MEMORIZATION DYNAMICS IN MASKED DIFFUSION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work on continuous diffusion models suggests that generalization and memorization can emerge on distinct training timescales, with population-level structure learned before individual training examples are memorized. In parallel, masked diffusion language models (DLMs) have become competitive language models while replacing autoregressive generation with iterative denoising. This raises a natural question: *does population learning also precede memorization in masked DLMs, or can the ordering change with the training regime?* We study this question using controlled Markov and hierarchical-grammar populations, trained canaries, never-trained controls, and Bayes-referenced population denoising. Contrary to a universal generalization-before-memorization picture, the relative timing of population learning and training-record extraction depends on the training regime. At low repetition, population learning can precede training-record extraction by thousands of optimization steps, whereas increasing repetition reverses the ordering and causes extraction to emerge first; in a contrasting Markov regime, extraction appears without an observed population-learning phase over the tested horizon. Across joint dataset-size and repetition interventions, normalized record frequency strongly organizes these dynamics, but controlled decoupling experiments show that it is not sufficient alone: both the sampling frequency of an individual repeated record and the aggregate training mass assigned to repeated records can alter the ordering. Mask-rate distribution shape further changes extraction onset even when its first two moments are matched, while pretrained LLaDA and MDLM reproduce the directional effect of repetition. Together, our results show that population learning and train-specific extraction are distinct phenomena whose relative timing is regime-dependent rather than universally ordered.

1 INTRODUCTION

Learning population structure and reproducing individual training records are distinct capabilities, yet both can emerge from the same training objective. The relevant question is therefore not only whether a model eventually memorizes, but *when* train-specific information becomes recoverable relative to population learning. Understanding this timing is also relevant to privacy auditing and machine unlearning, where the conditions under which individual records become extractable may matter.

Continuous-diffusion analyses provide a natural starting hypothesis. In several studied regimes, useful generation emerges before substantial memorization, with memorization delayed as dataset size grows (Bonnaire et al., 2025; Favero et al., 2025). Masked diffusion language models (DLMs), however, are trained through conditional denoising problems whose structure depends both on which records are sampled and on how tokens are masked. It is therefore unclear whether the same generalization-before-memorization ordering should persist.

We ask:

> *Does population learning precede training-record extraction in masked DLMs, and what training factors determine their relative timing?*

We study this question using controlled Markov and hierarchical-grammar populations, trained canaries, never-trained controls, and Bayes-referenced population denoising. This lets us separate

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.

population-level prediction from train-specific extraction rather than equating successful reconstruction with memorization.

Our central finding is a regime-dependent reversal. In the hierarchical setting, population learning precedes extraction at low canary repetition, whereas increasing repetition makes extraction emerge first. A contrasting Markov regime exhibits extraction without an observed meaningful population-learning phase over an extended training horizon. We then ask what organizes this reversal. Across a joint dataset-size and repetition sweep, normalized record frequency strongly structures the observed ordering. Controlled decoupling experiments go further: neither the sampling probability of an individual repeated record nor the aggregate training mass assigned to repeated records is sufficient alone, and varying either while holding the other fixed can reverse the ordering.

Masking provides a complementary DLM-specific intervention. Two training mask-rate distributions matched in mean, variance, and support still yield different extraction onsets, showing that their first two moments are insufficient to determine the dynamics. Pretrained LLaDA and MDLM experiments reproduce the directional repetition effect, while mask-schedule transfer is less consistent. Finally, we derive an exact finite-sample characterization of context uniqueness and find that it adds essentially no predictive value for neural extraction beyond reveal fraction under our tested criterion.

Contributions.

1. **Regime-dependent learning extraction ordering.** We demonstrate both population-learning-before-extraction and extraction-before-population-learning regimes, including a repetition-driven reversal within the same controlled model and data family.
2. **Controlled analysis of what governs the reversal.** A joint dataset-size repetition study shows that normalized record frequency strongly organizes the ordering, while controlled decoupling experiments show that both individual-record sampling probability and aggregate repeated-data mass matter.
3. **Masking and pretrained-model stress tests.** Matched-moment mask distributions produce different extraction onsets, and pretrained LLaDA and MDLM reproduce the directional repetition effect, while schedule transfer is mixed.
4. **Exact conditional analysis.** We characterize population and empirical masked-token conditionals and derive an exact finite-sample context-uniqueness result, while finding no incremental neural predictive value beyond reveal fraction under the tested criterion.

2 Related work

Memorization and training-data extraction. Training-data extraction has been demonstrated in autoregressive language models (Carlini et al., 2021) and image diffusion models (Carlini et al., 2023), with duplication known to influence memorization and copying (Somepalli et al., 2023). Extraction, however, depends on the supplied context, decoding procedure, and recovery criterion, so successful reconstruction need not by itself imply train specific memorization. We address this distinction using trained canaries, never-trained matched controls, and a separate measure of population denoising.

Generalization and memorization during diffusion training. (Bonnaire et al., 2025) relate delayed memorization in continuous diffusion to implicit dynamical regularization, with memorization emerging on a later timescale that grows with dataset size in their studied regimes. (Favero et al., 2025) analyze the competition between generalization and memorization in image and language diffusion, including hierarchical-grammar settings and early stopping. Our focus is different: we study whether the relative onset of population learning and training-record extraction is preserved across training regimes, and how repetition and masking interventions can alter that ordering.

Discrete diffusion and conditional prediction. Discrete diffusion models define corruption processes over categorical states (Austin et al., 2021), with masked and score-based formulations enabling language modeling (Lou et al., 2024; Sahoo et al., 2024). (Ou et al., 2025) connect absorbing diffusion objectives to clean-data conditionals, while LLaDA scales masked diffusion to large language models (Nie et al., 2025). This conditional-prediction view motivates our distinction between population

conditionals and the empirical conditionals induced by a finite training sample, as well as our analysis of when a revealed context uniquely identifies a training completion.

Extraction and associative memory in diffusion language models. Recent work has adapted training-data extraction to the bidirectional conditioning available in DLMs. (Luo et al., 2026) study generalized extraction under arbitrary masking patterns and sampling resolutions; (Wang & Asokan, 2026) consider infilling attacks beyond prefix-only extraction; and . (Pham et al., 2026) analyze discrete diffusion through an associative-memory perspective. These works primarily characterize *whether* and *how* information can be recovered from trained DLMs. We instead study *when* train-specific extraction emerges during optimization relative to population learning, and how its timing changes with record frequency, training-mixture composition, and the mask-rate distribution.

3 Setup and metrics

Data and masking. Let P_0 be a distribution over sequences in $\mathcal{V}^L$. A training set contains n background sequences and K canaries, each repeated R times; matched controls are never included in training. Under uniform sampling from the resulting multiset, each background record has probability $1/(n+KR)$, while a particular canary identity has sampling probability

$$p_{\text{can}} = \frac{R}{n+KR}. \tag{1}$$

The aggregate training mass assigned to all canaries is

$$f_{\text{can}} = \frac{KR}{n+KR} = Kp_{\text{can}}. \tag{2}$$

Thus changing repetition generally alters both the frequency of an individual record and the overall canary/background mixture. Later interventions vary K to probe these effects separately. The case $R = 0$ serves as an untrained-target control.

For each sequence, a masking rate $t \sim \pi$ determines independent Bernoulli masks m_i, producing the corrupted sequence X_t. Padding positions are excluded from masking targets and loss computation. Training-time masking and extraction-time conditioning are specified separately.

Population and empirical objectives. We consider the masked-denoising population risk

$$\mathcal{L}(\theta) = \mathbb{E}_{X\sim P_0,\, t\sim\pi,\, m|t}\left[\frac{1}{t}\sum_{i=1}^{L} m_i\big(-\log p_\theta(X_i \mid X_t, t)\big)\right]. \tag{3}$$

Theorem 1 (Bayes-optimal masked denoiser)**.** *Assume independent masking conditional on t, with the masked status of position i observable from X_t. For every position i, an unrestricted minimizer of Eq. (3) satisfies*

$$p^*(X_i = \cdot \mid X_t, t) = P_0(X_i = \cdot \mid X_t, t)$$

for $P(X_t, t, m_i = 1)$-almost every positive-weight masked context. Predictions on contexts receiving zero objective weight are unconstrained.

Proposition 1 (Empirical conditional)**.** *Replacing P_0 by the empirical training distribution yields the corresponding empirical conditional as the unrestricted minimizer, including the weights induced by repeated records.*

Both statements follow from conditional cross-entropy minimization (Appendix A). They characterize the target conditionals of unrestricted risk minimization, not the convergence or timescales of a finite neural model. The controlled implementation uses the normalized minibatch loss

$$\widehat{\mathcal{L}}_{\mathcal{B}} = \frac{\sum_{b,i} w_{bi}[-\log p_\theta(x_i^{(b)} \mid x_t^{(b)}, t_b)]}{\max(1, \sum_{b,i} w_{bi})}, \qquad w_{bi} = \frac{m_{bi}}{\max(t_b, t_{\min})}. \tag{4}$$

We do not assume that this ratio estimator has identical optimization dynamics to the population objective.

Population-learning onset. For the hierarchical-grammar experiments, held-out denoising is evaluated at fixed masking rate $t_0 = 0.15$. At training step s, we define

$$G(s) = \frac{\mathcal{L}_{\text{chance}} - \mathcal{L}_{\text{val}}(s)}{\mathcal{L}_{\text{chance}} - \mathcal{L}_{\text{Bayes}}}, \tag{5}$$

where all losses use aligned conventions. On evaluation grid $\{s_j\}_{j=1}^{J}$, the population-learning onset is

$$\tau_{\text{pop}}(g) = \min_{1 \le j \le J-2} \left\{ s_j : \min_{\ell \in \{0,1,2\}} G(s_{j+\ell}) \ge g \right\}. \tag{6}$$

We use $g \in \{0.25, 0.50\}$ on a common 40-step grid, with $g = 0.50$ as the primary threshold. Validation determines the onset; held-out test evaluations are reported separately. This quantity measures population denoising rather than generative sample quality.

Extraction onset. Let $E_{\text{can}}(s)$ and $E_{\text{ctrl}}(s)$ denote mean exact-recovery rates for trained canaries and matched controls, and let $D_{\text{can}}(s)$ denote mean normalized edit distance. We define

$$\tau_{\text{mem}}^{\text{exact}} = \min\{s : E_{\text{can}}(s) \ge 0.05\}, \tag{7}$$

$$\tau_{\text{mem}}^{\text{near}} = \min\{s : D_{\text{can}}(s) \le 0.10\}, \tag{8}$$

$$\tau_{\text{mem}}^{\text{adv}} = \min\{s : E_{\text{can}}(s) - E_{\text{ctrl}}(s) \ge 0.10\}. \tag{9}$$

The advantage endpoint is our primary train-specific extraction measure. All endpoints use first crossings; unobserved onsets are treated as right-censored rather than assigned to the final checkpoint. Conditioning, decoding, query budgets, and scored positions are specified for each experimental setting. These endpoints are not always independent: for example, with four canaries, one greedy reconstruction per canary, and zero control extraction, the exact and advantage thresholds coincide.

Timescale question. Prior diffusion analyses motivate a population-learning-before-memorization picture in which record-specific fitting emerges on a later timescale (Bonnaire et al., 2025; Favero et al., 2025). We test whether this ordering persists in masked DLMs as dataset size, repetition, and the training mask-rate distribution vary. The random-feature analysis in Appendix B provides only conditional motivation for separated optimization timescales; it does not identify its fast and slow modes with the empirical population-learning and extraction onsets measured here.

Context uniqueness. For revealed positions S and target sequence x, define $q_S(x_S) = P_0(X_S = x_S)$.

Theorem 2 (Finite-sample context uniqueness)**.** *Given a target context x_S, let $N-1$ competing sequences be independent draws from P_0, independent of the target, with S fixed independently of these draws. Then*

$$\Pr[x_S \textit{ is unique} \mid x_S, S] = (1 - q_S(x_S))^{N-1}. \tag{10}$$

Corollary 1 (Empirical-conditional collapse)**.** *If a revealed context identifies a single distinct training sequence, its empirical conditional completion distribution is a point mass. Each unrestricted token-conditional optimum consequently assigns unit probability to the corresponding target token.*

For $N > 1$, uniqueness probability at least $p \in (0, 1)$ is equivalent to $I(S) = -\log q_S(x_S) \ge I_p$, where

$$I_p = -\log\left(1 - p^{1/(N-1)}\right) = \log(N-1) - \log(-\log p) + o(1) \tag{11}$$

as $N \to \infty$ for fixed p. Here N counts the target and independent competitors, excluding deterministic repetitions. Proofs appear in Appendix A; whether this exact data-level quantity predicts extraction from a trained neural model is tested separately in Sec. 6.

4 Experiments

Experimental settings. We use two controlled populations: a first-order Markov chain and a hierarchical grammar, both generating length-64 sequences. The Markov baseline has vocabulary size 256, $n = 128$ background sequences, and eight canaries repeated $R = 16$ times. The hierarchical setting uses a depth-3 grammar with branching factor 4 and 16 terminal symbols. Both settings use

Table 1: **Repetition reverses the population-learning extraction ordering.** Triplets correspond to seeds 101/202/303. Population onset uses $G \geq 0.50$ and extraction uses the first canary–control advantage crossing at 0.10. For $R = 0$, the extraction onset is undefined because the target canaries are not included in training; the dash therefore denotes "not applicable," not a right-censored onset.

R	$\tau_{\text{pop}}(0.50)$	$\tau_{\text{mem}}^{\text{adv}}$	Ordering
0	400/440/440	–	not applicable
1	400/400/480	2740/2800/2900	population first (3/3)
4	640/920/640	900/1160/920	population first (3/3)
16	1480/1520/1560	300/200/220	extraction first (3/3)

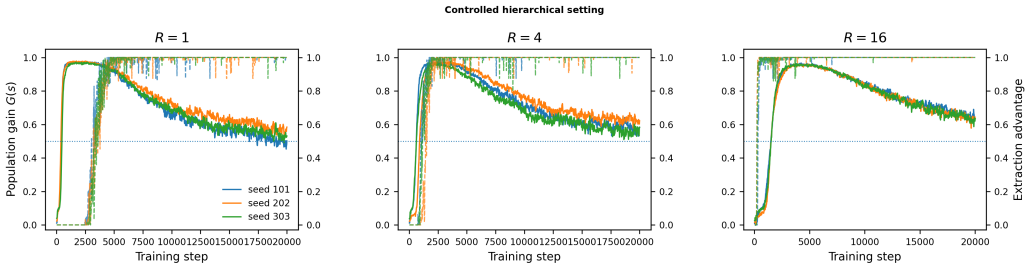

Figure 1: **Repetition reverses the observed learning-extraction ordering.** Solid curves show Bayes-referenced population gain $G(s)$; dashed curves show canary-control extraction advantage. At $R = 1$ and $R = 4$, population learning precedes extraction in every seed; at $R = 16$, extraction precedes population learning in every seed. Formal onsets use Sec. 3, rather than curve intersections.

the same 10.87M-parameter pre-LN Transformer (6 layers, width 384, 6 heads) trained with AdamW at learning rate 3×10^{-4}.

In the hierarchical setting, canaries and controls are i.i.d. random 64-token strings over the 16-symbol alphabet and are therefore out-of-distribution with respect to the grammar; controls follow the same construction but are never included in training. Unless otherwise specified, experiments use seeds 101, 202, and 303. Population and extraction metrics follow Sec. 3; complete configurations are given in Appendix H.

4.1 Repetition reverses learning-extraction ordering

We first ask whether population learning necessarily precedes training-record extraction. Using the hierarchical population, we hold the model, optimizer, and uniform training mask distribution fixed and vary $R \in \{0, 1, 4, 16\}$ across three paired seeds. Population onsets are evaluated on a common 40-step grid and extraction onsets on the native 20-step grid.

The ordering reverses with repetition. At $R = 1$, population learning precedes extraction in every seed: $\tau_{\text{pop}}(0.50) = 400$–$480$, whereas $\tau_{\text{mem}}^{\text{adv}} = 2740$–$2900$ (Table 1). At $R = 4$, population learning remains first in all three seeds, but the gap narrows substantially. At $R = 16$, the ordering reverses: extraction occurs at steps 200–300, while population learning is not observed until steps 1480–1560. The same qualitative ordering holds under the lower $G \geq 0.25$ threshold (Table 4).

Replaying the validation-defined population-onset checkpoints on the held-out test split gives $G_{\text{test}} = 0.48$–0.59 at $R = 1$ and 0.49–0.56 at $R = 4$, while extraction advantage remains zero at all six checkpoints. Thus the population-first separation is also present on unseen data. Increasing R, however, changes both the sampling probability of a particular canary and the overall canary/background mixture. We next separate these effects.

Table 2: **Learning-extraction ordering across dataset size and repetition.** P denotes population first and E extraction first; every entry is seed-unanimous.

n	$R = 1$	$R = 4$	$R = 16$	$R = 64$
128	P	P	E	E
256	P	P	E	E
512	P	P	P	E

Table 3: **Controlled decoupling of record-specific frequency and aggregate repeated-data mass.** Entries give mean $\Delta = \log(\tau^{\mathrm{adv}}_{\mathrm{mem}}/\tau_{\mathrm{pop}}(0.50))$ followed by the seed-level ordering. P denotes population first and E extraction first. Only A1 and B1 exhibit an ordering reversal; the interventions therefore show that changing either quantity *can*, rather than must, reverse the ordering.

Family	Held fixed	$K = 4$	$K = 8$	$K = 16$
A1	$f_{\mathrm{can}} = 0.20$	-0.796 (E 3/3)	$+0.317$ (P 3/3)	$+1.325$ (P 3/3)
A2	$f_{\mathrm{can}} = 1/3$	-1.484 (E 3/3)	-1.079 (E 3/3)	-0.096 (mixed)
B1	$p_{\mathrm{can}} = 0.025$	$+0.584$ (P 3/3)	$+0.355$ (P 3/3)	-0.855 (E 3/3)
B2	$p_{\mathrm{can}} = 1/24$	-0.429 (E 3/3)	-1.059 (E 3/3)	-2.168 (E 3/3)

4.2 Normalized record frequency organizes the ordering

We expand the hierarchical experiment jointly over background dataset size and repetition. For $K = 8$, we evaluate $n \in \{128, 256, 512\}$ and $R \in \{1, 4, 16, 64\}$ using the same three seeds. Define the signed onset separation

$$\Delta = \log \frac{\tau^{\mathrm{adv}}_{\mathrm{mem}}}{\tau_{\mathrm{pop}}(0.50)}. \tag{12}$$

Thus $\Delta > 0$ denotes population-first and $\Delta < 0$ extraction-first behavior.

The raw repetition level at which the ordering changes therefore depends on n. In contrast, normalizing repetition by dataset size substantially aligns the regimes. We add five intermediate conditions to construct matched R/n families across $n \in \{128, 256, 512\}$. For $R/n \in \{1/128, 1/32, 1/16, 1/8\}$, the corresponding mean onset separations are approximately $+1.91$, $+0.31$, -1.04, and -1.91, respectively. At $R/n = 1/32$, one run ($(n, R) = (256, 8)$, seed 303) lands exactly at $\Delta = 0$ on the evaluation grid and is treated as a tie.

Across the 17 conditions (51 seed-level observations), a descriptive regression on $\log p_{\mathrm{can}}$ explains 93.8% of the variation in Δ and 97.7% of the variation in $\log \tau^{\mathrm{adv}}_{\mathrm{mem}}$; adding $\log n$ produces negligible additional fit. At the resolution of the tested grid, the sign change is consistently bracketed by

$$p_{\mathrm{can}} \in (0.025,\ 0.0417), \qquad \text{equivalently} \qquad R/n \in (1/32,\ 1/16)$$

for $K = 8$. This is a bracket rather than a point estimate. Record-specific sampling frequency therefore strongly organizes the observed dynamics, but does not identify their cause: at fixed K, $f_{\mathrm{can}} = K p_{\mathrm{can}}$, so individual-record frequency and aggregate repeated-data mass remain deterministically linked.

4.3 Record-specific frequency and repeated-data mass both matter

We next vary the number of distinct canaries K to separate these two quantities. Two families hold f_{can} exactly fixed while changing p_{can}: A1 fixes $f_{\mathrm{can}} = 0.20$ with $p_{\mathrm{can}} \in \{0.05, 0.025, 0.0125\}$, and A2 fixes $f_{\mathrm{can}} = 1/3$ with $p_{\mathrm{can}} \in \{1/12, 1/24, 1/48\}$. Two complementary families hold p_{can} fixed while changing aggregate mass: B1 fixes $p_{\mathrm{can}} = 0.025$ with $f_{\mathrm{can}} \in \{0.10, 0.20, 0.40\}$, and B2 fixes $p_{\mathrm{can}} = 1/24$ with $f_{\mathrm{can}} \in \{1/6, 1/3, 2/3\}$. All conditions use the same nested canary pool, and extraction is evaluated on the shared anchors C1–C4 against U1–U4 so that changing K does not change the primary extraction resolution.

At fixed aggregate canary mass (A1), reducing the sampling probability of each individual canary changes the regime from extraction-first to population-first. Conversely, at fixed individual-canary sampling probability (B1), increasing the aggregate mass assigned to repeated records changes the

regime from population-first to extraction-first. The remaining two families shift the onset separation substantially without crossing zero. Matched controls remain unextracted throughout.

These interventions rule out either p_{can} or f_{can} as a sufficient one-variable description of the ordering. In the tested families, changes in p_{can} primarily shift extraction time, whereas changes in f_{can} primarily shift population-learning time. We do not interpret this as a complete causal decomposition: because $f_{\mathrm{can}} = Kp_{\mathrm{can}}$, holding one quantity fixed while varying the other necessarily also changes the number of distinct repeated records K.

For completeness, the earlier one-dimensional dataset-size and repetition sweeps, including finite-range log-log fits, are reported in Appendix D. They support the same directional dependence but are not used as universal scaling laws.

4.4 A CONTRASTING MARKOV REGIME

The hierarchical reversal is not universal across data distributions. In the first-order Markov setting, extraction emerges without an observed meaningful population-learning phase. Over 12,000 optimization steps, held-out loss never falls below the chance reference in any of the three seeds, while near-verbatim extraction occurs at step 240 and extraction-advantage onset at step 280 in every seed. Matched-control extraction remains zero, and the best held-out losses remain 46.5–48.2% above their Bayes references.

Thus, extraction can emerge and persist even when the measured population-learning criterion remains censored over an extended horizon. One-factor variations of vocabulary size, learning rate, model capacity, and mask-rate support, together with additional data-regime stress tests, show the same qualitative contrast (Appendix D).

4.5 MASKING-DISTRIBUTION SHAPE AFFECTS EXTRACTION ONSET

Record sampling is not the only factor governing extraction. Masked DLMs are trained on a distribution of conditional prediction problems induced by the masking schedule, so we next ask whether extraction timing is determined by coarse properties of that distribution.

Across four controlled schedules in the Markov setting, mean extraction-advantage onset ranges from 213 to 393 steps while best held-out losses remain similar. A broader 13-schedule analysis likewise shows that mean and variance alone poorly predict onset; because several shape statistics are correlated, however, this analysis does not identify a unique responsible feature.

Prospective matched-moment comparison. To directly test whether the first two moments are sufficient, we compare two mask-rate distributions $P^\star$ and $Q^\star$ with identical mean (0.51), variance $(7609/120000)$, and support $[0.02, 1]$, but different shapes. Matched runs share initialization and data-sampling streams, and record-presentation counts coincide at corresponding steps.

Extraction advantage emerges earlier under $P^\star$ than under $Q^\star$ in all three paired seeds: $(240, 200, 200)$ versus $(280, 280, 280)$. Near-verbatim onset agrees in direction in all three pairs, while exact onset agrees in two pairs and ties in one.

Matching the first two moments of the training mask-rate distribution is therefore insufficient to determine extraction onset. This comparison does not identify a unique higher-order mechanism: $P^\star$ and $Q^\star$ also differ in their tail masses, and population-learning onsets remain censored in all six runs.

5 PRETRAINED-MODEL STRESS TESTS

We next test whether the controlled trends persist when adapting pretrained masked DLMs. We consider LLaDA-8B-Base with rank-8 LoRA and MDLM with full-parameter fine-tuning. These experiments are portability tests rather than matched architectural comparisons: the setups differ in model scale, adaptation method, tokenization, batch size, and input processing. Additional protocol details and per-seed results are given in Appendix F.

Repetition consistently accelerates extraction. Both pretrained models reproduce the directional repetition effect. For MDLM, mean extraction-advantage onset decreases from 253.3 to 93.3 to 40.0

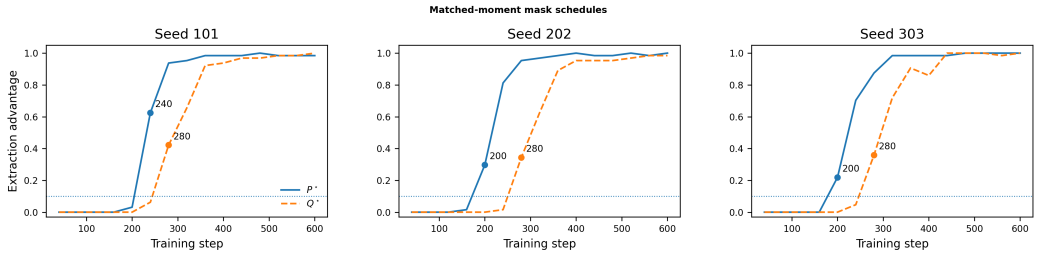

Figure 2: **Matched-moment mask distributions shift extraction onset.** $P^\star$ and $Q^\star$ have identical mask-rate mean, variance, and support but different shapes. In every paired seed, $P^\star$ reaches the extraction-advantage threshold first.

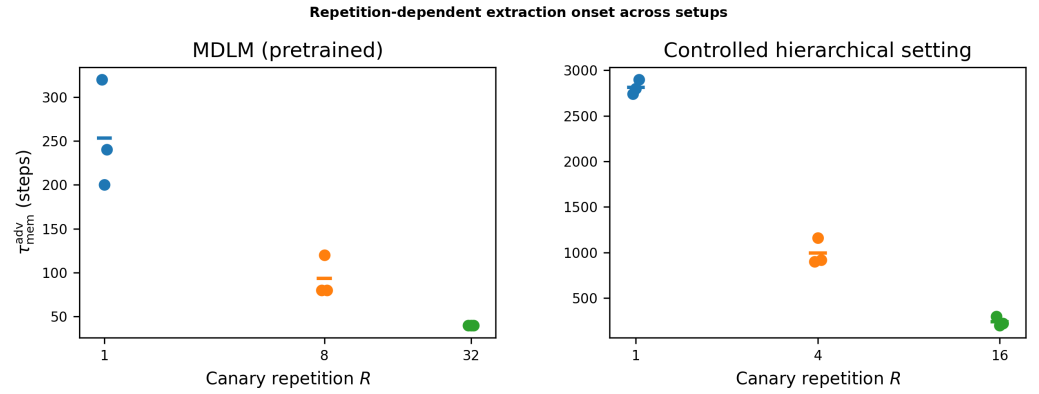

Figure 3: **Greater repetition leads to earlier extraction across distinct setups.** Extraction-advantage onset decreases with repetition in full-parameter MDLM adaptation and in the controlled hierarchical-grammar experiment. Points denote seeds and horizontal bars denote condition means; absolute training-step magnitudes are not comparable across setups.

steps as R increases from 1 to 8 to 32, with the direction holding in every seed. LLaDA shows the same qualitative trend over $R \in \{1, 8, 32\}$ on both the template background and WikiText-2, while matched controls remain unextracted. Given only three repetition levels, we interpret these results as directional portability rather than a universal scaling law.

Other trends transfer less uniformly. LLaDA also shows later extraction as the background dataset grows, consistent in direction with the controlled dataset-size experiments, although fitted finite-range slopes depend on the onset convention (Appendix F). Mask-schedule effects are less robust: LLaDA shows broadly compatible mean-level trends but substantial seed-level variation, while the primary MDLM comparison is inconclusive under its prespecified resolution criterion. We therefore regard repetition as the most consistently reproduced controlled trend and do not claim a universal schedule ordering across pretrained models.

6 What context uniqueness can and cannot explain

The result in Eq. 10 characterizes a property of the finite training sample: a revealed context may uniquely identify one empirical training completion. This does not imply that a finite neural denoiser will learn or exploit that context at a particular point in training.

We test this distinction directly using 2,160 canary-checkpoint-geometry instances. Adding context self-information $I(S) = -\log q_S(x_S)$ to a predictor based on reveal fraction yields essentially no improvement in held-out prediction ($\Delta R^2 \approx 0$), and the corresponding bootstrap coefficient interval contains zero (approximately $[-0.0026, 0.0003]$). Thus, under our tested protocol, context

uniqueness provides essentially no incremental predictive value for neural extraction beyond reveal fraction.

This negative result should not be interpreted as a general rejection of uniqueness-based explanations: the evaluated contexts are already empirically unique and $I(S)$ is highly correlated with reveal fraction. Rather, it separates an exact property of the empirical training distribution from evidence that a trained neural model actually exploits that property. Full calibration and predictive diagnostics are reported in Appendix G.

7 Discussion and limitations

The learning-extraction ordering is regime dependent. Our experiments show that population learning need not precede train-specific extraction in masked DLMs. Within the same hierarchical model and data family, increasing repetition changes the ordering from population-first to extraction-first, while the Markov setting exhibits extraction without an observed meaningful population-learning phase. Thus, a single generalization-before-memorization ordering does not describe all of the regimes studied here.

Record frequency organizes the dynamics, but does not fully explain them. The joint dataset-size-repetition study shows that individual-record sampling probability strongly organizes extraction timing and the learning-extraction ordering at fixed K. The controlled decoupling experiments further show that neither individual-record frequency nor aggregate repeated-data mass is sufficient alone: varying either while holding the other fixed can reverse the ordering. These interventions do not identify a unique mechanism, however. Since $f_{\mathrm{can}} = Kp_{\mathrm{can}}$, changing one quantity while fixing the other also changes the number of distinct repeated records K. Moreover, the hierarchical canaries are out-of-distribution relative to the grammar, so increasing f_{can} reallocates training mass away from the background population. The decoupling experiments therefore characterize competition induced by the training mixture rather than an intrinsic architecture-specific memorization clock. Similarly, the matched-moment masking experiment establishes that mean and variance are insufficient to determine extraction onset, but does not isolate the responsible higher-order property.

Measurement, portability, and scope. Most comparisons use three seeds, and extraction onset depends on the conditioning protocol, decoding rule, query budget, evaluation cadence, and threshold. Pretrained LLaDA and MDLM reproduce the directional repetition effect, whereas mask-schedule transfer is less consistent. Synthetic canaries provide controlled probes of train-specific recoverability rather than deployment-scale estimates of privacy leakage, and zero observed control extraction is not a privacy guarantee. Nevertheless, understanding when and under which training regimes individual records become extractable may inform privacy auditing and the evaluation of data-removal or machine-unlearning interventions.

8 Conclusion

We studied the relative emergence of population learning and training-record extraction in masked diffusion language models. In a controlled hierarchical setting, increasing record repetition reverses the ordering from population-learning-first to extraction-first, while a contrasting Markov regime exhibits extraction without an observed meaningful population-learning phase. Joint dataset-size and repetition experiments show that normalized record frequency strongly organizes these dynamics, but controlled decoupling reveals that both record-specific sampling frequency and aggregate repeated-data mass matter. Extraction timing also depends on the training mask-rate distribution beyond its first two moments, and pretrained LLaDA and MDLM reproduce the directional repetition effect. Overall, population learning and train-specific extraction are distinct training phenomena whose relative timing is regime dependent rather than governed by a universal memorization timescale.

Ethics statement

This work studies memorization and train-specific extraction, which can inform privacy auditing, but it does not claim deployment-scale privacy leakage. Controlled experiments use synthetic populations;

public text corpora such as WikiText-2 and text8 are used only as background text with synthetic inserted canaries. No private or personally identifying training data are used.

REPRODUCIBILITY STATEMENT

Appendices A-C provide complete proofs and the assumptions behind conditional theory. Appendices D-H report threshold robustness, geometry, pretrained-model sweeps, context-uniqueness tests, data generation, hyperparameters, onset definitions, censoring rules, and compute details. The analysis uses fixed seeds and predeclared thresholds; matched controls are evaluated with the same masks and samplers as canaries.

AI USE STATEMENT

Generative AI tools were used as research assistants for code drafting and debugging, literature-oriented brainstorming, manuscript editing, and consistency checks. Experimental claims and numerical results were verified against saved artifacts; theoretical statements and proofs were reviewed and revised by the authors. The authors take responsibility for all content in the submission.

REFERENCES

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

Tony Bonnaire, Rapha"el Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *Advances in Neural Information Processing Systems*, 2025.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium*, 2021.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium*, pp. 5253–5270, 2023.

Alessandro Favero, Antonio Sclocchi, and Matthieu Wyart. Bigger isn't always memorizing: Early stopping overparameterized diffusion models, 2025. URL https://arxiv.org/abs/2505.16959.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution, 2024. URL https://arxiv.org/abs/2310.16834.

Xiaoyu Luo, Wenrui Yu, Qiongxiu Li, and Johannes Bjerva. Characterizing memorization in diffusion language models: Generalized extraction and sampling effects, 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*, 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*, 2025.

Bao Pham, Mohammed J. Zaki, Luca Ambrogioni, Dmitry Krotov, and Matteo Negri. Language diffusion models are associative memories capable of retrieving unseen data, 2026.

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

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Yihan Wang and N. Asokan. Extracting training data from diffusion language models via infilling, 2026.

A Proofs of Exact Results

A.1 Proof of Theorem 1

Proof. Let $X = (X_1, \dots, X_L) \sim P_0$, let $t \sim \pi$, and let X_t denote the corrupted sequence generated by the masking process. Because the mask token is distinct from the clean vocabulary, whether position i is masked is observable from X_t.

For distributions P and Q on a finite set,

$$-\sum_x P(x) \log Q(x) = H(P) + \mathrm{KL}(P \| Q). \tag{13}$$

Hence cross-entropy is minimized uniquely by $Q = P$ on the support of P.

Fix a position i and a positive-weight masked context (x_t, t). Conditioning the contribution of Eq. (3) on this context gives

$$\begin{aligned}&\mathbb{E}[-\log p(X_i \mid x_t, t) \mid X_t = x_t, t, m_i = 1] \\ &\quad = H(P_0(X_i \mid x_t, t, m_i = 1)) + \mathrm{KL}(P_0(X_i \mid x_t, t, m_i = 1) \,\|\, p(\cdot \mid x_t, t)) \,.\end{aligned} \tag{14}$$

Since $m_i = 1$ is already determined by x_t,

$$P_0(X_i \mid x_t, t, m_i = 1) = P_0(X_i \mid x_t, t).$$

The KL term is nonnegative and vanishes exactly when

$$p(\cdot \mid x_t, t) = P_0(X_i = \cdot \mid x_t, t).$$

The factor $1/t$ in Eq. (3) is strictly positive for $t > 0$ and therefore does not change this contextwise minimizer.

Thus every unrestricted minimizer agrees with the population conditional at almost every positive-weight masked context. Contexts receiving zero objective weight impose no constraint. □

A.2 Proof of Proposition 1

Proof. Let $x^{(1)}, \dots, x^{(M)}$ denote the distinct sequences in the empirical support and write the empirical training distribution as

$$\widehat{P}_a = \sum_{j=1}^{M} a_j \, \delta_{x^{(j)}}, \qquad a_j \geq 0, \qquad \sum_{j=1}^{M} a_j = 1,$$

where the weights a_j include multiplicities induced by repeated records.

Fix a corrupted context x_t and masking rate $t > 0$, and let

$$S = \{k : x_{t,k} \neq \texttt{[MASK]}\}$$

be the revealed positions, with $u_S = x_{t,S}$. Under independent Bernoulli masking conditional on t, the probability of the fixed mask pattern is independent of the clean sequence and is therefore common to all sequences compatible with $X_S = u_S$. This factor cancels when conditioning on the corrupted context.

Consequently, whenever the denominator is positive,

$$\widehat{P}_a(X_i = v \mid X_t = x_t, t) = \frac{\sum_{j=1}^{M} a_j \, \mathbf{1}\{x_S^{(j)} = u_S, \; x_i^{(j)} = v\}}{\sum_{j=1}^{M} a_j \, \mathbf{1}\{x_S^{(j)} = u_S\}}, \tag{15}$$

for every masked position $i \notin S$. In particular,

$$\widehat{P}_a(X_i = v \mid X_t = x_t, t) = \widehat{P}_a(X_i = v \mid X_t = x_t)$$

under this masking mechanism.

Conditioning the empirical masked-loss objective on a positive-weight context therefore gives a cross-entropy between the empirical conditional in Eq. (15) and the candidate prediction $p(\cdot \mid x_t, t)$. By Eq. (13), its unrestricted minimizer is

$$p^\star(\cdot \mid x_t, t) = \widehat{P}_a(\cdot \mid x_t, t).$$

Contexts with zero empirical objective weight impose no constraint. □

A.3 Context uniqueness and empirical-conditional collapse

We now prove Theorem 2 and Corollary 1, and derive the associated information threshold.

Proof of Theorem 2. Fix the revealed set S and values x_S, and write

$$q = q_S(x_S) = P_0(X_S = x_S).$$

Each of the $N - 1$ independent competing sequences matches x_S on S with probability q, and therefore fails to match with probability $1 - q$. Independence gives

$$\Pr(\text{no independently sampled competitor matches } x_S \,|\, x_S, S) = \prod_{\nu=1}^{N-1} (1 - q) = (1 - q)^{N-1}.$$

□

Proof of Corollary 1. Suppose the revealed context x_S is compatible with exactly one distinct sequence $x^\star$ in the support of $\widehat{P}_a$. Then Eq. (15) assigns all conditional mass to $x^\star$:

$$\widehat{P}_a(X = x^\star \mid X_S = x^\star_S) = 1.$$

Hence every hidden-token conditional is a point mass at the corresponding token of $x^\star$.

Deterministic repetitions of $x^\star$ change its empirical weight but introduce no competing completion, so they do not affect this conclusion. □

Exact information threshold. Let

$$I(S) := -\log q_S(x_S).$$

For a desired uniqueness probability $p \in (0, 1)$, define q_p through

$$(1 - q_p)^{N-1} = p.$$

Solving gives

$$q_p = 1 - p^{1/(N-1)}, \tag{16}$$

so uniqueness probability at least p is equivalent to

$$q_S(x_S) \le q_p \qquad \Longleftrightarrow \qquad I(S) \ge I_p,$$

where

$$I_p = -\log\left(1 - p^{1/(N-1)}\right). \tag{17}$$

For fixed $p \in (0, 1)$, let $c = -\log p > 0$ and $m = N - 1$. Since

$$q_p = 1 - e^{-c/m} = \frac{c}{m} + O\left(\frac{1}{m^2}\right),$$

we obtain

$$I_p = -\log\left[\frac{c}{m}\left(1 + O\left(\frac{1}{m}\right)\right)\right]$$
$$= \log m - \log c + O\left(\frac{1}{m}\right). \tag{18}$$

Therefore

$$I_p = \log(N-1) - \log(-\log p) + O\left(\frac{1}{N-1}\right), \tag{19}$$

and hence

$$I_p = \log(N-1) - \log(-\log p) + o(1) \qquad \text{as } N \to \infty.$$

Here N counts the target and its independently sampled competitors; deterministic repetitions of the target are excluded from this count.

Scope. These results characterize the population and empirical targets of unrestricted masked-denoising risk minimization and when a revealed context uniquely identifies an empirical completion. They do not imply that a finite neural model realizes or exploits these conditionals at a particular point in training. That separate question is tested empirically in Sec. 6.

B Conditional Random-Feature Analysis

Scope. This appendix develops a conditional random-feature analysis motivated by continuous-diffusion memorization theory (Bonnaire et al., 2025) and adapted to masked corruption. The analysis uses a squared-loss random-feature surrogate together with an embedding-level corruption assumption and an explicit spectral assumption. It is not a mechanistic theory of the Transformer denoisers studied experimentally.

The exact part of the analysis characterizes the gradient-flow dynamics of the surrogate. Under the additional two-scale spectral assumption, it yields two separated relaxation times whose ratio satisfies

$$\frac{\tau_{\text{slow}}^{\text{RF}}}{\tau_{\text{fast}}^{\text{RF}}} = \Theta(\psi_n), \qquad \psi_n := \frac{n}{d}.$$

Interpreting the fast and slow random-feature modes as population learning and record-specific fitting, respectively, is a transferred heuristic rather than a consequence proved here.

Roadmap. We first introduce an embedding-level surrogate for masked corruption and the random-feature denoiser. We then derive its exact gradient-flow dynamics using the per-example feature second moment, obtain the modewise relaxation times, and finally impose a two-scale spectral assumption to derive the conditional separation of those timescales.

B.1 Setup and assumptions

We analyze a fixed masking level $t \in (0, 1]$ and a fixed target position i. Throughout this appendix we condition on position i being masked; for notational simplicity, $q_t(\cdot \mid x)$ denotes this conditional corruption law. The analysis is position-wise, so we suppress i when no ambiguity arises.

Training examples. Let $\{(x^\nu, y^\nu)\}_{\nu=1}^n$ be iid training examples, where x^ν is a clean sequence. Here n denotes the surrogate sample size; it is not identified with the main-text background-record count without an additional modeling choice. Let

$$x_t^\nu \sim q_t(\cdot \mid x^\nu)$$

be its corrupted version at masking level t. Let

$$z^\nu := \text{Enc}(x^\nu) \in \mathbb{R}^d, \qquad z_t^\nu := \text{Enc}(x_t^\nu) \in \mathbb{R}^d$$

denote centered clean and corrupted context representations. The target $y^\nu \in \mathbb{R}^m$ is the regression target at the masked position; for categorical denoising it may be the one-hot encoding of x_i^ν, with $m = |\mathcal{V}|$.

High-dimensional regime. Define the aspect ratios

$$\psi_n := \frac{n}{d}, \qquad \psi_p := \frac{p}{d}.$$

For each fixed pair (ψ_n, ψ_p), we consider the proportional limit

$$n, p, d \to \infty, \qquad \frac{n}{d} \to \psi_n, \qquad \frac{p}{d} \to \psi_p.$$

We subsequently compare these limiting descriptions across aspect-ratio regimes, with particular interest in

$$\psi_p \gg \psi_n \gg 1.$$

Thus, statements involving $\Theta(\psi_n)$ below describe dependence across such high-dimensional aspect-ratio regimes; they do not imply a fixed-d, $n \to \infty$ scaling law.

Random-feature surrogate. Let $W \in \mathbb{R}^{p\times d}$ have iid $N(0,1)$ entries, independently of the data, and let $\sigma : \mathbb{R} \to \mathbb{R}$ act element-wise. We assume only the moment and regularity conditions needed for the random-feature quantities below to be well defined. Define

$$\phi_t(z_t) := \sigma\left(\frac{Wz_t}{\sqrt{d}}\right) \in \mathbb{R}^p. \tag{20}$$

The random-feature denoiser is

$$f_A(z_t) := \frac{1}{\sqrt{p}} A\phi_t(z_t) \in \mathbb{R}^m, \qquad A \in \mathbb{R}^{m\times p}. \tag{21}$$

We analyze f_A with a squared-loss surrogate rather than a categorical softmax likelihood. Consequently, conclusions in this appendix concern the optimization dynamics of this surrogate and do not by themselves establish corresponding properties of the categorical Transformer denoisers used in the experiments.

Squared-loss surrogate. At the fixed masking level t and target position i, let $y(x,i) \in \mathbb{R}^m$ denote the regression target associated with the clean token at position i. We analyze the population squared-loss surrogate

$$\mathcal{L}_{\mathrm{sq}}(A) := \frac{1}{2}\,\mathbb{E}_{x\sim P_0}\mathbb{E}_{x_t\sim q_t(\cdot|x)}\left[\left\|y(x,i) - f_A(z_t)\right\|_2^2\right], \tag{22}$$

where, as above, $q_t(\cdot \mid x)$ denotes the corruption law conditioned on the target position being masked. For iid training sequences $x^1, \dots, x^n$, with $y^\nu := y(x^\nu, i)$, the corresponding empirical objective is

$$\mathcal{L}_{\mathrm{sq},n}(A) := \frac{1}{2n}\sum_{\nu=1}^{n} \mathbb{E}_{x_t^\nu\sim q_t(\cdot|x^\nu)}\left[\left\|y^\nu - f_A(z_t^\nu)\right\|_2^2\right]. \tag{23}$$

We study continuous-time full-batch gradient flow on this surrogate,

$$\dot{A}(\tau) = -\nabla_A \mathcal{L}_{\mathrm{sq},n}(A(\tau)). \tag{24}$$

Any positive multiplicative factor depending only on the fixed masking level t, such as the $1/t$ weighting in the masked-denoising objective, merely rescales τ and therefore does not affect the relative mode-relaxation times derived below.

Importantly, the corruption expectation in Eq. (23) acts on the squared loss:

$$\mathbb{E}_{x_t^\nu}\left[\|y^\nu - f_A(z_t^\nu)\|_2^2\right],$$

rather than on the feature vector before the square is formed. As a result, the quadratic term in A depends on the per-example feature second moment $\mathbb{E}[\phi_t(z_t^\nu)\phi_t(z_t^\nu)^\top]$, not only on the outer product of the mean feature vector. This distinction is used explicitly in the gradient-flow derivation below.

Embedding-level corruption assumption. To obtain a tractable high-dimensional model of masked corruption, we introduce an embedding-level additive-noise surrogate. This is an analytical assumption rather than a property established for the Transformer representations used in our experiments.

Assumption 1 (Embedding-level corruption model). For each fixed masking level t, there exist deterministic scalars $\Gamma_t \in [0,1]$ and $\Delta_t > 0$ such that, conditional on a clean embedding z^ν,

$$z_t^\nu \overset{d}{=} \Gamma_t z^\nu + \sqrt{\Delta_t}\,\xi^\nu, \tag{25}$$

where the corruption variables $\{\xi^\nu\}_{\nu=1}^n$ are independent across examples and satisfy

$$\mathbb{E}[\xi^\nu] = 0, \qquad \mathbb{E}[\xi^\nu(\xi^\nu)^\top] = I_d, \tag{26}$$

with ξ^ν sub-Gaussian and independent of z^ν. The clean embeddings are mean-zero sub-Gaussian with covariance

$$\Sigma := \mathbb{E}[zz^\top], \tag{27}$$

where $\|\Sigma\|_{\mathrm{op}} = O(1)$ and the empirical spectral distribution of Σ converges to a deterministic limit in the high-dimensional regime.

Remark 1 (Scope of the embedding surrogate). Equation (25) approximates token-space masking by an attenuation of the clean representation plus independent isotropic embedding noise. Such a form can arise in simplified additive or linearized representation models, but we do not establish it for the attention-based Transformer denoisers studied experimentally. Consequently, Assumption 1 should be viewed as part of the random-feature surrogate, not as a mechanistic description of LLaDA, MDLM, or the controlled Transformer.

B.2 Feature second moments and exact gradient-flow dynamics

The empirical squared-loss objective depends on the full second moment of the random features under corruption. We make this dependence explicit before deriving the gradient-flow dynamics.

Feature moments. For each training example, define the corruption-averaged feature vector

$$\phi^\nu := \mathbb{E}_{x_t^\nu \sim q_t(\cdot|x^\nu)}[\phi_t(z_t^\nu)] \in \mathbb{R}^p, \tag{28}$$

and collect the mean features and targets as

$$\Phi := [\phi^1, \dots, \phi^n] \in \mathbb{R}^{p\times n}, \qquad Y := [y^1, \dots, y^n] \in \mathbb{R}^{m\times n}.$$

Define the empirical feature-target cross-correlation

$$V := \frac{1}{n}\Phi Y^\top = \frac{1}{n}\sum_{\nu=1}^n \phi^\nu (y^\nu)^\top \in \mathbb{R}^{p\times m}, \tag{29}$$

and the empirical feature second-moment matrix

$$U := \frac{1}{n}\sum_{\nu=1}^n \mathbb{E}_{x_t^\nu \sim q_t(\cdot|x^\nu)}\left[\phi_t(z_t^\nu)\phi_t(z_t^\nu)^\top\right] \in \mathbb{R}^{p\times p}. \tag{30}$$

All expectations above are over corruption randomness, conditional on the clean training sample and the fixed random-feature matrix W.

For each example,

$$\mathbb{E}[\phi\phi^\top] = \mathbb{E}[\phi]\mathbb{E}[\phi]^\top + \mathrm{Cov}(\phi),$$

and therefore

$$U = \frac{1}{n}\Phi\Phi^\top + \frac{1}{n}\sum_{\nu=1}^n \mathrm{Cov}_{x_t^\nu|x^\nu}[\phi_t(z_t^\nu)]\,. \tag{31}$$

Hence

$$U \succeq \frac{1}{n}\Phi\Phi^\top,$$

with the difference given exactly by the average corruption-induced feature covariance. In particular, replacing the second moment by an outer product of mean features would generally change the quadratic term of the training objective.

Lemma 1 (Exact gradient-flow dynamics)**.** *For the empirical squared-loss objective Equation (23), the gradient-flow dynamics satisfy*

$$\dot{A}(\tau) = -\frac{1}{p}A(\tau)U + \frac{1}{\sqrt{p}}V^\top. \tag{32}$$

More generally, if the fixed-t objective is multiplied by a positive scalar c_t, then

$$\dot{A}(\tau) = -\kappa_t A(\tau)U + \beta_t V^\top, \qquad \kappa_t := \frac{c_t}{p}, \qquad \beta_t := \frac{c_t}{\sqrt{p}}. \tag{33}$$

Such a common scalar weighting rescales gradient-flow time but does not change relative modal relaxation times.

Proof. For a fixed training example ν, write

$$\phi_\nu := \phi_t(z_t^\nu).$$

Expanding the squared loss before taking the corruption expectation,

$$\left\| y^\nu - \frac{1}{\sqrt{p}}A\phi_\nu \right\|_2^2 = \|y^\nu\|_2^2 - \frac{2}{\sqrt{p}}(y^\nu)^\top A\phi_\nu + \frac{1}{p}\phi_\nu^\top A^\top A\phi_\nu. \tag{34}$$

Taking expectation over the corruption randomness gives

$$\begin{aligned}\mathbb{E}_{x_t^\nu}\left[\left\| y^\nu - \frac{1}{\sqrt{p}}A\phi_\nu \right\|_2^2\right] = \|y^\nu\|_2^2 - \frac{2}{\sqrt{p}}(y^\nu)^\top A\phi^\nu \\ + \frac{1}{p}\operatorname{Tr}\left(A^\top A\,\mathbb{E}_{x_t^\nu}[\phi_\nu\phi_\nu^\top]\right).\end{aligned} \tag{35}$$

Thus the quadratic term depends on the feature second moment $\mathbb{E}[\phi_\nu\phi_\nu^\top]$, rather than only on the mean feature $\mathbb{E}[\phi_\nu]$.

Averaging over $\nu = 1, \dots, n$ and using Equations (29) and (30) yields

$$\mathcal{L}_{\mathrm{sq},n}(A) = C - \frac{1}{\sqrt{p}}\operatorname{Tr}(AV) + \frac{1}{2p}\operatorname{Tr}(A^\top A\,U), \tag{36}$$

where

$$C := \frac{1}{2n}\sum_{\nu=1}^{n}\|y^\nu\|_2^2$$

does not depend on A.

Since U is symmetric,

$$\nabla_A \operatorname{Tr}(A^\top AU) = 2AU, \qquad \nabla_A \operatorname{Tr}(AV) = V^\top.$$

Therefore

$$\nabla_A\mathcal{L}_{\mathrm{sq},n}(A) = -\frac{1}{\sqrt{p}}V^\top + \frac{1}{p}AU,$$

and gradient flow gives

$$\dot{A}(\tau) = -\frac{1}{p}A(\tau)U + \frac{1}{\sqrt{p}}V^\top.$$

If the objective is multiplied by $c_t > 0$, the gradient and hence the entire right-hand side are multiplied by c_t, yielding Equation (33). □

Implication for the spectral analysis. The optimization dynamics depend on corruption through the spectrum of the second-moment matrix U. Consequently, any later assumption about separated spectral scales must apply to the full matrix in Equation (30), including the corruption-induced covariance term in Equation (31). The gradient-flow derivation itself does not imply that such a spectral separation exists; this is stated separately as an assumption below.

B.3 Closed-form solution and modewise relaxation

Lemma 2 (Closed-form gradient-flow solution)**.** *Assume $A(0) = 0$ and consider the scaled gradient-flow equation Equation (33),*

$$\dot{A}(\tau) = -\kappa_t A(\tau)U + \beta_t V^\top, \qquad \kappa_t, \beta_t > 0.$$

Then

$$A(\tau) = \frac{\beta_t}{\kappa_t} V^\top U^\dagger \left(I_p - e^{-\kappa_t U \tau}\right), \tag{37}$$

where $U^\dagger$ denotes the Moore-Penrose pseudoinverse.

Equivalently, if

$$U = \sum_{j=1}^{p} \lambda_j v_j v_j^\top$$

is an eigendecomposition, then

$$A(\tau) = \frac{\beta_t}{\kappa_t} \sum_{j:\lambda_j > 0} \frac{1 - e^{-\kappa_t \lambda_j \tau}}{\lambda_j} V^\top v_j v_j^\top. \tag{38}$$

Proof. The linear ODE has the variation-of-constants solution

$$A(\tau) = \beta_t V^\top \int_0^\tau e^{-\kappa_t U s}\, ds. \tag{39}$$

Because $U \succeq 0$, let P_0 denote the orthogonal projector onto $\ker(U)$. We first note that

$$V^\top P_0 = 0. \tag{40}$$

Indeed, for any $v \in \ker(U)$,

$$0 = v^\top U v = \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{E}_{x_t^\nu} \left[\left(v^\top \phi_t(z_t^\nu) \right)^2 \right]. \tag{41}$$

Every term is nonnegative, so $v^\top \phi_t(z_t^\nu) = 0$ almost surely for every ν. Consequently $v^\top \phi^\nu = 0$ and hence $V^\top v = 0$.

Using the spectral decomposition of U,

$$e^{-\kappa_t U s} = P_0 + \sum_{j:\lambda_j > 0} e^{-\kappa_t \lambda_j s} v_j v_j^\top.$$

Substituting into Equation (39), the nullspace term vanishes by Equation (40), while for $\lambda_j > 0$,

$$\int_0^\tau e^{-\kappa_t \lambda_j s}\, ds = \frac{1 - e^{-\kappa_t \lambda_j \tau}}{\kappa_t \lambda_j}.$$

Therefore

$$A(\tau) = \frac{\beta_t}{\kappa_t} \sum_{j:\lambda_j > 0} \frac{1 - e^{-\kappa_t \lambda_j \tau}}{\lambda_j} V^\top v_j v_j^\top,$$

which is Equation (38). The pseudoinverse identity

$$U^\dagger = \sum_{j:\lambda_j > 0} \lambda_j^{-1} v_j v_j^\top$$

then gives Equation (37). □

Rate versus amplitude. For a mode with eigenvalue $\lambda_j > 0$, the time dependence enters through

$$1 - e^{-\kappa_t \lambda_j \tau}.$$

Hence its relaxation rate is determined by $\kappa_t \lambda_j$, whereas its limiting coefficient also depends on $\beta_t/(\kappa_t \lambda_j)$ and on the alignment $V^\top v_j$. The timescale analysis below concerns only the former: it compares modal relaxation rates and makes no quantitative claim about the amplitudes of the learned coefficients or predictions.

B.4 Modewise relaxation and characteristic timescales

From Equation (38), the coefficient associated with an eigenvector v_j of U with eigenvalue $\lambda_j > 0$ evolves through the factor

$$1 - e^{-\kappa_t \lambda_j \tau}.$$

For an active mode, i.e. one with $V^\top v_j \neq 0$, the time required to reach any fixed fraction $\eta \in (0, 1)$ of its limiting coefficient is

$$\tau_j(\eta) = \frac{-\log(1-\eta)}{\kappa_t \lambda_j}. \tag{42}$$

Hence, up to a constant depending only on the chosen fraction η,

$$\tau(\lambda_j) \asymp \frac{1}{\kappa_t \lambda_j}. \tag{43}$$

Thus larger eigenvalues correspond to faster relaxation and smaller positive eigenvalues to slower relaxation. This is a statement about optimization rates only: the limiting amplitude of a mode additionally depends on $V^\top v_j$ and on λ_j^{-1}. In particular, separated eigenvalue scales in U imply separated relaxation timescales, but do not by themselves identify those modes with population learning or record extraction.

B.5 A conditional two-scale spectral assumption

The exact gradient-flow analysis above does not determine the spectrum of U. Moreover, because U contains the corruption-induced within-example covariance term in Equation (31), spectral results for an outer product of mean feature vectors do not automatically apply to the second-moment matrix considered here. We therefore state explicitly the spectral separation required by the subsequent timescale argument.

Prior random-feature analyses of continuous diffusion (Bonnaire et al., 2025) motivate the possibility of separated spectral scales in overparameterized regimes. We use this analogy only as motivation; we do not establish the corresponding spectral structure for masked corruption.

Assumption 2 (Two-scale spectrum of U). Consider the proportional high-dimensional regime

$$\psi_n = \frac{n}{d}, \qquad \psi_p = \frac{p}{d}, \qquad \psi_p \gg \psi_n \gg 1.$$

Assume that the nonzero spectrum of U relevant to the learned predictor separates into two disjoint bands, $\mathcal{B}_{\text{fast}}$ and $\mathcal{B}_{\text{slow}}$, such that

$$\inf \mathcal{B}_{\text{fast}} = \Theta(\psi_p), \qquad \sup \mathcal{B}_{\text{fast}} = \Theta(\psi_p), \tag{44}$$

$$\inf \mathcal{B}_{\text{slow}} = \Theta(\psi_p/\psi_n), \qquad \sup \mathcal{B}_{\text{slow}} = \Theta(\psi_p/\psi_n). \tag{45}$$

Zero-eigenvalue directions are permitted; they do not contribute to the trajectory because $V^\top$ vanishes on $\ker(U)$, as shown in the proof of Lemma 2.

B.6 Timescale separation and scope

For an active eigenmode with eigenvalue $\lambda > 0$, Equation (43) gives the characteristic relaxation scale $\tau(\lambda) \asymp (\kappa_t \lambda)^{-1}$. Define

$$\lambda_{\text{fast}}^{\min} := \inf \mathcal{B}_{\text{fast}}, \qquad \lambda_{\text{slow}}^{\min} := \inf \mathcal{B}_{\text{slow}}, \tag{46}$$

$$\tau_{\text{fast}}^{\text{RF}} := \frac{1}{\kappa_t \lambda_{\text{fast}}^{\min}}, \qquad \tau_{\text{slow}}^{\text{RF}} := \frac{1}{\kappa_t \lambda_{\text{slow}}^{\min}}. \tag{47}$$

Using another fixed quantile or characteristic eigenvalue within each band changes these definitions only by constant factors under Assumption 2.

Lemma 3 (Conditional separation of relaxation times)**.** *Under Assumption 2, in the aspect-ratio regime $\psi_p \gg \psi_n \gg 1$,*

$$\tau_{\text{fast}}^{\text{RF}} = \Theta\left(\frac{1}{\kappa_t \psi_p}\right), \qquad \tau_{\text{slow}}^{\text{RF}} = \Theta\left(\frac{\psi_n}{\kappa_t \psi_p}\right), \tag{48}$$

and consequently

$$\frac{\tau_{\text{slow}}^{\text{RF}}}{\tau_{\text{fast}}^{\text{RF}}} = \Theta(\psi_n) = \Theta\Big(\frac{n}{d}\Big). \tag{49}$$

Proof. By Assumption 2,

$$\lambda_{\text{fast}}^{\min} = \Theta(\psi_p), \qquad \lambda_{\text{slow}}^{\min} = \Theta(\psi_p/\psi_n).$$

Substituting these scales into the definitions above gives the two relaxation times. Their ratio cancels the common gradient-flow normalization κ_t, yielding Equation (49). □

Asymptotic interpretation. Equation (49) is a conditional statement across proportional high-dimensional regimes. For each fixed pair (ψ_n, ψ_p), the underlying limit is

$$n, p, d \to \infty, \qquad \frac{n}{d} \to \psi_n, \qquad \frac{p}{d} \to \psi_p.$$

Thus the $\Theta(\psi_n)$ dependence describes variation across aspect-ratio regimes; it is not a fixed-d, $n \to \infty$ scaling law.

For the normalization in Equation (32), where $\kappa_t = 1/p$, the same spectral scales give

$$\tau_{\text{fast}}^{\text{RF}} = \Theta(d), \qquad \tau_{\text{slow}}^{\text{RF}} = \Theta(n),$$

while their ratio remains $\Theta(n/d)$. A global rescaling of gradient-flow time changes these absolute scales but not their ratio.

Relation to the empirical quantities. The labels "fast" and "slow" refer only to relaxation rates in the random-feature surrogate. Motivated by continuous-diffusion analyses (Bonnaire et al., 2025), one may heuristically associate faster, population-structured modes with population learning and slower, sample-dependent modes with record-specific fitting. This correspondence is not established by the spectral calculation.

In particular, the analysis does not prove that fast-mode relaxation produces low population denoising loss, that slow-mode relaxation causes training-record extraction, or that either random-feature timescale coincides with an empirical onset. Establishing such a connection would require additional assumptions about feature-target alignment and population risk. The squared-loss surrogate likewise does not imply properties of the categorical generative sampler.

Scope. The exact results above characterize gradient-flow dynamics once the feature second-moment matrix U is fixed. The timescale separation in Equation (49) additionally requires the embedding-level corruption assumption (Assumption 1) and the two-scale spectral assumption (Assumption 2). Neither assumption is established for the attention-based Transformer denoisers used in our experiments.

Accordingly, this analysis provides conditional motivation for separated optimization timescales, not a theory that fixes the empirical ordering of population learning and training-record extraction. That ordering is measured independently in our experiments.

C Conditional Analysis of Mask-Schedule Effects

Scope. This appendix studies how the masking-time distribution enters the gradient-flow dynamics of the random-feature surrogate introduced in Section B. The analysis is conditional: it characterizes the schedule-averaged feature operator exactly and then introduces additional structural assumptions when comparing characteristic relaxation rates across masking schedules.

The analysis provides theoretical motivation for the schedule dependence studied empirically in Section 4.5. It does not, by itself, establish that truncating or otherwise reshaping a masking schedule must accelerate training-record extraction in the Transformer models used in our experiments. In particular, an upper bound on a characteristic relaxation time under one schedule is not sufficient to establish an ordering relative to another schedule unless the two schedules are related through assumptions that permit such a comparison.

C.1 SETUP AND NOTATION

We use the random-feature surrogate of Section B. For a fixed target position i, let

$$f_\Theta(z_t) = \frac{1}{\sqrt{p}}\Theta^\top \phi_t(z_t) \in \mathbb{R}^m, \qquad \Theta \in \mathbb{R}^{p\times m}, \tag{50}$$

where $\phi_t(z_t) \in \mathbb{R}^p$ is the random-feature representation of a corrupted context. As in Appendix B, the corruption law is understood to condition on the target position being masked.

Let π_{raw} denote the mask-rate sampling law used to generate training corruptions. The surrogate objective may additionally weight a sampled masking level by a positive deterministic factor (for example, the $1/t$ factor in Eq. (3)). We therefore write ρ for the resulting *effective normalized weighting* over masking levels. In general $\rho \neq \pi_{\mathrm{raw}}$; an overall positive multiplicative constant only rescales gradient-flow time. We write

$$\mathcal{L}_\rho(\Theta) = \mathbb{E}_{t\sim\rho}\left[\mathcal{L}_t(\Theta)\right], \tag{51}$$

where the empirical fixed-t loss is

$$\mathcal{L}_t(\Theta) = \frac{1}{2n}\sum_{\nu=1}^{n} \mathbb{E}_{x_t^\nu\sim q_t(\cdot|x^\nu)}\left[\left\| y^\nu - \frac{1}{\sqrt{p}}\Theta^\top\phi_t(z_t^\nu)\right\|_2^2\right]. \tag{52}$$

Here y^ν is the clean-token regression target at the fixed target position.

For each masking level t, define the corruption-averaged feature second-moment matrix

$$U(t) := \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{x_t^\nu\sim q_t(\cdot|x^\nu)}\left[\phi_t(z_t^\nu)\phi_t(z_t^\nu)^\top\right] \in \mathbb{R}^{p\times p}, \tag{53}$$

and the feature-target cross-moment

$$B(t) := \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{E}_{x_t^\nu\sim q_t(\cdot|x^\nu)}\left[\phi_t(z_t^\nu)\right](y^\nu)^\top \in \mathbb{R}^{p\times m}. \tag{54}$$

These expectations are over corruption randomness conditional on the training dataset and on the fixed random-feature matrix W.

The schedule-averaged operators are

$$U_\rho := \mathbb{E}_{t\sim\rho}[U(t)], \tag{55}$$
$$B_\rho := \mathbb{E}_{t\sim\rho}[B(t)]. \tag{56}$$

C.2 THE SCHEDULE-AVERAGED GRADIENT-FLOW OPERATOR

Lemma 4 (Gradient flow under masking-time mixtures)**.** *For the objective in Equation (51), gradient flow satisfies*

$$\frac{d\Theta(\tau)}{d\tau} = -\frac{1}{p}U_\rho\Theta(\tau) + \frac{1}{\sqrt{p}}B_\rho. \tag{57}$$

Thus the masking schedule enters the linear relaxation operator through the averaged second-moment matrix U_ρ.

Proof. Fix t. Expanding Equation (52) before taking the corruption expectation gives, for each example ν,

$$\begin{aligned}
&\mathbb{E}_{x_t^\nu}\left[\left\| y^\nu - \frac{1}{\sqrt{p}}\Theta^\top\phi_t(z_t^\nu)\right\|_2^2\right] \\
&\quad = \|y^\nu\|_2^2 - \frac{2}{\sqrt{p}}(y^\nu)^\top\Theta^\top\mathbb{E}_{x_t^\nu}[\phi_t(z_t^\nu)] \\
&\qquad + \frac{1}{p}\operatorname{Tr}\big(\Theta^\top\mathbb{E}_{x_t^\nu}[\phi_t(z_t^\nu)\phi_t(z_t^\nu)^\top]\Theta\big).
\end{aligned} \tag{58}$$

Averaging over the training examples and differentiating with respect to Θ yields

$$\nabla_\Theta \mathcal{L}_t(\Theta) = \frac{1}{p} U(t)\Theta - \frac{1}{\sqrt{p}} B(t). \tag{59}$$

Averaging over $t \sim \rho$ and exchanging differentiation with the finite or integrable expectation gives

$$\nabla_\Theta \mathcal{L}_\rho(\Theta) = \mathbb{E}_{t\sim\rho}\left[\nabla_\Theta \mathcal{L}_t(\Theta)\right] \tag{60}$$

$$= \frac{1}{p}\mathbb{E}_{t\sim\rho}[U(t)]\Theta - \frac{1}{\sqrt{p}}\mathbb{E}_{t\sim\rho}[B(t)] \tag{61}$$

$$= \frac{1}{p} U_\rho \Theta - \frac{1}{\sqrt{p}} B_\rho. \tag{62}$$

Therefore gradient flow $d\Theta/d\tau = -\nabla_\Theta \mathcal{L}_\rho(\Theta)$ gives Equation (57). □

Consequence. For a fixed schedule ρ, the relaxation rate of an eigenmode of U_ρ with eigenvalue $\lambda > 0$ is proportional to λ/p. Hence its characteristic relaxation time satisfies

$$\tau_\rho(\lambda) \asymp \frac{p}{\lambda}. \tag{63}$$

This statement is exact at the level of the linear random-feature surrogate. It does not yet imply any ordering between two different masking schedules: such a comparison requires information about how their respective operators U_{ρ_1} and U_{ρ_2} differ spectrally.

C.3 Spectral comparison of schedule mixtures

We next ask what can be inferred about the spectrum of

$$U_\rho = \mathbb{E}_{t\sim\rho}[U(t)]$$

from the fixed-t operators $U(t)$. We first use matrix inequalities that hold for arbitrary positive-semidefinite $U(t)$ and then state separately any additional assumptions needed to interpret a particular spectral band as relevant to record-specific fitting.

The exact result above shows that a masking schedule affects the random-feature relaxation dynamics through

$$U_\rho = \mathbb{E}_{t\sim\rho}[U(t)].$$

For arbitrary positive-semidefinite matrices $U(t)$, one always has the global inequality

$$\lambda_{\min}(U_\rho) \geq \mathbb{E}_{t\sim\rho}\left[\lambda_{\min}(U(t))\right], \tag{64}$$

because $\lambda_{\min}$ is concave on the cone of symmetric matrices. This bound concerns the global smallest eigenvalue and, by itself, does not identify a spectral component associated with record-specific fitting.

To obtain a directional schedule comparison for such a component, we therefore impose an additional structural assumption.

Assumption 3 (Shared schedule-sensitive spectral component). There exists a fixed subspace $\mathcal{H}_{\text{slow}} \subseteq \mathbb{R}^p$, with orthogonal projector P_{slow}, that is reducing for every $U(t)$:

$$P_{\text{slow}} U(t) = U(t) P_{\text{slow}} = P_{\text{slow}} U(t) P_{\text{slow}}. \tag{65}$$

On $\mathcal{H}_{\text{slow}}$, assume

$$U(t)|_{\mathcal{H}_{\text{slow}}} = s(t) I_{\mathcal{H}_{\text{slow}}} + v(t) S, \tag{66}$$

where $s(t), v(t) \geq 0$ are deterministic scalar functions and S is a fixed positive-definite operator on $\mathcal{H}_{\text{slow}}$ with

$$\lambda_S^- := \lambda_{\min}(S) > 0. \tag{67}$$

Assumption 3 is an analytical surrogate rather than a consequence of the masked-DLM architecture. Its substantive content is that the same invariant schedule-sensitive subspace and the same operator S apply across masking levels, while $s(t)$ and $v(t)$ change their strengths. Without such joint alignment,

spectral edges of an average of matrices cannot in general be obtained by averaging the spectral edges of the individual matrices.

Under Assumption 3, define

$$s_\rho := \mathbb{E}_{t\sim\rho}[s(t)], \qquad v_\rho := \mathbb{E}_{t\sim\rho}[v(t)]. \tag{68}$$

Then

$$P_{\mathrm{slow}} U_\rho P_{\mathrm{slow}} = s_\rho P_{\mathrm{slow}} + v_\rho S, \tag{69}$$

and the smallest active eigenvalue of this component is

$$\lambda^-_{\mathrm{slow}}(\rho) = s_\rho + v_\rho \lambda^-_S. \tag{70}$$

C.4 Schedules truncated away from nearly clean contexts

We first derive an absolute bound for schedules whose support excludes a neighborhood of $t = 0$.

Assumption 4 (Nondegeneracy away from $t = 0$). For every $t_{\min} > 0$, there exist constants $\underline{s}(t_{\min}) > 0$ and $\underline{v}(t_{\min}) \geq 0$, uniform in the high-dimensional limit, such that

$$s(t) \geq \underline{s}(t_{\min}), \qquad v(t) \geq \underline{v}(t_{\min}), \qquad t \in [t_{\min}, 1]. \tag{71}$$

If ρ is supported on $[t_{\min}, 1]$, then

$$s_\rho \geq \underline{s}(t_{\min}), \qquad\qquad v_\rho \geq \underline{v}(t_{\min}), \tag{72}$$

and therefore

$$\lambda^-_{\mathrm{slow}}(\rho) \geq \underline{s}(t_{\min}) + \underline{v}(t_{\min})\lambda^-_S =: c(t_{\min}) > 0. \tag{73}$$

Under the normalization of Equation (57), an eigenmode with eigenvalue $\lambda > 0$ relaxes on the characteristic time p/λ. Hence define

$$\tau^{\mathrm{RF}}_{\mathrm{slow}}(\rho) := \frac{p}{\lambda^-_{\mathrm{slow}}(\rho)}. \tag{74}$$

Equation (73) then gives

$$\tau^{\mathrm{RF}}_{\mathrm{slow}}(\rho) \leq \frac{p}{c(t_{\min})}. \tag{75}$$

This is an absolute bound for schedules truncated away from $t = 0$. It does not, by itself, prove that such a schedule has a shorter relaxation time than a particular untruncated schedule. A directional comparison requires additional information about the behavior of $s(t)$ and $v(t)$ near $t = 0$.

C.5 Adding schedule mass near $t = 0$

We next consider an explicit pairwise comparison. Let

$$\rho_2 = (1-\alpha)\rho_1 + \alpha\rho_0, \qquad \alpha \in (0, 1), \tag{76}$$

where ρ_0 is supported on $[0, t_0]$. Thus ρ_2 is obtained from ρ_1 by allocating an additional fraction α of its mass to nearly clean masking levels.

Assumption 5 (Near-zero decay). The functions $s(t)$ and $v(t)$ are continuous at $t = 0$ and satisfy

$$s(t) \to 0, \qquad v(t) \to 0, \qquad t \downarrow 0. \tag{77}$$

This assumption is not derived from the masking process or from the Transformer architecture; it is an additional structural condition on the random-feature surrogate.

Let

$$\lambda_1 := \lambda^-_{\mathrm{slow}}(\rho_1), \qquad \lambda_0 := \lambda^-_{\mathrm{slow}}(\rho_0).$$

If $\lambda_1 > 0$, then by Assumption 5, t_0 can be chosen sufficiently small that

$$\lambda_0 \leq \lambda_1.$$

Table 4: **Population-threshold robustness of the hierarchical ordering.** The ordering obtained with the lower population-learning threshold $G \geq 0.25$ is unchanged from the primary $G \geq 0.50$ analysis. Counts denote the number of seeds exhibiting the indicated ordering.

Repetition R	Ordering at $G \geq 0.25$
1	Population first (3/3)
4	Population first (3/3)
16	Extraction first (3/3)

Since both s_ρ and v_ρ are linear in the schedule,

$$\begin{aligned}\lambda^-_{\text{slow}}(\rho_2) &= s_{\rho_2} + v_{\rho_2}\lambda^-_S \\ &= (1-\alpha)\lambda^-_{\text{slow}}(\rho_1) + \alpha\lambda^-_{\text{slow}}(\rho_0) \\ &\leq \lambda^-_{\text{slow}}(\rho_1). \end{aligned} \tag{78}$$

Consequently,

$$\tau^{\text{RF}}_{\text{slow}}(\rho_2) \geq \tau^{\text{RF}}_{\text{slow}}(\rho_1). \tag{79}$$

Thus, under Assumptions 3 and 5, adding sufficiently concentrated probability mass near $t = 0$ slows the relaxation of the schedule-sensitive RF component. Equivalently, removing such mass can accelerate this component for schedule pairs related by Equation (76).

This is a restricted pairwise result. It does not order arbitrary masking distributions solely from their means, variances, or amount of mass in a particular interval.

C.6 Interpretation and scope

The schedule analysis yields two conditional conclusions. First, under Assumptions 3 and 4, schedules supported away from $t = 0$ admit an absolute upper bound on the characteristic relaxation time of the schedule-sensitive RF component. Second, under the additional near-zero decay assumption Assumption 5, adding sufficiently concentrated mass near $t = 0$ within the mixture construction of Equation (76) weakly increases that relaxation time.

Neither statement identifies $\tau^{\text{RF}}_{\text{slow}}(\rho)$ with the empirical training-record extraction onset. Such an interpretation would require additional assumptions connecting the schedule-sensitive RF subspace to record-specific fitting, and these assumptions are not established for the Transformer denoisers studied experimentally.

Accordingly, this appendix provides a conditional mechanism through which mask-schedule shape can alter optimization timescales; it does not establish a general ordering over arbitrary masking distributions. The empirical schedule experiments test schedule dependence directly and do not rely on this surrogate analysis for their validity.

D Controlled-Study Robustness and Long-Horizon Results

Threshold sensitivity. We recompute onset statistics over the prespecified threshold grid without retraining. Across the controlled dataset-size sweep, the fitted extraction-onset exponents remain stable: approximately 0.933-0.957 across the evaluated extraction criteria. Across the repetition sweep, the corresponding exponents remain approximately -0.462–-0.482. No qualitative sign changes occur. For near-verbatim extraction, we additionally evaluate normalized-edit-distance thresholds 0.10, 0.05, and 0.01.

As shown in Table 4, lowering the population-learning threshold from the primary $G \geq 0.50$ criterion to $G \geq 0.25$ leaves the qualitative ordering unchanged: population learning precedes extraction for $R = 1$ and $R = 4$, whereas extraction precedes population learning for $R = 16$ in all three seeds.

Held-out confirmation of the hierarchical ordering. Population-learning onsets in the hierarchical experiment are selected using validation data. To verify that the population-first result at low repetition

Table 5: Held-out test population gain at the validation-defined $G \geq 0.50$ population-onset checkpoints in the hierarchical experiment. Canary-control extraction advantage is zero at all six replayed checkpoints.

	G_{test}		
R	Seed 101	Seed 202	Seed 303
1	0.480	0.590	0.536
4	0.560	0.553	0.493

Table 6: Finite-range extraction-onset fits. Entries are descriptive exponents α and 95% bootstrap intervals for $\tau \propto n^{\alpha}$ or $\tau \propto R^{\alpha}$.

Setup / varied quantity	Endpoint	α	95% CI
Controlled / dataset size	Exact	0.943	[0.809, 1.061]
	Near	0.957	[0.807, 1.089]
	Advantage	0.940	[0.807, 1.061]
Controlled / repetition	Exact	-0.462	$[-0.515, -0.416]$
	Near	-0.474	$[-0.530, -0.426]$
	Advantage	-0.468	$[-0.515, -0.429]$
LLaDA / dataset size	Exact	0.670	[0.470, 0.868]
	Near	0.800	[0.602, 0.995]
	Advantage	0.670	[0.470, 0.868]

is not specific to that split, we replay the validation-defined $G \geq 0.50$ onset checkpoints on the held-out test set.

As shown in Table 5, the held-out test population gain remains substantial at the validation-selected onset checkpoints: $G_{\text{test}} = 0.480$–0.590 for $R = 1$ and 0.493–0.560 for $R = 4$. Canary-control extraction advantage is zero at all six checkpoints. Thus, the low-repetition population-first separation is also present on unseen test data rather than being an artifact of the validation split used for onset selection.

Finite-range one-dimensional scaling fits. For completeness, Table 6 reports the original one-dimensional onset fits that motivated the later joint intervention. They summarize only the tested ranges and do not isolate record presentation frequency.

For the controlled fits, the dataset-size sweep uses $n \in \{128, 256, 512, 1024, 2048, 4096\}$ at fixed $R = 16$, and the repetition sweep uses $R \in \{1, 2, 4, 8, 16, 32, 64\}$ at fixed $n = 128$, with three seeds per condition. The reported intervals use the frozen paired-observation bootstrap with 2,000 resamples. They should not be read as evidence from 18 or 21 independent seed trajectories.

Long-horizon Markov evaluation. We evaluate the baseline Markov configuration for 12,000 optimization steps using seeds 101, 202, and 303. Across 900 evaluated checkpoints, the Bayes-referenced population-learning criterion is never reached: held-out loss does not fall below the chance reference in any seed. In contrast, near-verbatim extraction occurs at step 240 in all three seeds, while exact and extraction-advantage onsets occur at step 280. Matched-control extraction remains zero throughout.

The best held-out losses remain 46.5–48.2% above their corresponding Bayes references. Thus, within the tested horizon, this configuration exhibits early train-specific extraction without an observed meaningful population-learning phase. We interpret this as a contrasting regime, not as evidence that population learning could never emerge at a longer training horizon.

One-factor robustness of the Markov regime. We test whether the extraction-without-population-learning behavior of the baseline Markov configuration is sensitive to individual modeling or training choices. Starting from the same $n = 128$, $R = 16$ reference configuration, we vary one factor at a time: vocabulary size, learning rate, model capacity, or mask-rate support. Table 7 recomputes

Table 7: One-factor robustness of the contrasting Markov regime. Extraction entries are $\tau_{\text{mem}}^{\text{adv}}$ for seeds 101/202/303 under the current canary-control advantage threshold of 0.10. Best held-out loss is the minimum attained during each run; matched-control extraction remains zero throughout all 15 runs.

Configuration	$\tau_{\text{mem}}^{\text{adv}}$	Best held-out loss	Chance
Baseline	$280/280/280$	$5.581/5.582/5.574$	5.545
Vocabulary 64	$260/260/210$	$4.133/4.157/4.153$	4.159
Learning rate 5×10^{-5}	$700/700/700$	$5.601/5.626/5.611$	5.545
Small model	$1140/1130/1180$	$5.554/5.564/5.550$	5.545
Low-mask-support schedule	$230/240/220$	$5.607/5.625/5.599$	5.545

Table 8: Hierarchical mask-schedule robustness at fixed $R = 16$. Entries are seed-level onsets for seeds 101/202/303. A dash denotes that the $G \geq 0.50$ population-learning criterion is not reached within 20,000 optimization steps.

Training mask schedule	$\tau_{\text{mem}}^{\text{adv}}$	$\tau_{\text{pop}}(0.50)$
$\text{Beta}(4, 4)$	$160/180/220$	$1080/1160/1120$
$\text{Beta}(5, 1)$	$400/420/340$	$-/-/-$

extraction onset from the saved trajectories using the same current canary-control advantage threshold of 0.10 used elsewhere in the paper.

As shown in Table 7, train-specific extraction is observed in every configuration and seed, although its onset can shift substantially with learning rate and model capacity. Matched controls remain unextracted in all 15 runs. At the same time, none of these interventions produces a meaningful preceding population-learning phase: the best held-out losses remain at or near the corresponding chance reference. In the vocabulary-64 condition, the small numerical decrease below chance remains far from the independently computed Bayes-optimal population loss, so it does not constitute substantial population learning. Thus the qualitative Markov contrast is not specific to the baseline vocabulary size, learning rate, model capacity, or masking support.

Additional data-regime stress tests. We additionally test whether the extraction-before-population-learning behavior of the first-order Markov configuration persists under other data regimes.

For a second-order Markov population, all three seeds remain censored for both the $G \geq 0.25$ and $G \geq 0.50$ Bayes-referenced population-learning criteria over 20,000 optimization steps. Extraction-advantage onset occurs at step 280 in all three seeds, while exact-extraction onset occurs at steps 280, 240, and 280 for seeds 101, 202, and 303, respectively. Matched-control extraction remains zero throughout, while trained-canary extraction eventually reaches one.

We also evaluate a text8 background as a natural-text stress test. Because a tractable Bayes population reference is unavailable in this setting, this experiment is not directly comparable to the controlled Markov and hierarchical populations. Under the prespecified relative held-out-loss improvement criterion, no onset is observed in any of the three seeds over 20,000 steps. Importantly, this relative criterion is anchored at the first recorded evaluation (step 40), not at the untrained network: held-out loss is already below the uniform-chance reference at that point and remains below chance for thousands of steps. Thus the censored relative onset must not be read as absence of population competence. Extraction-advantage onset occurs at steps 240, 280, and 240, while near-verbatim extraction occurs at steps 200, 240, and 200. Final trained-canary extraction reaches one and matched-control extraction remains zero. We therefore treat text8 only as a stress test showing early extraction alongside an already-competent held-out model, not as a Bayes-referenced ordering experiment.

Hierarchical mask-schedule robustness. We additionally evaluate two Beta masking schedules in the hierarchical-grammar setting at fixed $R = 16$, where both population learning and extraction can be measured using the same Bayes-referenced protocol as in the primary ordering experiment.

Table 9: Prespecified schedule-shape regressions. The baseline using only $\mathbb{E}[t]$ and $\mathrm{Var}(t)$ has LOSO $R^2 = -0.034$. Each row adds one candidate to that same baseline.

Candidate	LOSO R^2	ΔR^2	95% coefficient CI
M_{low}	0.667	0.700	[1.890, 2.364]
M_{mid}	0.762	0.796	[−1.081, −0.835]
M_{high}	0.175	0.209	[1.314, 1.712]
Mean uniqueness	−1.522	−1.489	[−1.462, −1.027]
Ambiguous uniqueness mass	0.217	0.250	[7.423, 9.598]
Conditional information	−0.621	−0.587	[−2.621, −1.971]

As shown in Table 8, extraction precedes population learning in all three $\mathrm{Beta}(4,4)$ runs, with a substantial separation between the two onsets. Under $\mathrm{Beta}(5,1)$, extraction is again observed in every seed, whereas the $G \geq 0.50$ population-learning criterion remains right-censored in all three runs over the 20,000-step horizon. Thus, changing the training mask distribution alters the learning-extraction separation even in the hierarchical regime where Bayes-referenced population learning is measurable.

D.1 Mask-schedule shape analysis and matched-moment protocol

Thirteen-schedule analysis. The schedule-shape analysis contains 13 conditions, each evaluated with seeds 101/202/303. Nine new conditions use

$$t = 0.02 + 0.98Z, \qquad Z \sim \mathrm{Beta}(a,b),$$

with

$$(a,b) \in \{(0.5,0.5),(2,2),(4,4),(1,3),(1,5),(3,1),(5,1),(2,5),(5,2)\}.$$

The four reused anchor schedules are uniform, low-mask-heavy, high-mask-heavy, and truncated-away-from-clean. The reference uniform schedule is $\mathrm{Unif}[0.02,1]$ and the truncated schedule is $\mathrm{Unif}[0.3,1]$; the low/high-mask-heavy production samplers use $t_{\min} = 0.02$, $t_{\max} = 1$, and skew parameter $k = 3$. All 39 schedule-seed rows have observed extraction-advantage onsets.

For each schedule we compute the first two mask-rate moments together with

$$M_{\mathrm{low}} = \Pr(t < 0.1), \qquad M_{\mathrm{mid}} = \Pr(0.2 < t < 0.8), \qquad M_{\mathrm{high}} = \Pr(t > 0.9). \tag{80}$$

We additionally estimate three population/context statistics: the mean finite-sample uniqueness probability, the probability that uniqueness lies in $(0.1, 0.9)$, and the mean conditional-information quantity $\mathrm{KL}(P_0(X_i \mid X_S) \| P_0(X_i))$ over masked positions. These Monte Carlo statistics use 200,000 draws per schedule and seed.

The prespecified response is $\log \tau^{\mathrm{adv}}_{\mathrm{mem}}$. A baseline linear model uses $\mathbb{E}[t]$ and $\mathrm{Var}(t)$; each candidate shape statistic is added one at a time. Evaluation is leave-one-schedule-out, holding out all three seeds of a schedule together. The coefficient uncertainty uses 10,000 seed-within-schedule bootstrap resamples with RNG seed 20260908. A candidate passes the prespecified criterion only if it improves LOSO R^2 by more than 0.10 and its coefficient interval excludes zero.

Several correlated schedule statistics pass the predictive criterion, with M_{mid} giving the largest LOSO improvement among those tested. This analysis therefore establishes predictive schedule-shape dependence within the tested family, not a unique causal statistic.

Prospective matched-moment schedules. The two prospectively specified schedules in Figure 2 are mixtures of uniforms:

$$P^\star = 0.475\,\mathrm{Unif}[0.25,0.27] + 0.475\,\mathrm{Unif}[0.75,0.77] + 0.05\,\mathrm{Unif}[0.02,1], \tag{81}$$

$$Q^\star = 0.6567901235\,\mathrm{Unif}[0.50,0.52] + 0.1466049383\,\mathrm{Unif}[0.05,0.07] \tag{82}$$

$$+ 0.1466049383\,\mathrm{Unif}[0.95,0.97] + 0.05\,\mathrm{Unif}[0.02,1]. \tag{83}$$

Both have mean 0.51, variance $7609/120000$, and support $[0.02,1]$. The comparison uses the Markov $n = 128$, $K = 8$, $R = 16$ configuration, 20,000 training steps, batch size 32, and seeds 101/202/303. Extraction is evaluated every 40 steps, and paired runs share initialization and data-sampling streams.

Extraction-advantage onsets are $(240, 200, 200)$ for $P^\star$ and $(280, 280, 280)$ for $Q^\star$. Both population-learning criteria remain censored in all six runs. Because the mixtures also differ in low- and high-mask tail mass, this experiment shows that the first two moments are insufficient but does not identify the responsible higher-order feature.

E Conditioning-Geometry Evaluation

The controlled geometry study evaluates five conditioning geometries over six reveal fractions, three checkpoints, and three seeds. Considering the 72 step-reveal-fraction comparisons, 40 cells are all-zero ties across seeds, 19 contain both positive and negative seed-level differences, four contain positive differences and ties but no negative differences, and four contain negative differences and ties but no positive differences. Among the remaining five cells with strictly sign-consistent differences across all three seeds, four favor the comparator and one favors scattered conditioning.

Thus no conditioning geometry is uniformly the most extractive across training times, reveal fractions, and seeds. The corresponding LLaDA checkpoint-reuse analysis uses only four canaries per run and is treated as supplementary evidence; it likewise does not support a universal geometry ordering.

F Pretrained-Model Stress Tests

F.1 LLaDA-8B

All experiments in this section use LLaDA-8B-Base with parameter-efficient LoRA fine-tuning. The primary adapter has rank 8, LoRA $\alpha = 16$, and modifies the attention query, key, value, and output projections, yielding 8,388,608 trainable parameters, approximately 0.1045% of the roughly 8.02B model parameters. The model is loaded in bf16 and optimized at learning rate 10^{-4} for 2,500 steps with batch size 2. Extraction is evaluated every 40 training steps. The template-background reference uses $n = 128$, four canaries, $R = 8$, evaluation masking rate $t_0 = 0.3$, reveal fraction 0.5, and scattered conditioning. Unless otherwise stated, the seeds are 101, 202, and 303.

Dataset-size scaling. For

$$n \in \{32, 64, 128, 256, 512\}$$

and three seeds, all 15 runs have observed first-crossing extraction onsets, whereas only 3/15 have an uncensored population-learning onset. The corresponding finite-range first-crossing exponents are approximately

$$0.670$$

for exact extraction and extraction advantage, and

$$0.800$$

for near-verbatim extraction. These values are onset-definition sensitive: requiring a sustained extraction crossing increases the exact/advantage dataset-size exponent to approximately 0.86. We therefore use this experiment only for the directional conclusion that extraction tends to occur later as n increases, not for a universal cross-model scaling exponent.

Repetition. At $n = 128$ with

$$R \in \{1, 8, 32\},$$

increasing repetition produces earlier extraction in every seed for the primary extraction-advantage criterion on the template background. The fitted onset exponent is approximately -0.70 to -0.79, depending on the extraction criterion.

We repeat the three-seed design using WikiText-2 as background text with synthetic canaries. The fitted exponent is

$$\alpha = -0.896$$

for exact/extraction-advantage onset (95% CI $[-1.043, -0.714]$, $R^2 = 0.951$) and

$$\alpha = -0.751$$

for near-verbatim onset (95% CI $[-0.864, -0.539]$, $R^2 = 0.955$). Matched controls remain unextracted. Thus the direction of the repetition effect transfers across both pretrained backgrounds.

Table 10: MDLM repetition stress test. Entries are onset steps for seeds 101/202/303. Exact extraction and extraction advantage coincide under this protocol because matched-control extraction remains zero.

R	$\tau_{\text{mem}}^{\text{adv}}$	$\tau_{\text{mem}}^{\text{near}}$
1	320/200/240	640/480/480
8	80/120/80	200/240/160
32	40/40/40	200/80/80

Mask schedule. For the LLaDA schedule sweep, the mean extraction-advantage onsets are approximately

$$547,\ 627,\ 893,\ 933$$

for uniform, truncated-away-from-clean, high-mask-heavy, and low-mask-heavy schedules, respectively. Five of the six mean-level pairwise directions agree with the controlled experiment, including low-mask-heavy as the slowest condition.

The transfer is nevertheless weak at the seed level: ranges overlap substantially across arms, and with three seeds per schedule no individual pairwise contrast provides strong evidence of separation. We therefore interpret LLaDA as suggestive of schedule dependence, not as replication of the controlled ordering.

LoRA rank. Across ranks

$$\{4, 8, 16, 32\},$$

extraction is detected in all 12 rank-seed runs and matched controls remain unextracted. The fitted log-onset slope against $\log_2$ rank is approximately

$$-0.084 \qquad (95\%\ \text{CI}\ [-0.266, 0.098]),$$

providing no detectable rank dependence over the evaluated range.

Matched dataset-size range. Restricting the controlled and LLaDA dataset-size fits to their shared range

$$n \in \{128, 256, 512\}$$

makes the apparent exponent difference unstable and produces overlapping uncertainty intervals. We therefore do not attribute the difference between the full-range controlled and pretrained fits causally to pretraining, model scale, or LoRA fine-tuning.

F.2 MDLM

We additionally test full-parameter adaptation of the pretrained MDLM model. The reported repetition and schedule experiments run for 2,500 training steps with seeds 101/202/303 and use four trained canaries together with matched never-trained controls. Extraction uses one greedy reconstruction per canary. Under this protocol, control extraction is zero throughout; consequently, the exact-extraction and canary-control-advantage thresholds coincide, and we do not treat these two endpoints as independent corroborating measurements. Optimizer and batch-size details are left to the accompanying released configuration rather than inferred here from the result artifacts.

Repetition. Under uniform masking, increasing canary repetition consistently accelerates extraction. Table 10 reports the per-seed onsets.

The mean extraction-advantage onset decreases from 253.3 to 93.3 to 40.0 steps as R increases from 1 to 8 to 32, and the direction holds separately in every seed. Near-verbatim onset also decreases in mean from 533.3 to 200.0 to 120.0 steps; seed 101 ties between $R = 8$ and $R = 32$ on this secondary endpoint. Thus the directional repetition effect transfers robustly to full-parameter MDLM adaptation, while the three repetition levels are insufficient to establish a general scaling law.

Table 11: Primary MDLM mask-schedule comparison. Entries are onset steps for seeds 101/202/303. Exact extraction coincides with extraction advantage under the evaluated protocol.

Schedule	$\tau^{\text{adv}}_{\text{mem}}$	$\tau^{\text{near}}_{\text{mem}}$
Uniform	$80/120/80$	$200/240/160$
Interior-centered	$80/80/80$	$160/160/160$
Low-mask-heavy	$120/80/80$	$320/160/200$
High-mask-heavy	$80/120/80$	$280/200/200$

Primary mask-schedule comparison. At fixed $R = 8$, we compare uniform, interior-centered, low-mask-heavy, and high-mask-heavy training-mask distributions. Table 11 gives the seed-level onsets under the prespecified evaluation cadence.

The corresponding mean extraction-advantage onsets are 93.3, 80.0, 93.3, and 93.3 steps, respectively. Hence three of the four condition means are tied under the primary evaluation grid. Only three of the six pairwise mean comparisons are therefore resolved, below the prespecified requirement of at least four resolved comparisons. We consequently retain the primary MDLM schedule result as *inconclusive*; the absence of censoring does not remove the lack of pairwise resolution.

Post-hoc dense-evaluation sensitivity. We additionally perform a supplementary reconstruction with extraction evaluations every five training steps over the early-training region. The resulting mean extraction-advantage onsets are 70.0, 66.7, 76.7, and 75.0 steps for uniform, interior-centered, low-mask-heavy, and high-mask-heavy schedules, respectively. At this denser evaluation, all six mean-level pairs are numerically resolved and five of six follow the controlled-study direction.

This analysis is strictly post hoc and does not replace the primary result. Increasing the evaluation cadence also changes the sequence of stochastic extraction queries, so the dense evaluation is not a clean isolation of temporal resolution. Indeed, some seed-level comparisons that were already non-tied under the primary evaluation change sign under the dense evaluation. We therefore interpret the dense result as a sensitivity analysis showing substantial query-level variability, rather than as evidence that the primary ties were solely caused by coarse evaluation cadence. The prespecified MDLM schedule conclusion remains inconclusive.

G Context Uniqueness: Calibration and Neural Extractability

The context-uniqueness theorem is a statement about iid sampling from the population and the resulting empirical conditional. It does not imply that a finite neural denoiser will exploit a unique context, nor that uniqueness determines when extraction appears during training.

Finite-sample calibration. We evaluate the closed-form uniqueness probability using 50,000 probes for each of $n \in \{128, 1024, 4096\}$, concentrating the reveal grid on nontrivial predicted probabilities. In the resulting hard bands, mean predicted and empirical uniqueness probabilities are closely aligned: $0.599/0.603$, $0.727/0.723$, and $0.924/0.923$, respectively. The corresponding Brier scores are 0.236, 0.084, and 0.067. We use these results as calibration evidence only; threshold accuracy becomes strongly base-rate dependent, particularly at the largest dataset size, and is not interpreted as discrimination performance.

Neural predictive test. We construct 2,160 observations spanning three seeds, three checkpoints, five conditioning geometries, six reveal fractions, and eight canaries. The prespecified models ask whether context self-information $I(S) = -\log q_S(x_S)$ improves leave-one-seed-out prediction of neural extraction beyond reveal fraction. It does not: adding $I(S)$ changes held-out R^2 by approximately -0.0002 without geometry covariates and $+0.00004$ with them, and the bootstrap coefficient interval $[-0.00264, 0.00025]$ contains zero.

This negative result has important identifiability limits. Every one of the 2,160 evaluated contexts is empirically unique, so an empirical- uniqueness indicator has zero variance and cannot support an association test. Moreover, $I(S)$ is almost perfectly correlated with reveal fraction ($r \approx 0.998$),

leaving little independent variation, and the prespecified regressions do not model training step even though extraction changes sharply across the three checkpoints. Under this test, $I(S)$ therefore adds essentially no predictive value beyond reveal fraction. Because the tested contexts are already unique and the covariates are nearly collinear, this negative result is not a general falsification of uniqueness-based neural mechanisms.

H Experimental and Reproducibility Details

Controlled populations and model. All controlled sequences have length 64. The first-order Markov population uses vocabulary size 256 and transition concentration 0.3. The hierarchical population uses a depth-3 grammar with branching factor 4 and 16 terminal symbols. Its canaries and controls are i.i.d. random 64-token strings over the same 16-symbol alphabet and are therefore out-of-distribution relative to the grammar. Controls are generated by the same canary process but are never included in training.

Both controlled settings use the same bidirectional pre-LN Transformer: six layers, $d_{\text{model}} = 384$, six attention heads, feed-forward width 1536, and approximately 10.87M parameters. Optimization uses AdamW with learning rate 3×10^{-4} and batch size 32. The reference uniform training schedule is $t \sim \text{Unif}[0.02, 1]$. Primary seeds are 101, 202, and 303. Unless stated otherwise, controlled intervention runs use a 20,000-step horizon. Population learning in the hierarchical experiments is evaluated on the common 40-step grid defined in Sec. 3; extraction uses the saved native evaluation grid.

Extraction protocol. Exact, near-verbatim, and canary-control advantage onsets use the thresholds in Sec. 3. Matched controls receive the same conditioning and decoding protocol as trained canaries. Unobserved crossings are right-censored and are never replaced by the final checkpoint. The reported Phase-G and Phase-K controlled runs have zero matched-control exact extraction at the relevant crossings.

Phase G: joint dataset-size-repetition grid. The base grid fixes $K = 8$ and evaluates

$$n \in \{128, 256, 512\}, \qquad R \in \{1, 4, 16, 64\},$$

with all three seeds, for 36 runs. Five additional conditions, $(n, R) \in \{(256, 2), (256, 8), (128, 8), (512, 32), (256, 32)\}$, again use all three seeds. Together these form 17 conditions and 51 seed-level runs. Only n, R, and the seed change relative to the frozen hierarchical base configuration; the model, optimizer, grammar, masking schedule, and evaluation protocol are held fixed.

The added cells complete four exactly matched normalized-frequency families:

R/n	p_{can}	(n, R) cells
$1/128$	$1/136 \approx 0.00735$	$(128, 1), (256, 2), (512, 4)$
$1/32$	0.025	$(128, 4), (256, 8), (512, 16)$
$1/16$	$1/24 \approx 0.04167$	$(128, 8), (256, 16), (512, 32)$
$1/8$	0.0625	$(128, 16), (256, 32), (512, 64)$

The primary ordering statistic is $\Delta = \log(\tau^{\text{adv}}_{\text{mem}}/\tau_{\text{pop}}(0.50))$. One run, $(n, R) = (256, 8)$ with seed 303, has $\Delta = 0$ at the evaluation-grid resolution and is treated as a tie. Descriptive regressions use all 51 seed rows; uncertainty for the Phase-G regression analysis resamples whole (n, R) cells, keeping their three seeds together.

Phase K: decoupling interventions. Phase K varies K using nested canary identities C1–C16 and nested never-trained controls U1–U16. The primary cross-K extraction endpoint is always computed on C1–C4 versus U1–U4, so changing the total number of distinct canaries does not change the endpoint resolution. Each cell uses seeds 101/202/303. The four intervention families are

Family	n	K	R	p_{can}	f_{can}
A1	256	4	16	0.05000	0.20
	256	8	8	0.02500	0.20
	256	16	4	0.01250	0.20
A2	256	4	32	0.08333	1/3
	256	8	16	0.04167	1/3
	256	16	8	0.02083	1/3
B1	288	4	8	0.02500	0.10
	288	8	9	0.02500	0.20
	288	16	12	0.02500	0.40
B2	320	4	16	1/24	1/6
	320	8	20	1/24	1/3
	320	16	40	1/24	2/3

Thus A1/A2 hold aggregate repeated-data mass exactly fixed while changing per-record sampling probability, whereas B1/B2 hold per-record sampling probability exactly fixed while changing aggregate repeated-data mass. Six seed-level runs are reused from Phase G where the configurations are identical; the remaining runs use the same frozen hierarchical apparatus. Because $f_{\text{can}} = Kp_{\text{can}}$, every intervention that varies one quantity while fixing the other also changes K; Phase K therefore does not uniquely identify a mechanism attributable to p_{can} or f_{can} alone.

One-dimensional and masking interventions. The earlier controlled dataset-size sweep uses $n \in \{128, 256, 512, 1024, 2048, 4096\}$ at fixed $R = 16$, and the repetition sweep uses $R \in \{1, 2, 4, 8, 16, 32, 64\}$ at fixed $n = 128$, with three seeds per condition. The 13-schedule shape analysis and the exact $P^\star$, $Q^\star$ mixtures are specified in Appendix D.1. The hierarchical Beta-schedule robustness experiments use the same hierarchical population and $R = 16$ protocol as the primary ordering study.

Compute. Controlled full-parameter experiments were run on 48GB NVIDIA RTX 6000- and L40S-class GPUs. LLaDA-8B-Base experiments use bf16 model loading with LoRA fine-tuning on 48GB GPUs. The reported pretrained sweep for which direct accounting was available consumed approximately 2.54 measured GPU-hours. Context-collision, bootstrap, regression, and other frozen-artifact analyses were performed on CPU where applicable.