Diversity, Veracity, Independence: Diagnosing and Reversing Model Collapse

Anonymous authors
Paper under double-blind review

Abstract

Training a generative model on data produced by earlier models degrades it—*model collapse*. The field frames this as telling synthetic data apart from real data, but that distinction is ill-defined and undetectable at scale. We instead treat collapse as a loss of **distributional integrity**, measured on three provenance-free axes: **Diversity** (are the outputs still varied?), **Veracity** (are they still valid?), and **Independence** (does each new sample carry fresh information, or merely echo earlier ones?). A single error law ties collapse to all three axes at once, and the data-processing inequality forces the *Independence* term to shrink under repeated self-training. We prove this in tractable (discrete and Gaussian) settings and show it offers a unifying lens on several prior results—casting *Strong Model Collapse*'s synthetic-data residual as a Veracity (distribution-mismatch) effect, complementary to the distinct Independence residual our own decomposition isolates. Building on this, we give a theory of *recovering* a collapsed model, by analogy with database recovery: a model can be restored only against a durable, append-only log of real data, and there is a sharp **point of no return** once lost content leaves that log. We provide a model-agnostic metrics suite and validate the framework on recursive loops in text, image, and audio, as well as on the original model-collapse authors' own Gaussian-mixture and OPT-125M setups. Two predictions from prior work that our diversity sweep first appeared to miss—the golden-ratio optimum (recovered as a self-consistency check of the corrected weighted-estimator statement) and recovery hysteresis—are borne out once tested against the correct knob and objective: the optimal real-data *weight* is $1/\varphi$ at matched sample sizes, and a continuous recursion exhibits a recovery-hysteresis loop that persists as the ramp slows. The provable core is linear/discrete; deep-network behavior is stated as falsifiable conjectures and measured.

1 Introduction

More and more training data is itself produced by generative models. When a model is trained on its own outputs, generation after generation, it degrades—a phenomenon called *model collapse*: rare cases disappear first, and the model's outputs grow steadily narrower and less faithful to reality (Shumailov et al., 2023; 2024; Alemohammad et al., 2024). This happens even when only a small fraction of the training data is model-generated (Dohmatob et al., 2025), so it is a practical concern for any model trained on today's web.

The field usually frames the problem as telling *synthetic* data apart from *real* data. We argue this is the wrong question. There is no agreed definition of what counts as collapse (Schaeffer et al., 2025), detecting whether a given text or image came from a model is unsolved at web scale, and watermarks are easily removed. A defense that first has to label each datum's origin cannot be built.

We propose to ask about *integrity* instead. A database never asks whether a row is "real"; it asks whether the row breaks a constraint. In the same spirit, we treat collapse as a silent violation of three measurable properties of a distribution—no provenance labels required (Table 1 lists all notation): **Diversity** (are the outputs still varied, or collapsing to a few repeated modes?), **Veracity** (is each output a *valid* instance the real process could have produced—on the real support—rather than an artifact or fabrication? not necessarily

factual truth, and not closeness to any particular training example), and **Independence** (does each new sample carry fresh information, or is it an echo of samples the model has already produced?).

These three are not separate knobs: they combine in one error law for the expected test error E_t, $E_t \approx \overline{V}_t + \kappa_t / N_{\mathrm{eff},t}$, whose terms are exactly Veracity (the bias $\overline{V}_t$, the Veracity *deficit*, which grows as veracity falls), Diversity (the variance budget κ_t—a Diversity quantity because it is the spread the model can still express, which shrinks as modes collapse), and Independence (the effective number of independent samples $N_{\mathrm{eff},t}$). So two datasets can share the same N_{eff} yet collapse at different speeds if their diversity or truthfulness differ—no single axis tells the whole story. Independence is the term that recursion *must* erode (a data-processing inequality), which is why it recurs across several previously separate analyses; but it is only part of the picture. Most importantly, this integrity view lets us ask a question prior work leaves almost untouched: once a model has collapsed, *can it be recovered, and at what cost?*

Contributions.

- **(C1) We reframe model collapse as a loss of *integrity*, not of provenance.** Instead of asking whether data is real or model-generated, we measure a distribution on three provenance-free axes—Diversity, Veracity, Independence (DVI)—and define collapse as their degradation. A veracity-conditioned diversity measure ($D_{|V}$) then separates a genuinely diverse model from a *confabulator*—a model that fluently produces many *different but false* outputs, so it scores high on Diversity yet low on Veracity—which the raw axes, taken separately, cannot.

- **(C2) We give a single error law that ties collapse to all three axes.** The law $E \approx \overline{V} + \kappa / N_{\mathrm{eff}}$ has one term per axis: the bias $\overline{V}$ is the Veracity deficit, the variance budget κ is Diversity, and the effective number of independent samples N_{eff} is Independence. Two datasets with the *same* N_{eff} but different diversity or truthfulness therefore collapse at different rates—no single axis explains collapse. We prove (in discrete and Gaussian settings) that recursion drives N_{eff} down; this unifies "accumulate vs. replace," clarifies that the golden-ratio result is an error-minimizing real-data *weight* at matched sample sizes (not a mixing fraction, and outside our discrete diversity regime; §4), and places Strong Model Collapse's synthetic-data residual on the *Veracity* axis (a distribution-mismatch term), complementary to the *Independence* deficit our own decomposition contributes. Empirically N_{eff} is a real but modest driver ($\sim$2–3% of the variance), confirming Diversity and Veracity as co-causes.

- **(C3) We develop a theory of *recovering* a collapsed model, by analogy to database recovery.** Recovery is possible only against a durable, append-only log of real data (a "write-ahead log of reality"): each axis has its own recovery law, there is a sharp *point of no return* (once lost content leaves the log, no amount of new data restores it), and an admission gate blocks corrupt updates. The guarantee is statistical—it restores distributional properties, not exact weights: *you can recover exactly what the durable log still covers.*

- **(C4) We release a model-agnostic DVI metrics suite**[1] (an integrity monitor and a recovery controller) with concrete estimators and built-in guards against common measurement pitfalls, and show it detects collapse not only in our testbeds but on the GMM and OPT-125M recursive setups of Shumailov et al. (2024) (§7.4).

- **(C5) We validate the framework across three modalities** from one codebase—text, image+text, and image+audio+text recursive loops. Two prior-work predictions our diversity sweep first appeared to miss are *reproduced* once tested against the correct knob and objective (§7): the golden-ratio optimal real-data *weight* $1/\varphi$ at matched sample sizes, and a recovery-hysteresis loop in a continuous recursion whose area persists as the ramp slows.

[1] Anonymized code and configurations: `https://anonymous.4open.science/r/collapse-anon`. De-anonymized on acceptance.

Table 1: Notation. V is Veracity and $\overline{V}$ its deficit; Diversity is D, and the diversity/variance budget in the error law is the distinct symbol κ (kappa); N_{eff} has one definition; $D_{|V}$ is never conflated with D.

Symbol	Meaning
t	Generation index (recursion step); p_t is the model's distribution at generation t, with $p_0 = p^\star$
E_t	Expected test error at generation t (the quantity the error law decomposes)
ϵ	Injection fraction: the fresh-real share of each pool ($= \alpha$; the persistent-excitation knob)
w	Real-data *weight* in the weighted estimator of Prop. 1 (distinct from the mixing *fraction* α/ϵ); optimal at $1/\varphi$ when $n{=}m$
D	Diversity (variety / breadth of outputs: mode count, coverage, distinct-n, Vendi)
V	Veracity (validity: fraction of outputs on the true support $p^\star$, $\Pr_{x\sim p}[x \in \operatorname{supp} p^\star]$; not factual truth per se)
$I \equiv N_{\text{eff}}$	Independence = effective independent sample size = effective rank of the cross-sample covariance; $N_{\text{eff}} = N/(1+(N-1)\rho)$ for mean correlation ρ
$D_{\|V}$	Veracity-conditioned diversity (first-class; $D - D_{\|V}$ diagnoses a confabulator)
$\overline{V}_t$	Veracity deficit (bias term in the error law $E_t \approx \overline{V}_t + \kappa_t/N_{\text{eff},t}$)
κ_t	Diversity/variance budget in the error law (a *distinct* symbol from the Diversity functional D)
ζ_t	Independence deficit in *our* error law, $\kappa_t(N - N_{\text{eff},t})/(N N_{\text{eff},t})$ (vanishes iff $\rho_t{=}0$); distinct from Strong Collapse's mismatch residual (App. B)
$\mathcal{M}(p;\alpha,\beta,\gamma)$	Mixing operator: fresh-real (α), accumulated (β), synthetic (γ)
V_{ver}	Verifier / grounding oracle (distinct from the Veracity functional V)
C	Cost-of-reversal (fresh independent veracious samples to restore integrity)

2 Related Work

Phenomenon and linear theory. Shumailov et al. (2024) named collapse and gave three error sources (statistical approximation, expressivity, functional approximation); Alemohammad et al. (2024) framed it as model autophagy. A regression/free-probability line characterizes it sharply: Dohmatob et al. (2024a) give closed-form high-dimensional error linear in iterations; Dohmatob et al. (2024b) cast it as a change of scaling laws; Dohmatob et al. (2025) prove it persists at synthetic fraction 1/1000, isolating a residual ζ; Seddik et al. (2024) show tail tokens truncate first; Bertrand et al. (2024) require a good initial model and a large-enough real fraction; Barzilai & Shamir (2025) and Borkar (2025) delineate when collapse is avoidable.

DVI as a unifying lens. Much of this literature can be re-read in one vocabulary (the unifying-view table, App. D in the supplementary material): each phenomenon or mitigation names, or acts on, a specific DVI axis, and several prior results can be re-read on specific DVI axes—for instance, Strong Collapse's synthetic-data residual falls on the *Veracity* axis (it is a distribution-mismatch term), complementary to the *Independence* residual our own decomposition isolates (App. B). We do *not* claim total unification: the scaling-law characterizations (Dohmatob et al., 2024b) sit largely outside the decomposition, Veracity is undefined in open-ended domains, and the axes are not orthogonal (§8). Concretely, mitigations each target a *dominant* DVI axis (some, e.g. curation, act on two); our framework names the axis each primarily moves and exposes the coupling the single-axis view hides.

Position and frontier. Schaeffer et al. (2025) show the Gerstgrasser–Dohmatob "contradiction" is a workflow difference (accumulate vs. replace)—in our terms, opposite sides of the $\beta > 0$ independence boundary. Borji (2024) argues collapse is a generic KDE inevitability. The 2026 frontier moves to networks and structure: interaction-graph topology (Wu et al., 2026), web-scale network dynamics (Wang et al., 2025), spectral detection (Gu et al., 2026), knowledge collapse (Keisha et al., 2025), and unlearning-as-collapse (Scholten et al., 2026). Our N_{eff} (spectral effective rank) is a detection signal in the same family as SIGMA's (Gu et al., 2026) but is embedded in a recovery theory rather than used only diagnostically; "knowledge collapse" (Keisha et al., 2025) (fluency survives, facts fail) is, in our terms, Veracity collapse with Diversity intact—the

degenerate-couple regime we isolate in App. A. Of the eight research gaps in this literature,[2] we address detection-as-integrity (#2,#4), the multimodal regime (#8), and—least crowded—**recovery** (#7).

3 The Distributional-Integrity (DVI) Framework

We call the framework *Distributional Integrity* by analogy with *data integrity* in databases. A database has integrity when its contents satisfy a set of validity constraints; a trained model has distributional integrity when its *output distribution* still satisfies analogous constraints—it stays varied, truthful, and built from independent information. Model collapse is the gradual loss of this integrity, and because it violates no explicit rule and throws no error, we describe it as silent corruption of the learned distribution.

A recursive training loop is a Markov chain on distributions: $p_{t+1} = \mathsf{L} \circ \mathsf{G}_N(p_t)$, where G_N draws N samples and L refits. The mixing operator $\mathcal{M}(p_t; \alpha, \beta, \gamma)$ blends fresh-real (α), accumulated (β), and synthetic (γ) mass; *replace* (α=β=0), *fixed-fraction*, and *accumulate* (β>0) are its three regimes.[3]

Let p be a distribution over an output space $\mathcal{X}$, optionally equipped with a similarity kernel $k(\cdot, \cdot)$ (the identity kernel for discrete tokens or triples), and let $p^\star$ be the real distribution. We measure three properties of p.

Definition 1 (Diversity)**.** Diversity $D(p)$ is the *effective support size*—the number of distinct outcomes over which p effectively spreads its mass. For discrete p it is a Hill number of order q, $D_q(p) = \left(\sum_x p(x)^q\right)^{1/(1-q)}$: order 0 is the mode count (richness) and order 1 is $\exp\left(-\sum_x p(x) \log p(x)\right)$, the exponential of the Shannon entropy. For structured or continuous spaces we use its similarity-aware form, the Vendi score (Friedman & Dieng, 2023) $D(p) = \exp\left(-\sum_j \lambda_j \log \lambda_j\right)$, where $\{\lambda_j\}$ are the eigenvalues of the normalized Gram matrix K/n with $K_{ij} = k(x_i, x_j)$. Mode count, distinct-n, coverage entropy, and Vendi are thus one functional at different orders or kernels, and collapse drives $D \to 1$ (a single effective mode).

Definition 2 (Veracity)**.** Veracity $V(p) = \Pr_{x \sim p}\left[x \in \operatorname{supp} p^\star\right]$ is the probability mass that p places on *valid* states, i.e. inside the support of the real distribution $p^\star$. It measures *validity*, not factual truth in a strong sense and not proximity to any particular training example: a novel sample that a real process could have produced is veracious, whereas a plausible-looking fabrication is not. When no ground-truth oracle is available (open-ended text), V cannot be computed exactly and is replaced by a fidelity proxy (§8).

Independence is $I \equiv N_{\text{eff}}$, defined in §4. The three axes are *not* orthogonal: a confabulator maximizes D while destroying V. We therefore add a fourth functional (Def. 3).

Definition 3 (veracity-conditioned diversity $D_{|V}$)**.** Let $v(x) \in [0, 1]$ be a per-sample veracity score, $p^V(x) \propto v(x)p(x)$. Then $D_{|V}(p) = D(p^V)$ (weighted, canonical), or $D^\tau_{|V}(p) = D(p \mid v \geq \tau)$ (thresholded, robustness). A confabulator is diagnosed by $D - D_{|V}$ growing while D stays flat—the collapse of *grounded* diversity.

The integrity bridge. The three axes mirror the three classical kinds of database integrity constraint, which is what motivates the framing. A *domain* constraint requires every stored value to be legal for its column (for instance, an age must lie in $[0, 150]$); the learning analogue is **Veracity**, since every generated sample should be a valid state. *Entity* integrity requires the database to keep a distinct row for each real entity it represents; the analogue is **Diversity**, since the model should still cover the modes of the real distribution rather than merge or drop them. A *referential* constraint requires a value in one table to point to a valid key in another, keeping the tables mutually consistent; the analogue is **Independence** and, in the multimodal case, the correspondence between channels. Collapse violates these constraints silently—it raises no error—which is why we call it silent corruption. Formally, a model satisfies a *DVI integrity constraint* $\tau = (\tau_D, \tau_V, \tau_I)$ when $D \geq \tau_D$, $V \geq \tau_V$, and $I \geq \tau_I$; these thresholds are what the recovery laws of §5 restore.

[2] Following the related-work taxonomy: (1) nonlinear theory, (2) no agreed collapse metric, (3) oracle-dependent solutions, (4) detection/provenance unsolved, (5) real-world prevalence unmeasured, (6) regime-specific mixing ratios, (7) recovery/reversibility, (8) multimodal/agentic ecosystems.

[3] Our (α, β, γ) assign α to fresh-real mass, so pure recursion is α=β=0. This relabels the coefficients of Shumailov et al. (2024), whose pure-recursion case is β_i=γ_i=0 (per their Author Correction (Shumailov et al., 2025)); the regimes coincide, only the labeling differs.

4 The Tri-Axis Error Law and the N_{eff}/DPI Independence Term

The collapse rate is governed by the error law $E_t \approx \overline{V}_t + \kappa_t/N_{\text{eff},t}$ of Eq. (1), in which all three DVI axes appear: the bias $\overline{V}_t$ is the *Veracity* deficit, the variance budget κ_t is the *Diversity* term, and $N_{\text{eff},t}$ is *Independence*. No single axis is sufficient—two pools at equal N_{eff} but different D, V collapse at different rates (the Phase-2 real-but-shuffled-caption arm collapses grounding with N_{eff} intact; §7). N_{eff} is the axis whose decay the data-processing inequality *forces*—the only provably-monotone one; but it is not the largest effect (empirically a partial $R^2 \approx 0.028$ of tail-extinction, §7), and Diversity and Veracity are the larger co-causes.

One definition. We take $N_{\text{eff}} = \text{erank}(\Sigma)$, the effective rank of the cross-sample covariance Σ, as the primary, measured quantity. For interpretability we summarize it by a single mean pairwise correlation ρ through $N_{\text{eff}} = N/(1+(N-1)\rho)$; thus $\rho = (N - N_{\text{eff}})/\big((N-1)N_{\text{eff}}\big)$ is *derived* from the measured erank, not the reverse—there is no circularity. Limits: $\rho{=}0 \Rightarrow N_{\text{eff}}{=}N$ (i.i.d.); $\rho{\to}1 \Rightarrow N_{\text{eff}}{\to}1$ (one effective sample). The empirical erank is bounded by $\min(n, d)$ and preserves *relative* decay—the quantity the independence term uses—even when it under-reads absolute magnitude. (These are standard identities, not a new result.)

Error decomposition. Augmenting the clean bias–variance form with an independence residual, we write the test error after t self-training rounds as

$$E_t \;\approx\; \overline{V}_t \;+\; \frac{\kappa_t}{N} \;+\; \zeta_t, \qquad \zeta_t = \frac{\kappa_t\,(N - N_{\text{eff},t})}{N\,N_{\text{eff},t}} \;\geq 0, \tag{1}$$

where κ_t/N is the naive (raw-count) variance and $\zeta_t = 0$ iff $\rho_t = 0$: the residual ζ_t is the surplus variance incurred when N correlated samples are counted as if independent—an *Independence* deficit. Here $\kappa_t/N + \zeta_t = \kappa_t/N_{\text{eff},t}$, so the three-term form expands the two-term law $E_t \approx \overline{V}_t + \kappa_t/N_{\text{eff},t}$ into naive variance plus independence surplus. This is *distinct* from the residual of *Strong Model Collapse* (Dohmatob et al., 2025), whose i.i.d. two-sample analysis instead isolates a distribution-*mismatch* term (a *Veracity* effect, nonzero even at $\rho{=}0$); the two are complementary DVI-axis pathways rather than the same object (App. B). For deep nonlinear learners our decomposition is a conjecture we verify empirically.

Starting hypothesis (the injection fraction ϵ). We parameterize the loop by the fraction $\epsilon \in [0, 1]$ of each pool that is *fresh, independent, on-support real data* (ϵ is the fresh-real weight α of $\mathcal{M}$). This is the persistent-excitation knob: at $\epsilon{=}0$ the loop is *closed* (pure replace), p_{t+1} is a function of p_t alone, and the data-processing inequality applies; at $\epsilon{>}0$ a fresh edge $p^\star \to p_{t+1}$ reopens the loop each generation, so the DPI no longer binds. Accordingly $N_{\text{eff},t} \to N_{\text{eff}}^{\infty}(\epsilon)$ with $N_{\text{eff}}^{\infty}(0){=}1$ in the limit (collapse; finite-budget runs floor above it, §7) and $N_{\text{eff}}^{\infty}(\epsilon)$ increasing in ϵ (bounded). Two caveats keep this from being white-noise dithering: the injected ϵ must carry *Veracity* (on-support) and *Diversity—plausible-but-false* injection raises N_{eff} while destroying V (a confabulator: it is learnable, so the fitter absorbs it), whereas off-support noise is discarded by the fitter and merely wastes the budget (§7)—and the bulk floor does not by itself save the *tail*: a rare mode at probability p_{rare} is repopulated only if $\epsilon N \gtrsim 1/p_{\text{rare}}$ within the horizon, else the point of no return persists (§5). Theorem 1 is the $\epsilon{=}0$ boundary; the $\epsilon{>}0$ regime is Thm. 2.

Theorem 1 (Independence collapse under pure recursion; discrete-MLE and Gaussian settings)**.** *Under replace ($\alpha{=}\beta{=}0$), the chain $p^\star \to p_1 \to p_2 \to \cdots$ post-processes p_t at every step, so the independent information about $p^\star$ is non-increasing: $I(p^\star; p_{t+1}) \leq I(p^\star; p_t)$ (data-processing inequality). In the* discrete-MLE *setting the mode count k_t is non-increasing and $k_t \to 1$ a.s.; in the* Gaussian *setting $\mathbb{E}[\det \Sigma_{t+1}] \leq \det \Sigma_t$ and $\Sigma_t \to 0$ a.s. Hence $N_{\text{eff},t} \to 1$ in both settings.*

Discrete argument. For finite $\mathcal{X}$ with empirical-MLE refit and multinomial resampling, $\{p_t\}$ is a Wright–Fisher chain: extinction is absorbing (a token not drawn cannot be resampled), so mode count is non-increasing and the chain is absorbed a.s. in a single-atom state $\delta_{x^\star}$. *Gaussian argument.* Refitting the covariance of N draws makes $\det \Sigma_t$ a supermartingale ($\mathbb{E}[\det \Sigma_{t+1}] \leq \det \Sigma_t$ by concavity of $\det^{1/d}$ plus MLE shrinkage), so $\Sigma_t \to 0$ a.s. and the mean random-walks (Xu et al., 2025), accumulating veracity deficit $\overline{V}_t \propto t$. *Scope.* The DPI bounds the between-generation quantity $I(p^\star; p_t)$; it does *not* by itself force step-wise monotonicity of the within-generation $\text{erank}(\Sigma_t)$—redistributing mass across surviving modes can transiently *raise* erank. What

is monotone is the mode count (discrete) and the expected log-volume $\det \Sigma_t$ (Gaussian); $N_{\text{eff},t}$'s decrease is the asymptotic consequence ($N_{\text{eff},t} \to 1$), whose empirical (near-monotone) trajectory the experiments track.

Theorem 2 (bounded error with a fresh source; linear setting)**.** *With $\beta > 0$ and a never-overwritten pool the count of independent real draws is non-decreasing, so $N_{\text{eff},t} \uparrow$ and the variance term tends to its irreducible floor; with a fixed real fraction the recursion has a finite-error fixed point. (Proof sketch in App. C; the general/deep case is conjectural.)*

Corollary 1 (replace-vs-accumulate)**.** *Replace $\Rightarrow N_{\text{eff}} \to 1$, error diverges (Thm. 1); accumulate $\Rightarrow$ error bounded (Thm. 2).* ***Accumulation is formally an independence intervention****, not a veracity or diversity fix—which is why Gerstgrasser et al. (2024) and Dohmatob et al. (2025) "differ by workflow, not contradiction" (Schaeffer et al., 2025).*

Proposition 1 (golden-ratio weighting; restated from prior work, regression regime)**.** *He et al. (2025) prove that in a* weighted *objective, $w \cdot$ (real loss) $+ (1-w) \cdot$ (synthetic loss), the optimal* weight *on real data is $w^\star = 1/\varphi \approx 0.618$* when the real and synthetic sample sizes are equal *(n=m; the k=1 case of $w^\star = (\sqrt{k^2+4k} - k)/2$), minimizing the asymptotic estimation error $\lim_t \mathbb{E}\|\hat{\theta}_t - \theta^\star\|^2$ (Gaussian/GLM/nonparametric; the overparameterized extension (Garg et al., 2025) carries $w^\star = 1/\varphi$ to min-norm interpolation and ridge). Collapse occurs as $w \to 0$ and is averted for w above $\approx$1/3. Crucially, $1/\varphi$ is an error-minimizing* weight at matched sample sizes, not *a diversity-maximizing mixing fraction. We restate it and do not re-derive it.*

Scope (Phase-1 neural). Our sweep varies the mixing *fraction* α and reads *diversity* (mode count), which is monotone in the real fraction and *saturates* (final mode-count 11 at α=0 $\to$ $\sim$300 at $\alpha\approx$0.38 $\to$ plateau $\sim$330–363 for $\alpha \in [0.5, 1]$); collapse is averted above a *threshold* ($\alpha\approx$0.4), with no interior optimum. This does *not* test Prop. 1: $1/\varphi$ is the optimal real-data *weight* (at n=m) minimizing *estimation error*, whereas a diversity/mode-count objective is monotone in the real fraction and has no interior optimum *by construction*. Tested against the correct knob (weight) and objective (MSE), the optimum *is* reproduced at $w^\star = 1/\varphi$ (§7, Fig. 3A).

5 Recovery Theory: Collapse as a Database-Recovery Problem

We organize recovery through a database-recovery analogy. The results below are structural—an internal Markov map cannot add exogenous support or independence—so the contribution is the *framework* (per-axis recovery laws, a point-of-no-return, and an admission gate), not the difficulty of any single proof. For readers outside databases: a **write-ahead log (WAL)** is an append-only record of every real datum, a **checkpoint** is a saved model state, **rollback** reverts to a checkpoint, **roll-forward** replays the log to rebuild lost content, **point-in-time recovery (PITR)** restores the model's distribution to an earlier logged state, and a **two-phase-commit** admission gate validates a candidate update before committing it. We map these to recovery results on the DVI axes.

Theorem 3 (Independence is not self-healing)**.** *No internal map R raises N_{eff}: independence is endogenously irreversible and recoverable only by an exogenous independent source.*

Theorem 4 (Diversity recovery is mode-gated; discrete setting)**.** *A lost mode is recoverable iff it survives in the residual/durable source: you can restore only modes the log still covers.*

Theorem 5 (Veracity recovery needs a grounding signal)**.** *With a verifier V_{ver} independent of how p_T was produced, bias is correctable without the original data by rejecting samples that fail V_{ver}. With error rate δ the residual bias is $O(\delta)$ and gains plateau once δ dominates (Feng et al., 2024; Yi et al., 2025). The perfect-verifier case is exact; the $O(\delta)$ rate is a proof sketch (App. C).*

Definition 4 (cost-of-reversal C)**.** The minimal number of fresh, independent, veracious samples needed to lift a degraded p_T back above an integrity threshold $\tau = (\tau_D, \tau_V, \tau_I)$.

Theorem 6 (point of no return; discrete/absorbing setting)**.** *Once a mode is extinct and absent from the durable log, no budget restores it: $C = \infty$. Recovery is possible iff a sufficient checkpoint/log lies within the recoverable horizon.*

Conjecture 1 (recovery hysteresis)**.** *Recovery is path-dependent: the diversity recovered on the up-path lags its loss on the down-path, so the recovery loop encloses positive area. Empirically supported in the*

continuous-Gaussian regime—the loop area does not decay as the ramp slows (§7, Fig. 3B)—and unproven for deep nonlinear learners, hence stated as a conjecture; the discrete point of no return is the limiting case in which the down-path never returns.

Theorem 7 (recoverability ≡ logged checkpoint within horizon; discrete/tractable setting)**.** *A collapsed model is recoverable iff the durable log retains, within the horizon, the independent source, the lost modes, and a grounding signal required by each below-threshold axis.*

Remark 1 (the essential disanalogy)**.** Recovery is *statistical, not bit-exact*: PITR restores DVI *properties*, not weights. Recovered models must be checked by task accuracy, not only DVI match. The guarantee is a curated-pipeline result; the open-web commons has no controllable WAL.

6 Metrics Suite

We provide per-axis estimators: D (mode count, coverage entropy, distinct-n, Vendi); V (KG-truth / QA-EM, verifier agreement, hallucination rate); $I = N_{\text{eff}}$ (effective rank of the Gram/covariance, near-duplicate rate, graph-spectral connectivity), *triangulated* by three estimators of the *one* N_{eff}. The $D_{|V}$ estimator (veracity-weighted Vendi, canonical; thresholded restriction as robustness) is the built-in confound control. Fragility flags are stated explicitly: embedder dependence of N_{eff}, verifier circularity for V, and—most important—that I measures *correlation*, not *ancestry* (an open problem, §8).

7 Experiments

All testbeds use a synthetic knowledge graph (E=200 entities, R=20 relations, a Zipfian heavy-tailed set of true triples) rendered into text through templates; because we know the true triples, we can score Veracity exactly. Generators range from simple categorical samplers to a small from-scratch GPT.

Setup and terminology. Each experiment runs a recursive loop and compares four *training regimes*: *replace* (train only on the model's own last-generation outputs—the pure-collapse case), *accumulate* (keep all past real data and add synthetic), *real_only* (keep training on fresh real data—the healthy control), and *mix*(α) (a fraction α of real data). For recovery, we compare *roll-forward* (keep injecting real data from a durable log, as a database replays its write-ahead log), *naive* (a single one-shot injection of real data), and an *endogenous control* (attempt recovery using only the model's own outputs). We read collapse on the DVI axes—Diversity (mode count, and *tail-coverage*: how many of the rare real modes survive), Veracity, and Independence (N_{eff}). Three principles govern every claim: **(L1)** a mediator such as N_{eff} is tested only where the manipulated variable is held fixed, never in a sweep; **(L2)** each claim is measured with the metric that encodes its meaning (e.g. the extinct tail for recovery), not a gameable summary; and **(L3)** a data-leakage audit precedes every run. Phase numbers index modality (Phase 1 text; Phase 2 image+text; Phase 3 image+audio+text). Within these we report labeled experiments: *E1* (text collapse and N_{eff} mediation), *E1-bi* (bimodal per-channel N_{eff}, a negative control), *E2* (bimodal channel dependence), and *E4-bi* (bimodal graded re-grounding); the suffix "-bi" denotes bimodal, and the numbering follows our internal register (we report only these).

7.1 Phase 1 — text (single modality)

Under *replace*, the DVI dashboard degrades monotonically (Fig. 1: N_{eff} 9.0→3.4, mode-count 443→9, tail-coverage→0) while *real_only* stays flat; collapse is detected before any task metric floors. At matched synthetic-fraction, lagged $1/N_{\text{eff}}$ predicts next-step *tail-extinction* (GEE-AR(1); generalized estimating equations with AR(1) working correlation; p=2.7×10^{-14}, partial R^2≈0.028): N_{eff} is a *real but modest* mediator of the diversity/tail axis; the strong "out-predicts synthetic-fraction" form is correctly refuted in a sweep (L1). At matched synthetic-fraction = 1, *replace* collapses to mode-count 9 while accumulate holds at 267—accumulation is an *independence* intervention.

Reversibility (the headline). On **tail-coverage** (the theorem's content, L2; Fig. 2): sustained exogenous roll-forward recovers the extinct tail **20/20**, one-shot naive **0/20**, endogenous control **0/20** (Thms. 3–4); on

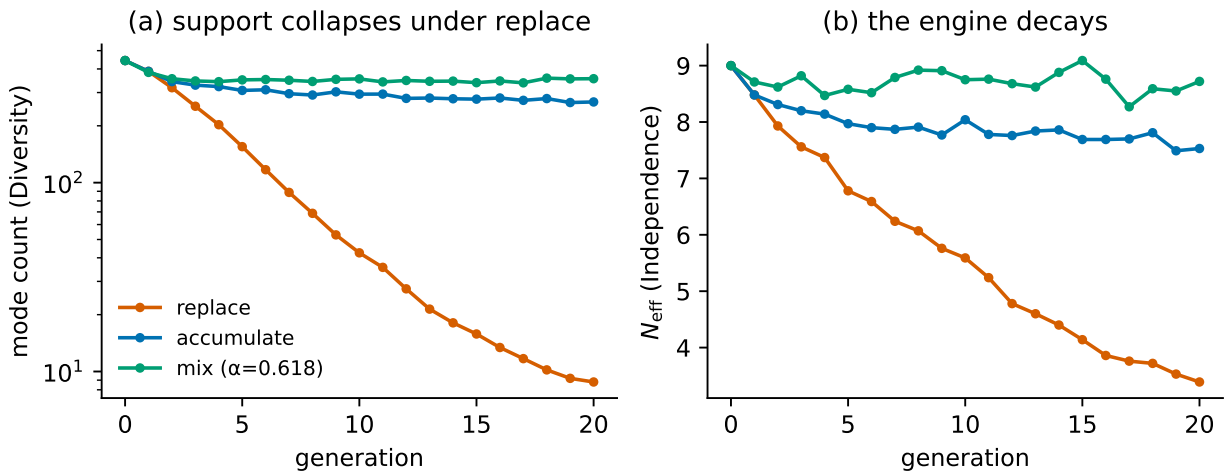

Figure 1: **Phase-1 collapse, 20 seeds.** Under *replace*, Diversity (mode count, log scale) and the N_{eff} term decay monotonically; *accumulate* stays bounded and *mix*(α=0.618) is stable. Detection precedes any task-metric floor.

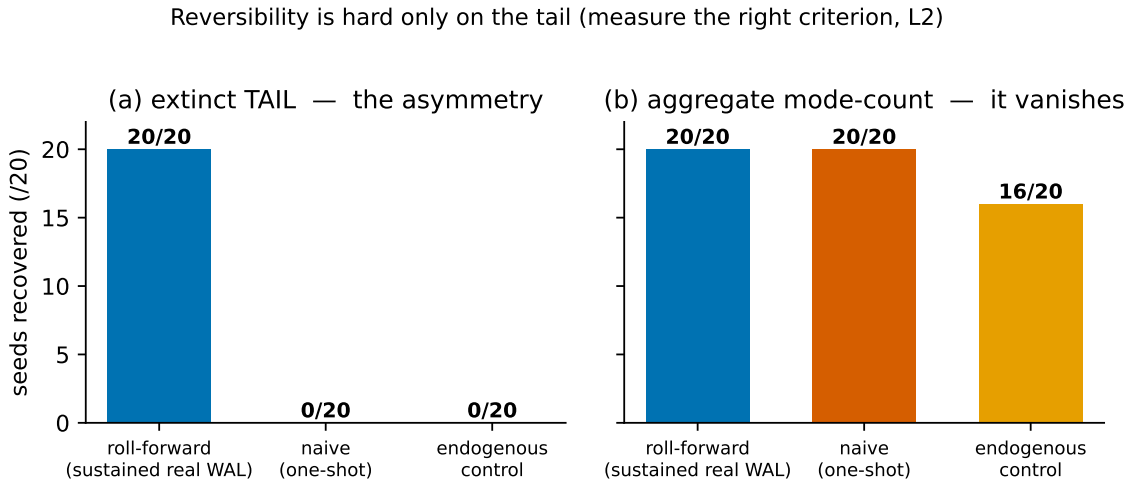

Figure 2: **Reversibility (Thms. 3–4), 20 seeds.** On the extinct *tail* (a), only a sustained exogenous roll-forward recovers (20/20 vs. 0/20); on aggregate *mode-count* (b) the asymmetry vanishes, because the neural model re-expresses common modes endogenously. Measure the criterion the claim is about (L2).

aggregate mode-count the asymmetry vanishes (the neural model re-expresses *common* modes endogenously). Point of no return (Thm. 6): full-WAL recovers the tail **10/10** (C=800); a partial-WAL with the tail dropped recovers **0/10**, C=∞. The two-phase-commit gate holds the pool under replace vs. ungated collapse.

Two phenomena, tested against the correct knob and objective (Fig. 3). Our earlier diversity sweep neither reproduced the golden ratio nor showed hysteresis; both were the wrong measurement, and both *do* appear once tested correctly. **Golden ratio (Eφ).** In the weighted estimator of Prop. 1 at matched sample sizes (n=m), the stationary estimation error is minimized at the real-data *weight* $w^\star = 1/\varphi$: the closed-form minimizer is 0.6180 (matching $w^\star(k)$=$(\sqrt{k^2+4k}-k)/2$ at k=$1, 2, \frac{1}{2}$: 0.618/0.732/0.500) and a 800-seed simulation tracks the closed form to 3.8% (Fig. 3A). The diversity/mode-count objective of our fraction-sweep is monotone in the real fraction and has no interior optimum, which is why the sweep saturated. **Hysteresis (EH).** In a continuous small-N Gaussian recursion, ramping the real fraction ϵ down (1→0,

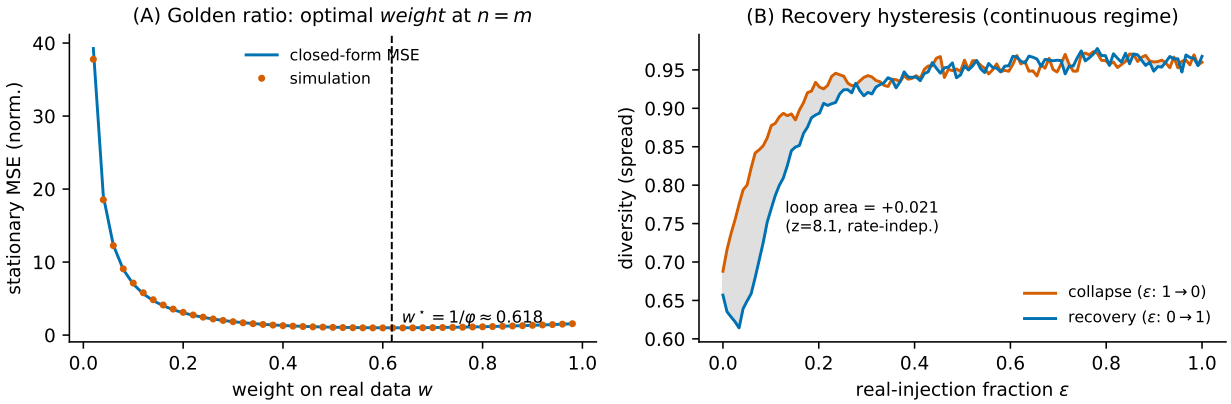

Figure 3: **The two prior-work predictions, tested against the correct knob/objective. (A) Golden ratio:** in a weighted recursive estimator at equal real/synthetic sample sizes (n=m), the stationary estimation error (MSE) is minimized at the real-data *weight* $w^\star = 1/\varphi \approx 0.618$ (vertical line); closed form and an 800-seed simulation agree. This is a *weight*, not a mixing fraction, and it minimizes *error*, not diversity—which is why our diversity fraction-sweep did not show it. **(B) Recovery hysteresis:** in a continuous small-N Gaussian recursion, the diversity (spread) recovered while ramping ϵ up (blue) lags its collapse while ramping ϵ down (orange); the enclosed loop (area +0.021, z=8.1 at R=120; 500 seeds; x-axis ϵ, y-axis diversity/spread) does not decay as the ramp slows (area $\in [0.013, 0.021]$ across R=30–240), i.e. genuine hysteresis rather than transient lag. The discrete point of no return is its limiting case.

collapse) then up (0→1, recovery) traces a loop: the recovery path stays below the collapse path across ϵ, with loop area +0.021±0.003 over 500 seeds (z=8.1; at ramp length R=120). Crucially the area *does not decay as the ramp slows*—it stays in $[0.013, 0.021]$ (all z>6) across R=30, 60, 120, 240 at fixed seeds—so the loop is genuine hysteresis, not the transient lag that would vanish quasi-statically (Fig. 3B). The discrete tail's point of no return (C=∞) is the limiting case in which the down-path never returns (the extinct mode is absorbing: it can never be resampled); a persistent-memory (accumulate) variant does not collapse and shows no loop (area +0.000±0.001), locating hysteresis in the collapsing regime. *Synthesis: the tail is the locus*—collapse is tail-extinction, N_{eff} mediates it, reversibility is hard only on the tail, and the tail dies abruptly.

The injection fraction ϵ governs a phase transition (Eϵ). We test the starting hypothesis of §4 directly, sweeping the fresh-real fraction ϵ, the pool size N, and the injected content (Fig. 4; categorical generator, 20 seeds). **(A) Magnitude.** Steady-state N_{eff}^{∞} increases monotonically in ϵ and is minimized at ϵ=0 (9.65±0.78 → 13.1±0.2): the bounded floor rises with injection, exactly the ϵ>0 regime of Thm. 2, with ϵ=0 the pure-replace boundary of Thm. 1. Full collapse to N_{eff}=1 is asymptotic and not reached in 60 generations at N=2000. **(B) Tail (a negative result).** Extinct-tail coverage rises with the injected volume ϵN, but N has an *independent* effect—at matched ϵN, a larger pool recovers more tail—so the clean "survives on ϵN alone" hypothesis is *disconfirmed*; $\epsilon N \gtrsim 1/p_{\text{rare}}$ remains a *necessary* but not sufficient condition. **(C) Kind is decisive.** At fixed ϵ=0.1, real injection keeps V=1.0; *off-support noise* is unlearnable, so the fitter discards it (V=1.0, but fewer modes—wasted budget); *plausible-but-false* injection is learnable and absorbed, driving V to 0.05 and ∼590 false modes *while* N_{eff} *rises to* 1.18× *gen-0.* This is the confabulator, and it is the sharpest demonstration that N_{eff} (Independence) is blind to Veracity (§8)—injection helps only when it carries Veracity.

7.2 Phase 2 — bimodal (image + text)

The KG is rendered to 32 × 32 glyph images; a frozen CNN oracle decodes images to triples, and a learned captioner is the cross-modal bridge. Grounding is scored against the *known* rendered triple (an L3 leakage

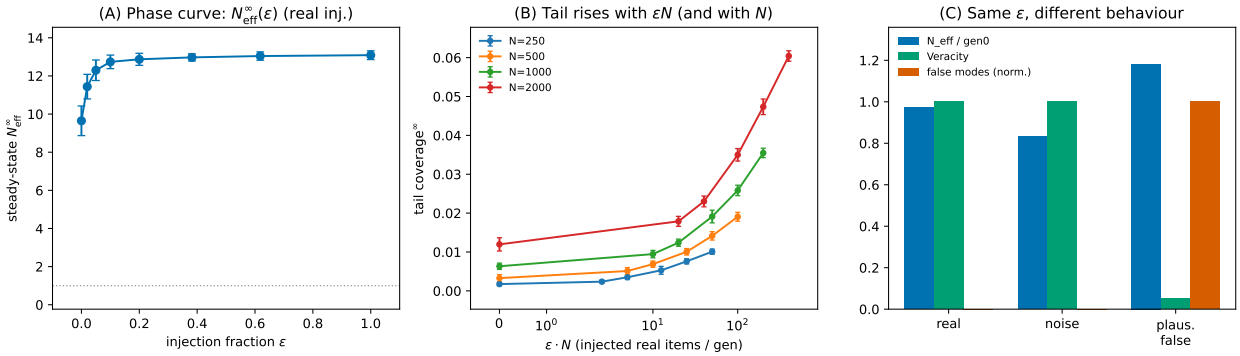

Figure 4: **The injection fraction ϵ (20 seeds, ± 1 s.d.). (A)** Steady-state N_{eff}^{∞} (mean of the last five generations) rises monotonically with real-injection ϵ, from 9.65 at ϵ=0 to 13.1 at ϵ=1 (60 generations); ϵ=0 is the minimum. Collapse to N_{eff}=1 is asymptotic (Thm. 1) and not reached at this finite budget—the dotted line marks N_{eff}=1. **(B)** Extinct-tail coverage rises with the injected volume ϵN *and, independently, with* N: the four pool sizes do *not* collapse onto a single ϵN curve. **(C)** At a fixed ϵ=0.1, the injection *kind* decides the outcome: real data keeps Veracity at 1.0; off-support noise is unlearnable and discarded (Veracity intact, budget wasted, fewer modes); plausible-but-false data is *learnable and absorbed*—N_{eff} looks healthiest (1.18× gen-0) while Veracity collapses to 0.05 and false modes proliferate. Independence is blind to this (§8).

Table 2: Phase-2 dependence arms (E2): grounded correspondence at generation 15 (g_{15}) and within-channel diversity (img_modes), 6 seeds. All arms start near $g_0 \approx 0.39$.

arm	grounded g_{15}	img_modes g_{15}	what the arm isolates
real_only (anchor)	0.41	176	healthy baseline
co_pair (self-consuming)	**0.26**	47	self-consuming collapse
co_shuf (shuffled captions)	**0.030**	45	channel dependence alone
co_indep (independent caption)	**0.075**	42	channel dependence alone
real_shuf (real + shuffled)	**0.12**	178	dependence, no synthetic data or diversity loss

audit found and corrected an oracle-as-judge confound; the oracle agreed with the known triple only ∼0.80 of the time, yet every number moved < 0.05).

Dependence is load-bearing; integrity, not provenance (E2). The decisive contrast between the self-consuming and shuffled-caption arms =0.233 (0.263±0.059 vs. 0.030±0.073, mean±s.d. over 6 seeds; Welch t=6.1, 95% CI $[0.13, 0.33]$ at df= 5, conservative) holds at *identical* marginals and *identical* within-channel diversity collapse (Table 2, Fig. 5). The real-but-shuffled arm—all-real, diverse images with shuffled real captions—crashes grounding to 0.12 with *no* syntheticness and *no* diversity loss. So cross-modal collapse is a **referential-integrity** violation: the channel *dependence*, not the data *provenance*, governs grounding.

N_{eff} is axis-specific (E1-bi, negative; 20 seeds). At matched ratio, lagged per-channel $1/N_{\text{eff}}$ does *not* predict grounding loss (caption channel p=0.07, wrong sign; image p=0.67), even as both grounding and N_{eff} decline. This is a single-arm matched (self-consuming) discriminant null, not a powered two-arm test—the caption-channel N_{eff} and the grounding score both read the caption channel (lagged one generation)—so it *bounds* the mediator's reach rather than proving it. Read with the real-but-shuffled arm (its converse), the two are *consistent with* a within-modality N_{eff} mechanism that governs diversity and does not leak into cross-modal grounding.

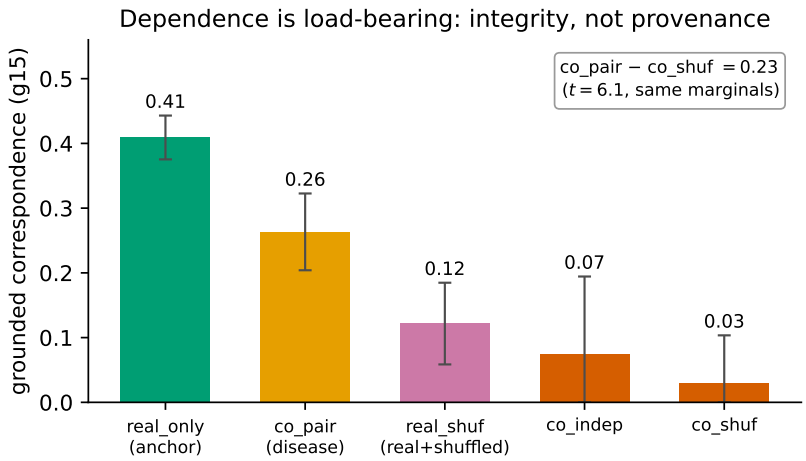

Figure 5: **Phase-2 dependence (E2), 6 seeds; error bars are ± 1 s.d.** The gap between the self-consuming and shuffled-caption arms (0.233; Welch t=6.1, 95% CI $[0.13, 0.33]$, at df= 5) holds at identical marginals; the real-but-shuffled arm (all-real, diverse) still floors grounding. Cross-modal collapse is a referential-integrity violation, not a provenance one.

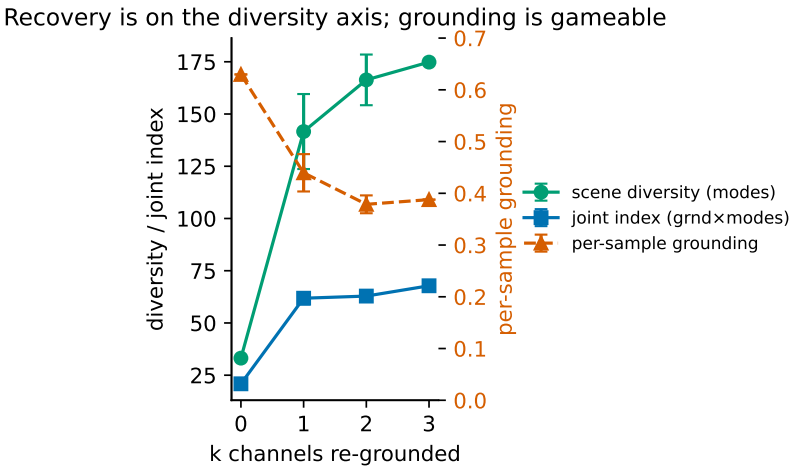

Figure 6: **Phase-3 graded re-grounding, 6 seeds, full 8-subset factorial (mean per k; error bars span the channel subsets at each k).** Diversity rises monotonically with k while per-sample grounding *falls*; the joint index (grounded×modes) is monotone but *saturates*—a quorum effect (the first re-grounded channel does the work).

Graded re-grounding is channel-asymmetric, and grounding is gameable (E4-bi). Re-grounding the caption channel (k=1) *overshoots* the both-real anchor (k=2) on grounding—because the image channel has collapsed to 47 modes and the captioner overfits them. Only k=2 restores grounding *and* diversity. Per-sample grounding is therefore gameable by mode-collapse and must be read jointly with diversity (the Phase-2 echo of L2).

7.3 Phase 3 — trimodal (image + audio + text)

A third deterministic channel—a spectrogram-style "audio" render of the same triple (verified label-injective)—is added; image/audio are read by fallible conv captioners, text by an exact parser. We re-ground k of the 3 channels from the real WAL (full 8-subset factorial, 6 seeds).

Recovery lives on the diversity axis; per-sample grounding is anti-monotone (Fig. 6). Scene diversity rises monotonically with k (mode count $33 \to 142 \to 166 \to 175$), while per-sample grounding *falls* ($0.63 \to 0.44 \to 0.38 \to 0.39$): k=0 collapses hardest yet scores highest grounding by overfitting its tiny survivor set. The joint index (grounded×modes) is monotone but *saturates* ($21 \to 62 \to 63 \to 68$)—the *first* re-grounded channel does the work; re-grounding is a **threshold/quorum** effect, not a linear dose (the multimodal echo of the Phase-1 α-sweep saturation). The k=2→k=3 increments fall within the seed s.d., so we read the quorum qualitatively, not as a calibrated dose curve. Per-channel asymmetry is modest (image is the cheapest single channel; the text channel uses an exact parser—a known-easy reference whose low decode error is partly by construction); the two fallible perceptual channels align worst with each other, as independent captioner errors compound.

Continuous-embedding extension (App. A). In the discrete render, false triples were alien to the model, so the raw↑/grounded↓ degenerate-couple paradox could not arise. Once false content is made plausible—in a continuous block-embedding study (8 seeds)—the paradox appears: raw alignment stays high while grounding floors (raw−grounded≈+0.58), the mechanism is the captioner's representational reach (a manifold-pinned captioner instead *decouples*, reproducing Phase-2), and raw alignment is *blind to grounding* (identical 0.575 for grounded-true and ungrounded-false content). This uses a *constructed* block embedding rather than a pretrained encoder; a real-frozen-embedder (CLIP/BLIP) confirmation is stated as optional and unrun.

7.4 External validity — the DVI dashboard on prior collapse setups

To test that the metrics suite (C4) generalizes beyond our testbeds, we instrument two prior recursive-collapse setups (Shumailov et al., 2024; 2023) with the DVI dashboard. We reuse their models, data, recursive protocol, and evaluation faithfully; the training loop is a clean reimplementation of their protocol, not their exact published code.

Recursive GMM (their toy; real veracity oracle, 10 seeds). In a faithful reproduction of their fit→sample→refit loop, all three axes track collapse alongside *their own* L_2-to-gen-0 diagnostic: Diversity (spread) $0.93 \to 0.39$, N_{eff} $9.7 \to 4.7$, and Veracity—scored against the *known* generating distribution, a real oracle—degrades monotonically (W_1 to truth $0 \to 0.55$).

Recursive OPT-125M on wikitext-2 (their headline LM; 5 seeds per arm, 7 generations). On this LM setup, the DVI dashboard tracks collapse on the *replace* (self-consuming) arm while a *real_only* control stays flat (Fig. 7; mean±s.d. over 5 seeds, ddof= 1): real-test perplexity $53\pm0.7 \to 326\pm38$, N_{eff} $278\pm1.7 \to 134\pm16$, distinct-2 $0.55 \to 0.06\pm0.013$, and **rare-token tail-coverage** $0.022 \to 0.0005\pm0.00015$—directly measuring Shumailov et al.'s "the tails of the original content distribution disappear," with the *real_only* control flat on every axis (perplexity $53 \to 53\pm0.6$, tail-coverage $0.022 \to 0.022$). The replace–control gap exceeds the seed spread on all four axes. Veracity is a *fidelity proxy* here (perplexity on a held-out real split; open text has no truth oracle), and N_{eff} uses a *frozen* external encoder (not the collapsing model's drifting embeddings)—a drift-leak guard (L3). Distinct-2, tail-coverage, and N_{eff} are all read from the generated corpus, so they co-move partly by construction; real-test perplexity (Veracity) is the one axis anchored on held-out real data, and it moves independently.

8 Discussion and Limitations

We state the open risks plainly. **(1) Non-orthogonality of** D, V, I: the marginal functionals cannot separate a confabulator from a healthy diverse model; $D_{|V}$ and a joint functional are the partial answer, but the cross-terms in Eq. (1) are unquantified [OPEN]. **(2) Independence readmits provenance**: I = erank measures correlation, not the *artificial* correlation of shared ancestry; separating the two without provenance labels is unresolved [OPEN]. **(3) Veracity is undefined in open-ended domains**: V_{ver} degrades to plausibility/consistency, weakening Thm. 5 [OPEN]. **(4) Proof gap**: Thms. 1,3,4,5 (perfect verifier), 7 and Cor. 1 are theorems in the discrete-MLE/Gaussian settings; Thm. 2, Prop. 1, Thm. 6 (general), and the hysteresis conjecture (Conj. 1) are proof sketches or conjectures for deep nonlinear learners—bridging to deep practice is the principal open problem. **(5) Scope**: the *recovery* theory and the multimodal testbeds are

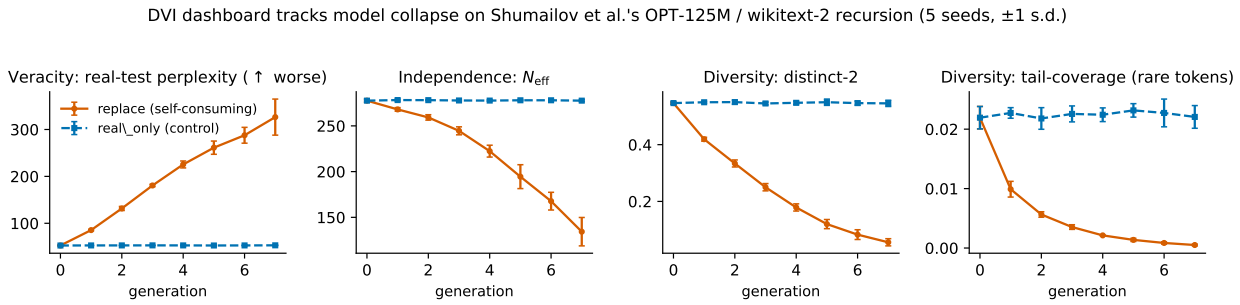

Figure 7: **External validity: the DVI dashboard on an OPT-125M / wikitext-2 recursion** (faithful reimplementation of Shumailov et al. (2024)'s recursive protocol; 5 seeds, ±1 s.d., 7 generations). The self-consuming *replace* arm collapses on every axis while the *real_only* control stays flat—isolating recursion, not drift, as the cause; the arms separate by more than the seed spread throughout.

curated pipelines (text, a controlled bimodal render, a controlled trimodal render with synthetic audio); no web-scale, real-audio, or agentic-ecosystem runs. The *metrics* are additionally validated on prior real recursive setups (GMM, OPT-125M; §7.4), but recovery itself is not yet tested at that scale. Multimodal arms use 6–8 seeds (vs. 20 for text; the external GMM and OPT-125M runs use 10 and 5 seeds respectively); only the headline E2 contrast carries a formal significance test (Welch t, 95% CI), with the remaining multimodal and external claims supported by the reported ±s.d. (for OPT-125M the replace–control gap exceeds the seed spread on all four axes, §7.4). **(6) Oracle circularity**: veracity is scored by the same template parser that defines the ground-truth triples, so V is internally consistent by construction; the Phase-2 L3 audit shows this confound is real and correctable in the controlled setting, but on open data—where no such oracle exists—V degrades to the plausibility proxy of (3). A cross-phase methodological finding: *per-sample grounding is gameable by mode-collapse*, so recovery is read on the diversity/joint axis, never grounding alone.

9 Conclusion

Recursion is a Markov chain on distributions; health is the triple (D, V, I) coupled in one error law $E \approx \overline{V} + \kappa/N_{\text{eff}}$, with N_{eff} its independence driver; and recovery obeys an axis-structured law—provably in the discrete/Gaussian regime, conjectured for deep learners—*you can statistically recover what the durable log still covers*. Shumailov et al. (2024) showed collapse happens; Dohmatob et al. (2025) showed it is unavoidable even at trace synthetic fractions; we add that collapse is a *manageable integrity violation*—measurable along DVI, partially provable through N_{eff}/DPI, and recoverable as far as your write-ahead log of reality reaches. The open-web commons, where that log is no one's to control, is the forward frontier.

Supplementary material. Proof sketches, the derivation relating our Independence residual to Strong Model Collapse's mismatch residual (App. B), the continuous-embedding degenerate-couple study (App. A), and the unifying view of the literature (App. D) are provided in the separate supplementary material.

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. URL https://arxiv.org/abs/2307.01850. arXiv:2307.01850.

Daniel Barzilai and Ohad Shamir. When models don't collapse: On the consistency of iterative MLE. *arXiv preprint arXiv:2505.19046*, 2025. URL https://arxiv.org/abs/2505.19046.

Quentin Bertrand, Avishek Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.00429. arXiv:2310.00429.

Ali Borji. A note on shumailov et al. (2024): AI models collapse when trained on recursively generated data. *arXiv preprint arXiv:2410.12954*, 2024. URL https://arxiv.org/abs/2410.12954.

Vivek S. Borkar. A theoretical basis for model collapse in recursive training. *arXiv preprint arXiv:2506.09401*, 2025. URL https://arxiv.org/abs/2506.09401.

Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model collapse demystified: The case of regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. URL https://arxiv.org/abs/2402.07712. arXiv:2402.07712.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. In *International Conference on Machine Learning (ICML)*, 2024b. URL https://arxiv.org/abs/2402.07043. arXiv:2402.07043.

Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.04840. arXiv:2410.04840.

Yunzhen Feng, Elvis Dohmatob, Pu Yang, Francois Charton, and Julia Kempe. Beyond model collapse: Scaling up with synthesized data requires verification. *arXiv preprint arXiv:2406.07515*, 2024. URL https://arxiv.org/abs/2406.07515.

Dan Friedman and Adji Bousso Dieng. The Vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2210.02410.

Anvit Garg, Sohom Bhattacharya, and Pragya Sur. Preventing model collapse under overparametrization: Optimal mixing ratios for interpolation learning and ridge regression. *arXiv preprint arXiv:2509.22341*, 2025. URL https://arxiv.org/abs/2509.22341.

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. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2404.01413. arXiv:2404.01413.

Yi Gu, Lingyou Pang, Xiangkun Ye, Tianyu Wang, Jianyu Lin, Carey E. Priebe, and Alexander Aue. SIGMA: Scalable spectral insights for LLM model collapse. *arXiv preprint arXiv:2601.03385*, 2026. URL https://arxiv.org/abs/2601.03385.

Hengzhi He, Shirong Xu, and Guang Cheng. Recursive learning without collapse: A weighting-based stabilization framework. *arXiv preprint arXiv:2502.18049*, 2025. URL https://arxiv.org/abs/2502.18049.

Figarri Keisha, Zekun Wu, Ze Wang, Adriano Koshiyama, and Philip Treleaven. Knowledge collapse in LLMs: When fluency survives but facts fail under recursive synthetic training. *arXiv preprint arXiv:2509.04796*, 2025. URL https://arxiv.org/abs/2509.04796.

Rylan Schaeffer, Joshua Kazdan, Apratim Cyrus Arulandu, and Sanmi Koyejo. Position: Model collapse does not mean what you think. *arXiv preprint arXiv:2503.03150*, 2025. URL https://arxiv.org/abs/2503.03150.

Yan Scholten, Sophie Xhonneux, Leo Schwinn, and Stephan Günnemann. Model collapse is not a bug but a feature in machine unlearning for LLMs. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2507.04219. arXiv:2507.04219.

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. How bad is training on synthetic data? a statistical analysis of language model collapse. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2404.05090. arXiv:2404.05090.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023. URL https://arxiv.org/abs/2305.17493.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024. doi: 10.1038/s41586-024-07566-y.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Author correction: AI models collapse when trained on recursively generated data. *Nature*, 640:E6, 2025. doi: 10.1038/s41586-025-08905-3.

Tianyu Wang, Akira Horiguchi, Lingyou Pang, and Carey E. Priebe. LLM web dynamics: Tracing model collapse in a network of LLMs. *arXiv preprint arXiv:2506.15690*, 2025. URL https://arxiv.org/abs/2506.15690.

X. Wu, Y. Zhou, and W. Su. When does model collapse occur in structured interactive learning? *arXiv preprint arXiv:2605.20151*, 2026. URL https://arxiv.org/abs/2605.20151.

Shirong Xu, Hengzhi He, and Guang Cheng. A probabilistic perspective on model collapse. *arXiv preprint arXiv:2505.13947*, 2025. URL https://arxiv.org/abs/2505.13947.

Ke Yi, Tianyu Liu, Yu Cheng, and Zhihao Xu. Escaping model collapse via synthetic data verification. *arXiv preprint arXiv:2510.16657*, 2025. URL https://arxiv.org/abs/2510.16657.