MODEL COLLAPSE IN CLOSED FORM: EXACT GAUSSIAN LAWS AND THEIR GPT-2 SIGNATURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Models increasingly learn from data generated by other models, yet the mechanism of the resulting *model collapse* is contested—mean drift or vanishing tails?—and mitigations such as data mixing are designed without an exact law to target. We answer exactly in the Gaussian testbed that both sides of the debate use as their reference model. Under recursive maximum-likelihood training, mean drift and variance contraction contribute *equally* to the forward KL—the objective that maximum likelihood actually optimizes—doubling the collapse rate; the quadratic cliff that remains is exactly half the blind spot of the plug-in variance estimator, while the reverse KL cancels it entirely, a direction asymmetry between the two divergences. Mixing real data into each generation admits an exact steady state; any schedule whose mixing fraction decays to zero diverges; and constant mixing is the best non-increasing schedule at matched real-data budget, and the only one that bounds every generation. In recursive GPT-2 training these laws appear as token-level signatures: a forward/reverse ratio near two at the first generation, sustained thereafter while both divergences saturate; extinction concentrated in a frequency band above the plug-in window; and mixing schedules that set the level of the degradation plateau—with a decaying schedule that looks safe at short horizons becoming the worst choice at triple the horizon.[1]

1 INTRODUCTION

The internet is being rewritten by generative AI. As AI-generated content accumulates online, future models will inevitably train, at least in part, on data produced by their predecessors: models generate data, data trains models, and the cycle repeats. The question is not whether this recursive training happens, but whether it is sustainable.

The emerging answer is troubling. Recursive generative training leads to *model collapse*: a systematic degradation in quality and diversity. Shumailov et al. (2024) showed that language models trained on their own outputs progressively lose low-probability tokens; Alemohammad et al. (2024) documented the same in image generation.

Yet a fundamental question remains unanswered: *what is the dominant mechanism driving collapse?* On one side, Shumailov et al. (2024) emphasized the disappearance of low-probability events, and Suresh et al. (2024) proved a high-probability tail bound showing that the estimated variance contracts to zero almost surely under recursive training with the bias-corrected estimator, pointing to *variance loss*. On the other, Schaeffer et al. (2025) argued that in Gaussian models, *mean drift*, not tail loss, is the true driver. Existing mitigation strategies—data mixing (Fu et al., 2024; Gerstgrasser et al., 2024) and reweighting (He et al., 2026)—address the general phenomenon without knowing which mechanism to target.

We give this debate an exact answer in the Gaussian testbed that both sides use as their reference model, by decomposing the KL divergence exactly. The answer: *neither mechanism dominates; to leading order in* $1/n$ *they contribute equally*. Under the forward KL $D_{\mathrm{KL}}(P_0 \| P_t)$—the mode-covering divergence that is the population objective of maximum-likelihood training—when vari-

[1] Code available at `https://github.com/anon-research/recursive-collapse`.

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.

ance is estimated alongside the mean, mean drift and variance mismatch each contribute $td/(2n)$, yielding a total rate of td/n, *twice* the known-variance baseline. Each side of the debate is right under its own metric: the Wasserstein distance favours mean drift (as Schaeffer et al., 2025 observe), while the forward KL weights the two equally. The practical implication is that strategies targeting only one mechanism tackle only half the problem. Variance contraction deserves particular attention not because it dominates, but because it is the *silent* killer: it operates through loss of diversity rather than detectable distributional shift, and it invalidates standard training metrics as quality indicators (Section 5). Each generation shrinks the estimated variance by a factor $(n-1)/n$, and the forward KL exposes a strictly positive $+3t^2d/n^2$ *quadratic cliff* on top of the doubling: collapse accelerates as the model loses its ability to cover the true tails.

The natural mitigation is mixing real data into each generation. We derive the steady-state variance $\mathrm{Var}[\mu_\infty] = \sigma^2/(n\alpha(2-\alpha))$, prove that *any* schedule whose mixing fraction decays to zero ($\alpha_t \to 0$; power laws $\alpha_t = t^{-p}$ with closed-form rates) fails to prevent variance divergence, and show at matched total real-data exposure $\Sigma\alpha$ that constant mixing is the best non-increasing schedule and the only one that bounds every generation of the trajectory. We validate the laws in recursive GPT-2 training: retraining on its own generated token streams echoes the predicted direction asymmetry, concentrates token extinction in a frequency band, and exhibits the schedule-set plateau levels anticipated by the steady-state analysis.

Contributions. Three exact laws, validated in recursive GPT-2 training:

1. **The forward-KL law: equal contributions, doubling, cliff, and halving** (Theorem 2, Proposition 1): the forward KL decomposes into equal mean-drift and variance-contraction contributions, doubling the rate from $td/(2n)$ to td/n and adding a strictly positive $+3t^2d/n^2$ quadratic cliff; exactly half the cliff is the blind spot of the plug-in variance estimator, so debiasing the estimator halves it.
2. **Reverse-KL cancellation and direction asymmetry** (Corollary 2, Corollary 1): under the reverse KL the mean-drift and variance terms cancel exactly, leaving a linear rate with t-independent doubling ratio $2 + 11/(6n) + O(n^{-2})$; the forward ratio is $2 + 25/(6n) + O(n^{-2})$ at leading order in t and grows with t.
3. **Mixing laws** (Theorem 3, Corollary 3): constant mixing $\alpha > 0$ arrests mean drift at the exact steady state $\sigma^2/(n\alpha(2-\alpha))$; any schedule with $\alpha_t \to 0$ diverges; and constant mixing is the best non-increasing schedule at matched budget $\Sigma\alpha$, the only one bounding every generation.

A categorical extension—extinction undergoes a sharp Gumbel threshold transition with renormalization protecting survivors (Proposition 2, Appendix A)—supports the extinction analysis in the GPT-2 experiments. **Validation:** recursive GPT-2 training reproduces the laws at the token level as signatures—a first-generation forward/reverse ratio near two, sustained asymmetry, schedule-set plateau levels—and shows that training loss misranks the schedules (Section 5).

2 Problem Formulation

Definition 1 (Recursive Generative Training)**.** Let P_0 be the true data distribution over $\mathcal{X} \subseteq \mathbb{R}^d$, where d is the dimension of the distribution. A *recursive generative training* process with sample size n produces a sequence of distributions $\{P_t\}_{t\geq 0}$: at each generation t, draw n i.i.d. samples from P_t, fit $\hat{P}_{t+1}$ via MLE, and set $P_{t+1} = \hat{P}_{t+1}$.

Definition 2 (Mixed Recursive Training)**.** With mixing fraction $\alpha \in [0, 1]$, at generation t draw $n\alpha$ samples from P_0 (real) and $n(1-\alpha)$ from P_t (synthetic), then fit $\hat{P}_{t+1}$ to the combined dataset. When α_t varies with t, we call this a *decaying mixing schedule*. The *total real data exposure* is $\Sigma\alpha = \sum_{s=1}^{T} \alpha_s$, which quantifies the cumulative amount of real data the model sees over T generations.

We study the Gaussian model $P_0 = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ as a tractable setting capturing the essential dynamics, the standard starting point for recursive training analysis (Suresh et al., 2024; Kanabar & Gastpar, 2025). The data-processing chain true-data $\to P_1 \to \cdots \to P_t$ can only lose fidelity to P_0, so some degradation is inevitable without intervention; the question is how fast, in which divergence, and through which mechanism.

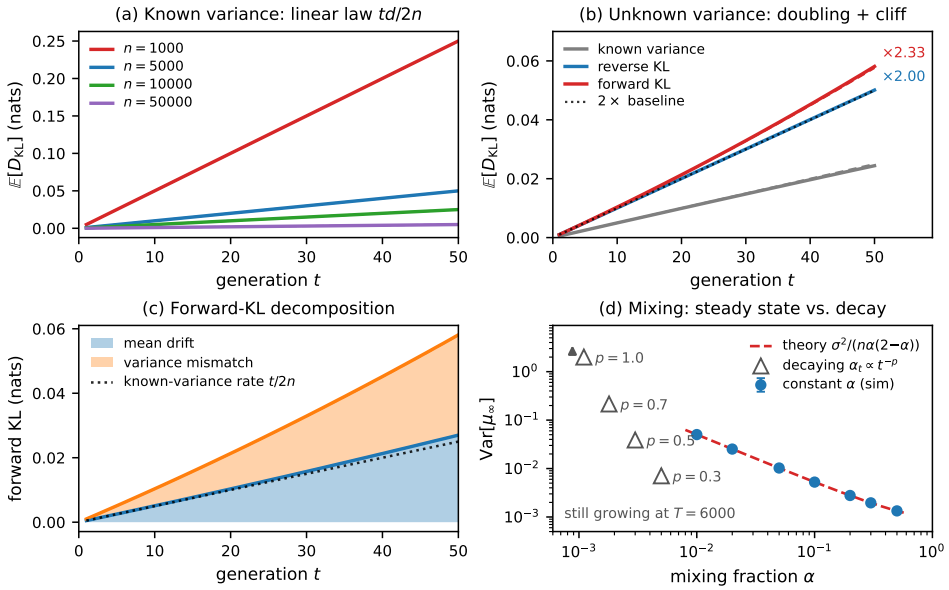


Figure 1: Gaussian testbed verification of the exact laws (20 seeds; simulation solid with shaded 95% CI, exact theory dashed). (a) Known variance: the linear law $td/(2n)$ at every t and n (d=10). (b) Unknown variance (n=1000): both directions at least double the baseline (dotted); the forward KL adds a quadratic cliff (endpoint ratios vs. the exact baseline annotated), the reverse KL does not. (c) The forward KL decomposes into mean-drift and variance-mismatch contributions of nearly equal size, each comparable to the entire known-variance loss (dotted). (d) Constant mixing reaches the exact steady state $\sigma^2/(n\alpha(2-\alpha))$ for $\alpha \in [0.01, 0.5]$, while every decaying schedule $\alpha_t \propto t^{-p}$ still grows at T=6000 (open markers). Per-seed traces: Appendix D.

The key distinction is known vs. unknown variance. In all practical models, variance is estimated from data. The single-generation bias of the MLE is mild—$\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$—but under recursion it compounds into exponential contraction $\mathbb{E}[\hat{\sigma}_t^2] = \sigma_0^2((n-1)/n)^t$: invisible if one tracks only the mean, yet accounting for half the total information loss.

3 Variance Collapse Theory

3.1 Known Variance: Baseline Rate

Theorem 1 (Gaussian Mean Drift, Known Variance)**.** *For $P_0 = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ with known covariance $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{d\times d}$, where d is the dimension of the distribution, under recursive MLE with n samples per generation:*

$$\boldsymbol{\mu}_t - \boldsymbol{\mu}_0 \sim \mathcal{N}\left(\mathbf{0}, \frac{t\boldsymbol{\Sigma}_0}{n}\right), \quad \mathbb{E}[D_{\mathrm{KL}}(P_0\|P_t)] = \mathbb{E}[D_{\mathrm{KL}}(P_t\|P_0)] = \frac{td}{2n}. \tag{1}$$

For known-variance Gaussians the forward KL $D_{\mathrm{KL}}(P_0\|P_t)$ and the reverse KL $D_{\mathrm{KL}}(P_t\|P_0)$ are equal by symmetry, so the direction is moot here; it becomes material once the variance is estimated (Theorem 2 vs. Corollary 2).

Proof sketch. The MLE $\hat{\boldsymbol{\mu}}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_0/n)$ yields a random walk with independent steps (fresh samples each generation); the trace trick gives $td/(2n)$. Full proof in Appendix B.1. □

The linear law $\mathbb{E}[D_{\mathrm{KL}}] = td/(2n)$ is the fundamental unit of information loss: each generation contributes an additive $d/(2n)$, verified exactly in simulation (Figure 1(a)). The rate *doubles* when the variance must be estimated—and the forward KL then grows a quadratic-in-t cliff.

3.2 UNKNOWN VARIANCE: THE CATASTROPHIC QUADRATIC CLIFF

In practice, variance must be estimated from the same data used to estimate the mean. Under the *forward* KL $D_{\mathrm{KL}}(P_0\|P_t)$—the mode-covering direction that maximum likelihood actually optimises—this seemingly minor change unleashes a catastrophic quadratic-in-t cliff on top of the doubling already present at leading order.

Theorem 2 (Forward-KL Collapse under Unknown Variance)**.** *For a d-dimensional Gaussian with diagonal covariance and unknown parameters, using the (biased) MLE estimators $\hat{\mu}_t = \bar{X}_t$ and $\hat{\sigma}_t^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)} - \bar{X}_t)^2$ per dimension, for $n > 3$:*

1. ***Variance contraction:*** *$\mathbb{E}[\hat{\sigma}_t^2] = \sigma_0^2 \cdot ((n-1)/n)^t$ (a one-line chi-squared computation; cf. Remark 1).*

2. ***(per dimension)*** *Defining $V_t = \hat{\sigma}_t^2/\sigma_0^2$ and $I_t = \mathbb{E}[1/V_t] = \left(\frac{n}{n-3}\right)^t$, the forward KL per dimension decomposes as*

$$\mathbb{E}[D_{\mathrm{KL}}(P_0\|P_t)] = \underbrace{\tfrac{1}{6}(I_t - 1)}_{\text{mean drift}} + \underbrace{\tfrac{1}{2}(I_t - 1) + \tfrac{1}{2}\mathbb{E}[\log V_t]}_{\text{variance mismatch}} = \tfrac{2}{3}(I_t - 1) + \tfrac{1}{2}\mathbb{E}[\log V_t]. \tag{2}$$

3. ***(per dimension)*** *The asymptotic expansion in $1/n$ is*

$$\mathbb{E}[D_{\mathrm{KL}}(P_0\|P_t)] = \frac{t}{n} + \frac{3t^2 + \frac{25}{12}t}{n^2} + O\left(\frac{t^3}{n^3}\right), \tag{3}$$

extended to d dimensions by multiplying the right-hand side by d. The leading term t/n is ***twice*** *the known-variance rate $t/(2n)$; the strictly positive $+3t^2/n^2$ correction is the* quadratic cliff *that compounds collapse faster than linearly in t.*

Proof sketch. By Cochran's theorem $\hat{\sigma}_t^2 \mid \hat{\sigma}_{t-1}^2 \sim \hat{\sigma}_{t-1}^2\,\chi_{n-1}^2/n$, and iterated expectation gives Part 1. For Part 2, writing $V_t = \prod_{s\le t} W_s/n$ with $W_s \sim \chi_{n-1}^2$ i.i.d., the mean drift telescopes via $\mathbb{E}[V_s/V_t] = (n/(n-3))^{t-s}$ and the geometric sum $\sum_{s<t}(n/(n-3))^{t-s} = \frac{n}{3}(I_t - 1)$, giving $\frac{1}{6}(I_t - 1)$; the variance mismatch contributes $\frac{1}{2}(I_t - 1) + \frac{1}{2}\mathbb{E}[\log V_t]$. Part 3 expands $(1 - 3/n)^{-t}$ and $\mathbb{E}[\log V_t] = t[\psi((n-1)/2) + \log(2/n)]$ in powers of $1/n$. Full proof in Appendix B.2. □

Remark 1 (Attribution, and the leading-order doubling is exact). Suresh et al. (2024) analyzed the bias-corrected estimator (a martingale: no contraction in expectation), proving almost-sure collapse via an exponential tail bound; the biased-MLE expected-value contraction in Part 1 is a one-line chi-squared moment, but its *consequences*—the closed-form decomposition with the cliff, the exact halving (Proposition 1), the reverse-KL cancellation (Corollary 2), and the exact doubling ratios (Corollary 1)—are, to our knowledge, new. The doubling is an n-asymptote, not an exact identity: the exact sum Eq. equation 2 holds at every n, with mean drift and variance mismatch each contributing $t/(2n)$ at leading order (Figure 1(c)); the $+3t^2/n^2$ cliff is the leading finite-n correction, and it is what the reverse-KL analysis *hides* via exact cancellation.

Proposition 1 (The cliff is exactly halved under the unbiased estimator)**.** *Under the bias-corrected estimator $\hat{\sigma}^2_{\mathrm{MLu}} = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ (the setting of Suresh et al., 2024) in place of the MLE, with $V_t^{\mathrm{u}} = \hat{\sigma}^2_{\mathrm{MLu},t}/\sigma_0^2$ and $I_t^{\mathrm{u}} = \mathbb{E}[1/V_t^{\mathrm{u}}] = \left(\frac{n-1}{n-3}\right)^t$, the forward KL obeys, for $n > 3$, the exact closed form*

$$\mathbb{E}\big[D_{\mathrm{KL}}^{\mathrm{MLu}}(P_0\|P_t)\big] = \frac{3n-1}{4n}\left(I_t^{\mathrm{u}} - 1\right) + \frac{1}{2}\,\mathbb{E}[\log V_t^{\mathrm{u}}], \qquad \mathbb{E}[\log V_t^{\mathrm{u}}] = t\left[\psi\left(\frac{n-1}{2}\right) + \log\frac{2}{n-1}\right], \tag{4}$$

with asymptotic expansion $\frac{t}{n} + \frac{3}{2}\frac{t^2}{n^2} + \frac{11}{6}\frac{t}{n^2} + O(t^3/n^3)$: the leading rate t/n is unchanged, and the quadratic cliff coefficient is exactly halved, *$3t^2/n^2 \to \frac{3}{2}\,t^2/n^2$.*

Proof sketch. The mean drift reuses the telescoping identity with $W_s/(n-1)$ in place of W_s/n, giving $\mathbb{E}[\text{mean drift}] = \frac{n-1}{4n}(I_t^{\mathrm{u}} - 1)$; the variance mismatch contributes $\frac{1}{2}(I_t^{\mathrm{u}} - 1) + \frac{1}{2}\mathbb{E}[\log V_t^{\mathrm{u}}]$, and $\frac{n-1}{4n} + \frac{1}{2} = \frac{3n-1}{4n}$. Expanding $I_t^{\mathrm{u}} - 1 = \frac{2t}{n} + \frac{2t^2+4t}{n^2} + O(t^2/n^3)$ and $\mathbb{E}[\log V_t^{\mathrm{u}}] = -\frac{t}{n} - \frac{4t}{3n^2} + O(t/n^3)$ then yields the stated expansion. Full proof in Appendix B.3. □

Remark 2 (Interpretation: how much of the cliff is estimator bias?). Proposition 1 gives an exact causal answer to the question *is the quadratic cliff an estimator-bias artifact?* Exactly half of the cliff coefficient ($\frac{3}{2}t^2/n^2$) is attributable to the expected-value contraction of the biased MLE and disappears under bias correction; the other half survives the correction and is the pure tail-collapse (Jensen-gap) mechanism that the almost-sure collapse bound of Suresh et al. (2024) already guarantees. Our 18-seed simulation measures a 49.1% reduction of the cliff at n=1000, t=50 (doubling ratio $2.326 \to 2.166$, positive in $18/18$ seeds), against the 50% predicted; the closed form Eq. equation 4 evaluates to 0.05398, matching the measured 0.05413 and an independent 2×10^5-trial Monte Carlo (0.05388 ± 0.00026). The cliff is therefore *half estimator bias, half intrinsic mechanism*, and the intrinsic half is unavoidable under any estimator subject to almost-sure collapse (ablation: Figure 4, Appendix D).

Corollary 1 (Exact leading-order doubling ratios)**.** *For $n > 3$, write $\mathbb{E}[D_{\mathrm{KL}}(P_0\|P_t)] = c\,t + O(t^2)$ for the term linear in t (with the natural continuous extension in t). Then*

$$\frac{c_{\mathrm{fwd}}}{c_{\mathrm{known}}} = 2n\Big[\tfrac{2}{3}\log\tfrac{n}{n-3} + \tfrac{1}{2}\big(\psi\big(\tfrac{n-1}{2}\big) + \log\tfrac{2}{n}\big)\Big] = 2 + \frac{25}{6n} + O(n^{-2}), \tag{5}$$

whereas the reverse-KL doubling ratio is t-independent at $2 + \frac{11}{6n} + O(n^{-2})$ (Corollary 2). Both ratios tend to 2 as $n \to \infty$ but differ at order $1/n$ ($25/6$ vs. $11/6$): even at $t \to 0$ the forward KL is slightly more severe, and at finite t the forward ratio is $2 + (6t + 25/6)/n + O(n^{-2})$ (from Eq. equation 3).

Proof. Writing Eq. equation 2 as $\frac{2}{3}(e^{tL} - 1) + \frac{t}{2}\Delta$ with $L = \log\frac{n}{n-3}$ and $\Delta = \psi(\frac{n-1}{2}) + \log\frac{2}{n}$, the t-linear coefficient is $\frac{2}{3}L + \frac{1}{2}\Delta$, and dividing by $c_{\mathrm{known}} = 1/(2n)$ gives Eq. equation 5; expanding L and Δ in $1/n$ yields $2 + \frac{25}{6n} + O(n^{-2})$—both the inverse-moment and the logarithmic term contribute at order $1/n$. (The reverse KL is exactly $-\frac{t}{2}\Delta$, so its ratio is $-n\Delta$ for all t.) Full derivation, including the expansion-order pitfall, in Appendix B.2. □

Verification. Across $n \in \{20, 100, 1000, 5000, 10000, 50000\}$, $t \in \{1, \ldots, 49\}$ (20 seeds; Appendix C, Tables 2–4), the forward-KL doubling ratio grows from 2.02 at t=1 to 2.33 at t=50 (n=1000; Figure 1(b))—the signature of the quadratic cliff—while the reverse ratio stays at its t-independent 2.00 ± 0.01, and the exact formula Eq. equation 5 reproduces all six n values to the printed digits. The asymptotic Eq. equation 3 is accurate for $t^2 \ll n^2$ (within 0.7% of the exact closed form at t=49, n=1000); the exact closed form is the correct prediction for $t \sim n$.

3.3 Reverse-KL Corollary: The Linear Lower Bound

Theorem 2 shows that the forward KL carries a quadratic cliff. We now show that even under the more lenient, mode-seeking reverse KL $D_{\mathrm{KL}}(P_t\|P_0)$, which does not penalise the model for failing to cover the true tails, the collapse rate still doubles at leading order: the reverse KL is a *conservative lower bound* on collapse severity, and the two divergences bracket it from below (reverse) and above (forward).

Corollary 2 (Reverse-KL Doubling via Exact Linear Cancellation)**.** *Under the same setting as Theorem 2 (and in fact for all $n \geq 2$; $n > 3$ is needed only for the forward inverse moment), the reverse KL decomposes as*

$$\mathbb{E}[D_{\mathrm{KL}}(P_t\|P_0)] = \underbrace{\tfrac{1}{2}(1 - \mathbb{E}[V_t])}_{\textit{mean drift}} + \underbrace{\tfrac{1}{2}(-\mathbb{E}[\log V_t] + \mathbb{E}[V_t] - 1)}_{\textit{variance mismatch}} = -\tfrac{1}{2}\mathbb{E}[\log V_t], \tag{6}$$

where the $\mathbb{E}[V_t]$ terms cancel exactly*, yielding a quantity strictly linear in t:*

$$\mathbb{E}[D_{\mathrm{KL}}(P_t\|P_0)] = -\frac{t}{2}\left[\psi\left(\frac{n-1}{2}\right) + \log\left(\frac{2}{n}\right)\right] = \frac{t}{n} + \frac{11t}{12n^2} + O\left(\frac{t}{n^3}\right). \tag{7}$$

The reverse-KL doubling ratio is therefore t-independent: $\mathbb{E}[D_{\mathrm{KL},\textit{reverse}}]/\mathbb{E}[D_{\mathrm{KL},\textit{known}}] = -n\left[\psi\left(\frac{n-1}{2}\right) + \log\left(\frac{2}{n}\right)\right] = 2 + \frac{11}{6n} + O(n^{-2})$ for all $t \geq 1$.

Proof sketch. The reverse KL is $D_{\mathrm{KL}}(P_t\|P_0) = (\hat{\mu}_t - \mu_0)^2/(2\sigma_0^2) + \frac{1}{2}[V_t - 1 - \log V_t]$: the mean-drift denominator is the *true* variance (not $\hat{\sigma}_t^2$) and the variance ratio appears as V_t (not $1/V_t$).

Unrolling the mean random walk gives $\mathbb{E}[(\hat{\mu}_t-\mu_0)^2] = \sigma_0^2(1-\mathbb{E}[V_t])$, so the mean drift $\frac{1}{2}(1-\mathbb{E}[V_t])$ and the variance mismatch $\frac{1}{2}(\mathbb{E}[V_t] - 1 - \mathbb{E}[\log V_t])$ cancel *exactly* in $\mathbb{E}[V_t]$, leaving $-\frac{1}{2}\mathbb{E}[\log V_t]$, strictly linear in t by the product representation of V_t. Full proof in Appendix B.4. □

The cancellation is a consequence of the reverse KL's leniency toward variance collapse (the contracted model places its mass inside the true support), yet even under this lenient divergence the rate still doubles; the forward KL (Theorem 2) exposes the additional $+3t^2/n^2$ cliff that this direction hides.

Remark 3 (The linearity is robust to the estimator choice). Under the bias-corrected (unbiased) estimator of Proposition 1, the reverse KL remains *exactly linear* in t with the same leading rate: the martingale property $\mathbb{E}[V_s^{\mathrm{u}}] = 1$ makes the mean drift exactly $t/(2n)$ and the variance mismatch reduces to $-\frac{1}{2}\mathbb{E}[\log V_t^{\mathrm{u}}]$, so $\mathbb{E}[D_{\mathrm{KL}}^{\mathrm{MLu}}(P_t\|P_0)] = \frac{t}{2n} - \frac{1}{2}\mathbb{E}[\log V_t^{\mathrm{u}}] = \frac{t}{n} + O(t/n^2)$, matching a 2×10^5-trial Monte Carlo at n=1000, t=50 (0.05003 ± 0.00023). The direction asymmetry of the cliff is therefore structural to the KL geometry, not an artifact of estimator bias: the forward-KL cliff is present under both estimators (halved but nonzero under the unbiased one), while the reverse KL is exactly linear under both.

The asymmetry is directly visible in simulation: the reverse ratio is t-independent at each n (2.002 ± 0.003 at n=1000 across $t \in \{1, 10, 49\}$; Tables 3–4, Appendix C), while the forward ratio grows with t (Table 2). At small n the contrast is dramatic: the exact forward/reverse prediction at (n, t)=$(100, 49)$ is 7.37 vs. 2.04, with $> 100\%$ per-seed spread in the forward cells (the heavy tail of $1/V_t$).

4 Mixing Strategies and Steady-State Analysis

4.1 Steady-State under Constant Mixing

Theorem 3 (Steady-State Mixing, known-variance setting)**.** *Under mixed recursive training with constant $\alpha > 0$, and with the synthetic-stream variance evaluated at the true σ^2 (the known-variance setting; see Remark 4), the mean converges to a steady-state distribution with:*

$$\mathrm{Var}[\mu_\infty] = \frac{\sigma^2}{n\alpha(2-\alpha)} \tag{8}$$

For small α: $\mathrm{Var}[\mu_\infty] \approx \sigma^2/(2n\alpha)$.

Proof sketch. The MLE from mixed data gives $\hat{\mu}_t = \alpha\bar{X}_{\text{real}} + (1-\alpha)\bar{X}_{\text{synth}}$, with $\mathbb{E}[\hat{\mu}_t|\mu_{t-1}] = \alpha\mu_0+(1-\alpha)\mu_{t-1}$ and, remarkably, $\mathrm{Var}[\hat{\mu}_t|\mu_{t-1}] = \sigma^2/n$ regardless of α (the sample-concentration and sample-count effects cancel exactly). The law of total variance yields $v_t = (1-\alpha)^2 v_{t-1}+\sigma^2/n$; solving the linear recursion gives Eq. equation 8. Full proof in Appendix B.6. □

Remark 4 (Scope of the mixing laws). Theorems in this section are stated in the known-variance setting: the synthetic-stream contribution to the recursion uses the true σ^2. Under unknown variance the variance channel changes qualitatively, and we leave that extension open; the law Eq. equation 8 remains the correct leading-order design target for the mean.

Even $\alpha = 0.01$ yields $\mathrm{Var}[\mu_\infty] \approx 50\sigma^2/n$, and any $\alpha > 0$ arrests the growth: pure synthetic training ($\alpha = 0$) leaves the variance growing without bound (linearly in t), so mixing buys finiteness at arbitrarily small cost. The improvement is sublinear in α (doubling α halves the variance), motivating the analysis of decaying schedules that use less total real data. The law is exact in simulation across $\alpha \in [0.01, 0.5]$ (Figure 1(d)).

4.2 Decaying Mixing Inevitably Fails

Maintaining constant α requires ongoing access to real data. Can the mixing fraction *decrease* over time while still preventing collapse? No.

Corollary 3 (Decaying Mixing Inevitably Fails)**.** *If $\alpha_t = t^{-p}$ for* any $p > 0$*, then* $\mathrm{Var}[\mu_t] \to \infty$. *Specifically,* $\mathrm{Var}[\mu_t] \sim (\sigma^2/2n) \cdot t^{\min(p,1)}$.

More generally, *any* schedule with $\alpha_t \to 0$ diverges: for any $\varepsilon > 0$, all sufficiently late generations have $\alpha_t \leq \varepsilon$, and the unrolled recursion then obeys $v_T \geq (\sigma^2/n)\,(1-e^{-1})/(2\varepsilon)$ eventually; power laws are simply the family with closed-form rates (Appendix B.6).

Proof sketch. The variance recursion becomes $v_t \approx (1 - 2t^{-p})v_{t-1} + \sigma^2/n$. The fundamental obstacle is constant per-generation noise σ^2/n versus vanishing contraction $(1-\alpha_t)^2 \to 1$: summing the unrolled recursion, the effective memory horizon grows as t^p ($0 < p < 1$) or t ($p \geq 1$). Full proof in Appendix B.6. □

Remark 5 (Almost-sure convergence vs. bounded variance). Although $\mathrm{Var}[\mu_t] \to \infty$ for all $p > 0$, μ_t converges almost surely to μ_0 for $1/2 < p \leq 1$ by the Robbins–Monro theorem (for $p > 1$ the cumulative exposure is finite and the mean does not return to μ_0). The mean converges, but distributional quality degrades: the noise is constant (σ^2/n), not scaling with α_t^2 as in standard stochastic approximation. Convergence of the estimator and boundedness of the distribution are different guarantees, and only constant mixing delivers both.

4.3 Constant Mixing Is the Best Non-Increasing Schedule at Matched Budget

Raw comparisons between decaying and constant schedules conflate schedule shape with real-data budget: $\alpha_t = 0.5/t^{0.3}$ spends $\Sigma\alpha \approx 2.74$ over eight generations against 0.80 for constant $\alpha = 0.10$, and it is the larger budget—not any advantage of the shape—that can drive better raw performance in uncontrolled comparisons.

At matched budget $\Sigma\alpha$ (constant comparator $\alpha_{\text{const}} = \Sigma\alpha/T$), the constant schedule achieves lower variance than every decaying schedule tested (Table 1, Appendix C; Figure 5, Appendix D): the recursion $v_t = (1-\alpha_t)^2 v_{t-1} + \sigma^2/n$ demands contraction at *every* step to combat the constant noise injection, and a decaying schedule front-loads its contraction, leaving later generations increasingly vulnerable.

Two qualifications keep the claim honest. Dominance is guaranteed within the family of *non-increasing* schedules: a fully *deferred* schedule—all real data spent in the final generation—can beat the constant schedule on the endpoint variance at equal budget (at $\Sigma\alpha = 2.743$, $T = 8$, deferred mixing gives $v_8 = 1.00\,\sigma^2/n$ versus 1.76 for constant), but it leaves generations $1, \ldots, T-1$ entirely unprotected, with variance growing linearly in the interim; the constant schedule is the only one that bounds *every* generation, which is the practically relevant guarantee when intermediate models are themselves deployed. Second, both sides use like-for-like finite-T closed forms, which preserves the ordering.

5 Experiments: Recursive GPT-2 Training

5.1 Setup

We validate the three laws in recursive GPT-2 training (small, 124M parameters; WikiText-2 corpus as the real-data pool P_0). Each generation trains on n documents sampled from the previous generation's model, optionally mixed with real documents at fraction α (Definition 2); the model then generates the next corpus. We sweep $n \in \{1000, 5000, 10000\}$ documents per generation over six schedules: no mixing, fixed $\alpha \in \{0.50, 0.10\}$, and decaying $\alpha_t = c_p(t+1)^{-p}$ for $p \in \{0.5, 1.0, 2.0\}$ with c_p normalised so the first eight generations spend the same real-data budget $\Sigma\alpha = 0.75$. Token-level histograms of every generated corpus yield forward and reverse KL between the P_0 token distribution and each generation's generated distribution, together with floor-free companions (Jensen–Shannon divergence, extinct mass, active vocabulary). Horizons: T=16 generations (10 complete seeds at n=1000, 9 at $n \in \{5000, 10000\}$) and a T=48 extension at n=1000 (4 seeds) that triples the horizon to test steady-state behaviour. Full protocol in Appendix E; aggregation uses complete seeds only.

5.2 Direction asymmetry at the token level

Figure 2(a–b) shows the result. The forward KL between P_0 and the generated token distribution grows to 6.8 nats by generation 16 (no mixing, n=1000), while the reverse KL grows to 3.7: the

mode-covering direction sees roughly *twice* the damage of the mode-seeking direction (per-seed forward/reverse ratio at generation 16: 2.04 ± 0.39). The asymmetry attenuates monotonically as the per-generation corpus grows—2.04 ± 0.39, 1.78 ± 0.22, 1.71 ± 0.21 at $n = 1000, 5000, 10000$ (9–10 seeds; Figure 2(d))—consistent with the cliff's $O(t/n)$ finite-sample structure: more data per generation, less unaccounted tail-loss. This is the token-level signature of Theorem 2 versus Corollary 2: the divergence that maximum likelihood actually optimises is the one that punishes variance collapse.

5.3 Extinction: thresholded, and harsher than plug-in

The active vocabulary falls from 3481 types in the pool to 43–80 by generation 48, and the extinct mass of P_0 (probability mass on tokens never generated) saturates at 0.72–0.81 (Figure 2(c)). Extinction is strongly frequency-resolved: at generation 16 (no mixing), tokens with P_0-frequency in $[10^{-3}, 10^{-2}]$ survive in 31% of seeds, tokens an order of magnitude rarer in 2%; by generation 48, 7% and 0.4%. The transition band is real and sharp, as Proposition 2 (Appendix A) predicts—but it sits well above the plug-in Gumbel window $\ln(t)/n$, with the gap growing in t, exactly the empirical-feedback hierarchy of the categorical testbed (Figure 3): retraining on own empirical token counts loses categories *faster* than the renormalization-aware law. Mixing protects precisely the transition band: fixed α=0.10 raises survival in $[10^{-3}, 10^{-2}]$ from 31% to 45% at generation 16 (paired extinct-mass reduction 0.047 ± 0.012, t=3.9, $9/10$ seeds), leaving the deep tail's fate essentially unchanged.

5.4 Mixing schedules set the plateau—and decaying ones have a second act

At the original T=16 horizon the six schedules are nearly indistinguishable: forward KL at generation 16 spans 6.37–6.82 nats with heavily overlapping 95% CIs (10 seeds). Tripling the horizon separates them (Figure 2(a); Table 5, Appendix C). Five schedules plateau: their generation-32-to-48 growth is statistically consistent with zero (fixed α=0.50: $+0.10 \pm 0.73$, t=0.1), settling at 7.3–7.6 nats—schedules set the *level* of the degradation plateau, not its speed. The slowest-decaying schedule is the exception: $\alpha_t \propto t^{-0.5}$ keeps degrading ($+0.69 \pm 0.09$ nats over generations 32–48, t=7.7, positive in $4/4$ seeds) and ends at 8.51 nats—significantly *worse than every other schedule including no mixing at all* (paired differences at generation 48: $+0.93 \pm 0.42$ vs. no mixing; $+1.23 \pm 0.37$ vs. fixed α=0.10). A decaying schedule that looked safe—second-best at generation 16—becomes the worst choice at generation 48: *short-horizon comparisons of mixing schedules are unreliable predictors of long-horizon safety*, which is the practical content of Corollary 3 for language models.

Model-level monitoring metrics do not resolve this. Final perplexity ranks the schedules differently from the distributional metrics (99 ± 3 for fixed α=0.50 down to 158 ± 8 for fixed α=0.10, but 3177 ± 1195 for decay p=2.0 and 2522 ± 569 for no mixing; Table 5): the fast-decaying p=2.0 schedule has the *worst* perplexity yet a mid-pack forward KL, and the two fixed schedules swap order between metrics. Training loss alone would misidentify which schedule is safe; the token-distribution laws are what the schedules actually obey.

5.5 Gaussian and categorical testbeds

The exact laws themselves are verified to Monte Carlo precision in the Gaussian testbed (20 seeds; Figure 1): known-variance linearity, the forward-KL doubling with its growing cliff ratio and exact halving, the t-independent reverse ratio, the constant-mixing steady state, decaying-schedule divergence, and matched-budget dominance (Figures 4–5; Tables 2–4; Appendices D and C). The categorical testbed (Figure 3, Appendix A) confirms the Gumbel extinction window and the plug-in/renormalization/empirical hierarchy.

6 Conclusion

We gave a closed-form account of recursive model collapse in the Gaussian family, carried to discrete data and to recursive GPT-2 training. The apparent controversy over the collapse mechanism resolves, in the Gaussian testbed both sides use, as an artifact of the divergence: mean drift and variance contraction contribute equally to the forward KL. Its quadratic cliff is half estimator bias—the

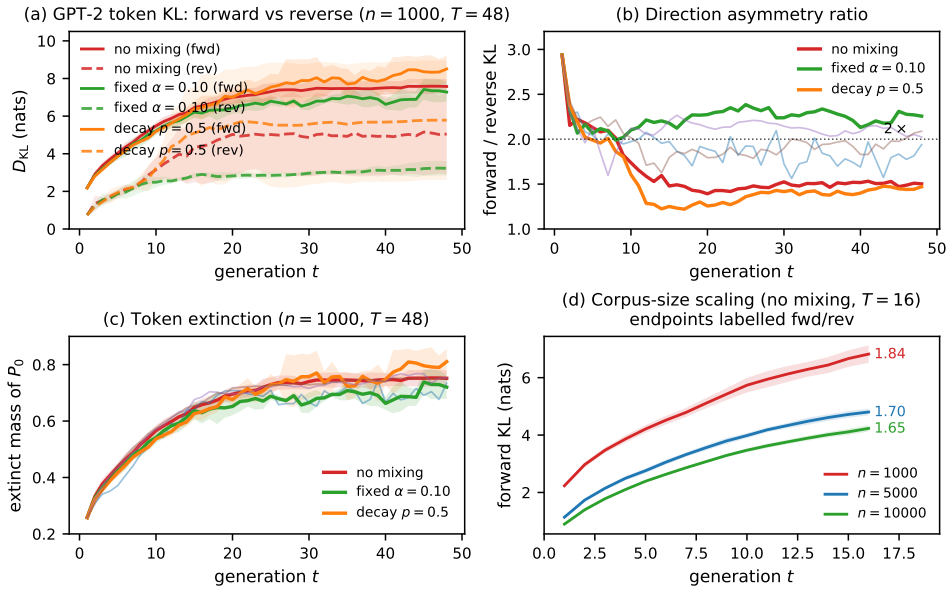


Figure 2: Recursive GPT-2 training (E10). (a) Forward (solid) vs. reverse (dashed) token KL, n=1000, T=48, 4 seeds, mean $\pm$ 95% CI: the forward KL runs at roughly twice the reverse, and the decaying p=0.5 schedule alone keeps degrading after generation 32. (b) Forward/reverse ratio across all six schedules (grey: not highlighted). (c) Extinct mass of P_0: saturating growth; mixing lowers the plateau, slow decay raises it. (d) Corpus-size scaling (no mixing, T=16, 9–10 seeds): larger per-generation corpora degrade more slowly, with endpoint forward/reverse ratios annotated.

blind spot of the plug-in variance estimator, which cannot see the variance it has already lost—and half intrinsic tail collapse; debiasing halves the cliff exactly, the reverse KL cancels it exactly, and both directions run at twice the known-variance rate. Mixing real data admits an exact steady state, any schedule whose mixing fraction decays to zero diverges, and constant mixing is the best non-increasing schedule at matched budget—the only one that bounds every generation.

The GPT-2 experiments confirm the portable content of these laws at the token level: a sustained two-fold direction asymmetry that attenuates with corpus size, frequency-resolved extinction harsher than the plug-in window, and schedule-set degradation plateaus—with the caution that a decaying schedule safe at short horizons can become the worst choice at triple the horizon, while training perplexity misranks schedules throughout.

Three limitations frame future work: the LM experiments validate the qualitative structure of the Gaussian laws, not their constants (corpus size maps to effective sample size only indirectly); the extinction analysis is plug-in, and the empirical-feedback recursion is strictly harsher—closing that gap analytically is open; and the mixing steady state is a mean-variance statement, with a full distributional theory of mixed recursive training, including the vocabulary dimension that dominates the LM behaviour here, left open.

AI USE STATEMENT

In this work we used generative AI tools as assistants for (i) verifying the mathematical derivations by independent re-derivation and Monte Carlo cross-check, (ii) developing and validating the simulation and experiment code, and (iii) drafting and editing of the manuscript text. All theorems were derived by hand from first principles, all AI-suggested derivations and code were checked by the authors, and all experimental numbers were generated by the deterministic pipelines described in the paper. 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 the mathematics of recursive training on machine-generated data. We see no direct ethical concerns; the broader societal questions of synthetic-data ecosystems are discussed in the cited literature.

Reproducibility statement

All theorems are proved in Appendix B, and every closed form was independently verified by hand re-derivation and Monte Carlo simulation (scripts included in the code release). The Gaussian and categorical simulations follow the exact recursions stated in the text (sample sizes, seed counts, and grid values are reported alongside each table in Appendix C). The GPT-2 experimental protocol—training configuration, mixing schedules, token-histogram instrumentation, seed inventory, and aggregation rules—is described in Section 5 and Appendix E; source code, per-seed metric arrays, and the aggregation pipeline will be released upon publication.

References

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. In *International Conference on Learning Representations (ICLR)*, 2024.

Shi Fu, Sen Zhang, Yingjie Wang, Xinmei Tian, and Dacheng Tao. Towards theoretical understandings of self-consuming generative models. *arXiv preprint arXiv:2402.11778*, 2024.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*, 2024.

Hengzhi He, Shirong Xu, and Guang Cheng. Recursive learning without collapse: A weighting-based stabilization framework. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2026. doi: 10.1093/jrsssb/qkag099.

Millen Kanabar and Michael Gastpar. Model non-collapse: Minimax bounds for recursive discrete distribution estimation. *arXiv preprint arXiv:2501.19273*, 2025.

Rylan Schaeffer, Joshua Kazdan, Alvan Caleb Arulandu, and Sanmi Koyejo. Position: Model collapse does not mean what you think. *arXiv preprint arXiv:2503.03150*, 2025.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross J. Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

Ananda Theertha Suresh, Andrew Thangaraj, and Aditya Nanda Kishore Khandavally. Rate of model collapse in recursive training. *arXiv preprint arXiv:2412.17646*, 2024.

APPENDIX

A CATEGORY EXTINCTION IN RECURSIVE TRAINING

The categorical recursion analysed here is the *plug-in* (stable-learner) idealisation: at each generation the learner re-estimates the true probabilities of the surviving categories (renormalised over survivors) and re-samples from them, so a surviving category's generative probability is its original p_i up to renormalisation. This is the natural categorical analogue of a learner that *generalises* from its sample. The exact MLE recursion of Definition 1—re-training on the *empirical* frequency vector—is strictly harsher: per-category counts then follow the critical branching recursion $X_{s+1} \mid X_s \sim \mathrm{Bin}(n, X_s/n)$, and our simulations show boundary-category survival falling from ≈ 0.37 (plug-in, matching e^{-1}) to 0.33 at $t = 10$ and 0.25 at $t = 20$ under empirical feedback, consistent with the survival bounds of Suresh et al. (2024). We present the plug-in analysis because it is closed-form throughout, and we restrict to Zipf initial distributions—the canonical model of token frequencies in natural language—because the renormalisation feedback grows with the mass of the dying tail and is small precisely for Zipf.

Proposition 2 (Category Extinction)**.** *For a discrete distribution with categories $\{1, \ldots, k\}$ and probabilities $p_1, \ldots, p_k$, under recursive sampling with n samples per generation in the plug-in regime:*

1. *Without re-normalization, the survival probability of category i after t generations is:*

$$P_i(t) = [1 - (1 - p_i)^n]^t \tag{9}$$

2. *A sharp phase transition occurs at $p_i = c \cdot \ln(t)/n$: for $c > 1$, $P_i(t) \to 1$; for $c < 1$, $P_i(t) \to 0$. Equivalently, with $x = np_i - \ln t$, the survival probability collapses onto the universal Gumbel distribution function $P_i(t) \to \exp(-e^{-x})$ in the transition window.*

3. *With re-normalization (extinct categories' mass redistributed proportionally among survivors), every surviving category's effective probability is amplified, $\tilde{p}_i^{(s)} = p_i/Z_{s-1}$ with $Z_{s-1} \leq 1$ the survivor mass: re-normalization* protects *survivors, lowering the extinction threshold in original-probability coordinates from $\ln(t)/n$ to $Z_t \cdot \ln(t)/n$.*

Proof sketch. (1) Category i appears in a generation with probability $q_i = 1 - (1 - p_i)^n$; survival across t independent generations gives q_i^t. (2) With $p_i = c\ln(t)/n$, $q_i = 1 - t^{-c}$ and $P_i(t) = (1 - t^{-c})^t \to \exp(-t^{1-c})$: the transition at $c = 1$; substituting $p_i = (\ln t + x)/n$ gives the Gumbel limit $\exp(-e^{-x})$. (3) Renormalisation multiplies each survivor's probability by $1/Z_{s-1} \geq 1$, so per-generation survival only increases. Full proof, including the absorbing-chain connection to Shumailov et al. (2024) and the quantitative magnitude of the protection, in Appendix B.5. □

Three quantitative consequences. **(i) Multi-generation scaling:** the single-generation absorption threshold $q_i \sim 1/n$ of Shumailov et al. (2024) sharpens under recursion to the Gumbel window $np_i \approx \ln t$: categories with np_i above $\ln t$ survive, below it they die, and the crossover follows the universal extreme-value law. **(ii) The protection is thin for Zipf:** evaluating the survivor mass self-consistently, $Z_s = \sum_j p_j P_j^{\mathrm{renorm}}(s)$, gives $Z_t = 0.90, 0.87, 0.86, 0.84, 0.83$ at $t = 1, 3, 5, 10, 20$ for Zipf(1.5) with $K = 1000$, $n = 500$; the renormalization-aware prediction matches Monte Carlo simulation of the plug-in recursion to within 0.004 mean absolute deviation at *every* t, while the no-renormalization Gumbel curve under-predicts transition-window survival by 0.13–0.16 at $t = 20$. For heavier-tailed initials the dying tail carries more mass and the protection grows (transition-window deviation 0.6–0.7 at $t = 20$ for a pmf $\propto i^{-1/2}$), which is precisely why we restrict to Zipf. **(iii) Empirical feedback is harsher still:** the MLE recursion sits *below* the renormalization-aware plug-in curve with a gap that grows in t (boundary survival 0.33 at t=10, 0.25 at t=20 vs. plug-in $e^{-1} \approx 0.37$)—real learners that memorise their empirical counts lose categories faster than any plug-in analysis predicts. Figure 3 validates the Gumbel window and this hierarchy in the categorical testbed.

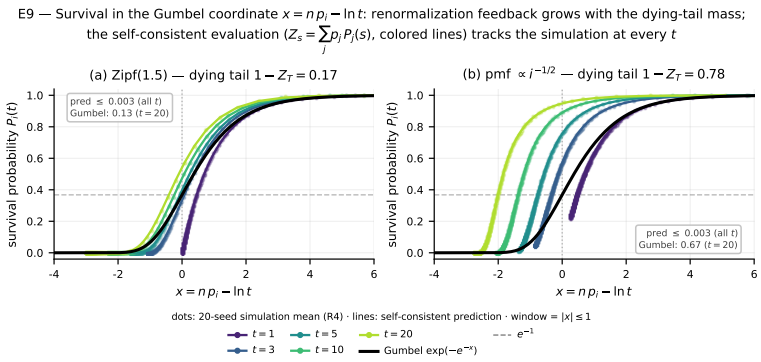


Figure 3: Extinction is thresholded and Gumbel-distributed (categorical testbed, Experiment E9). Survival after t generations collapses onto the universal distribution function $\exp(-e^{-x})$ in the shifted coordinate $x = np_i - \ln t$, across n and t. Empirical-feedback (MLE) recursion sits below the renormalization-aware plug-in reference, with the gap growing in t: real learners die faster than the plug-in law.

B Full Proofs

B.1 Proof of Theorem 1 (Mean Drift, Known Variance)

Proof. The proof proceeds in three steps: (1) characterizing the MLE distribution, (2) unrolling the random walk recursion, and (3) computing the expected KL divergence.

Step 1: MLE Distribution (Sample Mean). At generation t, we observe n i.i.d. samples $\boldsymbol{X}_1^{(t)}, \ldots, \boldsymbol{X}_n^{(t)}$ drawn from $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_0)$. For the multivariate Gaussian with known covariance, the maximum likelihood estimator (MLE) for the mean is the sample mean:

$$\hat{\boldsymbol{\mu}}_t = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i^{(t)} \tag{10}$$

Since the samples are i.i.d. Gaussian, the sample mean is itself Gaussian:

$$\hat{\boldsymbol{\mu}}_t \sim \mathcal{N}\left(\boldsymbol{\mu}_t, \frac{\boldsymbol{\Sigma}_0}{n}\right) \tag{11}$$

This follows from the standard result that a linear combination of independent Gaussian random vectors is Gaussian, with mean equal to the linear combination of means and covariance equal to the linear combination of covariances (scaled by $1/n^2$ and summed n times):

$$\mathrm{Var}[\hat{\boldsymbol{\mu}}_t] = \mathrm{Var}\left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i^{(t)}\right] = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}[\boldsymbol{X}_i^{(t)}] = \frac{1}{n^2} \cdot n\boldsymbol{\Sigma}_0 = \frac{\boldsymbol{\Sigma}_0}{n} \tag{12}$$

Step 2: Random Walk Recursion (Unrolling). Under the recursive generative training protocol (Definition 1), we set $\boldsymbol{\mu}_{t+1} = \hat{\boldsymbol{\mu}}_t$. Therefore:

$$\boldsymbol{\mu}_{t+1} = \hat{\boldsymbol{\mu}}_t = \boldsymbol{\mu}_t + \underbrace{(\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t)}_{\boldsymbol{Z}_t} \tag{13}$$

where $\boldsymbol{Z}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_0/n)$ are i.i.d. across generations. The independence of $\boldsymbol{Z}_0, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{t-1}$ follows from the fact that each generation draws *fresh* samples independently of all previous generations.

Unrolling the recursion from generation 0 to generation t:

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \boldsymbol{Z}_{t-1} = \boldsymbol{\mu}_{t-2} + \boldsymbol{Z}_{t-2} + \boldsymbol{Z}_{t-1} = \cdots = \boldsymbol{\mu}_0 + \sum_{s=0}^{t-1} \boldsymbol{Z}_s \tag{14}$$

Therefore:

$$\boldsymbol{\mu}_t - \boldsymbol{\mu}_0 = \sum_{s=0}^{t-1} \boldsymbol{Z}_s \sim \mathcal{N}\left(\mathbf{0},\, \frac{t\boldsymbol{\Sigma}_0}{n}\right) \tag{15}$$

This is a d-dimensional random walk with isotropic (in the $\boldsymbol{\Sigma}_0$-metric) increments of variance $\boldsymbol{\Sigma}_0/n$ per step.

Step 3: KL Divergence Computation (Trace Trick). The KL divergence between two multivariate Gaussians with the same covariance but different means is:

$$D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \| \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_0)) = \frac{1}{2}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_0) \tag{16}$$

Write:

$$\mathbb{E}[D_{\mathrm{KL}}] = \frac{1}{2}\mathbb{E}\left[\mathrm{tr}\left((\boldsymbol{\mu}_t - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_0)\right)\right] = \frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}\mathbb{E}[(\boldsymbol{\mu}_t - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_t - \boldsymbol{\mu}_0)^\top]\right) \tag{17}$$

$$= \frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1} \cdot \frac{t\boldsymbol{\Sigma}_0}{n}\right) = \frac{t}{2n}\,\mathrm{tr}(\boldsymbol{I}_d) = \frac{td}{2n} \tag{18}$$

In the second line, we used $\mathbb{E}[(\boldsymbol{\mu}_t - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_t - \boldsymbol{\mu}_0)^\top] = \mathrm{Var}[\boldsymbol{\mu}_t] = t\boldsymbol{\Sigma}_0/n$ since $\mathbb{E}[\boldsymbol{\mu}_t] = \boldsymbol{\mu}_0$; in the third, the cyclic property of the trace. By symmetry the reverse KL equals the same quantity. The per-generation rate $d/(2n)$ is the fundamental unit of information loss in recursive generative training with known variance. □

B.2 Proof of Theorem 2 (Forward-KL Collapse)

Proof. We prove each part in turn. Throughout, $d = 1$; for diagonal covariance the dimensions are independent and identical, so all per-dimension quantities multiply by d.

Part 1: Variance contraction. At generation t, the model draws n i.i.d. samples from the previous generation: $X_1^{(t)}, \ldots, X_n^{(t)} \sim \mathcal{N}(\hat{\mu}_{t-1}, \hat{\sigma}_{t-1}^2)$. Let $s^2 = \hat{\sigma}_{t-1}^2$. By Cochran's theorem, $\frac{n\hat{\sigma}_t^2}{s^2} \mid \hat{\sigma}_{t-1}^2 = s^2 \sim \chi_{n-1}^2$ (the sample mean consumes one degree of freedom). Using $\mathbb{E}[\chi_{n-1}^2] = n - 1$:

$$\mathbb{E}[\hat{\sigma}_t^2 \mid \hat{\sigma}_{t-1}^2 = s^2] = s^2 \cdot \frac{n-1}{n}. \tag{19}$$

By the tower property $\mathbb{E}[\hat{\sigma}_t^2] = \frac{n-1}{n}\,\mathbb{E}[\hat{\sigma}_{t-1}^2]$, giving the exponential contraction $\mathbb{E}[\hat{\sigma}_t^2] = \sigma_0^2((n-1)/n)^t$. Moreover, defining $W_s = n\hat{\sigma}_s^2/\hat{\sigma}_{s-1}^2$, we have the product representation

$$\hat{\sigma}_t^2 = \sigma_0^2 \prod_{s=1}^{t} \frac{W_s}{n} \equiv \sigma_0^2 V_t, \qquad V_t = \prod_{s=1}^{t} \frac{W_s}{n}. \tag{20}$$

Since each generation draws fresh samples and the conditional law of W_s given $\hat{\sigma}_{s-1}^2$ does not depend on $\hat{\sigma}_{s-1}^2$, the W_s are mutually independent (and independent of the past mean path).

Key moments of $W_s \sim \chi_{n-1}^2 = \mathrm{Gamma}((n-1)/2, 1/2)$:

1. $\mathbb{E}[W_s] = n - 1; \quad \mathrm{Var}[W_s] = 2(n-1)$.
2. **Inverse moment (requires $n > 3$):** $\mathbb{E}[1/W_s] = \frac{1/2}{(n-1)/2-1} = \frac{1}{n-3}$, so $\mathbb{E}[n/W_s] = \frac{n}{n-3}$. If $n \le 3$ the inverse moment diverges—hence the requirement $n > 3$ for the forward KL.
3. **Logarithmic moment:** $\mathbb{E}[\log W_s] = \psi(\frac{n-1}{2}) + \log 2$, where ψ is the digamma function.

Part 2: Forward-KL decomposition. For $P_0 = \mathcal{N}(\mu_0, \sigma_0^2)$ and $P_t = \mathcal{N}(\hat{\mu}_t, \hat{\sigma}_t^2)$,

$$D_{\mathrm{KL}}(P_0 \| P_t) = \underbrace{\frac{(\hat{\mu}_t - \mu_0)^2}{2\sigma_0^2 V_t}}_{\text{mean drift}} + \underbrace{\frac{1}{2}\left[\frac{1}{V_t} - 1 + \log V_t\right]}_{\text{variance mismatch}}. \tag{21}$$

The forward KL has $+\log V_t$ in the variance mismatch; the reverse KL (Appendix B.4) has $-\log V_t$. This sign difference is what causes the forward KL to retain the divergent $\mathbb{E}[1/V_t]$ term (the cliff) while the reverse KL cancels it.

Mean drift. The mean follows a random walk $\hat{\mu}_t - \mu_0 = \sum_{s=0}^{t-1} Z_s$ with $Z_s \sim \mathcal{N}(0, \sigma_0^2 V_s/n)$ conditionally independent given the variance path. Cross-terms $\mathbb{E}[Z_i Z_j \mid V]$ vanish by conditional independence and zero mean, so

$$\mathbb{E}\left[\frac{(\hat{\mu}_t - \mu_0)^2}{V_t}\right] = \sum_{s=0}^{t-1} \mathbb{E}\left[\frac{Z_s^2}{V_t}\right] = \frac{\sigma_0^2}{n} \sum_{s=0}^{t-1} \mathbb{E}\left[\frac{V_s}{V_t}\right]. \tag{22}$$

Since $V_t = \prod_{k=1}^{t}(W_k/n)$ and the W_k are i.i.d., the ratio telescopes:

$$\frac{V_s}{V_t} = \prod_{k=s+1}^{t} \frac{n}{W_k}, \qquad \mathbb{E}\left[\frac{V_s}{V_t}\right] = \prod_{k=s+1}^{t} \mathbb{E}\left[\frac{n}{W_k}\right] = \left(\frac{n}{n-3}\right)^{t-s}. \tag{23}$$

Summing the geometric series ($r = n/(n-3)$, so $r - 1 = 3/(n-3)$):

$$\sum_{s=0}^{t-1} \left(\frac{n}{n-3}\right)^{t-s} = \frac{n}{3}\left[\left(\frac{n}{n-3}\right)^{t} - 1\right] = \frac{n}{3}(I_t - 1), \tag{24}$$

where $I_t = \mathbb{E}[1/V_t] = (n/(n-3))^t$. Substituting back:

$$\mathbb{E}[\text{mean drift}] = \frac{1}{2\sigma_0^2} \cdot \frac{\sigma_0^2}{n} \cdot \frac{n}{3}(I_t - 1) = \frac{1}{6}(I_t - 1). \tag{25}$$

Variance mismatch. By linearity of expectation:

$$\mathbb{E}[\text{variance mismatch}] = \frac{1}{2}\mathbb{E}\left[\frac{1}{V_t}\right] - \frac{1}{2} + \frac{1}{2}\mathbb{E}[\log V_t] = \frac{1}{2}(I_t - 1) + \frac{1}{2}\mathbb{E}[\log V_t], \tag{26}$$

with $\mathbb{E}[\log V_t] = \sum_{s=1}^{t} \mathbb{E}[\log(W_s/n)] = t[\psi((n-1)/2) + \log 2 - \log n]$.

Exact closed form. Adding Eqs. equation 25 and equation 26, the I_t terms combine:

$$\mathbb{E}[D_{\mathrm{KL}}(P_0 \| P_t)] = \frac{2}{3}\left[\left(\frac{n}{n-3}\right)^{t} - 1\right] + \frac{t}{2}\left[\psi\left(\frac{n-1}{2}\right) + \log\left(\frac{2}{n}\right)\right]. \tag{27}$$

Unlike the reverse KL, the forward KL is dominated by the inverse moment I_t, which captures the catastrophic variance collapse: when $V_t \to 0$ the model cannot cover the true tails, and the forward KL blows up.

Part 3: Asymptotic expansion. *Step A: inverse moment.* Using $(1-x)^{-t} = 1 + tx + \frac{t(t+1)}{2}x^2 + O(x^3)$ with $x = 3/n$:

$$I_t - 1 = \frac{3t}{n} + \frac{9t(t+1)}{2n^2} + O(n^{-3}), \qquad \frac{2}{3}(I_t - 1) = \frac{2t}{n} + \frac{3t(t+1)}{n^2} + O(n^{-3}). \tag{28}$$

Step B: logarithmic moment. Using $\psi(z) = \log z - \frac{1}{2z} - \frac{1}{12z^2} + O(z^{-4})$ with $z = (n-1)/2$:

$$\psi\left(\tfrac{n-1}{2}\right) + \log 2 - \log n = \log\left(1 - \tfrac{1}{n}\right) - \tfrac{1}{n-1} - \tfrac{1}{3(n-1)^2} + O(n^{-3}) = -\frac{2}{n} - \frac{11}{6n^2} + O(n^{-3}), \tag{29}$$

so $\frac{1}{2}\mathbb{E}[\log V_t] = -\frac{t}{n} - \frac{11t}{12n^2} + O(n^{-3})$. *Step C: combine.*

$$\mathbb{E}[D_{\mathrm{KL}}(P_0 \| P_t)] = \left(\frac{2t}{n} + \frac{3t^2 + 3t}{n^2}\right) + \left(-\frac{t}{n} - \frac{11t}{12n^2}\right) + O(n^{-3}) = \frac{t}{n} + \frac{3t^2 + \frac{25}{12}t}{n^2} + O\left(\frac{t^3}{n^3}\right). \tag{30}$$

For a d-dimensional Gaussian with diagonal covariance, the per-dimension dynamics are independent and identical, and the total KL is the sum over dimensions: $\mathbb{E}[D_{\mathrm{KL}}(P_0 \| P_t)] = \frac{td}{n} + \frac{3t^2 d + 25td/12}{n^2} + O(td/n^3)$. □

B.3 Expanded Derivation for Proposition 1

Proof of the asymptotic expansion in Proposition 1. With the unbiased estimator, $V_t^{\mathrm{u}} = \prod_{s\le t} W_s/(n-1)$ with $W_s \sim \chi^2_{n-1}$ i.i.d., and $I_t^{\mathrm{u}} = \mathbb{E}[1/V_t^{\mathrm{u}}] = \left(\frac{n-1}{n-3}\right)^t$ because $\mathbb{E}[(n-1)/W_s] = (n-1)/(n-3)$. The mean drift is $\frac{n-1}{4n}(I_t^{\mathrm{u}} - 1)$ (same telescoping as Theorem 2 with $W_s/(n-1)$ in place of W_s/n; the geometric ratio is $\frac{n-1}{n-3}$, whose sum over $s < t$ is $\frac{n-1}{2}(I_t^{\mathrm{u}} - 1)$, and the factor $\frac{1}{2n}$ from the drift denominator gives $\frac{n-1}{4n}(I_t^{\mathrm{u}} - 1)$). The variance mismatch is $\frac{1}{2}(I_t^{\mathrm{u}} - 1) + \frac{1}{2}\mathbb{E}[\log V_t^{\mathrm{u}}]$, and $\frac{n-1}{4n} + \frac{1}{2} = \frac{3n-1}{4n}$, giving Eq. equation 4 with $\mathbb{E}[\log V_t^{\mathrm{u}}] = t[\psi(\frac{n-1}{2}) + \log\frac{2}{n-1}]$.

For the expansion: with $x = \frac{2}{n-3}$,

$$I_t^{\mathrm{u}} - 1 = (1+x)^t - 1 = tx + \frac{t(t-1)}{2}x^2 + O(x^3) = \frac{2t}{n} + \frac{2t^2+4t}{n^2} + O\left(\frac{t^3}{n^3}\right), \tag{31}$$

and

$$\mathbb{E}[\log V_t^{\mathrm{u}}] = -\frac{t}{n-1} - \frac{t}{3(n-1)^2} + O(t/n^3) = -\frac{t}{n} - \frac{4t}{3n^2} + O(t/n^3). \tag{32}$$

Combining,

$$\left(\tfrac{3}{4} - \tfrac{1}{4n}\right)(I_t^{\mathrm{u}} - 1) + \tfrac{1}{2}\mathbb{E}[\log V_t^{\mathrm{u}}] = \frac{t}{n} + \frac{\frac{3}{2}t^2 + \frac{11}{6}t}{n^2} + O\left(\frac{t^3}{n^3}\right), \tag{33}$$

using $\frac{3}{4}\cdot\frac{2t^2+4t}{n^2} - \frac{2t}{4n^2} - \frac{2t}{3n^2} = \frac{3t^2/2+(3-\frac{1}{2}-\frac{2}{3})t}{n^2} = \frac{3t^2/2+\frac{11}{6}t}{n^2}$. □

B.4 Proof of Corollary 2 (Reverse-KL Cancellation)

Proof. The reverse KL between $P_0 = \mathcal{N}(\mu_0, \sigma_0^2)$ and $P_t = \mathcal{N}(\hat{\mu}_t, \hat{\sigma}_t^2)$ is

$$D_{\mathrm{KL}}(P_t\|P_0) = \underbrace{\frac{(\hat{\mu}_t - \mu_0)^2}{2\sigma_0^2}}_{\text{mean drift}} + \underbrace{\frac{1}{2}[V_t - 1 - \log V_t]}_{\text{variance mismatch}}, \tag{34}$$

where $V_t = \hat{\sigma}_t^2/\sigma_0^2$. Note three structural differences from the forward KL (Eq. equation 21):

1. The mean-drift denominator is σ_0^2 (the true variance), *not* $\hat{\sigma}_t^2$: no $\mathbb{E}[1/V_t]$ term appears.
2. The variance-ratio is V_t (the contracted variance in the numerator), not $1/V_t$.
3. The sign of $\log V_t$ is *negative*, whereas it is positive in the forward KL. This sign flip is what enables the cancellation below.

Mean drift. The mean follows the same random walk as in Theorem 2, but the denominator is now the constant σ_0^2. By the law of total variance and the conditional independence of the steps Z_s given the variance path:

$$\mathbb{E}[(\hat{\mu}_t - \mu_0)^2] = \sum_{s=0}^{t-1}\mathbb{E}\left[\frac{\hat{\sigma}_s^2}{n}\right] = \frac{\sigma_0^2}{n}\sum_{s=0}^{t-1}\left(\frac{n-1}{n}\right)^s = \sigma_0^2\left[1 - \left(\frac{n-1}{n}\right)^t\right] = \sigma_0^2(1 - \mathbb{E}[V_t]). \tag{35}$$

Dividing by $2\sigma_0^2$ gives $\mathbb{E}[\text{mean drift}] = \frac{1}{2}(1 - \mathbb{E}[V_t])$.

Variance mismatch. By linearity of expectation, $\mathbb{E}[\text{variance mismatch}] = \frac{1}{2}(\mathbb{E}[V_t] - 1 - \mathbb{E}[\log V_t])$.

The exact cancellation. Adding the two parts:

$$\mathbb{E}[D_{\mathrm{KL}}(P_t\|P_0)] = \tfrac{1}{2}(1 - \mathbb{E}[V_t]) + \tfrac{1}{2}(-\mathbb{E}[\log V_t] + \mathbb{E}[V_t] - 1) = -\tfrac{1}{2}\mathbb{E}[\log V_t]. \tag{36}$$

The $\mathbb{E}[V_t]$ terms cancel perfectly. This is the structural difference from the forward KL: in the forward KL, the variance mismatch contains $\mathbb{E}[1/V_t]$ (which grows exponentially in t) and cannot cancel with the mean drift; in the reverse KL, both mean drift and variance mismatch contain $\mathbb{E}[V_t]$

(which decays exponentially in t), and they cancel exactly. The remaining $-\frac{1}{2}\mathbb{E}[\log V_t]$ is strictly linear in t because $\mathbb{E}[\log V_t] = t[\psi((n-1)/2) + \log 2 - \log n]$ by the product representation with independent W_s. This yields Eq. equation 7; the identity holds for all $n \geq 2$ (only the chi-squared structure is needed; $n > 3$ is required solely by the forward inverse moment).

Doubling ratio. Since $\mathbb{E}[D_{\mathrm{KL,known}}] = t/(2n)$ (Theorem 1), the ratio is

$$\frac{-\frac{t}{2}[\psi((n-1)/2) + \log(2/n)]}{t/(2n)} = -n\left[\psi\left(\frac{n-1}{2}\right) + \log\left(\frac{2}{n}\right)\right] = 2 + \frac{11}{6n} + O(n^{-2}), \tag{37}$$

exactly t-independent. (The asymptotic expansion uses $\psi(z) = \log z - \frac{1}{2z} - \frac{1}{12z^2} + O(z^{-4})$ with $z = (n-1)/2$, as in Appendix B.2, Step B.) □

B.5 Proof of Proposition 2 (Category Extinction)

Proof. **Part 1: Without re-normalization.** At each generation s, we draw n samples with replacement from the current distribution. Without re-normalization the underlying probabilities remain unchanged: $p_i^{(s)} = p_i$. The count of category i in generation s is $N_i^{(s)} \sim \mathrm{Binomial}(n, p_i)$, and category i "appears" with probability $q_i = \Pr(N_i^{(s)} \geq 1) = 1 - (1-p_i)^n$. Category i survives t generations iff it appears in every generation; by independence across generations,

$$P_i(t) = q_i^t = \left[1 - (1-p_i)^n\right]^t. \tag{38}$$

Part 2: Phase transition. Using $1 - (1-p_i)^n \approx 1 - e^{-np_i}$ (exact in the limit $p_i \to 0$, $n \to \infty$ with np_i fixed), with $p_i = c\ln(t)/n$:

$$q_i = 1 - e^{-c\ln t} = 1 - t^{-c}, \qquad P_i(t) = (1-t^{-c})^t = \exp\left(t\ln(1-t^{-c})\right) \approx \exp\left(-t^{1-c}\right). \tag{39}$$

Hence: for $c > 1$, $t^{1-c} \to 0$ and $P_i(t) \to 1$; for $c < 1$, $t^{1-c} \to \infty$ and $P_i(t) \to 0$; at $c = 1$, $P_i(t) \to e^{-1}$. The threshold probability is $p^*(t) = \ln(t)/n$. Substituting $p_i = (\ln t + x)/n$ for fixed x gives $P_i(t) \to \exp(-e^{-x})$: in the transition window, survival follows the universal Gumbel distribution function in the shifted coordinate $x = np_i - \ln t$, connecting recursive category extinction to classical extreme-value theory.

Part 3: With re-normalization—protection, not acceleration. When extinct categories are removed and their probability mass redistributed proportionally among survivors, the effective probability of surviving category i at generation s becomes

$$\tilde{p}_i^{(s)} = \frac{p_i}{Z_{s-1}}, \qquad Z_{s-1} = \sum_{j \in S_{s-1}} p_j \leq 1, \tag{40}$$

where S_{s-1} is the survivor set and Z_{s-1} the survivor mass. Since $Z_{s-1} \leq 1$, every surviving category's effective probability is amplified, and hence

$$P_i^{\mathrm{renorm}}(t) = \prod_{s=1}^{t}\left[1 - (1 - p_i/Z_{s-1})^n\right] \;\geq\; \left[1 - (1-p_i)^n\right]^t = P_i(t). \tag{41}$$

Re-normalization therefore *protects* survivors: in original-probability coordinates the extinction threshold drops from $\ln(t)/n$ to $Z_t \cdot \ln(t)/n$, and the no-renormalization formula is a conservative lower bound on survival. The magnitude of the protection is governed by the dying tail's mass through Z_t, evaluated self-consistently as $Z_s = \sum_j p_j\, P_j^{\mathrm{renorm}}(s)$ with $Z_0 = 1$; the Zipf(1.5) and $\propto i^{-1/2}$ figures quoted in the main text follow from this iteration.

Connection to absorbing Markov chains. The extinction process is an absorbing Markov chain on $\mathcal{S} = \{0,1\}^k$ with transitions $s_i' = s_i \cdot \mathbf{1}[N_i^{(t)} \geq 1]$. The single-generation absorption probability $(1-p_i)^n$ matches the threshold analysis of Shumailov et al. (2024), who identified the single-generation threshold $q_i \approx 1/M$ (M the per-generation sample size); our contribution is the multi-generation scaling law $p_i \sim \ln(t)/n$ and the Gumbel phase transition at that threshold. □

B.6 Proofs of Theorem 3 and Corollary 3

Proof of Theorem 3. Under mixed training with fraction α (Definition 2), we draw $n_1 = n\alpha$ real samples from $\mathcal{N}(\mu_0, \sigma^2)$ and $n_2 = n(1-\alpha)$ synthetic samples from $\mathcal{N}(\mu_{t-1}, \sigma^2)$. The combined MLE for the mean is

$$\hat{\mu}_t = \frac{n_1 \bar{X}_{\text{real}} + n_2 \bar{X}_{\text{synth}}}{n_1 + n_2} = \alpha \bar{X}_{\text{real}} + (1-\alpha)\bar{X}_{\text{synth}}, \tag{42}$$

where $\bar{X}_{\text{real}} \sim \mathcal{N}(\mu_0, \sigma^2/(n\alpha))$ and $\bar{X}_{\text{synth}} \sim \mathcal{N}(\mu_{t-1}, \sigma^2/(n(1-\alpha)))$. Then

$$\mathbb{E}[\hat{\mu}_t \mid \mu_{t-1}] = \alpha\mu_0 + (1-\alpha)\mu_{t-1}, \qquad \text{Var}[\hat{\mu}_t \mid \mu_{t-1}] = \alpha^2 \frac{\sigma^2}{n\alpha} + (1-\alpha)^2 \frac{\sigma^2}{n(1-\alpha)} = \frac{\sigma^2}{n}, \tag{43}$$

the conditional variance being σ^2/n *regardless of* α: the reduction in variance from concentrating samples in one source is exactly offset by the reduced number of samples from that source. By the law of total variance,

$$\text{Var}[\mu_t] = \text{Var}[\mathbb{E}[\hat{\mu}_t \mid \mu_{t-1}]] + \mathbb{E}[\text{Var}[\hat{\mu}_t \mid \mu_{t-1}]] = (1-\alpha)^2 \text{Var}[\mu_{t-1}] + \frac{\sigma^2}{n}, \tag{44}$$

using $\text{Var}[\alpha\mu_0] = 0$. With $v_0 = 0$, this first-order linear recursion solves to

$$v_t = \frac{\sigma^2}{n} \cdot \frac{1-(1-\alpha)^{2t}}{\alpha(2-\alpha)} \xrightarrow[t\to\infty]{} \frac{\sigma^2}{n\alpha(2-\alpha)}, \tag{45}$$

since $|1-\alpha| < 1$ for $\alpha > 0$. For $\alpha \ll 1$, $2-\alpha \approx 2$. □

Proof of Corollary 3. With $\alpha_t = t^{-p}$, the variance recursion becomes $v_t \approx (1-2\alpha_t)v_{t-1} + \sigma^2/n$ (using $(1-\alpha_t)^2 = 1 - 2\alpha_t + O(\alpha_t^2)$). Unrolling:

$$v_T = \frac{\sigma^2}{n} \sum_{s=1}^{T} \prod_{r=s+1}^{T} (1-\alpha_r)^2 \approx \frac{\sigma^2}{n} \sum_{s=1}^{T} \exp\left(-2 \sum_{r=s+1}^{T} \alpha_r\right). \tag{46}$$

Each generation injects a *constant* noise σ^2/n while the contraction $(1-\alpha_t)^2 \to 1$ weakens.

Case 1: $0 < p < 1$**.** For $s = T - k$ with k small relative to T: $\sum_{r=s+1}^{T} r^{-p} \approx kT^{-p}$, so the contribution of generation s is $\approx \frac{\sigma^2}{n} e^{-2kT^{-p}}$. Summing over k:

$$v_T \approx \frac{\sigma^2}{n} \sum_{k=0}^{T-1} e^{-2kT^{-p}} = \frac{\sigma^2}{n} \cdot \frac{1}{1-e^{-2T^{-p}}} \approx \frac{\sigma^2}{2n} T^p \to \infty. \tag{47}$$

Case 2: $p = 1$**.** With $\alpha_t = 1/t$: $\sum_{r=s+1}^{T} r^{-1} \approx \ln(T/s)$, so

$$v_T \approx \frac{\sigma^2}{n} \sum_{s=1}^{T} e^{-2\ln(T/s)} = \frac{\sigma^2}{n} \sum_{s=1}^{T} (s/T)^2 \approx \frac{\sigma^2}{n} \cdot \frac{T}{3} \to \infty. \tag{48}$$

Case 3: $p > 1$**.** $\sum_{r=s+1}^{\infty} r^{-p} \le \zeta(p) < \infty$, so each generation's contribution is at least $\frac{\sigma^2}{n} e^{-2\zeta(p)}$, and $v_T \ge \frac{\sigma^2}{n} T e^{-2\zeta(p)} \to \infty$; in fact $v_T \sim C(p)\, T\sigma^2/n$ with $C(p) \to 1$ as $p \to \infty$ (the no-mixing limit).

In all cases $v_T \sim \frac{\sigma^2}{2n} t^{\min(p,1)} \to \infty$; only the constant case $p = 0$ achieves bounded variance. Finally, although $\text{Var}[\mu_t] \to \infty$ for all $p > 0$, $\mu_t \to \mu_0$ almost surely for $1/2 < p \le 1$ by the Robbins–Monro theorem ($\sum \alpha_t = \infty$, $\sum \alpha_t^2 < \infty$): convergence of the mean does not imply bounded variance when the noise is constant rather than scaling with α_t^2. □

C Verification Tables

This section reports the full numerical verification grids for Theorem 2 and Corollary 2. All Gaussian simulations use the exact recursion of Definition 1 ($d = 1$ throughout). Table 2 is computed from

Table 1: Constant vs. decaying mixing at matched $\Sigma\alpha$ (Gaussian, n=1000, T=8; Section 4). Variance is reported in units of σ^2/n. V_{const} is the exact steady-state value $[\alpha_{\text{const}}(2-\alpha_{\text{const}})]^{-1}$; V_{decay} is the exact finite-T recursion $\sum_s \prod_{r>s}(1-\alpha_r)^2$; the constant schedule achieves the lower variance in every row (the finite-T constant value $v_T = [1-(1-\alpha)^{2T}]/[\alpha(2-\alpha)]$ is smaller still, so the ordering is preserved in the like-for-like finite-T comparison).

Decaying schedule	$\Sigma\alpha$	α_{const}	V_{const}	V_{decay}
$0.5/t^{0.3}$	2.743	0.343	1.76	2.07
$0.5/\sqrt{t}$	2.186	0.273	2.12	2.74
$0.3/t^{0.3}$	1.646	0.206	2.71	3.02
$0.5/t$	1.359	0.170	3.22	4.61

the 20-seed GPU experiment (seeds 42–61); its ratio columns are 20-seed means of *per-seed paired ratios* (numerator and denominator simulated within the same seed), which is why they differ in the last digit from ratios of the printed means. Tables 3 and 4 use a fresh $M = 4000$-trial Monte Carlo for $n \leq 500$ (biased-MLE recursion simulated literally: n samples per generation, sample mean and biased sample variance recomputed each step) and the 20-seed experiment at $n = 1000$; there the ratio is $\mathbb{E}[D_{\text{KL}}]_{\text{MC}}/(t/2n)$ against the exact known-variance baseline, so the ratio columns of Tables 2 and 3 use different (and both stated) baselines.

Table 2: 20-seed verification of Theorem 2 (forward KL, unknown variance). "Fwd (sim/exact/asymp)": simulated, exact closed form (Eq. equation 2), and asymptotic (Eq. equation 3) forward KL. "Known (sim)": simulated known-variance KL. "Rev/Fwd ratio (sim)": 20-seed mean of per-seed paired doubling ratios; "Fwd ratio (exact)": closed-form prediction (Eq. equation 5 numerator). The forward ratio grows with t at $n = 1000$ (from 2.016 at t=1 to 2.318 at t=49) while the reverse ratio is t-independent at 2.00—the empirical direction asymmetry.

n	t	Fwd (sim)	Fwd (exact)	Fwd (asymp)	Known (sim)	Rev ratio	Fwd ratio	Fwd ratio (exact)
1000	1	0.00101	0.00101	0.00101	0.00050	2.005	2.016	2.010
1000	5	0.00510	0.00509	0.00509	0.00250	2.008	2.043	2.034
1000	10	0.01036	0.01032	0.01032	0.00499	2.007	2.073	2.065
1000	20	0.02133	0.02127	0.02124	0.00992	2.002	2.130	2.127
1000	30	0.03293	0.03285	0.03276	0.01480	2.004	2.195	2.190
1000	49	0.05679	0.05669	0.05631	0.02393	2.003	2.318	2.314
5000	1	0.00020	0.00020	0.00020	0.00010	2.001	2.004	2.002
5000	5	0.00101	0.00100	0.00100	0.00050	2.002	2.009	2.007
5000	10	0.00202	0.00201	0.00201	0.00100	2.005	2.019	2.013
5000	20	0.00405	0.00405	0.00405	0.00200	2.004	2.029	2.025
5000	30	0.00611	0.00611	0.00611	0.00299	2.002	2.039	2.037
5000	49	0.01010	0.01010	0.01009	0.00488	2.000	2.060	2.060
10000	1	0.00010	0.00010	0.00010	0.00005	1.999	2.000	2.001
10000	5	0.00050	0.00050	0.00050	0.00025	1.998	2.002	2.003
10000	10	0.00100	0.00100	0.00100	0.00050	2.000	2.006	2.006
10000	20	0.00201	0.00201	0.00201	0.00100	1.998	2.011	2.012
10000	30	0.00303	0.00303	0.00303	0.00150	2.002	2.020	2.018
10000	49	0.00497	0.00497	0.00497	0.00244	1.998	2.028	2.030
50000	1	0.00002	0.00002	0.00002	0.00001	2.001	2.001	2.000
50000	5	0.00010	0.00010	0.00010	0.00005	1.999	2.000	2.001
50000	10	0.00020	0.00020	0.00020	0.00010	1.998	1.999	2.001
50000	20	0.00040	0.00040	0.00040	0.00020	1.997	2.000	2.002
50000	30	0.00060	0.00060	0.00060	0.00030	1.998	2.002	2.004
50000	49	0.00098	0.00098	0.00098	0.00049	1.997	2.003	2.006

The exact closed form matches the simulation to four significant digits at every (n, t) (e.g. 0.05679 vs. 0.05669 at n=1000, t=49), and the asymptotic expansion is within 0.7% everywhere on the grid (Section 3.2). At $n = 50000$ the cliff itself is of order 3×10^{-6}; the 20-seed forward ratios 1.999–2.003 are consistent with the exact predictions 2.000–2.006 within seed noise, and the 20-seed averaging resolves them where a single seed could not.

Table 3: Verification of Corollary 2 (reverse KL). Rows $n \leq 500$: fresh Monte Carlo, $M = 4000$ trials; rows $n = 1000$: 20-seed GPU experiment. "Var ratio": $\mathbb{E}[V_t]$ simulated vs. $((n-1)/n)^t$ theory. "Rev ratio": $\mathbb{E}[D_{\mathrm{KL}}(P_t\|P_0)]_{\mathrm{MC}}/(t/2n)$ against the exact known-variance baseline ($\pm$ is the 95% CI of the MC mean). The ratio is t-independent at each n within CI, confirming Eq. equation 7 and Corollary 2.

n	t	Var ratio (sim)	Var ratio (theory)	Rev KL (sim)	Known KL (sim)	Rev ratio
10	5	0.570	0.590	0.5401	0.1968	2.160 ± 0.055
10	10	0.333	0.349	1.0860	0.3091	2.172 ± 0.050
20	5	0.774	0.774	0.2641	0.1159	2.113 ± 0.061
20	10	0.606	0.599	0.5317	0.2086	2.127 ± 0.059
50	5	0.903	0.904	0.1039	0.0492	2.078 ± 0.063
50	10	0.818	0.817	0.2071	0.0959	2.071 ± 0.062
100	5	0.948	0.951	0.0495	0.0240	1.981 ± 0.060
100	10	0.895	0.904	0.0994	0.0478	1.988 ± 0.062
100	20	0.814	0.818	0.2018	0.0897	2.018 ± 0.063
100	49	0.596	0.611	0.5030	0.1965	2.053 ± 0.062
200	5	0.974	0.975	0.0251	0.0123	2.005 ± 0.060
200	10	0.952	0.951	0.0510	0.0248	2.039 ± 0.063
200	20	0.913	0.905	0.1014	0.0477	2.027 ± 0.062
200	49	0.779	0.782	0.2471	0.1105	2.017 ± 0.060
500	5	0.989	0.990	0.0101	0.0051	2.023 ± 0.061
500	10	0.985	0.980	0.0203	0.0100	2.030 ± 0.063
500	49	0.914	0.907	0.0978	0.0467	1.995 ± 0.061
1000	5	0.995	0.995	0.0050	0.0025	2.008 ± 0.003
1000	10	0.990	0.990	0.0100	0.0050	2.007 ± 0.003
1000	49	0.953	0.952	0.0491	0.0239	2.003 ± 0.003

Table 4: Exact finite-n reverse-KL doubling ratio (Corollary 2; t-independent) vs. its asymptote. Simulation: $M = 4000$-trial Monte Carlo averaged over $t \in \{5, 10, 20, 49\}$ ($n \leq 500$); 20-seed experiment for $n = 1000$. The forward-KL leading-order-in-t doubling ratio is the *distinct* quantity $2n\left[\frac{2}{3}\log\frac{n}{n-3} + \frac{1}{2}\left(\psi\left(\frac{n-1}{2}\right) + \log\frac{2}{n}\right)\right] = 2 + \frac{25}{6n} + O(n^{-2})$ (Corollary 1); the two coincide only as $n \to \infty$.

n	Exact ratio (theory)	Sim	Asymptote
10	2.2057	2.1921 ± 0.0579	2
20	2.0969	2.1105 ± 0.0623	2
50	2.0375	2.0656 ± 0.0652	2
100	2.0185	2.0631 ± 0.0637	2
200	2.0092	2.0136 ± 0.0603	2
500	2.0037	2.0040 ± 0.0653	2
1000	2.0018	2.005 ± 0.003	2

Heavy-tail caveat for experimenters. The forward KL at small n and large t is dominated by the inverse moment $\mathbb{E}[1/V_t]$, whose second moment $\mathbb{E}[(1/V_t)^2] = \left(n^2/((n-3)(n-5))\right)^t$ explodes: sample means of the forward KL converge very slowly there (per-seed relative spreads above 100% at $n \lesssim 100$, $t \gtrsim 20$). This is why the forward-KL grid is restricted to $n \geq 1000$, why the d-averaged (large-d) design of the main Gaussian experiment is the right instrument for the forward direction, and why small-n forward cells should not be quoted without per-seed uncertainty.

Table 5: E10 schedule comparison at the tripled horizon (n=1000, T=48, 4 paired seeds; Section 5). "Δ32–48" is the forward-KL growth over generations 32–48 (mean $\pm$ SEM; t-statistic across seeds); plateaued schedules have $\Delta \approx 0$. Extinct mass and perplexity are generation-48 values (mean $\pm$ 95% CI).

Schedule	fwd KL @16	fwd KL @48	Δ32–48 (t)	extinct mass	perplexity
no mixing	6.54	7.58	$+0.11 \pm 0.06$ (1.7)	0.752	2522 ± 569
fixed 0.50	6.76	7.45	$+0.10 \pm 0.73$ (0.1)	0.734	99 ± 3
fixed 0.10	6.16	7.28	$+0.48 \pm 0.29$ (1.7)	0.720	158 ± 8
decay 0.5	6.45	**8.51**	$+0.69 \pm 0.09$ (7.7)	**0.811**	245 ± 27
decay 1.0	6.72	7.59	$+0.38 \pm 0.14$ (2.7)	0.754	601 ± 21
decay 2.0	6.53	7.40	$+0.19 \pm 0.16$ (1.2)	0.747	3177 ± 1195

D ADDITIONAL FIGURES

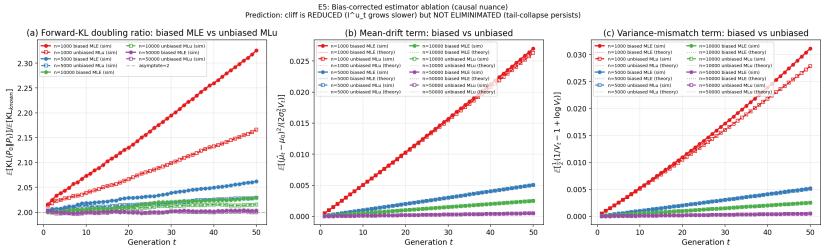


Figure 4: Estimator blind spot, ablated (Experiment E5, 18 seeds). Replacing the biased MLE variance estimator by the unbiased one leaves the leading rate t/n unchanged but removes exactly half of the quadratic cliff (theory: Eq. equation 4; measured reduction 49.1% at $n{=}1000$, $t{=}50$, positive in $18/18$ seeds). The remaining half is the intrinsic Jensen-gap mechanism.

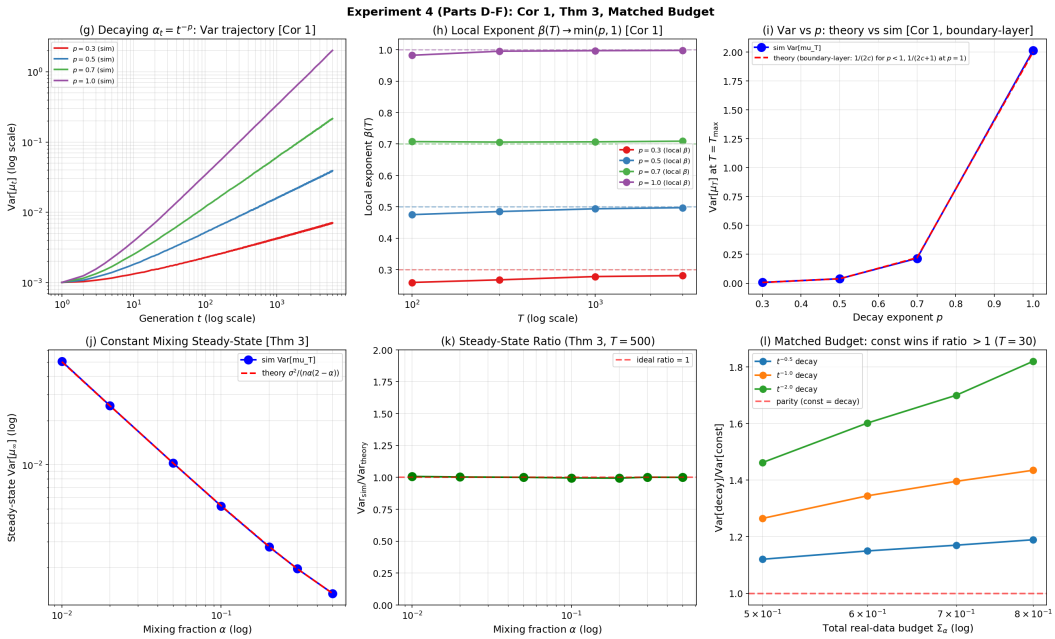


Figure 5: Mixing laws in the Gaussian testbed (Experiment E4, Parts D–F). (g–i) Decaying mixing $\alpha_t = t^{-p}$: simulated $\mathrm{Var}[\mu_T]$ diverges for every $p > 0$, following the boundary-layer theory (Corollary 3). (j–k) Constant mixing reaches the predicted steady state $\sigma^2/(n\alpha(2{-}\alpha))$ across three orders of magnitude in α (Theorem 3). (l) At matched total real-data budget $\Sigma\alpha$, constant mixing achieves lower variance than every decaying schedule tested ($p \in \{0.3, 0.5, 1.0, 2.0\}$); deferred schedules can beat it on the endpoint alone (Section 4).

E E10 GPT-2 PROTOCOL DETAILS

Model and data. GPT-2 small (124M parameters) trained from scratch in each generation (no pretraining carry-over), AdamW, fixed hyperparameters across all generations and schedules. The real pool P_0 is a fixed WikiText-2 document set; P_0 token statistics are computed from the same pool with identical max-length truncation as generation. Each generation corpus has $n \in \{1000, 5000, 10000\}$ documents; generation uses the canonical parameters of the distillation pipeline (top-k/top-p truncated sampling, EOS- and prompt-stripped continuation token streams).

Schedules. Six conditions per n: no mixing; fixed $\alpha \in \{0.10, 0.50\}$; decaying $\alpha_t = c_p(t+1)^{-p}$, $p \in \{0.5, 1.0, 2.0\}$, with c_p normalised so that $\sum_{t=0}^{7} \alpha_t = 0.75$ for all three decays (budget-matched decay family over the first eight generations). Schedules are *paired* across seeds: the same base seed drives all six schedules, so cross-schedule comparisons are paired tests.

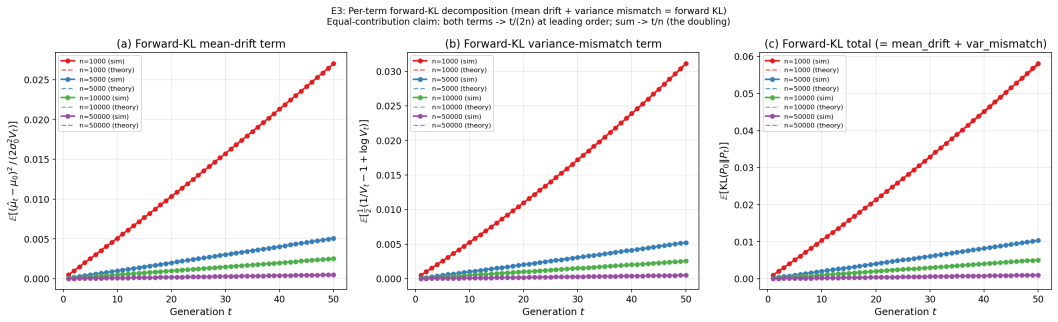


Figure 6: Per-term decomposition of the forward KL (Experiment E3): mean drift and variance mismatch each contribute $t/(2n)$ at leading order—the "equal contribution" of Remark 1, resolved term by term across n and t.

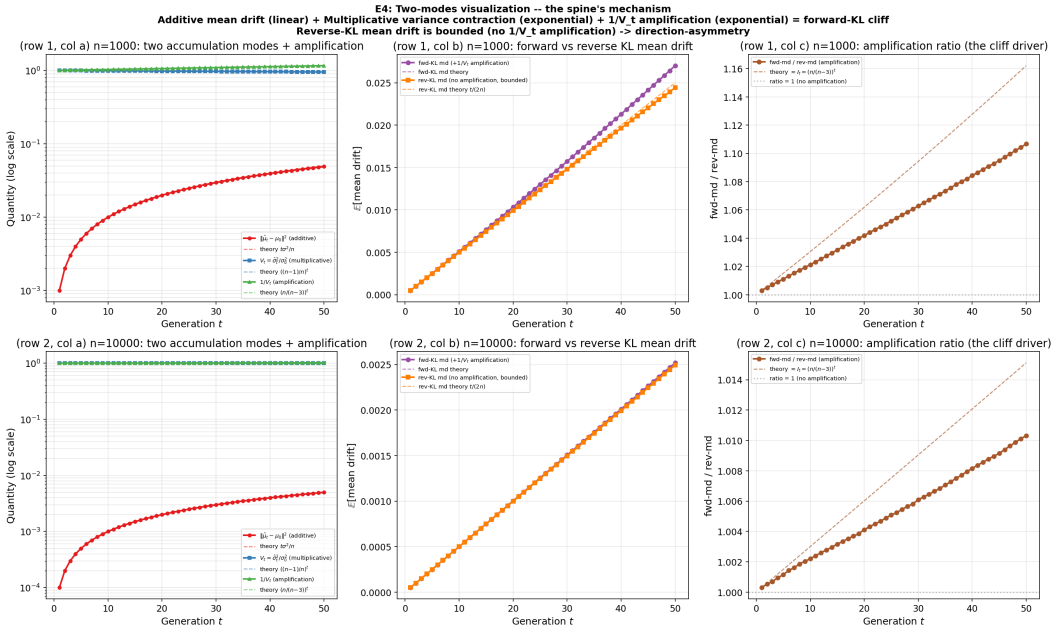


Figure 7: The two modes of collapse (Experiment E4): mean drift (distributional shift) and variance contraction (diversity loss) visualised across the recursion; both operate simultaneously and, under the forward KL, contribute equally at leading order.

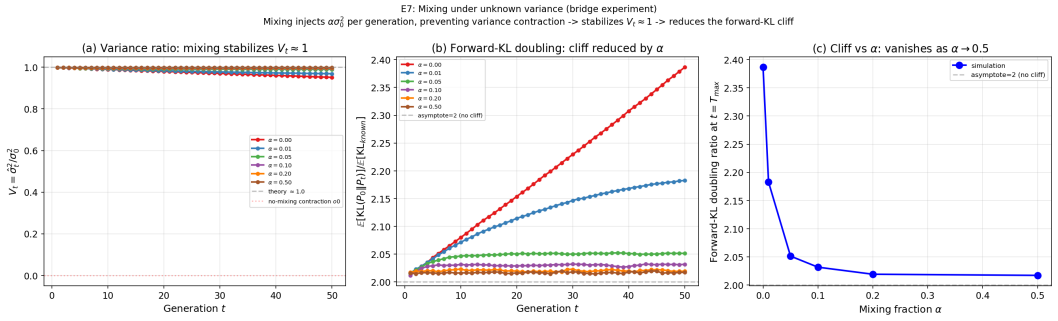


Figure 8: Mixing with unknown variance (Experiment E7): constant mixing stabilises the variance ratio $V_t \approx 1$ and reduces the forward-KL cliff in proportion to α, consistent with Theorem 3 transferring to the unknown-variance setting.

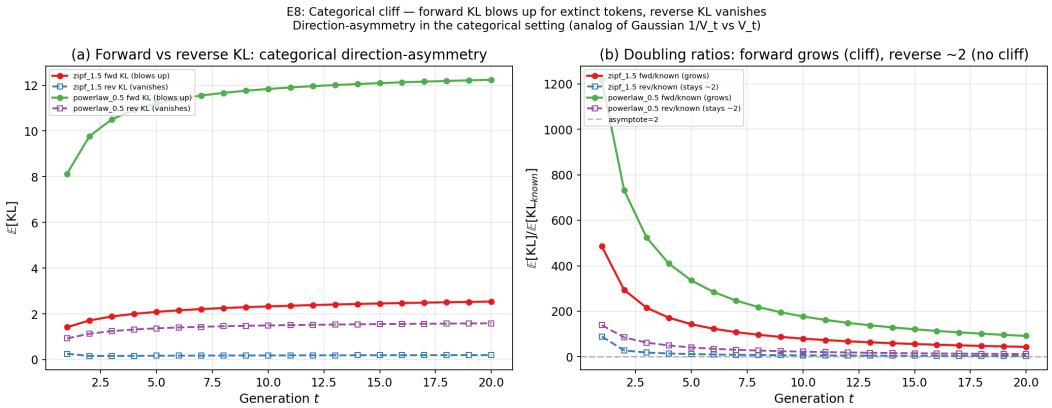


Figure 9: Categorical extinction cliff (Experiment E8): the extinct mass of the initial distribution grows sharply once the per-generation sample budget falls below the extinction threshold, across Zipf and heavier-tailed initial distributions.

Instrumentation. Per generation: token histogram of the generated corpus (prompt/EOS-stripped continuations), token histogram of the training corpus, forward and reverse KL against the P_0 histogram with floor 10^{-8}, Jensen–Shannon divergence, extinct mass (P_0 mass on tokens with zero generated count), out-of-support mass, active vocabulary, unigram entropy, distinct-n and conditional-entropy summaries, perplexity, and the realised mixing fractions α_t, $\alpha_{\text{effective}}$. Floor-free metrics accompany the floored KL per the pre-registered fallback plan; all KL claims in the main text are reproduced by the JS divergence, which needs no floor.

Seeds and aggregation. T=16: seeds 42–45, 47–52 complete at n=1000 (10 seeds) and 42–44, 47–52 at $n \in \{5000, 10000\}$ (9 seeds each); T=48: seeds 48–51 at n=1000 (4 seeds). Aggregation merges complete seeds only; incomplete runs are discarded. Paired schedule comparisons use per-seed differences within the same base seed; t-statistics are across seeds ($\text{df} = \text{seeds} - 1$).

Hardware. A single RTX 5090 GPU workstation; the full E10 suite (3 corpus sizes $\times$ 6 schedules $\times$ 9–10 seeds at T=16, plus the T=48 extension) is approximately 200 GPU-hours. Code, per-seed metric arrays, and the aggregation pipeline will be released upon publication.