THE PARALLEL-DECODING SCORE OF MASKED DIFFUSION LANGUAGE MODELS IS NOT A MUTUAL INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-free parallel decoders for masked diffusion language models decide how many positions to commit by bounding the dependence among them. The score they bound, $D(i{\to}j) = \mathbb{E}_{x_i \sim p_\theta(x_i|c)}[\mathrm{KL}(p_\theta(x_j \mid c, x_i) \,\|\, p_\theta(x_j \mid c))]$, is read as a conditional mutual information, a reading that presumes the model's any-order conditionals are those of a single joint. We give an exact model-internal identity, $D = I_q + \Delta$: I_q is the mutual information of the model's own one-at-a-time rollout and $\Delta \geq 0$ is a *self-marginalization defect*, the divergence between the marginal the model reports directly and the marginal its rollout induces. The identity splits any set's parallel error into a total correlation plus a sum of defects, and yields three one-sided certificates of incoherence. With estimators whose bias direction is known, the defect's mean is certified positive over pairs on four models and over scattered sets of two to eight positions on three, and on peaked anchors a conservative estimate puts it at a median 0.37–0.45 of the score. A set's parallel error also depends on the order in which it is revealed, by several times replicate noise. Where the two marginals disagree, the rollout's predicts the documents' own tokens better on average, so a parallel step samples from the less accurate one. Inside EB-Sampler's GSM8K decodes on LLaDA-8B-Instruct, half the decodes overrun the entropy allowance while the dependence it is written against stays within it: on the positions measured, the overrun comes from the defect. All code and the records behind every number are in the supplementary material and will be released at `https://github.com/xxx/xxx` upon acceptance.

1 INTRODUCTION

Masked diffusion language models (dLLMs) generate text by unmasking many positions in one forward pass, and their speed depends on how many positions can be committed together (Nie et al., 2025; Ye et al., 2025; Sahoo et al., 2024). Committing a set S in one step samples each position from its own marginal, which is exact only when the positions are conditionally independent given the context. One family of samplers measures that dependence with extra forward passes or learns to predict it (Azangulov et al., 2026; Wu & Huang, 2026; Sharma et al., 2026; Ringel et al., 2026), and a cheaper family commits by confidence or an entropy bound (Wu et al., 2026; Ben-Hamu et al., 2025; Luxembourg et al., 2026), justified by an analysis that identifies the parallel-decoding error with a conditional total correlation (Kang et al., 2026; Chen et al., 2026a; Liu et al., 2025; Zhang et al., 2026). The per-pair score $D(i{\to}j)$ above, how far revealing an *anchor* position i moves the model's marginal at a *target* j, is the model's counterpart of what these works call, for the data, "precisely joint mutual information" and fit neural estimators to (Ben-Hamu et al., 2025; Wu & Huang, 2026; Sharma et al., 2026).

These identifications rest on an unmeasured premise: that the conditionals a dLLM returns for different masking patterns are those of a single joint distribution. That a family of conditionals need not belong to any joint is classical (Besag, 1974; Heckerman et al., 2000), and masked language models are known to violate it (Wang & Cho, 2019; Goyal et al., 2022; Yamakoshi et al., 2022; Young et al., 2024); Kim (2026) named the dLLM counterpart, the *order-induced pseudo-joint*, without measuring it. When it fails, the score has no joint to be a mutual information of, "error = total correlation" needs a further condition (§2), and a set's total correlation can depend on the order.

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.

This paper carries out that measurement, and asks what it changes for decoding. We contribute:

- An exact identity, $D(i{\to}j) = I_q(X_i; X_j) + \Delta(j \mid i)$, with Δ the *self-marginalization defect*, its k-position form $\mathrm{KL}(q_\pi \| m_S) = \mathrm{TC}(q_\pi) + \sum_{i \in S} \Delta_\pi(i)$, three one-sided certificates of incoherence, two of which a score probe has already paid for, and a zero-cost defect-corrected estimator (§2). The algebra is the classical reverse I-projection decomposition (Amari, 2001; Cook, 2025); the contribution is a measurement framework and what it says about the score.
- That measurement, with an estimator that bounds Δ from below and budget, precision and placebo controls: the defect is certified positive over pairs on four models and over sets on three, is a large share of the score by a conservative estimate, marks the one-pass marginal as the less accurate against the text, and makes a set's parallel error depend on more than the set (§3–4.1).
- The form that matters operationally: the share is at least $1 - H(p_i)/D$, and empirically rises from 0.21–0.24 to 0.96 across quartiles of the anchor's peakedness on all three architectures.
- The decoding consequence (§5): D is the parallel step's error exactly *for a pair*, relative to revealing the anchor first, and between candidates it forgoes less of the realised gain than $\hat{I}_{\mathrm{cor}}$; over a set it barely beats a random choice on an enumerable model, as a decoder it collapses, and inside deployed samplers a large part of the parallel error is defect, which their budgets omit.

2 The score, the pseudo-joint, and the identity

Setup. A sequence is $x = (x_1, \ldots, x_L) \in V^L$. A partially revealed state is a context c: the revealed tokens together with the set $\mathcal{M}(c) \subseteq [L]$ of positions still masked. One forward pass returns, for every $i \in \mathcal{M}(c)$, a *direct marginal* $p_i(\cdot) := p_\theta(x_i = \cdot \mid c)$; writing $c \oplus (i{:}v)$ for the context with position i filled by v, a second returns $p_\theta(x_j \mid c \oplus (i{:}v))$. The network provides this family of local conditionals and nothing more, a generalized *dependency network* (Heckerman et al., 2000), and nothing in finite training forces the family to be compatible with any joint (Arnold & Press, 1989).

Pseudo-joint and induced marginal. Fix a set $S = \{i_1, \ldots, i_k\} \subseteq \mathcal{M}(c)$ and an order π of S. Revealing S one position at a time in that order samples from the *order-induced pseudo-joint*

$$q_\pi(x_S \mid c) := \prod_{t=1}^{k} p_\theta\big(x_{\pi_t} \mid c \oplus (\pi_{<t}{:}x_{\pi_{<t}})\big), \tag{1}$$

which Kim (2026) named and which is in general a different distribution for every π. Its marginal on position i, the *induced marginal* $q_\pi^{(i)}(\cdot) := \sum_{x_{S\setminus i}} q_\pi(x_S \mid c)$, need not equal the direct marginal p_i. Parallel unmasking of S, on the other hand, samples from the product of the *direct* marginals, $m_S(x_S \mid c) := \prod_{i \in S} p_i(x_i)$, which is what one forward pass makes available. Throughout, V is finite and every conditional the network returns is a softmax and therefore strictly positive, so $q_\pi \ll m_S$, every divergence below is finite, and every expectation splits.

Definition 1 (Coherence and the self-marginalization defect)**.** *The model is* coherent *on (c, S) if a joint $P(x_S)$ exists whose conditionals agree with the model's for every $i \in S$, every subset $A \subseteq S \setminus \{i\}$ and every x_A with $P(x_A) > 0$: $P(x_i \mid x_A) = p_\theta(x_i \mid c \oplus (A{:}x_A))$. The* self-marginalization defect *of position i under order π is $\Delta_\pi(i) := \mathrm{KL}(q_\pi^{(i)} \,\|\, p_i) \geq 0$.*

Coherence implies $q_\pi = P$ for every π, hence $\Delta_\pi(i) = 0$; the converse fails (Appendix B), so zero defect is strictly weaker than coherence, and is exactly the condition under which (2) reduces to "error = total correlation" — of the order-π rollout, which can still depend on π (Appendix B).

Proposition 1 (Parallel error decomposes into dependence plus defects)**.** *For every S and every order π,*

$$\mathrm{KL}\big(q_\pi \,\|\, m_S\big) = \mathrm{TC}(q_\pi) + \sum_{i \in S} \Delta_\pi(i), \qquad \mathrm{TC}(q_\pi) := \mathrm{KL}\Big(q_\pi \,\Big\|\, \prod_{i \in S} q_\pi^{(i)}\Big), \tag{2}$$

and all three quantities are non-negative.

The proof (Appendix A) is the Pythagorean relation for the reverse I-projection of q_π onto the product family (Amari, 2001; Cook, 2025). Prior statements of "error = total correlation" concern the data distribution, an ideal denoiser, the ELBO or a coherent model (Liu et al., 2025; Ben-Hamu et al.,

2025; Kang et al., 2026; Chen et al., 2026a; Zhang et al., 2026); here both sides are the *deployed model's* own quantities, and the second term is what those statements assume away.

Proposition 2 (What the score measures). *Let $S = \{i, j\}$, $\pi = (i, j)$, and let $D(i{\to}j) := \mathbb{E}_{v\sim p_i}\big[\mathrm{KL}(p_\theta(x_j \mid c \oplus (i{:}v)) \,\|\, p_j)\big]$ be the score. Then*

$$(a)\quad D(i{\to}j) = \mathrm{KL}(q_{(i,j)} \,\|\, m_S), \tag{3}$$

$$(b)\quad D(i{\to}j) = I_q(X_i; X_j) + \Delta(j \mid i), \qquad \Delta(j \mid i) = \mathrm{KL}(\bar{q}_j \,\|\, p_j), \tag{4}$$

where $\bar{q}_j := \mathbb{E}_{v\sim p_i}[p_\theta(\cdot \mid c{\oplus}(i{:}v))]$ and $I_q = \mathbb{E}_v[\mathrm{KL}(p_\theta(\cdot \mid c{\oplus}(i{:}v)) \| \bar{q}_j)]$ is the mutual information of the pseudo-joint $q_{(i,j)}$, and $0 \le \Delta(j \mid i) \le D(i{\to}j)$.

Part (a) says the score *is* the parallel-step error for the pair, relative to revealing it in the order (i, j): the reverse reference order gives $D(j{\to}i)$, and Table 1's asymmetry row measures what that choice is worth. Part (b) is Proposition 1 at $k = 2$, where $\Delta_\pi(i) = 0$, $\Delta(j \mid i) = \Delta_{(i,j)}(j)$ and $\mathrm{TC} = I_q$; it is also Topsøe's compensation identity $\mathbb{E}_v[\mathrm{KL}(P_v \| R)] = \mathbb{E}_v[\mathrm{KL}(P_v \| \bar{P})] + \mathrm{KL}(\bar{P} \| R)$ (Topsøe, 2001) at $P_v = p_\theta(\cdot \mid c \oplus (i{:}v))$, $R = p_j$. The score is the pseudo-joint's mutual information *plus* a term that vanishes exactly when $\bar{q}_j = p_j$, which coherence implies and which is strictly weaker than it (Appendix B); (a) and (b) pull in different directions for decoding (§5).

Proposition 3 (Three one-sided certificates). *Let $S = \{i, j\}$ and $\pi = (i, j)$, and write $q_{(i,j)}$ for the pseudo-joint of Proposition 2. (A) If the model is coherent on $(c, \{i, j\})$ with witnessing joint P, then $D(i{\to}j) = D(j{\to}i) = I_P(X_i; X_j)$, the mutual information under P, so $A(i, j) := D(i{\to}j) - D(j{\to}i) \neq 0$ certifies incoherence. (B) Let $I_{\mathrm{ent}}(i{\to}j) := H(p_j) - \mathbb{E}_{v\sim p_i}[H(p_\theta(\cdot \mid c{\oplus}(i{:}v)))]$ be the entropy-difference form of mutual information. Under coherence $I_{\mathrm{ent}} = I_P \ge 0$, so $I_{\mathrm{ent}}(i{\to}j) < 0$ certifies incoherence. In general $I_{\mathrm{ent}} = I_q + [H(p_j) - H(\bar{q}_j)]$: the two differ by how much entropy the direct marginal has in excess of the induced one, and that difference has either sign. (C) $q_{(i,j)}$'s marginal on i is p_i by construction, so $I_q(X_i; X_j) \le H(p_i)$ and, for $D(i{\to}j) > 0$, the defect's share obeys*

$$\rho(i{\to}j) := \frac{\Delta(j \mid i)}{D(i{\to}j)} \;\ge\; 1 - \frac{H(p_i)}{D(i{\to}j)}. \tag{5}$$

In particular $\rho \to 1$ along any sequence with $H(p_i)/D(i{\to}j) \to 0$: at an anchor whose entropy is small against the score, the score is almost entirely defect, and $D(i{\to}j) > H(p_i)$ certifies $\Delta(j \mid i) > 0$.

All three are one-sided ($A = 0$, $I_{\mathrm{ent}} \ge 0$ and a vacuous (5) do not imply coherence). (B) and (C) need no estimator beyond the forward passes a D probe already runs, if it estimates the full expectation over $v \sim p_i$ (a truncated probe can fire (B) on a coherent model); (A) needs the reverse direction probed as well, which our probe does only for pairs of anchors, so Table 1's asymmetry row is measured on anchor–anchor pairs, peaked at both ends (Appendix F). (A) holds for any expected f-divergence in place of the KL, total variation included, since under coherence each is symmetric in i and j. (B) is the quantity Sharma et al. (2026) use as a supervised target. (C) matters because a confidence-ranked decoder anchors on its most nearly deterministic position, where $I_q \le H(p_i)$ is smallest; §5 measures the share such a decoder meets.

Proposition 4 (Zero-cost corrected estimator). *From m samples $v_s \sim p_i$ and the conditionals $P_s = p_\theta(\cdot \mid c \oplus (i{:}v_s))$ that any D estimator already computes, $\hat{D} = \frac{1}{m}\sum_s \mathrm{KL}(P_s \| p_j)$, $\bar{P} = \frac{1}{m}\sum_s P_s$, $\hat{\Delta} = \mathrm{KL}(\bar{P} \| p_j)$ and $\hat{I}_{\mathrm{cor}} = \hat{D} - \hat{\Delta} = \frac{1}{m}\sum_s \mathrm{KL}(P_s \| \bar{P}) \ge 0$. The correction costs no additional forward pass. Below, $\hat{\Delta}$ is jackknifed and $\hat{I}_{\mathrm{cor}}$ clipped at zero (Appendix F).*

$\hat{I}_{\mathrm{cor}}$ estimates I_q and is non-negative; I_{ent} equals I_q only when $H(p_j) = H(\bar{q}_j)$. Every negativity result below is about I_{ent}, and the decoding rules of §5 rank by $\hat{I}_{\mathrm{cor}}$ or $\hat{D}$.

3 Estimating the defect

The defect is a KL between a mixture and a marginal over a 50k–152k-token vocabulary; its plug-in estimate is biased up by $O(1/m)$: on a 400-token enumerable example with a 10% head (Appendix F) it averages 0.334 against a true 0.250, the jackknife 0.269 and the split-sample bound 0.077.

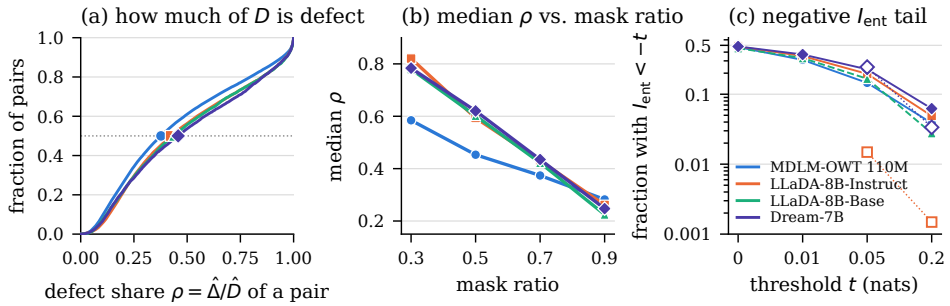

Figure 1: The defect across four models. (a) Cumulative distribution of the per-pair defect share $\rho = \hat{\Delta}/\hat{D}$; markers are the jackknife medians of Table 1. (b) Median ρ against the fraction of positions still masked: the fewer remain masked, the larger the share of the score that is defect. (c) Fraction of pairs with $I_{\text{ent}} < -t$, zero under coherence; open markers are replicate nulls (Table 1).

Estimator. We stratify the expectation over $v \sim p_i$: the top-K tokens of p_i are enumerated exactly and the tail is sampled, discarding anchors whose top-K mass is below 0.80, where the tail dominates the variance. $\hat{\Delta}$ is jackknifed over the tail draws; sums accumulate in float64 (Appendix F).

A lower bound in expectation. Writing $\langle u, w\rangle := \sum_v u(v)w(v)$, the samples are split into halves A and B and the defect is estimated as $\hat{\Delta}_{\text{split}} = \frac{1}{2}[\langle \bar{P}_A, \log \bar{P}_B - \log p_j\rangle + \langle \bar{P}_B, \log \bar{P}_A - \log p_j\rangle]$. Since A and B are independent, $\mathbb{E}\langle \bar{P}_A, \log \bar{P}_B\rangle \le \sum_v \bar{q}(v) \log \bar{q}(v)$ by Jensen, hence $\mathbb{E}[\hat{\Delta}_{\text{split}}] \le \Delta$, and a positive one-sided bound on its mean over pairs or sets *certifies* that their mean defect is positive: the finite-sample bias points the other way. Columns 2–4 of Table 8 are one-sided bounds on the mean of this quantity, and it is the numerator of Table 1's conservative rows. The *share* ρ is a ratio, and a bound on the numerator's mean bounds neither the ratio's expectation nor a median over pairs; where we report a median share with that numerator we call it conservative, not certified, and check it on the enumerable model of Appendix I, where it reads 0.520 against a true median of 0.526 (0.353 against 0.363 after Table 1's $\hat{D}$ filter, which selects on the same draws) while exceeding the truth on 36% of single pairs. The jackknife numerator is not conservative even in aggregate.

Controls. (i) A 3.9–5.2×-budget re-estimate on subsets of 1,374, 138 and 780 pairs, which moves the median share by $+0.002$ on MDLM-OWT, -0.025 on LLaDA-8B-Instruct and -0.006 on Dream-7B, each interval containing zero (Appendix G). (ii) A count-matched placebo, since $c \oplus (i{:}v)$ has one fewer mask than c (§4.2). (iii) A numerical floor: at $k = 1$ the two marginals coincide by construction, so any Δ there is float noise, and the magnitude of the mean split-sample reading there is the model's floor (9.0×10^{-9}, 4.9×10^{-4} and 8.1×10^{-4} nats on MDLM-OWT, LLaDA-8B-Instruct and Dream-7B, the last two set by bf16 batch reductions) and a defect counts as positive only if its lower bound clears both $10\times$ the floor and 10^{-5} nats (Appendix F). (iv) Before probing, real text must beat the same tokens shuffled and an all-masked context by 0.5 nats; every model clears both by 6.0–8.2 nats, and on MDLM-OWT, LLaDA-8B-Instruct and Dream-7B the true token scores highest at its own masked slot (Appendix F).

Exact and synthetic checks. The identities hold to 7×10^{-16} on enumerable specifications, and a small masked diffusion model trained on an enumerable joint with zero defect has a median ρ of 0.18 over pairs with $D > 10^{-4}$ (0.035 above 10^{-3}) that does not fall as it converges (Appendix I).

Models and data. We probe MDLM-OWT (110M; Sahoo et al., 2024), LLaDA-8B-Instruct and LLaDA-8B-Base (Nie et al., 2025), and Dream-7B-Instruct (Ye et al., 2025) (Dream-7B below), adapted from an autoregressive model, with a 152k vocabulary. Contexts are OpenWebText documents (Gokaslan & Cohen, 2019) truncated to $L = 128$ with a fraction $r \in \{0.3, 0.5, 0.7, 0.9\}$ of positions masked; anchors pass the head-mass filter and targets are other masked positions stratified by distance. LLaDA and Dream run in bf16, MDLM in float32. Table 1 lists the sample sizes; in total the paper rests on 61,090 records, one per probed anchor row, rollout cell, decoded problem or infilled block.

Table 1: Main probe, $\rho = \hat{\Delta}/\hat{D}$. The conservative rows use the split-sample numerator of §3, with $\mathbb{E}[\hat{\Delta}_{\text{split}}] \leq \Delta$ on every pair: its mean over all pairs is a one-sided certificate and its mass-weighted share $\sum \hat{\Delta}_{\text{split}} / \sum \hat{D}$ is conservative as a ratio of expectations; the median is checked in Appendix I. Brackets on ρ and the bound resample documents, then anchor rows (Appendix F); other rows resample pairs. The ρ medians use the 46,863, 19,736, 5,295 and 5,160 pairs with $\hat{D} \geq 10^{-3}$ nats, the other ρ and I_{ent} rows every pair. Asymmetry, $|\hat{D}(i{\to}j) - \hat{D}(j{\to}i)|$ over their mean, is on anchor–anchor pairs. Replicate nulls pair equal-budget runs, except MDLM-OWT's (Appendix G).

	MDLM-OWT 110M	LLaDA-8B-Instruct	LLaDA-8B-Base	Dream-7B
rows / pairs / documents	2,224 / 68,944 / 70	1,774 / 26,610 / 75	456 / 6,840 / 20	456 / 6,840 / 20
median ρ (jackknife)	0.377 [.351,.407]	0.423 [.397,.454]	0.437 [.374,.504]	0.458 [.408,.518]
median ρ, **conservative**	**0.370** [.344,.401]	**0.414** [.388,.446]	**0.431** [.365,.500]	**0.450** [.397,.514]
mass-weighted ρ, cons.	0.456 [.405,.509]	0.292 [.270,.318]	0.254 [.210,.308]	0.305 [.235,.396]
mean $\hat{\Delta}_{\text{split}}$, 95% LB (nats)	0.026	0.017	0.013	0.013
median asymmetry	0.783 [.754,.820]	0.730 [.695,.764]	0.704 [.655,.773]	0.708 [.635,.784]
$I_{\text{ent}} < 0$ / -0.05 / -0.2 nats	47.2 / 14.5 / 3.9%	48.1 / 19.6 / 4.9%	45.6 / 16.7 / 2.8%	48.2 / 22.7 / 6.2%
replicate null, -0.05 / -0.2	0.0 / 0.0%	1.5 / 0.1%	—	24.5 / 3.4%

4 THE DEFECT ON REAL MODELS

The defect is a large share of the score, and the share is not one number. On all four models the mean split-sample defect over every pair is certified positive, its lower bound clearing $10\times$ the floor ($1.6\times$ that on Dream-7B, and on LLaDA-8B-Base against LLaDA-8B-Instruct's floor), and the conservative median share is 0.37–0.45 (jackknife 0.38–0.46; mass-weighted 0.25–0.46; Table 1), with 91–95% of pairs above 0.1 (Table 7); running LLaDA-8B-Instruct in float32 moves its median share by only -0.005, so its defect is not an artefact of bf16 arithmetic (Appendix F); on GSM8K reference solutions in the models' own chat template it is 0.425 and 0.450 on LLaDA-8B-Instruct and Dream-7B (Appendix G). But ρ is a ratio of two model-internal quantities and moves systematically with three things. (i) *The anchor.* The share rises toward 1 as the anchor sharpens, consistent with Proposition 3(C), on all three architectures: over quartiles of the anchor's top-K mass the median share of the quartile's pairs runs $0.21 \to 0.32 \to 0.54 \to 0.96$ on MDLM-OWT, $0.24 \to 0.37 \to 0.63 \to 0.96$ on LLaDA-8B-Instruct and $0.24 \to 0.41 \to 0.62 \to 0.96$ on Dream-7B (row-level Spearman $+0.76$, $+0.81$, $+0.76$). This is not a ratio collapsing to $0/0$: over the same quartiles the median $\hat{D}$ falls $5.0\times$, $7.8\times$ and $5.8\times$ while the median $\hat{\Delta}$ falls only $1.1\times$, $2.1\times$ and $1.7\times$. What vanishes at a sharp anchor is the mutual information, which (5) caps at $H(p_i)$; that the defect stays is empirical. Binned by the anchor's entropy itself, with or without the $\hat{D}$ filter, the median share runs from 0.19–0.22 above 3 nats to 0.99–1.00 below 0.05 nats on all three (Appendix P). We probe anchors with top-K mass at least 0.80, which keeps 51–56% of randomly drawn masked positions — the peaked half; §5 measures the positions a decoder commits. Without that filter the jackknifed share over all pairs is 0.179, 0.242 and 0.245 on the three architectures, against 0.446, 0.582 and 0.548 for the filter's own pairs in those runs (Appendix F). (ii) *The mask ratio* (Figure 1b): ρ is 0.22–0.28 at ratio 0.9 and 0.58–0.82 at 0.3. (iii) *The summary*: weighting pairs by their score lowers the share on the 7–8B models and raises it on MDLM-OWT (Appendix K).

How often the certificates fire. On anchor–anchor pairs the score is asymmetric by a median 70–78% of its own magnitude, and by more than half on 62–65% of them, where a coherent model would give exactly zero and estimator noise alone gives 0.02–0.12 (Appendix G). The entropy-difference I_{ent}, the supervised target of Sharma et al. (2026), is negative on 46–48% of pairs; since a pair whose true value is near zero reads negative about half the time under any noise, the informative statistic is the tail: 14.5% and 19.6% fall below -0.05 nats on MDLM-OWT and LLaDA-8B-Instruct and 3.9% and 4.9% below -0.2, against replicate-noise nulls of 0.0% at both and of 1.5% and 0.1%; Dream-7B's replicate is noisier, and its tail clears it only at -0.2 (6.2% against 3.4%), while LLaDA-8B-Base has no null (Appendix G). Certificate (C) is the weakest: $H(p_i)$ usually dwarfs D, so the bound is non-vacuous on 1.7–4.2% of pairs, and, as a check on the jackknifed estimator (the plug-in satisfies it almost by construction), violated on none of the 71,759 checked on the three models of the quartile analysis.

Relation to confidence, and the effect on selection. The defect is poorly predicted from one forward pass: regressing $\log \hat{\Delta}$ on the anchor's and target's maximum probability and entropy explains $R^2 = 0.03$–0.10, and 0.25–0.34 with log distance added; $\log \hat{D}$, which needs the probe and shares $\hat{\Delta}$'s samples, explains 0.77–0.80. By $\hat{D}$ or by $\hat{I}_{\mathrm{cor}}$, a row's safest third differs in 17–25% of its members (Table 7).

4.1 Beyond pairs: the defect at every k, and the order-dependence of "the" total correlation

Pairs are the least interesting case of Proposition 1: the first position of an order never has a defect, so at $k = 2$ only one term contributes. We reveal 8-position sets one position at a time under four orders each (two random, left-to-right, confidence-descending), so the size-k set of an order is its first k positions, with 128 rollouts per order whose per-step KLs are computed exactly given the sampled prefix (Appendix F), on 161 sets on MDLM-OWT, 110 on LLaDA-8B-Instruct and 128 on Dream-7B, and estimate every term of (2) at every prefix length. Half the sets are contiguous blocks, as block-diffusion decoders use, and half scattered.

Certified at every k. The one-sided 95% lower bound on the mean summed defect over scattered sets (Table 8, Figure 3b) is positive at every k from 2 to 8 on all three models (0.011 to 0.287 nats on MDLM-OWT, 0.009 to 0.200 on LLaDA-8B-Instruct, 0.009 to 0.140 on Dream-7B) and grows with k, as it must over nested prefixes. The estimator's sampling bias is downward, so these certify the mean over set–order cells, not each set; every bound clears its threshold, the two $k = 2$ cells of the 7–8B models only narrowly, and Dream-7B's falls to $0.94\times$ under a stricter floor convention (Appendix F). On contiguous blocks the mean jackknifed estimate is 5–10$\times$ the scattered one at $k = 8$, but the bound clears its threshold only at $k \leq 3$ on MDLM-OWT, at $k = 2$ on LLaDA-8B-Instruct and at no k on Dream-7B: the split-sample penalty, largest on blocks, falls only about as $n^{-0.55}$ with the rollout count on all three models, so blocks need a larger budget than we spent (Appendix H). That is a property of the estimator on heavy-tailed rollouts, not evidence that the defect is small there.

Order dependence of the set-level error. Under coherence, q_π and hence $\mathrm{KL}(q_\pi \| m_S)$ do not depend on π. Only at $k = 8$ do the four orders reveal the *same* set. There, the parallel error's relative spread (range over mean) across the four orders has a median of 0.260 $[0.211, 0.319]$ over 161 sets on MDLM-OWT (0.62 nats absolute), 0.224 $[0.187, 0.292]$ over 110 on LLaDA-8B-Instruct and 0.202 $[0.160, 0.267]$ over 128 on Dream-7B (Figure 3c). Its null rebuilds that statistic from noise alone: every cell was re-run with fresh rollouts at the same budget, and four of a set's run-to-run differences, scaled by $1/\sqrt{2}$ and randomly signed, stand in for its four orders, which give a median of 0.069, 0.081 and 0.083 on the three models; the spread is 3.8, 2.8 and 2.4 times that ($P < 0.001$ on each; Appendix H). A spread in $\mathrm{KL}(q_\pi \| m_S)$ could sit entirely in the defect sum, so we measure each term on the same sets against its own null: $\mathrm{TC}(q_\pi)$ alone spreads by 0.294, 0.248 and 0.196, 3.4, 2.5 and 2.3 times its null, and the defect sum by 0.47, 0.48 and 0.38, 3.4, 3.6 and 2.3 times its null ($P < 0.001$ on each). "The conditional total correlation of S" in which the schedule-level analyses are written (Kang et al., 2026; Chen et al., 2026a; Zhang et al., 2026) is therefore not one quantity for a deployed dLLM: both the parallel error it stands for and the total correlation alone depend on the order, well beyond replicate noise, on all three models.

Contiguous blocks against scattered sets. At equal k a contiguous block incurs a median parallel error 38–57$\times$ that of a scattered set at $k = 2$, falling to 19–30$\times$ at $k = 8$ on the three models (Table 2). Dilated schedules (Luxembourg et al., 2026) rest on dependence decaying with distance; both terms of (2) do, so a block decoder pays the defect as well.

4.2 What the defect is not

Two narrower sources could account for the defect; our controls exclude one and do not isolate the other. *Mask-count sensitivity*: filling i removes one mask, but a count-matched placebo that re-masks a far-away revealed position does not remove the defect (the summed defect is $4.98\times$ $[3.54, 7.24]$ its main-condition value on MDLM-OWT, $2.95\times$ $[2.28, 3.80]$ on LLaDA-8B-Instruct and $2.45\times$ $[1.27, 5.30]$ on Dream-7B, resampling documents) and re-masking the *closest* revealed

position amplifies it further. Re-masking also withholds information, and on an exactly coherent joint it reads 20% of the score, so this control does not isolate the count, and mask-count sensitivity is not excluded (Appendix E). *Instruction tuning* does not create the defect: LLaDA-8B-Base and -Instruct share their pretrained initialization, and on the 1,722 pairs with $\hat{D} > 10^{-4}$ nats in both the median paired difference in $\hat{\Delta}$ is -2×10^{-5} nats, while tuning raises the summed defect by 18–27%, carried by the upper tail. Appendix E gives both controls in full. *Nor is it the rollout's error.* Scored at the document's own token x_j^*, a proper score, $\log \bar{q}_j(x_j^*) - \log p_j(x_j^*)$ averages $+0.015$ $[+0.006, +0.025]$ nats per pair on LLaDA-8B-Instruct and $+0.015$ $[+0.003, +0.031]$ on Dream-7B over 20 new documents each: 0.82 $[0.35, 1.28]$ and 0.96 $[0.18, 1.84]$ times the mean defect, where an exact $\bar{q}_j$ gives the defect and an exact p_j a negative value. Where they disagree, the one-pass marginal a parallel step samples from is on average the less accurate, and tempering it to $\bar{q}_j$'s entropy closes almost none of the gap; a generic benefit of averaging over uninformative perturbations is not excluded (Appendix E).

5 Consequences for decoding

Proposition 2 suggests two uses. By (a), $D(i{\to}j)$ is exactly the KL cost of committing j in parallel with i, relative to revealing i first, so between two candidates for the same anchor it is the right quantity to *minimize*; by (b), $\hat{I}_{\text{cor}}$ estimates the rollout's own mutual information I_q, the closest model-internal analogue of the quantity the literature invokes. Both concern a pair, and a set's error is not a sum of pairwise terms. Scored on the rollout-estimated parallel error of the set an anchor and each score's two safest companions form (Appendix J), $\hat{D}$ is not resolved from $\hat{I}_{\text{cor}}$ in the mean ($\hat{I}_{\text{cor}}$ minus $\hat{D}$: -0.020 $[-0.131, +0.083]$ and $+0.084$ $[-0.007, +0.186]$ nats on LLaDA-8B-Instruct and Dream-7B), its sets beat random companions' on 78% and 73% of anchor rows though costly outliers leave the mean unresolved, the lowest-entropy or most confident companions' are at least as cheap, and I_{ent}'s are the dearest on average; on the enumerable model D beats a random companion by only 0.035 nats. The literature uses D for the first purpose and justifies it by the second. We test the first on the realised likelihood gain $G(i{\to}j) := \log p_\theta(x_j^\star \mid c \oplus (i{:}x_i^\star)) - \log p_\theta(x_j^\star \mid c)$ of revealing the true token at i. G is the log-ratio whose model expectation is D, at the data's tokens, so this is not independent evidence; it tests which score's choice forgoes less on real text (Appendix M).

Between $\hat{D}$ and $\hat{I}_{\text{cor}}$, $\hat{D}$ forgoes less. For each probed anchor row we choose the third of its targets that a score calls safest to unmask alongside the anchor and charge the $|G|$ it thereby forgoes. $\hat{D}$ forgoes less than $\hat{I}_{\text{cor}}$ on all three models, by $+0.0045$ to $+0.0065$ nats (6–13% of what $\hat{D}$ forgoes; $P(\hat{I}_{\text{cor}}$ better$) \leq 0.001$ under either resampling), and by more on pairs above 10^{-3} nats; I_{ent} forgoes 0.10–0.12 nats more than $\hat{D}$, while the signed gain does not separate the two (Table 9, Appendix M). That is the direction Proposition 2(a) predicts for a pair, for a reason the score's justification omits; at the set level the two are not resolved.

The score is not a decoder on its own. We decode 120 GSM8K problems (Cobbe et al., 2021) with LLaDA-8B-Instruct under its semi-autoregressive block decoding (Nie et al., 2025), greedily within a step, sweeping each rule's knob and counting every forward pass including probes (Figure 2, Appendix N). Against seven baselines (random, confidence, entropy, a confidence-gated dilated schedule after Luxembourg et al. (2026), Fast-dLLM (Wu et al., 2026), EB-Sampler (Ben-Hamu et al., 2025) and PUNT (Azangulov et al., 2026)), a rule that commits the most confident position and fills the rest of the step with the lowest-$\hat{D}$ positions, read from a two-point probe of the anchor's top two tokens, in T steps per 32-token block at three passes a step, does not merely lose, it collapses: 0.083, 0.100 and 0.425 at 48, 96 and 192 passes ($T = 2, 4, 8$), against 0.608 for Fast-dLLM at 34 and 0.825 for EB-Sampler at 93; its paired differences from the best baseline costing no more are -0.400 to -0.725 (-0.400 to -0.675 if cheaper on every problem). Neither reading of the score anticipates this, since neither sees the target's own uncertainty: by Proposition 2(a), $D(i{\to}j)$ is the cost of committing j *alongside* i, not evidence the model knows what to put there. The collapse is a property of these cheapest instantiations, not evidence that no defect-aware decoder can work (Appendix L). On the probe's pairs the score and the target's own entropy are only partly rank-correlated, and the third the score calls safest carries about twice the entropy of the third confidence picks (Appendix N): a position the anchor barely moves is admitted however uncertain it is on its own. At the same

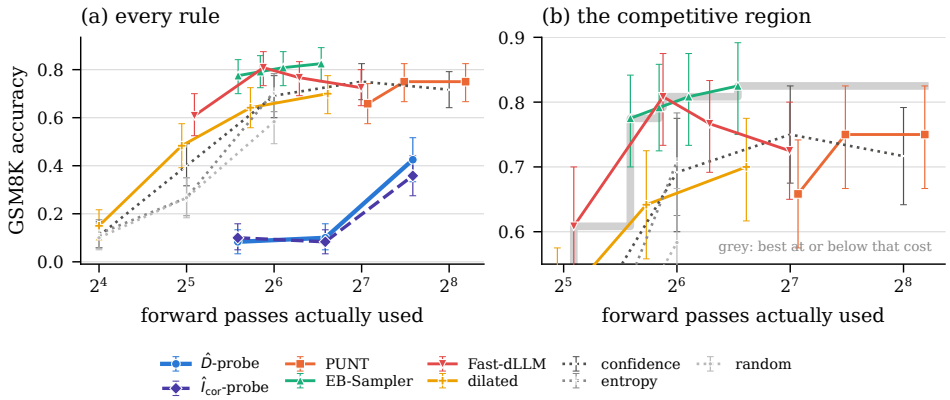

Figure 2: GSM8K under semi-autoregressive block decoding with LLaDA-8B-Instruct, 120 problems, generation length 256, block 32. Every gate threshold is swept, the fixed-step rules at $T \leq 8$ and confidence to $T = 32$, and the horizontal axis counts *every* forward pass a rule issues, probes included. (a) all nine rules; the two that fill a step by the score after a confidence-chosen anchor collapse. (b) the region where the published samplers compete, with the envelope of the best accuracy any rule reaches at or below each cost.

T, where only the choice of positions differs, the rule trails confidence decoding by 0.27–0.30 at $T = 4$ and 8 and falls below random order there (Holm-adjusted $P \leq 0.05$; $T = 2$ is unresolved). The same sweep on the first 80 problems with Dream-7B-Instruct replicates the collapse against the samplers (0.125–0.287 against EB-Sampler's 0.762 and 0.838 at 39.4 and 56.5 passes, differences negative by 0.550–0.650, and by 0.113–0.575 strictly), and 11 of the 12 strict comparisons (both rules and models) give $P < 0.05$. On HumanEval (Chen et al., 2021) LLaDA-8B-Instruct repeats the pattern against the samplers though not against random order (0.049–0.110 pass@1 against 0.220 for Fast-dLLM at 37.8 passes and 0.421 and 0.427 for EB-Sampler at 93.9 and 125), and all six strict comparisons (both rules) give $P < 0.05$; so does Dream-7B on 60 problems, where four of six do (Appendix N).

What the frontier says about the samplers' own guarantees. The rules that work all fill the step by confidence or entropy. On LLaDA-8B-Instruct EB-Sampler (0.775 at 48.1 passes to 0.825 at 92.9) and Fast-dLLM form the envelope and beat one-token-per-step confidence decoding at 23–36% of its compute, which is about the confidence *order* rather than parallelism; on Dream-7B they only match it (Appendix N). EB-Sampler admits a set U while $\sum_{j \in U} h_j - \max_{j \in U} h_j \leq \gamma$ with $h_j = H(p_j)$, which bounds the total correlation of a coherent joint with those marginals. On the deployed model the step's cost is $\mathrm{TC}(q_\pi)$ *plus* the defect sum (Proposition 1), and γ, like the schedule budgets of Kang et al. (2026) and Chen et al. (2026a), is written against the first term only; for $k \geq 3$, or $k = 2$ with the higher-entropy position first, it does not provably bound even that (Appendix D). Rather than transfer Table 8's 15–21% from probe text, we measure all three terms on the sets these gates admit inside the samplers' own decodes of the first 20 problems, in confidence order and on each set's first 8 positions (Appendix D). At $\gamma = 0.4$ and 128 rollouts the defect is 0.568 $[0.474, 0.644]$ of the parallel error, between the split-sample 0.386 and the plug-in 0.597; at 32 rollouts, which read 0.03 higher there, it is 0.347 $[0.302, 0.394]$ at $\gamma = 1.6$ and 0.358 $[0.306, 0.417]$ for Fast-dLLM at $\tau = 0.8$; on Dream-7B it is 0.353 $[0.318, 0.396]$ at $\gamma = 0.4$ and 128 rollouts, with a split-sample 0.041 unresolved from zero, and 0.293 and 0.308 at the other settings at 32, and there the allowance covers the error in aggregate (Appendix D). On LLaDA-8B-Instruct, summed along the decode, the realised error is 1.14 $[0.90, 1.40]$ of the allowance EB-Sampler charges itself at $\gamma = 0.4$ and 0.70 at $\gamma = 1.6$; the dependence it is written against takes only 0.49 and 0.46 of it, and the defect 0.65 and 0.24. At $\gamma = 0.4$ the error overruns the allowance on 10 of the 20 decodes, by 1.06–2.57×, and on every one of them the measured dependence alone stays within it (at most 0.65×, or 0.94× with the conservative defect): on the measured prefixes, the overrun comes from the defect.

The share a decoder can read for itself. The pairwise share can also be read during a decode, since an anchor probe leaves the decode untouched (360 generations under six settings reproduce

exactly). Per generation, the stratified jackknifed estimator of Appendix F puts the defect summed over the positions the rule commits at a median 0.264 to 0.339 of the summed score across settings on LLaDA-8B-Instruct and 0.240 to 0.316 on Dream-7B, and 0.315 and 0.269 pooled into one ratio; being ratios of sums, these compare with the probe's jackknifed mass-weighted 0.327 $[0.305, 0.351]$ on probe text (Appendix K), not with its higher median. On Dream-7B the 14% of committed pairs scoring above their anchor's entropy have a median share of 0.86, as Proposition 3(C) requires; the rest, 86% of the summed score, 0.22. The two-point probe a sampler could afford reads $1.71\times$ larger, partly from truncating the expectation over $v \sim p_i$ (Appendix L). It buys no detectable gain, though: the accumulated score predicts failure only through confidence, and as a filter on Fast-dLLM or EB-Sampler it beats the best baseline costing no more at none of ten operating points and a value-permuting placebo at none of five, with gains up to about 0.13 not excluded (Appendix L).

What changes for the literature. The dependence that Ben-Hamu et al. (2025) describe as "precisely joint mutual information" is one for a coherent model; for deployed ones it is the rollout's total correlation plus a defect, a jackknifed 15–21% of the parallel error on probed sets and 0.29–0.57 on the sets EB-Sampler and Fast-dLLM admit. The entropy-difference target of Sharma et al. (2026) falls below -0.05 nats on 14.5% and 19.6% of pairs on MDLM-OWT and LLaDA-8B-Instruct, where replicate noise gives 0.0% and 1.5%, so an estimator fitted to it learns something that is not an information. The temporal-stability score of Wu & Huang (2026), an expected KL of the form of D that their Lemma A.1 equates with a conditional mutual information, carries the defect too. Schedule-level analyses budgeting against a total correlation (Kang et al., 2026; Chen et al., 2026a; Zhang et al., 2026) are exact for a deployed model only with the defect sum added back, and the error they budget moves with the reveal order. None of this says the samplers are wrong; what changes is what the score measures.

6 Related work

That the parallel-decoding cost equals a conditional total correlation was stated for masked diffusion (Liu et al., 2025; Ben-Hamu et al., 2025; Kang et al., 2026; Chen et al., 2026a; Zhang et al., 2026), for the data distribution, an ideal denoiser or a coherent model, and training-free samplers bound or test it (Wu et al., 2026; Luxembourg et al., 2026; Azangulov et al., 2026; Wu & Huang, 2026). That conditionals need not belong to any joint is classical (Besag, 1974; Arnold & Press, 1989) and documented for BERT-style models (Goyal et al., 2022; Yamakoshi et al., 2022; Young et al., 2024); any-order models address it at training time (Shih et al., 2022), and marginalization self-consistency, whose violation Δ measures, is a training target in Liu et al. (2024). Kim (2026) carried the framing to dLLMs with a two-order curl diagnostic and no measurements. We propose no sampler (Appendix C).

7 Limitations and conclusion

The certificates are one-sided, (B) and (C) miss incoherence that moves no marginal (Appendix B), and the contiguous-block bound certifies only short blocks. On the bf16 models 43–62% of pairs sit within $10\times$ the numerical floor, so we make no per-pair claims. Contexts are OpenWebText (MDLM-OWT's training text) at $L = 128$; anchors are filtered for peakedness. The downstream evidence is two tasks, decoded greedily though the identity describes sampling, with in-decode measurements on two models; its nulls leave gains up to 0.13 unexcluded. Appendix Q lists the rest.

The dependence score behind training-free parallel decoders is not a mutual information of the models that compute it: Proposition 2 splits it into the rollout's own dependence and a self-marginalization defect, a median 0.37–0.45 of the pairwise score on peaked probed anchors and 0.25–0.46 of its mass by a conservative estimate, and Proposition 1 splits a set's error the same way, the defect being a jackknifed 15–21% of it on probed sets and 0.29–0.57 on the sets deployed gates admit, where the conservative estimator resolves it only on LLaDA-8B-Instruct (0.39). Between two candidates the score is still the right comparison in principle, since Proposition 2(a) makes it the step's error relative to revealing the anchor first, not because it measures dependence. Over a set it is neither, and a budget written against the dependence alone leaves the defect sum out, and the documents' own tokens, on average, place that defect in the one-pass marginal. On an enumerable toy a 600-fold better fit shrank the defect and the score together but not the share (Appendix I); whether an explicit training penalty would (Liu et al., 2024) is left open.

REPRODUCIBILITY STATEMENT

Propositions 1–4 are proved in Appendix A, and the counterexample separating zero defect from coherence is in Appendix B. Appendix F specifies the estimator, the stratification and jackknife, the split-sample lower bound and its derivation, the float64 accumulation, the numerical-floor calibration and the model-conditioning check; Appendices E and G report the placebo, budget and replicate-noise controls and the document-token comparison in full, §4.1 with Appendices H and K the rollout experiments, Appendices I and J the enumerable models, and Appendices D, L, M, N and O the decoding and selection experiments. The supplementary material contains the complete code, the 61,090 records, the analysis scripts that produce every table and figure from those records, and a unit-test suite for the estimators. In the supplementary archive, `code/reproduce.sh` followed by `code/figs_paper.py` recomputes every table, figure and number in the paper on a CPU from those records alone, with no network access and no credentials; its toy-model stage retrains five small transformers and takes hours. Three inputs come from reloading a model rather than from a record, and their outputs are shipped: the high-budget re-measurement of Appendix G, released as per-pair estimates; the float32-against-float64 accumulation errors of Appendix F, which need a GPU; and the slot-alignment check of Appendix F, which needs the model checkpoints.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing, editing, and debugging code, including code used to implement our methods and experiments. We did not use generative AI tools for any of the other tasks that require disclosure: in particular, we did not use them to generate synthetic data, to develop theoretical models or conceptual frameworks, to formulate mathematical claims or assist in proofs, to propose or refine hypotheses, to design the research methodology or experiments, to assist with translation, to clean or reformat datasets, or to support data analysis or the interpretation of results. Additionally, we used generative AI tools to edit the manuscript to improve readability and clarity, and to obtain simulated peer-review-style feedback on drafts, which informed revisions to the writing and presentation. All AI-assisted code was reviewed and tested for correctness by the authors, and all AI-suggested edits and feedback were reviewed and adopted at the authors' discretion; the scientific content, claims, and results are entirely the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Shun-ichi Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001.

Barry C. Arnold and D. V. Gokhale. Distributions most nearly compatible with given families of conditional distributions. *TEST*, 7(2):377–390, 1998.

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

Marianne Arriola, Aaron Gokaslan, Justin T. Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In *International Conference on Learning Representations (ICLR)*, 2025.

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 (ICLR)*, 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 (NeurIPS)*, 2025.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, 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 *Conference on Learning Theory (COLT)*, 2026a.

Zigeng Chen, Gongfan Fang, Xinyin Ma, Ruonan Yu, and Xinchao Wang. dParallel: Learnable parallel decoding for dLLMs. In *International Conference on Learning Representations (ICLR)*, 2026b.

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.

William Cook. A hierarchical decomposition of Kullback–Leibler divergence: Disentangling marginal mismatches from statistical dependencies. *arXiv preprint arXiv:2504.09029*, 2025.

Guhao Feng, Yihan Geng, Jian Guan, Wei Wu, Liwei Wang, and Di He. Theoretical benefit and limitation of diffusion language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-Predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6112–6121, 2019.

Aaron Gokaslan and Vanya Cohen. OpenWebText corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. Exposing the implicit energy networks behind masked language models via Metropolis–Hastings. In *International Conference on Learning Representations (ICLR)*, 2022.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. Non-autoregressive neural machine translation. In *International Conference on Learning Representations (ICLR)*, 2018.

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.

Matthew D. Hoffman and Matthew J. Johnson. ELBO surgery: Yet another way to carve up the variational evidence lower bound. In *NIPS Workshop on Advances in Approximate Bayesian Inference*, 2016.

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 (ICLR)*, 2022.

Daniel Israel, Guy Van den Broeck, and Aditya Grover. Accelerating diffusion LLMs via adaptive parallel decoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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 (ICLR)*, 2026.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. In *International Conference on Machine Learning (ICML)*, 2025.

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

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

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

Yuchen Li, Alexandre Kirchmeyer, Aashay Mehta, Yilong Qin, Boris Dadachev, Kishore Papineni, Sanjiv Kumar, and Andrej Risteski. Promises and pitfalls of generative masked language modeling: Theoretical framework and practical guidelines. In *International Conference on Machine Learning (ICML)*, 2024.

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

Sulin Liu, Peter J. Ramadge, and Ryan P. Adams. Generative marginalization models. In *International Conference on Machine Learning (ICML)*, 2024.

Omer Luxembourg, Haim Permuter, and Eliya Nachmani. Plan for speed: Dilated scheduling for masked diffusion language models. In *International Conference on Machine Learning (ICML)*, 2026.

Amr Mohamed, Guokan Shang, and Michalis Vazirgiannis. LESS is more: Mutual-stability sampling for diffusion language models. *arXiv preprint arXiv:2606.16908*, 2026.

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 (NeurIPS)*, 2025.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *International Conference on Learning Representations (ICLR)*, 2025.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning (ICML)*, 2019.

Yuji Ren, Chenkai Xu, Zhuocheng Gong, Jianguo Li, and Zhijie Deng. CForce: Boosting parallel decoding for dLLMs via consistency forcing. *arXiv preprint arXiv:2608.13925*, 2026.

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

Yusuf Sahin, Ahmed Rockey Saikia, Volkan Cevher, and Paolo Favaro. Attention-discounted adaptive sampler for masked diffusion language models. *arXiv preprint arXiv:2606.10829*, 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 (NeurIPS)*, 2024.

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

Andy Shih, Dorsa Sadigh, and Stefano Ermon. Training and inference on any-order autoregressive models the right way. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Flemming Topsøe. Basic concepts, identities and inequalities — the toolkit of information theory. *Entropy*, 3(3):162–190, 2001.

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 (Short Papers)*, 2023.

Benigno Uria, Iain Murray, and Hugo Larochelle. A deep and tractable density estimator. In *International Conference on Machine Learning (ICML)*, 2014.

Eitan Wagner, Yuli Slavutsky, and Omri Abend. ConTestS: A framework for consistency testing of span probabilities in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Alex Wang and Kyunghyun Cho. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 30–36, 2019.

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 (ICLR)*, 2026.

Yue Wu and Jian Huang. Stability-weighted decoding for diffusion language models. *arXiv preprint arXiv:2604.17068*, 2026.

Minkai Xu, Tomas Geffner, Karsten Kreis, Weili Nie, Yilun Xu, Jure Leskovec, Stefano Ermon, and Arash Vahdat. Energy-based diffusion language models for text generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Takateru Yamakoshi, Thomas L. Griffiths, and Robert D. Hawkins. Probing BERT's priors with serial reproduction chains. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022.

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.

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.

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. In *International Conference on Learning Representations (ICLR)*, 2025.

A PROOFS

Proposition 1. Write $\tilde{m} := \prod_{i \in S} q_\pi^{(i)}$. Then

$$\mathrm{KL}(q_\pi \| m_S) = \mathbb{E}_{q_\pi}\Big[\log \frac{q_\pi}{\tilde{m}}\Big] + \mathbb{E}_{q_\pi}\Big[\log \frac{\tilde{m}}{m_S}\Big] = \mathrm{TC}(q_\pi) + \sum_{i \in S} \mathbb{E}_{q_\pi}\Big[\log \frac{q_\pi^{(i)}(x_i)}{p_i(x_i)}\Big].$$

Each summand depends on x_i alone, so its expectation under q_π equals its expectation under the marginal $q_\pi^{(i)}$, which is $\mathrm{KL}(q_\pi^{(i)} \| p_i) = \Delta_\pi(i)$. Non-negativity of the three terms is that of KL. □

Proposition 2. (a) $q_{(i,j)}(v,w) = p_i(v)\, p_\theta(w \mid c \oplus (i{:}v))$ and $m_S(v,w) = p_i(v)p_j(w)$; the p_i factors cancel inside the logarithm and $\mathrm{KL}(q_{(i,j)} \| m_S) = \mathbb{E}_v[\mathrm{KL}(p_\theta(\cdot \mid c \oplus (i{:}v)) \| p_j)] = D(i{\to}j)$. (b) For any family P_v with mixture $\bar{P} = \mathbb{E}_v P_v$ and any R, $\mathbb{E}_v[\mathrm{KL}(P_v \| R)] = \mathbb{E}_v[\mathrm{KL}(P_v \| \bar{P})] + \mathrm{KL}(\bar{P} \| R)$, since $\mathbb{E}_v\langle P_v, \log \bar{P} - \log R\rangle = \langle \bar{P}, \log \bar{P} - \log R\rangle$. Take $P_v = p_\theta(\cdot \mid c \oplus (i{:}v))$, $R = p_j$; the first term is $I_q(X_i; X_j)$ because $\bar{P} = \bar{q}_j$ is the pseudo-joint's marginal on j and P_v its conditional. The trailing bound $0 \le \Delta(j \mid i) \le D(i{\to}j)$ of the statement is then immediate from (b) and $I_q \ge 0$. □

Proposition 3. (A) Under coherence $q_{(i,j)} = q_{(j,i)} = P$, so both scores equal $\mathrm{KL}(P \| P_i \otimes P_j) = I_P$ and $A = 0$, and likewise for any expected f-divergence, since $\mathbb{E}_{v \sim P_i} D_f(P(\cdot \mid v) \| P_j) = D_f(P \| P_i \otimes P_j)$. (B) Under P, $I_{\mathrm{ent}} = H(X_j) - H(X_j \mid X_i) = I_P \ge 0$. The general identity follows by writing $I_q = \mathbb{E}_v[\mathrm{KL}(P_v \| \bar{q}_j)] = H(\bar{q}_j) - \mathbb{E}_v H(P_v)$ and $I_{\mathrm{ent}} = H(p_j) - \mathbb{E}_v H(P_v)$; subtracting gives $I_{\mathrm{ent}} - I_q = H(p_j) - H(\bar{q}_j)$. (C) $q_{(i,j)}(v,w) = p_i(v)p_\theta(w \mid c \oplus (i{:}v))$, so $q_{(i,j)}$'s marginal on i is exactly p_i and $I_q(X_i; X_j) = H_q(X_i) - H_q(X_i \mid X_j) \le H(p_i)$. Dividing $\Delta = D - I_q \ge D - H(p_i)$ by $D > 0$ gives (5). The bound tends to 1 along any sequence with $H(p_i)/D \to 0$ ($H(p_i) = 0$ itself is excluded by strict positivity), and $\rho \le 1$ always. □

Proposition 4. $\hat{D} - \hat{\Delta} = \frac{1}{m}\sum_s \langle P_s, \log P_s - \log p_j\rangle - \langle \bar{P}, \log \bar{P} - \log p_j\rangle = \frac{1}{m}\sum_s \langle P_s, \log P_s - \log \bar{P}\rangle$, using $\frac{1}{m}\sum_s P_s = \bar{P}$; each term is a KL and the sum is non-negative. The same holds for any non-negative weights summing to one, including the stratified weights of Appendix F. The jackknifed $\hat{I}_{\mathrm{cor}}$ is non-negative too: the jackknife leaves out one tail draw at a time with the head fixed, the leave-one-out mixtures average to $\bar{P}$, so by convexity of the KL in its first argument their mean defect is at least $\hat{\Delta}$, the jackknifed $\hat{\Delta}$ is at most the plug-in, and $\hat{D}$, being linear in the draws, is its own jackknife. The clip at zero therefore guards only against rounding, and binds on 8 of 77,054 pairs (Appendix F). □

B ZERO DEFECT DOES NOT IMPLY COHERENCE

Take $V = \{0, 1\}$, rows indexing x_i and columns x_j, and two joints $J = \left(\begin{smallmatrix} 0.38 & 0.02 \\ 0.32 & 0.28 \end{smallmatrix}\right)$, $J' = \left(\begin{smallmatrix} 0.23 & 0.17 \\ 0.47 & 0.13 \end{smallmatrix}\right)$, both with marginals $p_i = (0.4, 0.6)$ and $p_j = (0.7, 0.3)$. Define a model by $p_\theta(x_j \mid c \oplus (i{:}v)) = J[v, \cdot]/p_i(v)$, $p_\theta(x_i \mid c \oplus (j{:}w)) = J'[\cdot, w]/p_j(w)$ and direct marginals p_i, p_j. Then the induced marginal of j under (i, j) is p_j and that of i under (j, i) is p_i, so both defects are zero, yet $q_{(i,j)} = J \ne J' = q_{(j,i)}$ and $\mathrm{TC}(J) = 0.117$ nats against $\mathrm{TC}(J') = 0.025$: the model is incoherent and its total correlation is order-dependent. Coherence implies zero defect; the converse is false, and zero defect is exactly the condition under which (2) reduces to "error = total correlation", of an order-dependent rollout. Of the three certificates only (A) fires here: $D(i{\to}j) = \mathrm{TC}(J)$ and $D(j{\to}i) = \mathrm{TC}(J')$ differ, while $I_{\mathrm{ent}} = I_q \ge 0$ in both directions and (5) is vacuous. (B) and (C) certify $\Delta > 0$ itself, since $I_{\mathrm{ent}} < 0$ forces $H(\bar{q}_j) > H(p_j)$ and hence $\bar{q}_j \ne p_j$; (A) certifies incoherence whether or not it moves a marginal. None of the three is complete in its KL form: replacing J' by $J'' = \left(\begin{smallmatrix} a & 0.4-a \\ 0.7-a & a-0.1 \end{smallmatrix}\right)$ with $a = 0.172211$ (the same marginals, a chosen so that $\mathrm{TC}(J'') = \mathrm{TC}(J)$) gives a model with $q_{(i,j)} \ne q_{(j,i)}$ on which the KL asymmetry is zero, $I_{\mathrm{ent}} = I_q \ge 0$ in both directions and (5) is vacuous. The total-variation form of (A) does fire on it (0.200 against 0.216), as does the χ^2 form, so running (A) with more than one divergence catches cases the KL form misses.

C Related work in full

Parallel decoding and its error. Committing several positions from their own marginals, with the dependence among them as the cost, is the multimodality problem of non-autoregressive translation (Gu et al., 2018); confidence-ranked iterative unmasking goes back to Mask-Predict and MaskGIT (Ghazvininejad et al., 2019; Chang et al., 2022), and the semi-autoregressive block schedule used with LLaDA (Nie et al., 2025) is trained into the model by block diffusion (Arriola et al., 2025). That the cost equals a conditional total correlation was stated for masked diffusion (Liu et al., 2025; Ben-Hamu et al., 2025; Kang et al., 2026; Chen et al., 2026a; Zhang et al., 2026; Lavenant & Zanella, 2025), for the data distribution, an ideal denoiser or a coherent model. Feng et al. (2025) analyse how many parallel steps a masked diffusion model needs, Li & Cai (2025) bound the sampling error of a diffusion language model by mutual-information terms of the data distribution plus the mask predictor's training error, a term a positive defect implies is positive, since exact conditionals are coherent, and Kim et al. (2025) study how the unmasking order interacts with training and inference-time planning. Samplers commit by confidence (Wu et al., 2026), by entropy sums (Ben-Hamu et al., 2025), by spreading co-committed positions apart (Luxembourg et al., 2026), by KL-based independence tests (Azangulov et al., 2026), by stability across steps (Wu & Huang, 2026; Mohamed et al., 2026), by attention (Sahin et al., 2026) or by verifying the block with a small autoregressive model (Israel et al., 2025); others learn a dependence predictor (Ringel et al., 2026) or a neural estimator of pairwise mutual information (Sharma et al., 2026), and training-based variants distill certainty (Chen et al., 2026b) or consistency (Ren et al., 2026) into the model. A sequence-level energy model can also correct the product-of-marginals step directly (Xu et al., 2025). We propose no sampler (the probe rules of §5 are diagnostics, not proposals) and ask what the quantity these methods share measures.

The algebra and its estimation. Proposition 1 is the Pythagorean relation of the reverse I-projection onto product distributions (Amari, 2001), used to separate marginal mismatch from dependence by Cook (2025); Proposition 2(b) is the compensation identity (Topsøe, 2001), the same split that carves the ELBO's average KL (Hoffman & Johnson, 2016). In the language of variational bounds on mutual information, D is the upper bound $\mathbb{E}_v \mathrm{KL}(P_v \| r) \geq I_q$ with the direct marginal as the variational marginal r, and Δ is the bound's gap (Poole et al., 2019). What is new is the reference measure: both sides are the deployed model's own quantities, and the gap is a property of the model rather than of a chosen r. The plug-in estimator's $O(1/m)$ bias and its first-order correction are classical (Paninski, 2003); the split-sample bound of §3 is what makes the direction of the remaining bias certain, for that estimator; the jackknife has no guaranteed direction.

Incoherent conditionals. That conditionals need not be compatible with any joint is classical (Besag, 1974; Arnold & Press, 1989; Arnold & Gokhale, 1998; Heckerman et al., 2000). For masked language models, sampling from the conditionals as if they defined a Markov random field began with Wang & Cho (2019); the incompatibility was exposed by running a Metropolis–Hastings sampler on the model's own energy (Goyal et al., 2022), connected to order-dependent pseudo-Gibbs stationary distributions (Yamakoshi et al., 2022), measured pairwise (Torroba Hennigen & Kim, 2023; Young et al., 2024), tested across model sizes (Wagner et al., 2024), and flagged as a generation hazard (Li et al., 2024). In the any-order autoregressive line, order-ensembling averages over the order at test time (Uria et al., 2014); Shih et al. (2022), observing that such models "define the same distribution in multiple different ways", restrict the conditionals trained on; and autoregressive diffusion models (Hoogeboom et al., 2022), like the absorbing-state models measured here, train on all orders without enforcing coherence. Marginalization self-consistency (a model's marginals agreeing with the marginals its own conditionals imply, the property whose violation Δ measures) is a training target in Liu et al. (2024). Kim (2026) carried the framing to dLLMs: it names the order-induced pseudo-joint, relates it to a Bayes-optimality condition, and proposes a curl diagnostic that needs two orders per pair, without measurements; our defect needs one order and sits inside an exact identity with the score itself.

D What a deployed gate spends

Method. We run our reimplementations of EB-Sampler's and Fast-dLLM's decodes (Appendix N) on the first 20 GSM8K problems with LLaDA-8B-Instruct and, at every step, measure the three

terms of (2) on the set U that gate admitted, with the k-sweep's rollout estimator (§4.1, Appendix F). Selection reads only the forward pass and the argmax and the rollouts write into copies, so the measurement cannot steer the decode. Against the unmeasured sweep of Appendix N, the 60 LLaDA-8B-Instruct decodes measured at 32 rollouts are identical in answer and pass count, and Dream-7B's $\gamma = 0.4$ decodes measured at 32 and at 128 rollouts are identical to each other on all 20 problems. Runs made in other sessions, LLaDA-8B-Instruct's at 128 rollouts and Dream-7B's against the earlier unmeasured sweep, keep the answer on 19 of 20 and 45 of 60 problems but the pass count on only 10 and 14, which we attribute to bf16 differences between sessions that a greedy, threshold-gated decode amplifies: the uninstrumented sweep script itself, re-run in a later session on the first four problems, reproduces the text of only 19 and 30 of its 64 decodes on the two models (`code/check_session_repro.py`). The measurements describe the decodes as they ran. The reference order π is confidence-descending, the order a sequential decoder would use and close to the entropy order in which EB-Sampler accumulates its own allowance; §4.1 is why that choice has to be named, and every number below is for it. A step that commits $|U|$ positions costs $|U|$ times the rollout count to measure, so sets above 8 positions are measured on their first 8 in that order. Since a set's parallel error is, by the chain rule, a sum of non-negative per-position terms along the order, a capped reading is in expectation a *lower* bound on the set's; the cap binds on 11%, 13% and 26% of measured steps at $\gamma = 0.4$, $\gamma = 1.6$ and $\tau = 0.8$. $\mathrm{TC}(q_\pi)$ is the residual of the other two terms (Appendix F), so (2) holds in every generation by construction and the dependence inherits the defect estimator's bias. Intervals resample problems.

What the three terms are. At $\gamma = 0.4$ and 128 rollouts, summed along each decode, the parallel error has median 9.0 $[8.1, 10.8]$ nats per generation, of which the dependence is 4.4 and the defect 4.9; at 32 rollouts it is 37.3 at $\gamma = 1.6$ and 18.0 at $\tau = 0.8$, of which the dependence is 23.8 and 11.6 and the defect 11.4 and 6.3. These are medians over generations, and medians of three quantities do not add, which is why the shares below are ratios of sums. They are sums along one greedy path, not divergences between distributions over generations. The gate's accumulated allowance has median 9.3 and 51.6 nats, and as a ratio of sums the realised error is 1.14 $[0.90, 1.40]$ of it at the tight setting and 0.70 $[0.66, 0.74]$ at the loose one: the allowance roughly covers the error at $\gamma = 0.4$ and covers it at $\gamma = 1.6$. What changes is the composition: the dependence alone, the term the allowance is written against, is 0.49 $[0.45, 0.53]$ and 0.46 $[0.43, 0.48]$ of the allowance, and the defect 0.65 of it at $\gamma = 0.4$. Generation by generation, the error overruns the allowance on 10 of the 20 decodes at $\gamma = 0.4$, by 1.06–2.57×, while the dependence alone stays within it on all ten, at most 0.65× with the jackknifed defect and 0.94× with the split-sample one, which leaves the most to TC; the unmeasured tails of capped sets would add to both terms, so this is a statement about the measured prefixes. Across all 20 decodes the dependence alone stays within the allowance, at most 0.67× (0.97× with the split-sample defect), while the defect alone exceeds it on 3. Measured at 32 rollouts, the same decodes overrun on 8 of 20, with the dependence within the allowance on all eight by the jackknife but on only four by the split-sample defect, whose penalty is larger at that budget. The allowance is summed over the whole admitted set and the error over its measured prefix, so on capped steps the realised error is understated relative to the allowance. Fast-dLLM declares no budget; its threshold rule is justified by a result for a single joint, that marginals above $1 - \epsilon$ with $(n + 1)\epsilon \leq 1$ make the product's argmax the joint's (Wu et al., 2026, Theorem 1), and on an incoherent model the order-π rollouts need not share one argmax, so its premise fails in the same way. As ratios of sums the defect is 0.568 $[0.474, 0.644]$ of the parallel error at $\gamma = 0.4$ (128 rollouts) and 0.347 $[0.302, 0.394]$ and 0.358 $[0.306, 0.417]$ at $\gamma = 1.6$ and $\tau = 0.8$ (32 rollouts), the figures §5 quotes. The cap truncates numerator and denominator alike, so these shares are estimates on the first 8 positions, not bounds; and the entropy allowance is not a proven bound on $\mathrm{TC}(q_\pi)$ once $k \geq 3$, nor at $k = 2$ when the higher-entropy position is revealed first, because it uses the direct entropies $H(p_j)$ while $\mathrm{TC}(q_\pi)$ is controlled by the induced ones, which Proposition 3(B) shows can be larger.

On Dream-7B. The same three settings on the same 20 problems with Dream-7B-Instruct, at 32 rollouts, put the defect at 0.423 $[0.387, 0.465]$, 0.293 $[0.264, 0.320]$ and 0.308 $[0.273, 0.341]$ of the parallel error at $\gamma = 0.4$, $\gamma = 1.6$ and $\tau = 0.8$, with plug-in readings 0.495, 0.355 and 0.390. The split-sample readings are negative at this budget, −0.173, −0.109 and −0.390, as on LLaDA-8B-Instruct at $\gamma = 1.6$ and $\tau = 0.8$, so at 32 rollouts the conservative estimator certifies nothing on these sets. Re-measured at 128 rollouts on all 20 $\gamma = 0.4$ problems, the jackknifed share falls from 0.423 to 0.353 $[0.318, 0.396]$ and the plug-in from 0.495 to 0.397, while the split-sample reading rises from −0.173 to 0.041 $[-0.021, +0.115]$: the bracket narrows as on LLaDA-8B-Instruct, but on

Dream-7B's in-decode sets the conservative estimator does not resolve the defect from zero even at 128 rollouts, and the share there rests on the jackknife. The 32-rollout jackknifed readings at $\gamma = 1.6$ and $\tau = 0.8$ are therefore likely to sit several hundredths high. Here the allowance covers the error: at 32 rollouts the realised error is 0.73 $[0.63, 0.82]$ and 0.67 $[0.63, 0.72]$ of it at $\gamma = 0.4$ and 1.6, the dependence 0.42 and 0.48 of it and the defect 0.31 and 0.20, and the error overruns the allowance on 3 of the 20 decodes at $\gamma = 0.4$; at 128 rollouts the ratio is 0.70 $[0.65, 0.75]$, with the dependence 0.45 and the defect 0.25 of the allowance, and one decode overruns it. The cap binds on 14%, 15% and 28% of measured steps.

The rollout count, and why the conservative estimator needs it. These sets lie within one 32-token block, where rollouts are more dispersed than on scattered sets and the split-sample penalty falls only about as $n^{-0.55}$ (Appendix H), so at 32 rollouts that estimator is underpowered here. Re-measuring all 20 of the $\gamma = 0.4$ problems at 128 rollouts shows it is a power problem and not a small defect: the split-sample share rises from 0.129 $[-0.114, +0.309]$ to 0.386 $[+0.240, +0.495]$, excluding zero, while the jackknife falls from 0.595 to 0.568 and the plug-in, which is biased the other way, from 0.643 to 0.597. The bracket narrows from $[0.13, 0.64]$ to $[0.39, 0.60]$ from both ends, as a shrinking penalty requires; per generation, the jackknifed share falls on 16 of the 20 problems and the split-sample share rises on 17. The $\gamma = 1.6$ and $\tau = 0.8$ shares were not re-measured, and at 32 rollouts their jackknifed readings are likely to sit a few hundredths high in the same way. §5 quotes the jackknifed reading, the estimator of column 5 of Table 8, and the 128-rollout split-sample figure is the conservative one.

E THE CONTROLS IN FULL, AND WHICH MARGINAL THE TEXT FAVOURS

The mask-count control. Filling i removes one mask, and an absorbing-state denoiser is in theory time-independent (Ou et al., 2025; Zheng et al., 2025) but a trained one need not be. Under the count-matched placebo, in which a far-away revealed position is re-masked so that the number of masks is unchanged, the defect does not vanish: summed over matched pairs it is $4.98\times$ $[3.54, 7.24]$ its main-condition value on MDLM-OWT (19,015 matched pairs in 40 documents), $2.95\times$ $[2.28, 3.80]$ on LLaDA-8B-Instruct (8,756 in 25) and $2.45\times$ $[1.27, 5.30]$ on Dream-7B (2,195 in 20), with intervals that resample documents and then anchor rows. Matching the count amplifies the defect rather than removing it. Re-masking also withholds a revealed token, which moves even a coherent model's conditionals, so this does not isolate the count: on the exactly coherent joint of Appendix I, whose defect is zero, the same placebo reads a positive value on 35% of contexts with signal and 20% of the score in aggregate (`code/toy_placebo_exact.py`). What it shows is that the count-matched condition does not make the defect go away. Re-masking the *closest* revealed position amplifies it further: on the 629 anchors of the MDLM-OWT placebo run (Appendix G), the median defect over pairs above 10^{-3} nats in both conditions is $12.7\times$ its main-condition value, against $5.8\times$ for the far placebo (a ratio of medians, where the $4.98\times$ above is a ratio of sums over all matched pairs). Both placebos are harder conditions than the main one, not neutral ones (these models are sensitive to which positions carry masks, the more so the closer) and neither supports the hypothesis, because the *share* rises too, from 0.345 to 0.700 and 0.823 on the matched MDLM-OWT anchors and from 0.395 to 0.682 on LLaDA-8B-Instruct. Every number we report is the least-perturbed condition.

Base against instruction-tuned weights. LLaDA-8B-Base and -Instruct share their pretrained initialization, so we probed both on the same documents, ratios, and (anchor, target) indices; an anchor enters only if it passes each model's own head-mass filter, which leaves 2,089 pairs present in both, and we compare the 1,722 of them with $\hat{D} > 10^{-4}$ nats in both. The typical pair does not move: the median paired difference in $\hat{\Delta}$ is -2×10^{-5} nats $[-1.0 \times 10^{-4}, +2.8 \times 10^{-5}]$ against a median defect of about 6×10^{-3}, with Instruct the larger on 48.8% of pairs; restricted to the 783 pairs whose $\hat{D}$ clears $10\times$ the floor in both, the median difference is $+7 \times 10^{-5}$ $[-5 \times 10^{-4}, +7 \times 10^{-4}]$ and Instruct is the larger on 50.4%. The upper tail does move: the summed defect is $1.18\times$ $[1.05, 1.32]$ its Base value while the *share* falls slightly, by 0.005 $[0.001, 0.011]$ in ρ; at a $1.5\times$ larger probe budget on 2,412 cells it is $1.27\times$ $[1.13, 1.41]$, with the median cell again unmoved and Instruct the larger on 49.8% of cells. Instruction tuning therefore does not create the defect, though it raises its aggregate by 18–27% through the upper tail, at least in this model family; that it is laid down in pretraining is the natural reading but not one these weights can test. Whether it is specific to the masked-diffusion

objective or the ordinary incoherence of any imperfect fit is separate; on the enumerable model of Appendix I ρ does not fall as the conditionals converge, over the range tested. The two models also agree on how often I_{ent} is negative but disagree on *which* pairs 41.6% of the time, against about 50% for independent signs, so the negative pairs are largely different between the two.

Which marginal the text favours. The defect $\mathrm{KL}(\bar{q}_j \| p_j)$ records that two of the model's marginals for one position in one context disagree; it does not say which is wrong. Each context, however, is cut from a document with a realised token x_j^* at j, and $\log \bar{q}_j(x_j^*) - \log p_j(x_j^*)$ compares the two with a proper score: its expectation over documents is $\mathrm{KL}(P_j \| p_j) - \mathrm{KL}(P_j \| \bar{q}_j)$ for the data's marginal P_j given the context, which is $+\Delta(j \mid i)$ exactly if $\bar{q}_j = P_j$ and $-\mathrm{KL}(p_j \| \bar{q}_j) < 0$ if $p_j = P_j$. We reran the probe of §3 unchanged on 20 OpenWebText documents per model that no other probe run uses (documents 100–119; 9,195 LLaDA-8B-Instruct and 8,820 Dream-7B pairs), recording the log-probability of the document's token under both marginals (`code/analyze_truth.py`). As a replication, the conservative median share on these documents is 0.412 and 0.445, against Table 1's 0.414 and 0.450. The mean difference is $+0.0154$ $[+0.0057, +0.0255]$ nats per pair on LLaDA-8B-Instruct and $+0.0147$ $[+0.0031, +0.0306]$ on Dream-7B, resampling documents and then anchor rows, and $+0.038$ $[+0.013, +0.062]$ and $+0.056$ $[+0.009, +0.108]$ on the 3,599 and 1,992 pairs whose defect clears $10\times$ the floor. Over all pairs, as a multiple of the mean jackknifed defect, it is 0.82 $[0.35, 1.28]$ and 0.96 $[0.18, 1.84]$. By tercile of the defect it is -0.0001, $+0.004$ and $+0.043$ on LLaDA-8B-Instruct and -0.0001, $+0.002$ and $+0.042$ on Dream-7B, so it is carried by the pairs whose defect is large; pair by pair $\bar{q}_j$ scores higher on only 49–53%, and the gain lies in the size of the wins rather than their number. $\log \bar{q}_j$ is estimated from the probe's finite draws and is biased down by Jensen, against $\bar{q}_j$; scoring each of the two half-sample mixtures, which use half the draws and so carry about twice that bias, and averaging gives $+0.0145$ and $+0.0138$. Nor is the gain a generic effect of the mixture's entropy: tempering p_j, pair by pair, to the entropy of $\bar{q}_j$ changes its score by $+0.0003$ $[-0.0022, +0.0028]$ and $+0.0005$ $[-0.0020, +0.0031]$ nats, and $\bar{q}_j$ beats the tempered marginal by $+0.0151$ $[+0.0064, +0.0245]$ and $+0.0142$ $[+0.0033, +0.0278]$, almost all of its gain. A single reveal, with no averaging over the anchor's value, scores 0.054 and 0.060 nats below p_j, as a conditional on a guessed value should; the gain is in the mixture over the anchor's values that defines $\bar{q}_j$. On the unfiltered anchors of Appendix F the difference is larger, $+0.028$ $[+0.015, +0.042]$ and $+0.029$ $[+0.009, +0.051]$. Where the two marginals disagree, then, the one-pass marginal a parallel step samples each position from is on average the less accurate. The gain is not uniform: as a multiple of the mean defect it is 0.41 and 0.30 at adjacent pairs (distance ≤ 3), neither resolved from zero, and 1.48 and 1.79 at distances of 21 or more; on Dream-7B at mask ratios 0.3–0.5 it is -0.002 $[-0.008, +0.004]$, against $+0.035$ at 0.7–0.9; and it is positive on 15 and 12 of the 20 documents. The rollout's marginal is the better of the two on average, not an exact one. What these controls do not exclude is a generic benefit of averaging the model's predictions over perturbations of the context that carry no information about j; a placebo mixture over such perturbations, at equal compute, would separate the two, and we have not run it. The documents are OpenWebText, which may overlap the models' pretraining data; that would favour whichever marginal fits the training text better, not either one by construction.

F Estimator details and numerical checks

Stratified anchor. For an anchor i we enumerate the K most probable tokens of p_i with exact weights and draw n_{tail} samples, with replacement, from the renormalized tail; the estimate of any expectation over $v \sim p_i$ is the head sum plus the tail mass times the tail average. K is capped at 32 with $n_{\text{tail}} = 12$ on MDLM-OWT and at 24 with $n_{\text{tail}} = 10$ on the three 7–8B models, and is adaptively reduced when the head already carries 0.95 of the mass. Anchors whose head mass is below 0.80 are discarded before any statistic is computed. The threshold is the same for every model, but the head it is measured over is that model's own top-K, so the filter is slightly stricter on the 7–8B models. Targets are 32 masked positions per anchor on MDLM-OWT and 16 on the 7–8B models, stratified by distance so that adjacent, short-range, and long-range pairs are all represented; each list includes the anchor itself, which is dropped, so the pair counts are 31 and 15 times the row counts. Because the discarded anchors are the flat ones and ρ rises with peakedness (§4), this filter is not neutral for the share, and §4 reports its effect rather than leaving it implicit. Rerunning the probe on MDLM-OWT with no filter, where 51.2% of randomly drawn masked positions pass it, reads 0.446 above the threshold, 0.114 below it and 0.179 $[0.178, 0.181]$ (resampling pairs) over

all 100,540 pairs; below the threshold the tail term dominates and, as in the example below, the jackknife can still read high, so 0.114 and 0.179 are if anything too large. On LLaDA-8B-Instruct and Dream-7B an unfiltered run on 10 further documents each (documents 120–129), drawing anchors the same way, passes 56.2% and 53.4% of anchors and reads 0.242 $[0.211, 0.283]$ and 0.245 $[0.217, 0.282]$ over all pairs, 0.582 and 0.548 above the threshold and 0.161 and 0.154 below it, resampling documents and then rows (`code/analyze_share_population.py`); reweighted to the main probe's mask-ratio mix, the above-threshold readings become 0.441 and 0.442, against 0.423 and 0.458 in the main probe, for the reason given next. The MDLM-OWT unfiltered run draws its anchors uniformly and filters afterwards, and on the 70 documents both cover it reads 0.448 above threshold against the main probe's 0.377. The gap is the mask-ratio mix, not sampling noise: the main probe draws its anchors among filter-passing positions in equal numbers at every ratio, whereas in the unfiltered run few anchors pass at ratio 0.9 and many at 0.3, where ρ is highest (10% against 35% of pairs at 0.9, 27% against 13% at 0.3); in the main probe the equal anchor counts become unequal pair counts through the $\hat{D}$ filter, which removes more pairs at low ratios, where more context leaves less dependence. The pooled median therefore depends on both design choices. Ratio by ratio the two runs read 0.628 and 0.585 at 0.3, 0.461 and 0.453 at 0.5, 0.357 and 0.374 at 0.7, and 0.295 and 0.282 at 0.9, and reweighting the unfiltered run's pairs to the main probe's ratio mix gives 0.370 (`code/analyze_share_robustness.py`). Table 1's intervals resample documents first and anchor rows within each drawn document, giving $[0.344, 0.401]$, $[0.388, 0.446]$, $[0.365, 0.500]$ and $[0.397, 0.514]$ for the conservative row; resampling anchor rows alone gives intervals 1.3–1.6× narrower, and resampling pairs, the wrong unit because a row's pairs share an anchor and ρ rises with its peakedness, 2.4–4.1× narrower still. Because the pooled median depends on the ratio mix, shares are compared within a run and never across, and we do not claim the ordering of the per-model shares in Table 1 — only that the defect is a large share of the score on each.

The bias example of §3. `code/e1c_split_check.py` draws a random enumerable pair of conditionals over 400 tokens, whose 8 most probable anchor tokens carry 9.7% of the mass, and runs the stratified estimator with $K = 8$ and 12 tail draws 300 times; the figures are means over those runs. The example sits outside the head-mass filter on purpose, to show the plug-in's bias where it is largest; at $K = 32$ with 24 tail draws (26% head) the plug-in, jackknife and split-sample readings are 0.281, 0.264 and 0.196 against the same true 0.250.

Jackknife. $\hat{\Delta}$ is the plug-in $\mathrm{KL}(\bar{P}\|p_j)$ with $\bar{P}$ the weighted mixture; its bias is removed to first order by the leave-one-out jackknife over the tail draws, with the enumerated head held fixed (over rollouts in the k-sweep), and the result is clipped at zero, as is $\hat{I}_{\mathrm{cor}}$ (below).

Split-sample bound. With n rollouts (or tail samples) split into halves, $\hat{\Delta}_{\mathrm{split}}$ is as in §3. Its expectation is at most Δ by Jensen's inequality applied to the concave logarithm in $\mathbb{E}\log\bar{P}_B(v) \le \log\mathbb{E}\bar{P}_B(v) = \log\bar{q}(v)$; the two halves must be independent, which holds because they are separate draws. For the k-sweep, per-step split estimates are summed, so the sum is a lower bound on $\sum_i \Delta_\pi(i)$ in expectation. The one-sided 95% bound reported is the 5th percentile of the bootstrap distribution of the mean over set–order cells; resampling whole sets instead is reported below.

Set-level estimator. For an order π of a set, each rollout reveals the set one position at a time, sampling each token from the model's conditional given the prefix drawn so far. The parallel error is estimated by the chain rule, $\mathrm{KL}(q_\pi\|m_S) = \sum_t \mathbb{E}[\mathrm{KL}(p_\theta(x_{\pi_t} \mid c \oplus \text{prefix}) \,\|\, p_{\pi_t})]$, with each inner KL computed exactly over the vocabulary and the outer expectation averaged over rollouts, which is unbiased (a Rao-Blackwellised estimate). The defect of the t-th position is the KL from the rollout-averaged conditional to p_{π_t}, estimated as for pairs, and $\mathrm{TC}(q_\pi)$ is the residual $\mathrm{KL}(q_\pi\|m_S) - \sum_t \hat{\Delta}_\pi(\pi_t)$, so it inherits the defect estimator's bias.

Bias ladder. Plug-in ≥ jackknife is guaranteed by convexity (Appendix A) and holds in every cell, as a check on the code; jackknife ≥ split, which is not guaranteed, holds for every k in aggregate, and 11 of 646 MDLM set–order cells violate it at $k = 8$ and are reported, not removed.

Float64 accumulation. $\hat{D}$ is computed as the cross-entropy $-\langle P_s, \log p_j\rangle$ minus the entropy $H(P_s)$, two vocabulary sums of order nats, while D can be 10^{-3}. Accumulating the vocabulary sum

in float32 gives an absolute error of 1.9×10^{-4} nats on D against a float64 reference; accumulating in float64 gives 2.3×10^{-7}. $\hat{\Delta}$ is summed termwise as a log-ratio and I_{ent} is a difference of two entropies each accumulated in float64; their errors are 2.4×10^{-7} and 7.9×10^{-7}. All reported numbers use float64 accumulation.

Asymmetry. D is measurable in both directions only for a pair both of whose positions were probed as anchors, and each cell's target list is seeded with the cell's own anchors, so the bidirectional set is exactly the anchor–anchor pairs. Both of their endpoints therefore passed the head-mass filter, which the peakedness trend of §4 makes a non-neutral condition. Table 1's asymmetry row is the 5,185, 3,277, 884 and 843 such unordered pairs whose two directions average at least 10^{-3} nats. Being filtered on the mean of the two directions rather than on each, they are not a subset of the 46,863, 19,736, 5,295 and 5,160 directed pairs the ρ rows use, and each stands for two directed evaluations.

Estimator used in the measurements. $\hat{\Delta}$ is the jackknifed estimate except where the split-sample one is named (Table 1's conservative row, Table 8's bounds), and $\hat{I}_{\text{cor}} = \hat{D} - \hat{\Delta}$ is clipped at zero. The jackknifed $\hat{I}_{\text{cor}}$ is non-negative in exact arithmetic (Appendix A), so the clip guards against rounding alone (the head mixture is accumulated in float32 before the float64 reduction) and it binds on 8 of the 77,054 pairs whose $\hat{D}$ clears 10^{-3} nats.

Numerical floor. At $k = 1$ the induced marginal is the direct marginal by construction. The magnitude of the mean split-sample reading there over scattered cells, the floor, is 9.0×10^{-9} nats for MDLM (float32 weights), 4.9×10^{-4} for LLaDA-8B-Instruct and 8.1×10^{-4} for Dream-7B (bf16 weights; a batch of 32–128 rollouts and a single direct forward reduce in different orders). LLaDA-8B-Base, which shares LLaDA-8B-Instruct's architecture and precision, has no k-sweep, and we use LLaDA-8B-Instruct's floor for it. The floor is measured in the rollout pipeline and transferred to the pair probe, whose batches are of the same kind. The same floors bound how much of the *pair* pool is resolved, which matters because Table 1's inclusion rule is $\hat{D} \geq 10^{-3}$ nats and not a floor rule. The median $\hat{\Delta}$ over included pairs is 5.2×10^5, 13 and 6 times the model's floor on MDLM-OWT, LLaDA-8B-Instruct and Dream-7B; 0.0%, 43.4% and 62.5% of individual pairs lie within $10\times$ of it and are therefore not resolved one at a time, and dropping them moves the median share from 0.377 to 0.377, 0.423 to 0.411 and 0.458 to 0.430. The medians in Table 1 are robust to the floor; individual bf16 pairs near it are not, and we make no per-pair claims on those models. The floor also confounds the cross-scale comparison, since MDLM-OWT runs in float32 and the 7–8B models in bf16, and at zero truth a KL's numerical contamination is positive. Injecting a floor-magnitude perturbation into MDLM's own $\hat{\Delta}$ and $\hat{D}$ moves its median share from 0.377 to 0.393 at LLaDA-8B-Instruct's floor and 0.399 at Dream-7B's; since shares are not compared across runs, we read this only as bounding what precision alone could add to a bf16 model's share, about a third of the difference between the measured values. We also settle it directly rather than by injection, by running LLaDA-8B-Instruct itself in float32 (with the weights split between GPU and host memory) and pairing it against a bfloat16 run on the same documents, ratios and (anchor, target) pairs. On the 2,279 pairs with $\hat{D} \geq 10^{-3}$ nats in both, the defect is unchanged: $\sum \hat{\Delta}_{\text{fp32}} / \sum \hat{\Delta}_{\text{bf16}} = 0.985\ [0.969, 1.001]$, and ρ moves by $-0.011\ [-0.015, -0.008]$ in the mean and -0.005 in the median. Against that, two bfloat16 runs of the same probe differ by $+0.000\ [-0.002, +0.002]$ in ρ, so the precision effect on the share is real but small against the share itself, and smaller than the injection above predicted. What precision does move is the per-pair *sign* of I_{ent}, which disagrees on 12.4% of pairs between precisions against 1.6% between reseeds, while the rate itself barely moves (0.496 against 0.512) — which is why we report the rate and the tail rather than per-pair signs on the bf16 models. The tails the certificate uses do not move: on these pairs I_{ent} falls below -0.05 nats on 25.7% in bf16 and 24.4% in float32, and below -0.2 on 7.5% in both, while the change of precision alone moves a pair's I_{ent} down by more than 0.05 on 9.2% of them and by more than 0.2 on 0.5%. The bf16 replicate nulls of Appendix G see reseed noise only, and this is the check on the arithmetic they share. That control is at the pair level, while the certificates of Table 8 are at the set level, so we repeat it there, with a float32 k-sweep over cells the bfloat16 run already covers — the same documents, mask ratios, structures and unmasking orders. At $k = 1$ the induced marginal *is* the direct marginal, so the truth is exactly zero and the reading is the floor itself: 4.7×10^{-9} nats in float32 against 4.6×10^{-4} in bfloat16 on the same 64 scattered cells, a factor of 9.8×10^4. The floor is arithmetic and not estimator bias, and in float32, where the absolute guard below would bind as it does on MDLM-OWT, the

$k = 2$ bound of Table 8 would clear its threshold by about $900\times$ rather than $1.88\times$. What precision does not move is the quantity we report: the share at $k = 2$ on those cells is 0.341 $[0.272, 0.435]$ in float32 against 0.349 $[0.259, 0.448]$ in bfloat16. The rollouts are redrawn between the two runs, so the absolute nats follow the draw and only the share is comparable across them. We certify a defect only if its lower bound exceeds $10\times$ the model's own floor, subject to an absolute guard of 10^{-5} nats, i.e. thresholds of 10^{-5}, 4.9×10^{-3} and 8.1×10^{-3} nats; the guard binds on MDLM-OWT, whose floor is far below it. Over all cells, contiguous included, the $k = 1$ control reads a mean $|\mathrm{KL}|$ of 9.2×10^{-8} nats on MDLM-OWT and 5.0×10^{-4} and 9.4×10^{-4} on the bf16 models — on the bf16 models that is the floor itself to within a factor of 1.2, and on MDLM-OWT it is two orders of magnitude below the absolute guard of 10^{-5}, which is the threshold that binds there. The floor is the magnitude of the *mean* reading at $k = 1$, because what is certified is a lower bound on a mean and that is the systematic component such a bound inherits; the scatter around it is already carried by the bound's own interval. Two stricter conventions are worth reporting. Taking instead the mean of the *absolute* split-sample readings at $k = 1$ over all cells (2.1×10^{-8}, 5.0×10^{-4} and 9.4×10^{-4}; on the bf16 models every rollout's first conditional is the same batched forward, so the split, jackknife and KL readings there are one number, while on MDLM-OWT they differ at 10^{-8}) every certified cell survives except Dream-7B at $k = 2$, which falls to $0.94\times$ its threshold. That is the narrower of the two cells Table 8 marks as marginal. The float32 control above shows, on LLaDA-8B-Instruct, that the bf16 floor is arithmetic rather than estimator bias; no float32 run exists for Dream-7B, so its $k = 2$ cell stays marked as marginal rather than settled. And since $\sum_i \Delta_\pi(i)$ adds $k - 1$ non-zero terms while the floor is measured on one, a floor scaled by $k - 1$ is the conservative reading at large k: under it every cell still clears, at $1.09\times$ to $2.47\times$ on Dream-7B, $1.88\times$ to $5.85\times$ on LLaDA-8B-Instruct and $10^3\times$ throughout on MDLM-OWT. Every scattered entry of Table 8 clears its threshold, but not by the same margin: Dream-7B at $k = 2$ clears it by only $1.09\times$ (0.00886 against 0.00811) and LLaDA-8B-Instruct at $k = 2$ by $1.88\times$ (0.00915 against 0.00488), while every cell at $k \geq 3$ clears by about $2.5\times$ or more and MDLM by $10^3\times$ throughout. The two $k = 2$ cells on the 8B-class models are therefore the marginal ones, and we flag them as such in the table rather than reporting them at the same confidence as the rest. Clustering the bootstrap on sets rather than cells (resampling the four orders of a set, and their nested prefixes, together) puts Dream-7B's $k = 2$ cell at $1.06\times$ its threshold and leaves every other cell clearing it, and a Holm correction over all 21 one-sided tests, each P the set-clustered bootstrap's share of resampled means at or below the threshold, rejects every null, the largest P being 0.029 for that same cell (`code/analyze_share_robustness.py`).

Model conditioning check. Before any probe, each loaded model is scored on three real documents with 16 of 128 positions masked: the negative log-likelihood of the true tokens under real context, under the same tokens shuffled, and under a fully masked context. Real context must beat both by at least 0.5 nats or the run aborts. The margins are 6.7 and 6.0 nats for MDLM-OWT, 7.3 and 8.0 for LLaDA-8B-Instruct, 7.4 and 8.2 for LLaDA-8B-Base, and 7.5 and 7.7 for Dream-7B. A run writes the margins it was produced under into its output file as a provenance record. The check exists because a model whose positional encoding is silently missing, so that it cannot tell positions apart, or whose logits are read one position off, still produces finite, plausible-looking scores; both failure modes are invisible to any statistic computed from the scores alone. The first is caught by the real-versus-shuffled margin; the second by a slot-alignment check (`code/slot_alignment_check.py`), which reads the true token's log-probability at its own masked index and at the two neighbouring ones. It does not ask the model to be right, only to be aligned, so it applies to a 110M model as well as an 8B one: a one-position offset scores the token highest at a neighbour. On MDLM-OWT the token scores highest at its own slot on 0.933 of 298 masked positions, against 0.044 and 0.023 at the left and right neighbours; on LLaDA-8B-Instruct and Dream-7B, over 592 masked positions of 16 documents each, on 0.961 and 0.963, against 0.005 and 0.034, and 0.007 and 0.030. Dream-7B keeps an autoregressive model's output convention, in which row t scores token $t + 1$, and our loader applies the shift its own sampler applies; this check is what confirms it.

G Budget, replicate and in-distribution controls

Budget replication. On the anchors of the first 20, 6 and 6 documents (1,374 MDLM-OWT pairs, 138 LLaDA-8B-Instruct pairs and 780 Dream-7B pairs) we re-estimate every quantity with the stratification widened to an exact head of up to 64 tokens and a 64-sample tail, $3.9\times$, $4.9\times$ and $5.2\times$ the anchor's forward passes of the main ("cheap") estimate. On Dream-7B the cheap and

widened ("gold") estimates correlate 0.998 for $\hat{D}$ and 0.998 for $\hat{\Delta}$, and the median share moves from 0.4593 to 0.4537. The median shift, gold minus cheap, is $+0.002$ $[-0.016, +0.012]$, -0.025 $[-0.046, +0.026]$ and -0.006 $[-0.027, +0.012]$ on the three models, resampling anchor rows; the LLaDA-8B-Instruct subset is small and its median (0.343) sits below Table 1's, and its mass-weighted share moves by only -0.009. Correlations between cheap and gold estimates are $0.9997/0.9989$ for $\hat{D}$, $0.9999/0.9974$ for $\hat{\Delta}$ and $0.9993/0.9987$ for I_{ent} (MDLM/LLaDA); the sign of I_{ent} agrees on $98.8\%/99.3\%$ of pairs; the median share ρ is 0.3815 (cheap) against 0.3833 (gold) on MDLM and 0.3435 against 0.3182 on LLaDA; the fraction of $I_{\text{ent}} < -0.05$ is 0.154 against 0.151 and 0.254 against 0.268.

Replicate-noise null. Re-running the estimator on the same pairs with independent samples gives a distribution of I_{ent} differences whose fraction below -0.05 nats is 0.000 and below -0.2 is 0.000; the fraction below 0 is ~ 0.49, as it must be for a difference of two independent estimates, so it says nothing about the sign of I_{ent} itself; the tail fractions are what this null is for. The same replication gives certificate (A) a null it otherwise lacks. Table 1's asymmetry row is $|D(i{\to}j) - D(j{\to}i)|$ over the mean of the two, and a noisy estimator makes that positive whether or not the model is incoherent; applied instead to two independent estimates of the *same* direction, at the two budgets, the same statistic reads a median of 0.022 on MDLM-OWT and 0.035 on LLaDA-8B-Instruct and exceeds 0.5 on 1.2% and 1.9% of the pairs with $\hat{D} \geq 10^{-3}$, against measured asymmetries of 0.70–0.78 and 62–65%. One of those estimates is at the widened budget, so this null somewhat understates the noise of the measured statistic, which contrasts two main-budget estimates, and it is taken on the replication subset rather than on anchor–anchor pairs; the same-budget null below is the matched one.

Replicate nulls on the bf16 models. The nulls above come from the budget replication. A second bf16 run of the LLaDA-8B-Instruct probe re-estimates $2{,}018$ pairs with independent tail draws at the same budget: the I_{ent} difference falls below -0.05 nats on 1.5% of them and below -0.2 on 0.1%, against measured tails of 19.6% and 4.9%, and 97% of the pairs below -0.05 in one run are below it in the other; the same-direction asymmetry null is 0.024 against a measured 0.730. The same replicate on Dream-7B (862 pairs) is far noisier: 24.5% of the differences fall below -0.05 and 3.4% below -0.2. So Dream-7B's -0.05 tail (22.7%) is within its own noise and only its -0.2 tail (6.2%) clears it; its asymmetry null, 0.117, stays far below the measured 0.708. These nulls use the raw difference of two estimates, whose variance is twice one estimate's; scaled by $1/\sqrt{2}$, as the order-spread null of Appendix H is, they read 1.0% and 0.1% on LLaDA-8B-Instruct and 17.9% and 1.4% on Dream-7B, and no conclusion changes. They see reseed noise only, not the arithmetic that two bf16 runs share; the float32 run of Appendix F leaves LLaDA-8B-Instruct's tails where they were. We therefore rest the negative-I_{ent} certificate on MDLM-OWT and LLaDA-8B-Instruct, and on Dream-7B only at -0.2; LLaDA-8B-Base has no replicate.

In-distribution contexts. Table 1 probes the instruction-tuned models on raw web text, which is not what they decode. We repeat the probe on GSM8K reference solutions, in the in-domain format a decode produces: each of 40 test problems, disjoint from the decoding experiments, in the model's own chat template with the prompt of Appendix N, followed by its reference solution with the calculator annotations removed, masking a fraction r of the solution tokens only. Solutions are far more predictable than web text (only 42% and 44% of pairs clear $\hat{D} \geq 10^{-3}$, so these medians are over a more selected subset) but the share is unchanged: the conservative median is 0.425 $[0.395, 0.461]$ on LLaDA-8B-Instruct and 0.450 $[0.410, 0.492]$ on Dream-7B, against 0.414 and 0.450 in Table 1, and 0.432 and 0.434 on the pairs whose defect is resolved above $10\times$ the bf16 floor. The asymmetry is 0.831 and 0.849, and I_{ent} falls below -0.05 nats on 8.8% and 10.4% of pairs, the latter within Dream-7B's replicate noise.

Placebo. Table 1's main condition and the count-matched placebo share documents, ratios, and index pairs; the aggregate ratio is $\sum \hat{\Delta}_{\text{placebo}} / \sum \hat{\Delta}_{\text{main}}$ over matched pairs. The MDLM-OWT placebo has 629 anchor rows, the LLaDA-8B-Instruct placebo 591 and the Dream-7B placebo 586, of which 317 share an anchor with the main run; resampling pairs instead of documents gives the narrower $[4.34, 5.74]$, $[2.72, 3.22]$ and $[2.01, 3.02]$.

Base versus Instruct. Full paired statistics, over pairs whose target is not the anchor itself: $2{,}089$ pairs present in both, $1{,}722$ with $\hat{D} > 10^{-4}$ in both; median paired difference in $\hat{\Delta}$

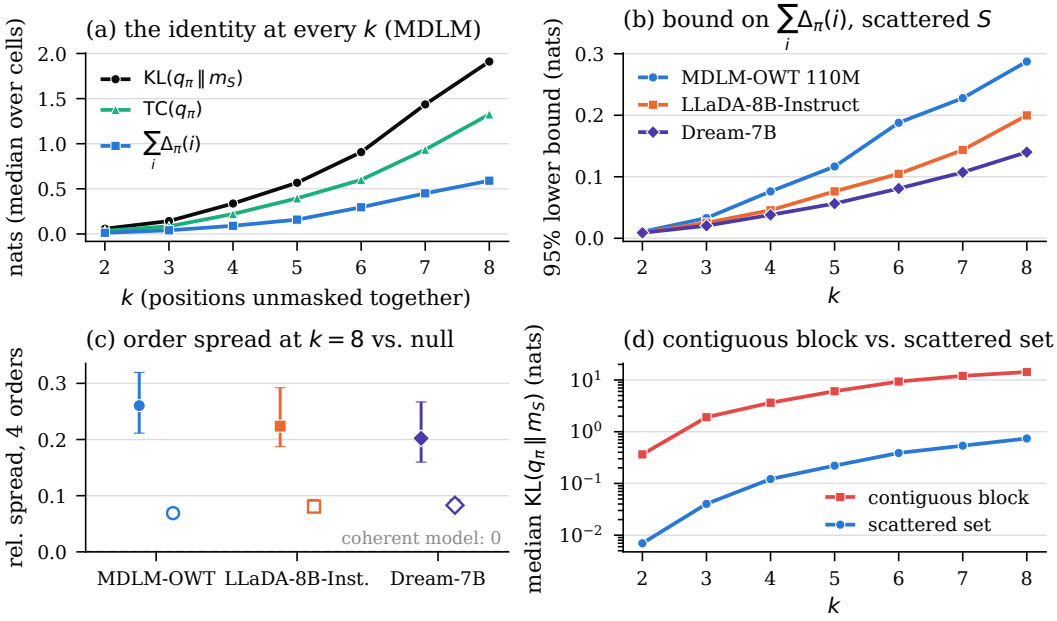

Figure 3: Sets of k positions. (a) The three terms of (2) on MDLM-OWT, each at its own median over 646 cells (medians do not sum); the defect sum is 19–21% of the parallel error at every k, mass-weighted as in Table 8. (b) One-sided 95% lower bound on $\sum_i \Delta_\pi(i)$ for scattered sets, from the split-sample estimator, on three models: positive at every k. (c) For a fixed 8-position set, the median relative spread (range over mean) of $\mathrm{KL}(q_\pi \| m_S)$ across four unmasking orders with a bootstrap interval (filled), against its range-matched replicate-noise null (open, with a bar to its 95th percentile; Appendix H); a coherent model would give zero. (d) Median parallel error of contiguous blocks against scattered sets of the same size on MDLM-OWT.

-0.00002 $[-0.00010, +0.00003]$ nats and mean $+0.00305$ $[+0.00075, +0.00544]$; fraction with Instruct larger 0.488; median paired difference in ρ -0.0053 $[-0.0112, -0.0014]$; aggregate ratio $\sum \hat{\Delta}_{\mathrm{Inst}} / \sum \hat{\Delta}_{\mathrm{Base}} = 1.18$ $[1.05, 1.32]$. The sum is carried by the upper tail; the median pair does not move and the share barely does.

H Split-sample penalty and contiguous blocks

Which order is cheapest. Over MDLM-OWT's 161 8-position sets the least-error order is confidence-descending 49 times, left-to-right 28 and each of the two random orders 42. A uniform split produces counts at least this uneven 12% of the time ($\chi^2 = 5.8$, 3 d.f.), so these data do not resolve which order minimizes the parallel error.

What the order comparison is measured on. MDLM-OWT's 646 set–order cells are 161 sets under four orders and one set, cut short, under two, which the order comparison leaves out; LLaDA-8B-Instruct has 110 sets and Dream-7B 128, half contiguous and half scattered in each. The four orders share a set only at $k = 8$: $161/161$ sets share it there and $0/161$ at $k = 2$ and $k = 4$, which is why the spread is reported at $k = 8$ alone and why Figure 3c plots that one point per model. The spread is a range over four orders divided by their mean, at 128 rollouts, so its null has to be a range too. Every cell of the three models was re-run with fresh rollouts at the same budget (on MDLM-OWT and LLaDA-8B-Instruct the replicate, run longer, also covers 7 and 2 further sets, which the comparison does not use). For each set, the differences between the two runs' estimates of the same order, divided by $\sqrt{2}$, match one estimate's noise in variance, and in distribution if that noise is symmetric; four draws with replacement from the set's own scaled differences, each given a random sign, form a noise-only four-order range, which is divided by the set's four-order mean as the spread is. The median of that ratio over sets, recomputed 2,000 times, is the null of §4.1,

Table 2: Contiguous blocks against scattered sets at the 128-rollout budget: median parallel error $\mathrm{KL}(q_\pi \| m_S)$ and median jackknifed defect sum $\sum_i \hat{\Delta}_\pi(i)$ in nats, over every set–order cell of each structure (cell counts in parentheses, contiguous then scattered). The ratio is of the two medians.

		parallel error			defect sum	
model (cells)	k	contiguous	scattered	ratio	contiguous	scattered
MDLM-OWT (324, 322)	2	0.364	0.007	52×	0.051	0.002
	4	3.631	0.121	30×	0.473	0.031
	8	14.3	0.739	19×	2.125	0.196
LLaDA-8B-Instruct (220, 220)	2	0.517	0.009	57×	0.062	0.003
	4	3.829	0.104	37×	0.526	0.031
	8	14.9	0.592	25×	2.322	0.173
Dream-7B (256, 256)	2	0.389	0.010	38×	0.048	0.004
	4	3.772	0.100	38×	0.486	0.027
	8	16.3	0.540	30×	2.153	0.127

and P is the fraction of those medians at or above the observed spread, so $1/2001$ is the smallest value it can take. For the parallel error the null is 0.069 (95th percentile 0.076) on MDLM-OWT, 0.081 (0.092) on LLaDA-8B-Instruct and 0.083 (0.092) on Dream-7B; for $\mathrm{TC}(q_\pi)$ it is 0.086, 0.099 and 0.086 against spreads of 0.294, 0.248 and 0.196; for the defect sum 0.140, 0.133 and 0.164 against spreads of 0.471, 0.479 and 0.382, ratios of 3.4, 3.6 and 2.3; every P is $1/2001$. Sets whose four-order mean is below 10^{-4} nats are skipped, which drops two sets from the TC comparison on MDLM-OWT. Because $\mathrm{TC}(q_\pi)$ and the defect sum are estimated through the defect's plug-in mixture, an order-dependent estimator bias that both runs share would pass through this null; the parallel error, estimated without bias (Appendix F), is the order statistic this concern does not touch. A cheaper replicate gives a conservative null in the same construction, since differences against a run with fewer rollouts overstate one estimate's noise: against Dream-7B's 32-rollout run, on the 64 sets whose orders both runs share, the null is 0.140 for the parallel error, 0.163 for TC and 0.373 for the defect sum, against spreads of 0.190, 0.196 and 0.415 on those sets. The cheaper run's null for the parallel error is 1.7, 1.9 and 1.7 times the same-budget one on the three models; independent Gaussian noise at a quarter, an eighth and a quarter of the budget, those runs' rollout counts, would give 1.58, 2.12 and 1.58. The relative difference between two single estimates of a cell, at 128 and 32 rollouts, reads 0.059 $[0.054, 0.066]$, 0.074 $[0.058, 0.094]$ and 0.075 $[0.059, 0.090]$ on the three models, but it is not the statistic the spread is, which is why the null above is built as a range. The TC spreads with their intervals are 0.294 $[0.240, 0.327]$, 0.248 $[0.170, 0.316]$ and 0.196 $[0.172, 0.266]$. The pooled parallel-error spread mixes the two structures: relative to their larger mean, contiguous blocks spread by a median 0.133, 0.118 and 0.125 on the three models (1.0, 1.3 and 1.3 nats in absolute terms), and scattered sets by 0.399, 0.360 and 0.299 (0.26, 0.20 and 0.12 nats); pooled, the absolute spread is 0.62, 0.52 and 0.37 nats. Analysed on its own, on the same sets, the same-budget replicate reproduces the parallel-error spread, at 0.247, 0.252 and 0.208 on the three models against 0.260, 0.224 and 0.202.

Contiguous against scattered. Table 2 gives both structures' median parallel error and defect sum at $k = 2$, 4 and 8 on each model: the contiguous block's parallel error is 38–57× the scattered set's at $k = 2$ and 19–30× at $k = 8$ (Figure 3d shows MDLM-OWT), and its defect sum 11–29×.

The split-sample penalty. Exactly, $\hat{\Delta}_{\mathrm{split}} = \frac{1}{2}[\mathrm{KL}(\bar{P}_A \| p_j) + \mathrm{KL}(\bar{P}_B \| p_j)] - \frac{1}{2}[\mathrm{KL}(\bar{P}_A \| \bar{P}_B) + \mathrm{KL}(\bar{P}_B \| \bar{P}_A)]$: the mean of the two half-sample plug-ins less half the symmetrized KL between the halves. We measure the penalty as the jackknife estimate minus $\hat{\Delta}_{\mathrm{split}}$; the last term dominates it, and it would fall as $1/n$ if the rollout conditionals concentrated. For a set, each position's own direct marginal takes the place of p_j. On 595 MDLM-OWT set–order cells measured at both 32 and 128 rollouts the penalty ratio is 0.44–0.47 at every k where a $4\times$ increase predicts 0.25; on 256 Dream-7B cells it is 0.47–0.49, and on 157 LLaDA-8B-Instruct cells measured at 16 and 128 it is 0.31–0.34 where 0.125 is predicted. Read as a power of n, the exponent is 0.53–0.57 on the three models, a descriptive two-point rate and not a law. For scattered sets the penalty is small enough at $n = 128$ to certify every k; for contiguous blocks, whose rollouts are far more dispersed, the

one-sided bound on MDLM-OWT is 0.021 and 0.024 nats at $k = 2$ and 3 and negative from $k = 4$; on LLaDA-8B-Instruct it is 0.023 at $k = 2$ ($4.8\times$ its threshold) and 0.002 at $k = 3$, below it; on Dream-7B it is 0.003 at $k = 2$, below its threshold. How many more rollouts would certify longer blocks depends on the true defect, which is what is unknown. The mean jackknifed contiguous defect on MDLM-OWT is 0.13, 0.33, 0.61, 0.96, 1.32, 1.72 and 2.27 nats at $k = 2, \ldots, 8$, against 0.45 for scattered sets at $k = 8$; at $k = 8$ it is 2.19 against 0.25 on LLaDA-8B-Instruct and 1.94 against 0.19 on Dream-7B.

I TOY MODEL

A three-layer transformer masked diffusion model (width 96) is trained on an enumerable joint over $|V| = 6$, $L = 6$ until its conditionals match the data's to a mean KL of 0.007 nats. Because the joint is enumerable, the data's own defect can be computed exactly and is $\leq 7.4 \times 10^{-16}$ for every pair and masking pattern. The trained model's defect is not: over the 887 of 900 probed pairs whose score exceeds 10^{-4} nats its median share ρ is 0.18 (0.035 over the 619 above 10^{-3}, where the toy's larger scores are mostly dependence), and 16.1% of its I_{ent} values are negative (16.3% over all 900 pairs). The identities themselves are checked separately on random enumerable conditional specifications (`code/e1a_synthetic_exact.py`): Proposition 1 holds to a residual of 6.7×10^{-16} across all 3! orders at $k = 3$, and the identities of Propositions 2–3 to 2.2×10^{-16} on pairs. Incoherence is created by training, not inherited from the data, which agrees with the Bayes-optimality result of Kim (2026).

Approximation error, or the objective? A model that has not converged has wrong conditionals, and wrong conditionals are generically incoherent, so one trained toy showing $\rho > 0$ cannot separate "the masked-diffusion objective induces incoherence" from "any imperfect fit is incoherent". A sweep bears on it, over the range it covers. We train the same architecture at five capacities and training lengths (26k to 2.7M parameters, 2k to 24k steps) and compute the exact ρ against each model's exact conditional fit to the data. The fit improves about 600-fold, from a mean KL to the data's conditionals of 0.203 nats down to 0.00034; the median share does not follow it down, running 0.183 $[0.169, 0.195]$, 0.290 $[0.265, 0.313]$, 0.209 $[0.186, 0.226]$, 0.214 $[0.158, 0.238]$ and 0.257 $[0.193, 0.327]$ over 1,500 exactly-measured pairs each. Over this range the *share* does not fall with fit quality: it moves non-monotonically between 0.183 and 0.290, with one model per setting, and ends no lower at the best fit than at the worst. It must fall eventually wherever the data's pairs are dependent, since $\Delta \to 0$ as the conditionals become exact while D tends to the data's mutual information; over a 600-fold range of fit it has not started to. The absolute defect does fall, as it must, since the data's conditionals are coherent and Δ is continuous in the conditionals; it falls roughly in step with D, whose scale drops $102\times$ over the same sweep. Better fitting therefore shrinks the defect and the dependence together without making the score any more of a mutual information. Nor is the direct marginal the more accurate of the two: against the data's own marginal, over the 1,454 of 3,000 random pairs whose defect exceeds 10^{-3} nats, the rollout's induced marginal is the closer on average, at a KL of 0.0088 against 0.0110 nats (difference $+0.0022$ $[+0.0007, +0.0037]$), though on only 47% of pairs singly (`code/toy_marginal_reference.py`), as the documents' own tokens also say on LLaDA-8B-Instruct and Dream-7B (Appendix E). We stop short of claiming the defect is unique to this objective, which would need a comparison against a differently-trained family of any-order conditionals rather than a sweep within one.

What the reported share statistic does on a model whose truth is known. Because Δ, D and ρ are all exact here, this model also settles what the estimator of §3 does to the statistic Table 1 reports. We run the same stratified head-plus-tail estimator on 4,000 pairs at the toy's own budget ($K = 2$, $n_{\text{tail}} = 4$, anchors filtered at head mass 0.80) and compare against the exact values. There $K = 2$ already covers a third of the six-token vocabulary, unlike the $K \leq 32$ of a real one, so this checks the ratio's bias and not the tail regime; the numerator's conservatism in expectation is a theorem (§3) for pairs not selected on the estimate. On this filtered population the true median share is 0.526. The split-sample numerator gives 0.520, below it (gap -0.006 $[-0.021, +0.002]$, resampling pairs); the jackknife numerator gives 0.535, above it. Adding Table 1's inclusion rule, $\hat{D} \geq 10^{-3}$ on the same draws, keeps 3,084 pairs with a true median of 0.363 and a split-sample median of 0.353 (gap -0.010 $[-0.025, +0.003]$). For the tail regime we build a synthetic enumerable family with the 7–8B models' budget: a 2,000-token anchor marginal of Zipf form whose top-24 tokens carry

0.80–0.95 of the mass, conditionals $\mathrm{softmax}(b + sz_v)$ around a shared base b, and a direct target marginal $\mathrm{softmax}(b + tz_0)$ with its own perturbation, so the family is incoherent by an amount t drawn independently of s ($\hat{D}$ spans about 10^{-4} to 1 nats). With $K = 24$ and $n_{\mathrm{tail}} = 10$ over 3,000 pairs the split-sample median is 0.468 against a true 0.470 (gap -0.002 $[-0.006, +0.003]$), 0.426 against 0.428 after the $\hat{D}$ filter (gap -0.002 $[-0.009, +0.001]$), and the jackknife's 0.470 and 0.431 sit at or just above the truth (`code/toy_bound_check.py --synthetic`). The median of the split-sample ratio thus sits at or below the truth in all four checks, though no check resolves it from the truth. So the split-sample version is conservative in aggregate and the jackknife version is not, which is why the first is the one the paper calls conservative and the second is reported as the primary estimate. Neither is a per-pair bound on the ratio: the split-sample estimate exceeds the exact ρ on 36% of individual pairs, as it may, since Jensen constrains the numerator's mean and says nothing about a ratio.

J SELECTION SCORED ON THE PARALLEL-STEP ERROR ITSELF

G is the model's own log-ratio realised at the data's tokens, so a test that charges a score the $|G|$ it forgoes is not independent of D. On the enumerable model of Appendix I no proxy is needed: for a set S revealed in an order π we compute $\mathrm{KL}(q_\pi \| m_S)$ exactly and ask which score picks the cheaper set. A context is drawn by revealing 0, 1 or 2 positions, uniformly, with values from the data joint, which reaches 577 contexts with four to six masked positions; we score every one of them and weight it by its probability, so the numbers below are exact expectations rather than estimates. The anchor is the most confident masked position, as in confidence-ordered samplers; each score, computed exactly, ranks the remaining candidates by its value relative to the anchor and takes the two lowest; the set is revealed in the order anchor first, then the two companions by position. The expected exact cost in nats is 0.455 for an oracle over every pair of companions, 0.671 for I_{ent}, 0.756 for D, 0.791 for a companion pair drawn at random (its exact mean over pairs), 0.857 for the I_q that $\hat{I}_{\mathrm{cor}}$ estimates, 0.946 for confidence and 0.955 for entropy.

Three things follow. The paper's claim survives the removal of the proxy: I_q costs 0.101 more than D, so D is the better of the two selectors on the quantity itself, consistent with Proposition 2(a) at the pair level. Confidence and entropy cost 0.191 and 0.200 more; on this joint committing the most nearly deterministic positions commits the costliest sets. But (a) does not extend to a set: D beats a random companion pair by only 0.035, and I_{ent}, the entropy-difference form whose negative values certify incoherence, selects the cheapest sets of all, 0.084 below D. With three to five candidates per context this is a small testbed and we do not generalise from it; what it shows is that the ordering of D above I_q in §5 also appears when sets are scored exactly rather than against G, and that being a valid mutual information is neither necessary nor sufficient for being a useful selector. On the real models the order is different: there I_{ent} is the worst of the three on probed pairs (Appendix M).

The same test on the real models, scored on the set. On LLaDA-8B-Instruct and Dream-7B the set's cost cannot be enumerated, but the rollout estimator of §4.1 gives it without bias, so the test runs without the proxy G (`code/gpu_set_select.py`). A context is one of 20 OpenWebText documents no main probe uses (documents 100–119), truncated to 128 tokens, at each of the four mask ratios; in each we take two anchors, the most confident masked position (median top probability 0.999) and a random position the probe of §3 would admit (median 0.70 and 0.61). The candidates are the 12 masked positions nearest the anchor, a block decoder's neighbourhood; one stratified anchor row gives every candidate's $\hat{D}$, $\hat{I}_{\mathrm{cor}}$ and I_{ent}, each rule picks two companions as above, and the set is revealed anchor first and then by position, its parallel error estimated from 128 rollouts. Over 160 and 159 anchor rows the mean cost in nats is, on LLaDA-8B-Instruct and Dream-7B, 0.471 and 0.464 for $\hat{D}$, 0.451 and 0.548 for $\hat{I}_{\mathrm{cor}}$, 0.775 and 0.640 for I_{ent}, 0.627 and 0.570 for random companions, 0.444 and 0.355 for confidence and 0.364 and 0.379 for entropy. The costs are heavy-tailed, with medians of 0.006 and 0.009 for $\hat{D}$ against 0.163 and 0.162 for random companions, so we report paired differences in the mean with intervals that resample documents and then rows. $\hat{I}_{\mathrm{cor}}$ minus $\hat{D}$ is -0.020 $[-0.131, +0.083]$ and $+0.084$ $[-0.007, +0.186]$, with $\hat{D}$ the cheaper of the two on 31% and 30% of rows and $\hat{I}_{\mathrm{cor}}$ on 18% and 14%; I_{ent} minus $\hat{D}$ is $+0.305$ $[+0.105, +0.511]$ and $+0.176$ $[-0.061, +0.409]$, with I_{ent} the dearer on 81% and 77% of rows; random minus $\hat{D}$ is $+0.156$ $[-0.015, +0.338]$ and $+0.106$ $[-0.114, +0.330]$, random the dearer on 78% and 73% of rows; and

entropy minus $\hat{D}$ is -0.107 $[-0.250, +0.025]$ and -0.085 $[-0.215, +0.035]$. So on real models I_{ent} is the worst selector on average at the set level, as it is on probed pairs and unlike on the toy; $\hat{D}$ is not resolved from $\hat{I}_{\text{cor}}$ in the mean and beats random companions on most rows; and choosing the companions by their own entropy or confidence is at least as cheap as choosing them by the score. Where $\hat{D}$ has an advantage it is over random and I_{ent}-chosen companions, not over the targets' own entropy or confidence.

K WHY THE ROLLOUT EXPERIMENT REPORTS A LOWER SHARE

Column 5 of Table 8 measures the same algebraic quantity as Table 1 on a different population with a different summary, so the gap between 0.209 and 0.377 is a fact about ρ rather than a discrepancy. The two summaries themselves differ in both directions: median-of-ratios against the mass-weighted $\sum \hat{\Delta} / \sum \hat{D}$ is 0.377 against 0.478 on MDLM-OWT, 0.423 against 0.327 on LLaDA-8B-Instruct and 0.458 against 0.340 on Dream-7B: on MDLM-OWT the large-D pairs carry more than the typical share, and on the 7–8B models less. Three things separate them. It pools contiguous with scattered sets and weights by mass, so it is dominated by the contiguous ones, whose parallel error is $42\times$ larger in the median over the 559 cells of that column, those whose parallel error at $k = 2$ exceeds 10^{-4} nats (0.521 against 0.0124 nats; over all 646 cells it is $52\times$, Table 2) and whose share is 0.161 against 0.636. It draws positions at random rather than from the head-mass-filtered pool. And it is a ratio of sums: the median-of-ratios over the same cells is 0.229. Restricted to the confidence-descending order that confidence-ordered samplers use, the $k = 2$ share is 0.356 mass-weighted, inside the 0.327–0.478 range of the pair probe's mass-weighted shares, and 0.508 by median, above Table 1's medians. These numbers come from `results/share_population.json`.

L WHAT THE DEFECT DOES NOT BUY A DECODER

Four negative results support §5's statement that reading the share off the model buys no detectable gain; with 120 problems, gains up to about 0.13 are not excluded. Figure 4 plots the filters, their placebos, the failure predictor and the I_{ent} ranking; the paragraphs below give every number, including the budget correction.

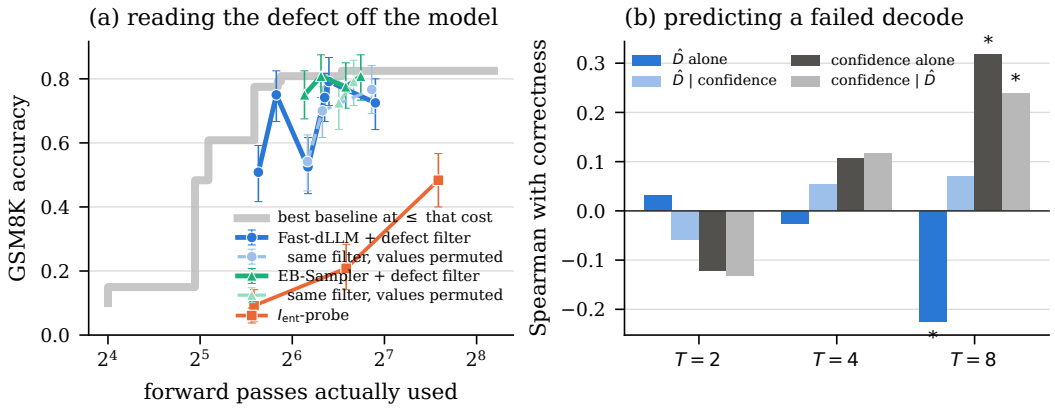

Figure 4: What the defect does not buy a decoder, on GSM8K with LLaDA-8B-Instruct. (a) The defect filters and the I_{ent} ranking against the grey envelope of the best accuracy any published or standard sampler reaches at or below each cost (Figure 2); the dashed curves are the permutation placebos, which shuffle the defect values across positions and commit the same number of positions per step. (b) Rank correlation between a per-problem quantity and whether the decode got the problem right, each shown alone and after the other is regressed out; stars mark $P < 0.05$ by permutation. The accumulated score $\hat{D}$ correlates with correctness at $T = 8$ and stops doing so once confidence is conditioned on.

The share inside three deployed decoders. Adding an anchor probe to a rule changes nothing it does. The two-pass probe was checked on all 720 generations of six decoder settings (confidence decoding at $T = 2, 4, 8$, EB-Sampler at $\gamma = 0.4$ and 1.6, Fast-dLLM at $\tau = 0.8$), and the stratified probe (an exact head of $K = 8$ tokens and 10 tail draws, 18 passes per probed step) on the first 60 problems of each setting, 360 generations: every decode reproduces its uninstrumented counterpart's text, prediction and cheap score exactly, and only the forward-pass count differs — which is what the probe costs. Two estimators run on the same committed positions of the same decodes. Each gives, per generation, the ratio of the summed defect to the summed score over the positions committed in parallel. The stratified jackknifed one gives medians over generations of 0.282 $[0.257, 0.294]$, 0.264 $[0.247, 0.278]$ and 0.294 $[0.265, 0.316]$ for confidence at $T = 2, 4, 8$; 0.303 $[0.282, 0.354]$ at $\gamma = 1.6$ and 0.298 $[0.282, 0.348]$ at $\gamma = 0.4$ for EB-Sampler; and 0.339 $[0.305, 0.370]$ at $\tau = 0.8$ for Fast-dLLM. Pooled into one ratio of sums within each setting they are 0.298, 0.271, 0.320, 0.333, 0.427 and 0.377; over all 360 generations the median is 0.291 and the pooled ratio 0.315. The two-point probe, at 2 passes, gives medians of 0.586, 0.524, 0.528, 0.527, 0.513 and 0.553, 0.536 over all generations, and a pooled ratio of 0.558. *What the two-point probe costs* below explains why they differ.

On Dream-7B, pair by pair. The same six settings on the first 60 problems with Dream-7B-Instruct give stratified medians of 0.275, 0.253 and 0.240 for confidence at $T = 2, 4, 8$, 0.253 and 0.268 for EB-Sampler at $\gamma = 0.4$ and 1.6, and 0.316 for Fast-dLLM, pooled ratios of 0.281, 0.259, 0.253, 0.255, 0.265 and 0.343, and 0.263 and 0.269 over all 360 generations; the two-point probe reads $1.89\times$ larger. This run also logged every committed pair: 78,256 in all, 42,430 with $\hat{D} \geq 10^{-3}$ nats, with a median share of 0.251 $[0.237, 0.267]$, resampling problems and then generations. The decode's anchors are sharp, a median entropy of 0.19 nats over the pairs with signal, and most committed scores are smaller still. On the 14.4% of pairs whose score exceeds the anchor's entropy, where Proposition 3(C) has force, the median share is 0.862, and the bound is undercut on 0.3% of them, within estimation noise; the other pairs carry 85.6% of the summed score and have a median share of 0.215. The in-decode share is low because most committed pairs score below their anchor's entropy, where (C) is silent, not because the bound fails.

These pairwise shares and the set-level shares of Appendix D are different quantities measured on the same kind of decode: the first charges each committed position's pairwise defect relative to one anchor, against its pairwise score, and the second the defect sum of Proposition 1 along the confidence order, against the set's parallel error, so they need not agree (0.427 against 0.568 at $\gamma = 0.4$).

What the two-point probe costs. The decode probe spends two forward passes, so it enumerates only the anchor's top two tokens and truncates the expectation over $v \sim p_i$ to a two-point support. Truncation changes both $I_q = \mathbb{E}_v[\mathrm{KL}(P_v \| \bar{q}_j)]$ and D, and neither moves in a fixed direction in general, so no one-sided theorem follows; the size and sign of the effect are a measurement, and we make it twice. On MDLM-OWT, where a 32-token head enumeration is affordable on a CPU, `code/top2_truncation_check.py` measures it over anchors drawn uniformly from the masked positions rather than by confidence, so that the whole peakedness range is covered: on 778 pairs the two-point share is the larger of the two on 98.5% of them, and the median ratio by the anchor's top-two mass is 3.05 below 0.5, 1.97 on $[0.5, 0.7)$, 1.54 on $[0.7, 0.85)$, 1.46 on $[0.85, 0.95)$ and 1.08 above 0.95 ($n = 452$, 144, 77, 41, 64). On LLaDA-8B-Instruct it is measured directly, by running both estimators on the same committed positions of the same decodes: the two-point share is larger on 100% of the 360 generations, by a median factor of 1.71 $[1.66, 1.76]$. That factor sits inside the MDLM calibration's range but well above what its top bucket, where the decode's confident anchors fall, would have predicted, so the calibration does not transfer across models and contexts and we use it only for the direction; §5 quotes the stratified reading and treats the two-point one only as what a two-pass probe would have said.

As a failure predictor. Over the 120 GSM8K problems decoded by confidence at $T = 2, 4, 8$, the two-point $\hat{D}$ accumulated along the decode per position committed in parallel differs between problems the decode got right and wrong by $+0.012$ ($P = 0.34$), -0.008 ($P = 0.28$) and -0.023 ($P = 0.011$) at $T = 2, 4, 8$ (difference of means, permutation P); confidence differs by -0.013, $+0.029$ and $+0.038$ ($P = 0.68$, 0.08, 0.002). So at $T = 8$ both predict marginally, and of the six tests only confidence's survives a Holm correction. At $T = 8$ the two are Spearman-correlated -0.83 across problems, and once each is regressed out of the other the score's partial correlation with

correctness is $+0.071$ ($P = 0.45$) while confidence's is $+0.239$ ($P = 0.009$). The score predicts failure only through confidence, though with 120 problems a partial correlation below about 0.25 could go undetected (the threshold for 80% power).

As a filter on a working sampler. The defect map is computed once per block from the block's most confident masked position (two forward passes per block) and used to keep a fraction κ of the candidates the gate admits that it ranks safest. On Fast-dLLM's confidence gate the six operating points $\tau \in \{0.3, 0.5, 0.8\}$, $\kappa \in \{0.5, 0.8\}$ reach 0.508–0.792 accuracy at 50–120 forward passes ($\tau = 0.3$ extends below the baseline grid); on EB-Sampler's entropy bound the four points $\gamma \in \{0.8, 1.6\}$, $\kappa \in \{0.5, 0.8\}$ reach 0.750–0.808 at 70–107. Against the same gate unfiltered at the same threshold, the filter gains $+0.133$ $[+0.058, +0.217]$ and $+0.142$ $[+0.058, +0.225]$ at $\tau = 0.5$ while spending 1.7–$2.4\times$ the passes, which an unfiltered $\tau = 0.8$ spends to reach 0.808; at the six other points that have an unfiltered counterpart it ties or loses (-0.083 to $+0.017$), and $\tau = 0.3$ has none. None of the ten beats the best baseline costing no more; the closest is $(\gamma, \kappa) = (0.8, 0.8)$, which ties Fast-dLLM at $\tau = 0.8$ while spending 79.4 passes against 58.8. The permutation placebo shuffles the same defect values across the same positions inside the block, so it commits the same number of positions with no signal; the filtered rule minus the placebo, both at $\kappa = 0.5$, is -0.017 $[-0.117, +0.075]$, $+0.042$ $[-0.042, +0.133]$ and -0.042 $[-0.108, +0.025]$ on the confidence gate at $\tau = 0.3, 0.5, 0.8$, and $+0.017$ $[-0.033, +0.067]$ and $+0.050$ $[-0.025, +0.133]$ on the entropy gate at $\gamma = 0.8, 1.6$. Every interval crosses zero.

As a correction to the gate's own budget. The paragraph above tests the defect as a re-ranking inside someone else's gate, which is not what the identity claims: it says EB-Sampler's *bound* is incomplete, not that its ranking is wrong. The test the identity implies charges both terms against the same budget. Its affordable proxy admits while $\sum_{j \in U} h_j - \max_{j \in U} h_j + \sum_{j \in U} \hat{\Delta}(j \mid a) \le \gamma$, where a is the block's most confident position and $\hat{\Delta}(j \mid a)$ comes from the same two forward passes per block the filter rule spends; it charges pairwise defects relative to one anchor rather than the $\Delta_\pi(j)$ of Proposition 1. Since $\hat{\Delta} \ge 0$ the corrected bound never admits more than the bound it corrects. Swept over $\gamma \in \{0.4, 0.8, 1.6, 3.2\}$ on the same 120 problems it reaches 0.817, 0.750, 0.750 and 0.717 at 118, 93, 72 and 60 forward passes. Against the best baseline costing no more that is -0.008 $[-0.058, +0.042]$, -0.075 $[-0.133, -0.017]$, -0.058 $[-0.133, +0.008]$ and -0.092 $[-0.175, -0.017]$: two ties and two losses, no wins. Against uncorrected EB-Sampler at the same γ, the direct test of charging the defect, it is $+0.008$ $[-0.033, +0.050]$, -0.042 $[-0.108, +0.025]$ and -0.025 $[-0.092, +0.050]$ at $\gamma = 0.4, 0.8, 1.6$, at 1.5–$1.7\times$ the passes: the charge makes the gate commit less per step without making it more accurate. Against the permutation placebo, which charges shuffled defect values against the same budget, it is $+0.033$ $[-0.025, +0.092]$ at $\gamma = 0.8$ and $+0.008$ $[-0.058, +0.075]$ at $\gamma = 1.6$. Both intervals cross zero. Charging the defect where the identity says it belongs, through this proxy, does not buy a decoder either. It is the sharper null of the two, because it tests what the identity implies rather than a re-ranking it never licensed, but the proxy is two-point, pairwise and computed once per block, and the two-point reading overstates the defect (above), so it tests the cheapest instantiation only.

As a ranking on its own, by $\hat{I}_{\mathrm{cor}}$ or I_{ent}. Ranking by $\hat{I}_{\mathrm{cor}}$ on GSM8K gives 0.100, 0.083 and 0.358 at 48, 96 and 192 forward passes, and by I_{ent}, the entropy-difference form whose negativity certificate this paper reports, 0.092 $[0.042, 0.142]$, 0.208 $[0.142, 0.283]$ and 0.483 $[0.400, 0.567]$. The I_{ent} rule is ahead of the $\hat{D}$ rule by $+0.108$ $[+0.025, +0.200]$ at $T = 4$ ($P = 0.029$) and unresolved from it at $T = 2$ and 8, and it trails confidence decoding at the same T by 0.19–0.21 at $T = 4$ and 8 ($P \le 0.002$). All three collapse for the reason $\hat{D}$ does. We report these as properties of the cheapest instantiations, which use one anchor and a two-point probe per step, and not as evidence that no defect-aware decoder can work.

M Calibration of the score against the realised gain

Does the reference order change which candidate wins? Proposition 2(a) makes $D(i \to j)$ the parallel step's error relative to the order (i, j), and Table 1 shows the reverse direction differs by 70–78% of the pair's mean, so the ranking a decoder reads off could in principle be an artefact of that choice. The anchor-first reference is the one a confidence-anchored decoder implies, since

the question is which targets to add to its most confident position; in a probe rule the anchor and its companions are committed in the same step, though, so it is a convention and not a fact about the decoder. The ranking's stability is measurable on the targets whose reverse direction was also probed, which are the anchor–anchor pairs of Table 1, peaked at both ends. Within an anchor row the two rankings have a Spearman correlation of $+0.599$ $[+0.579, +0.617]$, $+0.587$ $[+0.561, +0.613]$, $+0.602$ $[+0.551, +0.649]$ and $+0.573$ $[+0.524, +0.620]$ on the four models, and the safest third the two would commit overlaps on 0.45–0.53 of positions against a chance $1/3$. So the orders agree on much of the ranking and not all of it: the score is a statement about a named order, and about half of the borderline selections would change under the other one, on this doubly peaked population.

As a least-squares regressor for the signed G, on the pairs with $\hat{D} \geq 10^{-3}$, neither $\log(1+\hat{D})$ nor $\log(1+\hat{I}_{\mathrm{cor}})$ is better on any model: their R^2 are 0.107 and 0.080 on MDLM-OWT, 0.152 and 0.151 on LLaDA-8B-Instruct and 0.034 and 0.071 on Dream-7B, and the paired difference, resampling anchor rows, is -0.027 $[-0.071, +0.011]$, -0.002 $[-0.012, +0.008]$ and $+0.037$ $[-0.001, +0.060]$; and the ratio of $\hat{D}$ to the mean realised gain is 1.32, 0.95 and 1.29, an over-prediction on two models and an under-prediction on the third. So the score does not convert into an expected likelihood gain by a fixed factor, and it is not numerically the same object as G, which is the same log-ratio evaluated at the data's tokens rather than averaged under the model. That is why §5 uses G only to adjudicate which score's choice forgoes less, and not as evidence that either is calibrated. The paired ΔR^2 is that of $\log(1+\hat{I}_{\mathrm{cor}})$ minus that of $\log(1+\hat{D})$.

How robust the selection result is. Table 9's head-to-head holds under two changes and vanishes under a third. Restricted to pairs with $\hat{D} \geq 10^{-3}$ nats, where the numerical floor cannot rank them, $\hat{D}$ forgoes less than $\hat{I}_{\mathrm{cor}}$ by more, $+0.0073$ $[+0.0038, +0.0116]$, $+0.0070$ $[+0.0031, +0.0108]$ and $+0.0090$ $[+0.0051, +0.0129]$ nats on the three models. I_{ent} as a third selector forgoes 0.10–0.12 nats more than $\hat{D}$ on every model, the reverse of its standing on the enumerable toy (Appendix J). Charged the *signed* gain instead of $|G|$, the two do not separate ($\hat{I}_{\mathrm{cor}}$ minus $\hat{D}$ of $+0.0013$ $[-0.0025, +0.0047]$, -0.0010 $[-0.0041, +0.0020]$ and $+0.0012$ $[-0.0029, +0.0057]$): what $\hat{D}$ avoids better is the magnitude of the dependence it breaks, in either direction. The P values of §5 are fractions of $5{,}000$ bootstrap resamples; Dream-7B's document-level one rests on 20 documents.

N THE GSM8K FRONTIER IN FULL

Table 3 gives every operating point of Figure 2. Decoding is semi-autoregressive with block 32 and generation length 256, greedy within a step, on the first 120 problems of the GSM8K test split (Cobbe et al., 2021), each given in the model's chat template followed by "Reason step by step and put the final numeric answer after '#### '."; answers are graded by exact match on the number after the marker. The baselines are implemented from their own algorithm boxes: Fast-dLLM unmasks every position whose maximum probability exceeds τ, and the most confident one if none does (without its key-value cache, which changes speed but not passes); EB-Sampler sorts by entropy and grows the set while $\sum_{j \in U} h_j - \max_{j \in U} h_j \leq \gamma$; PUNT runs $\lceil \log_2 |M| \rceil$ bit-partition rounds on the confidence rank, committing the anchor half and pruning any j whose conditional moves by more than ϵ in KL; our dilated baseline, a confidence-gated variant of dilated scheduling, filters the confidence order by a minimum index gap $\max(1, \lfloor |M_t| g_0 / B \rfloor)$, where $|M_t|$ is the number of positions still masked at step t, B the block size and g_0 the schedule's initial gap. T is the number of steps per block of a fixed-step rule. The two probe rules spend two extra passes per step on the most confident masked position's top two tokens, renormalized, commit that anchor, and fill the step's remaining slots with the positions of lowest two-point $\hat{D}$ (or $\hat{I}_{\mathrm{cor}}$) relative to it. Compute is the number of forward passes actually issued, so PUNT's pruning rounds and our probes are charged in full; charged that way, PUNT at $\epsilon = 0.05$ costs more than one-token decoding, and Fast-dLLM, run without its approximate key-value cache, issues the passes its threshold rule implies with exact logits. These are reimplementations from the algorithm boxes, not the authors' code. Confidence decoding at $T = 32$ is one token per step, but in a *confidence* order rather than reading order, and on this task both EB-Sampler and Fast-dLLM beat it by a significant paired margin at 23–36% of the compute ($+0.108$ $[+0.033, +0.183]$ and $+0.092$ $[+0.025, +0.158]$). A left-to-right reference at the same budgets explains why: reading order reaches 0.817 at 256 forward passes, $+0.100$ $[+0.025, +0.175]$ over confidence order at the same cost, and is indistinguishable from EB-Sampler at 93 passes (-0.008 $[-0.067, +0.050]$) and from Fast-dLLM at

59 ($+0.008$ $[-0.050, +0.067]$). It gives 0.625 at 64 and 0.258 at 32. EB-Sampler is on the envelope at every cost it occupies, at 68.8 passes only by tying Fast-dLLM's 0.808 at 58.8. Paired against Fast-dLLM's 0.808 at 59 passes, EB-Sampler's 0.825 at 93 is $+0.017$ $[-0.042, +0.075]$. So the parallel samplers are not beating sequential decoding; they are matching it at 23–36% of the compute, which is what they claim, and the $T = 32$ row is a statement about the confidence order rather than a sequential upper bound. We report it as such. Matched at the step count, the two probe rules are not resolved from each other, with $\hat{D}$ minus $\hat{I}_{\text{cor}}$ at -0.017 $[-0.092, +0.050]$ for $T = 2$, $+0.017$ $[-0.058, +0.083]$ for $T = 4$ and $+0.067$ $[-0.042, +0.175]$ for $T = 8$, so at this sample size the ordering the controlled selection test of §5 resolves is not visible once the score, rather than a gate, chooses what to commit.

Matched at the step count. At the same T the probe rules and the fixed-step baselines commit the same number of positions per step, so the paired difference isolates which positions fill it. On LLaDA-8B-Instruct the $\hat{D}$ rule minus confidence decoding is -0.025 $[-0.083, +0.042]$, -0.300 $[-0.392, -0.200]$ and -0.267 $[-0.375, -0.150]$ at $T = 2, 4, 8$, and minus random order -0.017 $[-0.092, +0.058]$, -0.167 $[-0.250, -0.083]$ and -0.158 $[-0.267, -0.050]$ (Holm-adjusted $P = 0.002$ and 0.046 at $T = 4$ and 8, over the eighteen comparisons of both rules against the three baselines). On Dream-7B-Instruct only $T = 8$ resolves, -0.300 $[-0.438, -0.163]$ against confidence and -0.212 $[-0.338, -0.087]$ against entropy, and the rule is not below random order. On HumanEval it is below confidence at $T = 8$ by 0.079 $[0.030, 0.134]$ ($P = 0.004$, 0.075 after Holm) and above random order, by 0.018–0.037, at every T (`code/analyze_decode_paired.py`).

What the score's ranking actually selects. On the probe pairs, within an anchor row, $\hat{D}$ and the target's own entropy are rank-correlated $+0.43$ to $+0.45$ on the three models, so the third the score would commit overlaps the third a confidence gate would commit on 63–64% of positions against a chance 33%, but carries about twice the entropy (1.98 against 1.01 nats on MDLM-OWT, 1.74 against 0.85 on LLaDA-8B-Instruct and 2.00 against 0.94 on Dream-7B).

A second model. We repeat the sweep on Dream-7B-Instruct over the first 80 of the same problems, holding the decoding configuration, the knob grid and the grading fixed; Table 4 gives every operating point. What replicates is the claim of §5. The two probe rules give 0.125 and 0.100 at 48.0 passes, 0.188 and 0.138 at 96.0, and 0.287 and 0.212 at 192.0, each losing to the best baseline costing no more by 0.55–0.70 with every interval excluding zero; matched at the step count they are again indistinguishable from each other, with $\hat{D}$ minus $\hat{I}_{\text{cor}}$ at $+0.025$ $[-0.025, +0.075]$, $+0.050$ $[-0.013, +0.125]$ and $+0.075$ $[-0.050, +0.200]$. What does not replicate is the margin the parallel gates hold over confidence order: EB-Sampler's 0.838 at 56.5 passes is $+0.013$ $[-0.037, +0.062]$ over confidence order's 0.825 at 256.0, and Fast-dLLM at $\tau = 0.9$ reaches the same 0.825 at 64.1 with a paired difference of exactly zero on every problem. Restricting the LLaDA run to these 80 problems leaves its margins where they were, $+0.113$ $[+0.025, +0.212]$ for EB-Sampler and $+0.087$ $[+0.000, +0.175]$ for Fast-dLLM against $+0.108$ and $+0.092$ on the full 120, so this is the model and not the sample, although Fast-dLLM's interval, clear of zero on all 120, touches it on these 80. On Dream the gates match confidence order at a quarter of its compute instead of beating it at a third. That bears on the samplers rather than on the score, whose collapse as a ranking is the same on both models and is therefore not a property of one model's confidence calibration.

What "costing no more" means. The comparison above reads it as the mean passes an operating point spends, which is the frontier's own axis. The adaptive gates spend a different number on every problem, though, so a baseline that is cheaper on average can be dearer on some: EB-Sampler at $\gamma = 1.6$ spends between 20 and 78 passes on Dream against the $T = 2$ probe's fixed 48. Requiring the baseline to cost no more on *every* problem changes which baseline eight of the twelve comparisons are measured against and leaves all twelve negative, at 0.38–0.69 on LLaDA-8B-Instruct and 0.11–0.62 on Dream-7B-Instruct. On 80 binary differences a bootstrap quantile can only land on multiples of $1/80$, and the cheapest point on Dream sits on that grid, so we read the strict comparisons with the exact McNemar test on the discordant pairs instead, which does not depend on a resampling seed: eleven of the twelve give $P < 0.05$, and the twelfth is that cheapest point, where the strictly cheaper baseline is entropy at $T = 4$ rather than EB-Sampler, $\hat{D}$ at $T = 2$ is worse on 15 problems and better on 6 for a difference of -0.113, and $P = 0.078$. The comparator is chosen by maximising accuracy over the baseline points that qualify, on the same problems the test then uses, which biases

the comparison against the probe rules; we report both readings and claim only what survives the strict one.

A second task. We repeat the sweep on HumanEval (Chen et al., 2021), 164 Python function-completion problems graded by their own unit tests. The model continues the task's prompt, the function's signature and docstring, as raw text; the continuation is cut at the first line that opens a new top-level definition, a print, an assertion, a comment or a code fence, and prompt plus completion is run against the task's tests in a separate process with a 10-second limit, which gives pass@1 under greedy decoding. Decoding, rules and knobs are those above, except that PUNT's costliest setting is not run. Table 5 and Figure 5 give every operating point. The probe rules lose to the samplers again, though here not to random order: $\hat{D}$ reaches 0.049, 0.079 and 0.110 at 48, 96 and 192 passes, against 0.220 for Fast-dLLM at 37.8 and 0.421 and 0.427 for EB-Sampler at 93.9 and 125.0, the best baselines costing no more, for differences of -0.17, -0.34 and -0.32; under the strict reading they are -0.06, -0.14 and -0.32, and all six strict comparisons of the two probe rules give $P < 0.05$ by the exact McNemar test ($P = 0.031$ for both at $T = 2$ and $P < 10^{-4}$ for the other four). The two probe rules are again not resolved from each other ($\hat{D}$ minus $\hat{I}_{\text{cor}}$ between -0.006 and $+0.012$). The gates match one-token-per-step confidence decoding at about a third of its compute, as they do on Dream-7B's GSM8K at a quarter: EB-Sampler at $\gamma = 0.4$ reaches 0.421 at 94 passes against 0.390 at 256, a paired $+0.030$ $[-0.025, +0.091]$, and Fast-dLLM at $\tau = 0.8$ the same 0.390 at 77.

On Dream-7B-Instruct, over the first 60 problems (Table 6), the probe rules again lose to the samplers: $\hat{D}$ reaches 0.050, 0.133 and 0.317 at 48, 96 and 192 passes against 0.433 for Fast-dLLM at 46.5 and 0.717 for EB-Sampler at 93.4, differences of -0.38, -0.58 and -0.42 from the best baseline costing no more. Under the strict reading the six comparisons of the two probe rules run from -0.08 to -0.57, and four give $P < 0.05$; the exceptions are both rules at $T = 2$ ($P = 0.065$ and 0.23), where the strictly cheaper comparator is confidence decoding at 32 passes. At $T = 8$, $\hat{D}$ beats $\hat{I}_{\text{cor}}$ by $+0.167$ $[+0.050, +0.300]$. Confidence decoding reaches the best accuracy, 0.733, only at 256 passes; Fast-dLLM and EB-Sampler match it at 124 and 113.

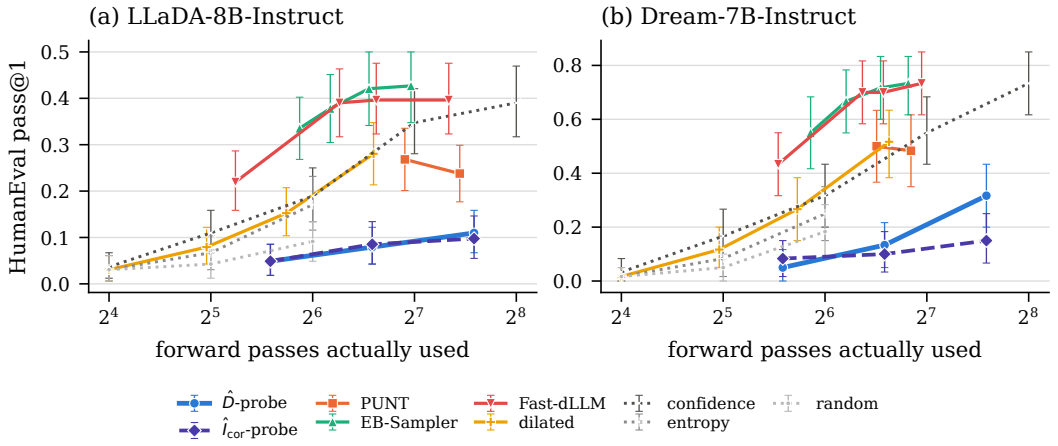

Figure 5: HumanEval pass@1 against the forward passes each rule actually used, every operating point of Tables 5 and 6: (a) LLaDA-8B-Instruct on all 164 problems, (b) Dream-7B-Instruct on the first 60, with 95% intervals over problems. The two probe rules, which fill a step by the score after a confidence-chosen anchor, run along the bottom.

O MATCHED-BUDGET INFILLING

Figure 6 plots the result. We mask a contiguous block of 32 tokens of real text and fill it under five selection rules at identical total forward-pass budgets, scoring the recovered tokens against the truth: random, confidence, entropy, and two probe rules that spend two extra passes per step on the most confident masked position's top-two tokens and then unmask the positions with the lowest $\hat{D}$ or the lowest $\hat{I}_{\text{cor}}$ relative to that anchor. Table 10 gives every cell. At a budget $B \geq 4$ a probe

Table 3: Every operating point of Figure 2: 120 GSM8K problems, LLaDA-8B-Instruct, block 32, generation length 256. "passes" counts every forward the rule issued. Brackets are 95% bootstrap intervals over problems. Bold marks the best accuracy overall.

rule	setting	passes	accuracy
random	$T = 2$	16.0	0.100 [0.050,0.158]
	$T = 4$	32.0	0.267 [0.183,0.350]
	$T = 8$	64.0	0.583 [0.492,0.667]
confidence	$T = 2$	16.0	0.108 [0.058,0.167]
	$T = 4$	32.0	0.400 [0.316,0.492]
	$T = 8$	64.0	0.692 [0.600,0.775]
	$T = 16$	128.0	0.750 [0.675,0.825]
	$T = 32$	256.0	0.717 [0.642,0.792]
entropy	$T = 2$	16.0	0.117 [0.058,0.175]
	$T = 4$	32.0	0.267 [0.192,0.350]
	$T = 8$	64.0	0.708 [0.625,0.783]
dilated (Luxembourg et al., 2026)	$g_0 = 2$	16.0	0.150 [0.091,0.217]
	$g_0 = 4$	30.8	0.483 [0.392,0.575]
	$g_0 = 8$	53.1	0.642 [0.558,0.725]
	$g_0 = 16$	97.9	0.700 [0.617,0.775]
Fast-dLLM (Wu et al., 2026)	$\tau = 0.5$	34.0	0.608 [0.525,0.700]
	$\tau = 0.8$	58.8	0.808 [0.733,0.875]
	$\tau = 0.9$	78.2	0.767 [0.692,0.833]
	$\tau = 0.99$	127.5	0.725 [0.650,0.800]
EB-Sampler (Ben-Hamu et al., 2025)	$\gamma = 1.6$	48.1	0.775 [0.700,0.842]
	$\gamma = 0.8$	57.4	0.792 [0.725,0.858]
	$\gamma = 0.4$	68.8	0.808 [0.733,0.875]
	$\gamma = 0.1$	92.9	**0.825** [0.750,0.892]
PUNT (Azangulov et al., 2026)	$\epsilon = 0.5$	134.3	0.658 [0.575,0.742]
	$\epsilon = 0.2$	179.4	0.750 [0.667,0.825]
	$\epsilon = 0.05$	291.6	0.750 [0.667,0.825]
$\hat{D}$-probe	$T = 2$	48.0	0.083 [0.033,0.133]
	$T = 4$	96.0	0.100 [0.050,0.158]
	$T = 8$	192.0	0.425 [0.333,0.517]
$\hat{I}_{\mathrm{cor}}$-probe	$T = 2$	48.0	0.100 [0.050,0.158]
	$T = 4$	96.0	0.083 [0.033,0.133]
	$T = 8$	192.0	0.358 [0.275,0.442]

rule takes $\lfloor B/3 \rfloor$ nominal steps, fewer when the block empties early (24 passes at $B = 32$), and therefore spends at most the baseline's passes; at $B = 1$ and $B = 2$ it cannot, and spends 3 against the baseline's 1 and 2, which is why those two cells are excluded from the matched comparison.

Read at matched budget, the probe rules appear to win everywhere: on MDLM-OWT (300 documents) $\hat{D}$ beats confidence by $+0.024$ $[+0.018, +0.029]$ accuracy at budget 4 and $+0.020$ $[+0.012, +0.028]$ at 8; on LLaDA-8B-Instruct (100 documents) by $+0.024$ $[+0.016, +0.034]$ and $+0.013$ $[+0.001, +0.024]$; on Dream-7B (200 documents) by $+0.012$ $[+0.006, +0.018]$ and $+0.010$ $[+0.001, +0.019]$; on that model the true-token negative log-likelihood improves by -0.638 $[-0.818, -0.464]$ and -1.463 $[-1.672, -1.246]$ nats. Each of these eight intervals excludes zero. **That reading is not sound, and we do not use it**: the budget is a pass count, a probe rule spends part of it on the probe and so takes fewer, larger steps, and in this testbed accuracy *falls* with budget for every rule, apart from small reversals (confidence runs .103 → .085 → .080 → .081 → .076 across budgets 2 to 32 on MDLM). The budget-1 control settles it: there every rule commits the whole block in one step and makes no selection at all, and all five agree exactly — 0.1084 on MDLM-OWT, 0.1306 on LLaDA-8B-Instruct, 0.1297 on Dream-7B. Those are the numbers the probe rules reach at budget 4, for three forward passes instead of one: the matched-budget "win" is one-step decoding bought at triple price.

The comparison that does isolate the rule is the frontier over passes actually spent, against the best baseline costing no more. On that comparison the experiment supports nothing: on MDLM the best probe point is 0.108 at 3 passes, which confidence matches at 1, and every probe point above

Table 4: The same sweep on Dream-7B-Instruct, over the first 80 of the same problems with the decoding configuration, the knob grid and the grading unchanged. "passes" counts every forward the rule issued. Brackets are 95% bootstrap intervals over problems. Bold marks the best accuracy overall.

rule	setting	passes	accuracy
random	$T = 2$	16.0	0.050 [0.013,0.100]
	$T = 4$	32.0	0.125 [0.050,0.200]
	$T = 8$	64.0	0.362 [0.263,0.475]
confidence	$T = 2$	16.0	0.113 [0.050,0.188]
	$T = 4$	32.0	0.225 [0.138,0.312]
	$T = 8$	64.0	0.588 [0.487,0.700]
	$T = 16$	128.0	0.787 [0.700,0.875]
	$T = 32$	256.0	0.825 [0.738,0.912]
entropy	$T = 2$	16.0	0.075 [0.025,0.138]
	$T = 4$	32.0	0.237 [0.150,0.338]
	$T = 8$	64.0	0.500 [0.400,0.613]
dilated (Luxembourg et al., 2026)	$g_0 = 2$	16.0	0.075 [0.025,0.138]
	$g_0 = 4$	30.9	0.225 [0.138,0.325]
	$g_0 = 8$	52.7	0.537 [0.425,0.637]
	$g_0 = 16$	97.4	0.812 [0.725,0.900]
Fast-dLLM (Wu et al., 2026)	$\tau = 0.5$	30.0	0.537 [0.425,0.650]
	$\tau = 0.8$	51.2	0.812 [0.725,0.900]
	$\tau = 0.9$	64.1	0.825 [0.738,0.912]
	$\tau = 0.99$	99.3	0.825 [0.738,0.912]
EB-Sampler (Ben-Hamu et al., 2025)	$\gamma = 1.6$	39.4	0.762 [0.662,0.850]
	$\gamma = 0.8$	48.6	0.725 [0.625,0.825]
	$\gamma = 0.4$	56.5	**0.838** [0.750,0.912]
	$\gamma = 0.1$	77.5	0.812 [0.725,0.900]
PUNT (Azangulov et al., 2026)	$\epsilon = 0.5$	109.2	0.625 [0.512,0.725]
	$\epsilon = 0.2$	141.3	0.800 [0.713,0.887]
	$\epsilon = 0.05$	223.7	0.750 [0.650,0.838]
$\hat{D}$-probe	$T = 2$	48.0	0.125 [0.050,0.200]
	$T = 4$	96.0	0.188 [0.100,0.275]
	$T = 8$	192.0	0.287 [0.188,0.388]
$\hat{I}_{\mathrm{cor}}$-probe	$T = 2$	48.0	0.100 [0.037,0.175]
	$T = 4$	96.0	0.138 [0.062,0.212]
	$T = 8$	192.0	0.212 [0.125,0.300]

that loses (-0.0082 $[-0.0119, -0.0047]$ at 6 passes, -0.0141 $[-0.0198, -0.0081]$ at 15, -0.0270 $[-0.0335, -0.0206]$ at 24); on LLaDA-8B-Instruct no multi-step probe cell is positive, and on Dream-7B every paired difference has an interval crossing zero. We report the experiment and its failure mode, and take the decoding evidence in §5 from the controlled selection test on realised gain and from the GSM8K frontier instead, where compute is counted per forward pass and every rule's knob is swept.

Table 5: The same sweep on HumanEval (Chen et al., 2021) (164 problems, pass@1; the task's prompt is continued and graded by its unit tests), LLaDA-8B-Instruct, block 32, generation length 256. PUNT at $\epsilon = 0.05$, which on GSM8K costs more than one-token decoding, is not run. "passes" counts every forward the rule issued. Brackets are 95% bootstrap intervals over problems. Bold marks the best accuracy overall.

rule	setting	passes	pass@1
random	$T = 2$	16.0	0.030 [0.006,0.061]
	$T = 4$	32.0	0.043 [0.012,0.073]
	$T = 8$	64.0	0.091 [0.049,0.134]
confidence	$T = 2$	16.0	0.037 [0.012,0.067]
	$T = 4$	32.0	0.110 [0.067,0.159]
	$T = 8$	64.0	0.189 [0.134,0.250]
	$T = 16$	128.0	0.348 [0.280,0.421]
	$T = 32$	256.0	0.390 [0.317,0.470]
entropy	$T = 2$	16.0	0.030 [0.006,0.061]
	$T = 4$	32.0	0.067 [0.030,0.104]
	$T = 8$	64.0	0.171 [0.116,0.232]
dilated (Luxembourg et al., 2026)	$g_0 = 2$	16.0	0.030 [0.006,0.061]
	$g_0 = 4$	31.1	0.079 [0.043,0.122]
	$g_0 = 8$	53.5	0.152 [0.104,0.207]
	$g_0 = 16$	96.9	0.280 [0.213,0.348]
Fast-dLLM (Wu et al., 2026)	$\tau = 0.5$	37.8	0.220 [0.159,0.287]
	$\tau = 0.8$	76.8	0.390 [0.317,0.463]
	$\tau = 0.9$	98.7	0.396 [0.323,0.476]
	$\tau = 0.99$	161.6	0.396 [0.323,0.476]
EB-Sampler (Ben-Hamu et al., 2025)	$\gamma = 1.6$	58.6	0.335 [0.268,0.402]
	$\gamma = 0.8$	72.1	0.378 [0.305,0.451]
	$\gamma = 0.4$	93.9	0.421 [0.341,0.500]
	$\gamma = 0.1$	125.0	**0.427** [0.348,0.500]
PUNT (Azangulov et al., 2026)	$\epsilon = 0.5$	119.9	0.268 [0.201,0.335]
	$\epsilon = 0.2$	174.2	0.238 [0.177,0.299]
$\hat{D}$-probe	$T = 2$	48.0	0.049 [0.018,0.085]
	$T = 4$	96.0	0.079 [0.043,0.122]
	$T = 8$	192.0	0.110 [0.067,0.159]
$\hat{I}_{\mathrm{cor}}$-probe	$T = 2$	48.0	0.049 [0.018,0.085]
	$T = 4$	96.0	0.085 [0.043,0.134]
	$T = 8$	192.0	0.098 [0.055,0.146]

Table 6: The HumanEval sweep of Table 5 on Dream-7B-Instruct, over the first 60 problems, with the same prompt format, grading, decoding configuration and knob grid. "passes" counts every forward the rule issued. Brackets are 95% bootstrap intervals over problems. Bold marks the best accuracy overall.

rule	setting	passes	pass@1
random	$T = 2$	16.0	0.017 [0.000,0.050]
	$T = 4$	32.0	0.050 [0.000,0.100]
	$T = 8$	64.0	0.183 [0.083,0.283]
confidence	$T = 2$	16.0	0.033 [0.000,0.083]
	$T = 4$	32.0	0.167 [0.083,0.267]
	$T = 8$	64.0	0.317 [0.200,0.433]
	$T = 16$	128.0	0.550 [0.433,0.683]
	$T = 32$	256.0	**0.733** [0.617,0.850]
entropy	$T = 2$	16.0	0.017 [0.000,0.050]
	$T = 4$	32.0	0.083 [0.017,0.167]
	$T = 8$	64.0	0.250 [0.150,0.350]
dilated (Luxembourg et al., 2026)	$g_0 = 2$	16.0	0.017 [0.000,0.050]
	$g_0 = 4$	31.1	0.117 [0.050,0.200]
	$g_0 = 8$	53.0	0.267 [0.150,0.383]
	$g_0 = 16$	99.0	0.517 [0.383,0.634]
Fast-dLLM (Wu et al., 2026)	$\tau = 0.5$	46.5	0.433 [0.317,0.550]
	$\tau = 0.8$	82.5	0.700 [0.583,0.817]
	$\tau = 0.9$	95.2	0.700 [0.583,0.817]
	$\tau = 0.99$	123.5	**0.733** [0.617,0.850]
EB-Sampler (Ben-Hamu et al., 2025)	$\gamma = 1.6$	58.0	0.550 [0.417,0.683]
	$\gamma = 0.8$	73.9	0.667 [0.550,0.783]
	$\gamma = 0.4$	93.4	0.717 [0.600,0.833]
	$\gamma = 0.1$	112.8	**0.733** [0.617,0.833]
PUNT (Azangulov et al., 2026)	$\epsilon = 0.5$	90.9	0.500 [0.367,0.633]
	$\epsilon = 0.2$	114.9	0.483 [0.350,0.617]
$\hat{D}$-probe	$T = 2$	48.0	0.050 [0.000,0.117]
	$T = 4$	96.0	0.133 [0.050,0.217]
	$T = 8$	192.0	0.317 [0.200,0.433]
$\hat{I}_{\mathrm{cor}}$-probe	$T = 2$	48.0	0.083 [0.017,0.150]
	$T = 4$	96.0	0.100 [0.033,0.183]
	$T = 8$	192.0	0.150 [0.067,0.250]

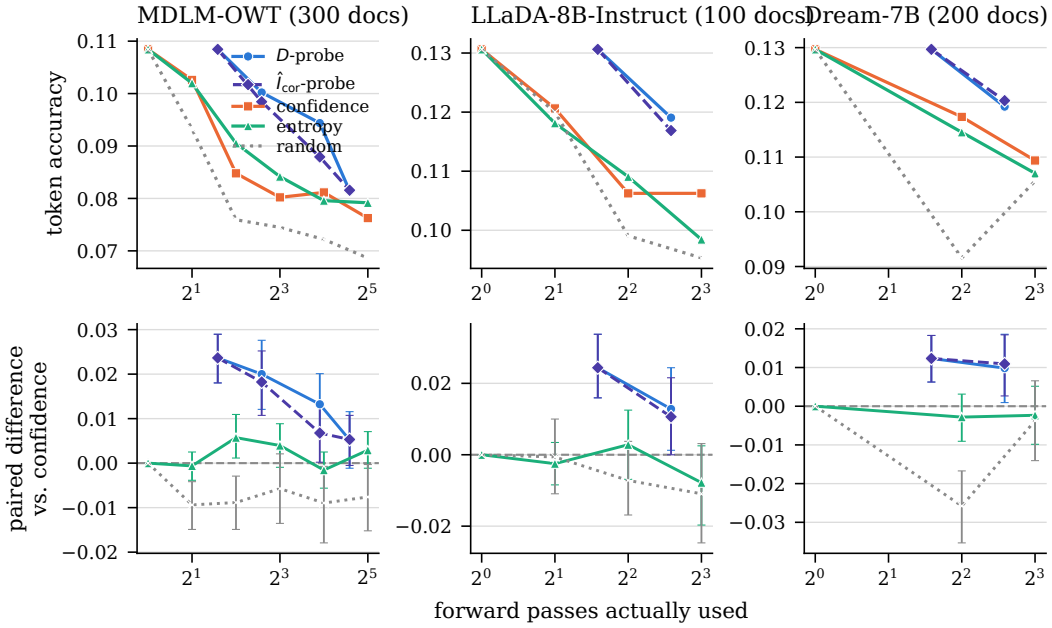

Figure 6: Matched-compute infilling of a 32-token block, scored against the true tokens. Top: accuracy against the number of forward passes each rule actually used; probe rules spend three passes per step and therefore take fewer, larger steps. Bottom: the paired difference to the confidence rule at the same nominal budget, with 95% bootstrap intervals — the matched-budget reading that Appendix O explains is unsound, shown so the failure mode is visible. Cells in which a probe rule would have spent more passes than its baseline are excluded.

Table 7: Per-model detail for the pair probe, on the same pairs as Table 1. The first two rows describe how widespread the defect and the asymmetry are; the last five are the regression and ranking diagnostics (confidence features: maximum probability and entropy of anchor and target) of §4. The last row is the fraction of the third of each row's candidates with the *lowest* score, the set a decoder would commit, on which $\hat{D}$ and $\hat{I}_{\mathrm{cor}}$ disagree; over the highest-scoring third instead they disagree on 13.6%, 17.0%, 16.8% and 16.3%.

	MDLM-OWT 110M	LLaDA-8B-Instruct	LLaDA-8B-Base	Dream-7B
pairs with $\rho > 0.1$	91.1%	94.7%	93.1%	93.8%
pairs with asymmetry > 0.5	64.9%	62.8%	62.2%	61.8%
R^2 of $\log \hat{\Delta}$ on confidence features	0.027	0.096	0.100	0.066
the same, $+ \log$ distance	0.246	0.300	0.340	0.301
$\log \hat{D}$ alone	0.794	0.796	0.767	0.777
Kendall τ between $\hat{D}$ and $\hat{I}_{\mathrm{cor}}$	0.858	0.829	0.829	0.810
safest-third selections that differ	17.3%	20.9%	20.1%	24.8%

P Additional tables

The share by the anchor's own entropy. §4 bins anchors by top-K head mass, a proxy for the variable of Proposition 3(C). The quartile edges are the empirical quartiles of the rows' head mass, which round to 0.80, 0.90, 0.95, 0.98 and 1.00 on all three models, over the 2,203, 1,768 and 456 rows with at least one pair above 10^{-3} nats (of Table 1's 2,224, 1,774 and 456), giving 551, 442 and 114 anchor rows per quartile on MDLM-OWT, LLaDA-8B-Instruct and Dream-7B (a row on a shared edge counts in both); the shares quoted there are jackknifed medians over the quartile's pairs (14,600 to 7,976, 5,996 to 3,555 and 1,533 to 994 from the flattest quartile to the sharpest), over the same quartiles the median $\hat{D}$ falls 5.0×, 7.8× and 5.8× from the flattest to the sharpest and the median $\hat{\Delta}$ only 1.1×, 2.1× and 1.7×; certificate (C) is non-vacuous on 4.2%, 2.1% and 1.7% of the 46,863, 19,736 and 5,160 pairs and violated on none of them; the median over rows of each row's median runs $0.22 \to 0.33 \to 0.57 \to 0.96$, $0.25 \to 0.41 \to 0.68 \to 0.97$ and $0.24 \to 0.43 \to 0.68 \to 0.97$, and the Spearman correlation is between a row's head mass and its median share. The adaptive head is truncated once it carries 0.95 of the mass, so the top two quartiles are partly defined by that stopping rule. Binned instead by $H(p_i)$ itself, with edges 0, 0.05, 0.2, 0.5, 1, 2 and 3 nats, the median share over all pairs with $\hat{D} > 0$ falls monotonically on each of the three models, from 0.987, 0.990 and 0.994 below 0.05 nats to 0.198, 0.219 and 0.214 above 3 nats (MDLM-OWT, LLaDA-8B-Instruct, Dream-7B). Restricted to Table 1's $\hat{D} \geq 10^{-3}$ the end points are 0.993, 0.994, 0.997 and 0.191, 0.215, 0.209, so the inclusion rule is not what produces the rise; on MDLM-OWT, in float32, neither is the numerical floor (`code/analyze_share_robustness.py`).

Q Limitations in full

§7 names the main boundaries; the remaining ones are these. *Estimation.* Table 1's shares and lower bounds and the placebo ratios resample documents; other intervals resample within one run (anchor rows, pairs or set–order cells), so the asymmetry row's and Table 8's are narrower than a document-level interval would be, and no interval covers a change of protocol such as the mask-ratio mix. The median shares use pairs selected on $\hat{D}$ from the same draws as their numerator, which the unselected mean certificate and mass-weighted share of Table 1 avoid and the toy and synthetic checks of Appendix I probe without settling. The budget replication covers 138 LLaDA-8B-Instruct pairs, where its median shift of -0.025 is not resolved from zero. The k-sweep draws positions without the head-mass filter, so its set-level shares describe a different population from the pair probe's, and the top two peakedness quartiles are partly set by the adaptive head's stopping rule. The replicate-noise null for the I_{ent} tail is measured on MDLM-OWT, LLaDA-8B-Instruct and Dream-7B; Dream-7B's is noisy enough that only its -0.2 tail clears it, LLaDA-8B-Base has none, on the bf16 models the per-pair sign of I_{ent} depends on precision, and the bf16 nulls see reseed noise but not the arithmetic both runs share, which only LLaDA-8B-Instruct's float32 run checks. The asymmetry is measured only on anchor–anchor pairs, both ends peaked. Per-pair fractions on the bf16 models (Table 7) include pairs within 10× the floor. The in-distribution probe masks reference

Table 8: Sets of k positions. Columns 2–4: one-sided 95% lower bounds on $\sum_{i\in S}\Delta_\pi(i)$ in nats on *scattered* sets; a bound above the certification threshold of Appendix F (10^{-5}, 4.9×10^{-3}, 8.1×10^{-3} nats) certifies a strictly positive defect. Every entry clears its threshold; the two $k=2$ cells on the 7–8B models clear it by only $1.9\times$ (†) and $1.1\times$ (underlined) and are the marginal ones. Contiguous blocks are certified only at $k\le3$ on MDLM-OWT and at $k=2$ on LLaDA-8B-Instruct (§4.1). Column 5: the share $\rho=\sum_i\Delta_\pi(i)/\mathrm{KL}(q_\pi\|m_S)$ on MDLM-OWT, mass-weighted over all cells of both structures and all four orders, with its 95% interval. Columns 2–4 are split-sample bounds; column 5 uses the jackknifed numerator and is an estimate, not a bound. On LLaDA-8B-Instruct it is .179, .164, .165, .169, .173, .176, .183 and on Dream-7B .150, .152, .152, .154, .155, .160, .164 from $k=2$ to $k=8$, so the range over all three models and all k is .150–.209.

k	MDLM-OWT (128 rollouts)	LLaDA-8B-Instruct (128)	Dream-7B (128)	ρ on MDLM-OWT
2	0.0106	0.0092†	<u>0.0089</u>	.209 [.159,.286]
3	0.033	0.025	0.020	.192 [.165,.230]
4	0.076	0.046	0.038	.192 [.171,.216]
5	0.117	0.076	0.056	.194 [.174,.219]
6	0.188	0.105	0.081	.195 [.178,.214]
7	0.228	0.143	0.107	.193 [.180,.209]
8	0.287	0.200	0.140	.199 [.186,.215]

Table 9: Realised-gain experiment: means over probed pairs, and the selection test on probe rows (mean $|G|$ forgone by the safest third; lower is better). The experiment reuses the pair probe's anchors, and covers all of them on MDLM-OWT and Dream-7B; on LLaDA-8B-Instruct it covers 1,742 of the 1,774, the 32 it misses lying in two documents, one of which it misses entirely, so its pair count is 26,130 against Table 1's 26,610. The paired difference in forgone $|G|$, $\hat I_{\mathrm{cor}}$ minus $\hat D$, is $+0.0045$ nats $[+0.0016,+0.0083]$ on MDLM-OWT over 2,224 rows, $+0.0052$ $[+0.0022,+0.0081]$ on LLaDA-8B-Instruct over 1,742 and $+0.0065$ $[+0.0035,+0.0097]$ on Dream-7B over 456, each with $P(\hat I_{\mathrm{cor}}$ better$)<0.001$; resampling documents rather than rows gives $[+0.0014,+0.0082]$, $[+0.0021,+0.0086]$ and $[+0.0038,+0.0095]$ over 70, 74 and 20 documents, $P\le0.001$ on each. The oracle forgoes a further 66–77% less than $\hat D$.

	MDLM-OWT	LLaDA-8B-Instruct	Dream-7B
pairs with a realised-gain measurement	68,944	26,130	6,840
mean $\hat D$ / $\hat I_{\mathrm{cor}}$ / $\hat\Delta$ / G (nats)	.103 / .054 / .049 / .078	.089 / .060 / .029 / .094	.080 / .053 / .027 / .062
$\hat D/G$	1.32	0.95	1.29
forgone $\|G\|$: random	0.205	0.191	0.150
confidence / entropy	0.179 / 0.178	0.134 / 0.129	0.099 / 0.096
$\hat D$ / $\hat I_{\mathrm{cor}}$	**0.076** / 0.081	**0.062** / 0.068	**0.050** / 0.056
I_{ent}	0.199	0.168	0.148
oracle	0.018	0.019	0.017
$\hat I_{\mathrm{cor}}-\hat D$, paired	+.0045 [+.0016,+.0083]	+.0052 [+.0022,+.0081]	+.0065 [+.0035,+.0097]

solutions, not the states of a real decode. The count-matched placebo withholds information as well as matching the count, so it is not a clean null for mask-count sensitivity (Appendix E). Dream-7B has no float32 control, and its $k=2$ certificate is marginal and fails under the stricter floor convention of Appendix F. The slot-alignment check covers MDLM-OWT, LLaDA-8B-Instruct and Dream-7B but not LLaDA-8B-Base. *Set level.* The parallel error is measured against the model's own order-π rollout, the sequential decode a parallel step replaces, not against the data; on the toy neither marginal is the more accurate pair by pair (Appendix I), and on the real models the rollout's is the more accurate only on average (Appendix E). The certificates bound a mean over sets, not each set. The order-spread null assumes the four orders' noise is exchangeable and cannot see an order-dependent bias that both runs share in $\mathrm{TC}(q_\pi)$ and the defect sum, and the order that minimizes the parallel error is not resolved (Appendix H). Contiguous blocks are certified only at small k, and on Dream-7B not at all. LLaDA-8B-Base is not in the k-sweep. *Decoding.* The in-decode shares of Appendix D come from 20 problems per gate and model, at 128 rollouts at $\gamma=0.4$ and 32 otherwise, with sets capped at 8 positions and a named reference order; the selection test of Table 9 runs on probe rows rather than decoding states and omits LLaDA-8B-Base; its $|G|$ criterion is not independent of D; the probe rules rank by a two-point estimate; the budget correction charges a pairwise proxy for $\sum_i\Delta_\pi(i)$; decoding

Table 10: Matched-budget infilling, all cells, from the released records. Accuracy is exact-match on the 32 masked tokens. At budget 1 every rule commits the whole block in one step, so the five rules must agree exactly, and do. "passes" is the number of forward passes actually issued at budgets 4 and 8; a probe rule spends three per step and so takes fewer, larger steps at the same budget, which is the confound Appendix O describes.

model	rule	budget 1	2	4	8	16	32	passes at 4 / 8
MDLM-OWT	confidence	**.108**	.103	.085	.080	.081	.076	4 / 8
(300 docs)	entropy	**.108**	.102	.091	.084	.080	.079	4 / 8
	random	**.108**	.093	.076	.074	.072	.069	4 / 8
	$\hat{D}$-probe	**.108**	**.108**	**.108**	**.100**	**.094**	.081	3 / 6
	$\hat{I}_{\mathrm{cor}}$-probe	**.108**	**.108**	**.108**	.098	.088	**.082**	3 / 6
LLaDA-8B-Instruct	confidence	**.131**	.121	.106	.106			4 / 8
(100 docs)	entropy	**.131**	.118	.109	.098			4 / 8
	random	**.131**	.120	.099	.095			4 / 8
	$\hat{D}$-probe	**.131**	**.131**	**.131**	**.119**			3 / 6
	$\hat{I}_{\mathrm{cor}}$-probe	**.131**	**.131**	**.131**	.117			3 / 6
Dream-7B	confidence	**.130**		.117	.109			4 / 8
(200 docs)	entropy	**.130**		.115	.107			4 / 8
	random	**.130**		.092	.106			4 / 8
	$\hat{D}$-probe	**.130**		**.130**	.119			3 / 6
	$\hat{I}_{\mathrm{cor}}$-probe	**.130**		**.130**	**.120**			3 / 6

is greedy, while Proposition 1 describes sampling, on the first 120 (80 on Dream-7B) GSM8K test problems and 164 (60) HumanEval problems, at one length and block size; and the strict comparator is chosen on the problems it is tested on. The matched-budget infilling experiment was uninformative (Appendix O). *Toy.* The toy model has $|V| = 6$ and $L = 6$, the capacity sweep has one model per setting, and nothing there is generalised beyond what the real models show.