compton. Efficient $\varepsilon$-approximate minimum-entropy couplings. *arXiv preprint arXiv:2509.19598*, 2025.

Spencer Compton, Dmitriy Katz, Benjamin Qi, Kristjan Greenewald, and Murat Kocaoglu. Minimum-entropy coupling approximation guarantees beyond the majorization barrier. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10445–10469, 2023.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38, 1977.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

Mark Fannes. A continuity property of the entropy density for spin lattice systems. *Communications in Mathematical Physics*, 31(4):291–294, 1973.

Jinyuan Feng, Xin Yu, Yiqun Chen, Xiaochi Wei, Yan Gao, Yi Wu, Yao Hu, and Zhiqiang Pu. Focus on the core: Empowering diffusion large language models by self-contrast. *arXiv preprint arXiv:2605.01373*, 2026.

Hengyu Fu, Baihe Huang, Virginia Adams, Charles Wang, Venkat Srinivasan, and Jiantao Jiao. From bits to rounds: Parallel decoding with exploration for diffusion language models. *arXiv preprint arXiv:2511.21103*, 2025.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005.

Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. SSD-LM: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023.

David Heckerman, David Maxwell Chickering, Christopher Meek, Robert Rounthwaite, and Carl Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1:49–75, 2000.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Emiel Hoogeboom, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. In *International Conference on Learning Representations*, 2022.

Yuezhou Hu, Harman Singh, Monishwaran Maheswaran, Haocheng Xi, Coleman Hooper, Jintao Zhang, Aditya Tomar, Michael W. Mahoney, Sewon Min, Mehrdad Farajtabar, Kurt Keutzer, Amir Gholami, and Chenfeng Xu. Residual context diffusion language models. *arXiv preprint arXiv:2601.22954*, 2026.

Fei Huang, Tianhua Tao, Hao Zhou, Lei Li, and Minlie Huang. On the learning of non-autoregressive transformers. In *International Conference on Machine Learning*, pp. 9356–9376. PMLR, 2022.

Zemin Huang, Yuhang Wang, Zhiyang Chen, and Guo-Jun Qi. Don't settle too early: Self-reflective remasking for diffusion language models. In *International Conference on Learning Representations*, 2026.

Moongyu Jeon, Sangwoo Shin, Dongjae Jeon, and Albert No. Information-theoretic discrete diffusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Wonjun Kang, Kevin Galim, Seunghyuk Oh, Minjae Lee, Yuchen Zeng, Shuibai Zhang, Coleman Hooper, Yuezhou Hu, Hyung Il Koo, Nam Ik Cho, and Kangwook Lee. ParallelBench: Understanding the trade-offs of parallel decoding in diffusion LLMs. In *International Conference on Learning Representations*, 2026.

Siva Rajesh Kasa, Yasong Dai, Sumit Negi, and Hongdong Li. Fast-dLLM++: Fréchet profile decoding for faster diffusion LLM inference. *arXiv preprint arXiv:2606.02955*, 2026.

Bumjun Kim, Dongjae Jeon, Moongyu Jeon, and Albert No. Dependency-aware parallel decoding via attention for diffusion LLMs. *arXiv preprint arXiv:2603.12996*, 2026.

Jeonseong Kim. Path-dependent denoising: A non-conservative field perspective on order collapse in diffusion language models. *arXiv preprint arXiv:2605.09303*, 2026.

Murat Kocaoglu, Alexandros G. Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In *AAAI Conference on Artificial Intelligence*, 2017.

Xianghao Kong, Rob Brekelmans, and Greg Ver Steeg. Information-theoretic diffusion. In *International Conference on Learning Representations*, 2023.

Mladen Kovačević, Ivan Stanojević, and Vojin Šenk. On the entropy of couplings. *Information and Computation*, 242:369–382, 2015.

Hugo Lavenant and Giacomo Zanella. Error bounds and optimal schedules for masked diffusions with factorized approximations. *arXiv preprint arXiv:2510.25544*, 2025.

Sanghyun Lee, Seungryong Kim, Jongho Park, and Dongmin Park. Lookahead unmasking elicits accurate decoding in diffusion language models. *arXiv preprint arXiv:2511.05563*, 2025.

Gen Li and Changxiao Cai. Breaking AR's sampling bottleneck: Provable acceleration via diffusion language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems*, 2022.

Anji Liu, Oliver Broadrick, Mathias Niepert, and Guy Van den Broeck. Discrete copula diffusion. In *International Conference on Learning Representations*, 2025.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning*, 2024.

Albert W. Marshall and Ingram Olkin. *Inequalities: Theory of Majorization and Its Applications*. Academic Press, 1979.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978.

Robert G. Newcombe. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine*, 17(22):2635–2650, 1998.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 50608–50646, 2025.

Shen Nie, Qiyang Min, Shaoxuan Xu, Zihao Huang, Yuxuan Song, Yong Shan, Yankai Lin, Wayne Xin Zhao, Chongxuan Li, and Ji-Rong Wen. Improved large language diffusion models. *arXiv preprint arXiv:2606.25331*, 2026.

OpenCompass Contributors. OpenCompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

Fred Zhangzhi Peng, Zachary Bezemek, Sawan Patel, Jarrid Rector-Brooks, Sherwood Yao, Avishek Joey Bose, Alexander Tong, and Pranam Chatterjee. Path planning for masked diffusion model sampling. *arXiv preprint arXiv:2502.03540*, 2025.

Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Houxing Ren, Mingjie Zhan, Zimu Lu, Ke Wang, Yunqiao Yang, Haotian Hou, Junting Pan, and Hongsheng Li. Edit-based refinement for parallel masked diffusion language models. *arXiv preprint arXiv:2605.09603*, 2026.

Liran Ringel, Ameen Ali, and Yaniv Romano. Dependency-guided parallel decoding in discrete diffusion language models. *arXiv preprint arXiv:2604.02560*, 2026.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024.

Shreshth Saini, Avinab Saha, Balu Adsumilli, Neil Birkbeck, Yilin Wang, and Alan C. Bovik. TABES: Trajectory-aware backward-on-entropy steering for masked diffusion models. *arXiv preprint arXiv:2602.00250*, 2026.

Jai Sharma, Yifan Wang, and Bryan Li. Neural estimation of pairwise mutual information in masked discrete sequence models. *arXiv preprint arXiv:2605.20187*, 2026.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Lucas Torroba Hennigen and Yoon Kim. Deriving language models from masked language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1149–1159, 2023.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Greg Ver Steeg and Aram Galstyan. Discovering structure in high-dimensional data through correlation explanation. In *Advances in Neural Information Processing Systems*, volume 27, pp. 577–585, 2014.

Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82, 1960.

Chuling Wen, Weijie Liang, and Jian Lu. Conditional total correlation and the serial depth of adaptive parallel sampling. *arXiv preprint arXiv:2608.25505*, 2026.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Andreas Winter. Tight uniform continuity bounds for quantum entropies: Conditional entropy, relative entropy distance and energy constraints. *Communications in Mathematical Physics*, 347 (1):291–313, 2016.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dLLM: Training-free acceleration of diffusion LLM by enabling KV cache and parallel decoding. In *International Conference on Learning Representations*, 2026.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7B: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

Yushi Ye, Xu Chen, Haoyun Jiang, Jinsong Lan, Haihong Tang, Bo Han, Ivor Tsang, Yanfeng Wang, Bo Zheng, and Jiangchao Yao. Ripple-pivot search: Active parallel decoding for diffusion large language models. *arXiv preprint arXiv:2608.11742*, 2026.

Tom Young, Yunan Chen, and Yang You. Inconsistencies in masked language models. *arXiv preprint arXiv:2301.00068*, 2024.

Shaorong Zhang, Longxuan Yu, Rob Brekelmans, Luhan Tang, Salman Asif, and Greg Ver Steeg. Generation order and parallel decoding in masked diffusion models: An information-theoretic perspective. *arXiv preprint arXiv:2602.00286*, 2026.

Shuibai Zhang, Fred Zhangzhi Peng, Yiheng Zhang, Jin Pan, and Grigorios G. Chrysos. Corrective diffusion language models. *arXiv preprint arXiv:2512.15596*, 2025.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. *arXiv preprint arXiv:2409.02908*, 2024.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

# A PROOFS FOR SECTIONS 2 AND 3

Appendices A to C prove the results of Sections 2 to 5 in the order in which they appear, and Appendix D contains the trajectory-level statements. Results cited from prior work are used as lemmas and are identified where they are used.

## A.1 PROOF OF THEOREM 3

*Proof.* Write $x = (x_i)_{i\in S}$. By Assumption 1, every $\widehat{P}_i$ is strictly positive, so all quantities below are finite. Marginalizing over the coordinates other than $i$ yields $\sum_x P(x)\log_2 \widehat{P}_i(x_i) = -\operatorname{CE}(P_i, \widehat{P}_i)$, and therefore

$$\operatorname{KL}(P\|Q) = \sum_{i\in S} \operatorname{CE}(P_i, \widehat{P}_i) - H(P). \tag{8}$$

(i) Substituting $\operatorname{CE}(P_i, \widehat{P}_i) = H(P_i) + \operatorname{KL}(P_i\|\widehat{P}_i)$ and $\operatorname{TC}(P) = \sum_{i\in S} H(P_i) - H(P)$ gives (i).

(ii) Nonnegativity of TC is the subadditivity of entropy. For any $a \in S$, the chain rule gives $H(P) = H(P_a) + H(X_{S\setminus\{a\}} \mid X_a) \geq H(P_a)$, with equality if and only if $X_S$ is $P$-almost surely a function of $X_a$. Choosing $a \in \arg\max_{i\in S} H(P_i)$ gives $\operatorname{TC}(P) \leq e^\star(S)$ and the equality condition. Conversely, if $X_S$ is a function of $X_a$ for some $a$, then $H(P) = H(P_a) \geq H(P_i)$ for every $i$, so this $a$ attains the maximum.

(iii) Pinsker's inequality (Cover & Thomas, 2006, Ch. 11) gives $\operatorname{TV}(Q, P) \leq \sqrt{\operatorname{KL}(P\|Q)_{\text{nats}}/2}$, and $\operatorname{KL}_{\text{nats}} = \ln 2 \cdot \operatorname{KL}_{\text{bits}}$; combining this with (i) and (ii) yields the stated bound.

(iv) If $\widehat{P}_i = P_i$, every $\operatorname{KL}(P_i\|\widehat{P}_i)$ vanishes and $\widehat{h}_i = H(P_i)$, so $e(S \mid z) = e^\star(S)$ and (iii) reduces to the stated bound. Moreover, $e(S \mid z) = 0$ if and only if $H(P_i) = 0$ for every $i$ other than a maximizer; then $\operatorname{TC}(P) \leq e^\star(S) = 0$, and $\operatorname{TC}(P) = 0$ means $P = \bigotimes_{i\in S} P_i = Q$.

(v) By (8) and $H(P) \geq H(P_a)$, $\operatorname{KL}(P\|Q) \leq \sum_{i\in S\setminus\{a\}} \operatorname{CE}(P_i, \widehat{P}_i) + [\operatorname{CE}(P_a, \widehat{P}_a) - H(P_a)]$, and the bracket equals $\operatorname{KL}(P_a\|\widehat{P}_a)$; equality holds if and only if $H(P) = H(P_a)$, that is, if and only if $X_S$ is a function of $X_a$. □

## A.2 THE MAJORIZATION MEET, AND THE PROOFS OF PROPOSITION 6 AND COROLLARY 7

**Lemma 19** (Majorization meet)**.** *Let $\widehat{P}_1, \ldots, \widehat{P}_m$ be probability vectors, zero-padded to a common length $N$, and let $g = \bigwedge_i \widehat{P}_i$ be defined by $\sum_{l\leq k} g(l) = \min_i \sum_{l\leq k} \widehat{P}_i^{\downarrow}(l)$ for $k = 1, \ldots, N$. Then:*

*(a) $g$ is a probability vector with nonincreasing entries;*
*(b) $g \prec \widehat{P}_i$ for every $i$, and hence $H(g) \geq \max_i H(\widehat{P}_i)$;*
*(c) every coupling $P$ of $\widehat{P}_1, \ldots, \widehat{P}_m$ satisfies $P \prec g$, and hence $H(P) \geq H(g)$.*

*Proof.* (a) For each $i$, the map $k \mapsto G_i(k) := \sum_{l\leq k} \widehat{P}_i^{\downarrow}(l)$ is nondecreasing, satisfies $G_i(0) = 0$ and $G_i(N) = 1$, and is concave in the sense that its increments are nonincreasing. The pointwise minimum $G(k) = \min_i G_i(k)$ is therefore nondecreasing and concave, because the piecewise-linear interpolants of the $G_i$ are concave on $[0, N]$ and their minimum agrees with $G$ at the integers. Hence the increments $g(k) = G(k) - G(k-1)$ are nonnegative and nonincreasing, and $G(0) = 0$, $G(N) = 1$.

(b) By construction, $G(k) \leq G_i(k)$ for every $k$ and $i$, which is $g \prec \widehat{P}_i$; entropy is Schur-concave (Marshall & Olkin, 1979, Ch. 3).

(c) Fix $i$, a coupling $P$, and $k \geq 1$, and let $A$ be a set of $k$ atoms attaining $\sum_{l\leq k} P^{\downarrow}(l)$. The atoms of $A$ lie in at most $k$ of the blocks $B_v = \{x : x_i = v\}$, say $B_{v_1}, \ldots, B_{v_j}$ with $j \leq k$, so $\sum_{l\leq k} P^{\downarrow}(l) \leq \sum_{t=1}^{j} \widehat{P}_i(v_t) \leq \sum_{l\leq k} \widehat{P}_i^{\downarrow}(l)$. Hence $P \prec \widehat{P}_i$ for every $i$, the partial sums of $P^{\downarrow}$ are dominated by $G$, and $P \prec g$; Schur-concavity gives $H(P) \geq H(g)$. □

*Proof of Proposition 6.* Let $P$ be any joint law with marginals $\widehat{P}_i$, $i \in S$. By (1), $\mathrm{TC}(P) = \sum_{i\in S} \widehat{h}_i - H(P)$. By the definition of a minimum-entropy coupling, $H(P) \geq H_{\mathrm{MEC}}$, so $\mathrm{TC}(P) \leq e_{\mathrm{opt}}(S)$, and the bound is attained by a minimum-entropy coupling, so no smaller quantity bounds $\mathrm{TC}(P)$ for every such $P$. By Lemma 19(c), $H_{\mathrm{MEC}} \geq H(g)$, whence $e_{\mathrm{opt}}(S) \leq e_{\mathrm{maj}}(S)$; by Lemma 19(b), $H(g) \geq \max_{i\in S} \widehat{h}_i$, whence $e_{\mathrm{maj}}(S) \leq e(S \mid z)$. The profile satisfies $H(\mathrm{PROF}_S) \leq H_{\mathrm{MEC}}$ (Compton et al., 2023, Thm. 3.4), and the major-profile, the minimum-entropy distribution whose sketch lies below the profile, satisfies $\max(H(\mathrm{PROF}_S), H(g)) \leq H(\mathrm{MPROF}_S) \leq H_{\mathrm{MEC}}$ (Compton et al., 2023, Def. 5.4 and Thm. 5.5); since a larger lower bound on the coupling entropy yields a smaller certificate, $e_{\mathrm{opt}}(S) \leq e_{\mathrm{MP}}(S) \leq \min(e_{\mathrm{prof}}(S), e_{\mathrm{maj}}(S))$. The greedy coupling $G_S$ is a coupling of the marginals, so $H_{\mathrm{MEC}} \leq H(G_S)$, which gives $e_{\mathrm{G}}(S) \leq e_{\mathrm{opt}}(S)$.

For the constants, Compton et al. (2023, Thms. 4.1 and 4.2) prove $H(G_S) \leq H(\mathrm{PROF}_S) + \kappa_m$, which is $e_{\mathrm{prof}}(S) - e_{\mathrm{G}}(S) \leq \kappa_m$; Compton (2022) proves $H(G_S) \leq H(g) + \log_2 e$, which is $e_{\mathrm{maj}}(S) - e_{\mathrm{G}}(S) \leq \log_2 e$; and for $m = 2$, Cicalese et al. (2019) construct a coupling whose entropy exceeds $H(g)$ by at most one bit, which gives $e_{\mathrm{maj}}(S) - e_{\mathrm{opt}}(S) \leq 1$. Since $e_{\mathrm{MP}} \leq e_{\mathrm{prof}}$, these give $e_{\mathrm{G}} \leq e_{\mathrm{opt}} \leq e_{\mathrm{MP}} \leq e_{\mathrm{G}} + \kappa_m$.

For the computational claim, sorting each $\widehat{P}_i$ costs $O(|\mathcal{V}| \log |\mathcal{V}|)$, the meet costs $O(m|\mathcal{V}|)$, and the profile and the major-profile take $\widetilde{O}(m|\mathcal{V}|)$ time (Compton et al., 2023, App. F.3). The greedy coupling has at most $m|\mathcal{V}| - (m-1)$ atoms (Kocaoglu et al., 2017; Compton et al., 2023), and each step updates $m$ sorted lists in $O(m \log |\mathcal{V}|)$ time. For $3 \leq m \leq 11$, Compton et al. (2023, App. D) give constants smaller than $\kappa_m$, approximately 0.77, 0.90 and 0.99 for $m = 3, 4, 5$, which may replace $\kappa_m$ in (2) and in Corollary 11. □

*Proof of Corollary 7.* Taking $P$ to be a minimum-entropy coupling gives $\beta(S) \geq \mathrm{TC}(P) = e_{\mathrm{opt}}(S)$, and Proposition 6 gives the remaining inequalities. Compton (2022) shows that the gap $H_{\mathrm{MEC}} - H(g)$ can approach $\log_2 e$ as the number of marginals grows, which is the final claim. □

**Neither $e_{\mathrm{prof}}$ nor $e_{\mathrm{maj}}$ dominates the other.** With $H(\mathrm{PROF}_S) = \int_0^1 \log_2(1/\mathrm{PROF}_S(x))\, dx$ as defined by Compton et al. (2023), their pair $p = (\frac{1}{2}, \frac{1}{2})$ and $q = (\frac{3}{4}, \frac{1}{20}, \frac{1}{20}, \frac{1}{20}, \frac{1}{20}, \frac{1}{20})$ (Compton et al., 2023, App. G.3, Remark 4) gives $H(\bigwedge) = 1.885475 > H(\mathrm{PROF}_S) = 1.830482$, so $e_{\mathrm{maj}} = 0.506285 < e_{\mathrm{prof}} = 0.561278$, whereas their running example $p = (0.5, 0.4, 0.1)$, $q = (0.6, 0.2, 0.2)$ (Compton et al., 2023, Fig. 1) gives $H(\bigwedge) = 1.485475 < H(\mathrm{PROF}_S) = 1.660964$, so $e_{\mathrm{prof}} = 1.070951 < e_{\mathrm{maj}} = 1.246439$. For two marginals, the gap between the minimum-entropy coupling and the meet can reach approximately 0.66 bits (Compton et al., 2023, App. H).

### A.3 Proofs of Propositions 5 and 8, and the binary example

*Proof of Proposition 5.* Every joint law $P$ with these entropies satisfies $\mathrm{TC}(P) \leq \sum_i h_i - \max_i h_i$ by Theorem 3(ii), so it suffices to exhibit a law that attains the bound. Put $k_j = \lceil 2^{\delta_j} \rceil$. The entropy of a law on $k_j$ symbols ranges continuously from 0 for a point mass to $\log_2 k_j \geq \delta_j$ for the uniform law, so there is a law $\rho_j$ on $k_j$ symbols with $H(\rho_j) = \delta_j$. Let $Z_1, \ldots, Z_m$ be independent with $Z_j \sim \rho_j$. The vector $Y_j = (Z_1, \ldots, Z_j)$ takes at most $\prod_{l\leq j} k_l \leq |\mathcal{V}|$ values, so it can be encoded injectively into $\mathcal{V}$; assign the encoded $Y_j$ to the position of rank $j$. Then $H(Y_j) = \sum_{l\leq j} \delta_l = h_{(j)}$, every coordinate is a function of $Y_m$, and the joint entropy equals $H(Y_m) = h_{(m)}$, so $\mathrm{TC} = \sum_i h_i - \max_i h_i$. For the sufficient condition, $\lceil x \rceil < 2x$ for $x \geq 1$ gives $\prod_j k_j < 2^{\sum_j \delta_j + m} = 2^{\max_i h_i + m} \leq |\mathcal{V}|$. □

*Proof of Proposition 8.* Fix $s \geq 2$, write $m = 2^s$ and $N = m^2 + 1$, and view a coupling of $P_s$ and $Q_s$ as a nonnegative $N \times N$ matrix $\Pi$ with row sums $P_s$ and column sums $Q_s$. Only the $m$ rows in the support of $P_s$ carry mass, each with row sum $1/m$; one column, the heavy column, has sum $\frac{1}{2}$, and each of the $m^2$ light columns has sum $1/(2m^2)$.

*(i) Lower bound.* Each entry $\pi$ of the heavy column satisfies $\pi \leq 1/m$, so $-\pi \log_2 \pi \geq \pi \log_2 m$, and the heavy column contributes at least $\frac{1}{2} \log_2 m = \frac{s}{2}$ to $H(\Pi)$. Each entry $\pi$ of a light column satisfies $\pi \leq 1/(2m^2)$, so the light columns, which carry total mass $\frac{1}{2}$, contribute at least $\frac{1}{2} \log_2(2m^2) = \frac{2s+1}{2}$. Hence $H(\Pi) \geq \frac{3s+1}{2}$.

*(i) Attainment.* Assigning the row mass $1/m$ of each of the first $m/2$ support rows to the heavy column fills it, and distributing the row mass of each of the remaining $m/2$ rows evenly across $2m$ distinct light columns uses all $m^2$ light columns, with each receiving its column sum $1/(2m^2)$. The coupling has entropy $\frac{1}{2}\log_2 m + \frac{1}{2}\log_2(2m^2) = \frac{3s+1}{2}$, so $H_{\mathrm{MEC}} = \frac{3s+1}{2}$ and $e_{\mathrm{opt}} = s + (s+1) - \frac{3s+1}{2} = \frac{s+1}{2}$.

*(ii) Entropy-only information.* With $\delta_1 = s$ and $\delta_2 = 1$, $\lceil 2^s \rceil \lceil 2 \rceil = 2^{s+1} \le N$, so Proposition 5 gives the tightest entropy-only certificate $e(S) = s$; the construction is the pair of a uniform variable on $2^s$ symbols and an independent fair bit. The difference from (i) is $s - \frac{s+1}{2} = \frac{s-1}{2}$.

*Full support.* For $\varepsilon \in (0,1)$, let $P_s^\varepsilon = (1-\varepsilon)P_s + \varepsilon U_N$ and $Q_s^\varepsilon = (1-\varepsilon)Q_s + \varepsilon U_N$, where $U_N$ is uniform on the $N$ symbols. For all marginals, $e - e_{\mathrm{opt}} = H_{\mathrm{MEC}} - \max\{H(p), H(q)\} \ge H(p \wedge q) - \max\{H(p), H(q)\}$ by Lemma 19(c), and the right-hand side is continuous in $(p, q)$. At $\varepsilon = 0$, the partial sums are $G_P(k) = \min\{k2^{-s}, 1\}$ and $G_Q(k) = \frac{1}{2} + (k-1)2^{-(2s+1)}$, the minimum equals $G_P$ for $k \le 2^{s-1}$ and $G_Q$ otherwise, and the meet consists of $2^{s-1}$ entries $2^{-s}$, one entry $2^{-(s+2)}$ and $2^{2s} - 2^{s-1}$ entries $2^{-(2s+1)}$, with entropy $\frac{3s+1}{2} - \frac{s-1}{2^{s+2}}$. The right-hand side therefore equals $(s-1)(\frac{1}{2} - 2^{-(s+2)})$ at $\varepsilon = 0$, and it exceeds $(s-1)(\frac{1}{2} - 2^{-(s+2)}) - \delta$ for all sufficiently small $\varepsilon$. The smoothed pair also keeps the entropy-only optimum of (ii): as $\varepsilon \to 0$, $\delta_1 = H(P_s^\varepsilon) \to s$ and $\delta_2 = H(Q_s^\varepsilon) - H(P_s^\varepsilon) \to 1$, so for small $\varepsilon$ we have $\lceil 2^{\delta_1} \rceil \le 2^s + 1$ and $\lceil 2^{\delta_2} \rceil \le 3$, and $3(2^s + 1) \le 2^{2s} + 1 = N$ for $s \ge 2$. By Proposition 5, the tightest entropy-only certificate of the smoothed pair is its excess bits, so the separation between the entropy-only and the marginal-only optimum equals $e - e_{\mathrm{opt}}$ and satisfies the claim.

*Fixed alphabet.* For two marginals $p, q$ on $d$ symbols, the independent coupling has entropy $H(p) + H(q) \ge H_{\mathrm{MEC}}$, so $e_{\mathrm{opt}} \ge 0$ and $e - e_{\mathrm{opt}} \le e = \min\{H(p), H(q)\} \le \log_2 d$. In the family above, $e - e_{\mathrm{opt}} = \frac{s-1}{2}$ and $\log_2 N \le 2s + 1$, so the largest separation is of order $\log d$. □

**The binary example of Section 3.** On two symbols, the binary entropy $h_2$ is strictly increasing on $[0, \frac{1}{2}]$, so an entropy value determines the marginal up to relabeling, which does not change the maximal total correlation. The tightest entropy-only certificate at entropies $(h_2(0.1), 1)$ is therefore $e_{\mathrm{opt}}$ for the marginals $(0.1, 0.9)$ and $(0.5, 0.5)$. Their couplings are $\left(\begin{smallmatrix} t & 0.1-t \\ 0.5-t & 0.4+t \end{smallmatrix}\right)$ with $t \in [0, 0.1]$; entropy is concave in $t$, so its minimum is attained at $t \in \{0, 0.1\}$, where it equals $0.1\log_2 10 + 0.5 + 0.4\log_2 2.5 \approx 1.360964$. Hence the tightest certificate is $h_2(0.1) + 1 - 1.360964 \approx 0.108032$, whereas $e = h_2(0.1) \approx 0.468996$. Here $\lceil 2^{h_2(0.1)} \rceil \lceil 2^{1-h_2(0.1)} \rceil = 4 > 2$, so the condition of Proposition 5 fails.

# B Proofs for Section 4

## B.1 Proof of Theorem 9

*Proof.* Order $S$ as $a_1, \ldots, a_{|S|}$, extending the given anchors arbitrarily. The chain rule gives $H(P) = H(X_{A_T}) + H(X_{S \setminus A_T} \mid X_{A_T})$ and $H(X_{A_T}) = \sum_{t \le T} H(X_{a_t} \mid X_{a_1}, \ldots, X_{a_{t-1}})$. Substituting into $\mathrm{TC}(P) = \sum_{i \in S} H(P_i) - H(P)$ gives (3), so $e_T$ depends only on $A_T$ and is exact if and only if $X_S$ is $P$-almost surely a function of $X_{A_T}$. Monotonicity in $T$ follows from $H(X_{A_T}) \le H(X_{A_{T+1}})$. For $|S| = 2$ and $T = 2$, $e_2(S) = I(X_{a_1}; X_{a_2}) = \mathrm{TC}(P)$. Finally, each conditional entropy in (3) is a sum of nonnegative terms over the values of the conditioning coordinates; retaining only the $k$ most probable values decreases it and increases the right-hand side of (3), so the bound remains valid. □

## B.2 Greedy anchor selection

**Corollary 20** (Optimal anchors and greedy guarantees)**.** *Fix $z$ and $S$, and set $F(A) = H(X_A \mid z)$ for $A \subseteq S$, so that $e_T(S; A) = \sum_{i \in S} H(P_i) - F(A)$ by (3). Then $F(\emptyset) = 0$ and $F$ is monotone and submodular, so the tightest depth-$T$ certificate maximizes $F$ over $|A| = T$. Write $F^\star = \max_{|A|=T} F(A)$.*

*(a)* Exact queries. *The greedy set $A^{\mathrm{g}}$, which appends $a_t \in \arg\max_{i \notin A_{t-1}} H(X_i \mid X_{A_{t-1}}, z)$ at every level, satisfies*

$$e_T\big(S; A^{\mathrm{g}}\big) - \min_{|A|=T} e_T(S; A) \;=\; F^\star - F(A^{\mathrm{g}}) \;\le\; F^\star / e. \tag{9}$$

*(b)* Truncated queries. *If at every level the score used in place of $H(X_i \mid X_{A_{t-1}}, z)$ is a quantity $\widehat{H}$ with $0 \le H(X_i \mid X_{A_{t-1}}, z) - \widehat{H} \le \varepsilon$ for every candidate $i$, the resulting set $A^{\mathrm{tg}}$ satisfies*

$$e_T\big(S; A^{\mathrm{tg}}\big) - \min_{|A|=T} e_T(S; A) \;\le\; F^\star/e + T\varepsilon. \tag{10}$$

*For the top-$k$ truncation of Theorem 9, evaluated on retained tuples that carry probability mass at least $w$ at every level, $\varepsilon \le (1-w)\log_2 |\mathcal{V}|$.*

*Proof.* All entropies are conditional on $z$, which we suppress. $F(\emptyset) = 0$; the chain rule gives $F(A \cup \{i\}) - F(A) = H(X_i \mid X_A) \ge 0$; and for $A \subseteq B$ and $i \notin B$, $H(X_i \mid X_A) \ge H(X_i \mid X_B)$ because conditioning does not increase entropy, which is submodularity. Part (a) is the guarantee of Nemhauser et al. (1978) for monotone submodular functions with $F(\emptyset) = 0$, since the marginal gain of $i$ at level $t$ is $H(X_i \mid X_{A_{t-1}})$.

For (b), let $g_t(i) = H(X_i \mid X_{A_{t-1}})$ be the true gain and $\widehat{g}_t(i)$ the score, with $0 \le g_t - \widehat{g}_t \le \varepsilon$. If $\widehat{a}_t$ maximizes $\widehat{g}_t$ and $i_t^\star$ maximizes $g_t$, then $g_t(\widehat{a}_t) \ge \widehat{g}_t(\widehat{a}_t) \ge \widehat{g}_t(i_t^\star) \ge g_t(i_t^\star) - \varepsilon$. Put $\delta_t = F^\star - F(A_t)$ and let $A^\star$ attain $F^\star$. Monotonicity and submodularity give $F^\star \le F(A^\star \cup A_t) \le F(A_t) + T \max_i g_{t+1}(i)$, so $\delta_{t+1} \le \delta_t(1-1/T) + \varepsilon$ and $\delta_T \le F^\star(1-1/T)^T + \varepsilon \sum_{j<T}(1-1/T)^j \le F^\star/e + T\varepsilon$. For the truncation constant, the discarded tuples carry mass at most $1-w$ and each discarded conditional entropy is at most $\log_2 |\mathcal{V}|$. The scores are the quantities that the depth-$T$ evaluation of (3) computes, so neither variant requires additional queries. □

*Remark* 21 (Failure of the multiplicative guarantee under truncation). Let $A$ be uniform on four values, $C = A$, and $B = 0$ when $A = 0$ and uniform on four values otherwise, so that $H(A) = H(C) = 2$, $H(B) \approx 1.880$, $H(A, B) = 3.5$ and $H(A, C) = 2$. With $T = 2$, the first greedy anchor is $A$; retaining only its most probable value, $A = 0$ after the index tie-break, both remaining positions have conditional entropy 0, so the truncated scores tie and the rule may append $C$. Then $F(\{A, C\})/F^\star = 4/7 < 1 - 1/e$, whereas the exact greedy rule appends $B$ and attains $F^\star$.

Corollary 20 applies to exact conditional entropies of a single joint law. When the clamped conditionals of a learned denoiser are not coherent (Definition 12), they need not be the conditionals of any joint law, and $F$, hence the guarantee, is not defined.

### B.3 Depth-2 certificates and the marginal-only optimum

*Remark* 22 (Depth-2 certificates and the marginal-only optimum). Let $S = \{a, b, c\}$ with $X_a = X_b$ uniform on four values and $X_c$ uniform on three values and independent of $X_a$. Then $\mathrm{TC}(P) = 2$, $e(S) = e_{\mathrm{maj}}(S) \approx 3.585$ and $e_{\mathrm{opt}}(S) \approx 3.189$. The anchors of largest marginal entropy, $a$ and $b$, give $e_2(S) = e(S) > e_{\mathrm{opt}}(S)$, whereas the greedy rule selects $c$ and gives $e_2(S) = \mathrm{TC}(P)$.

We verify the claims.

Let $X_a = X_b$ be uniform on $\{1, 2, 3, 4\}$ and $X_c$ uniform on $\{1, 2, 3\}$ and independent of $X_a$. Then $\sum_i H(P_i) = 4 + \log_2 3 \approx 5.585$, $H(P) = 2 + \log_2 3$, $\mathrm{TC}(P) = 2$ and $e(S) = 2 + \log_2 3 \approx 3.585$. The partial sums of the sorted marginals are $k/4$, $k/4$ and $\min\{k/3, 1\}$, whose minimum is $k/4$, so the meet is uniform on four symbols and $e_{\mathrm{maj}}(S) = e(S)$.

Every coupling of $(U_4, U_4, U_3)$, where $U_d$ is uniform on $d$ symbols, projects onto a coupling of $(U_4, U_3)$ of no larger entropy, and every coupling of $(U_4, U_3)$ extends to one of $(U_4, U_4, U_3)$ with the same entropy by repeating the first coordinate, so $H_{\mathrm{MEC}}(U_4, U_4, U_3) = H_{\mathrm{MEC}}(U_4, U_3)$. Entropy is concave, so its minimum over the transportation polytope of $(U_4, U_3)$ is attained at a vertex; enumerating the vertices shows that the minimum equals $\frac{3}{2} + \frac{1}{4}\log_2 12 \approx 2.396$, attained by the coupling with mass $\frac{1}{4}$ on $(i, i)$ for $i \le 3$ and mass $\frac{1}{12}$ on $(4, j)$ for $j \le 3$. Hence $e_{\mathrm{opt}}(S) = 4 + \log_2 3 - \frac{3}{2} - \frac{1}{4}\log_2 12 \approx 3.189 < e(S)$.

The two positions of largest marginal entropy are $a$ and $b$, and $H(X_a, X_b) = 2$, so $e_2(S; a, b) = e(S)$. The greedy rule keeps $a$ and appends $c$, because $H(X_b \mid X_a) = 0 < H(X_c \mid X_a) = \log_2 3$, and $e_2(S; a, c) = 2 = \mathrm{TC}(P)$. Given $X_a = v$, the conditional marginal of $X_b$ is a point mass and that of $X_c$ is uniform on three values, so the conditional coupling and meet both have entropy $\log_2 3$, and Proposition 10 gives $e_a^{\mathrm{opt}}(S) = e_a^{\mathrm{maj}}(S) = 2 = \mathrm{TC}(P)$ from the same passes that give $e_2(S; a, b) = e(S)$.

### B.4 Proof of Proposition 10

*Proof.* (i) Let $\mathcal{P}_a$ be the set of joint laws $P'$ of $X_S$ with $P'(X_a = \cdot) = P_a$ and $P'(X_i = \cdot \mid X_a = v) = P_{i|a=v}$ for every $i \neq a$ and every $v$ with $P_a(v) > 0$. Every $P' \in \mathcal{P}_a$ has marginals $\sum_v P_a(v) P_{i|a=v} = P_i$, and by the chain rule $H(P') = H(P_a) + \sum_v P_a(v)\, H(P'(X_{S\setminus\{a\}} = \cdot \mid X_a = v))$, where each conditional law is a coupling of $\{P_{i|a=v}\}_{i\neq a}$. Hence $\mathrm{TC}(P') \leq e_a^{\mathrm{opt}}(S)$, and choosing a minimum-entropy coupling at each $v$, which exists because the set of couplings is compact and entropy is continuous, attains the bound. Every law in $\mathcal{P}_a$ has the marginals $P_i$, so $e_a^{\mathrm{opt}}(S) \leq e_{\mathrm{opt}}(S)$ by Proposition 6. The entropy of a coupling is at least the entropy of each of its marginals, so averaging over $v$ gives $e_a^{\mathrm{opt}}(S) \leq e_2(S; a, b)$ for every $b \neq a$. Lemma 19(c) gives $e_a^{\mathrm{maj}} \geq e_a^{\mathrm{opt}}$, Lemma 19(b) gives $e_a^{\mathrm{maj}} \leq e_a^{\mathrm{env}}$, and $\sum_v P_a(v) \max_{i\neq a} H(P_{i|a=v}) \geq H(X_b \mid X_a)$ gives $e_a^{\mathrm{env}} \leq e_2(S; a, b)$. The constant follows by applying Proposition 6 to the $|S| - 1$ conditional marginals at each $v$; the meet of a single vector is that vector, so $c_1 = 0$.

(ii) Write $D = \mathcal{V} \setminus R$, and let $\mathcal{P}_R$ be the set of joint laws $P'$ of $X_S$ with marginals $P_i$ for every $i \in S$ and $P'(X_i = \cdot \mid X_a = v) = P_{i|a=v}$ for $i \neq a$ and $v \in R$. For $P' \in \mathcal{P}_R$, let $C_v$ be the conditional law of $X_{S\setminus\{a\}}$ given $X_a = v$, and let $T$ be the conditional law of $X_S$ given $X_a \in D$. Each $C_v$, $v \in R$, is a coupling of $\{P_{i|a=v}\}_{i\neq a}$, and $T$ has $X_a$-marginal $q_a$ and, because $P_i = \sum_{v\in R} P_a(v) P_{i|a=v} + r T_i$, $X_i$-marginal $q_i$. Conversely, for any couplings $C_v$ of the queried conditionals and any coupling $T$ of $(q_a, \{q_i\}_{i\neq a})$, the law $\sum_{v\in R} P_a(v)\, \delta_v \otimes C_v + r\, T$ belongs to $\mathcal{P}_R$. By the chain rule,

$$H(P') = H(P_a) + \sum_{v\in R} P_a(v)\, H(C_v) + \sum_{v\in D} P_a(v)\, H\big(P'(X_{S\setminus\{a\}} = \cdot \mid X_a = v)\big),$$

and since $H(T) = H(q_a) + \sum_{v\in D} q_a(v)\, H(P'(X_{S\setminus\{a\}} = \cdot \mid X_a = v))$, the last sum equals $r[H(T) - H(q_a)]$. The terms are minimized independently by minimum-entropy couplings, which gives $\min_{P'\in\mathcal{P}_R} H(P')$ and hence (4). For $R = \emptyset$, $r = 1$, $q_a = P_a$ and $q_i = P_i$, so $\beta_\emptyset^\star = \sum_i H(P_i) - H_{\mathrm{MEC}}(\{P_i\}_{i\in S}) = e_{\mathrm{opt}}(S)$. The feasible families are nested, $\mathcal{P}_a \subseteq \mathcal{P}_R \subseteq \mathcal{P}_\emptyset$, so $e_a^{\mathrm{opt}} \leq \beta_R^\star \leq e_{\mathrm{opt}}$. Replacing $H_{\mathrm{MEC}}$ by any lower bound on it can only increase the right-hand side of (4), which preserves validity. □

*Proof of Corollary 11.* Every $H_{\mathrm{MEC}}$ in (4) enters with a nonpositive coefficient, $-P_a(v)$ or $-r$. The major-profile entropy is a lower bound and the greedy entropy an upper bound on the entropy of a minimum-entropy coupling (Appendix A.2). Replacing a term by its lower bound can only increase the right-hand side, and by its upper bound only decrease it, which gives $L_R \leq \widetilde{L}_R \leq \beta_R^\star \leq \widetilde{U}_R \leq U_R$; $U_R$ and $\widetilde{U}_R$ are certificates because $\beta_R^\star$ is. For a single marginal, both replacements are exact: the greedy coupling of one law is that law, and its profile entropy equals its entropy (Compton et al., 2023, Remark 1), so $H(\mathrm{PROF}) \leq H(\mathrm{MPROF}) \leq H_{\mathrm{MEC}}$ forces equality; this is $\kappa_1 = 0$. For $j \geq 2$ marginals, $H(G) - H(\mathrm{MPROF}) \leq H(G) - H(\mathrm{PROF}) \leq \kappa_j$ by Proposition 6. The conditional couplings at $v \in R$ involve $m - 1$ marginals with total weight $1 - r$, and the residual coupling involves $m$ marginals with weight $r$. Summing the gaps of all replaced terms gives $U_R - L_R \leq (1 - r)\kappa_{m-1} + r\kappa_m$, and summing only the residual gap gives $\widetilde{U}_R - \widetilde{L}_R \leq r\kappa_m$. For $m = 2$, $\kappa_1 = 0$, so the brackets coincide and have width at most $r\kappa_2 = r \log_2 e/e$. For $R = \emptyset$, where $r = 1$, (4) reduces to $\sum_i H(P_i) - H_{\mathrm{MEC}}(\{P_i\}_{i\in S})$, whose brackets are $[e_{\mathrm{G}}, e_{\mathrm{MP}}]$.

For the last claim, let $X_a$ be independent of $(X_b, X_c)$ with $P_a = (\frac{1-r}{2}, \frac{1-r}{2}, r)$, $P_b = (0.7, 0.25, 0.05)$ and $P_c = (0.6, 0.2, 0.2)$, and query the first two values of $X_a$. Every retained conditional pair is $(P_b, P_c)$, whose greedy coupling $(0.6, 0.2, 0.1, 0.05, 0.05)$ and major-profile $(0.6, 0.2, 0.15, 0.05)$ differ in entropy by $0.137744$ bits. The residual laws are a point mass and $(P_b, P_c)$, whose greedy coupling and major-profile are the same as before, so $U_R - L_R = 0.137744$ for every $r \in (0, 1)$, whereas $\widetilde{U}_R - \widetilde{L}_R = 0.137744\, r$. □

If exactly one anchor value $d$ is unqueried, $q_a$ is a point mass and $q_i = P_{i|a=d}$, so the residual term equals $P_a(d) H_{\mathrm{MEC}}(\{P_{i|a=d}\}_{i\neq a})$ and $\beta_R^\star = e_a^{\mathrm{opt}}$: the marginals determine the last conditional.

For two positions with exact outputs and $R = V_k(a)$, the discarding certificate $b_{a,k}^{\mathrm{mec}}$ defined below satisfies $b_{a,k}^{\mathrm{mec}} - \mathrm{TC}(P) = \sum_{v\notin R} P_a(v) H(P_{b|a=v}) = r\, H(X_b \mid X_a,\ X_a \notin R)$, so the truncation gap depends on the unqueried mass and on the conditional entropy of the unqueried part, and the residual term of (4) removes at most this amount.

**Truncated certificates without the residual constraint.** Retaining only $v \in V_k(a)$ and omitting the unqueried terms from the subtracted sum in (i) yields $b^{\mathrm{mec}}_{a,k} \le b^{\mathrm{maj}}_{a,k} \le b^{\mathrm{env}}_{a,k} \le \min_{b \neq a} e_{2,k}(S; a, b)$, all of which bound $\mathrm{TC}(P)$, with $b^{\mathrm{maj}}_{a,k} - b^{\mathrm{mec}}_{a,k} \le w_k(a)\, c_{|S|-1}$. They bound the residual term of (4) below by zero, so $\beta^{\star}_{V_k(a)} \le b^{\mathrm{mec}}_{a,k}$, and they need not lie below $e_{\mathrm{opt}}(S)$. Let $X_a$ and $X_b$ be independent with $P_a = (\frac{1}{2}, \frac{1}{2})$ and $P_b = (0.1, 0.9)$, and query one value of $X_a$. The residual law $q_b$ equals $P_b$ and $q_a$ is a point mass, so $\beta^{\star}_R = \mathrm{TC}(P) = 0$, whereas $b^{\mathrm{mec}}_{a,1} = \frac{1}{2} H(P_b) \approx 0.2345$ exceeds $e_{\mathrm{opt}}(S) \approx 0.1080$ (Appendix A.3).

**The envelope does not admit a constant gap.** Let $X_a$ be independent of two further coordinates whose conditional marginals are the $P_s$ and $Q_s$ of Proposition 8 at every value of $X_a$. Then $e^{\mathrm{env}}_a(S) = s$ and $e^{\mathrm{opt}}_a(S) = \frac{s+1}{2}$, so the envelope exceeds the optimum by $\frac{s-1}{2}$; in general only $e^{\mathrm{env}}_a - e^{\mathrm{opt}}_a \le (e^{\mathrm{env}}_a - e^{\mathrm{maj}}_a) + c_{|S|-1}$ holds.

### B.5 Proofs of Propositions 13 to 15

*Proof of Proposition 13.* (i) Entropy is concave, so Bayes-consistency gives

$$C_{\mathrm{full}}(u,t) = \sum_v \widehat{P}_u(v)\, H(\widehat{P}_{t|u=v}) \le H\Big(\sum_v \widehat{P}_u(v) \widehat{P}_{t|u=v}\Big) = \widehat{h}_t.$$

Every term of $C_{\mathrm{full}}(u,t)$ is nonnegative, so $C_k(u,t) \le C_{\mathrm{full}}(u,t)$, and substituting into (5) gives $d^{u\to t}_{2,k}(S) \ge \sum_{i \in S \setminus \{u,t\}} \widehat{h}_i \ge 0$. (ii) Bayes-consistency gives $\widehat{P}_t(x) = \sum_{v \in \mathcal{V}} \widehat{P}_u(v) \widehat{P}_{t|u=v}(x)$, a sum of nonnegative terms, and dropping the terms with $v \notin R$ gives the inequality. The marginals and conditionals of the target joint satisfy the law of total probability, so exact outputs are Bayes-consistent, and a violation of (i) or (ii) rules out exactness. □

*Proof of Proposition 14.* For $(u,t) \in \{(a,b),(b,a)\}$,

$$C_{\mathrm{full}}(u,t) - C_k(u,t) = \sum_{v \notin V_k(u)} \widehat{P}_u(v)\, \widehat{h}_{t|u=v} \in \big[0, (1 - w_k(u)) \log_2 |\mathcal{V}|\big],$$

and $d^{u\to t}_{2,\mathrm{full}} = d^{u\to t}_{2,k} - (C_{\mathrm{full}}(u,t) - C_k(u,t))$, so the complete score lies in $I^{u\to t}_k(S)$. If $(a,b)$ is jointly coherent with pair law $Q_{ab}$, the chain rule in both orders gives $\widehat{h}_a + C_{\mathrm{full}}(a,b) = H(Q_{ab}) = \widehat{h}_b + C_{\mathrm{full}}(b,a)$, so both complete scores equal $\sum_{i \in S} \widehat{h}_i - H(Q_{ab})$, and this common value lies in both intervals. For averages with common nonnegative weights $\omega_S$, if the pair is jointly coherent at every set of positive weight, a common point $y_S$ of the two intervals exists at each such set, and $\sum_S \omega_S y_S$ lies in both averaged intervals, because averaging preserves inequalities between endpoints; by contraposition, disjoint averaged intervals imply incoherence at some set of positive weight. □

The two formulations of joint coherence in Definition 12 agree: if $Q^{a\to b} = Q^{b\to a} = Q$, then $Q$ has marginals $\widehat{P}_a, \widehat{P}_b$ and both families of conditionals, and conversely such a law equals both products. Assumption 1 makes all conditionals well defined.

*Proof of Proposition 15.* (i) The statistic $D_k$ is a partial sum of the nonnegative terms $\frac{1}{2}|Q^{a\to b}(v,w) - Q^{b\to a}(v,w)|$, whose total is $\mathrm{TV}(Q^{a\to b}, Q^{b\to a})$. Every term is computable from the retained passes: for $v \in A_k$, the pass at $z[a \leftarrow v]$ returns $\widehat{P}_{b|a=v}(w)$ for every $w$, and for $w \in B_k$, the pass at $z[b \leftarrow w]$ returns $\widehat{P}_{a|b=w}(v)$ for every $v$. Under joint coherence every term vanishes. (ii) With $\widehat{P}_a = \widehat{P}_b = (\frac{1}{2}, \frac{1}{2})$, $Q^{a\to b}$ has entries $0.45$ on the diagonal and $0.05$ off the diagonal, and $Q^{b\to a}$ reverses these entries. Both have uniform marginals, so both directions are Bayes-consistent; every conditional has entropy $h_2(0.1)$, so the complete scores coincide and $\mathrm{IncLB}_k = 0$ for every $k$. Nevertheless $\mathrm{TV}(Q^{a\to b}, Q^{b\to a}) = 0.8$, and $D_k > 0$ whenever $A_k$ and $B_k$ share at least one value. □

### B.6 CERTIFICATES FOR A LEARNED DENOISER

Theorem 3(v) replaces Theorem 3(iv) when the denoiser marginals are inexact. For a singleton $S = \{a\}$, it reduces to $\mathrm{KL}(P_a \| \widehat{P}_a)$, whereas $e(S \mid z) = 0$, so the exact-marginal bound fails exactly at those singletons for which $\widehat{P}_a \neq P_a$. Because it involves the target marginals, it is an offline validation certificate: a single continuation does not identify $\mathrm{CE}(P_i, \widehat{P}_i)$ or $\mathrm{KL}(P_a \| \widehat{P}_a)$ at the state it visits, but these quantities can be estimated in expectation over a specified distribution of states, or pointwise where repeated conditional samples or a reference law are available. An inference-time version requires a calibration budget: if $\sum_{i \in S} |H(P_i) - H(\widehat{P}_i)| \leq \delta$, then $e^\star(S) \leq e(S \mid z) + 2\delta$.

The depth-2 certificate of Theorem 9 subtracts a conditional entropy that is read from the denoiser, which yields a valid bound only when that entropy is exact. The next two results restore validity under calibration budgets. We write $h_2(t) = -t \log_2 t - (1-t)\log_2(1-t)$, with $h_2(0) = h_2(1) = 0$, and $f_d(t) = t \log_2(d-1) + h_2(t)$ for an integer $d \geq 2$.

**Lemma 23** (Continuity of entropy under a distance budget)**.** *The function $f_d$ is nondecreasing on $[0, 1 - 1/d]$, and $f_d(1 - 1/d) = \log_2 d$. Put $\varphi_d(\varepsilon) = f_d(\min\{\varepsilon, 1 - 1/d\})$. For probability vectors $p$ and $q$ on $d$ symbols with $\mathrm{TV}(p, q) \leq \varepsilon$, we have $|H(p) - H(q)| \leq \varphi_d(\varepsilon)$.*

*Proof.* For $t \in (0, 1)$, $f_d'(t) = \log_2(d-1) + \log_2 \frac{1-t}{t} \geq 0$ if and only if $t \leq 1 - 1/d$, and a direct computation gives $f_d(1 - 1/d) = \log_2 d$. Let $T = \mathrm{TV}(p, q)$. The sharp continuity estimate of Audenaert (2007, Thm. 1), which holds for every $T \in [0, 1]$, gives $|H(p) - H(q)| \leq f_d(T)$. If $\varepsilon \leq 1 - 1/d$, monotonicity gives $f_d(T) \leq f_d(\varepsilon)$; otherwise $\varphi_d(\varepsilon) = \log_2 d$, which bounds the difference because both entropies lie in $[0, \log_2 d]$. This budget form also appears in Winter (2016, Lemma 1). □

**Proposition 24** (Calibration-aware depth-2 certificate)**.** *Fix a state $z$, a reveal set $S$ with $|S| \geq 2$, and distinct anchors $a, b \in S$. Let $\widehat{P}_{ab} = \widehat{P}_a \otimes \widehat{P}_{b|a}$ be the pair law with marginal $\widehat{P}_a$ and clamped conditionals $\widehat{P}_{b|a=v}$, and let $\widehat{H}_\theta(X_b \mid X_a) = \mathbb{E}_{v \sim \widehat{P}_a}[H(\widehat{P}_{b|a=v})]$. Let $\bar{\varepsilon}_{ab} \geq \mathrm{TV}(\widehat{P}_{ab}, P_{ab})$ and $\bar{\varepsilon}_a \geq \mathrm{TV}(\widehat{P}_a, P_a)$ be calibration budgets, and set*

$$\bar{\Delta} = \varphi_{|\mathcal{V}|^2}(\bar{\varepsilon}_{ab}) + \varphi_{|\mathcal{V}|}(\bar{\varepsilon}_a), \qquad L_+ = \max\left\{0,\ \widehat{H}_\theta(X_b \mid X_a) - \bar{\Delta}\right\}.$$

*Then*

$$\mathrm{KL}(P \,\|\, Q) \ \leq \ \sum_{i \in S \setminus \{a\}} \mathrm{CE}(P_i, \widehat{P}_i) + \mathrm{KL}(P_a \,\|\, \widehat{P}_a) - L_+, \tag{11}$$

*and $\mathrm{TV}(Q, P)$ is bounded by applying (15) to the right-hand side. The bound is never looser than Theorem 3(v), and it is strictly tighter if and only if $\widehat{H}_\theta(X_b \mid X_a) > \bar{\Delta}$. With exact outputs and zero budgets, it reduces to $\mathrm{TC}(P) \leq e_2(S)$ of Theorem 9, with equality if and only if $H(X_{S \setminus \{a,b\}} \mid X_a, X_b) = 0$. Replacing $\widehat{H}_\theta(X_b \mid X_a)$ by its truncation to the $k$ most probable values of $X_a$ preserves validity.*

*Proof.* By (8) and the chain rule $H(P) \geq H(P_a) + H(X_b \mid X_a)$, together with $H(P_a) = \mathrm{CE}(P_a, \widehat{P}_a) - \mathrm{KL}(P_a \| \widehat{P}_a)$,

$$\mathrm{KL}(P \| Q) \ \leq \ \sum_{i \in S \setminus \{a\}} \mathrm{CE}(P_i, \widehat{P}_i) + \mathrm{KL}(P_a \| \widehat{P}_a) - H(X_b \mid X_a). \tag{12}$$

Conditional entropy is nonnegative, and $H(X_b \mid X_a) = H(P_{ab}) - H(P_a)$ while $\widehat{H}_\theta(X_b \mid X_a) = H(\widehat{P}_{ab}) - H(\widehat{P}_a)$, so Lemma 23 on $|\mathcal{V}|^2$ and $|\mathcal{V}|$ symbols gives $H(X_b \mid X_a) \geq \widehat{H}_\theta(X_b \mid X_a) - \bar{\Delta}$; hence $H(X_b \mid X_a) \geq L_+$, which gives (11). Since $L_+ \geq 0$, the bound never exceeds that of Theorem 3(v). With exact outputs and zero budgets, $L_+ = H(X_b \mid X_a)$ and the right-hand side equals $e_2(S)$, while the left-hand side equals $\mathrm{TC}(P)$; the difference is $H(X_{S \setminus \{a,b\}} \mid X_a, X_b)$ by Theorem 9. Truncation discards nonnegative terms from $\widehat{H}_\theta$, which can only decrease $L_+$. □

For a learned denoiser, the clamped conditionals need not be coherent, so $\widehat{P}_a \otimes \widehat{P}_{b|a}$ and $\widehat{P}_b \otimes \widehat{P}_{a|b}$ may differ. Proposition 24 is valid for either order, with $\bar{\varepsilon}_{ab}$ referring to the pair law actually used; the uncorrected depth-2 substitution, which omits $\bar{\Delta}$, is not valid (Table 6).

# C Proofs for Section 5 and the sampler

## C.1 Proofs of Proposition 16 and Corollary 17

*Proof of Proposition 16.* Write $\widehat{h}_{(1)} \le \cdots \le \widehat{h}_{(m)}$ for the sorted entropies of $M = M(z)$, with ties broken by index, and $J = J(e)$.

*(a), upper bound.* Let $S \subseteq M$ be nonempty with $e(S \mid z) \le e$, put $t = |S| - 1$, and choose $a \in \arg\max_{i \in S} \widehat{h}_i$. Then $e(S \mid z) = \sum_{i \in S \setminus \{a\}} \widehat{h}_i \ge \sum_{l=1}^{t} \widehat{h}_{(l)}$, because the sum of any $t$ elements of $M$ is at least the sum of the $t$ smallest. Since $t \le m - 1$, the definition of $J$ gives $t \le J$, hence $|S| \le J + 1$.

*(a), attainment.* The sets $B$ and $\{(m)\}$ are disjoint because $J \le m - 1$, the largest entropy in $S^\star = B \cup \{(m)\}$ is $\widehat{h}_{(m)}$, and $e(S^\star \mid z) = \sum_{l=1}^{J} \widehat{h}_{(l)} \le e$.

*(b).* The prefix $U_t = \{(1), \ldots, (t)\}$ has $e(U_t \mid z) = \sum_{l=1}^{t-1} \widehat{h}_{(l)}$, which is at most $e$ if and only if $t - 1 \le J$. The longest admissible prefix therefore has length $J + 1$ and equals $B \cup \{(J + 1)\}$; its entropy mass is $\sum_{l=1}^{J+1} \widehat{h}_{(l)}$, whereas that of $S^\star$ is $\widehat{h}_{(m)} + \sum_{l=1}^{J} \widehat{h}_{(l)}$.

*(c).* Let $S$ have cardinality $J+1$ and largest-entropy position $a$, and put $T = S \setminus \{a\}$, so that $|T| = J$ and $\sum_{i \in T} \widehat{h}_i \le e$. If $J = 0$, then $T = B = \emptyset$. Let $J \ge 1$ and let the ranks of the elements of $T$ be $t_1 < \cdots < t_J$, so that $t_i \ge i$. If $T \ne B$, then $t_J \ge J + 1$, and since the entropies are distinct, $\sum_{i \in T} \widehat{h}_i \ge \sum_{i=1}^{J-1} \widehat{h}_{(i)} + \widehat{h}_{(J+1)} = \sum_{l=1}^{J} \widehat{h}_{(l)} + \widehat{h}_{(J+1)} - \widehat{h}_{(J)} > e$, a contradiction. Hence $T = B$ and $a$ has rank larger than $J$. Conversely, for every $j \in M \setminus B$, the set $B \cup \{j\}$ has $J + 1$ elements, largest entropy $\widehat{h}_j$, and excess bits $\sum_{l=1}^{J} \widehat{h}_{(l)} \le e$. □

*Proof of Corollary 17.* Applying (1) to the conditional law of $X_S$ given $z$ gives $H(X_S \mid z) = \sum_{i \in S} \widehat{h}_i - \mathrm{TC}(X_S \mid z)$ under exact marginals. For $j \in M \setminus B$, the set $S_j = B \cup \{j\}$ has $e(S_j \mid z) = \sum_{i \in B} \widehat{h}_i \le e$, so Theorem 3(ii) gives $\mathrm{TC}(X_{S_j} \mid z) \le e$, while $\mathrm{TC} \ge 0$ always. Hence $H(X_{S_j} \mid z) \ge \sum_{i \in B} \widehat{h}_i + \widehat{h}_j - e$ and $H(X_U \mid z) \le \sum_{i \in B} \widehat{h}_i + \widehat{h}_{(J(e)+1)}$, and subtracting gives the claim; $j = (m)$ gives (7). □

*Remark* 25 (Anchoring can increase dependence with exact marginals). Let $B_1, B_2$ be independent fair bits, $C = (B_1, B_2)$, and $D$ uniform on three values and independent of the rest. At $e = 2$, the anchored set $\{B_1, B_2, C\}$ commits 2 bits with one-step error $\mathrm{TV} = 0.75$, whereas the prefix $\{B_1, B_2, D\}$ commits $2 + \log_2 3$ bits with $\mathrm{TV} = 0$ .

**Verification of Remark 25.** The marginal entropies are 1, 1, 2 and $\log_2 3$ for $B_1$, $B_2$, $C$ and $D$. At $e = 2$, $\widehat{h}_{(1)} + \widehat{h}_{(2)} = 2 \le e < 2 + \log_2 3$, so $J(e) = 2$ and $B = \{B_1, B_2\}$. Since $C$ is a function of $(B_1, B_2)$, the joint law of $\{B_1, B_2, C\}$ assigns probability mass $\frac{1}{4}$ to each of four outcomes, whereas the product of the exact marginals assigns mass $\frac{1}{16}$ to each of 16 outcomes, so $\mathrm{TV} = 1 - 4 \cdot \frac{1}{16} = 0.75$. The prefix consists of independent positions, so the product of its marginals is exact. Finally, $\widehat{h}_{(m)} - \widehat{h}_{(J(e)+1)} = 2 - \log_2 3 < e$.

## C.2 Proof of Proposition 18

*Proof.* Fix a reveal policy that selects one masked position per step as a function of the current state, possibly with randomization independent of the committed values. For a terminal sequence $x$ and a realization of the randomization, the revealed positions form a permutation $(i_1, \ldots, i_n)$ determined by $x$ and that realization, and the probability that the sampler commits $x$ along it is $\prod_{t=1}^{n} Q(X_{i_t} = x_{i_t} \mid X_{i_1} = x_{i_1}, \ldots, X_{i_{t-1}} = x_{i_{t-1}})$, which equals $Q(x)$ by the chain rule in the order $(i_1, \ldots, i_n)$. Averaging over the randomization gives $Q(x)$ for every such policy. Calibration error of a coherent denoiser changes $Q$ but not this invariance. □

### C.3 Information revealed per step under an adaptive reveal policy

**Theorem 26** (Information accounting under an adaptive reveal policy)**.** *Let $Z_0$ be fully masked and $Z_{k+1} = (Z_k, X_{S_k})$, where $S_k \subseteq M(Z_k)$ is $\sigma(Z_k, C)$-measurable. Then, with $\Delta I_k := I(X; Z_{k+1} \mid C) - I(X; Z_k \mid C)$,*

$$\Delta I_k = \mathbb{E}\big[H(X_{S_k} \mid Z_k, C)\big] = \mathbb{E}\Big[\textstyle\sum_{i \in S_k} H(X_i \mid Z_k, C)\Big] - \mathbb{E}\big[\,\mathrm{TC}(X_{S_k} \mid Z_k, C)\big]. \tag{13}$$

*Proof.* Since $S_k$ is $\sigma(Z_k, C)$-measurable, $\sigma(Z_{k+1}) = \sigma(Z_k, X_{S_k})$ conditionally on $C$, so $\Delta I_k = I(X; X_{S_k} \mid Z_k, C) = H(X_{S_k} \mid Z_k, C)$, because $X_{S_k}$ is a deterministic function of $X$ and $Z_k$. The second equality follows from (1). □

Equation (13) shows that the gap between the sum of marginal entropies, which an entropy-sum sampler constrains, and the information that a step reveals is exactly the total correlation of the reveal set, so the round-count lower bound of Fu et al. (2025) cannot be evaluated from marginal entropies alone. Information identities for masked processes with a fixed schedule (Jeon et al., 2025) assume that the reveal indicators are independent of $X$; an adaptive sampler violates this assumption, and (13) applies instead.

### C.4 The anchored excess-bit sampler

**Algorithm 1** Anchored excess-bit sampling (monotone; certified under exact marginals).

**Require:** prompt $c$; denoiser $p_\theta$; excess-bit budgets $\{e_k\}$
$z_0 \leftarrow (\texttt{[MASK]}, \ldots, \texttt{[MASK]})$
**for** $k = 0, 1, \ldots$ until no masked position remains **do**
  for every masked $i$: $p_{i,k}(\cdot) \leftarrow p_\theta(X_i = \cdot \mid z_k, c)$, $\widetilde{q}_{i,k}(v) \leftarrow p_{i,k}(v)/(1 - p_{i,k}(\texttt{[MASK]}))$ for $v \neq \texttt{[MASK]}$, and $\widehat{h}_{i,k} \leftarrow H(\widetilde{q}_{i,k})$
  $a \leftarrow \arg\max_{i \in M(z_k)} \widehat{h}_{i,k}$; $S_k \leftarrow \{a\}$; $b \leftarrow e_k$ ▷ the free position is unconstrained
  **for** $i \in M(z_k) \setminus \{a\}$ in ascending order of $\widehat{h}_{i,k}$ **do**
    **if** $b \geq \widehat{h}_{i,k}$ **then**
      $S_k \leftarrow S_k \cup \{i\}$; $b \leftarrow b - \widehat{h}_{i,k}$
    **end if**
  **end for**
  $z_{k+1,i} \sim \widetilde{q}_{i,k}(\cdot)$ independently for $i \in S_k$
**end for**
**return** $\widehat{x} = z_K$

The commit law $\widetilde{q}_{i,k}$ is the output law restricted to non-mask tokens and renormalized; it equals $p_{i,k}$ when the output vocabulary contains no mask token, and the entropies that define the budget are those of $\widetilde{q}_{i,k}$. By Proposition 16(a), $S_k$ has maximum cardinality among the sets with $e(S_k \mid z_k) \leq e_k$, and by Theorem 3(iv), the step satisfies $\mathrm{TV} \leq \sqrt{(\ln 2/2)\, e_k}$ when $\widetilde{q}_{i,k}$ are the target marginals. Selecting the position $(J(e_k) + 1)$ of the entropy order instead of the most uncertain position recovers the entropy-ordered prefix of Ben-Hamu et al. (2025) with the entropy proxy. Every commit reveals its positions, so the sampler terminates after at most $n$ steps, each costing one forward pass plus $O(n|\mathcal{V}|)$ arithmetic. The categorical experiments of Section 6.4 instead sample the full output vocabulary and use its entropies (Appendix E.5).

## D Trajectory-level accounting and matched-cost comparisons

### D.1 The trajectory decomposition

The following identity is the error decomposition of Ben-Hamu et al. (2025), whose expectation is also taken under the target chain; for exact marginals it appears in Zhang et al. (2026) and, for general adaptive policies, in Wen et al. (2026). We restate it with the total-variation consequence used in the supplementary material.

**Theorem 27** (Trajectory decomposition)**.** *Fix $c$ and a deterministic, state-measurable reveal policy $\sigma$ that reveals at least one position per step, and let Assumption 1 hold. Let $\nu = p^\star(\cdot \mid c)$ be the law*

*of $X$ under exact joint commits $X_{S_k} \sim p^\star(\cdot \mid Z_k, c)$, and let $\mu$ be the law under product commits $X_{S_k} \sim \bigotimes_{i \in S_k} p_\theta(X_i = \cdot \mid Z_k, c)$. Writing $P_{k,i}$ and $\widehat{P}_{k,i}$ for the target and model marginals at $Z_k$,*

$$\mathrm{KL}(\nu \,\|\, \mu) = \sum_{k \geq 0} \mathbb{E}_{Z_k \sim \nu}\Big[\,\mathrm{TC}\left(X_{S_k} \mid Z_k\right) + \sum_{i \in S_k} \mathrm{KL}\left(P_{k,i} \,\|\, \widehat{P}_{k,i}\right)\Big], \tag{14}$$

*and* $\mathrm{TV}(\mu, \nu) \leq \left(\frac{\ln 2}{2} \sum_{k \geq 0} \mathbb{E}_{Z_k \sim \nu}\big[e^\star(S_k \mid Z_k) + \sum_{i \in S_k} \mathrm{KL}(P_{k,i} \| \widehat{P}_{k,i})\big]\right)^{1/2}$.

*Proof.* Because $\sigma$ depends only on the state and the reveal process is monotone, the trajectory $(Z_0, \ldots, Z_n)$ is a deterministic function of the terminal vector and conversely, so the divergence between the path laws of the exact and product-commit chains equals $\mathrm{KL}(\nu \| \mu)$. The path laws factorize into the per-step kernels, and the chain rule for divergence gives $\sum_k \mathbb{E}_{Z_k \sim \nu}[\mathrm{KL}(\mathcal{P}(Z_k, \cdot) \| \mathcal{Q}(Z_k, \cdot))]$, where $\mathcal{P}$ and $\mathcal{Q}$ are the exact and product kernels; Assumption 1 ensures finiteness. Applying Theorem 3(i) to each summand gives (14), and Theorem 3(ii) with Pinsker's inequality gives the total-variation bound. □

*Remark* 28 (Comparison of total-variation conversions). A divergence bracket $\mathcal{B}$ in bits can be converted by Pinsker's inequality or by the Bretagnolle–Huber inequality (Bretagnolle & Huber, 1979; Canonne, 2022):

$$\mathrm{TV} \;\leq\; \min\left\{\sqrt{(\ln 2/2)\,\mathcal{B}},\; \left(1 - 2^{-\mathcal{B}}\right)^{1/2}\right\}. \tag{15}$$

The two agree at $\mathcal{B}_\times = 1.5936/\ln 2 \approx 2.2991$ bits, the unique positive solution of $x/2 = 1 - e^{-x}$ with $x = \mathcal{B} \ln 2$; Pinsker's inequality is sharper below $\mathcal{B}_\times$, the Bretagnolle–Huber inequality is sharper above it, and only the latter remains below 1 for every finite $\mathcal{B}$. A weaker variant, $\mathrm{TV} \leq 1 - \frac{1}{2} \cdot 2^{-\mathcal{B}}$, appears in Tsybakov (2009, Ch. 2). For the calibration-aware brackets of Table 6, with mean $4.319586$ bits, the per-set Pinsker and Bretagnolle–Huber conversions average $1.193560$ and $0.949555$, and their per-set minimum averages $0.942859$, compared to a mean exact total variation of $0.302712$.

### D.2 Matched-cost comparison by randomized mixtures

Reveal rules are indexed by a hyperparameter and attain only a discrete set of expected costs. To compare two rules at a common cost, we interpolate between adjacent operating points; interpolation in $\log \mathrm{TV}$ carries no guarantee, whereas linear interpolation in total variation does.

**Lemma 29** (Attainable interpolation)**.** *Let two reveal policies induce sampler laws $\mu_1, \mu_2$ with expected numbers of denoiser evaluations $K_1, K_2$. For $\lambda \in [0, 1]$, the randomized policy that runs the first policy with probability $\lambda$ and the second otherwise induces $\mu_\lambda = \lambda\mu_1 + (1 - \lambda)\mu_2$, uses $K_\lambda = \lambda K_1 + (1 - \lambda)K_2$ evaluations in expectation, and satisfies $\mathrm{TV}(\mu_\lambda, \nu) \leq \lambda\ \mathrm{TV}(\mu_1, \nu) + (1 - \lambda)\ \mathrm{TV}(\mu_2, \nu)$ for every target law $\nu$.*

*Proof.* The output law of the randomized policy is $\mu_\lambda$ by the law of total probability, the expected cost is linear, and $\mathrm{TV}(\cdot, \nu)$ is convex by the triangle inequality. □

*Remark* 30 (The interpolated distance is a guarantee, not an identity). The inequality of Lemma 29 can be strict by an arbitrary factor: for $\nu = (\frac{1}{2}, \frac{1}{2})$, $\mu_1 = (1, 0)$ and $\mu_2 = (0, 1)$, the interpolated value at $\lambda = \frac{1}{2}$ is $\frac{1}{2}$, whereas $\mu_{1/2} = \nu$. Each interpolated entry of Table 7 is therefore a value that the corresponding rule is guaranteed to attain at that cost, and comparing entries across rules compares guarantees rather than realized distances; a two-sided comparison requires the realized mixture law, as in Table 12.

## E Experimental setup

### E.1 Tractable sources and exact calculations

**Sources.** The *reference mixture* has a latent variable with $K = 4$ uniformly weighted values and a response of $n = 7$ positions over $|\mathcal{V}| = 4$ symbols; each emission is a softmax of logits drawn once from $\mathcal{N}(0, 1.6^2)$ under a fixed seed, giving $p(x \mid c) = \sum_j \pi_j \prod_i \theta_{j,i}(x_i)$ with $\mathrm{TC}(X \mid C) = 2.1510$ bits, and all $16{,}384$ sequences are enumerated. The *Markov source* has $n = 8$ positions and

$|\mathcal{V}| = 3$ symbols, with an initial distribution and seven transition matrices given by softmaxes of $\mathcal{N}(0, 1.4^2)$ logits; results are averaged over five seeds. The *hidden Markov models* of Table 4 have $(n, |\mathcal{V}|) = (24, 6)$ and $(40, 8)$ and the same emission construction. Conditional marginals are computed in closed form from the posterior over the latent variable (mixtures) or by forward–backward recursion (Rabiner, 1989), and reveal-set joints by enumeration over $|\mathcal{V}|^{|S|}$ assignments. Unless stated otherwise, the denoiser returns these exact conditional marginals, and no network is trained.

**Exact sampler law.** Each deterministic reveal rule yields a unique trajectory for each complete sequence, and the probability of the trajectory is the product of the commit probabilities along it. Traversing the reveal tree gives the induced law and the expected number $K$ of denoiser evaluations. Reference calculations discard branches of probability below $10^{-14}$; the operating-point calculations below retain every branch.

**Reveal rules and sweeps.** *Fixed-$k$* reveals the $k$ masked positions of smallest entropy, with $k \in \{1, 2, 3, 4, 5, 7\}$ (mixture) and $k \in \{1, 2, 3, 4, 8\}$ (Markov). *Confidence threshold* (Wu et al., 2026) reveals every position with $\max_v p_\theta(v) > \tau$, and at least the most confident one, with $\tau \in \{0.99, 0.95, 0.9, 0.8, 0.6, 0.4, 0.2\}$. *Entropy sum* adds positions in ascending order of entropy while the running sum does not exceed $\Delta \in \{0.02, 0.05, 0.1, 0.25, 0.5, 1, 2, 4\}$ bits. *Entropy-bounded* (Ben-Hamu et al., 2025) orders the masked positions by the entropy $\widehat{h}_i$ or by the confidence error $1 - \max_v p_\theta(v)$ and grows the prefix $U$ while $\sum_{i \in U} \widehat{h}_i - \max_{i \in U} \widehat{h}_i \leq \gamma$, with $\gamma \in \{0, 0.02, 0.05, 0.1, 0.25, 0.5, 1, 2\}$ bits. *Anchored* is Algorithm 1 on the same budget grid. *Depth*-2 admits the sets with $e_2(S) \leq e$, evaluated by enumerating subsets of the masked positions, and expands the anchor over its $k \in \{1, 2, |\mathcal{V}|\}$ most probable values.

**Operating-point instance.** The full sweep in the supplementary material uses the true mixture emissions of the depth-2 calibration experiment below ($n = 7$, $|\mathcal{V}| = 4$, four uniformly weighted components, logit standard deviation 1.6, source seed 0), which gives 45 settings. For the confidence threshold, equality is excluded and an empty selection uses the most confident position; the entropy-sum rule uses the lowest-entropy singleton if no position satisfies its budget constraint. Ties are broken by ascending position index, and by the largest index for tied anchors of maximal entropy. All calculations use float64 arithmetic.

**Certificate evaluation.** For Table 4, reveal sets of size two to five are drawn from reachable states by sampling a full sequence, revealing a uniformly random subset of positions, and selecting a reveal set among the masked positions; 12,000 reveal sets are evaluated. For each set, we compute the exact joint law, its total correlation, the excess bits and the meet, profile and major-profile bounds of Proposition 6, the latter two by the procedures of Compton et al. (2023); the minimum-entropy coupling at $|S| = 2$ is computed exactly by the backtracking algorithm of Compton et al. (2023). The depth-2 bound uses the anchor of largest entropy and, as the second anchor, the remaining position of largest marginal entropy; this is not the greedy rule of Corollary 20. To test Proposition 16(b), we draw 3,000 reachable states of the reference mixture and compare, at each budget $e \in \{0.05, 0.25, 1, 2\}$, the cardinality of the anchored set with that of the prefix, which gives 12,000 comparisons.

**Learned denoisers.** Three mixture-of-products denoisers are fitted by expectation maximization (Dempster et al., 1977) on samples from the true source, with a misspecified number of components or sample size: $K' = 2$ and $K' = 3$ with $2 \times 10^4$ samples, and $K' = 4$ with $2 \times 10^5$ samples. At 3,000 reachable states, we evaluate one singleton reveal set and one reveal set of size at least two, and record violations of the exact-marginal bound $\mathrm{TV} \leq \sqrt{(\ln 2/2)\, e(S \mid z)}$ and of Theorem 3(v).

**Depth-2 calibration experiment.** For Table 6, true emission logits are drawn from $\mathcal{N}(0, 1.6^2)$ with $n = 7$, $|\mathcal{V}| = 4$ and four uniformly weighted components, and independent $\mathcal{N}(0, 0.7^2)$ noise is added to each logit to define the model. For each of 4,000 samples, we draw a sequence from the true source, choose the observed positions uniformly from the 120 subsets of sizes zero to five, draw the set size $k$ uniformly from $\{2, \ldots, m\}$ for the $m$ remaining masked positions, and choose $S$ uniformly among the subsets of size $k$. The true source and the model compute their own posteriors, the anchors $a, b \in S$ are the two positions of largest model marginal entropy, and

$P = P_S(\cdot \mid z)$ and $Q = \bigotimes_{i \in S} \widehat{P}_i(\cdot \mid z)$ are computed exactly. We compare the uncorrected expression $\sum_{i \in S} \widehat{h}_i - \widehat{h}_a - \widehat{H}_\theta(X_b \mid X_a)$, the cross-entropy form (12) with the exact offline value of $H(X_b \mid X_a)$, and the calibration-aware form (11) with the exact total variations as budgets, evaluated without clipping, that is, with $L = \widehat{H}_\theta(X_b \mid X_a) - \Delta$ in place of $L_+$. The correction $\Delta$ is the term of Lemma 23 or, for comparison, the term suggested by Fannes' inequality (Fannes, 1973), $\Delta_{\mathrm{F}} = g_{16}(\varepsilon_{ab}) + g_4(\varepsilon_a)$ with $g_d(\varepsilon) = 2t \log_2 \frac{d}{2t}$, $t = \min\{\varepsilon, \frac{1}{2}\}$ and $g_d(0) = 0$. An underestimate means an expression below $\mathrm{KL}(P\|Q)$ or $\mathrm{TC}(P)$ by more than $10^{-9}$ bits.

**Anchored comparison under corrupted marginals.** The denoiser corrupts each marginal toward the uniform law in proportion to a power of its entropy, $\widehat{P}_i = (1 - \varepsilon_i)P_i + \varepsilon_i \,\mathrm{Unif}$ with $\varepsilon_i = 0.25\,(\widehat{h}_i / \log_2 |\mathcal{V}|)^\alpha$. At each reachable state and a fixed budget, we compare the exact total variation between the product of the model marginals and the true joint of the anchored set and of the prefix, for $\alpha \in \{0, 0.5, 1, 2\}$. The anchored set has the smaller distance in $98.3\%$ of states at $\alpha = 0$, where the corruption is uniform across positions, but in $68.6\%$ at $\alpha = 2$, where it concentrates on the uncertain positions that anchoring commits.

### E.2 Controlled experiments

The controlled experiments use a separately generated mixture of products with $n = 7$, $|\mathcal{V}| = 4$ and four uniformly weighted components, whose emission logits are drawn once from $\mathcal{N}(0, 1.6^2)$ with NumPy seed 2027; its total correlation is 2.1675 bits. The *exact* denoiser returns the conditional marginals of the source. The *coherent but miscalibrated* denoiser adds independent $\mathcal{N}(0, 0.7^2)$ noise to every emission logit and computes its own posterior, so all of its outputs are conditional marginals of one perturbed mixture. Three incoherent denoisers are derived from it: *pass-dependent noise* perturbs the log-probabilities returned by every forward pass with independent $\mathcal{N}(0, \sigma^2)$ noise, $\sigma \in \{0.1, 0.3\}$; *mismatched clamped passes* computes the clamped conditionals from a second, independently perturbed mixture; and the *position-split* denoiser returns the conditionals of the first perturbed mixture at even positions and of the second at odd positions.

A state is drawn by sampling a sequence, choosing the number of revealed positions and the revealed positions uniformly, and keeping enough masked positions for the set or pair to be drawn uniformly among them. The anchor experiment uses 1,000 reveal sets of each size 3, 4 and 5 (seed 11), and the power experiment uses 2,000 pairs per denoiser (seed 22). From the stored conditionals of one anchor, computing the envelope took 37 to 65 microseconds per set and the meet relaxation 99 to 142 on one processor core in float64 at $|\mathcal{V}| = 4$; these are aggregation costs on stored outputs, and only the meet grows with $|\mathcal{V}| \log |\mathcal{V}|$. Truncation retains the $k$ most probable values under the baseline pass, with ties broken by value index. For the sampler laws, the prefix and anchored rules are evaluated at budgets $\{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.5, 3\}$ bits, the randomized policy of Lemma 29 is evaluated exactly at matched costs, and at $e = 0$ the third order reveals positions from left to right.

**Residual certificates.** For Table 9, the supplementary script draws 2,000 reveal sets of size two and 1,000 of size three with generator seed 11 by the same state construction, and 1,000 sets each of sizes four and five for the marginal-level bracket. The anchor is the position of largest marginal entropy, and $R$ contains its $k$ most probable values. Minimum-entropy couplings of two (possibly partial) marginals are computed exactly by the backtracking algorithm of Compton et al. (2023, Alg. 2 and Lemma 5.3) with the meet as pruning bound, checked against vertex enumeration on random instances; the major-profile follows Compton et al. (2023, App. G.3) and the greedy coupling Kocaoglu et al. (2017). For two positions, $\beta_R^\star$ is computed exactly and the computable bracket replaces the residual coupling. For three positions, the exact-retained bracket $[\widetilde{L}_R, \widetilde{U}_R]$ uses exact conditional couplings of two marginals, and the general bracket $[L_R, U_R]$ replaces every coupling. For the vocabulary study (Table 10), the same construction with $n = 7$ and four components uses $|\mathcal{V}| \in \{4, 16, 64, 256\}$, logit standard deviation $1.6\,(\log_2 |\mathcal{V}|/2)^{1/2}$, which recovers the controlled instance at $|\mathcal{V}| = 4$, 500 reveal sets of size two per vocabulary (generator seed 13), and the smallest $k$ whose retained anchor mass reaches 0.5, 0.8 or 0.95. The vocabulary size and the logit scale change together, and the realized retained masses exceed the targets by different amounts, so the study describes trends in this source family rather than an isolated effect of the vocabulary size. A set is *certified* at budget $\tau$ if $U_R \le \tau$, *not certifiable* with the given information if $L_R > \tau$, since then

$\beta_R^\star > \tau$, and *undecided* otherwise. Because the true total correlation is known on these sources, we also count the sets that meet the budget but are not certifiable, $\mathrm{TC} \le \tau < L_R$; this is a lower bound on the sets that the given information fails to certify although they meet the budget. All inequalities of Propositions 6 and 10 and Corollary 11 were checked on every set with tolerance $10^{-9}$.

### E.3 Conditional queries to LLaDA

**Model and saved states.** We use LLaDA-8B-Instruct (Nie et al., 2025) in evaluation mode with bf16 weights and batch size one. The 35 GSM8K training-split prompts are selected with seed 260904, excluding the four in-context demonstrations. Saved argmax trajectories use seed 0, 256 decoding steps, 256 generated tokens, block length 8, temperature zero and no classifier-free guidance; the offline queries reuse these states at softmax temperature one, and test prompts are not used. With $q$ masked candidates in the active block, we use states with $q \ge m+1$ for $m \in \{2,3,4\}$, which gives 6,720, 5,600 and 4,480 state–$m$ cases.

**Feasible sets.** Membership and budgets are fixed by the saved entropies $\widehat{h}_i^{\mathrm{src}}$, sorted by increasing entropy and then by position:

$$e_m = \sum_{j=1}^{m-1} \widehat{h}_{(j)}^{\mathrm{src}}, \qquad \mathcal{F}_m(z) = \Big\{ S \subseteq M_{\mathrm{active}}(z) : |S| = m,\ \sum_{i\in S} \widehat{h}_i^{\mathrm{src}} - \max_{i\in S} \widehat{h}_i^{\mathrm{src}} \le e_m \Big\}.$$

The prefix contains the $m$ lowest-entropy positions, the anchored set contains the $m-1$ lowest-entropy positions and the highest-entropy position, and *uniform* is the equal-weight average over the distinct sets in $\mathcal{F}_m(z)$. There are 30,240, 22,400 and 15,680 feasible sets at $m = 2, 3, 4$.

**Clamped queries.** A fresh forward pass supplies the full-vocabulary probabilities $\widehat{P}_i(\cdot \mid z)$ and entropies $\widehat{h}_i = H(\widehat{P}_i(\cdot \mid z))$ over all $|\mathcal{V}| = 126{,}464$ tokens, including special tokens. Each set uses its two positions of largest fresh entropy, $a$ and then $b$, fixed across $k$. At each clamped position, tokens are ordered by decreasing logit, and $V_k(u)$ is the prefix of length $k \in \{1,2,4,8\}$ of this list; retained probabilities are not renormalized. For each retained token $v$, we start from the original state, set only $u$ to $v$, and read $\widehat{h}_{t|u=v}$ at the other anchor. Identical queries are shared across sets, rules and directions. For exact marginals and conditionals of one joint law, $\widehat{h}_a + C_{\mathrm{full}}(a,b) = H(X_a, X_b \mid z) = \widehat{h}_b + C_{\mathrm{full}}(b,a)$, so both orders give the same depth-2 certificate; the direction gap alone does not separate truncation from incoherence, whereas the truncation intervals of Proposition 14 do. Logits and entropies are computed in float32, and sums and aggregates in float64.

**Aggregation and coherence statistics.** At fixed $(m, k, \text{direction}, \text{rule})$, we average over sets within a state, over states within a prompt, and equally over the 35 prompts; standard errors use the sample standard deviation of the prompt means divided by $\sqrt{35}$, and the mean of ratios omits sets with $d_1 \le 10^{-6}$ bits. After query sharing, 6,720 baseline and 295,680 clamped passes give 302,400 evaluations, plus 40 warm-up passes. For Tables 3 and 13, the averaged endpoints $\bar{d}_{2,k}^{u\to t} - (1 - \bar{w}_k(u)) \log_2 |\mathcal{V}|$ and $\bar{d}_{2,k}^{u\to t}$ are the averages of the per-set endpoints under the common hierarchical weights, because both endpoints are affine in the per-set quantities, and $\mathrm{IncLB}_k$ is computed from these averaged intervals. A negative mean of $d_2/d_1$ implies that at least one set has $d_2 < 0$.

### E.4 Composition at common states on LLaDA

Both commit laws use the saved states, budgets $e_m$ and feasible families $\mathcal{F}_m(z)$ of Appendix E.3. We write $A$ for the anchored set, $P$ for the prefix, and $\mathcal{I}_m(z) = \mathcal{F}_m(z) \setminus \{A, P\}$ for the interior family. Let $C_m(z)$ be the masked positions left uncommitted in every feasible set and $B_{\mathrm{next}}(z)$ the next generation block. The primary region is $C_m(z) \cap B_{\mathrm{next}}(z)$ and the secondary region is $C_m(z)$; for a region $\mathcal{C}$ and a commit $x_S$,

$$D(S, x_S; \mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{j\in\mathcal{C}} \Big[ \widehat{h}_j^{\mathrm{before}} - \widehat{h}_j^{\mathrm{after}}(z[S \leftarrow x_S]) \Big],$$

with full-vocabulary entropies at temperature one. States without common positions are omitted, which leaves 6,510, 5,425 and 4,340 state–$m$ cases. Argmax commits use the pre-commit logits and

one new forward pass. Categorical commits use four repetitions of independent full-vocabulary draws at temperature one by Gumbel-max sampling in float64 (master seed 260906), without top-$k$ or top-$p$ truncation; no draw of the mask token was observed. $D_A$, $D_P$, $D_U$ and $D_{\mathcal{I}}$ are the anchored, prefix, uniform and interior measurements; state measurements are averaged within each prompt, and the 35 prompt means receive equal weight. Paired differences are formed on common eligible states before aggregation, and standard deviations use population normalization over the prompt means. These one-step runs neither generate full answers nor estimate task accuracy.

### E.5 Complete generations on LLaDA

**Model, data and rules.** All evaluations use LLaDA-8B-Instruct without additional training, bf16 weights, batch size one, generation length 256 and no classifier-free guidance. GSM8K uses all 1,319 test examples, the fixed four-shot prompt of the OpenCompass evaluation suite (OpenCompass Contributors, 2023), and block length 8. MBPP uses argmax decoding on all 500 tasks of the full-version test split, block length 256, and the official three-shot prompt with three public assertions per task. Fixed schedules use the official low-confidence position order and commit quota for their step count; the 256-step schedule is the official one. For the prefix and anchored rules, the budget $e$ is constant throughout each generation, the eligible positions are the masked positions of the active block, and only marginal entropies are used. GSM8K uses fixed schedules with 256, 128 and 64 steps and both entropy rules at $e \in \{1, 2\}$; MBPP uses fixed schedules with 256, 128, 64 and 32 steps and both entropy rules at $e \in \{0.1, 0.25, 0.5, 1, 2, 4\}$. All configurations are reported.

**Token commits.** Argmax commits use temperature zero and generation seed 0. The GSM8K categorical setting draws each committed token independently from the full output vocabulary at temperature one by Gumbel-max sampling in float64 (master seed 260911), with one completion per prompt and configuration; no draw of the mask token was observed. This chain differs from that of Algorithm 1, which renormalizes the output law over non-mask tokens at each step: conditioning on the event that no mask token is drawn anywhere in an adaptive trajectory is not the same as renormalizing at each step, and with $r_i$ the mass on the mask token, the entropies satisfy $H(q_i) = h_2(r_i) + (1 - r_i)H(\widetilde{q}_i)$. The reported entropies and budgets are those of the full output vocabulary.

**Scores, intervals and cost.** GSM8K accuracy is the fraction of completions scored correct by the recorded answer extraction; MBPP pass@1 is the fraction of completions that pass the three public tests after fixed code postprocessing, with a 2-GiB memory limit and inner and outer timeouts of 9 and 10 seconds; timeouts and resource-limit failures, 38 of 8,000 completions, count as incorrect. Per-configuration intervals are Wilson 95% intervals (Wilson, 1927), paired accuracy differences use method 10 of Newcombe (1998), and NFE and time differences use 10,000 paired bootstrap resamples of prompts (Efron, 1979). Intervals are descriptive and unadjusted for multiple comparisons, and the categorical study does not measure variation over sampling seeds. Generation NFE counts denoiser passes for completed answers. MBPP runs alternate computation windows of 0.5 seconds with rests of 0.5 seconds at step boundaries, so MBPP times include these rests and do not define a hardware speedup. Pairwise comparisons, time and memory statistics, and operational costs are in the supplementary material.

**Protocol differences for MBPP.** Ben-Hamu et al. (2025) evaluate entropy-bounded sampling on MBPP with LLaDA-8B-Base, four-shot prompts, a 512-token window with termination at a stop phrase, and, additionally, semi-autoregressive blocks of 64 tokens. Our protocol uses LLaDA-8B-Instruct, a single block of 256 tokens, and no stop-phrase termination. For LLaDA-8B-Instruct, Nie et al. (2025) report that low-confidence remasking terminates early because of end-of-sequence padding in the supervised fine-tuning data and that block-wise semi-autoregressive sampling performs better. These observations motivate our hypothesis that entropy ordering within a single 256-token block underlies the degradation: the lowest-entropy positions available to the rule may include end-of-sequence positions. Program lengths were not measured, and the current experiments do not isolate this mechanism.

### E.6 Further evaluations

The coherence tests of Section 4.2 can be evaluated per unique state–pair rather than on aggregates. Eight saved states per development prompt give 280 state–pairs, and the baseline pass and $k = 8$ clamped passes in both directions cost $280(1+2\cdot 8) = 4{,}760$ forward evaluations before query sharing. Reporting $\mathrm{IncLB}_k$, the statistic $D_k$ of Proposition 15, the residual test of Proposition 13(ii) and the retained masses, with a numerical guard band and a cluster interval over prompts, would convert the existence result of Section 6.2 into empirical detection rates with effect sizes, and recomputing a subset of states at higher precision would bound the contribution of bf16 inference. The same passes give the residual laws of Proposition 10(ii): a negative entry certifies that no joint law matches the outputs and is reported as incompatibility, and otherwise the general bracket of Corollary 11 applies to the outputs, whose reading as a bound for the target still presupposes exactness. For three or more positions, the exact-retained bracket is computationally impractical at this vocabulary size with our exact solver. The baseline passes give $e_{\mathrm{G}}$, $e_{\mathrm{maj}}$, $e_{\mathrm{prof}}$ and $e_{\mathrm{MP}}$ of Proposition 6. The fraction of feasible sets that satisfy the condition of Proposition 5 requires only the saved entropies; its simple form $\max_i \widehat{h}_i \le 16.9 - m$ allows largest entropies up to 14.9, 13.9 and 12.9 bits for $m = 2, 3, 4$, against a maximum of $\log_2 126{,}464 \approx 16.95$ bits. By Proposition 18, an effect of the reveal order on the output law of single-position decoding with categorical commits certifies incoherence; because this statement concerns output laws, such a comparison requires repeated draws and a distributional or task-level test. Complete generations can further be evaluated with the renormalized commit law of Algorithm 1 and several sampling seeds, with MBPP protocols that vary the block size and the reveal order separately, against confidence-threshold decoding (Wu et al., 2026; Kasa et al., 2026) and conditional-independence testing (Azangulov et al., 2026), on further tasks (Hendrycks et al., 2021; Chen et al., 2021; Suzgun et al., 2022; Zhou et al., 2023; Kang et al., 2026), and with other masked diffusion models (Ye et al., 2025).

## F Additional experimental results

This appendix reports the results that support the claims of Section 6. Auxiliary tables for the tractable sources (the advantage of the meet as a function of the vocabulary size, trajectory accounting, and the complete sweep of operating points) and the complete numerical tables for the LLaDA experiments are in the supplementary material.

### F.1 Exact verification on tractable sources

#### F.1.1 Certificate comparisons

On softmax marginals of $\mathcal{N}(0, 1.6^2)$ logits, the advantage of the meet over the excess bits decreases with the vocabulary size, from $1.84\%$ of the excess bits at $|\mathcal{V}| = 4$ and $|S| = 2$ to $0.07\%$ at $|\mathcal{V}| = 1000$ and $|S| = 16$ (supplementary material). On the hidden Markov models, the depth-2 bound removes $68.8\%$ and $69.6\%$ of the excess bits, against $10.0\%$ and $7.2\%$ for $e_{\mathrm{MP}}$.

#### F.1.2 Depth-2 query costs

A clamped pass at $z[a \leftarrow v]$ can serve as the forward pass of the next step only if the committed set is $\{a\}$ and the committed value is among the retained values: for a larger reveal set, the next state conditions on all committed values. For example, if $X_a$ and $X_b$ are independent fair bits and $X_c = X_b \oplus N$ with an independent $N \sim \mathrm{Bernoulli}(0.1)$, then $\mathbb{P}(X_c = 1 \mid X_a) = \frac{1}{2}$, whereas $\mathbb{P}(X_c = 1 \mid X_a, X_b = 1) = 0.9$. We therefore assess a cost of $1 + k$ forward passes per depth-2 step.

#### F.1.3 Learned and noisy denoisers

For the three denoisers fitted by expectation maximization, with model errors $\mathrm{TV}(p_\theta, p_{\mathrm{data}})$ of $0.3394$, $0.2039$ and $0.0074$, the exact-marginal bound $\mathrm{TV} \le \sqrt{(\ln 2/2)\, e(S \mid z)}$ is violated at all singleton reveal sets and at $0.0\%$ to $0.1\%$ of larger sets, whereas Theorem 3(v) is never violated, with mean slack $1.517$ to $1.537$ bits. Table 6 reports the depth-2 expressions under a noisy-logit denoiser.

Table 4: The certificate hierarchy of Proposition 6 and the depth-2 bound of Theorem 9 on reveal sets drawn from reachable states, with an exact denoiser; all quantities in bits. Here $e$ is the excess-bit bound (Ben-Hamu et al., 2025), and $e_{\mathrm{maj}}$, $e_{\mathrm{prof}}$ and $e_{\mathrm{MP}}$ are the meet, profile and major-profile bounds. The exact minimum-entropy-coupling optimum at $|S| = 2$ is $e_{\mathrm{opt}} = 1.051$ on the mixture and 0.838 on the Markov source, which removes 22% and 25% of the excess bits; at $|S| = 2$ on the mixture, the meet, profile and major-profile bounds remove 0.95%, 10.46% and 18.11%. The chain $\mathrm{TC} \le e_{\mathrm{MP}} \le \min(e_{\mathrm{prof}}, e_{\mathrm{maj}}) \le e$ never failed, and $\mathrm{TC} \le e_{\mathrm{opt}} \le e_{\mathrm{MP}}$ held wherever $e_{\mathrm{opt}}$ was computed. In these sets $e_2$ was below every marginal-only bound, which Remark 22 shows is not guaranteed in general. The hidden Markov rows use reveal sets of size two and three, and the profile bound was not computed for the Markov and hidden Markov sources.

| source | $\lvert S\rvert$ | $\mathrm{TC}(P)$ | $e(S)$ | $e_{\mathrm{maj}}(S)$ | $e_{\mathrm{prof}}(S)$ | $e_{\mathrm{MP}}(S)$ | $e_2(S)$ |
|---|---|---|---|---|---|---|---|
| mixture | 2 | 0.0844 | 1.3405 | 1.3278 | 1.2003 | 1.0977 | **0.0844** |
| | 3 | 0.2421 | 2.8653 | 2.8464 | 2.6436 | 2.5089 | 1.3481 |
| | 4 | 0.4998 | 4.5125 | 4.4915 | 4.2333 | 4.0866 | 2.8817 |
| | 5 | 0.8840 | 6.3719 | 6.3519 | 6.0587 | 5.9072 | 4.6651 |
| | all | 0.4282 | 3.7771 | 3.7590 | 3.5384 | 3.4045 | 2.2490 |
| | ratio to TC | 1.00 | 8.82 | 8.78 | 8.26 | 7.95 | 5.25 |
| Markov | 2 | 0.0663 | 1.1139 | 1.1017 | – | 0.8749 | **0.0663** |
| | 3 | 0.2111 | 2.3987 | 2.3843 | – | 2.0661 | 1.2042 |
| | 4 | 0.4213 | 3.7869 | 3.7750 | – | 3.4071 | 2.5565 |
| | 5 | 0.6997 | 5.2375 | 5.2259 | – | 4.8213 | 3.9804 |
| | all | 0.3505 | 3.1335 | 3.1210 | – | 2.7923 | 1.9524 |
| | ratio to TC | 1.00 | 8.94 | 8.90 | – | 7.97 | 5.57 |
| HMM, $n$=24, $\lvert\mathcal{V}\rvert$=6 | 2, 3 | 0.086 | 2.501 | – | – | 2.252 | 0.779 |
| HMM, $n$=40, $\lvert\mathcal{V}\rvert$=8 | 2, 3 | 0.064 | 3.044 | – | – | 2.826 | 0.926 |

Table 5: The cost of the depth-2 certificate on the mixture source. A depth-2 step expands the anchor over its $k$ most probable values and therefore costs $1 + k$ forward passes; the reduction in the number of steps does not offset the increase in forward passes.

| rule | steps at $\mathrm{TV} = 0.05$ | forward passes per step | forward passes at $\mathrm{TV} = 0.05$ |
|---|---|---|---|
| entropy-bounded, entropy proxy (Ben-Hamu et al., 2025) | 6.174 | 1 | **6.17** |
| anchored (Algorithm 1) | 5.767 | 1 | **5.77** |
| depth-2, top-1 anchor expansion | 5.485 | 2 | 10.97 |
| depth-2, top-2 anchor expansion | 5.502 | 3 | 16.51 |
| depth-2, exact anchor expansion | 5.436 | 5 | 27.18 |

Table 6: Depth-2 expressions under a noisy-logit denoiser on a separately generated mixture source. All rows use the same 4,000 state–set pairs with $|S| \ge 2$ and the same ordered anchors; $\mathcal{B}$ is the full expression, the divergence slack is $\mathcal{B} - \mathrm{KL}(P\|Q)$, and the mean true divergence and total correlation are 0.541539 and 0.160119 bits. No instance of $L > H(X_b \mid X_a) + 10^{-9}$ was observed. The calibration-aware rows are evaluated without clipping (Appendix E.1). Theorem 3(v) equals the cross-entropy row plus $H(X_b \mid X_a) \le \log_2 4$, so its mean is at most 3.705 bits; the unclipped calibration-aware brackets exceed this value on average, whereas the clipped form of Proposition 24 never exceeds Theorem 3(v).

| expression | below KL (%) | below TC (%) | mean slack (bits) | mean $\mathcal{B}$ (bits) |
|---|---|---|---|---|
| uncorrected depth-2 | 48.175 | 19.700 | 0.728763 | 1.270302 |
| cross-entropy form, exact $H(X_b \mid X_a)$ | 0.000 | 0.000 | 1.163197 | 1.704736 |
| calibration-aware, Fannes, unclipped | 0.000 | 0.000 | 4.709833 | 5.251372 |
| calibration-aware, Lemma 23, unclipped | 0.000 | 0.000 | 3.778047 | 4.319586 |

Table 7: The attainable trade-off of each reveal rule with an exact denoiser and one forward pass per step. Each entry interpolates linearly in total variation between the two adjacent exact operating points of that rule at a common expected cost $K$; by Lemma 29 it is a value that the rule is guaranteed to attain at that cost, and by Remark 30 it is an upper bound on the realized distance. The complete sweep of operating points is in the supplementary material.

| reference mixture ($\mathrm{TC}(X \mid C) = 2.1510$ bits) rule | $K = 5.0$ | $K = 6.0$ | $K = 6.5$ | $K$ at TV $=0.05$ | $K$ at TV $=0.02$ |
|---|---|---|---|---|---|
| fixed-$k$ | 0.1536 | 0.0768 | 0.0384 | 6.349 | 6.740 |
| confidence threshold (Wu et al., 2026) | 0.1492 | 0.0549 | 0.0268 | – | – |
| entropy sum ($\sum_{i \in S} \widehat{h}_i \le \Delta$) | 0.1425 | 0.0663 | 0.0282 | – | – |
| entropy-bounded (Ben-Hamu et al., 2025), entropy proxy | 0.1414 | 0.0636 | 0.0246 | 6.174 | 6.559 |
| entropy-bounded (Ben-Hamu et al., 2025), confidence proxy | 0.1671 | 0.0726 | 0.0253 | – | – |
| anchored (Algorithm 1) | **0.1308** | **0.0388** | **0.0169** | **5.767** | **6.428** |

| Markov source ($\mathrm{TC}(X \mid C) = 2.0412$ bits, five seeds) rule | $K = 4.0$ | $K = 5.0$ | $K = 6.0$ | $K$ at TV $=0.10$ | $K$ at TV $=0.05$ |
|---|---|---|---|---|---|
| fixed-$k$ | 0.2619 | 0.1965 | 0.1310 | 6.473 | 7.236 |
| confidence threshold (Wu et al., 2026) | 0.3345 | 0.2428 | 0.1511 | 6.591 | 7.287 |
| entropy sum | 0.2948 | 0.2137 | 0.1326 | 6.402 | 7.018 |
| entropy-bounded, entropy proxy (Ben-Hamu et al., 2025) | 0.2821 | 0.1857 | 0.0893 | 5.889 | 6.408 |
| anchored (Algorithm 1) | **0.1614** | **0.0958** | **0.0474** | **4.936** | **5.864** |

### F.1.4 Reveal rules at matched cost

### F.1.5 Controlled experiments

Table 8 compares anchor rules and the relaxations of Proposition 10(i), Tables 9 and 10 report the residual certificates of Proposition 10(ii) and the brackets of Corollary 11, and Tables 11 and 12 report the coherence tests and the exact sampler laws.

Table 8: Anchor selection and query-level certificates on the controlled instance with an exact denoiser: 1,000 reveal sets per size, means in bits (top) or fractions of sets and mean gains (bottom). The anchor $a$ is the position of largest marginal entropy; $e_2$ is evaluated at the marginal-entropy pair and at the best second anchor for $a$, $e_a^{\mathrm{env}}$ and $e_a^{\mathrm{maj}}$ are the relaxations of Proposition 10(i), and the column $e_a^{\mathrm{opt}} \ge$ applies its constant to the measured $e_a^{\mathrm{maj}}$. The last two columns of the top panel retain the $k = 2$ most probable anchor values without the residual constraint (Appendix B.4). The gain $\min_b e_2(S; a, b) - e_a^{\mathrm{maj}}$ is split into the part obtained from the largest conditional entropy at each anchor value (the envelope) and the remainder. Over all 12,000 pairs of a set and an anchor, $e_a^{\mathrm{maj}} \ge \mathrm{TC}$ and $e_a^{\mathrm{maj}} \le e_2(S; a, b)$ for every $b$ held without exception.

| | | | | $e_2$, complete queries | | Proposition 10(i) | | | $k = 2$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\lvert S \rvert$ | TC | $e$ | $e_{\mathrm{maj}}$ | marginal | best for $a$ | $e_a^{\mathrm{env}}$ | $e_a^{\mathrm{maj}}$ | $e_a^{\mathrm{opt}} \ge$ | $e_2$, greedy | $b_{a,2}^{\mathrm{maj}}$ |
| 3 | 0.268 | 2.966 | 2.954 | 1.435 | 1.428 | 1.399 | 1.384 | 0.384 | 1.937 | 1.914 |
| 4 | 0.565 | 4.693 | 4.681 | 3.053 | 3.037 | 2.990 | 2.969 | 1.527 | 3.629 | 3.596 |
| 5 | 0.942 | 6.525 | 6.515 | 4.817 | 4.792 | 4.732 | 4.709 | 3.266 | 5.444 | 5.404 |

| | anchor selection | | | | gain over the best second anchor | | |
|---|---|---|---|---|---|---|---|
| $\lvert S \rvert$ | greedy attains best pair | marginal attains best pair | top-1 greedy picks exact greedy anchor | top-2 greedy picks exact greedy anchor | mean | share from the envelope | 90th percentile |
| 3 | 98.5% | 88.8% | 84.8% | 91.3% | 0.0438 | 65.2% | 0.1182 |
| 4 | 98.4% | 75.9% | 70.4% | 83.2% | 0.0674 | 69.0% | 0.1421 |
| 5 | 97.5% | 62.9% | 52.8% | 74.6% | 0.0830 | 72.3% | 0.1534 |

Table 9: Residual certificates (Proposition 10(ii)) and computable brackets (Proposition 6, Corollary 11) on the controlled instance with an exact denoiser, in bits. The anchor is the position of largest marginal entropy, and its $k$ most probable values are queried with $k$ clamped passes; $k = 0$ is the marginal-only level and $k = 4$ queries every value ($e_a^{\mathrm{opt}}$). *Discarding* drops the unqueried values (Appendix B.4). For two positions $\beta_R^\star$ is exact and the computable bracket $[L_R, U_R]$ replaces the residual coupling. For three positions the exact-retained bracket $[\widetilde{L}_R, \widetilde{U}_R]$ uses exact conditional couplings and the general bracket $[L_R, U_R]$ replaces every coupling; entries give the upper end with the lower end in brackets. Budget rows give the percentage of sets whose certificate does not exceed $\tau$; a set is not certifiable with the given information if the lower end exceeds $\tau$ (Appendix E.2), and undecided otherwise; *meets budget, not certifiable* counts sets with $\mathrm{TC} \le \tau$ whose lower end exceeds $\tau$ (93.2% and 100% of the sets meet the budget for two and three positions). For three positions, even $k = 4$ leaves 14.2% of the sets uncertifiable, since $e_a^{\mathrm{opt}}$ exceeds TC by 0.86 bits on average; the missing information concerns other anchors. Bottom: the marginal-level bracket $e_{\mathrm{MP}} - e_{\mathrm{G}}$ against the constants of Proposition 6.

| | | $k=0$ | $k=1$ | $k=2$ | $k=3$ | $k=4$ |
|---|---|---|---|---|---|---|
| $\lvert S\rvert = 2$ | retained anchor mass | 0 | 0.444 | 0.724 | 0.901 | 1 |
| (2,000 sets, | discarding | – | 0.838 | 0.474 | 0.227 | 0.086 |
| $\mathrm{TC} = 0.086$) | residual $\beta_R^\star$ | 1.090 | 0.586 | 0.260 | 0.086 | 0.086 |
| | computable $U_R$ [$L_R$] | 1.138 [1.080] | 0.606 [0.581] | 0.271 [0.256] | 0.086 [0.086] | 0.086 |
| | $\le \tau = 0.25$: discarding | – | 4.7% | 24.8% | 58.8% | 93.2% |
| | residual / computable | 1.5% / 1.4% | 14.6% / 14.4% | 50.0% / 48.5% | 93.2% / 93.2% | 93.2% |
| | not certifiable / undecided | 98.5% / 0.1% | 85.3% / 0.4% | 49.5% / 2.0% | 6.9% / 0.0% | – |
| | meets budget, not certifiable | 91.7% | 78.4% | 42.7% | 0.0% | – |
| $\lvert S\rvert = 3$ | retained anchor mass | 0 | 0.391 | 0.678 | 0.877 | 1 |
| (1,000 sets, | discarding | – | 2.262 | 1.739 | 1.363 | 1.129 |
| $\mathrm{TC} = 0.273$) | exact-retained $\widetilde{U}_R$ [$\widetilde{L}_R$] | 2.615 [2.507] | 1.901 [1.838] | 1.406 [1.380] | 1.133 [1.127] | 1.129 |
| | general $U_R$ [$L_R$] | 2.615 [2.507] | 1.914 [1.835] | 1.430 [1.375] | 1.166 [1.121] | – |
| | $\le \tau = 1.5$: discarding | – | 11.9% | 32.6% | 56.0% | 85.8% |
| | exact-retained / general | 4.5% / 4.5% | 26.9% / 26.4% | 50.9% / 49.0% | 84.7% / 79.4% | – |
| | not certifiable / undecided (general) | 94.7% / 0.8% | 72.2% / 1.4% | 46.7% / 4.3% | 12.6% / 8.0% | – |
| | meets budget, not certifiable (general) | 94.7% | 72.2% | 46.7% | 12.6% | – |

| | $\lvert S\rvert = 2$ | $\lvert S\rvert = 3$ | $\lvert S\rvert = 4$ | $\lvert S\rvert = 5$ |
|---|---|---|---|---|
| mean (maximum) of $e_{\mathrm{MP}} - e_{\mathrm{G}}$ | 0.058 (0.329) | 0.109 (0.328) | 0.145 (0.391) | 0.173 (0.387) |
| constant $\kappa_{\lvert S\rvert}$ of Proposition 6 (small-$m$ value) | $\approx 0.531$ | $\approx 1.221$ ($\approx 0.77$) | $\approx 1.221$ ($\approx 0.90$) | $\approx 1.221$ ($\approx 0.99$) |

Table 10: Residual certificates for two positions in mixtures with larger vocabularies (Appendix E.2); the vocabulary size and the logit scale change together. For each target of retained anchor mass, $k$ is the smallest number of queried values that reaches it; the realized mass is reported. *Recovered* is $\sum_j (b_j - U_j) / \sum_j (b_j - \mathrm{TC}_j)$ over sets $j$, with $b_j$ the discarding certificate. $L_R - \mathrm{TC}$ lower-bounds the slack of the tightest certificate $\beta_R^\star$, which only more queries can reduce because the full-query optimum equals TC for two positions, and $U_R - L_R \le \kappa_2 r$ bounds the slack due to approximation. The last column gives the percentages of sets that are certified, not certifiable with this information (in parentheses, those that nevertheless meet the budget, $\mathrm{TC} \le \tau < L_R$), and undecided at $\tau = 0.25$ bits; all percentages, including those in parentheses, are relative to all 500 sets; 92.2%, 65.8%, 68.2% and 76.2% of the sets meet the budget for the four vocabularies. Computing the bracket took at most 18 ms per set in our Python implementation.

| $\lvert\mathcal{V}\rvert$ (mean TC) | target | realized mass | mean $k$ | discarding | $U_R$ | recovered | $L_R - \mathrm{TC}$ | $U_R - L_R$ | certified / not (meets budget) / undecided |
|---|---|---|---|---|---|---|---|---|---|
| 4 (0.093) | 0.5 | 0.657 | 1.8 | 0.552 | 0.333 | 47.8% | 0.222 | 0.018 | 35.8 / 62.8 (55.0) / 1.4% |
| | 0.8 | 0.883 | 2.8 | 0.255 | 0.109 | 90.0% | 0.016 | 0.001 | 90.6 / 9.2 (1.4) / 0.2% |
| | 0.95 | 0.992 | 3.7 | 0.103 | 0.093 | 97.7% | 0.000 | 0.000 | 92.2 / 7.8 (0.0) / 0.0% |
| 16 (0.218) | 0.5 | 0.585 | 2.6 | 1.088 | 1.016 | 8.4% | 0.766 | 0.032 | 4.4 / 95.4 (61.2) / 0.2% |
| | 0.8 | 0.833 | 5.8 | 0.573 | 0.522 | 14.4% | 0.294 | 0.011 | 18.8 / 78.8 (44.6) / 2.4% |
| | 0.95 | 0.960 | 9.6 | 0.306 | 0.276 | 33.7% | 0.057 | 0.002 | 62.2 / 37.8 (3.6) / 0.0% |
| 64 (0.281) | 0.5 | 0.550 | 4.2 | 1.576 | 1.551 | 1.9% | 1.239 | 0.031 | 5.4 / 94.6 (62.8) / 0.0% |
| | 0.8 | 0.815 | 10.9 | 0.821 | 0.808 | 2.4% | 0.514 | 0.012 | 5.4 / 94.4 (62.6) / 0.2% |
| | 0.95 | 0.953 | 23.1 | 0.422 | 0.415 | 4.6% | 0.131 | 0.003 | 64.0 / 36.0 (4.2) / 0.0% |
| 256 (0.350) | 0.5 | 0.544 | 4.7 | 1.652 | 1.646 | 0.5% | 1.285 | 0.012 | 7.6 / 92.4 (68.6) / 0.0% |
| | 0.8 | 0.809 | 15.8 | 0.896 | 0.894 | 0.3% | 0.540 | 0.004 | 8.4 / 91.6 (67.8) / 0.0% |
| | 0.95 | 0.951 | 46.0 | 0.492 | 0.491 | 0.8% | 0.140 | 0.001 | 72.0 / 28.0 (4.2) / 0.0% |

Table 11: Detection rates of the coherence tests on 2,000 pairs of masked positions of the controlled instance ($|\mathcal{V}| = 4$). The first two denoisers are coherent; the others are not (Appendix E.2). The mean $\mathrm{TV}(Q^{a\to b}, Q^{b\to a})$ measures the incoherence; tests trigger when a statistic exceeds $\tau = 10^{-9}$ or, in the last two columns, a guard band of $\tau = 0.01$, in bits for the entropy tests. The one-direction test is Proposition 13(i), the two-direction entropy test is Proposition 14, and the probability-level test is Proposition 15(i); *full* means $k = |\mathcal{V}|$.

| | | $\tau = 10^{-9}$ | | | | $\tau = 0.01$ | |
|---|---|---|---|---|---|---|---|
| denoiser | mean TV | one-direction full | two-direction full | two-direction $k = 2$ | probability $k = 2$ | two-direction full | probability $k = 2$ |
| exact | $< 10^{-15}$ | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| coherent, miscalibrated | $< 10^{-15}$ | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% |
| pass-dependent noise, $\sigma = 0.1$ | 0.056 | 33.8% | 100.0% | 2.5% | 100.0% | 88.1% | 95.2% |
| pass-dependent noise, $\sigma = 0.3$ | 0.167 | 49.7% | 100.0% | 17.6% | 100.0% | 96.5% | 99.9% |
| mismatched clamped passes | 0.317 | 60.1% | 100.0% | 32.3% | 100.0% | 96.8% | 100.0% |

Table 12: Exact sampler laws on the controlled instance ($\mathrm{TC}(X) = 2.1675$ bits). Top: total variation to the target at matched expected cost $K$ with an exact denoiser, for the randomized mixture of the two adjacent operating points of each rule; the interpolated value is the bound of Lemma 29, and the realized value is computed from the mixture law. Bottom: total variation at $e = 0$ under three reveal orders, namely lowest entropy first (the prefix at $e = 0$), highest entropy first (the anchored rule at $e = 0$), and left to right. The coherent but miscalibrated denoiser has the same error under every order, equal to the distance of its model joint from the target, as Proposition 18 requires.

| rule, quantity | $K = 6.5$ | $K = 6.0$ | $K = 5.5$ | $K = 5.0$ | $K = 4.5$ |
|---|---|---|---|---|---|
| prefix, interpolated bound | 0.0261 | 0.0616 | 0.1049 | 0.1687 | 0.2326 |
| prefix, realized | 0.0258 | 0.0568 | 0.0910 | 0.1410 | 0.2147 |
| anchored, interpolated bound | 0.0129 | 0.0418 | 0.0810 | 0.1160 | 0.1440 |
| anchored, realized | 0.0114 | 0.0408 | 0.0745 | 0.1044 | 0.1377 |

| denoiser at $e = 0$ | lowest entropy first | highest entropy first | left to right |
|---|---|---|---|
| exact | 0 | 0 | 0 |
| coherent, miscalibrated | 0.4818 | 0.4818 | 0.4818 |
| incoherent (position-split) | 0.4168 | 0.4482 | 0.4314 |

### F.2 CONDITIONAL QUERIES TO LLADA

Table 13 gives, for every combination of size, rule and truncation level, the averaged scores and retained masses in both directions, the incompatibility bound (6), and the mean ratios $\bar{r} = \overline{d_2/d_1}$. Standard errors and the ratio of means are in the supplementary material.

Table 13: Coherence statistics for all 36 combinations of size $m$, rule and truncation level $k$. Scores $\bar{d}_2$ and retained masses $\bar{w}$ are hierarchical means over 35 prompts; $\bar{d}_1 = 0.2009$, $0.6085$ and $1.2419$ bits at $m = 2, 3, 4$. IncLB is computed from the averaged truncation intervals with $\log_2 |\mathcal{V}| = 16.948$ (Appendix E.3); positive values, in bold, certify incoherence on a set of states of positive weight. Negative mean ratios $\bar{r}$, in bold, imply that sets with $d_2 < 0$ exist (Proposition 13(i)).

| | | | $a \to b$ | | $b \to a$ | | | $\bar{r}$ | |
|---|---|---|---|---|---|---|---|---|---|
| $m$ | rule | $k$ | $\bar{d}_2$ | $\bar{w}$ | $\bar{d}_2$ | $\bar{w}$ | $\mathrm{IncLB}_k$ | $a \to b$ | $b \to a$ |
| 2 | prefix | 1 | 0.1729 | 0.9131 | 0.2122 | 0.9572 | 0 | 0.568 | **−1.966** |
| 2 | prefix | 2 | 0.1603 | 0.9688 | 0.1930 | 0.9874 | 0 | 0.499 | **−2.007** |
| 2 | prefix | 4 | 0.1529 | 0.9877 | 0.1850 | 0.9958 | 0 | 0.443 | **−2.025** |
| 2 | prefix | 8 | 0.1486 | 0.9945 | 0.1821 | 0.9983 | **0.0045** | 0.396 | **−2.033** |
| 2 | anchored | 1 | 0.1768 | 0.7370 | 0.3141 | 0.9572 | 0 | 0.625 | **−3.165** |
| 2 | anchored | 2 | 0.1592 | 0.8603 | 0.2634 | 0.9874 | 0 | 0.319 | **−3.257** |
| 2 | anchored | 4 | 0.1382 | 0.9264 | 0.2465 | 0.9958 | **0.0374** | **−0.092** | **−3.289** |
| 2 | anchored | 8 | 0.1198 | 0.9626 | 0.2407 | 0.9983 | **0.0920** | **−0.405** | **−3.304** |
| 2 | uniform | 1 | 0.1747 | 0.8314 | 0.2604 | 0.9572 | 0 | 0.589 | **−7.580** |
| 2 | uniform | 2 | 0.1593 | 0.9197 | 0.2238 | 0.9874 | 0 | 0.418 | **−7.649** |
| 2 | uniform | 4 | 0.1450 | 0.9602 | 0.2108 | 0.9958 | 0 | 0.200 | **−7.674** |
| 2 | uniform | 8 | 0.1338 | 0.9803 | 0.2062 | 0.9983 | **0.0434** | 0.017 | **−7.687** |
| 3 | prefix | 1 | 0.5634 | 0.8713 | 0.6235 | 0.9151 | 0 | 0.768 | 0.431 |
| 3 | prefix | 2 | 0.5381 | 0.9453 | 0.5862 | 0.9686 | 0 | 0.723 | 0.393 |
| 3 | prefix | 4 | 0.5195 | 0.9751 | 0.5677 | 0.9874 | 0 | 0.690 | 0.377 |
| 3 | prefix | 8 | 0.5077 | 0.9882 | 0.5586 | 0.9944 | 0 | 0.668 | 0.369 |
| 3 | anchored | 1 | 0.5673 | 0.7186 | 0.7710 | 0.9151 | 0 | 0.821 | **−0.550** |
| 3 | anchored | 2 | 0.5363 | 0.8451 | 0.6949 | 0.9686 | 0 | 0.685 | **−0.622** |
| 3 | anchored | 4 | 0.4986 | 0.9166 | 0.6615 | 0.9874 | 0 | 0.496 | **−0.649** |
| 3 | anchored | 8 | 0.4642 | 0.9573 | 0.6467 | 0.9944 | **0.0878** | 0.364 | **−0.662** |
| 3 | uniform | 1 | 0.5642 | 0.8007 | 0.6938 | 0.9151 | 0 | 0.778 | **−0.525** |
| 3 | uniform | 2 | 0.5361 | 0.8991 | 0.6361 | 0.9686 | 0 | 0.690 | **−0.580** |
| 3 | uniform | 4 | 0.5079 | 0.9482 | 0.6095 | 0.9874 | 0 | 0.582 | **−0.602** |
| 3 | uniform | 8 | 0.4850 | 0.9741 | 0.5973 | 0.9944 | **0.0175** | 0.510 | **−0.612** |
| 4 | prefix | 1 | 1.1838 | 0.8325 | 1.2595 | 0.8740 | 0 | 0.869 | 0.874 |
| 4 | prefix | 2 | 1.1513 | 0.9172 | 1.2078 | 0.9447 | 0 | 0.836 | 0.842 |
| 4 | prefix | 4 | 1.1199 | 0.9587 | 1.1783 | 0.9743 | 0 | 0.810 | 0.827 |
| 4 | prefix | 8 | 1.0969 | 0.9797 | 1.1604 | 0.9878 | 0 | 0.794 | 0.819 |
| 4 | anchored | 1 | 1.1847 | 0.7006 | 1.4329 | 0.8740 | 0 | 0.884 | 2.921 |
| 4 | anchored | 2 | 1.1421 | 0.8289 | 1.3441 | 0.9447 | 0 | 0.801 | 2.868 |
| 4 | anchored | 4 | 1.0896 | 0.9056 | 1.2974 | 0.9743 | 0 | 0.712 | 2.845 |
| 4 | anchored | 8 | 1.0400 | 0.9513 | 1.2717 | 0.9878 | **0.0241** | 0.645 | 2.834 |
| 4 | uniform | 1 | 1.1823 | 0.7707 | 1.3417 | 0.8740 | 0 | 0.879 | 1.171 |
| 4 | uniform | 2 | 1.1438 | 0.8763 | 1.2703 | 0.9447 | 0 | 0.824 | 1.128 |
| 4 | uniform | 4 | 1.1018 | 0.9340 | 1.2312 | 0.9743 | 0 | 0.769 | 1.109 |
| 4 | uniform | 8 | 1.0656 | 0.9665 | 1.2089 | 0.9878 | 0 | 0.731 | 1.099 |

### F.3 COMPOSITION AT COMMON STATES ON LLADA

Table 14: Decrease of the mean model entropy at common states, in bits per position, averaged equally over 35 development prompts. $A$, $P$ and $U$ denote the anchored set, the prefix and the uniform average over all feasible sets; parentheses give population standard deviations over the prompt means. The next-block region contains the common positions of the next block, and the all-later region contains all subsequent common positions. Categorical values average four draws at the saved argmax states; categorical and argmax results come from separate runs. Interior-family comparisons are in the supplementary material.

| | | next block | | | | | all later |
|---|---|---|---|---|---|---|---|
| commit | $m$ | $D_A$ | $D_P$ | $D_U$ | $D_A - D_P$ | $D_A - D_U$ | $D_A - D_P$ |
| argmax | 2 | 0.8281 | 0.3230 | 0.5814 | 0.5051 (0.1326) | 0.2467 (0.0796) | 0.1242 (0.0394) |
| argmax | 3 | 0.8880 | 0.4682 | 0.6825 | 0.4198 (0.1385) | 0.2055 (0.0798) | 0.1035 (0.0439) |
| argmax | 4 | 0.9462 | 0.6054 | 0.7822 | 0.3407 (0.1343) | 0.1640 (0.0743) | 0.0829 (0.0433) |
| categorical | 2 | 0.6851 | 0.2617 | 0.4832 | 0.4235 (0.1163) | 0.2019 (0.0721) | 0.0987 (0.0344) |
| categorical | 3 | 0.6777 | 0.3386 | 0.5157 | 0.3392 (0.0983) | 0.1621 (0.0575) | 0.0775 (0.0315) |
| categorical | 4 | 0.6716 | 0.3989 | 0.5419 | 0.2727 (0.0990) | 0.1297 (0.0557) | 0.0561 (0.0335) |

### F.4 Complete generations on LLaDA

Table 15: Accuracy and mean NFE of complete generations. The prefix and anchored rules are compared at every common budget $e$ (bits). Accuracy is in percent, and $\Delta$ is anchored minus prefix in percentage points with paired 95% intervals (Newcombe, 1998). Fixed schedules are listed as steps: accuracy. All intervals are descriptive and unadjusted, and the categorical setting uses one sampling seed. Per-configuration intervals, all pairwise comparisons, and time and memory statistics are in the supplementary material.

| | prefix | | anchored | | anchored − prefix | |
|---|---|---|---|---|---|---|
| $e$ | acc. (%) | NFE | acc. (%) | NFE | $\Delta$ acc. | 95% interval |
| GSM8K, argmax ($n$ = 1,319); fixed: 256: 79.00%; 128: 77.10%; 64: 63.91% | | | | | | |
| 1 | 79.00 | 73.78 | 74.75 | 63.75 | −4.25 | [−6.59, −1.90] |
| 2 | 74.53 | 65.32 | 72.86 | 59.63 | −1.67 | [−4.16, 0.82] |
| GSM8K, categorical ($n$ = 1,319); fixed: 256: 68.54%; 128: 25.25%; 64: 7.05% | | | | | | |
| 1 | 52.24 | 116.13 | 42.53 | 118.16 | −9.70 | [−12.95, −6.43] |
| 2 | 31.31 | 137.46 | 30.71 | 118.22 | −0.61 | [−3.89, 2.68] |
| MBPP, argmax ($n$ = 500); fixed: 256: 39.60%; 128: 33.60%; 64: 20.00%; 32: 10.20% | | | | | | |
| 0.1 | 8.60 | 87.33 | 13.00 | 46.50 | 4.40 | [1.26, 7.62] |
| 0.25 | 8.80 | 65.86 | 12.20 | 37.35 | 3.40 | [0.22, 6.65] |
| 0.5 | 10.00 | 49.22 | 16.40 | 31.25 | 6.40 | [3.01, 9.87] |
| 1 | 16.80 | 36.00 | 14.40 | 25.01 | −2.40 | [−6.00, 1.18] |
| 2 | 17.60 | 25.78 | 13.40 | 18.86 | −4.20 | [−7.90, −0.53] |
| 4 | 17.80 | 17.90 | 15.40 | 13.30 | −2.40 | [−6.00, 1.18] |

## G Background and related work

**Masked diffusion and how much to reveal.** Discrete denoising diffusion (Ho et al., 2020; Austin et al., 2021a) generalizes diffusion to finite state spaces, and absorbing-state corruption connects it to masked language modeling (Devlin et al., 2019). Continuous-time and score-ratio formulations (Campbell et al., 2022; Lou et al., 2024), simplified objectives (Sahoo et al., 2024), an analysis of the time variable (Zheng et al., 2024), and continuous-space variants (Li et al., 2022; Han et al., 2023) followed. At sampling time, such a model is an any-order autoregressive model over reveal sets (Hoogeboom et al., 2022) whose order can be planned (Peng et al., 2025). Wu et al. (2026) unmask every position whose top-one probability exceeds a threshold, Kasa et al. (2026) select commit sets from the full sorted confidence profile, and Ben-Hamu et al. (2025) unmask the largest entropy-ordered prefix with bounded excess bits. Cai & Li (2026) study the iteration complexity of confidence rules, Chen et al. (2025) use trajectory-level consistency, Fu et al. (2025) lower bound the number of decoding rounds by the total information of the sample divided by the information revealed per round, and Kang et al. (2026) construct benchmarks on which the loss of token dependence under parallel decoding is pronounced.

**Dependency-aware, independence-testing and lookahead decoding.** Committing a set from a product of marginals is exact only under conditional independence. DCD combines the diffusion model with a pretrained copula model (Liu et al., 2025); DEMASK and neural pairwise mutual-information estimation use learned dependency predictors (Ringel et al., 2026; Sharma et al., 2026); DAPD induces a dependency graph from self-attention (Kim et al., 2026); and PUNT tests approximate conditional independence with additional passes of the base denoiser (Azangulov et al., 2026). LookUM scores candidate unmasking paths (Lee et al., 2025), RPS selects and assigns a pivot position of intermediate entropy (Ye et al., 2026), and TABES uses gradients to estimate the reduction of the remaining entropy (Saini et al., 2026). The certificates of Section 4 are complementary: they give numerical upper bounds on the total correlation of a reveal set that are valid whenever the queried outputs are exact, together with tests of the coherence that this interpretation requires.

**Compatibility of learned conditionals.** Whether a family of conditional distributions derives from one joint law is the classical compatibility problem (Arnold & Press, 1989), and dependency networks specify models through possibly incompatible conditionals (Heckerman et al., 2000). For masked language models, Torroba Hennigen & Kim (2023) derive pairwise joints from unary conditionals, find those conditionals to be incompatible, and construct near-compatible joints, which requires on the order of $|\mathcal{V}|$ forward passes per pair; Young et al. (2024) show that distributions from different masking patterns of the same model are inconsistent. For diffusion language models, Kim (2026) defines order-induced pseudo-joints, which for two positions are the laws $Q^{a\to b}$ and $Q^{b\to a}$ of Definition 12, characterizes order consistency by the vanishing of their log-ratios, and separates incompatibility from conditional dependence as sources of parallel-decoding error. Our tests certify incompatibility from the truncated queries and entropy summaries that a depth-2 certificate already produces, with interval and residual arguments that account for the omitted values.

**Minimum-entropy couplings.** The minimum-entropy coupling problem asks for the joint law of least entropy with given marginals (Kocaoglu et al., 2017; Cicalese et al., 2019). It is NP-hard (Kovačević et al., 2015), admits an additive approximation scheme for a fixed number of marginals (Compton, 2025), and approximation guarantees for the greedy coupling are measured relative to the majorization lower bound (Cicalese et al., 2019; Compton, 2022) or to the profile bounds (Compton et al., 2023). Minimum-entropy couplings maximize mutual information among couplings (Cicalese et al., 2019; Compton et al., 2023); Proposition 6 uses this correspondence and these guarantees to identify and bracket the best marginal-only certificate.

**Draft-conditioned refinement and information-theoretic identities.** Draft-then-refine pipelines remask low-confidence tokens (Huang et al., 2026), supervise visible errors (Zhang et al., 2025), refine a completed response by minimal edits (Ren et al., 2026), recycle discarded representations (Hu et al., 2026), or contrast predictions with those obtained after temporarily remasking high-information tokens (Feng et al., 2026). When a draft or other side information is supplied to the denoiser, our results apply with that information included in $C$; conditioning additionally on $Y$ changes the total correlation by $\mathrm{TC}(X \mid C) - \mathrm{TC}(X \mid C, Y) = \sum_i I(X_i; Y \mid C) - I(X; Y \mid C)$, the correlation explained by $Y$ in the sense of Ver Steeg & Galstyan (2014). The discrete analogs of the I-MMSE relation (Guo et al., 2005; Kong et al., 2023) for masked processes are due to Jeon et al. (2025), with continuous-time counterparts in Casado Noguerales et al. (2026); they assume reveal indicators that are independent of the data, which adaptive samplers violate (Theorem 26). The accumulation of conditional dependence across rounds is established by Li & Cai (2025); Lavenant & Zanella (2025); Chen et al. (2026); Zhang et al. (2026), and Wen et al. (2026) prove that the forward divergence of any adaptive policy equals the conditional total correlation accumulated over its reveal rounds.