

HOMEOSTASIS: Closed-Loop Cryptographic Provenance and Homeostatic Control for Detecting and Mitigating Generative Model Collapse

Maruf Sharker Munna

marufsharkermunna@gmail.com

University of Liberal Arts Bangladesh <https://orcid.org/0009-0008-2981-6624>

Research Article

Keywords: model collapse, synthetic data, data provenance, watermarking, homeostatic control, Lyapunov stability, PID control, generative artificial intelligence

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10373786/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

HOMEOSTASIS: Closed-Loop Cryptographic Provenance and Homeostatic Control for Detecting and Mitigating Generative Model Collapse

Md. Maruf Sharker Munna*

*Department of Computer Science and Engineering
University of Liberal Arts Bangladesh
Dhaka, Bangladesh
Email: marufsharkermunna@gmail.com*

Abstract— Generative models are increasingly trained on content produced by earlier generative models. This feedback loop can narrow the learned distribution and erase its tails, a failure known as model collapse. Research on provenance, which attributes content to its source, and research on collapse dynamics have so far developed largely apart. This paper presents HOMEOSTASIS, a closed-loop framework that couples the two. Generated content is signed and hash-chained at creation; a compact vector of model vital signs (predictive entropy, calibration error, and representational drift) is fused into a Lyapunov-style collapse-risk functional; and a proportional-integral-derivative (PID) controller acts on the data-mixing ratio and on quarantine decisions before degradation becomes visible in output quality. We release a reference implementation with real SHA-256 hashing, Ed25519 signatures, and an append-only ledger, and we evaluate five experimental arms on a recursive Gaussian testbed (30 seeds, 20 generations). Closed-loop control reduced the final collapse risk from 2.67 to 0.48 (paired Cohen's $d = -1.65$) and preserved 97.4% of tail coverage, against 66.2% without control. Provenance-gated quarantine, however, added no detectable benefit over control alone ($d = -0.11$, adjusted $p = 0.82$), partly because the controller drove the human-data fraction to almost 1 in this setting. We explain why, delimit what the testbed can and cannot show, and specify the evaluation required at language-model scale.

Index Terms— model collapse, synthetic data, data provenance, watermarking, homeostatic control, Lyapunov stability, PID control, generative artificial intelligence

I. INTRODUCTION

For most of the deep learning era the open web served as a noisy proxy for human-generated data, and models were trained to approximate that distribution. The assumption is eroding. Large language models, diffusion models [1], and multimodal generators now publish text, images, and code at a scale that makes machine-generated material a growing share of what the next crawl will collect. Because web corpora are assembled without systematic attribution of authorship, later models are partly trained on the outputs of earlier ones, often unknowingly.

The consequences of this recursion have been documented from several directions. Shumailov et al. showed that repeatedly fitting a model to samples drawn from its predecessor makes the tails of the original distribution disappear first and leads to irreversible defects [2]. Self-consuming loops degrade image quality and diversity unless fresh real data enter each round [3], [4], language models show the same tendency [5], and neural scaling laws change character once synthetic data are mixed in [6]. The severity depends on how data are combined: replacing the corpus with synthetic samples collapses, whereas accumulating synthetic data alongside the original corpus can keep the error bounded [7], [8]. A critical commentary cautions against extrapolating from the simplest settings to arbitrary pipelines [9].

A separate community approaches the same ecosystem from the standpoint of accountability. In the European Union,

Article 50(2) of the AI Act, applicable since August 2026, requires providers of generative systems to mark synthetic output in a machine-readable way [10], and standards such as C2PA attach signed provenance metadata to content [11], [12]. Watermarking research has progressed from token-level sampling biases [13] to schemes that leave the output distribution unchanged [14], [15]. These efforts are motivated by misinformation and intellectual property rather than by model collapse, and they stop at attribution: they say where content came from, but nothing in the training pipeline acts on that answer.

The collapse literature has the complementary weakness. Mitigation schemes that reweight real and synthetic samples [16] or select samples by a verifier [17], [18] presuppose that the pipeline knows, or can estimate, which samples are synthetic. The authors of the weighting analysis list this reliance on distinguishing real from synthetic samples as a key limitation [16]. Provenance is exactly the signal that would supply it.

This paper argues that the two problems are halves of one control problem. Verified provenance supplies the sensing half of a feedback loop; a homeostatic controller supplies the actuation half. We call the resulting architecture HOMEOSTASIS, after the biological principle of holding an internal variable near a set point [19], [20]. The contributions are as follows.

- **Architecture.** A closed-loop design in which signed, hash-chained provenance records and model vital signs jointly

drive a corrective controller over the training pipeline (Section IV).

- **Formalization.** A collapse-risk functional over a three-dimensional vital-sign space, and a drift condition under which the controlled risk is provably ultimately bounded (Proposition 1).
- **Reference implementation.** Open code with real SHA-256 hashing, Ed25519 signatures, a hash-chained ledger, closed-form vital signs, a PID controller, and 18 unit tests (Section V).
- **Evaluation with a negative result.** A five-arm experiment with 30 paired seeds, bootstrap confidence intervals, and false-discovery-rate correction (Section VI). Control prevents collapse in the testbed; provenance-gated quarantine adds no detectable benefit on top of it.
- **Delimitation.** A frank account of why the negative result arises, what the testbed cannot establish, and which experiments would settle the question (Section VII).

Section II reviews related work, Section III formalizes the problem, and Section IV describes the framework. Sections V and VI present the experimental setup and results, Section VII discusses them together with limitations and governance, and Section VIII concludes.

II. RELATED WORK

A. Model collapse: theory and mitigation

In the Gaussian setting, recursively refitting a model to its own samples drives the estimated variance toward zero while the mean performs a random walk [2]. Analogous results exist for linear and ridge regression, with explicit error rates and optimal regularization [21]. Diffusion models lose tails and diversity across self-training rounds, and the accumulated divergence can be written as a discounted sum of per-generation score errors [22]; a complementary view attributes collapse to a shift from generalization to memorization, accompanied by falling entropy of the training data [18]. Selection can itself cause collapse: a verifier that sees only a biased slice of the target distribution prunes globally relevant tail modes [17].

Mitigation work is dominated by data-mixing arguments. Accumulating real and synthetic data bounds the error [7], and optimal weighting of real and synthetic samples slows drift under limited data retention [16]. All of these analyses are diagnostic or static. None monitors a running pipeline, and all assume that the real or synthetic origin of each sample is known.

B. Provenance and watermarking

Two technical routes establish provenance. Signed metadata, as in C2PA, offers strong evidence while intact, but it can be stripped when content is re-encoded or re-uploaded [11], [12]. Statistical watermarks survive such removal better and degrade gracefully. For language models, green-list sampling embeds a detectable bias [13], undetectable constructions rely on a secret key and a one-way function [14], and distortion-free schemes tolerate substantial random edits [15]. Fine-tuning a model on watermarked text leaves a

detectable trace in the model, a property called radioactivity [23], although robustness and radioactivity can conflict in federated fine-tuning [24]. For images, lossless watermarks have been proposed for diffusion models [25], yet generative regeneration attacks can remove many invisible watermarks [26]. Adoption in practice remains uneven [27]. Provenance guarantees are therefore probabilistic, a premise the present framework adopts.

C. Homeostatic and control-theoretic learning

Biological homeostasis, such as synaptic scaling that stabilizes neuronal firing [19], has inspired regulation schemes for artificial networks, and homeostatic reinforcement learning treats reward as the reduction of deviation from a physiological set point [20]. In control theory, Lyapunov functions can be synthesized as neural networks with formal verification [28], [29], combined with model-predictive control (MPC) to obtain safety and stability guarantees [30], [31], and used to certify safe exploration in reinforcement learning [32]. Neural approximations make MPC laws cheap to evaluate [33]. These methods regulate internal state from internal signals. None of them consumes an external, verifiable signal about the lineage of the data being processed, so they cannot tell a genuine distribution shift from creeping contamination by the model's own descendants.

D. Positioning

Table I summarizes the gap. Collapse theory and weighting schemes assume lineage is known; provenance systems produce lineage but take no action; control methods act but lack lineage. HOMEOSTASIS combines the three ingredients.

TABLE I
Positioning of HOMEOSTASIS against the three research threads.

Thread	External lineage	Model state	Automatic action
Collapse theory	Assumed known	No	No
Weighting schemes	Assumed known	No	Static reweighting
Provenance, watermarks	Produced	No	No
Lyapunov / MPC control	No	Yes	Plant-level
HOMEOSTASIS	Verified, used	Yes	Mixing, quarantine

III. PROBLEM FORMULATION

Consider generative models $G_1, \dots, G_T$ trained over discrete generations. Model G_t is fitted to a corpus $D_t = D_t^H \cup D_t^S$ that mixes human-originated data D_t^H with synthetic data D_t^S produced by earlier models. In the *replace* regime, D_t is drawn solely from G_{t-1} ; in the *accumulate* regime, new synthetic data are added to the retained corpus [7]. Real pipelines lie between these extremes, with a synthetic share that grows with calendar time and is not directly observed.

We denote by $\lambda_t \in [0,1]$ the fraction of D_t that is certified human data. The question addressed here is the following. *Given only the content itself and whatever provenance signals*

can be embedded and verified, can a training pipeline detect an impending collapse trajectory early enough, and correct it automatically enough, to hold the learned distribution near its human reference without reverting to purely human-sourced data?

Primitives. The provenance layer uses a collision-resistant hash $H(\cdot)$ [34], a digital signature scheme (Ed25519 in our implementation [35]), and hash chaining in the style of Merkle's construction [36], which makes retroactive alteration of a record detectable. The control layer uses a Lyapunov-style nonnegative functional [37], the PID law, and optionally MPC [38], [39].

IV. THE HOMEOSTASIS FRAMEWORK

A. Overview

HOMEOSTASIS has four coupled subsystems (Fig. 1): (A) a *provenance substrate* that signs and chains every generated unit at creation; (B) a *vital-sign sensor* that extracts a compact statistic of model behavior; (C) a *collapse-risk estimator* that fuses the vital signs into a scalar; and (D) a *homeostatic controller* that turns risk into corrective actions on the next corpus. The verifier in the loop uses the ledger to decide which ingested samples are admitted.

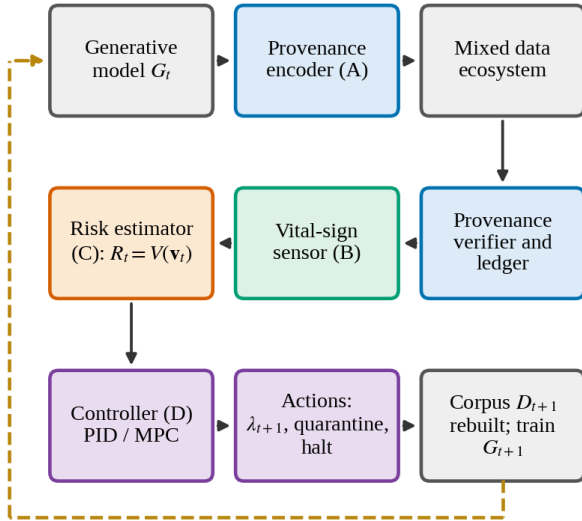


Fig. 1. Closed loop of HOMEOSTASIS. Content from G_t is stamped by the provenance substrate (A), enters the data ecosystem, and is verified against the ledger. The vital-sign sensor (B), the risk estimator (C), and the controller (D) determine the composition of the next corpus, which trains G_{t+1} .

B. Provenance substrate

When G_t emits an output x , the provenance encoder embeds a statistical watermark, giving $\tilde{x}$, and creates a record that is chained to its predecessor and signed:

$$\begin{aligned} r_t &= (\text{id}_{G_t}, H(\tilde{x}), t, h_{t-1}), \\ h_t &= H(h_{t-1} \parallel r_t), \\ \sigma_t &= \text{Sign}_{sk_{G_t}}(h_t). \end{aligned} \quad (1)$$

Records are appended to a shared, append-only ledger anchored at a genesis hash h_0 of a certified human corpus. A consumer verifies σ_t under the publisher's public key, recomputes h_t , and, if the ledger is organized as a Merkle tree, checks inclusion in $O(\log n)$ time [36]. If metadata attached to $\tilde{x}$ has been stripped, the watermark allows $\tilde{x}$ to be matched back to its ledger entry with probability less than one.

The scheme is probabilistic by design. Undetectable watermarks depend on a secret key [14], and invisible image watermarks can be removed by regeneration [26]. The substrate therefore raises the cost of undetected contamination and yields auditable evidence; it does not offer cryptographic certainty about lineage.

C. Vital signs and the collapse-risk functional

At generation t the sensor computes $\mathbf{v}_t = (H_t, \text{ECE}_t, \Delta_t) \in \mathbb{R}^3$:

- H_t is the predictive entropy of the model, whose decline is the statistical signature of variance shrinkage [2], [18]. In the Gaussian testbed it is the differential entropy of the fitted model; for classifiers or language models it is the mean Shannon entropy of the output distribution over a fixed probe set.
- ECE_t is the expected calibration error [40]–[42], a model-agnostic proxy for distribution mismatch that is cheap to compute at scale [43].
- Δ_t is a representational-drift statistic: a distance (Wasserstein, maximum mean discrepancy, or a Fréchet-type distance) between the model's feature distribution at generation t and a reference anchored at the last generation trained on verified human data.

Given a healthy reference entropy H^* established from human data, the collapse-risk functional is

$$\begin{aligned} V(\mathbf{v}) &= w_1 (H^* - H)_+^2 + w_2 \text{ECE}^2 + w_3 \Delta^2, \\ R_t &= V(\mathbf{v}_t), \end{aligned} \quad (2)$$

where $(\cdot)_+ = \max(\cdot, 0)$ penalizes only entropy below the healthy level, and $w_i \geq 0$ are weights. The functional is nonnegative, and it vanishes at the healthy equilibrium. Intervention is triggered when R_t exceeds a threshold τ . In the Gaussian testbed, $\Delta_t = 0$ forces the fitted model to equal the reference, and hence $H_t = H^*$, so $V = 0$ holds only at the equilibrium.

The following proposition states the property that the controller is meant to deliver. It is conditional on a drift assumption about the controller, which we neither prove for the controller of Section IV-D nor test in Section VI.

Proposition 1 (ultimate boundedness). Suppose there are constants $\alpha \in (0, 1)$ and $\eta \geq 0$ such that $R_{t+1} \leq (1 - \alpha)R_t + \eta$ whenever $R_t > \tau$, and $R_{t+1} \leq \tau + \eta$ whenever $R_t \leq \tau$. Let $M = \max\{\tau + \eta, \eta/\alpha\}$. Then (i) the set $\{R \leq M\}$ is forward invariant, and (ii) if $R_{t_0} > M$, then $R_{t_0+k} - \eta/\alpha \leq (1 - \alpha)^k (R_{t_0} - \eta/\alpha)$ for as long as R stays above τ .

Proof. (i) If $R_t \leq \tau$, then $R_{t+1} \leq \tau + \eta \leq M$. If $\tau < R_t \leq M$, then $R_{t+1} \leq (1 - \alpha)M + \eta \leq M$, because $M \geq \eta/\alpha$. (ii) Subtracting η/α from both sides of the drift inequality gives

$R_{t+1} - \eta/\alpha \leq (1 - \alpha)(R_t - \eta/\alpha)$, and the claim follows by induction. $\square$

The constant η absorbs the risk that new contamination injects in one generation. The proposition makes the informal requirement precise: correction must remove, on average, a fraction α of the excess risk faster than contamination adds η . It is a statement about the closed loop under an assumption, not a theorem about HCRC.

D. The Homeostatic Collapse-Risk Controller

The Homeostatic Collapse-Risk Controller (HCRC) is a discrete PID law on the threshold error $e_t = R_t - \tau$ [38]:

$$u_t = K_p e_t + K_i \sum_{s \leq t} e_s + K_d (e_t - e_{t-1}), \quad (3)$$

applied to the mixing ratio when intervention is triggered:

$$\lambda_{t+1} = \begin{cases} \text{clip}_{[0,1]}(\lambda_t + \text{clip}_{[-c,c]}(u_t)), & R_t > \tau, \\ \lambda_t, & R_t \leq \tau. \end{cases} \quad (4)$$

The action set also includes quarantining samples that fail verification and suspending retraining pending human-verified data. When risk is rising but has not yet crossed τ , a short-horizon MPC formulation can replace the PID law [39]:

$$u_{t:t+H-1} = \underset{u}{\operatorname{argmin}} \sum_{k=t}^{t+H-1} [V(\mathbf{v}_k) + \rho \|u_k\|^2] \quad (5)$$

s.t. $\mathbf{v}_{k+1} = f(\mathbf{v}_k, u_k)$,

where f is an identified model of how vital signs respond to actions and ρ discourages oscillatory over-correction. The MPC variant is part of the design but is not implemented in the reference code. Algorithm 1 gives the loop as implemented.

Algorithm 1: Homeostatic Collapse-Risk Controller (as implemented).

Input: corpus D_t , ledger L , gains (K_p, K_i, K_d) , threshold τ , weights $\mathbf{W}$, reference H^*
Output: corpus D_{t+1} , mixing ratio λ_{t+1}
1: $\mathbf{V}_t \leftarrow \text{ComputeVitalSigns}(G_t, \text{probe set}, \text{reference})$
2: $R_t \leftarrow V(\mathbf{V}_t)$; $e_t \leftarrow R_t - \tau$; $I \leftarrow I + e_t$
3: $u_t \leftarrow K_p e_t + K_i I + K_d (e_t - e_{t-1})$
4: if $R_t > \tau$ then $\lambda_{t+1} \leftarrow \text{clip}_{[0,1]}(\lambda_t + \text{clip}_{[-c,c]}(u_t))$
5: else $\lambda_{t+1} \leftarrow \lambda_t$
6: for each synthetic sample $x \in D_t$ do
7: $(\text{valid}, g) \leftarrow \text{VerifyProvenance}(x, L)$
8: if not valid then quarantine x and backfill from the certified human reservoir
9: return D_{t+1}, λ_{t+1}

E. Computational cost

Table II lists the cost of each component for a corpus of n samples, a probe batch of size b , and feature dimension d . The dominant term in a single-organization deployment is the corpus-wide provenance scan; it can be amortized by verifying a random subset of D_t per generation, trading completeness for cost.

TABLE II
Time complexity of the HOMEOSTASIS components.

Component	Cost
Sign and chain one record	$O(1)$
Verify one record (signature, hash)	$O(1)$
Merkle inclusion proof (proposed)	$O(\log n)$
Full-corpus verification scan	$O(n)$
Entropy and ECE from outputs	$O(bd)$
Drift: MMD / Gaussian closed form	$O(b^2 d) / O(d^3)$
Risk functional and PID step	$O(1)$
MPC, exhaustive over $ \mathcal{A} $ actions	$O(\mathcal{A} ^H)$

F. Explainability and governance

Because the controller changes what data a model sees, each intervention must be explainable. The design decomposes R_t into per-sensor Shapley contributions [44], treating the three vital signs as players, and cross-references quarantined samples with their ledger entries to produce an auditable report. Attribution intervals should be reported rather than point values, since Shapley estimates are sensitive to feature correlation. This module is specified but not evaluated here.

V. EXPERIMENTAL SETUP

A. Testbed

We use the minimal setting of the foundational collapse analyses [2], [21]. The reference distribution is $\mathcal{N}(\mathbf{0}, I_5)$. A finite reservoir of 4,000 certified human samples is drawn once and reused in every generation, which models the scarcity of verified human data. Each generation's corpus holds $n = 40$ samples: a fraction λ_t drawn with replacement from the reservoir and the remainder sampled from the previous model. Model G_t is the maximum-likelihood Gaussian fitted to the corpus, and G_0 is the reference. The initial human fraction is $\lambda_1 = 0.10$, a synthetic-dominated regime, and the small corpus makes finite-sample collapse visible within $T = 20$ generations. Each arm is run with seeds 0, ..., 29; the same seed yields the same human reservoir in every arm, which permits paired comparisons.

B. Experimental arms

Table III defines the five arms. Verification in A1 is logged but does not alter the corpus, which isolates attribution without action. Quarantine acts only when provenance and control are both active (A3, A4): synthetic samples that fail verification are removed and replaced by draws from the human reservoir. A4 lowers the watermark survival probability from 0.95 to 0.40 to emulate large-scale stripping.

TABLE III
Experimental arms.

Arm	Verification	Control	Watermark survival
A0	Off	Off	n/a
A1	Log only	Off	0.95
A2	Off	On	n/a
A3	Quarantine	On	0.95

A4	Quarantine	On	0.40
----	------------	----	------

C. Instantiation of the vital signs

For a Gaussian model $\mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t)$ in d dimensions, the entropy is $H_t = 1/2 \ln((2\pi e)^d |\Sigma_t|)$, and the healthy level is $H^* = 7.095$ nats. Drift is the exact 2-Wasserstein distance to the reference [45],

$$\Delta_t^2 = \|\boldsymbol{\mu}_t\|^2 + \text{Tr}(\Sigma_t + I - 2\Sigma_t^{1/2}), \quad (6)$$

which is computed in closed form rather than approximated. A classifier that observes its own label would trivially report $\text{ECE} = 0$, so calibration is measured with a noisy-channel classifier: samples are projected onto the reference's principal axis, observed through Gaussian noise with standard deviation equal to half the reference spread, and the classifier predicts whether the projection exceeds the reference mean using exact Bayesian updating under the reference prior. Its confidence is compared with the true label on 1,000 probe samples from G_t in ten bins. Calibration degrades only when G_t has drifted from the reference. Finally, *tail coverage* is the fraction of 4,000 points drawn from the reference's Mahalanobis shell [2.5, 3.5] at which G_t still assigns at least 1% of the reference density, a direct operationalization of tail loss.

D. Implementation of the provenance layer

Provenance uses the `cryptography` library: SHA-256 hashing and Ed25519 signatures, with a fresh key pair per generation, and an append-only hash-chained ledger anchored at a genesis hash. A record verifies if its hash recomputes, its signature validates under the claimed generation's public key, and its watermark flag is intact. Eighteen unit tests cover tamper detection, wrong-key rejection, stripped-watermark handling, controller triggering, action clipping, and integral accumulation. Two simplifications matter for interpretation. The content hash is computed from a record identifier rather than from the sample vector, and watermark survival is a Bernoulli draw. The cryptographic layer is thus exercised but not stressed.

E. Statistical analysis

Comparisons are paired across seeds at the final generation. Each paired difference is checked with the Shapiro–Wilk test [46]; a paired t -test is used if normality is not rejected at 0.05, and the Wilcoxon signed-rank test otherwise [47]. Effect sizes are paired Cohen's d (mean difference divided by the standard deviation of differences) [48] with 10,000-resample bootstrap 95% confidence intervals [49]. All 16 tests (four contrasts, four metrics) are corrected with the Benjamini–Hochberg procedure [50], and $|d| \geq 0.5$ is taken as the threshold of practical significance. The risk threshold $\tau = 0.35$ was chosen after inspecting the uncontrolled baseline so that A0 crosses it; its influence is examined in Section VI-D. All parameters appear in Table VI (Appendix).

VI. RESULTS

A. Closed-loop control prevents collapse

Without control, the mean risk rises from 0.47 at $t = 1$ to a maximum of 3.09 at $t = 18$ and ends at 2.67 (Fig. 2). All controlled arms hold the risk between roughly 0.36 and 0.53 from generation 5 onward. Tail coverage tells the same story (Fig. 3): the uncontrolled arm falls to a minimum of 0.60 and ends at 0.66, whereas the controlled arms stay between 0.96 and 0.99 from generation 5 onward. Arm A1 is trajectory-identical to A0, since verification alone changes nothing in the corpus. Its provenance recall of 0.947 matches the watermark survival rate by construction, which confirms that the verification logic behaves as designed.

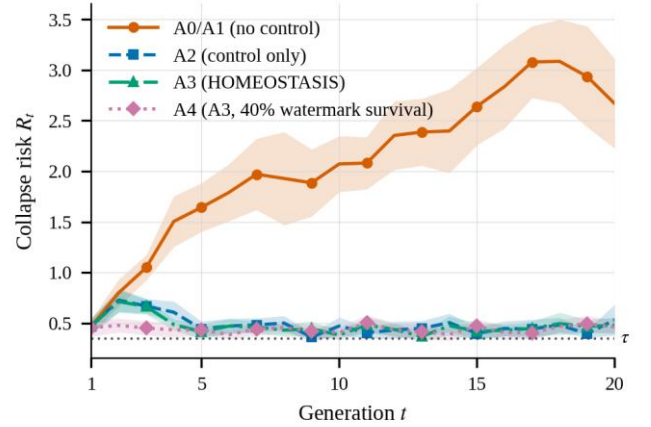


Fig. 2. Mean collapse risk R_t across 30 seeds (shaded: 95% confidence interval). The dotted line marks the threshold $\tau = 0.35$. A1 coincides with A0.

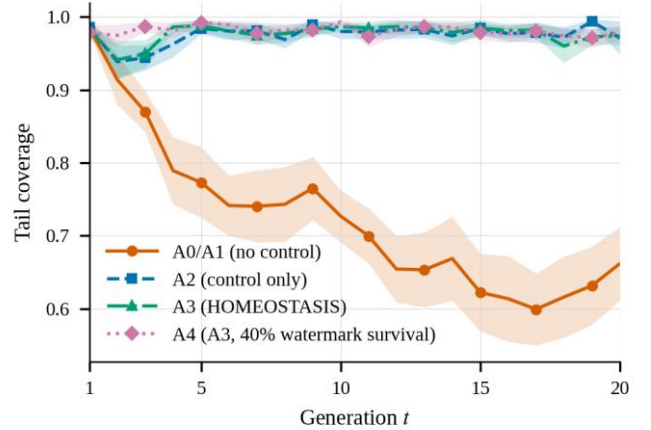


Fig. 3. Mean tail coverage across 30 seeds (shaded: 95% confidence interval). Uncontrolled recursion loses about a third of the reference tail.

TABLE IV

Final-generation ($t = 20$) means $\pm$ standard deviation over 30 seeds.

Arm	Risk R	Drift Δ	Entropy H	Tail cov.	λ_{20}
A0	2.67 ± 1.24	1.22 ± 0.30	6.23 ± 0.59	0.66 ± 0.14	0.10
A1	2.67 ± 1.24	1.22 ± 0.30	6.23 ± 0.59	0.66 ± 0.14	0.10
A2	0.53 ± 0.43	0.58 ± 0.09	6.84 ± 0.37	0.97 ± 0.06	1.00
A3	0.48 ± 0.22	0.58 ± 0.08	6.83 ± 0.26	0.97 ± 0.03	1.00
A4	0.46 ± 0.18	0.56 ± 0.07	6.80 ± 0.25	0.98 ± 0.03	0.98

Table IV gives the final-generation values, and Table V the paired tests. Comparing A3 with A0, all four effects are large and survive correction: risk ($d = -1.65$), drift ($d = -2.04$), entropy ($d = +0.86$), and tail coverage ($d = +2.12$), each with an adjusted $p < 0.001$. This supports the claim that closing the loop on vital-sign risk prevents the collapse that the uncontrolled recursion exhibits, in this testbed.

TABLE V

Paired contrasts at $t = 20$: Cohen's d with bootstrap 95% CI and BH-adjusted p (first arm minus second).

Contrast	Metric	d	95% CI	p_{adj}
A3 vs A0	Risk	-1.65	[-2.16, -1.33]	<0.001
	Drift	-2.04	[-2.69, -1.71]	<0.001
	Entropy	+0.86	[0.54, 1.29]	<0.001
	Tail cov.	+2.12	[1.77, 2.77]	<0.001
A3 vs A2	Risk	-0.11	[-0.40, 0.32]	0.82
	Drift	0.00	[-0.37, 0.37]	1.00
	Entropy	-0.02	[-0.39, 0.35]	0.95
	Tail cov.	+0.04	[-0.47, 0.32]	0.82
A4 vs A3	Risk	-0.07	[-0.42, 0.31]	0.82
	Drift	-0.22	[-0.66, 0.14]	0.43
	Entropy	-0.09	[-0.48, 0.27]	0.82
	Tail cov.	+0.12	[-0.25, 0.48]	0.82

B. Provenance adds no detectable benefit over control alone

The comparison that tests the unification claim is A3 against A2. Final risk is 0.48 against 0.53 ($d = -0.11$, adjusted $p = 0.82$), drift is 0.58 in both arms, and tail coverage is 0.974 against 0.971. No contrast approaches significance (Table V). With 30 pairs a paired test detects an effect of about $d = 0.5$ with 80% power, and the confidence

intervals exclude effects larger than about $|d| = 0.5$; a small benefit therefore cannot be ruled out, but a practically large one can.

The human-data trajectories explain the null result (Fig. 4). In both A2 and A3 the controller raises λ_t above 0.9 by generation 8 and to essentially 1 by generation 14. Once the corpus consists almost entirely of human samples, there is nothing left for the verifier to quarantine, and the quarantine rate in A3 at $t = 20$ is exactly zero. The residual risk of about 0.5 is also not collapse. A corpus of 40 pure-human samples has a finite-sample estimation error of its own: fitting 2,000 such corpora gives a mean risk of 0.42 (SD 0.20). Both controlled arms therefore sit at the floor set by $n = 40$, below which neither can move.

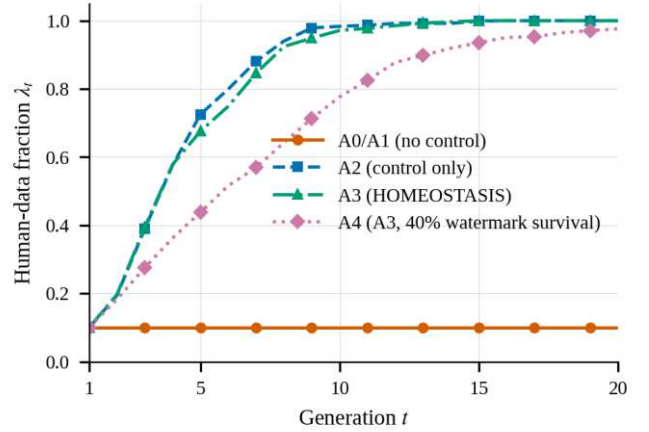


Fig. 4. Mean human-data fraction λ_t . Controlled arms converge to almost entirely human data; A4 converges more slowly.

C. Watermark stripping (A4)

With the watermark survival probability reduced to 0.40, A4 shows no significant degradation relative to A3: final risk is 0.46 against 0.48 ($d = -0.07$, adjusted $p = 0.82$). Averaged over generations, 35% of synthetic samples were quarantined in A4. This result must not be read as robustness. In the simulator, a stripped watermark makes verification fail, flagged samples are replaced by human data, and human data are precisely what prevents collapse here. A real adversary would instead try to have synthetic content accepted as human, a false negative that the simulator does not model.

D. Sensitivity to the risk threshold

Because τ was set after observing the baseline, we repeated A2 and A3 for $\tau \in \{0.2, 0.35, 0.6, 1.0, 2.0\}$ (Fig. 5). Both controlled arms remain far below the uncontrolled risk of 2.67 throughout; the final risk of A3 ranges from 0.40 to 1.12. The paired contrast A3 minus A2 has d between -0.11 and $+0.33$, below the practical threshold at every setting and changing sign, so there is no consistent benefit of provenance gating (these contrasts were not tested formally). At $\tau = 0.6$, 1.0, and 2.0 the final human fraction of A3 is 0.85, 0.50, and 0.28, so the null result does not depend on saturation alone. At $\tau = 2$ the risk of A3 (1.12) is higher than that of A2 (0.85); we have not isolated the cause.

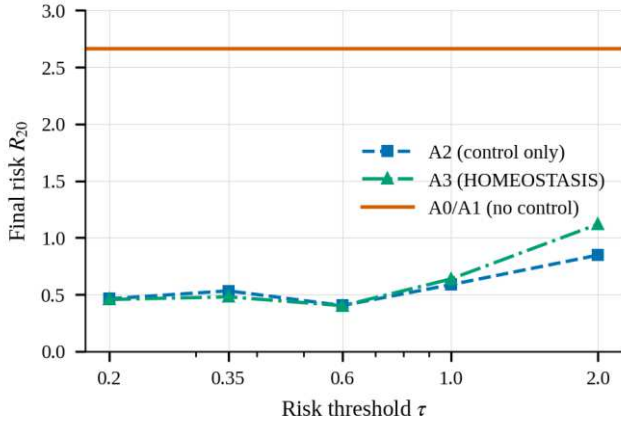


Fig. 5. Final risk R_{20} against the threshold τ . The horizontal line is the uncontrolled arm.

VII. DISCUSSION

A. What evidence supports

Two claims were put to the test. The first, that feedback on vital-sign risk slows or prevents collapse relative to no intervention, is supported in the testbed with large effects on all four metrics. The second, that provenance-gated action outperforms control that cannot see which samples are synthetic, is not supported: the point estimates are close to zero and the intervals are narrow enough to exclude a large benefit. The experiment was designed so that it could have gone either way, and we report the outcome as it came out.

The positive result should itself be read modestly. In this testbed the controller discovers that the cure for collapse is to stop using synthetic data, since the human reservoir is unlimited in the sense that it can be resampled indefinitely. The controller has no cost term, can only increase λ_t , and therefore says nothing about the cost–stability trade-off that motivates tolerating synthetic data in the first place.

B. When provenance could matter

We see four conditions under which per-sample provenance might add value beyond aggregate control. They are hypotheses, not findings, and each can be tested with the released code. First, a *bounded human-data budget* would stop the controller from saturating, leaving synthetic samples to be sorted. Second, *clustered or bursty contamination* would make aggregate signals lag behind localized damage that per-sample verification could catch. Third, *unflagged synthetic content* that an adversary passes as human would test whether verification fails safely. Fourth, pipelines that use weighting rather than replacement [16] need per-sample origin labels, which only provenance can supply.

C. Limitations

The first limitation is scope: a five-dimensional Gaussian fitted by maximum likelihood is not a language or diffusion model, and whether the three vital signs transfer to such models is open. Entropy and drift are closed-form here but must be estimated, noisily, at scale. Second, τ was chosen post hoc, so absolute risk values are not comparable across studies,

although the sensitivity analysis suggests that the conclusions are not an artifact of one setting. Third, the simulator abstracts the provenance layer: content hashes are computed from identifiers, watermark survival is a coin flip, verification has no false positives by construction, and paraphrase, translation, and partial excerption are absent. Fourth, the ledger is a linear hash chain, not a Merkle tree, and key management is reduced to one key pair per generation. Fifth, the MPC variant and the Shapley explanation module are specified but not implemented, the PID integrator has no anti-windup, and the drift assumption of Proposition 1 has not been estimated from data. Finally, the literature is moving quickly, and findings cited from recent preprints should be rechecked against their final versions before formal submission.

D. Governance and ethical considerations

Provenance infrastructure at ecosystem scale carries risks that the technical design cannot remove. A ledger could be repurposed to track the authorship of human content, so its scope should be restricted to model-generated material. Provenance mechanisms that only well-resourced providers can afford could entrench incumbents, which argues for open, standards-body governance of the kind that surrounds C2PA [11]. A false positive could attribute content wrongly, so attribution claims should carry calibrated confidence rather than binary verdicts. Verification regimes could be compelled to suppress particular content, which calls for transparency and independent audit rights. Because quarantine and retraining suspension change what future models learn, a human override should apply to any action affecting more than a pre-agreed share of a corpus. Sampling-based audits keep participation feasible for under-resourced groups.

VIII. CONCLUSION AND FUTURE WORK

We proposed HOMEOSTASIS, a framework that treats verified data lineage as the sensing input of a homeostatic controller for recursive generative training, and we evaluated a working reference implementation on a recursive Gaussian testbed. Closed-loop control on a Lyapunov-style risk functional prevented the collapse seen in the uncontrolled recursion, with large effects on risk, drift, entropy, and tail coverage. Provenance-gated quarantine added no detectable benefit once the controller had raised the human-data share to nearly one, and the experiment gives no grounds to claim robustness against a strategic adversary.

The next steps follow from the limitations. Experiments with language and diffusion models should use corpora with verified provenance and a held-out human-only probe set. Within the simulator, a capped human-data budget, clustered contamination, and a false-negative adversary are the most informative extensions. Beyond that lie learned vital-sign sensors, MPC with identified dynamics, and federated ledgers that share only risk scores, where the tension between watermark robustness and radioactivity [24] will constrain the design.

APPENDIX A: EXPERIMENTAL PARAMETERS AND REPRODUCIBILITY

Table VI lists the parameters of the reported experiment. The code, raw per-generation records (3,000 rows), statistical output, the threshold-sensitivity data, and all figure scripts are supplied with this paper in the directory `homeostasis_mvp/`. Running `run_experiment.py` and `analyze.py` reproduces every number in Table IV and Table V exactly, because seeds are fixed.

TABLE VI
Parameters of the reported experiment.

Parameter	Value
Dimension d ; reference	5; $\mathcal{N}(\mathbf{0}, I_5)$
Human reservoir size	4,000
Corpus size n ; generations T	40; 20
Initial human fraction λ_1	0.10
Seeds per arm	30 (0–29)
Risk weights (w_1, w_2, w_3)	(1, 1, 1)
PID gains (K_p, K_i, K_d)	(0.5, 0.05, 0.1)
Action clip C ; threshold τ	0.35; 0.35
Healthy entropy H^*	7.095 nats
ECE bins; probe size	10; 1,000
Tail shell; density ratio	[2.5, 3.5]; 0.01
Bootstrap resamples; $ d $ threshold	10,000; 0.5
Watermark survival (A1–A3; A4)	0.95; 0.40

REFERENCES

- [1] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, vol. 33, pp. 6840–6851.
- [2] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, “AI models collapse when trained on recursively generated data,” *Nature*, vol. 631, pp. 755–759, 2024.
- [3] S. Alemohammad et al., “Self-consuming generative models go MAD,” in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024.
- [4] Q. Bertrand, A. J. Bose, A. Duplessis, M. Jiralspong, and G. Gidel, “On the stability of iterative retraining of generative models on their own data,” in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024.
- [5] M. Briesch, D. Sobania, and F. Rothlauf, “Large language models suffer from their own output: An analysis of the self-consuming training loop,” arXiv:2311.16822, 2023.
- [6] E. Dohmatob, Y. Feng, P. Yang, F. Charton, and J. Kempe, “A tale of tails: Model collapse as a change of scaling laws,” in *Proc. Int. Conf. Machine Learning (ICML)*, 2024.
- [7] M. Gerstgrasser et al., “Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data,” arXiv:2404.01413, 2024.
- [8] J. Kazdan et al., “Collapse or thrive? Perils and promises of synthetic data in a self-generating world,” arXiv:2410.16713, 2024.
- [9] A. Borji, “A note on Shumailov et al. (2024): ‘AI models collapse when trained on recursively generated data,’” arXiv:2410.12954, 2024.
- [10] European Parliament and Council of the European Union, “Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act),” Official Journal of the European Union, L 2024/1689, Jul. 2024 (Art. 50 applicable from Aug. 2, 2026).
- [11] Coalition for Content Provenance and Authenticity, “C2PA technical specification, version 2.2,” 2025.
- [12] B. Chandra, J. Dunietz, K. Roberts, Y. Lee, P. Fontana, and G. Awad, “Reducing risks posed by synthetic content: An overview of technical approaches to digital content transparency,” National Institute of Standards and Technology, NIST AI 100-4, Nov. 2024, doi: 10.6028/NIST.AI.100-4.
- [13] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, “A watermark for large language models,” in *Proc. Int. Conf. Machine Learning (ICML)*, 2023, pp. 17061–17084.
- [14] M. Christ, S. Gunn, and O. Zamir, “Undetectable watermarks for language models,” in *Proc. Conf. Learning Theory (COLT)*, 2024.
- [15] R. Kudithipudi, J. Thickstun, T. Hashimoto, and P. Liang, “Robust distortion-free watermarks for language models,” *Trans. Machine Learning Research*, 2024.
- [16] H. He, S. Xu, and G. Cheng, “Recursive learning without collapse: A weighting-based stabilization framework,” arXiv:2502.18049, 2025.
- [17] X. Qiao et al., “When sample selection bias precipitates model collapse,” arXiv:2606.13732, 2026.
- [18] L. Shi, M. Wu, H. Zhang, Z. Zhang, M. Tao, and Q. Qu, “A closer look at model collapse: From a generalization-to-memorization perspective,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.
- [19] G. G. Turrigiano, “Homeostatic synaptic plasticity: Local and global mechanisms for stabilizing neuronal function,” *Cold Spring Harbor Perspectives in Biology*, vol. 4, no. 1, Art. no. a005736, 2012.
- [20] M. Keramati and B. Gutkin, “Homeostatic reinforcement learning for integrating reward collection and physiological stability,” *eLife*, vol. 3, Art. no. e04811, 2014.
- [21] E. Dohmatob, Y. Feng, and J. Kempe, “Model collapse demystified: The case of regression,” arXiv:2402.07712, 2024.
- [22] N. B. Khelifa, R. E. Turner, and R. Venkataramanan, “Error propagation and model collapse in diffusion models: A theoretical study,” arXiv:2602.16601, 2026.
- [23] T. Sander, P. Fernandez, A. Durmus, M. Douze, and T. Furon, “Watermarking makes language models radioactive,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, vol. 37.
- [24] L. Huang, Z. Shao, and T. Baluta, “Watermark robustness and radioactivity may be at odds in federated learning,” arXiv:2510.17033, 2025.
- [25] Z. Yang, K. Zeng, K. Chen, H. Fang, W. Zhang, and N. Yu, “Gaussian Shading: Provable performance-lossless image watermarking for diffusion models,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 12162–12171.
- [26] X. Zhao et al., “Invisible image watermarks are provably removable using generative AI,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, vol. 37, pp. 8643–8672.
- [27] B. Rijsbosch, G. van Dijck, and K. Kollnig, “Adoption of watermarking for generative AI systems in practice and implications under the new EU AI Act,” arXiv:2503.18156, 2025.
- [28] A. Abate, D. Ahmed, M. Giacobbe, and A. Peruffo, “Formal synthesis of Lyapunov neural networks,” *IEEE Control Systems Letters*, vol. 5, no. 3, pp. 773–778, 2021.
- [29] Y.-C. Chang, N. Roohi, and S. Gao, “Neural Lyapunov control,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, vol. 32.
- [30] Z. Wu, F. Albalawi, Z. Zhang, J. Zhang, H. Durand, and P. D. Christofides, “Control Lyapunov-barrier function-based model predictive control of nonlinear systems,” *Automatica*, vol. 109, Art. no. 108508, 2019.
- [31] M. Mittal, M. Gallieri, A. Quaglino, S. S. M. Salehian, and J. Koutník, “Neural Lyapunov model predictive control: Learning safe global controllers from sub-optimal examples,” arXiv:2002.10451, 2020.
- [32] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, “Safe model-based reinforcement learning with stability guarantees,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, vol. 30.

- [33] B. Karg and S. Lucia, "Efficient representation and approximation of model predictive control laws via deep learning," *IEEE Trans. Cybernetics*, vol. 50, no. 9, pp. 3866–3878, 2020.
- [34] National Institute of Standards and Technology, "Secure hash standard (SHS)," Federal Information Processing Standards Publication 180-4, 2015.
- [35] D. J. Bernstein, N. Duif, T. Lange, P. Schwabe, and B.-Y. Yang, "High-speed high-security signatures," *J. Cryptographic Engineering*, vol. 2, no. 2, pp. 77–89, 2012.
- [36] R. C. Merkle, "A digital signature based on a conventional encryption function," in *Advances in Cryptology (CRYPTO '87), Lecture Notes in Computer Science*, vol. 293, 1987, pp. 369–378.
- [37] H. K. Khalil, *Nonlinear Systems*, 3rd ed.. Prentice Hall, 2002.
- [38] K. J. Åström and R. M. Murray, *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton Univ. Press, 2008.
- [39] J. B. Rawlings, D. Q. Mayne, and M. M. Diehl, *Model Predictive Control: Theory, Computation, and Design*, 2nd ed.. Nob Hill Publishing, 2017.
- [40] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in *Proc. AAAI Conf. Artificial Intelligence*, 2015.
- [41] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2017, pp. 1321–1330.
- [42] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring calibration in deep learning," in *CVPR Workshops*, 2019, pp. 38–41.
- [43] M. Abdar et al., "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.
- [44] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, vol. 30.
- [45] D. C. Dowson and B. V. Landau, "The Fréchet distance between multivariate normal distributions," *J. Multivariate Analysis*, vol. 12, no. 3, pp. 450–455, 1982.
- [46] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3–4, pp. 591–611, 1965.
- [47] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.
- [48] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.. Lawrence Erlbaum Associates, 1988.
- [49] B. Efron, "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.
- [50] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Royal Statistical Society: Series B*, vol. 57, no. 1, pp. 289–300, 1995.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [HOMEOSTASISSupplementaryCode.zip](#)