SEMANTIC DIFFUSION LANGUAGE MODELS VIA LOW-RANK CORRUPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion language models are commonly trained to denoise text in which tokens have been replaced by a `[MASK]` or by tokens sampled uniformly from the vocabulary. Such replacements carry no information about the original token. To train on corruptions that more closely resemble plausible mistakes, we instead consider semantically related alternatives, which can be harder to identify as corrupted while preserving clues that aid recovery. We introduce Semantic Diffusion Language Models (SDLMs), a family of discrete diffusion models that draw these replacements from a nonnegative low-rank kernel, which permits exact training-loss evaluation at linear rather than quadratic cost in vocabulary size. Like uniform replacements, semantic ones are not marked as corrupted, so the model must learn not only what each token should become but also when to revise it. We show that with uniform noise, models time revisions by their confidence in the observed token alone: belief in a plausible alternative is indistinguishable from belief in an unrelated word. SDLMs instead use the full posterior, keeping a token settled when the model's belief falls on words similar to the observed token. SDLM exceeds the MAUVE of matched MDLM and GIDD baselines with $4\times$ fewer steps. We also introduce SHARP, a predictor–corrector sampler that reallocates revisions toward tokens the model is most likely to change. With SHARP, SDLM outperforms ReMDM at every sampling budget and Duo with Ψ-samplers at all but the lowest.

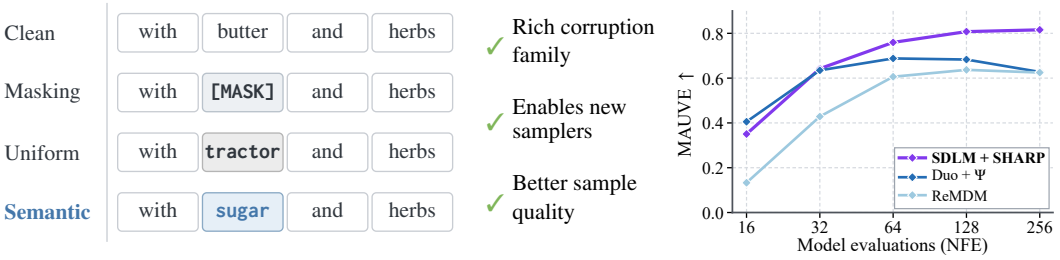

Figure 1: **(Left) Corruption examples:** semantic replacements retain clues about the original token, unlike masking or the unrelated uniform replacement shown. **(Right) Predictor–corrector sampling:** SDLM + SHARP achieves higher MAUVE than Duo + Ψ and ReMDM from 32 NFE.

1 INTRODUCTION

Diffusion language models offer an alternative to left-to-right generation, allowing tokens to be generated in parallel over multiple denoising steps (Lou et al., 2024; Sahoo et al., 2024). They learn by reversing a process that corrupts text, so the choice of corruption shapes what the model must learn. Masking removes tokens and asks the model to fill the gaps, but its exact reverse process cannot revise a token once it has been filled in (Sahoo et al., 2024; Shi et al., 2024). Uniform replacement lets the model revise tokens it has already generated by keeping noisy states in the vocabulary (Austin et al., 2021; von Rütte et al., 2025), and recent studies show encouraging results at scale (von Rütte et al., 2026; Sahoo et al., 2026).

Masked and uniform corruption share a useful simplification: each replacement draws independently of the token it replaces. This *one-shot forgetting* property makes corruption tractable without having to specify relationships between tokens. Information can still disappear gradually across a sequence, but an individual replacement retains no information about its source. Moreover, uniform replacements are often easy to identify as noise, making the task resemble masked reconstruction.

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.

Semantically aware replacements offer an alternative to uniform ones, retaining clues to the clean token while making corruptions less obvious. Semantic corruption has been explored through embedding-based token replacements in D3PM (Austin et al., 2021) and semantic hierarchies in HDLM (Zhou et al., 2025), whose cluster-level noise states retain partial source information but, like the designated masks of MultiMDM (Chen et al., 2026), still mark which positions are corrupted. Beyond choosing which tokens replace which, training and generation need the probability of each token after several successive replacements, summed over paths through intermediate tokens. Computing these probabilities for a large vocabulary can require expensive operations on dense vocabulary-by-vocabulary matrices. D3PM's semantic construction, for instance, uses matrix operations whose cost grows quadratically with vocabulary size. GDDS avoids these computations during training through a snapshot objective for clean-token prediction (Zekri et al., 2026), but approximating semantic transitions in its ancestral sampler is still costly and accumulates error. These approaches motivate a complementary question: can we make these probabilities tractable, so that training and generation can both use them exactly?

First, we propose **Semantic Diffusion Language Models (SDLMs)**, which combine replacements from a fixed semantic matrix with an independent reset to a known prior, and use a nonnegative low-rank factorisation of this matrix to compute these probabilities exactly at vocabulary scale. Second, we analyse how semantic corruption changes what the model must learn. Following Zekri et al. (2026), the continuous-time training objective (Campbell et al., 2022) decomposes into a term for what a token should change to and a term for how quickly it should change. Under uniform and GIDD corruption (von Rütte et al., 2025), how quickly a token is revised depends only on the model's confidence that the observed token matches the clean token, so at a given confidence the rate is the same whether the remaining posterior mass lies on similar words or on unrelated ones. Semantic corruption lets this rate use the full posterior. At the same confidence in the observed token, the model then revises it more slowly when its remaining belief favours similar words, which may be equally valid in context, and sooner when that belief favours unrelated words (Section 3). Third, since samplers revise tokens at these learned rates, we introduce SHARP, a predictor–corrector sampler that augments Ψ-samplers (Deschenaux et al., 2026) by reallocating revisions toward tokens the model is most likely to change, without additional network evaluations.

We evaluate SDLM against masked (MDLM; Sahoo et al., 2024), uniform (Duo; Sahoo et al., 2025) and mixed (GIDD; von Rütte et al., 2025) diffusion models of comparable size and training token budget. Without additional sampling corrections, SDLM achieves higher MAUVE than MDLM and GIDD across all evaluated sampling budgets and exceeds their best MAUVE with $4\times$ fewer steps. SDLM+SHARP achieves higher MAUVE than ReMDM (Wang et al., 2025) and Duo with Ψ-samplers at all but the lowest evaluated sampling budget (Figure 1); with only 64 steps, it exceeds the best MAUVE either reaches at any budget. We will release our code and checkpoints for SDLM and our retrained baselines.

2 Continuous-time discrete diffusion

Generation runs corruption in reverse, deciding at every step whether each token stays or is replaced. We formulate both corruption and generation as continuous-time Markov chains (CTMCs) (Campbell et al., 2022), in which tokens change at random times. The forward process specifies how quickly each token changes into other tokens, while a Transformer parameterises the corresponding reverse transition rates using the noisy sequence and time. These reverse rates are what the model learns. Let $\mathcal{V}$ be the token vocabulary, including any noise symbols, with $V = |\mathcal{V}|$. Given a clean sequence $X_0 = (x_0^1, \dots, x_0^L) \sim q_0$, we corrupt each position independently using the same CTMC. This gives noisy sequences $Z_t = (z_t^1, \dots, z_t^L)$ for $t \in [0, 1]$, with $Z_0 = X_0$ and

$$q_t(Z_t \mid X_0) = \prod_{k=1}^{L} q_t(z_t^k \mid x_0^k). \tag{1}$$

Since forward corruption factorises over positions, we describe it for one token, writing x_0 and z_t for its clean and noisy states. Let $Q_t \in \mathbb{R}^{V \times V}$ be the **forward rate matrix** for $0 \le t < 1$. Its off-diagonal entries are the rates of token changes, and its diagonal makes each row sum to zero:

$$Q_t(z, y) = \lim_{h \downarrow 0} \frac{\mathbb{P}\left(z_{t+h} = y \mid z_t = z\right)}{h} \quad (y \neq z), \qquad Q_t(z, z) = -\sum\nolimits_{y \neq z} Q_t(z, y). \tag{2}$$

Training and generation need the probability of each token after corruption has run for some time. For $0 \le s \le t < 1$, we write these **transition probabilities** as $q_{t|s}$ and collect them in the **forward kernel** $P_{t|s}$, the matrix form used in Section 4:

$$q_{t|s}(y \mid z) := \mathbb{P}\left(z_t = y \mid z_s = z\right), \qquad [P_{t|s}]_{yz} = q_{t|s}(y \mid z).$$

Each column of $P_{t|s}$ gives the destination probabilities for one source token z. Because Q_t indexes source tokens by rows instead, the Kolmogorov forward equation uses its transpose:

$$\partial_t P_{t|s} = Q_t^\top P_{t|s}, \qquad P_{s|s} = I, \tag{3}$$

Writing $P_t := P_{t|0}$, the **corruption marginal** is the x_0-th column of P_t, $q_t(\cdot \mid x_0) = P_t \mathbf{e}_{x_0}$, where $\mathbf{e}_{x_0}$ is the point mass at x_0. When the clean token is known, the reverse rates follow from the forward ones by Bayes' rule. For a clean token x_0 and $q_t(z \mid x_0) > 0$, define the **clean-conditioned reverse rates**, for $y \neq z$, by

$$\widehat{Q}_t(z, y \mid x_0) := \lim_{h \downarrow 0} \frac{q_{t-h|t,0}(y \mid z, x_0)}{h} = Q_t(y, z) \frac{q_t(y \mid x_0)}{q_t(z \mid x_0)}. \tag{4}$$

For a fixed forward process, these are known token-level rates. Once the clean token x_0 is given, they require neither the other positions nor the clean-data distribution q_0. They provide the training targets in the local loss below.

The clean token is unknown during generation, so the **model posterior** $d_\theta(x_0) := d_\theta(x_0^k \mid Z_t, t)$ at position k is trained to approximate the true posterior $q_{0|t}(x_0^k \mid Z_t)$ and averages the clean-conditioned rates into the **modelled reverse rates**:

$$\widehat{Q}_t^\theta(z, y) = \sum\nolimits_{x_0 : q_t(z|x_0) > 0} d_\theta(x_0)\, \widehat{Q}_t(z, y \mid x_0), \qquad y \neq z. \tag{5}$$

Our network outputs a clean-token distribution x_θ, from which d_θ follows by Bayes' rule with the likelihood of the observed token, $d_\theta(x_0) \propto x_\theta(x_0)\, q_t(z \mid x_0)$ (Casado Noguerales et al., 2026).

Any rate matrix, forward or reverse, splits into how quickly a token changes and what it changes to. Each off-diagonal rate of Q_t acts as an independent Poisson clock, and the first to ring decides when token z changes and what replaces it, so the clocks combine into one:

$$r_t(z) := \sum\nolimits_{y \neq z} Q_t(z, y), \qquad Q_t(z, y) = r_t(z)\, \mu_t(y \mid z) \quad (y \neq z), \tag{6}$$

Here $r_t(z)$, the **exit rate**, is the total rate at which z changes, and $\mu_t(\cdot \mid z)$, the **replacement law**, is the distribution of the new token, defined when $r_t(z) > 0$. Decomposing the reverse rates in (4) and (5) in the same way gives the clean-conditioned and modelled exit rates and replacement laws:

$$\widehat{Q}_t(z, y \mid x_0) = r_t(z \mid x_0)\, \mu_t(y \mid z, x_0), \qquad \widehat{Q}_t^\theta(z, y) = r_t^\theta(z)\, \mu_t^\theta(y \mid z). \tag{7}$$

For a training triple (t, x_0, z), the **local loss** measures how well the model reverses corruption of clean token x_0, observed as z at time t. It compares, per unit time, the clean-conditioned reverse transition with the model's reverse transition $p^\theta_{t-h|t}$, which is built from the modelled rates and uses only the noisy sequence and time:

$$\mathcal{J}_{t,z}^{x_0}(\theta) = \lim_{h \downarrow 0} \frac{1}{h} \mathrm{KL}\Big(q_{t-h|t,0}(\cdot \mid z, x_0) \,\Big\|\, p^\theta_{t-h|t}(\cdot \mid z) \Big). \tag{8}$$

Averaging this loss over corrupted examples, summing over positions, and integrating over time gives the path term of the negative evidence lower bound (ELBO) (Campbell et al., 2022). The full bound, including reconstruction and prior-matching terms, is given in Appendix A.

3 Learning to denoise: what changes and when

Each denoising step decides, for every token, *whether* to change it now and, if so, *what* to change it to. When corruption replaces tokens with other vocabulary tokens rather than a `[MASK]`, as with uniform or semantic noise, the first decision must be learned from context.

Question. How quickly should a model revise an observed token z? Its posterior d_θ holds both its confidence in z and which alternatives it favours. Which of these sets the exit rate $r_t^\theta(z)$, and how does the type of corruption change the answer?

Over a short reverse step, each possible token change has probability proportional to its rate, so the local KL (8) reduces to a sum of Poisson KL divergences between the clean-conditioned and modelled reverse rates (Casado Noguerales et al., 2026). Factorising both rates as in (7) splits it into a replacement term and a timing term (Zekri et al., 2026):

$$\mathcal{J}_{t,z}^{x_0}(\theta) = \sum\nolimits_{y \neq z} \Phi\Big(\widehat{Q}_t(z, y \mid x_0), \widehat{Q}_t^{\theta}(z, y)\Big) \tag{9a}$$

$$= r_t(z \mid x_0)\ \mathrm{KL}\big(\mu_t(\cdot \mid z, x_0) \,\big\|\, \mu_t^{\theta}(\cdot \mid z)\big) + \Phi\big(r_t(z \mid x_0), r_t^{\theta}(z)\big), \tag{9b}$$

where $\Phi(a, b) := a \log(a/b) + b - a$. The first term penalises errors in *what replaces the token*, weighted by how often the target process changes it, whereas the second penalises errors in the *exit rate*, which is therefore learned directly. Both r_t^{θ} and μ_t^{θ} are computed from the posterior d_θ, and we next show that the corruption decides which beliefs in d_θ drive the exit rate.

3.1 Which beliefs set the exit rate

When sampling, a short reverse step of length h keeps or replaces z according to

$$p_{t-h|t}^{\theta}(\cdot \mid z) = \big[1 - h\, r_t^{\theta}(z)\big]\delta_z + h\, r_t^{\theta}(z)\, \mu_t^{\theta}(\cdot \mid z) + o(h), \tag{10}$$

so $r_t^{\theta}(z)$ sets how quickly z is revised and μ_t^{θ} what replaces it. If the clean token were known to be x_0, z would instead be revised at the **clean-conditioned exit rate**

$$r_t(z \mid x_0) = \sum\nolimits_{y \neq z} \widehat{Q}_t(z, y \mid x_0). \tag{11}$$

The model does not know x_0, so it averages these rates under its posterior:

Proposition 3.1 (Posterior-averaged exit rate)**.** At fixed t, z, the clean-conditioned exit rates are fixed by the corruption process alone, and the modelled exit rate is their posterior average:

$$r_t(z \mid x_0) = \partial_t \log q_t(z \mid x_0) - Q_t(z, z), \qquad r_t^{\theta}(z) = \sum\nolimits_{x_0} d_\theta(x_0)\, r_t(z \mid x_0). \tag{12}$$

Only $\partial_t \log q_t(z \mid x_0)$ varies with x_0, so clean tokens whose corruption is still making z more likely at time t imply faster revision. We write a general corruption marginal as

$$q_t(\cdot \mid x_0) = \alpha_t \boldsymbol{\delta}_{x_0} + (1 - \alpha_t)\boldsymbol{\pi}_t(\cdot \mid x_0), \tag{13}$$

where α_t is the probability that no replacement has occurred and $\boldsymbol{\pi}_t(\cdot \mid x_0)$ is the noise law, which decides which beliefs set the exit rate.

Masked: the posterior plays no role. Masking sets $\boldsymbol{\pi}_t(\cdot \mid x_0) = \boldsymbol{\delta}_m$, with m the `[MASK]`. At a masked position, every clean token x_0 gives the same exit rate $r_t(m \mid x_0) = -\dot{\alpha}_t/(1 - \alpha_t)$. Averaging under any posterior leaves it unchanged, so the model cannot affect when masks are filled. Unmasked tokens have $r_t(z \mid x_0) = 0$ and are never revised.

Uniform and GIDD: only confidence matters. When the noise law is shared across clean tokens, $\boldsymbol{\pi}_t(\cdot \mid x_0) = \boldsymbol{\pi}_t$, as for uniform corruption ($\boldsymbol{\pi}_t = \mathbf{u}$) and GIDD (von Rütte et al., 2025), which mixes $\mathbf{u}$ with $\boldsymbol{\delta}_m$, Proposition 3.1 gives only two clean-conditioned exit rates, depending on whether the candidate clean token is the observed one:

$$r_t(z \mid x_0) = \begin{cases} r_t^0(z) := \partial_t \log\big[\alpha_t + (1 - \alpha_t)\pi_t(z)\big] - Q_t(z, z), & x_0 = z, \\ r_t^1(z) := \partial_t \log\big[(1 - \alpha_t)\pi_t(z)\big] - Q_t(z, z), & x_0 \neq z, \end{cases} \tag{14}$$

for candidates with positive likelihood. Posterior averaging therefore gives

$$r_t^{\theta}(z) = d_\theta(z)\, r_t^0(z) + \big[1 - d_\theta(z)\big] r_t^1(z). \tag{15}$$

For uniform corruption $r_t^1(z) > r_t^0(z)$, so confidence in z slows its revision. Every alternative clean token implies the same rate $r_t^1(z)$, so only $d_\theta(z)$ moves the rate, whether the remaining belief lies on similar words or on unrelated ones, and plausible alternatives can be treated like errors.

Semantic: similar alternatives slow revision. When the noise law depends on the clean token, as for the SDLMs of Section 4, different candidates imply different exit rates. A semantic kernel spreads noise from a word to similar words before unrelated ones, so a clean token similar to z implies a lower exit rate than an unrelated one, and for a fixed kernel the rate grows with the expected number of replacements separating the two (Appendix B). Belief in similar words therefore keeps z settled, while belief in unrelated words hastens its revision (Figure 2). Section 5.3 tests whether the kernel's semantic content matters in practice, and SHARP (Section 4.3) builds on these exit rates.

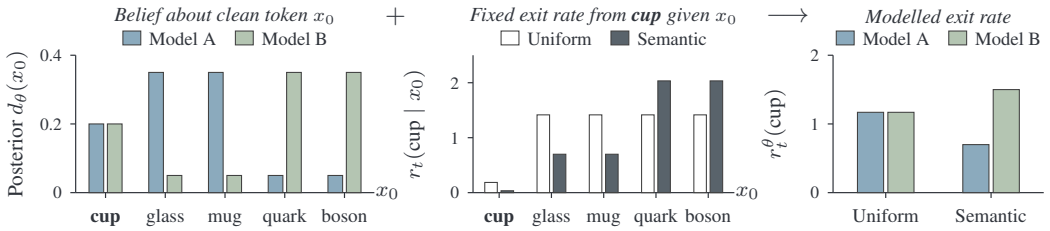

Figure 2: **Belief in related alternatives can slow revision.** For an observed sequence "She poured tea into the **cup**," assume two models put the same posterior mass on $z = \text{cup}$, but Model A places more probability on the plausible alternatives glass and mug than Model B (left). A semantic kernel treats these as similar to cup, so they imply lower exit rates (centre) and Model A is less likely to leave cup (right). Uniform corruption makes no such distinction.

4 SEMANTIC DIFFUSION

Section 3 showed that under semantic corruption the exit rate depends on the whole posterior d_θ, not only on the confidence $d_\theta(z)$. We build such a corruption by combining a fixed semantic replacement kernel, which replaces a token by a similar word at random times, with an independent reset to a known prior. By constraining this kernel to a low-rank form, we can compute corruption and training probabilities exactly, even at large vocabulary scale. We first explain why we design rates rather than marginals, then derive the corruption law and its low-rank computation (proofs in Appendix C).

Marginal-first design. A common starting point is to specify the corruption marginals $q_t(\cdot \mid x)$ directly. For the GIDD family, including masked and uniform corruption (Section 3.1), stacking these marginals as columns gives $P_t = \alpha_t I + (1 - \alpha_t)\boldsymbol{\pi}_t \mathbf{1}^\top$, an identity branch that keeps the clean token and a shared-noise branch that forgets it.

To let replacements retain clues to the clean token, we make the noise columns source-dependent. These marginals must still be connected by valid transition rates. For a differentiable path from $P_0 = I$, invertible before the terminal time, the Kolmogorov equation gives the candidate rate matrix

$$Q_t^\top = \dot{P}_t P_t^{-1}, \tag{16}$$

whose off-diagonal entries must be nonnegative. Nothing guarantees this for prescribed semantic marginals, and checking it requires inverting a $V \times V$ matrix at every time.

Rate-first design. We avoid this inverse problem by designing the rates and deriving their marginals. A rate matrix is valid when its off-diagonal entries are nonnegative and its rows sum to zero, which the classical uniformisation form guarantees. Any finite-state CTMC with locally integrable rates can be written in this form, by taking β_t at least as large as every exit rate and $K_t = I + Q_t/\beta_t$:

$$Q_t = \beta_t(K_t - I), \qquad \beta_t \geq 0, \quad K_t \in \mathbb{R}_{\geq 0}^{V \times V}, \quad K_t \mathbf{1} = \mathbf{1}. \tag{17}$$

Here β_t is the intensity of a state-independent Poisson clock, and K_t is the replacement kernel, which at each ring replaces the current token z by y with probability $K_t(z, y)$, where y may be z itself. The following characterisation identifies when replacements discard all source information.

Proposition 4.1. For a finite-state CTMC with continuous rates, the following are equivalent, where $\mathbf{1} \in \mathbb{R}^V$ is all ones, ρ_t is a probability vector over $\mathcal{V}$, and $N_{s,t}$ counts replacements in $(s, t]$.

GIDD marginals		Rank-1 representation		One-shot forgetting
$P_t = \alpha_t I + (1 - \alpha_t)\pi_t \mathbf{1}^\top$	$\iff$	$Q_t = \beta_t(\mathbf{1}\rho_t^\top - I)$	$\iff$	$Z_t \perp\!\!\!\perp Z_s \mid N_{s,t} \geq 1$

Rank-one replacements are thus exactly the shared-noise branch above, written as rates (formal statement in Appendix C). A fixed semantic kernel instead replaces a token by similar words, so repeated replacements spread noise outward over time, reaching similar words before unrelated ones.

4.1 CONSTRUCTING THE SEMANTIC CORRUPTION LAW

Fixing the replacement kernel. With a time-varying replacement kernel, rate matrices at different times need not commute, so the marginals depend on the order in which replacements act. Fixing

the kernel removes this ordering dependence and expresses repeated replacements as powers of a single matrix, which we will exploit for low-rank computation in Section 4.2. We therefore fix a row-stochastic kernel $K_t = K$ that favours semantically related replacements (Section 5.1), and write Λ_t for the expected number of replacements by time t:

$$Q_t = \beta_t(K - I), \qquad \Lambda_t := \int_0^t \beta_s \, ds. \tag{18}$$

Each replacement applies the same kernel, so $K^n(x_0, \cdot)$ gives the token distribution after n replacements. By time t, the replacement count $N_t \sim \mathrm{Poisson}(\Lambda_t)$ induces the transition matrix

$$P_t = \alpha_t I + (1 - \alpha_t)\Pi_t^\top, \qquad \Pi_t := \mathbb{E}\big[K^{N_t} \mid N_t \geq 1\big]. \tag{19}$$

Here $\alpha_t := \exp(-\Lambda_t)$ is the probability of no replacement, while $\Pi_t(x_0, \cdot)$ is the token distribution conditional on at least one replacement. This gives a source-dependent noise law $\boldsymbol{\pi}_t(\cdot \mid x_0) = \Pi_t(x_0, \cdot)$, obtained from valid transition rates.

Ensuring convergence to the prior. The semantic process (18) alone need not converge to a known prior. We therefore augment it with an independent reset, giving the SDLM forward process (21). To start reverse generation from a known prior π, the terminal law must match π and forget the clean input. For a clean token x_0 drawn from its data marginal, let q_1 denote the marginal law of z_1. The expected terminal prior term of the negative ELBO separates these requirements:

$$\mathbb{E}_{x_0}\big[\mathrm{KL}\big(q_1(\cdot \mid x_0)\|\pi\big)\big] = \underbrace{I_q(x_0; z_1)}_{\text{retained information}} + \underbrace{\mathrm{KL}(q_1\|\pi)}_{\text{marginal mismatch}}. \tag{20}$$

Both terms vanish when $q_1(\cdot \mid x_0) = \pi$ for every clean token x_0. Semantic replacements alone can reach this endpoint if the kernel converges to the prior and the schedule mixes enough. The independent *prior clock* guarantees convergence without tying it to the semantic mixing rate. It resets the token to π, giving the rate matrix below with nonnegative rates β_t^{sem} and β_t^{pr}:

$$Q_t = \beta_t^{\mathrm{sem}}(K - I) + \beta_t^{\mathrm{pr}}(\mathbf{1}\pi^\top - I), \qquad \text{s.t.} \quad K^\top \pi = \pi. \tag{21}$$

Stationarity of π under K means semantic replacements preserve the prior after a reset, and makes the semantic and reset components commute, yielding the simple marginal below. The prior can be a point mass on an absorbing `[MASK]` state or a distribution over vocabulary tokens.

Lemma 4.2 (Semantic corruption with prior reset)**.** For the dynamics in (21) and Π_t as in (19) for the semantic clock, the forward transition matrix is

$$P_t = \underbrace{\alpha_t^{\mathrm{pr}}\alpha_t^{\mathrm{sem}} I}_{\text{no replacement}} + \underbrace{\alpha_t^{\mathrm{pr}}(1 - \alpha_t^{\mathrm{sem}})\Pi_t^\top}_{\text{semantic only}} + \underbrace{(1 - \alpha_t^{\mathrm{pr}})\pi\mathbf{1}^\top}_{\text{prior reset}}, \tag{22}$$

where α_t^{sem} and α_t^{pr} are the probabilities of no semantic replacement and no prior reset by time t. If $\alpha_t^{\mathrm{pr}} \to 0$ as $t \uparrow 1$, then $q_t(\cdot \mid x_0) \to \pi$ for every x_0, regardless of the mixing speed of K. We next exploit low-rank structure to evaluate the marginals and training objective exactly.

4.2 Exact computation with a low-rank semantic kernel

Evaluating the ELBO in our clean-token parametrisation requires $P_t x_\theta$, which propagates the network prediction through the forward process. A generic dense kernel poses two challenges: **(i)** obtaining P_t at arbitrary training times is expensive, with tabulation requiring $O(V^2)$ storage per matrix and interpolation between stored times; **(ii)** computing $P_t x_\theta$ costs $O(V^2)$ per prediction. A low-rank semantic kernel makes both evaluating P_t and computing $P_t x_\theta$ practical at vocabulary scale.

Low-rank construction. Choose nonnegative $V \times R$ matrices A and B, with $R \ll V$, and set

$$K = AB^\top, \qquad \text{s.t.} \quad A\mathbf{1}_R = \mathbf{1}_V, \quad B^\top \mathbf{1}_V = \mathbf{1}_R, \quad K^\top \pi = \pi. \tag{23}$$

Each column of B defines a distribution over replacement tokens, a semantic prototype, and each row of A specifies how the current token mixes these prototypes, as in probabilistic latent semantic analysis (Hofmann, 1999). The kernel therefore defines which words count as similar.

Factoring K lets us evaluate the Poisson mixture Π_t in (19) through the reduced $R \times R$ matrix $M := A^\top B$. With $N_t \sim \mathrm{Poisson}(\Lambda_t)$ counting semantic replacements, the factorisation gives

$$\Pi_t^\top = \mathbb{E}\big[(K^\top)^{N_t} \mid N_t \geq 1\big] = B\, S_t\, A^\top, \tag{24}$$

$$S_t := \mathbb{E}\big[M^{N_t - 1} \mid N_t \geq 1\big] \in \mathbb{R}^{R \times R}. \tag{25}$$

We call S_t the reduced replacement mixture. The outer factors account for the first replacement, and each additional replacement contributes another application of M.

We compute $\Pi_t^\top x_\theta = B[S_t(A^\top x_\theta)]$ by projecting the prediction into the reduced space, applying the replacement mixture there, and mapping back to vocabulary probabilities (Figure 3a). The retention and prior-reset terms require only scaling and addition, so neither Π_t nor P_t needs to be materialised.

Proposition 4.3 (Tractable semantic diffusion)**.** With $K = AB^\top$, the marginal in Lemma 4.2 has the exact factored representation

$$P_t = \alpha_t^{\mathrm{pr}} \alpha_t^{\mathrm{sem}} I + \alpha_t^{\mathrm{pr}}(1 - \alpha_t^{\mathrm{sem}}) B S_t A^\top + (1 - \alpha_t^{\mathrm{pr}}) \pi \mathbf{1}^\top. \tag{26}$$

For a sampled triple (t, x_0, z), with $\Phi(a, b) = a \log(a/b) + b - a$, the local loss is

$$\mathcal{J}_{t,z}^{x_0}(\theta) = \sum_{y \neq z} \Big[\beta_t^{\mathrm{sem}} (AB^\top)_{yz} + \beta_t^{\mathrm{pr}} \pi(z)\Big] \Phi\left(\frac{[P_t \delta_{x_0}]_y}{[P_t \delta_{x_0}]_z}, \frac{[P_t x_\theta]_y}{[P_t x_\theta]_z}\right). \tag{27}$$

For a sequence of length L, evaluating $x_\theta \mapsto \mathcal{J}_{t,z}^{x_0}(\theta)$ costs $O(VR + R^3/L)$ operations per token.

Computational cost. Computing S_t costs $O(R^3)$, shared across all L positions in a sequence, and at inference reduced transitions are cached on the fixed sampling grid (Appendix C). In our implementation, training needs about 6% more compute than MDLM, whereas a dense semantic kernel needs about five times as much even with P_t precomputed, and generation is about $1.3\times$ slower than MDLM (Figure 3b).

4.3 SHARP: SHARPENING WHERE TO REVISE

At each step, a sampler revises every visible token with a probability set by its exit rate, as in (10), and draws the new token from its replacement law. SDLM's exit rates use the full posterior (Section 3), so they reflect which alternatives the model favours, not only its confidence in the observed token. SHARP concentrates each step's revisions on the tokens the model is most likely to revise. Like nucleus sampling (Holtzman et al., 2020), it keeps the smallest set of positions whose revision probabilities cover a fixed fraction of their total, sets the other probabilities to zero, and scales the kept ones by a common factor, capped at one, so that the expected number of revisions is unchanged (Appendix D). Applied on top of Ψ-Cap (Deschenaux et al., 2026), which adds revision opportunities, SHARP changes only where revisions happen and adds no network evaluations. Section 5.4 evaluates this allocation in generation and compares its effects on semantic and uniform diffusion.

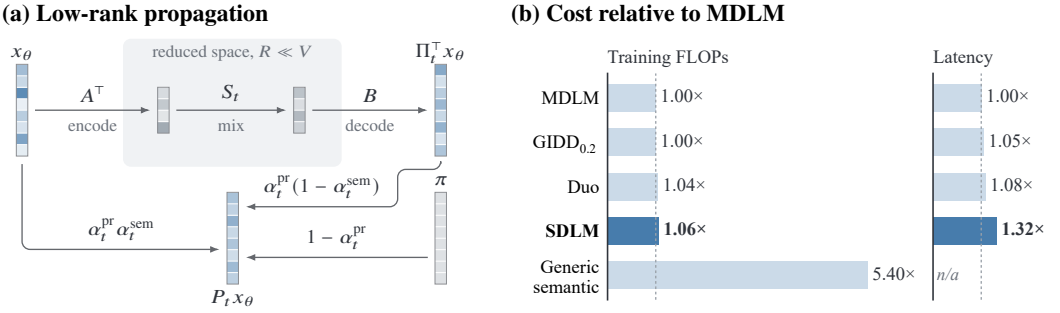

Figure 3: **Tractable semantic diffusion.** (a) Low-rank forward propagation. (b) Training FLOPs per 512×512-token batch (MDLM: 602 TFLOPs), and latency for 512 tokens at batch size one and 32 NFE on an RTX 4090 (MDLM: 0.553 s). Generic semantic assumes P_t is given.

5 EXPERIMENTAL RESULTS

We train a 425M-parameter SDLM on OpenWebText for 500k steps with batch size 512 and context length 512 (131B training tokens), the setup of GIDD and HDLM (von Rütte et al., 2025; Zhou et al., 2025). To compare generation at matched model size and training token budget, we retrain MDLM and Duo with the same backbone and budget, keeping Duo's Gaussian-relaxation curriculum (Sahoo et al., 2025), and evaluate released GIDD models trained with this setup. MAUVE is sensitive to sequence length (Pillutla et al., 2023), so we evaluate all models on 512-token sequences, matching the training context length used in these comparisons (protocol and full results in Appendix E).

Evaluation criteria. Our primary metric is MAUVE (Pillutla et al., 2023). We do not rank models by generative perplexity (GenPPL) alone, as small reductions in entropy can have an outsized impact on the metric (Acquaviva, 2026; Franca & Tong, 2026). Instead, we report GenPPL–entropy frontiers across temperatures, which show the trade-off between fluency and diversity (Pynadath et al., 2026a;b). We omit cross-process ELBO comparisons, whose unknown gaps to the likelihood differ (Pynadath et al., 2026b).

5.1 THE CORRUPTION PROCESS

Kernel construction. We use a rank-256 semantic kernel built from pretrained token embeddings. Tokens are assigned to prototypes, each defining a soft distribution over related replacements (Appendix E). We evaluate this fixed construction without tuning its rank or assignments.

What is the detection–recovery trade-off? Section 1 motivates semantic replacements as harder to identify as corrupted while retaining clues to their source. We measure both with one pretrained masked diffusion model, shared across corruption types, which predicts whether a token changed and, given a change, which clean token it replaced. At the same change probability, semantic corruption raises detection loss (.036 versus at most .0053 bits) but nearly halves recovery loss (.57 versus 1.06 bits) relative to mask, uniform and GIDD corruption (Appendix F).

5.2 GENERATION QUALITY

With ancestral sampling (no predictor–corrector steps or nucleus truncation), SDLM achieves higher MAUVE than MDLM and GIDD at every budget and exceeds their best MAUVE with $4\times$ fewer steps. It matches or beats Duo, the strongest baseline, from 32 NFE (Figure 4, top left). At 16 NFE, Duo has higher MAUVE, but SDLM has the better GenPPL–entropy frontier at 8 and 16 NFE (bottom). With predictor–corrector sampling (top middle), SDLM+SHARP outperforms ReMDM

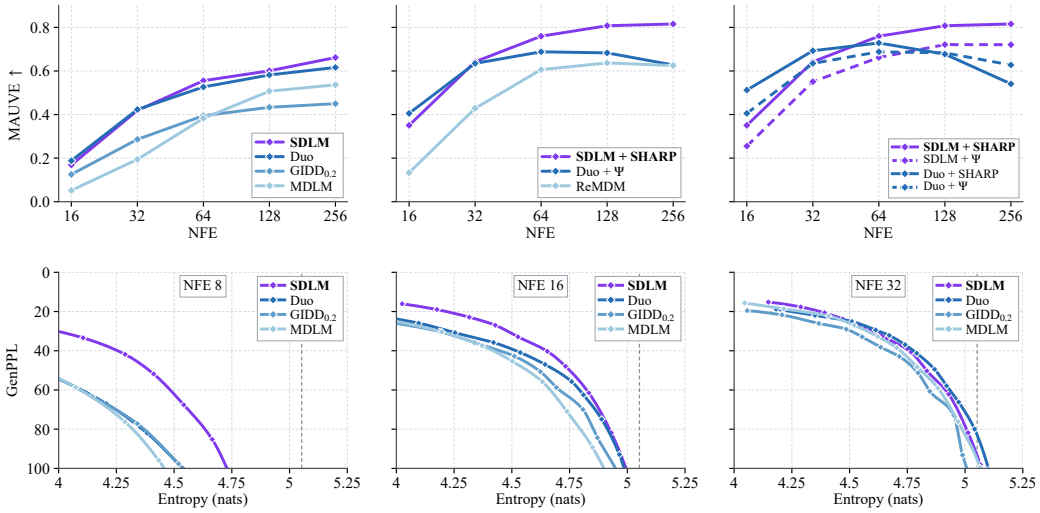

Figure 4: **Generation quality. Top, left to right:** MAUVE with ancestral sampling; predictor–corrector sampling (SDLM+SHARP, Duo+Ψ, and ReMDM); and the effect of SHARP on SDLM and Duo relative to their Ψ baselines (solid versus dashed). **Bottom:** ancestral-sampling GenPPL–entropy frontiers (colours as top left) with inverted GenPPL axis; dashed lines: reference entropy.

at every budget and Duo+Ψ at all but 16 NFE, and at 64 NFE exceeds the best MAUVE that either reaches at any budget.

5.3 What the kernel contributes

Does vocabulary alignment help? SDLM's gains could come from the semantic relationships encoded in K, or from source-dependent corruption of any kind. MultiMDM, for example, assigns tokens to designated masks by token index, a semantically arbitrary grouping, and remains competitive with or improves on MDLM (Chen et al., 2026). To separate the two, we train a control on the same kernel with its vocabulary indices randomly permuted (the mask fixed): prototypes keep their sizes and weights, but each now groups unrelated tokens. With initialisation and schedule also fixed, the semantic kernel improves MAUVE at every budget, from .31 to .52 at 128 NFE (Figure 5a). Both runs use 13B training tokens, about 10% of the main budget.

Does lower rank preserve quality? At the same small training budget, rank 64 gives the highest MAUVE among the three semantic kernels at all five NFE budgets (Figure 5b). Rank 16 remains within .04 of rank 256 across these budgets, but trails rank 64 throughout. This points to an optimal rank between 16 and 256 that may also outperform rank 256 at the full training budget, although gains at 10% of that budget need not carry over (Appendix F).

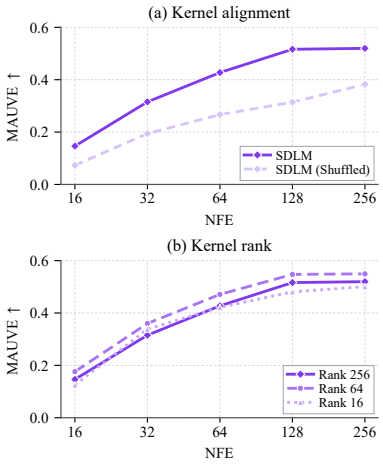

Figure 5: **Kernel alignment and rank.** Generation MAUVE after 13B training tokens, one training seed per model.

Does the exit rate use the alternatives? On 256 held-out passages, we compare SDLM's exit rate with a confidence-only rule that keeps the model's confidence in the observed token but ignores which alternatives it favours. At matched confidence, SDLM's rate separates injected unrelated substitutions from correct tokens with AUROC .88, against .72 for the confidence-only rule, and is $1.6\times$ higher on those substitutions than confidence alone implies (Appendix F).

5.4 SHARP across samplers and models

Gains over Ψ. We use one SHARP recipe, selected by MAUVE, for all reported budgets rather than tuning it per NFE: fraction .90 on Ψ-Cap .02, with nucleus .95 on the clean-token prediction. It improves SDLM+Ψ by .09–.10 at every budget while holding replacement laws fixed.

Semantic versus uniform rates. On matched Duo with the Ψ settings recommended by Deschenaux et al. (2026) (Appendix E), SHARP helps Duo through 64 NFE, and Duo+SHARP leads SDLM+SHARP at 16–32 NFE. Its gain vanishes by 128 NFE and is negative at 256 (Figure 4, top right), where repeated 8-token spans rise from 12% to 38% of documents (SDLM: 6% to 12%), which MAUVE penalises (Pillutla et al., 2023). At higher NFE, SHARP also reduces sample entropy more strongly for Duo than for SDLM: at 256 NFE, mean within-document unigram entropy falls by .172 versus .086 nats. This larger loss of diversity may help explain Duo+SHARP's MAUVE drop. One explanation consistent with Section 3 is that Duo's exit rates use only confidence, so concentrating revisions on them also targets ambiguous but correct tokens.

6 Conclusion

We introduced SDLMs, which corrupt text with semantically related replacements from a low-rank kernel, keep training exact at linear cost in vocabulary size, and time revisions using the full posterior. Using a kernel chosen without any design search, SDLM compares favourably with masked and uniform diffusion baselines, and SHARP improves it further. Beyond this first instantiation lies a large design space, including the kernel rank, noise schedules and induced token relationships, as well as further analysis of the learned revision rates and samplers built on them.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, design and provide feedback on experiments, interpret results, formulate mathematical claims and assist in writing proofs, and propose or refine hypotheses. We have not used generative AI tools to generate synthetic data sets or to develop theoretical models or conceptual frameworks, and translation, dataset cleaning and qualitative or thematic analysis are not applicable to this work. Additionally, we used generative AI tools to create and edit code and figures, suggest experimental parameters, identify and summarise related work, draft and edit text, and format references. All AI-assisted work was checked by multiple authors over multiple rounds of revision, and derivations were verified by hand and numerically. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies text generation using models trained on OpenWebText. Web-derived training data can contain personal information, copyrighted material, biases and harmful language; the methods studied here do not remove these risks. Generated text can reproduce harmful content or invent claims about real people and organisations. The qualitative samples in Appendix G illustrate model behaviour and are not factual assertions or endorsements. Improvements in MAUVE or generative perplexity do not establish factuality, fairness, privacy or safety. These properties would require separate evaluation and safeguards before deployment.

REPRODUCIBILITY STATEMENT

Appendices A–C provide the mathematical derivations and semantic-kernel construction. Appendix D specifies the SHARP sampler. Appendix E records the model and training setup, kernel parameters, generation budgets and seeds, metric definitions, evaluation protocol and sampler selection procedure. Appendix G identifies the settings used for the qualitative examples.

REFERENCES

Sam Acquaviva. Fixing flow model evals. Blog post, 2026. URL https://samacquaviva.com/projects/flow-evals/. Accessed September 22, 2026.

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

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

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

Sijin Chen, Yinuo Ren, Heyang Zhao, Ziheng Cheng, Quanquan Gu, and Lexing Ying. Multi-mask diffusion language models for few-step generation, 2026. URL https://arxiv.org/abs/2607.19686v3. Version 3.

Justin Deschenaux, Caglar Gulcehre, and Subham Sekhar Sahoo. The diffusion duality, chapter II: Ψ-samplers. In *International Conference on Learning Representations*, 2026.

Antonio Franca and Alexander Tong. Hacking generative perplexity: Why unconditional text evaluation needs distributional metrics. In *ICML 2026 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2026.

Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pp. 289–296, 1999.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

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

Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. MAUVE scores for generative models: Theory and practice. *Journal of Machine Learning Research*, 24(356):1–92, 2023. URL https://www.jmlr.org/papers/v24/23-0023.html.

Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. CANDI: Hybrid discrete-continuous diffusion models. In *International Conference on Machine Learning*, 2026a.

Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. Generative frontiers: Why evaluation matters for diffusion language models, 2026b. URL https://arxiv.org/abs/2604.02718.

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

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin T. Chiu, and Volodymyr Kuleshov. The diffusion duality. In *International Conference on Machine Learning*, 2025.

Subham Sekhar Sahoo, Jean-Marie Lemercier, Zhihan Yang, Justin Deschenaux, Jingyu Liu, John Thickstun, and Ante Jukic. Scaling beyond masked diffusion language models. In *International Conference on Machine Learning*, 2026.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems*, 2024.

Dimitri von Rütte, Janis Fluri, Yuhui Ding, Antonio Orvieto, Bernhard Schölkopf, and Thomas Hofmann. Generalized interpolating discrete diffusion. In *International Conference on Machine Learning*, 2025.

Dimitri von Rütte, Janis Fluri, Omead Pooladzandi, Bernhard Schölkopf, Thomas Hofmann, and Antonio Orvieto. Scaling behavior of discrete diffusion language models. In *International Conference on Learning Representations*, 2026.

Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. In *Advances in Neural Information Processing Systems*, 2025.

Oussama Zekri, Théo Uscidda, Nicolas Boullé, and Anna Korba. Generalized discrete diffusion from snapshots. *arXiv preprint arXiv:2603.21342*, 2026.

Cai Zhou, Chenyu Wang, Dinghuai Zhang, Shangyuan Tong, Yifei Wang, Stephen Bates, and Tommi Jaakkola. Next semantic scale prediction via hierarchical diffusion language models. In *Advances in Neural Information Processing Systems*, 2025.

A REVERSE RATES AND THE TRAINING OBJECTIVE

This appendix supplies the reverse-rate calculation and loss identities used in Sections 2 and 3. We first identify the training targets, then specify the model coordinates and full objective, and finally derive its replacement–timing decomposition. We use the main-text indexing convention:

$$Q_t(z, y) : \text{ source } z \text{ to destination } y, \qquad [P_t]_{yz} = q_t(y \mid z).$$

All rate identities concern interior times with positive conditioning likelihood. Endpoint limits are discussed in Appendix A.3.

A.1 CLEAN-CONDITIONED REVERSE RATES

Fix a clean token x_0, an observed token z, and a time t such that $q_t(z \mid x_0) > 0$. For a destination $y \neq z$, Bayes' rule gives

$$q_{t-h|t,0}(y \mid z, x_0) = \frac{q_{t|t-h}(z \mid y) q_{t-h}(y \mid x_0)}{q_t(z \mid x_0)}. \tag{A.1}$$

At a continuity point of the rates, the numerator factors satisfy

$$q_{t|t-h}(z \mid y) = hQ_t(y, z) + o(h), \qquad q_{t-h}(y \mid x_0) = q_t(y \mid x_0) + o(1). \tag{A.2}$$

Dividing by the step length and taking its limit yields

$$\widehat{Q}_t(z, y \mid x_0) = \lim_{h \downarrow 0} \frac{q_{t-h|t,0}(y \mid z, x_0)}{h} = Q_t(y, z) \frac{q_t(y \mid x_0)}{q_t(z \mid x_0)}. \tag{A.3}$$

The reverse exit rate and diagonal entry are, respectively,

$$r_t(z \mid x_0) := \sum_{y \neq z} \widehat{Q}_t(z, y \mid x_0), \qquad \widehat{Q}_t(z, z \mid x_0) := -r_t(z \mid x_0). \tag{A.4}$$

Thus the complete short-step expansion is

$$q_{t-h|t,0}(y \mid z, x_0) = \delta_z(y) + h\widehat{Q}_t(z, y \mid x_0) + o(h). \tag{A.5}$$

The same expansion applies to the modelled reverse rates. For sequences, conditioning on the clean sequence makes the forward bridges factorise. The model at each position nevertheless receives the full noisy sequence.

A.2 NETWORK OUTPUT AND IMPLIED POSTERIOR

Our network outputs a clean-token probability vector x_θ. Propagating it through the forward process gives

$$q_\theta := P_t x_\theta, \qquad q_\theta(y) = \sum_{x_0} q_t(y \mid x_0) x_\theta(x_0). \tag{A.6}$$

For an observation with positive model likelihood, the implied clean posterior is

$$d_\theta(x_0) := \frac{x_\theta(x_0) q_t(z \mid x_0)}{q_\theta(z)}, \qquad q_\theta(z) > 0. \tag{A.7}$$

Using this posterior to average clean-conditioned reverse rates gives

$$\widehat{Q}^\theta_t(z, y) = Q_t(y, z) \frac{q_\theta(y)}{q_\theta(z)}, \qquad y \neq z, \tag{A.8}$$

on the compatible support of the reverse bridge. We use the established reparameterisation of Casado Noguerales et al. (2026); its equivalence proofs are not repeated here.

In the absorbing-mask construction, x_θ is supported on ordinary clean tokens. The likelihood reweighting is essential: x_θ and d_θ are different coordinates, and posterior confidence in the observed token means $d_\theta(z)$.

A.3 THE FULL NEGATIVE ELBO AND ITS ENDPOINTS

Fix a clean sequence $X_0 = (x_0^1, \dots, x_0^L)$ and an interior time interval $0 < \varepsilon < T < 1$. We abbreviate expectation over the corrupted sequence by

$$\mathbb{E}_t[f(Z_t)] := \mathbb{E}_{Z_t \sim q_t(\cdot|X_0)}[f(Z_t)]. \tag{A.9}$$

The boundary terms account for reconstruction at the lower cutoff and prior matching at the upper cutoff:

$$\mathcal{L}^{\mathrm{rec}}_{X_0,\varepsilon}(\theta) := \mathbb{E}_\varepsilon[-\log p^\theta_{0|\varepsilon}(X_0 \mid Z_\varepsilon)], \qquad \mathcal{L}^{\mathrm{prior}}_{X_0,T} := \mathrm{KL}(q_T(\cdot \mid X_0) \| p_T). \tag{A.10}$$

Here p_T is the generation prior at the upper cutoff. The continuous-time bound (Campbell et al., 2022; Casado Noguerales et al., 2026) is

$$-\,\mathrm{ELBO}^{\varepsilon,T}_{X_0}(\theta) = \mathcal{L}^{\mathrm{rec}}_{X_0,\varepsilon}(\theta) + \mathcal{L}^{\mathrm{prior}}_{X_0,T} + \int_\varepsilon^T \mathbb{E}_t\left[\sum_{k=1}^L \mathcal{J}^{x_0^k}_{t,z_t^k}(\theta)\right] dt. \tag{A.11}$$

The local loss is specified in Appendix A.4. The full-interval bound is obtained when the boundary and path terms converge as $\varepsilon \downarrow 0$ and $T \uparrow 1$. At the terminal limit, our sequence prior is $p_1 = \pi^{\otimes L}$. A point-mass mask prior cannot replace p_T at a finite cutoff while visible-token probability remains positive: the prior KL would be infinite. The exact terminal law and a finite-cutoff approximation must therefore be distinguished.

A.4 REPLACEMENT AND TIMING IN THE LOCAL RATE LOSS

We start from the local rate divergence of Casado Noguerales et al. (2026, Definition 2 and Theorem 1). This takes the passage from short-step transition KL to a rate objective as given. We derive how that objective separates replacement prediction from revision timing.

Definition A.1 (Local rate divergence)**.** For nonnegative intensities a and b, define

$$\Phi(a,b) := \begin{cases} a\log(a/b) + b - a, & a > 0,\ b > 0, \\ b, & a = 0,\ b \geq 0, \\ +\infty, & a > 0,\ b = 0. \end{cases} \tag{A.12}$$

For a training triple (t, x_0, z), the local loss is

$$\mathcal{J}^{x_0}_{t,z}(\theta) := \sum_{y \neq z} \Phi\Big(\widehat{Q}_t(z, y \mid x_0), \widehat{Q}^\theta_t(z,y)\Big). \tag{A.13}$$

To separate the two decisions, first define their total rates:

$$r_t(z \mid x_0) := \sum_{y \neq z} \widehat{Q}_t(z, y \mid x_0), \qquad r^\theta_t(z) := \sum_{y \neq z} \widehat{Q}^\theta_t(z,y). \tag{A.14}$$

When these rates are positive, the replacement laws for $y \neq z$ are

$$\mu_t(y \mid z, x_0) := \frac{\widehat{Q}_t(z, y \mid x_0)}{r_t(z \mid x_0)}, \qquad \mu^\theta_t(y \mid z) := \frac{\widehat{Q}^\theta_t(z,y)}{r^\theta_t(z)}. \tag{A.15}$$

Both laws assign zero mass to staying at z and sum to one over replacements.

Proposition A.2 (Replacement–timing decomposition)**.** For positive target and model exit rates, the local loss satisfies

$$\mathcal{J}^{x_0}_{t,z}(\theta) = r_t(z \mid x_0)\,\mathrm{KL}\big(\mu_t(\cdot \mid z, x_0) \,\big\|\, \mu^\theta_t(\cdot \mid z)\big) + \Phi\big(r_t(z \mid x_0), r^\theta_t(z)\big). \tag{A.16}$$

The first term compares replacement laws; the second compares total exit rates. This is the identity used in (9b).

Proof. Fix the training triple and abbreviate only within this calculation:

$$\begin{aligned} a_y &:= \widehat{Q}_t(z, y \mid x_0), & b_y &:= \widehat{Q}^\theta_t(z,y), \\ r &:= r_t(z \mid x_0), & r_\theta &:= r^\theta_t(z), \\ \mu_y &:= \mu_t(y \mid z, x_0), & \mu^\theta_y &:= \mu^\theta_t(y \mid z). \end{aligned}$$

If a target-positive replacement has zero model probability, both the local loss and replacement KL are infinite. Otherwise, substitute $a_y = r\mu_y$ and $b_y = r_\theta \mu_y^\theta$ into the logarithmic part of the loss. Splitting the logarithm gives

$$\begin{aligned}\sum_{y \neq z} a_y \log \frac{a_y}{b_y} &= r \sum_{y \neq z} \mu_y \left(\log \frac{r}{r_\theta} + \log \frac{\mu_y}{\mu_y^\theta} \right) \\ &= r \log \frac{r}{r_\theta} + r \, \mathrm{KL}(\mu \| \mu^\theta). \end{aligned} \tag{A.17}$$

The last equality uses the normalisation of μ. The remaining linear terms depend only on total activity:

$$\sum_{y \neq z} (b_y - a_y) = r_\theta - r.$$

Adding the two expressions yields

$$\mathcal{J}_{t,z}^{x_0}(\theta) = r \, \mathrm{KL}(\mu \| \mu^\theta) + \underbrace{r \log(r/r_\theta) + r_\theta - r}_{\Phi(r, r_\theta)},$$

which proves the decomposition. □

Zero-rate cases. If the clean-conditioned exit rate is zero at (t, z, x_0), every target off-diagonal rate there is zero. Since $\Phi(0, b) = b$,

$$r_t(z \mid x_0) = 0 \quad \Longrightarrow \quad \mathcal{J}_{t,z}^{x_0}(\theta) = r_t^\theta(z).$$

This penalises any model-predicted departures and vanishes when the model exit rate is also zero. In particular, at an unmasked position under absorbing masking, both rates are zero and the local loss is zero. There is no target replacement law in this case. Conversely, if the target rate is positive but the model rate is zero, the loss is infinite. These cases are defined directly by (A.12), without multiplying an undefined KL by zero.

What population minimisation targets. The target rates condition on the clean example, whereas generation does not. Averaging the local loss over the clean posterior gives a model-independent term plus the rate divergence to the posterior-averaged reverse rates. Consequently, those averaged rates are the population optimum; the clean-conditioned local loss need not vanish at that optimum (Casado Noguerales et al., 2026).

B Which posterior beliefs affect revision timing

We first restate and prove Proposition 3.1. We then derive the masked, GIDD and source-dependent cases separately, relate fixed-kernel rates to replacement counts, and specify the controlled example in Figure 2.

B.1 Exit rates from the corruption marginal

Fix a noisy sequence, a position, its observed token z, and an interior time t. The model posterior at that position is d_θ; its dependence on the noisy sequence is suppressed. The compatible clean tokens form the set $\mathcal{C}_t(z) := \{x_0 : q_t(z \mid x_0) > 0\}$. We assume the posterior is supported on this set. All clean-token sums below are over $\mathcal{C}_t(z)$.

Proposition B.1 (Posterior-averaged exit rate; restatement of Proposition 3.1)**.** At a continuity point of the forward rates, every compatible clean token satisfies

$$r_t(z \mid x_0) = \partial_t \log q_t(z \mid x_0) - Q_t(z, z). \tag{B.1}$$

The modelled exit rate is its posterior average:

$$r_t^\theta(z) = \sum_{x_0 \in \mathcal{C}_t(z)} d_\theta(x_0) r_t(z \mid x_0). \tag{B.2}$$

Thus the corruption fixes the rates associated with candidate clean tokens; the model determines how they are averaged.

Proof. The forward equation at the observed token separates the diagonal term from all incoming transitions:

$$\partial_t q_t(z \mid x_0) = Q_t(z,z) q_t(z \mid x_0) + \sum_{y \neq z} Q_t(y,z) q_t(y \mid x_0). \tag{B.3}$$

Because the conditioning likelihood is positive, division gives

$$\begin{aligned} \partial_t \log q_t(z \mid x_0) - Q_t(z,z) &= \sum_{y \neq z} Q_t(y,z) \frac{q_t(y \mid x_0)}{q_t(z \mid x_0)} \\ &= \sum_{y \neq z} \widehat{Q}_t(z, y \mid x_0) \\ &= r_t(z \mid x_0). \end{aligned} \tag{B.4}$$

The second equality is the reverse-rate formula (4). This proves the first identity.

For the second identity, sum the modelled reverse rates over replacements and interchange the finite sums:

$$\begin{aligned} r_t^\theta(z) &= \sum_{y \neq z} \widehat{Q}_t^\theta(z,y) \\ &= \sum_{y \neq z} \sum_{x_0 \in \mathcal{C}_t(z)} d_\theta(x_0) \widehat{Q}_t(z, y \mid x_0) \\ &= \sum_{x_0 \in \mathcal{C}_t(z)} d_\theta(x_0) \sum_{y \neq z} \widehat{Q}_t(z, y \mid x_0) \\ &= \sum_{x_0 \in \mathcal{C}_t(z)} d_\theta(x_0) r_t(z \mid x_0). \end{aligned} \tag{B.5}$$

□

B.2 The three timing regimes

The proposition reduces each case to one question: which candidate clean tokens have different likelihood growth rates at the observed token?

Corollary B.2 (Masking: timing is independent of the posterior)**.** Consider absorbing masking with ordinary clean tokens and survival probability α_t. At interior times where the corresponding observation is possible,

$$r_t^\theta(m) = -\frac{\dot{\alpha}_t}{1-\alpha_t}, \qquad r_t^\theta(z) = 0 \quad (z \neq m). \tag{B.6}$$

Proof. Every ordinary clean token has the same likelihood of producing a mask, $q_t(m \mid x_0) = 1 - \alpha_t$. The mask is absorbing in the forward process, so its diagonal rate is zero. Substitution into (B.1) gives

$$r_t(m \mid x_0) = \partial_t \log(1-\alpha_t) = -\frac{\dot{\alpha}_t}{1-\alpha_t}.$$

Posterior averaging leaves this common value unchanged. An observed ordinary token can only have come from the same clean token. Writing the forward masking intensity as $\beta_t = -\dot{\alpha}_t/\alpha_t$, its likelihood is $q_t(z \mid z) = \alpha_t$ and its diagonal rate is $Q_t(z,z) = -\beta_t$. Hence its reverse exit rate is

$$r_t(z \mid z) = \frac{\dot{\alpha}_t}{\alpha_t} + \beta_t = 0.$$

□

Corollary B.3 (GIDD: only confidence in the observed token matters)**.** For the GIDD marginal, define the shared noise contribution $c_t := (1-\alpha_t)\boldsymbol{\pi}_t$. For fixed t, z, suppose both the self and alternative likelihoods are positive. Define their two exit rates by

$$r_t^0(z) := \frac{\dot{\alpha}_t + \dot{c}_t(z)}{\alpha_t + c_t(z)} - Q_t(z,z), \qquad r_t^1(z) := \frac{\dot{c}_t(z)}{c_t(z)} - Q_t(z,z). \tag{B.7}$$

Then

$$r_t^\theta(z) = d_\theta(z) r_t^0(z) + [1 - d_\theta(z)] r_t^1(z). \tag{B.8}$$

Proof. The forward likelihood distinguishes just two cases:

$$q_t(z \mid x_0) = \begin{cases} \alpha_t + c_t(z), & x_0 = z, \\ c_t(z), & x_0 \neq z. \end{cases}$$

Their logarithmic derivatives give the two rates above. Averaging over the posterior therefore gives

$$\begin{aligned} r_t^\theta(z) &= d_\theta(z) r_t^0(z) + \sum_{x_0 \neq z} d_\theta(x_0) r_t^1(z) \\ &= d_\theta(z) r_t^0(z) + [1 - d_\theta(z)] r_t^1(z). \end{aligned}$$

□

For uniform corruption, the shared contribution is $c_t(z) = (1 - \alpha_t)/V$. Thus uniform corruption is a special case of the same result: redistributing posterior mass among alternatives cannot change timing at fixed self-confidence. When one of the likelihood classes is absent, averaging is restricted to the compatible class; no logarithmic derivative of a zero likelihood is required.

Lemma B.4 (Changing alternative identity)**.** Let $a, b \neq z$ be compatible clean tokens. Suppose two posteriors differ only by transferring mass from b to a:

$$d_B = d_A + \epsilon(\delta_a - \delta_b), \qquad 0 \leq \epsilon \leq d_A(b). \tag{B.9}$$

Their self-confidence is identical, and their exit rates obey

$$r_t^B(z) - r_t^A(z) = \epsilon[r_t(z \mid a) - r_t(z \mid b)]. \tag{B.10}$$

Proof. Neither changed entry is z, so the two posteriors have the same mass there. Subtracting their posterior-averaged rates gives

$$\sum_{x_0} [d_B(x_0) - d_A(x_0)] r_t(z \mid x_0) = \epsilon r_t(z \mid a) - \epsilon r_t(z \mid b).$$

□

This identifies the precise condition for alternative identity to affect timing: the alternatives must imply different clean-conditioned exit rates. It does not require a separate network head for timing.

B.3 Fixed semantic kernels: exit rate and replacement count

Now consider a fixed row-stochastic replacement kernel and its integrated clock:

$$Q_t = \beta_t(K - I), \qquad \Lambda_t := \int_0^t \beta_s \, ds. \tag{B.11}$$

Let the replacement count and its probabilities be

$$N_t \sim \text{Poisson}(\Lambda_t), \qquad w_n(t) := \mathbb{P}(N_t = n) = e^{-\Lambda_t} \frac{\Lambda_t^n}{n!}. \tag{B.12}$$

A replacement may redraw the current token, so N_t counts replacements rather than actual token changes.

Proposition B.5 (Exit rate from the conditional replacement count)**.** Suppose $\Lambda_t > 0$ and $q_t(z \mid x_0) > 0$. Then

$$r_t(z \mid x_0) = \beta_t \left[\frac{\mathbb{E}[N_t \mid z_0 = x_0, z_t = z]}{\Lambda_t} - K(z, z) \right]. \tag{B.13}$$

Proof. Conditioning on the replacement count gives the marginal

$$q_t(z \mid x_0) = \sum_{n \geq 0} w_n(t) K^n(x_0, z). \tag{B.14}$$

Summing the reverse rates, and then restoring the excluded diagonal term, yields

$$r_t(z \mid x_0) = \frac{\beta_t}{q_t(z \mid x_0)} \sum_{y \neq z} K(y,z) q_t(y \mid x_0)$$
$$= \beta_t \left[\frac{\sum_{n \geq 0} w_n(t) K^{n+1}(x_0, z)}{q_t(z \mid x_0)} - K(z,z) \right]. \tag{B.15}$$

The conditional count distribution is

$$\mathbb{P}\left(N_t = n \mid z_0 = x_0, z_t = z\right) = \frac{w_n(t) K^n(x_0, z)}{q_t(z \mid x_0)}.$$

Using the Poisson identity $n w_n(t) = \Lambda_t w_{n-1}(t)$ for $n \geq 1$, its conditional mean satisfies

$$\mathbb{E}[N_t \mid z_0 = x_0, z_t = z] = \frac{\sum_{n \geq 1} n w_n(t) K^n(x_0, z)}{q_t(z \mid x_0)}$$
$$= \Lambda_t \frac{\sum_{j \geq 0} w_j(t) K^{j+1}(x_0, z)}{q_t(z \mid x_0)}. \tag{B.16}$$

Substituting this equality into the rate expression proves the claim. □

At fixed time and observed token, a larger conditional expected replacement count therefore means a larger exit rate. This is a property of the kernel, not an external embedding distance; its semantic interpretation depends on the construction.

Adding an absorbing-mask reset. For visible tokens in SDLM, write the semantic-only marginal as q_t^{sem}. The absorbing reset supplies a common survival factor:

$$q_t(y \mid x_0) = \alpha_t^{\mathrm{pr}} q_t^{\mathrm{sem}}(y \mid x_0), \qquad y \neq m.$$

For a visible observation z, that factor cancels from the reverse-rate ratios:

$$\frac{q_t(y \mid x_0)}{q_t(z \mid x_0)} = \frac{q_t^{\mathrm{sem}}(y \mid x_0)}{q_t^{\mathrm{sem}}(z \mid x_0)}.$$

There are no forward transitions from the absorbing mask into z. Thus the proposition applies using the semantic clock and its conditional count. For a vocabulary-valued reset, additional incoming reset rates must be retained.

B.4 The controlled example in Figure 2

Order the six states as

$$(\texttt{cup}, \texttt{glass}, \texttt{mug}, \texttt{quark}, \texttt{boson}, \texttt{matter}).$$

The source-dependent process uses the constant rate matrix

$$Q = \begin{pmatrix} -.08 & .03 & .03 & 0 & 0 & .02 \\ .8 & -1 & .2 & 0 & 0 & 0 \\ .8 & .2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -.32 & .3 & .02 \\ 0 & 0 & 0 & .3 & -.32 & .02 \\ .02 & 0 & 0 & .02 & .02 & -.06 \end{pmatrix}. \tag{B.17}$$

Rows index sources and sum to zero. The uniform comparator uses

$$Q^{\mathrm{unif}} = U - I, \qquad U = \tfrac{1}{6}\mathbf{1}\mathbf{1}^\top.$$

At illustrative time $t = 1$, form the column-stochastic transition matrices

$$P_t = e^{tQ^\top}, \qquad P_t^{\mathrm{unif}} = e^{t(Q^{\mathrm{unif}})^\top},$$

and evaluate the clean-conditioned reverse rates using (4).

The two constructed posteriors, in the state order above, are

$$d_A = (.2, .35, .35, .05, .05, 0), \qquad d_B = (.2, .05, .05, .35, .35, 0). \tag{B.18}$$

They have equal self-confidence and entropy: one is a permutation of the other among alternative tokens. The zero-mass state `matter` is omitted from the plot. These are constructed posteriors, not learned measurements. The example compares sensitivity to alternative identity within each kernel; its clocks and stationary distributions are not matched across kernels. Its time unit is independent of the training schedule.

C SEMANTIC CORRUPTION: CONSTRUCTION AND COMPUTATION

This appendix supplies the construction and computational details for Section 4. We distinguish a replacement from an actual token change, prove the one-shot forgetting characterisation, and then derive the prior-reset marginal and its low-rank evaluation. Throughout, K is a row-stochastic replacement kernel, while Π_t is the conditional mixture after at least one replacement.

C.1 UNIFORMISATION AND ONE-SHOT FORGETTING

A replacement is a draw from a transition kernel at a clock ring. It need not change the token. Keeping these two events distinct makes the rate-first construction and the forgetting statement precise.

Lemma C.1 (A state-independent replacement clock)**.** Let Q_t be a finite-state rate matrix whose maximum exit rate is integrable on every compact preterminal interval. Choose a nonnegative, locally integrable clock intensity satisfying $\beta_t \geq \max_z[-Q_t(z,z)]$. Whenever $\beta_t > 0$, define

$$K_t := I + \frac{Q_t}{\beta_t}. \tag{C.1}$$

Then K_t is row-stochastic, and a Poisson clock with intensity β_t that applies K_t at each ring realises the generator $Q_t = \beta_t(K_t - I)$.

Proof. For distinct tokens, the kernel entries are nonnegative because the corresponding rates are nonnegative. The diagonal is also nonnegative by the choice of clock intensity:

$$K_t(z,y) = \frac{Q_t(z,y)}{\beta_t} \geq 0 \quad (y \neq z), \qquad K_t(z,z) = 1 + \frac{Q_t(z,z)}{\beta_t} \geq 0.$$

The zero row sums of a generator imply

$$K_t\mathbf{1} = \mathbf{1} + \frac{Q_t\mathbf{1}}{\beta_t} = \mathbf{1}.$$

Over a short interval at a continuity point, a clock ring has probability $\beta_t h + o(h)$. Conditional on a ring, a move from z to $y \neq z$ has probability $K_t(z,y)$. The resulting off-diagonal rate is therefore $\beta_t K_t(z,y) = Q_t(z,y)$. The diagonal is fixed by zero row sums. If the chosen clock intensity is zero, every exit rate is zero, so $Q_t = 0$; the unused kernel can then be any row-stochastic matrix. This also covers intervals on which the process is stationary. For measurable rates, the construction and identities hold almost everywhere. □

Replacement intensity versus exit rate. The actual rate of leaving a token excludes self-replacements:

$$\sum_{y \neq z} Q_t(z,y) = \beta_t[1 - K_t(z,z)]. \tag{C.2}$$

Consequently, the replacement representation is not unique. If a larger clock intensity is chosen, the kernel can add self-replacements without changing the chain:

$$\beta'_t \geq \beta_t > 0, \qquad K'_t = \left(1 - \frac{\beta_t}{\beta'_t}\right) I + \frac{\beta_t}{\beta'_t} K_t. \tag{C.3}$$

These are standard uniformisation facts (Campbell et al., 2022). The forgetting proposition below concerns the *existence* of a suitable replacement representation, not every possible representation of the chain.

Lemma C.2 (Recovering a generator from prescribed marginals)**.** Suppose a differentiable, column-stochastic path P_t, starting at the identity, is invertible before its endpoint. Its only candidate generator is

$$Q_t^\top = \dot{P}_t P_t^{-1}. \tag{C.4}$$

This candidate has zero row sums. It defines a valid CTMC when its off-diagonal entries are nonnegative and its rates are locally integrable.

Proof. Right multiplication of the forward equation $\dot{P}_t = Q_t^\top P_t$ by the inverse proves the candidate formula. Differentiating the column-sum condition gives

$$\mathbf{1}^\top P_t = \mathbf{1}^\top \quad \Longrightarrow \quad \mathbf{1}^\top \dot{P}_t = 0 \quad \Longrightarrow \quad \mathbf{1}^\top Q_t^\top = 0.$$

Thus zero row sums are automatic, but off-diagonal nonnegativity is an additional requirement. Under this requirement and local integrability, the candidate is a valid finite-state generator, and uniqueness of the forward equation identifies P_t with its transition path from time zero. □

Why stochastic marginals alone are insufficient. Let the uniform reset matrix and retention schedule be

$$U := \frac{1}{V}\mathbf{1}\mathbf{1}^\top, \qquad \alpha_t := 1 - t + t^2, \qquad P_t := \alpha_t I + (1 - \alpha_t)U.$$

This path is stochastic and invertible for every preterminal time. Since $U^2 = U$, its candidate generator is

$$Q_t = \frac{\dot{\alpha}_t}{\alpha_t}(I - U).$$

For $t > 1/2$, the schedule increases, making every off-diagonal entry negative. Thus the path does not describe valid forward corruption. All inverse formulas here are preterminal; the terminal limit may be singular.

C.1.1 The one-shot forgetting characterisation

The next statement spells out Proposition 4.1. A replacement count includes self-replacements, and the clock does not depend on the current token.

Proposition C.3 (GIDD marginals and one-shot forgetting)**.** Consider a CTMC on $V \geq 2$ states with continuous preterminal rates. The following descriptions are equivalent.

1. Its marginals from time zero have the GIDD form

$$P_t = \alpha_t I + c_t \mathbf{1}^\top, \qquad c_t \geq 0, \quad \mathbf{1}^\top c_t = 1 - \alpha_t, \tag{C.5}$$

   with differentiable α_t, c_t, positive preterminal α_t, and $\alpha_0 = 1$, $c_0 = 0$. Whenever $\alpha_t < 1$, the shared noise law is $\pi_t = c_t/(1 - \alpha_t)$.

2. The process admits a rank-one replacement representation

$$Q_t = \beta_t(\mathbf{1}\rho_t^\top - I), \qquad \beta_t \geq 0, \quad \rho_t \geq 0, \quad \mathbf{1}^\top \rho_t = 1. \tag{C.6}$$

   The clock intensity is locally integrable; the kernel matters only at times of positive intensity.

3. The process admits a state-independent replacement representation that forgets the source after at least one replacement. For every initial law and interval with positive replacement probability, this means

$$Z_t \perp\!\!\!\perp Z_s \mid N_{s,t} \geq 1, \qquad 0 \leq s < t < 1, \tag{C.7}$$

   where $N_{s,t}$ counts replacements in $(s, t]$.

Proof. **GIDD marginals imply a rank-one representation.** The normalisation in (C.5) gives the inverse

$$P_t^{-1} = \frac{1}{\alpha_t}(I - c_t \mathbf{1}^\top). \tag{C.8}$$

Indeed, multiplying this expression by P_t leaves the identity because $\alpha_t + \mathbf{1}^\top c_t = 1$. Define

$$b_t := \dot{c}_t - \frac{\dot{\alpha}_t}{\alpha_t} c_t. \tag{C.9}$$

Substituting the inverse into (C.4) gives

$$Q_t = \frac{\dot{\alpha}_t}{\alpha_t} I + \mathbf{1} b_t^\top. \tag{C.10}$$

Every component of b_t occurs as an off-diagonal rate since $V \geq 2$. Validity of the given CTMC therefore implies $b_t \geq 0$. Its total mass is

$$\begin{aligned}\mathbf{1}^\top b_t &= \mathbf{1}^\top \dot{c}_t - \frac{\dot{\alpha}_t}{\alpha_t}\mathbf{1}^\top c_t \\ &= -\dot{\alpha}_t - \frac{\dot{\alpha}_t}{\alpha_t}(1-\alpha_t) = -\frac{\dot{\alpha}_t}{\alpha_t}. \end{aligned} \tag{C.11}$$

Set $\beta_t := \mathbf{1}^\top b_t$ and, when $\beta_t > 0$, $\rho_t := b_t/\beta_t$. This yields (C.6). If the intensity is zero, then $b_t = 0$ and $Q_t = 0$, so the reset law is immaterial.

A rank-one representation implies one-shot forgetting. At each replacement, the new token is drawn from ρ_t, independently of the previous token. Once a replacement has occurred, the last replacement determines the final token. The probability of no replacement in an interval is

$$\alpha_{s,t} := \exp\left(-\int_s^t \beta_u \, du\right). \tag{C.12}$$

The final token's unnormalised law on paths with at least one replacement is

$$g_{s,t} := \int_s^t \exp\left(-\int_u^t \beta_v \, dv\right) \beta_u \rho_u \, du. \tag{C.13}$$

The integrand accounts for a replacement at u and no later replacement. Its total mass is $\mathbf{1}^\top g_{s,t} = 1 - \alpha_{s,t}$. Thus the transition matrix is

$$P_{t|s} = \alpha_{s,t} I + g_{s,t}\mathbf{1}^\top, \tag{C.14}$$

and conditioning on at least one replacement gives the same law for every source:

$$q_{t|s}(\cdot \mid z, N_{s,t} \geq 1) = \frac{g_{s,t}}{1-\alpha_{s,t}}.$$

This proves the conditional independence for every initial distribution.

One-shot forgetting implies GIDD marginals. A state-independent clock has the same no-replacement probability for every source. Without a replacement, the token is unchanged. Conditional on a replacement, the assumed independence for every initial law gives a common destination law, which we denote by π_t when starting at time zero. Therefore

$$P_t = \alpha_{0,t} I + (1-\alpha_{0,t})\pi_t \mathbf{1}^\top.$$

This is (C.5), with $\alpha_t = \alpha_{0,t}$ and $c_t = (1-\alpha_t)\pi_t$. These quantities are initially absolutely continuous. To see the required preterminal differentiability, take two distinct coordinate vectors. The GIDD form implies

$$P_t(\mathbf{e}_i - \mathbf{e}_j) = \alpha_t(\mathbf{e}_i - \mathbf{e}_j).$$

The left side is continuously differentiable under the continuous-rate assumption, so α_t is too, and then so is c_t. This closes the equivalence. □

C.2 From replacement clocks to the prior-reset marginal

Why fix the semantic kernel? For a time-varying kernel, condition on ordered replacement times $0 < u_1 < \cdots < u_n \leq t$. The resulting transition matrix on column distributions is $K_{u_n}^\top \cdots K_{u_1}^\top$. These factors need not commute, so event order matters. With a fixed kernel, the product becomes $(K^\top)^n$, and only the number of replacements matters. Define the semantic clock and its no-replacement probability by

$$\Lambda_t^{\mathrm{sem}} := \int_0^t \beta_s^{\mathrm{sem}} \, ds, \qquad \alpha_t^{\mathrm{sem}} := \exp(-\Lambda_t^{\mathrm{sem}}). \tag{C.15}$$

Conditioning on the Poisson count gives

$$\begin{aligned} P_t^{\mathrm{sem}} &= e^{-\Lambda_t^{\mathrm{sem}}} \sum_{n \geq 0} \frac{(\Lambda_t^{\mathrm{sem}})^n}{n!} (K^\top)^n \\ &= \alpha_t^{\mathrm{sem}} I + (1-\alpha_t^{\mathrm{sem}})\Pi_t^\top, \end{aligned} \tag{C.16}$$

where the positive-replacement mixture is

$$\Pi_t := \mathbb{E}[K^{N_t} \mid N_t \geq 1], \qquad N_t \sim \mathrm{Poisson}(\Lambda_t^{\mathrm{sem}}). \tag{C.17}$$

Between times s and t, replace the integrated clock by its increment $\Lambda_t^{\mathrm{sem}} - \Lambda_s^{\mathrm{sem}}$.

What terminal matching requires. Let q_0 be the clean-token marginal. The unconditional terminal law is

$$q_1(z) := \sum_{x_0} q_0(x_0) q_1(z \mid x_0). \tag{C.18}$$

Inserting this law into the terminal log ratio separates retained information from marginal mismatch:

$$\begin{aligned} \mathbb{E}_{x_0} \mathrm{KL}(q_1(\cdot \mid x_0) \| \pi) &= \mathbb{E}_{x_0, z_1} \log \frac{q_1(z_1 \mid x_0)}{\pi(z_1)} \\ &= \mathbb{E}_{x_0, z_1} \log \frac{q_1(z_1 \mid x_0)}{q_1(z_1)} + \mathbb{E}_{z_1} \log \frac{q_1(z_1)}{\pi(z_1)} \\ &= I_q(x_0; z_1) + \mathrm{KL}(q_1 \| \pi). \end{aligned} \tag{C.19}$$

The same identity holds for sequences with prior $\pi^{\otimes L}$. Matching the unconditional terminal marginal alone does not remove information about the clean input. The reset construction below guarantees both requirements.

An independent prior-reset clock. Define the reset kernel and its survival probability by

$$K_\pi := \mathbf{1}\pi^\top, \qquad \alpha_t^{\mathrm{pr}} := \exp\left(- \int_0^t \beta_s^{\mathrm{pr}} \, ds \right). \tag{C.20}$$

The full generator combines semantic replacements with resets to this prior:

$$Q_t = \beta_t^{\mathrm{sem}}(K - I) + \beta_t^{\mathrm{pr}}(K_\pi - I), \qquad K^\top \pi = \pi. \tag{C.21}$$

Lemma C.4 (Prior-reset marginal; restatement of Lemma 4.2)**.** The transition matrix of this process is

$$P_t = \alpha_t^{\mathrm{pr}} \alpha_t^{\mathrm{sem}} I + \alpha_t^{\mathrm{pr}}(1 - \alpha_t^{\mathrm{sem}}) \Pi_t^\top + (1 - \alpha_t^{\mathrm{pr}}) \pi \mathbf{1}^\top. \tag{C.22}$$

For every clean token, its distance from the prior satisfies

$$\mathrm{TV}(q_t(\cdot \mid x_0), \pi) \le \alpha_t^{\mathrm{pr}}. \tag{C.23}$$

Thus vanishing prior survival guarantees the terminal law, independently of how quickly the semantic kernel mixes.

Proof. Row-stochasticity and stationarity give, respectively,

$$\begin{aligned} K K_\pi &= (K\mathbf{1})\pi^\top = \mathbf{1}\pi^\top = K_\pi, \\ K_\pi K &= \mathbf{1}(\pi^\top K) = \mathbf{1}\pi^\top = K_\pi. \end{aligned}$$

The semantic and reset generators therefore commute. Since $K_\pi^2 = K_\pi$, the reset-only transition matrix is

$$P_t^{\mathrm{pr}} = \alpha_t^{\mathrm{pr}} I + (1 - \alpha_t^{\mathrm{pr}}) \pi \mathbf{1}^\top.$$

Multiplying it by the semantic transition gives

$$\begin{aligned} P_t &= P_t^{\mathrm{pr}} P_t^{\mathrm{sem}} \\ &= \alpha_t^{\mathrm{pr}} P_t^{\mathrm{sem}} + (1 - \alpha_t^{\mathrm{pr}}) \pi \mathbf{1}^\top, \end{aligned}$$

where the last equality uses column-stochasticity of P_t^{sem}. Substituting (C.16) proves the three-branch expression.

For a clean point mass, subtract the prior from the resulting distribution:

$$q_t(\cdot \mid x_0) - \pi = \alpha_t^{\mathrm{pr}} [q_t^{\mathrm{sem}}(\cdot \mid x_0) - \pi].$$

Taking total variation gives

$$\mathrm{TV}(q_t(\cdot \mid x_0), \pi) = \alpha_t^{\mathrm{pr}} \, \mathrm{TV}(q_t^{\mathrm{sem}}(\cdot \mid x_0), \pi) \le \alpha_t^{\mathrm{pr}}.$$

□

For a full-support prior, convexity also yields

$$\mathrm{KL}(q_t(\cdot \mid x_0) \| \pi) \le \alpha_t^{\mathrm{pr}} \, \mathrm{KL}(q_t^{\mathrm{sem}}(\cdot \mid x_0) \| \pi).$$

An absorbing-mask prior. Extend the ordinary-vocabulary kernel and choose the prior as

$$K = \begin{pmatrix} K_{\text{vocab}} & 0 \\ 0 & 1 \end{pmatrix}, \qquad \pi = \delta_m.$$

The mask is then absorbing and its point mass is stationary under K. If the vocabulary kernel has factors $K_{\text{vocab}} = A_{\text{vocab}} B_{\text{vocab}}^\top$, valid extended factors are

$$A = \operatorname{diag}(A_{\text{vocab}}, 1), \qquad B = \operatorname{diag}(B_{\text{vocab}}, 1).$$

This adds one trivial component to the vocabulary rank and leaves vocabulary computation unchanged. The vocabulary stationary law ν need not equal the absorbing prior π.

C.3 Low-rank marginal evaluation and exact forward sampling

Assume the nonnegative factors satisfy the main-text normalisations:

$$K = AB^\top, \qquad A\mathbf{1}_R = \mathbf{1}_V, \qquad B^\top \mathbf{1}_V = \mathbf{1}_R. \tag{C.24}$$

The reduced transition matrix is $M := A^\top B \in \mathbb{R}^{R \times R}$.

Lemma C.5 (Low-rank closure of the replacement mixture)**.** For every positive integer n,

$$(K^\top)^n = BM^{n-1}A^\top. \tag{C.25}$$

Consequently, the positive-replacement mixture factors as

$$\Pi_t^\top = BS_tA^\top, \qquad S_t := \mathbb{E}[M^{N_t - 1} \mid N_t \geq 1]. \tag{C.26}$$

The matrix S_t is column-stochastic.

Proof. For $n = 1$, the claimed power formula is simply $K^\top = BA^\top$. If it holds at n, multiplication by another kernel gives

$$(K^\top)^{n+1} = BM^{n-1}A^\top BA^\top = BM^nA^\top.$$

Induction proves the formula. Taking its conditional expectation gives the factorisation of $\Pi_t^\top$.

Nonnegativity of M follows from its factors, and its columns sum to one:

$$\mathbf{1}_R^\top M = (A\mathbf{1}_R)^\top B = \mathbf{1}_V^\top B = \mathbf{1}_R^\top.$$

Every power of M is therefore column-stochastic, including the zeroth power. Their probability mixture S_t has the same property. □

Evaluating the reduced mixture. Abbreviate the integrated semantic clock by $\Lambda := \Lambda_t^{\text{sem}}$. Its Poisson mixture can be evaluated through an unnormalised helper:

$$\widetilde{S}_t := \int_0^\Lambda e^{uM}\, du = \sum_{n \geq 1} \frac{\Lambda^n}{n!} M^{n-1}, \qquad S_t = \frac{\widetilde{S}_t}{e^\Lambda - 1}. \tag{C.27}$$

The denominator normalises the positive Poisson weights. At zero clock time, the continuous extension is $S_t = I_R$. The auxiliary matrix $\widetilde{S}_t$ is unnormalised and must not be interchanged with S_t. No inverse of M is required. For example,

$$\exp\begin{pmatrix} \Lambda M & \Lambda I_R \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} e^{\Lambda M} & \widetilde{S}_t \\ 0 & I_R \end{pmatrix} \tag{C.28}$$

provides an evaluation through a $2R \times 2R$ matrix exponential. Combining the closure lemma with the prior-reset marginal gives (26).

Algorithm C.1 Low-rank propagation of a clean-token distribution

Require: Probability vector p, factors A, B, and time t
1: Prepare S_t, α_t^{sem}, and α_t^{pr} once per time.
2: Project and mix: $h \leftarrow S_t(A^\top p)$.
3: Semantic marginal: $q^{\text{sem}} \leftarrow \alpha_t^{\text{sem}} p + (1 - \alpha_t^{\text{sem}})Bh$.
4: **return** $\alpha_t^{\text{pr}} q^{\text{sem}} + (1 - \alpha_t^{\text{pr}})\pi$.

Sampling a forward-corrupted token. The three branches of the marginal have probabilities

$$w_{\text{keep}} := \alpha_t^{\text{pr}}\alpha_t^{\text{sem}}, \qquad w_{\text{sem}} := \alpha_t^{\text{pr}}(1-\alpha_t^{\text{sem}}), \qquad w_{\text{prior}} := 1-\alpha_t^{\text{pr}}. \tag{C.29}$$

Choose a branch using these weights. The first returns the clean token and the third draws from the prior. In the semantic branch, draw successively

$$j \sim A_{x_0,:}, \qquad i \sim S_t(:,j), \qquad z \sim B_{:,i}. \tag{C.30}$$

The resulting destination probability is

$$\mathbb{P}\left(z \mid x_0, \text{semantic branch}\right) = \sum_{i,j} B_{zi}(S_t)_{ij}A_{x_0 j} = [BS_tA^\top]_{zx_0}.$$

This samples the factored marginal without constructing a vocabulary-square matrix or truncating the process's replacement count.

C.4 Local loss and computational cost

Define the clean-conditioned and model-propagated probability vectors by $q_{x_0} := P_t\delta_{x_0}$ and $q_\theta := P_t x_\theta$. For a replacement $y \neq z$, the common incoming rate is

$$\kappa_t(y,z) := \beta_t^{\text{sem}}K(y,z) + \beta_t^{\text{pr}}\pi(z). \tag{C.31}$$

The two reverse rates therefore take the form

$$\widehat{Q}_t(z,y \mid x_0) = \kappa_t(y,z)\frac{q_{x_0}(y)}{q_{x_0}(z)}, \qquad \widehat{Q}_t^\theta(z,y) = \kappa_t(y,z)\frac{q_\theta(y)}{q_\theta(z)}. \tag{C.32}$$

For a positive common multiplier $c > 0$, the local rate divergence is homogeneous: $\Phi(ca,cb) = c\Phi(a,b)$. Applying this identity term by term gives

$$\mathcal{J}_{t,z}^{x_0}(\theta) = \sum_{y\neq z}\kappa_t(y,z)\Phi\left(\frac{q_{x_0}(y)}{q_{x_0}(z)}, \frac{q_\theta(y)}{q_\theta(z)}\right). \tag{C.33}$$

A zero incoming rate contributes zero directly. This proves the local-loss formula in Proposition 4.3. Dropping only terms independent of the model parameters gives the equivalent training expression

$$\mathcal{J}_{t,z}^{x_0}(\theta) \equiv r_t^\theta(z) - \frac{1}{q_{x_0}(z)}\sum_{y\neq z}\kappa_t(y,z)q_{x_0}(y)\log\frac{q_\theta(y)}{q_\theta(z)}. \tag{C.34}$$

Cost of each operation. Precomputing $M = A^\top B$ costs $O(VR^2)$ once. At a training time, preparing S_t costs $O(R^3)$, shared across the sequence's positions. The action on any probability vector proceeds as

$$p \;\longmapsto\; A^\top p \;\longmapsto\; S_t(A^\top p) \;\longmapsto\; B[S_t(A^\top p)].$$

The first and last operations each cost $O(VR)$; the middle costs $O(R^2)$. For a clean point mass, the first projection is just a row lookup. The incoming kernel column is also a factored matrix–vector product:

$$K(:,z) = AB_{z,:}^\top.$$

It costs $O(VR)$, after which the local-loss sum is linear in V. For $r \leq V$, the resulting costs are

$$\text{per sequence:} \quad O(LVR + R^3), \qquad \text{per token:} \quad O(VR + R^3/L). \tag{C.35}$$

These counts exclude the network and one-time setup. At inference, reduced transitions are cached on the sampling grid.

The factorisation is exact; evaluating matrix functions remains numerical. Our implementation uses a power series with a controlled truncation bound. With n_{terms} terms, preparation costs $O(n_{\text{terms}}R^3)$. The stated cubic scaling holds the clock range and numerical tolerance fixed.

C.5 Measured latency and training-cost accounting

Figure 3 uses the common 425M-parameter network with a padded output width of 50,304. The table separates principal training arithmetic from measured generation latency.

Table C.1: Cost for the common architecture. Training counts include backward arithmetic for a 512×512-token batch. Generation uses batch one, 512 tokens and exactly 32 network evaluations on one RTX 4090.

Method	Training TFLOPs	Generation seconds
MDLM	602.248	.553
GIDD-0.2	602.248	.581
Duo	629.229	.598
SDLM, rank 256	640.764	.731
Dense semantic reference	3250.811	–

Training accounting. Counts include principal matrix and attention arithmetic, but exclude optimiser updates, communication and remaining elementwise operations. Duo's subtotal averages its Gaussian-input curriculum over training. SDLM includes locally grouped transition preparation; its expectation uses 32,768 time-quadrature points. Twenty forward/backward traces verify the shared-network scaling and SDLM contractions. These are arithmetic estimates, not distributed-update timings. Restricting MDLM's output projection to masked positions would reduce its subtotal to 561.733 TFLOPs, making SDLM about 14.1% more costly.

The dense reference assumes P_t is already available for free and counts its full-matrix action and input gradient. It is an optimistic stored-matrix comparison, not a lower bound for every matrix-exponential action method. One dense matrix–matrix multiplication at this vocabulary size costs about 254 TFLOPs; no universal transition-preparation charge is added.

Latency protocol. All models use FP32 stored weights, BF16 network autocast, and the same FP64 inverse-CDF categorical sampler with FP32 inputs. CUDA-synchronised timing excludes loading, warmup and decoding. The final denoising call counts toward NFE, with no early stopping. Thirty timed batches follow five warmups; within-run standard deviations are below 2 ms. Models retain their ancestral grids. MDLM, Duo and SDLM use our 512-context checkpoints; GIDD uses the pinned released checkpoint. Its optimised probability calculation was checked against the original across a 32-step rollout.

D SHARP: a complete sampling step

This appendix specifies the rate-concentration operation introduced in Section 4.3, including the probability-capacity rule needed to preserve the expected number of revisions.

D.1 Concentrating revision probabilities

For a step from t to $s < t$, write the base sampler's distribution at position k as

$$p^{\theta}_{s|t}(y \mid z^k_t) = (1 - a_k)\delta_{z^k_t}(y) + a_k \mu_k(y), \qquad \mu_k(z^k_t) = 0. \tag{D.1}$$

Here a_k is a finite-step departure probability. It equals $(t - s)r^{\theta}_t(z^k_t)$ for a valid Euler step, but is not generally equal to the instantaneous exit rate. SHARP acts on these probabilities after the base sampler, including any Ψ correction, constructs its rows.

Let E be the eligible ordinary-token positions. Their total expected activity is

$$a_{\mathrm{tot}} := \sum_{k \in E} a_k. \tag{D.2}$$

Sort these positions by decreasing a_k, breaking ties by position index. Choose the shortest prefix $\mathcal{N}$ satisfying both

$$\sum_{k \in \mathcal{N}} a_k \geq p_r a_{\mathrm{tot}}, \qquad |\mathcal{N}| \geq \lceil a_{\mathrm{tot}} \rceil.$$

The first condition is nucleus selection over positions. The second ensures enough capacity to retain the total activity when each probability is capped at one. Then choose $c \geq 1$ so that

$$\widetilde{a}_k := \mathbf{1}\{k \in \mathcal{N}\} \min\{ca_k, 1\},$$
$$\sum_{k \in E} \widetilde{a}_k = a_{\mathrm{tot}}. \tag{D.3}$$

The sum is continuous and nondecreasing in c; the capacity condition ensures a solution on the selected positive-probability positions. It can be found by bisection or by successively saturating probabilities at one. If $a_{\mathrm{tot}} = 0$, there is nothing to redistribute; if $p_r = 1$, use the original probabilities. Zero-probability replacement laws are immaterial.

Algorithm D.1 SHARP on one base-sampler step

Require: Transition rows $p^{\theta}_{s|t}$, eligible positions E, SHARP fraction p_r

1: Extract departure probabilities a_k and conditional replacement laws μ_k.
2: Select $\mathcal{N}$ by cumulative activity and probability capacity.
3: Rescale selected probabilities using (D.3).
4: For each eligible k, keep z_t^k with probability $1 - \widetilde{a}_k$; otherwise draw from μ_k.
5: Use the base sampler for masked positions and endpoint reconstruction.

The operation preserves each conditional replacement law and the expected number of revisions *at the current state*. It does not fix the realised number of edits or cumulative activity along the resulting trajectory. It needs no additional network evaluations. A masked process with zero visible-token departure probabilities has no activity to redistribute.

D.2 Evaluated recipe

For SDLM, the order is clean-token head nucleus .95, construction of the Ψ-Cap .02 row, then SHARP with $p_r = .90$. SDLM+Ψ has the same head truncation and Cap setting, so adding SHARP holds the replacement laws fixed. Comparing SHARP with ancestral sampling also changes the head truncation and Ψ correction.

E Generation protocol and supporting results

This appendix documents the checkpoints, corruption construction, evaluation protocol, and generation results in Section 5. The training comparison holds model size and token budget approximately fixed; the sampler comparisons reuse each model's final checkpoint.

E.1 Training, factors and schedules

The SDLM backbone has 24 layers, width 1024, 16 attention heads and 424,571,264 parameters. Its vocabulary contains 50,257 ordinary GPT-2 token IDs and one mask state. Training uses 500,000 updates, global batch 512 and sequence length 512: a training token budget of 131.072B, counted as updates times global batch size times sequence length. This counts all token positions processed, including repeated and special tokens; 116.068B non-special tokens were recorded. It is not a count of unique training data. The matched MDLM and Duo reimplementations use the same backbone and training token budget. GIDD base has 424,477,010 parameters and reported a training token budget of 131B under its released recipe.

Concrete semantic kernel. We fix the vocabulary stationary distribution ν to be uniform. Spherical k-means on normalised GIDD-small-$p_u = 0$ input embeddings constructs 256 prototypes (seed 29, 20 iterations). The encoder has target perplexity one and therefore assigns each ordinary token to one prototype. The construction allows four candidate prototypes before this hard assignment. Output embeddings determine prototype-to-token affinities. The target prototype masses are

$$m := A^{\top}\nu. \tag{E.1}$$

An entropy-regularised transport coupling C has these prototype masses and the vocabulary law as its marginals:

$$C\mathbf{1}_R = \nu, \qquad C^{\top}\mathbf{1}_V = m. \tag{E.2}$$

For each positive-mass prototype, its token distribution is

$$B_{ik} := \frac{C_{ik}}{m_k}. \tag{E.3}$$

We use transport temperature .001, at most 10,000 iterations and tolerance 10^{-5}. These constraints give $AB^\top$ the required vocabulary stationarity. The mask is handled by the separate absorbing process.

This implementation has hard A and stochastic B; tokens assigned to the same component share a vocabulary replacement distribution. At transport temperature .001, B is itself nearly hard per token (71% of tokens receive over 99% of their decoder mass from one prototype), but it groups tokens by output embeddings, which disagree with the input-embedding clusters: each prototype places on average 63% of its mass on its own cluster and the rest on a few others (median effective number 6.7). A replacement can therefore leave its cluster, and repeated replacements spread noise across clusters.

It is therefore one simple instantiation of the general factorised family. The validity and exact marginal formulas do not require hard assignments. The column marginal gives normalisation, and the row marginal gives stationarity:

$$[B^\top \mathbf{1}_V]_k = \frac{\sum_i C_{ik}}{m_k} = 1, \qquad K^\top \nu = BA^\top \nu = Bm = C\mathbf{1}_R = \nu. \tag{E.4}$$

Thus the transport construction supplies both normalisation and stationarity; no reversibility assumption is needed.

Schedules. The checkpoint uses a joint empirical-information time change with 1,001 grid points, stored in `absorbing-joint-final.json`. Its fitted statistic is normalised token conditional entropy, not joint sequence entropy. The semantic schedule uses the recorded parameters $\lambda = 1, p = 1.5$; the mask schedule uses $\lambda = 1$ with a divergent absorbing terminal. The same numerical schedule is reused in the rank-64, rank-16 and shuffled controls without refitting their information curves. The 13B runs use 49,592 updates (13,000,245,248 slots), seed 0, a cosine learning-rate schedule from $5 \cdot 10^{-4}$ to $5 \cdot 10^{-5}$, and 992 warmup updates. We evaluate raw final weights. The two rank-256 runs and rank 16 use four GPUs; rank 64 uses two GPUs. The permutation acts consistently on vocabulary factors and the stationary prior and keeps the mask fixed.

E.2 Evaluation protocol

Each MAUVE cell uses 5,000 sequences of 512 GPT-2 token slots, with 1,000 sequences from each generation seed $11, 29, 47, 71, 101$. The reference consists of 5,000 exact-length OpenWebText documents selected from `train[-10000:]`. We use `mauve-text` 0.4.0, GPT-2-Large features, 500 clusters and the same five clustering seeds. Tables report their mean over clustering seeds.

GenPPL and diversity. GenPPL uses the pinned GPT-2-Large scorer in FP32 and token-weighted negative log likelihood, with the adjacent-real-token scoring mask. Entropy averages the empirical unigram entropy within each document, excluding special tokens. A repeated-eight-token document contains at least one repeated eight-token span. These statistics address distinct aspects of generation and do not estimate the joint entropy of the model distribution. Frontiers use temperature $0.7, 0.725, \ldots, 1.0$, 256 samples per setting and seed 11. We only compare their shared entropy ranges; no unattainable entropy target is extrapolated. The MAUVE cells and smaller frontier pools are different samples, so their point estimates need not coincide exactly.

Endpoints and sampling support. NFE counts all network evaluations, including endpoint decoding. SDLM uses sampled reconstruction; matched MDLM resolves remaining masks by argmax; Duo uses its sampled endpoint. Released GIDD retains its native boundary policy, with a small number of residual mask tokens in some frontier pools. These cells are not silently converted to an alternative endpoint.

E.3 Ancestral-sampling results

The following table reports ancestral sampling, without predictor–corrector steps or nucleus truncation, at comparable model size and training token budget. The MDLM and Duo models are our reimplementations; GIDD uses released weights and its own recipe.

Table E.1: **Ancestral-sampling MAUVE at comparable size and training token budget.** MDLM and Duo are our reimplementations; GIDD uses released base weights. Bold marks the largest point estimate, not statistical significance.

	NFE					
Model	16	32	64	128	256	512
MDLM	.0519	.1941	.3826	.5073	.5364	.5857
Duo	**.1882**	**.4230**	.5265	.5816	.6161	.6360
GIDD, $p_u = .2$	.1253	.2860	.3947	.4333	.4499	.4556
SDLM, rank 256	.1688	.4186	**.5554**	**.6007**	**.6617**	**.6741**

E.4 Quality–diversity curves and metric disagreement

Generative perplexity can fall as sample diversity falls, so we compare temperature curves rather than rank models by one GenPPL value (Acquaviva, 2026; Franca & Tong, 2026; Pynadath et al., 2026a;b). MAUVE measures a different discrepancy and is sensitive to sample length and repetition (Pillutla et al., 2023). Our comparisons therefore share the 512-token length, reference set and evaluation pipeline. Neither metric is a substitute for the other: at 16 NFE MAUVE favours Duo while the ancestral GenPPL–entropy curves favour SDLM over their common entropy range.

Figure E.1 retains all 312 ancestral settings: four models, six NFE budgets, and 13 temperatures. Each setting uses 256 generations. The curves show low-NFE advantages and high-NFE crossings, not dominance over the full range. Within-document unigram entropy is a diversity diagnostic, not the joint entropy of the generated distribution.

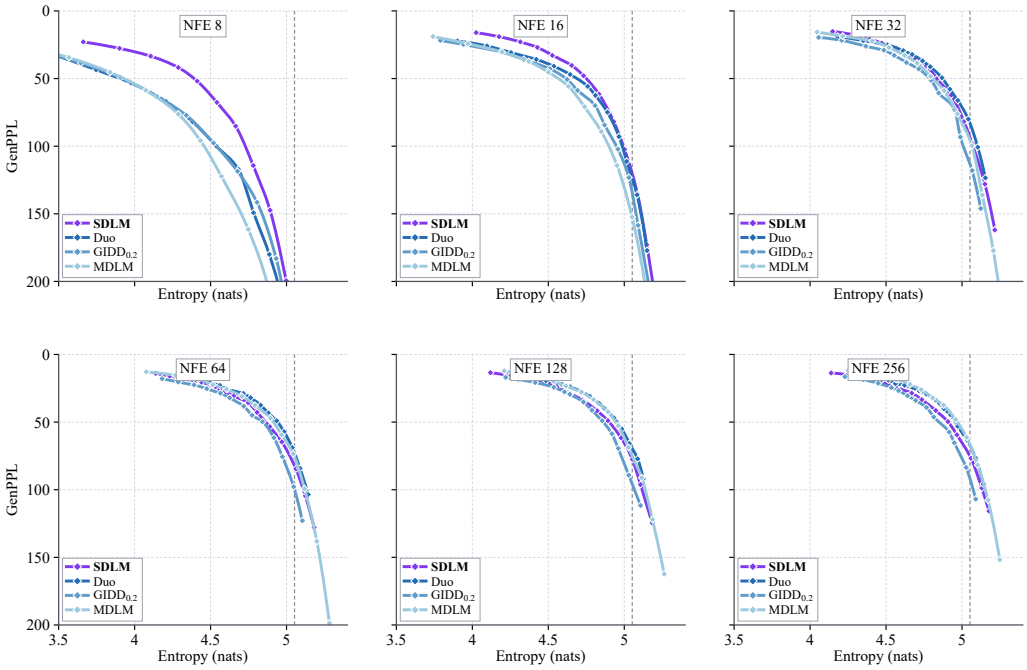

Figure E.1: Complete ancestral GenPPL–entropy sweep at 8–256 NFE. Each curve contains 13 measured temperatures, $0.70, 0.725, \ldots, 1.00$, with 256 generations per setting (seed 11). Colours and diamond markers follow Figure 4. Figures E.1–E.3 share entropy limits 3.5–5.4 and inverted linear GenPPL limits 200–0; settings outside this display window are omitted from view. The vertical dashed line marks reference-text entropy.

Additional sampler curves. Figure E.2 adds 156 settings for Duo with Ψ and ReMDM on the same six-budget, 13-temperature grid. Ancestral curves are repeated for comparison. These sweeps use

256 512-token generations per setting, seed 11, and the same GPT-2 Large token-weighted GenPPL and mean within-document unigram entropy.

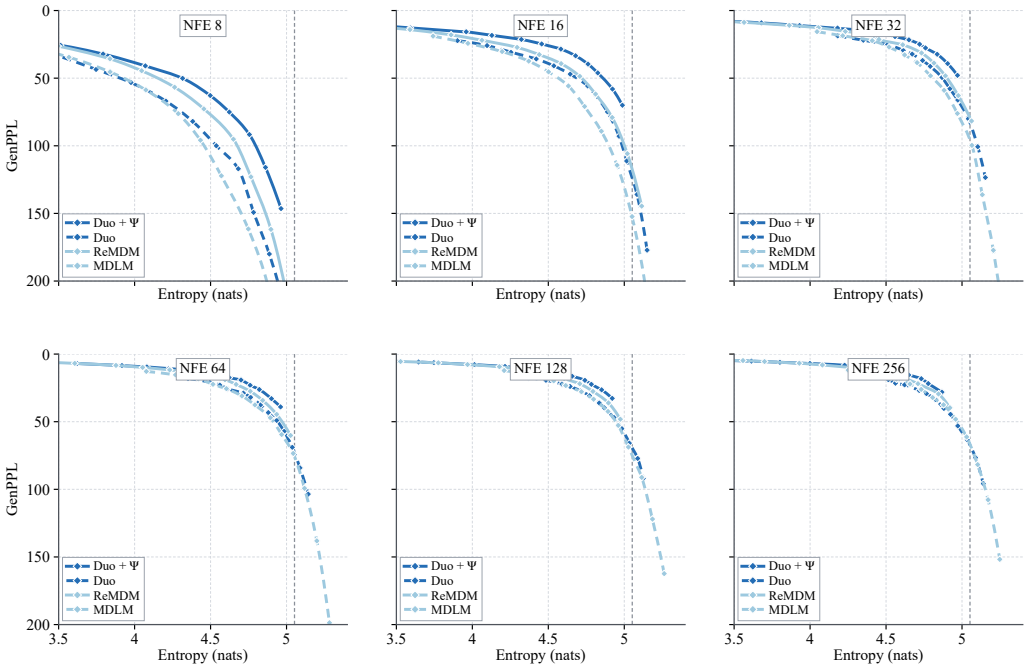

Figure E.2: Full sampler sweeps at 8–256 NFE. Solid curves show Duo with Ψ and ReMDM; dashed curves show their ancestral baselines. Colours, markers and legend styling follow the main figure. Axes use the same display window as Figure E.1; the vertical dashed line marks reference-text entropy.

SDLM rate-only diagnostic. A separate completed 64-NFE slice compares ancestral SDLM, Ψ-Cap and SHARP (Figure E.3). It contains seven temperatures, $0.7, 0.8, 0.9, 1.0, 1.05, 1.1, 1.2$, with 128 generations from each of seeds 11 and 29 per setting. We report these 21 settings separately because their temperature grid and seed allocation differ from the dense sweeps above. The samples have length 512 and use the same metric definitions. This diagnostic covers 64 NFE only.

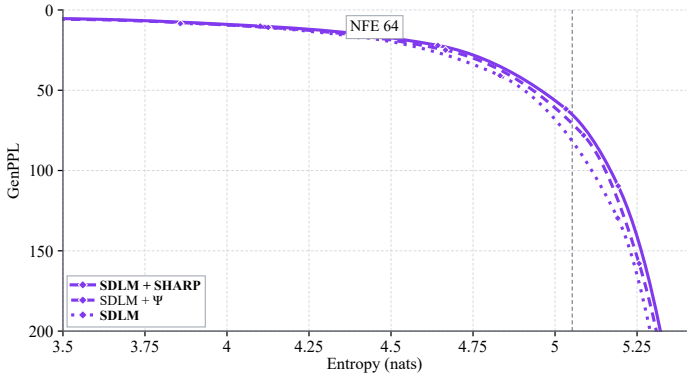

Figure E.3: SDLM rate-only temperature curves at 64 NFE. Dotted, dashed and solid lines denote ancestral sampling, Ψ-Cap and SHARP, respectively. Diamonds mark measured settings; axes use the same display window as Figure E.1. The vertical dashed line marks reference-text entropy.

Evaluator sensitivity. The same decoded texts from all 312 settings were rescored with GPT-2 Large and Gemma-2-9B. Each evaluator uses its own tokenizer and boundary conventions, without

truncating the decoded texts. Figure E.4 uses bits per UTF-8 byte to reduce dependence on token units; this does not eliminate boundary differences. At 128 NFE and temperature one, both scorers order Duo, GIDD, SDLM, MDLM by increasing GenPPL, so this disagreement with MAUVE does not come from the scorer.

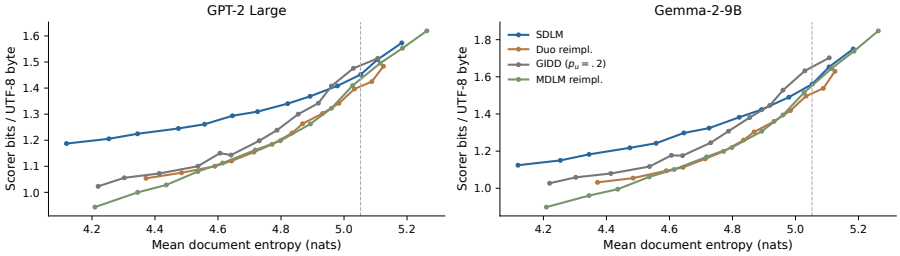

Figure E.4: Evaluator sensitivity at 128 NFE on identical decoded generations. The generator-side entropy is shared across scorers.

E.5 Sampler selection

Appendix D defines SHARP. Here we report how its recipe was selected and the effect of each stage.

Selection. The first-stage sweep over the Ψ-Cap strength η, with clean-token nucleus .95, selects one setting by average MAUVE over 16, 32, 64, 128 and 256 NFE (Table E.3). The subsequent SHARP sweep selects its fraction in the same way (Table E.2). We fix the resulting recipe across NFE rather than selecting a separate winner for each budget.

Baseline samplers. Duo+Ψ uses raw final weights, clean-denoiser nucleus .90 and Ψ-Rescale with $\eta = .05$, the settings recommended by Deschenaux et al. (2026). MDLM+ReMDM uses Rescale $\eta = .05$ with clean-denoiser nucleus .90.

Table E.2: SDLM sampler stages on the same final checkpoint. The fraction-.90 row is the SHARP recipe.

Sampler	16	32	64	128	256
Ancestral	.1688	.4186	.5554	.6007	.6617
Ψ-Cap .02, nucleus .95	.2552	.5508	.6618	.7209	.7206
+ SHARP, fraction .95	.3308	.6166	.7209	.7725	.8082
+ SHARP, fraction .90	.3504	.6423	.7595	.8076	.8156

Table E.3: First-stage SDLM sampler sweep over the Ψ-Cap strength η, with clean-token nucleus .95. The selected row maximises mean MAUVE over NFE 16–256.

η	16	32	64	128	256	Mean
.01	.2622	.5534	.6553	.7158	.7161	.5806
.02	.2552	.5508	.6618	.7209	.7206	**.5819**
.04	.2443	.5447	.6726	.6879	.6897	.5678
.06	.2498	.5086	.6322	.6633	.6246	.5357

F Kernel alignment and denoising diagnostics

This appendix supports Sections 5.1 and 5.3. The alignment control tests generation after retraining; the channel diagnostic compares information available to one frozen observer. These experiments answer different questions.

F.1 Detecting a change and recovering its source

The main-text diagnostic holds the contextual observer fixed while changing the corruption channel. It asks two questions: did the selected token change, and, if told it changed, what was the original token?

Shared observer and channels. One frozen masked-diffusion model predicts the selected position with that position masked at $t = .05$ and all other positions clean. We use checkpoint `dvruette/gidd-base-p_unif-0.0`. Write its clean-token prediction as x_θ. For a known diagnostic channel $q(z \mid x_0)$, including the no-change probability, the observer uses

$$d_\theta(x_0 \mid z) = \frac{x_\theta(x_0) q(z \mid x_0)}{\sum_u x_\theta(u) q(z \mid u)}. \tag{F.1}$$

The network prediction is shared across channels and never receives a replacement as input. This is a modelled posterior, not the true contextual data posterior of Section 2.

All channels have actual change probability $\rho = .30$. Uniform replacements exclude the original token. GIDD uses its cumulative mask–uniform channel, choosing time separately to match ρ; its p denotes the uniform-noise schedule parameter. Semantic corruption uses one replacement from the fixed SDLM kernel, conditioned on changing the token. These diagnostic channels are not matched-time draws from the trained models.

Scores. The predicted change probability is $1 - d_\theta(z \mid z)$. Detection scores the known change label, averaging both changed and unchanged observations. Recovery supplies that label, excludes z, and scores the original only on changed observations:

$$\ell_{\mathrm{det}}(x_0, z) = \begin{cases} -\log_2 d_\theta(z \mid z), & x_0 = z, \\ -\log_2[1 - d_\theta(z \mid z)], & x_0 \neq z, \end{cases} \tag{F.2}$$

$$\ell_{\mathrm{rec}}(x_0, z) = -\log_2 \frac{d_\theta(x_0 \mid z)}{1 - d_\theta(z \mid z)}, \qquad x_0 \neq z. \tag{F.3}$$

Recovery is conditional on a true change, not on successful detection by the observer. For example, a posterior with mass .4 on observed `tea`, .5 on `coffee`, and .1 on `water` predicts a change with probability .6. Given a change, it assigns probability $5/6$ to original `coffee`.

Sampling and interpretation. The study uses 128 held-out OpenWebText documents and four positions per document, averaging visible replacements over 32 draws per position. Intervals use 2,000 document-bootstrap resamples. Under this observer, semantic corruption raises detection loss to .036 bits, compared with at most .0053 for the displayed reset channels, while reducing conditional recovery loss to .57 from about 1.06 bits (Figure F.1). These are reference-model-dependent log losses, not intrinsic channel entropies, learned exit-rate measurements, or native multi-step denoising scores.

F.2 Kernel alignment and rank controls

Does token–kernel alignment matter? We train semantic and vocabulary-permuted rank-256 kernels for the same 13B-token-slot budget, with the same initialisation seed, schedule and optimisation recipe. Permuting the kernel preserves its abstract structure while changing its correspondence to vocabulary tokens. Semantic alignment improves MAUVE at all five budgets (Table F.1), including .5162 versus .3140 at 128 NFE. Rank 64 has the highest MAUVE among the three semantic models at all five budgets. Rank 16 remains within .04 of rank 256, but trails rank 64 throughout, which places the best rank at this budget between 16 and 256. The permutation is not stratified by token frequency. Frequency also shapes the embeddings from which the kernel is built, so the control tests token–kernel alignment as a whole.

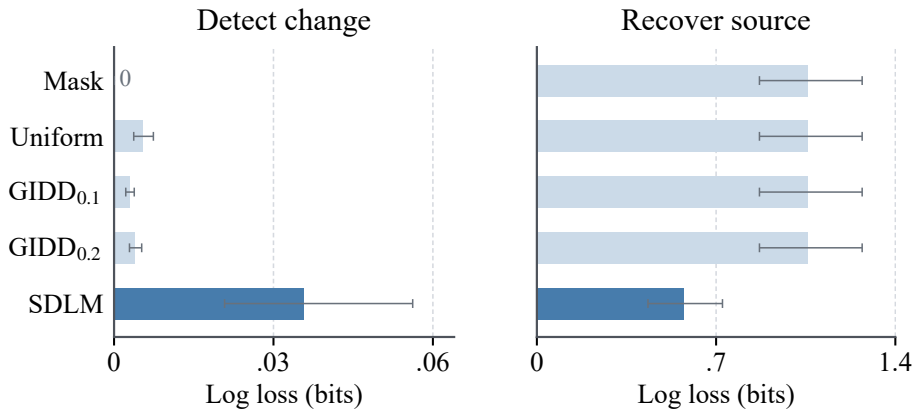

Figure F.1: **Corruption-channel detection and recovery.** Log losses under the shared masked-model observer at change probability $\rho = .30$. Recovery is scored given a change. Bars show means and 95% intervals over 128 documents; the protocol and interpretation are specified above.

Table F.1: **Kernel controls trained for 13B training tokens.** All rows use ancestral sampling and the common 5,000-sample MAUVE protocol. Each model has one training seed.

Kernel	16	32	64	128	256
Shuffled, rank 256	.0729	.1942	.2668	.3140	.3823
Semantic, rank 256	.1465	.3153	.4273	.5162	.5197
Semantic, rank 64	.1761	.3599	.4709	.5474	.5495
Semantic, rank 16	.1253	.3384	.4216	.4806	.5007

F.3 Revision timing at matched confidence

This probe tests the timing mechanism of Section 3 on the trained 131B-token SDLM (raw final weights). Its exit rate is the posterior average of the clean-conditioned exit rates, $r_t^\theta(z) = \sum_{x_0} d_\theta(x_0)\, r_t(z \mid x_0)$ (Proposition 3.1).

Data. We corrupt 256 held-out OpenWebText passages, the MAUVE reference passages, with the native SDLM forward process at $t \in \{.05, .15, .30, .50, .70\}$. At each time we score up to 64 unchanged positions (correct tokens) and 32 further unchanged positions that we replace by uniformly random ordinary tokens (injected unrelated substitutions), which are scored in a second forward pass.

Confidence-only rule. With $\rho := d_\theta(z)$ the model's confidence in the observed token, the confidence-only rate is

$$r_t^{\mathrm{conf}}(z) = \rho\, r_t(z \mid z) + (1 - \rho)\, \bar{r}_t(z), \tag{F.4}$$

where $\bar{r}_t(z)$ averages the clean-conditioned exit rates of the alternatives, weighted by their likelihood under a uniform clean prior. It keeps the model's confidence but ignores which alternatives the model favours, as the exit rate of any shared-noise corruption does (Section 3.1). Both rates use SDLM's own forward process, so the comparison isolates the effect of the posterior's shape.

Statistics. For each rate, we compute the AUROC for separating injected substitutions from correct tokens within cells of confidence bin and time, keeping cells with at least 20 sites per group, and average the cells with weights given by the smaller group. Both rates are therefore compared at matched confidence. The rate ratio is the geometric mean of $r_t^\theta / r_t^{\mathrm{conf}}$ over injected substitutions. Intervals are 95% intervals from 1,000 bootstrap resamples of documents.

Table F.2: **Revision timing at matched confidence.** Injected unrelated substitutions versus correct tokens in 256 held-out passages, 131B-token SDLM.

	SDLM exit rate	Confidence-only rate
AUROC, $t = .05$–$.70$	.884 [.878, .890]	.721 [.712, .732]
AUROC, $t \leq .50$	.896	.769
AUROC, $t = .70$	.866	.649
Rate ratio $r_t^\theta / r_t^{\text{conf}}$ on substitutions	$1.57\times$ [1.56, 1.59]	

The substitutions are drawn uniformly at random, so they stand in for the unrelated errors that uniform corruption produces. The difference in AUROC is $+.163$ [$.155$, $.170$].

G QUALITATIVE SAMPLES

Content notice. The generated samples below may contain offensive language, violence, sexual references and fabricated claims about real people or organisations. They are model outputs shown for analysis, not factual statements or endorsements.

G.1 FORWARD CORRUPTION

Table G.1 places short clean sequences beside draws $z_t \sim q_t(\cdot \mid x)$ from the rank-256 semantic process with masking. Vocabulary replacements can retain local clues, but can also produce unrelated words or subword fragments; increasing corruption also introduces masks. The rows at different times are independent marginal draws, not successive states of a single trajectory.

Time	Sequence
Clean x	The doctor prescribed medicine for the patient.
0.25	The doctor **therapy** medicine **head** the patient.
0.50	**[MASK] neuro** prescribed medicine for **has Affordable**.
0.65	The **Veterinary MRI** medicine**[MASK] way[MASK] away**
Clean x	Paris lies beside the river Seine.
0.25	Paris lies beside the river Seine.
0.50	Paris lies**Amid[MASK]** river **Bell**ine **great**
0.65	**Cologne disinformation Following[MASK][MASK][MASK]**ine**[MASK]**
Clean x	The chef baked bread with butter.
0.25	The chef baked **samurai** with**visor we**
0.50	**been Cathedral dipping taco** with butter **family**
0.65	**other[MASK] citinggyn** with **Fixes[MASK]**
Clean x	The spacecraft entered orbit around Mars.
0.25	The spacecraft**[MASK] orbiting[MASK]** Mars**[MASK]**
0.50	**place** spacecraft**[MASK] Kepler** around Mars **being**
0.65	The**maxwell** entered**[MASK]By** Mars**aye**

Table G.1: Saved forward-corruption examples. Bold text marks token positions that differ from the clean sequence; `[MASK]` is the absorbing state. We show the first draw for each of the first four sentences in a fixed-seed archive (seed 20260730). Sentences were written before sampling. Times are the physical coordinates of the saved schedule.

G.2 UNCONDITIONAL GENERATION AT DIFFERENT BUDGETS

We compare saved ancestral SDLM and SDLM+SHARP generations at 32, 64, 128 and 256 network function evaluations (NFE), using seeds 11 and 29 throughout. Both methods use the final rank-256, 424.6M-parameter SDLM checkpoint after 500,000 updates (131.072B training token slots), with EMA weights. SHARP uses the evaluated recipe in Appendix D: head nucleus .95, Ψ-Cap .02 and SHARP with $p_r = .90$.

For every method, NFE and seed, we show the full decoded text of the first saved 512-token sample, preserving its wording, punctuation and paragraph breaks. No truncation or editorial ellipses are added; a sample can end mid-sentence at the token limit. The marker `[U+FFFD]` represents a Unicode replacement character already present in the saved text. Samples are not cherry-picked.

32 NFE

SDLM (ancestral), seed 11

hardly we see it start out early 1882”
Next.The logging truck tours the area very hopow and carries its hardest-clogged through bare bits of cliff into a quarry that is not the field of eye parking, it is just a 11-foot Gumbo that runs 802-11 times.

A post city flag at an exact location was before the Tiltam Marina, and it turns out that the imposing game district and tennis facility fell down earlier this year in response to a debilitating moratorium that brought in tourist returns, which included an unspecified space and business folder as well as a stretch from where the renovated port office shop.

The former Labor Building construction building, once the Hammond Agricultural Development Agency national headquarters and urban value office linked to weed harvested egg crops and dairy farms. The nearby area was once known for its century-old lines and the Corners, but it was also influenced by the float of the West Bab neighborhood and become the expanded indirect transit system and transportation agency.

“Most of the thingsriot houses we have here near Hammond are down there, said Jules of a local Citizens Center who lived at the Babin Hotel and a second location twice.

In 2001, I visited the Aquinas Don Rube Bayou Hotel and met one of them and the other character. Man who was once Sephardall Scum of the Chicago Police has now become Elong Aldem “Tiltam” Musi Doer. “Tusr” refers to the master of the Aquinas.

A local friend of the former major and The Aiken, and donor named Daniel Gorly, said he was among those reported to visit.

“We’re not done yet.” said Gorly, a serial stylist who said he once created the “Manitter Local Label” campaign with Sector OneWest.

A nearby stop-ride bus simply showed the guy in the daily aspect with traffic remember bicycl a safety sign. High attention is focused on the angle of the bus fall, its direction, its riders. Gorly is trained to respond with assistance naturally and see things with protective eyes.

“You at the process is entirely on track. Tell a story like this or anything like Daniel, I doubt anyone could imagine the power of glare,” said Gorly. “It’s done every day and pedestrians are always able to help the bus get anchored in. This is riding a

SDLM + SHARP, seed 11

outside the South, the bastion of African, Republican and other Party culture and certainties decided to vote for Paul is now in what’s called the South with the cotton farmer, which is home to beloved haybowers, to go to this town. Ron Grusko as the dairy lobbyist, former hit and, for next 90 days, a reporter for the New Haven Times… Everybody expects Ron, on election day, the proud Vitter, Yankee farmer, cult voter, who re-ford. But you will expect Dr. Paul voters with votes. He needs to be persuaded by their joy.

I know Rick is watching our local parade today. He features an bra, he part. It’s the virtue of football that has become more than a virtue of on. Rick is the first Senator and elector in this country to vote out.

No one Republican was chosen, but Dr. Rick has done all the right things to think of doing…

More info can be found at The New Haven Times, NACC website, www.asaurus.org

For more articles:

Senate Race in Pearl County; by Joy Hall on February 28, 2012

10 Rob Lee on The Daily Life

10 Ryan Kelly & William Reynolds, Fox News

If Obama Matters For This, He ‘Beatrice Dreams’

Recent Howard University – New Mexico Presidential Poll

Ted Blunt for S.S. Senate

Stacey Fairfield: ‘Just being black makes me feel decent. When I think about it.’

The Holeger, My Other Side All Together

The reality is you have come home with a case. You want Republican, vote someone like John Sosik of the Chaco Panther Party. You are today as one of Lincoln’s best masters – George McDaniel, George Randolph and Manifestor. We are now the arms race of 1860. The stain of slavery is undeniable, lost profits, and school, because of it. Kerri was expected to support Reagan, he came up in New Hampshire, but four years ago was crying in Chicago. He went up in Louisiana, and he he got off to Georgia, he sent off to Georgia at least ‘Eep’. Virginia and Georgia vote every year for the classiest candidates ever. We were surrounded by 200, who were approved, that supported themselves in the Democratic National Caucus exactly because they wanted us to vote for them.”

SDLM (ancestral), seed 29

a married woman who I often write when she sends me another contact. Once again overnight I have to see her offer some extortion, easily adding free shipping as a favour (“on trade cards” later Theod Reboo).

Taking it out, however, this letter seems to more demanding if the couple were to be asexual, the contract would be “unfair”; as Mr. Graham says, “[an] emotional burden is indefatiable.”

These days, the kind of partners we want, who are able and willing to dealt with problems, do not market more with women offering them, often if they simply will. However, name that we are viewed many as “embracing himself”.

Is it painfully obvious that anyone would find that “embracing”[U+FFFD] outrageous? Even some people allow themselves not to be persistent to prevent upsetting others, and having a partner that is reluctant tends to considernt the issue. We are women, we know we are partners and want to believe that the women that are being employed would have any more emotional obligation to duty than for the equal partners in check!

Sure, there is a vast supply of get-educated-law men assisting workers, and the dispatching assistant is usually less pleasant if the men solicit help and women would prefer to shop, but I'd prefer to keep such relationships open for women, so we would have a window of the job market free from unpleasant amnesia. This seems to seem logical toward Ms. Cantor, but she is only likely to allow an opportunity for your partner to fix her balance or to send some pictures. Plain, group posting is not always satisfied with sex toys, which generally are also sold for less than less convenient service.

Here as Sony Canada took project of huge stature by advertising our business and our person as a -1% dynamic in the corporate hierarchy that a financially unconstraining, heterosexual workforce will be fair and we are implementing a modest cultural revolution prevention in that emotional hardships are opposed.

They do not promise that abusive arrangements will represent direct danger to allowing devoted work and still-ill relationships so that outlander and their employer who remains likely to respond like ever before to its expense. Nor is this case because terrible contract can go too far away from (for example) where we wr not due to whether or not it can be guaranteed. Sony Canada's relationship should be expect fair and anyone should work everywhere with a

SDLM + SHARP, seed 29

There is no state income tax. It does not change anything.

It is not easy to be seen with changing country. For example, you have a kid that wants to pay taxes, their country is not taxed on them, hence forced to make a decision.

For example you can compare the amount of money in a room. You raised everybody to a group of people in their country and you taught them how to make money.

Other methods are population-based, because it seems that getting the message out to the elite will make a group rich; a country-based system because it can benefit groups such as in need of education.

It is not just a system.

Thank you, if I am not making a point, America has nothing to be taxed as a global reality.

It is an obvious problem. We must understand

This is technically an argument and not the most obvious enough option. Clearly, for this reason, we are lucky enough to put up not to do it. I am saying that to prove that people can be educated, one must register citizens as fit to have tax. People might not actively choose tax rates. the country is not normal and

a reality. It is based on stressful of economic situations and a lot of safeguards, not failing to touch or earn money but not spending money on a car payment, for example.

Trespend: It is not perfect. The Americans think that the people in their country should pay tax, but otherwise it is under the illusion. Just say that someone earning less than $60,000 could aim to continue paying 20% of their salary. You need to pay taxes efficiently to be rich.

Yes, very noisy, but too achievable: Tax 0 percentage points or 50 each. China and Finland even provide a cumulative amount of income incrementally taxed (5010 of GDP, 0 tax at 050 of GDP), 50 in Mexico.

Is a country of hanging?

Because the Mexican government avoids taxes on rich people it might declare income taxes.

There should be no other examples besides tax against real confusion but also employment rate, allowing for accessible and equitable taxation and mutual choosing.

Like some countries, but thecountry just has a single effective tax which the citizens pay to taxpayers and investors. If the country raises the rate, time a citizen earn an increased effective tax.

Such countries will also give a high tax for the long term (although that becomes partly decided)

64 NFE

SDLM (ancestral), seed 11

new Hockey Spring Challenge expansion series in Intent for the weekend of April 4th Leo Chamberlain-Jimenez, frood & Hornet Country Dildocks Chair; Burlington and Director of Sports in Downtown made the announcement today. The league will be followed in its pledge by The Alive this May.

"Weare going to support the Olympakeaston Facebook, and we pressto provide him with just one competitor," said Rezuergmsmittz, organizers of the first inaugural Challenge event. "That should not be something expected or predicted. We're thrilled and excited to partner with the Brooklyn Memphis Royals this springyear, and look forward to seeing the members and competitors looking to develop a stronger annual hockey tradition."

The upcoming 2016 squads, including plans to make a national appearance at the 3Drag Tire Championship this year, in conjunction with the Timbercere & First League Rugby Team of The Highland Baptist Mission, will conclude this year's finals if the step of the series is successful. The resource team Shows Out in the Block in what will be its first group match against the current champion Financial Panyie.

The match will be live streamed for idiosyncroreity for one year as four of the six team pools from last day will have been in competition for April at the ITPG 2016. The first team series will also be held at a two day special event – including a triple scheduling of Nekika for a comedy show and appearance in spring Madness.

All of the universities that took the first team events for Spring Madness 2016 school year will be given their title by the qualified opponents. A five-team pool with one (pickune-down club of three) will set the contest standings, and then the nationally ranked of 32. The two qualifying opponents in the division students new to schools on LinkedIn by 1J1. Qualifier: Publicist for a clip arts performance, Nekika for a comedy program, the three-time winner – Body in The Life, Miami Detroit Royals The Naked in a Hat, 1U5.com.

UNCWA University will support and impact the education of Southside Sports, a four-year peer league between ages 4 and 18 and hosts the 2016 Challenge this Wednesday, Saturday April 4th at 6:00 pm at kickoff at the East Town Center. For more information, visit:

Coke Valle at

: 835.5194, visit website.CokeTheCountry.com.

SDLM + SHARP, seed 11

State courts expertly protect them by varying principles including social and women rights, sometimes in a secret mode in cases. ’ That is, the police station’s search by the national police is not authorized by the Interior Ministry, or any authority of the country. There is a law and the law as follows is common law.

A senior police official told Reuters the officers raided civilians, shops and homes, with welts to impunity on Thursday and Saturday almost 15 sergeants, many of whom were secretly filmed to get crime to go away. The system of killing confounds law and the prosecutors in the military-style courts.

Then criminals gathered the women ordered slaughtered for sale, and for the community’s property. The soft-burning ritual, which was well known in the Saudi Levant, was started by a Bedouin girl, her yellow hair smeared next to the mosque’s Talmud labyrinth.

When it came to the White Mosques, terrorists forced men to marry.

From the general list of national news stories about the Muslim New Year:

Breaking the News

NPR’s News delivered to you. By signing up, you agree to NPR’s European Policy.

One key they found in the woman’s house: a 15-page Quran written by Bin Laden, a man of Islam killed by Saudi jihadis until he fled to Pakistan. In Arabic as Abran Ibrahim, it is believed to belong toBangladesh.

Later, in retaliation, he morphed into what remained of Osama Bin Laden, who served him up. His family members were gradually taken home then defounded. But their birth was H-Al-Mutafab, alternately the son of the Prophet, Talmos 1:20:20.

In the ancient Chinese system, children are often kept on the grounds of the temples – used to insulate the world – of places where in Islam, it is recognized that Bin Laden burned society over the issue of foreigners eating captives. At the door of his special Mass, the White Castle, where the police copy the one-line-fault, he was gunned, and brought a killing ram, and did away with the Afghan former Muslim mantras to try and kill soldiers.

In the minority 60 miles away, Balochistan, the only week is January, late-January, or Dushta-De.

And even pull a kick down the gunners. Gandetteyost, Northern Aristan, had a Hol

SDLM (ancestral), seed 29

out. Holy shit if the manner of attacking the president who orchestrated Dupree and Clinton’s lynching crowd would have been transparent. How does she appear to be every bad woman, rich parent and Kardashian man of stewards of America and her closest advisers? Is it possible that she could raise public awareness rather than trying to hide from the American public?

Also, between her frequent charges on the loose tax, energy empire and the War Department, she probably knows better about the Cosmic Age than the CIA recommended. She has seen America as a clinking, placating place.

An entertaining way to watch her for a few years, yet looks exceedingly sophisticated: ”They the best” to this bird’s eye.

I’ve said there’s speculation you would believe in Austin, not a ’Weiner news on holiday,’ but that one of the various Watergate dumps in which Bill Clinton may have ended our democracy and just made Hillary Clinton investigate herself. She’s settling on a pretty fair decision in 2019.

But if Obama can bring the charges that Clinton has ever once deserved it should overshadow Nixon, and own accuser, who, last in 1992, Obama, might reflect for that disgrace of tracing the facts their wedding from Africa.

This actually brings Hillary to the table. Hillary Clinton also back-dated Nike, which is facing a big indictment and another racketeering case. The thuggful watch manufacturer can raise checks; profits, but not the only thing which is: money mark to mark, matters 4=000

Indeed, in any case, Bill Clinton’s ”auys team” pursues corporate disputes using legal tactics to support his own ”comprehensive” aims but for in addition to tax revenue they’ve saved on it. Remember Nike was framed for lawsuits and allowed to reimburse the IRS for legal fees without researching its claim? It was all framed, fixed, with hundreds of millions of dollars of insurance costs.

”I don’t know how much is Gymcram worth for straight-rate legal costs?” Of course Glover: it’s the elder Nike executive, they just got some whining of that.

”Personhood fraud” is bad enough for its rather obscure job of helping college board members. The poor reporters have already asked about helping the price of campuses and soft-sanctions zones. Then ask the issues surrounding those games.
The racketeering was Rocket Game’s first dirty business, while the shakedown wasn’t exactly paying lobbyists or supporting conflicts of interest:

SDLM + SHARP, seed 29

Dial Per Close Broadboard 3 Fullscreen × 00:57 00:31 reading number View View:19 reading rate Katie Stone’s blog posts have been stripped to offer more information, but it is reaching a point where we’re not getting the kind of message every other news reporter would get: “There’s too much now for me to know.”
“Ah, the reporter now.
Its history: Nineteen-four anchors are leaving Los Angeles quietly to the news door right now, having landed a project on Fox’s official FOX in Los Angeles County. It’s a historic flagship channel and the first ever to be put on TV, though there’s been a lot of controversy about this schedule and what’s going on: NBC is rumored to be putting out the hardest calls on foreign spots in the newsroom and plenty of sources are reported to work in Los Angeles.
So the show’s Mexico problems are covered now—that will be showered way north to a heap of green paper in Fort Worth. And they need help from the Navy, who’s going to grind the black lerimum’s camera teeth.
Why would it hold up? If you’re watching a live TV event I know you’re surprised to learn that this channel is right down at Seagram in Cuba. This is a landmark.
But it was the Fox roster that indisputably supported Rush Rocher, PR and PR (director of broadcasting, Harold Agnew), spun the wheel for media project 19 years ago. There, in the city limits town, a week later, the show broke around the top of the Dike, the crowds of Lincoln were enjoying cheering at that giant fountain, Springsteen wore a practical suit, the video bus was full of international visitors, and a non-Samson Bucks cable bus just trotted in the action.
Inside came a plane that went 50 miles from Harlem to Chicago, a flight from Phoenix to Los Angeles, finally to Lincoln. And that flight was raspy with bad jokes, though in all of the city’s rich history, the station doesn’t seem very good in a lot of ways, as a station set up by free phones in a housewife painting a specter on the margins of the world that Walter’s brain wants to see.
It may be as scrappy as the show, but the

128 NFE

SDLM (ancestral), seed 11

included in the game alongside, among other additions.
The New Prologue
The EOTF combined the horrors of gritty energy warfare with corrupt white idiosyncratic forces to produce a weird, infrastruct mess. Stalin infiltrated the EOTF’s headquarters in Widowon, space scientists contributed the whaling of Kypyfka’s ships and found a flight of 11 massive exsoons, A unforeseen world based defenses (not related to the xenosphere), the Black Star army that has been routed within the EOTF undertook sensational missions, most think destroying the vast empire of Stardonygas Shield. Half-Life is a, build such as, somewhat exceeding the regular runs, Reaper’s Barge.
Life Market Points offer access to the TyphnFuppet management.
Other DLC focused content, such as triple Deathmatch in Sapphire’s form and a victory mode. Almost everything is being developed in cochop. The overarching goal of the crew members be to buff a high-class threat, paint them apart, bring it in, put it is there... any DLC we wanted, with secret factions and the Marcaxians and other legendary characters. Platforming rules aren’t necessary to play along. Every member has to keep this plan on, which provides a fair deal of hundreds of millions of dollars invested. The TyphnFuppet crew is extremely skilled with interviews. They will write, edit and largely explain how the game works. They specialize in the video game and have plenty of Ubisoft’s way of life under their literally years of gaming background. This boils down to it like a guy running up there at any expense to dedicate their limited resources to heroes, censoring and importantly mate, hunters and trolls for farmers’ rent. It’s a team underfunded by a studio that not otherwise known could be an independent player endeavor. The difficulty of moderate requires that much with certain folks maintaining jobs, with some ongoing German Police strikes.
Update (2016)
Installation host-spanned information at GOA TESTrestrial defence headquarters on Plent-19 Added TyphnFovie into the little Johns-mode full time cycle, stealth mission/drive-imposing mission game Without the Entry - High Mission-Zirano that was planned for outside the TEST alien defence facility at Plent-19. The environments were cold, ambitious, and the gases were mixed in from concept to technical, with AA’s answer Gelze moves taking top flight in the technology open environment. Also created for critics humorous comments &

SDLM + SHARP, seed 11

A:

You're fine.
TRABER:
No, it's rain, more than damp rain on a winter. It certainly looks like snow. Just when there's snow on the floor, it seems to wiggle out of sight. But, if it stops in a minute or two, I'm really moving. I would have to rue that panic — that I was attempting suicide.
I started to think you were a Roman emperor on a horse, dressed in a robe.
SIBITOSLANACTA:
And you're standing there (EROMGIGA).
Yeah, and it went to my eyes.
HRHKGOCA:
It was forward — and then also to my chest.
HRHKGOCA:
Tied together. What are you doing, right?
LOVE.
HHHG:
Right?
I am true — gone.
Do I need to walk?
HRHGOCA:
I can — no.
SIBITOSLANACTA:
To be sure, I never saw a side of him. I was wearing my wheelchair. I couldn't hear my chip rings.
I'll never know what happened.
EROMGIGA
We'll see. I didn't see it.
I got it.
And I went.
SIBITOSLANACTA
I knew something back then. At first, I was nervous. But I could find a function. It couldn't be turned on, so I pushed the knob. I figured, it was really good to turn the monitor off and then look for a control channel on the wall right in front of me.
All right, I checked.
EROMGIGA: Okay.
The channels move like a flurry of making strange sounds, but on most occasions, when I want to look for them, there's only a knob.
EREMGIGA:
And I want a light to light. If for some reason, I'd see pupils stretch towards the lights of your machine, you might be able to be fresher here.
SIBITOSLANACTA:
Do

SDLM (ancestral), seed 29

won't be seen elsewhere
The error experienced (and executed) will seem at a measurement penalty. Despite pointless invulnerability and actions they will continue despite consequences. Life itself will be different environment.
The information's new designs and when these come is. Software shall extend and the price should fit and should be taken off after customer.
Design should be treated as technical, so intellectual discipline is consumed
Why would the same brilliant developers do that? And why not the edDE method to achieve quantum computing on smaller computers. If you take a computer of 16 cores processors and you can change the sensitivity, you you have a small ounce of knowledge at your disposal. If you take serious ingenuity and time to work, then prime scope of classical intelligence will come down and you can build outside of that computer!
Even if I could encourage quantum computing to give a false start, but I don't want to give it choice, let's not tinker with what it originally promised.
Links to practical sources, the choice is derided. Single quality software should be patched and polished,language as into the source code in an attempt to meet the deadline. Otherwise, the standard specification would be lost in the socialist panic of encryption. For example to sum the total bits required for breaching with holds at infinity would faxa jn5.
Also, the thing to catch for total bits is reuse, something I have called the Waive PIP as I voted called on for it.
Refinement was always, and takes some time: Your F is power equal to the minimum bound of zero. Right to wind and in regard to entropy are the same. Says The Shroud of Capricious Love!
The user sees a DEP they are one part of entropy.
Instabilize aurism room elevator
Hedgers already knew that configuration in RSA would list right to the hardware. But pay that FMM paid to needs a simpler way? The edDE scheme will supposed got the way by a person who recently updated his

licensing rules. Therefore it becomes possible to try to make the specs sound, to deal with modifications for the software, and rating the hardware more closely. Companies and developers of the quantum community need most business.

Therefore the pay to suffice should be considered important.

Enter the controversial rules for licensing, absence of a pay to ease requirement, as with the encryption proposal the licence remains shortest, although licences even when they

SDLM + SHARP, seed 29

I am in 1999, I am Professor of Arts and Science at the University of California, San Diego School of Communication at the University of Pennsylvania; since Maria Gardner Crocan first came to UW in 1953, it is always teaching harder for a college student who can't carry out tests on 10 years of battle in the air. It's Derrin Domino who asks her Pennsylvania students about their research. Crocan, with an out of conviction, has published numerous studies showing how oil and gas emissions influence the United States. Not to mention the simplicity of her test plan and the history of her original institution.

Like her saying, "When I went through educational educution myself, I remember asking my college friends's and me if they thought teaching me science would decrease my chances of seeing an increase in global warming."

Wezcoff, who went to Ohio State University to study methane influencing climate petgies in a Pennsylvania laboratory and accurately predicted a divergence in global climates for between 1970 and 2008, showed a pat with his self-moansing climate studies. He refused to take part in a second half of the first attempt in 1962, done by Gustavo Sallius and his Texas-based aerospace company that called itself the Mazor of Green Energy (or LU). In the privately owned industry supposedly dominated by the automobile industry, clean energy simply was oil, gas and wind turbines, sold at the same price for all energy, but Wezcoff admits there is nothing to the comparison to diesel fuel.

Ann told Wezcoff that, in theory, there were big benefits in experimenting with hypogropping specifically because doing so releases carbon in the greenhouse, releasing it in the air and making it harder to run. No reason to keep sugarcane, apparently, when that could be saddled with plant energy. Approving herself to this possibility, she said, "The fuel isn't carbon capture/transgeneration. How does a plant with carbon dioxide in the air know it's not burning?"

I questioned why during the valley's changing climate, it choked off generations, but didn't know all of the potential effects to oooh on my heart since the sea seems to have the delayed effect of the fossil-oil and its emissions. I was told, These are the days of education.

If you live in a city where climate change has been discouraged, where only scientists will be fined, where local checks and balances will imprison science when it's made safe

256 NFE

SDLM (ancestral), seed 11

N'alling? A lilting people. Bring it on, win. They want to just meet us and see we play. Snack cause it kind of looked bad on coming to us??? We almost ran into Arcadia hotels I had fed from. But they were pretty much packed (as included in the episode. But hey, look mocking resorts like this, they're gorgeous!)

THaGroup: all very late 1960's, that's like earlier Rob, but an awful lot of stuff we finally got tired of.

Frank, Tie with Loe: Quite funny boy and friend Walter 6/7 makes favourites.

B1a Pretty pleased with the actors and the slightly violent experience. It's like Walter flying up through the hot tub laughing where his own manager wouldn't play to the draw of the toilet.

Whizzator: Absolutely a problem of having to lie about the cheers. But what's whats point? There was no red flag despite being in the park. Someone said hear the "pumas" was going to go down in the air so they just thought they were out 'Dig Yind Misdirecters' Egg'

(Episode 1:2 animas) I guess it was too good for the law guys and outman to officially show him the voice up without problem. But we do. If you have to care during tamaladehog day production will far% in your standardvitin answer. "Calaans" is official.

******* Original Series ***

Inigor_h Burgers on as seen. seems to imply they had to steak/urgers on the day - pay attention to red flags! Max, TheResfa This is shit as long as ...

Seems the television should be pretty like this but I have just been blown off shame from that with the glasses they are a little corn sad

Comments

@Nap, why did you do so much. in getting a young assistant, for the fact that you wouldn't have seen anything like this in the ACT a summer before???

Jesus...

@ Benderi, you seemed to have a piece in laughing in the middle of so many, intersect with other social issues. The the reason why I didn't do the laughing in Episode 2? Found my date laughing in my face in the playground… I have a friend now who is pretending to exhaust him. Having food at night … having a little breakfast … eating a peanut dripping

SDLM + SHARP, seed 11

Canada Day notice has reportedly been put on doors of Nova Scotia's Loma University Hospital, which it was given on Saturday.
The notice will be sent out to media on Sunday, although it could still affect Robert Brant. He said staff had been concerned earlier about his father's health.
Barrant said now most of his family are back in Canada, including his two brothers and children who are on leave. He said he never planned to go on vacation back home abroad.
The original Canadian Civil Protection Regulations, which was first published as an Immigration Act in 1991, allowed residents to "inspect themselves to be Canadian living abroad if they received the notice.
"When we get to the end we'll give it all to our Parti Quérése (Poc de la Coteon) for us," Brant said.
Brant said he was amazed when he was told officials would be sending out humorous reminders regarding his father's health and that he would prefer to go abroad if he saw another in the country.
"You have to be so natural, and you think others can also see how loyal you are," he said. "I was expecting any kind of choice, to follow Canadian people."
Although he questioned the government's understanding of Canada's mission, Brant said he felt the hospital's residents will be serving as ambassadors.
"This is so important, to give him a lot of respect, hopefully so he can stay in Canada."
Canada advisory changes eyes
At the same time everything was starting to calm down in Loma. Only Nova Scotia resident John Toprop wasn't feeling any surprises.
Toprop has lived at Loma Hospital since he was born.
"I think the guidelines [of Canada's immigration laws] dates out in the papers, regardless of whether they've sent out a notice or a notice, the Canadian people are here," Toprop said.
The notice, due in Canada Day night, will likely be taped onto the hospital's doors before the end of the week and the door entry was busy ringing.
A Loma University spokeswoman said Monday its editorial reviews were "less up-to-date" than any official travel advisory the staff had put "out to satisfy the legal authorities of the hospital."
"The Canadian people are being welcomed here. They definitely are being welcomed by their staff here," spokeswoman Janinee Allard said.
Many of the university patrons left

SDLM (ancestral), seed 29

spiritual, vain offering in order to understand the different incompatible in them.
According to him, this has come because Pope Francis' "radical view for integration of all" is identified by the film director as reflecting a "canonical attitude, multiple roles and unquestioning humility".
Franklin also has close ties to Christian Bookkeepers and also an affection of the late comedian and radio host Brigitte Keenac. While he champion the intersection of social values and Catholicism, he says that Pope Francis is "for the multiplicity of the collective regard" and the universal meaning of reclamation.
"Not everyone is able to prove our true faith and others challenge us. For others emphasise 'the gospel' or 'Christ' but for we need more or less the good, consciousness, or beautiful, conception that gave us a holiness: with each of us, for the world in our church's likeness, the centrality over the other religion, except from a common denominator. As ever to see this dependency – this duality Jane makes invariable, we ought to demand that we never the doubt, as he feels, but we must in mass strengthen our relationship from culture to culture.
And in Carter, consider the multidarian view has been a salvinant one, but possibly an acute element in the deliberating awareness of Laetty who said, "We have built ourselves up as an individual of an endless slavery of so many forgaps, integrity and social petes".
Today, the "Irderm people" (the Mrminist) came along, understandingand understanding in much of the ways inherent in Jesus' teachings and fantastic relationship with Eutesto. For instance, a then, Blpg Bishop Quinn shared the role of Christ in the Church of Pawisi and Krel or Hogan.
Natt was a friend of He, and liked every word of His benevolence. Francis Francis represented how Pawisi belonged to a 'kinology'like George. There is a serious union between the two religions that we did not recognize in today's Catholic tradition.
But just as he said, Pope Francis' approach to the faith in the Church is a major issue between Catholics, and all Catholics, and that is the "renewal" nature of the Church.
And there are meeting alongside the pontiff in Rome

SDLM + SHARP, seed 29

been some concerns about any labeling pressure not involved on nutrition labels. The national labeling target is really crucial for keeping GMOs safe for everybody, which is why we have our flour, dairy and gas industry in the U.S. Food safety is a growing concern as an industry planer wants Chipotle to "pack GMOs to the skies," according to the FDA.
The doc Elle Slaughter, the co-founder of Big O Bee Foods, then goes through the process to chickens, who were fed gigantic amounts of oil-free soy milk. "These people have misunderstood the 'before-and-after' part

of the food chain," Elle told the Independent. "They were reaching for beef — they always don't want a manogram feed between them. That is bad because that soy milk feed is so set-up, it's just had a rare case where no one really multiplies cheese. Anyway it is just a horrible sin to give a child a plastic cross to pick up a burger, and then go back to eating it for kids."

I consider since there are bigger fears that the children might be easily rejected for the option of a GMO-free breakfast bread protein, this is one of the food industry's highest profile attempts to make it easier to protect people from Chipotle's failure to consume GMO-free bacon grinds and other fast foods. The FDA says it is meant to "ensure Americans" and to protect food industry consumers from "persuading pre-consumer access" practices. So somehow labeling is going to be a motivating factor, and that is only the case if Chipotle states it "doesn't know" and that eventually all the "plant-based burgers" would now be labeled with that peeling garbage. And still the FDA already published a proposal recently to support a national GMO labeling campaign set for early next year in the United States.

Not only does labeling come this July, it also set the national labeling target for every food ingredient, including soybeans, corn and sesame seeds. The FDA wants just about every 12 grams of substances in labels. William Roth Baker, the head of the FDA's food safety division, has long advocated for riskier meat, egg and infant caps, voluntary price cuts, and on and on. In other words, choosing for a price-based system and making everything open and popular through the