CAN SCALING UNDO AI-INDUCED DATA POLLUTION?

Anonymous authors
Paper under double-blind review

ABSTRACT

AI models are increasingly deployed while their scale increases over time, and deployment can change the evidence available to train future models. We model this interaction through coupled dynamics of AI adoption and evidence dilution, while learner scale follows an exogenous path. We study whether increasing model scale can undo degradation created by earlier deployment, and how the answer depends on when scaling occurs and on curation. At a fixed learner scale, we show that sufficiently high curation leaves only the clean equilibrium, while sufficiently low curation can also sustain a stable equilibrium with persistent degradation. We then prove that there are settings in which, from the same initial state and at the same curation rate, starting at a higher scale can lead to the clean equilibrium, while starting at a lower scale and switching to the same higher scale after sufficient time can instead lead to persistent degradation. We show numerically that the qualitative behavior persists under alternative specifications of how evidence degradation affects model quality and future production. Separately, controlled fine-tuning experiments with Qwen2.5-1.5B show that degradation propagates through the evolving evidence even when each generation starts from the same pretrained checkpoint, while curation substantially reduces the effect. Thus, the effect of scaling depends not only on the eventual model scale reached, but also on the evidence produced along the way and on the strength of curation.

1 INTRODUCTION

AI systems increasingly produce text, code, images, and other material that may later become part of the evidence used to train future models, including later versions of the same systems (Shumailov et al., 2024; Alemohammad et al., 2024). At the same time, scaling has become a central organizing idea in machine learning: increasing model capacity, compute, and data can substantially improve performance on the distributions used for training and evaluation (Kaplan et al., 2020; Hoffmann et al., 2022). Standard scaling analyses typically treat the training and evaluation distribution as given rather than altered by deployment. If the evidence continues to faithfully represent the world, then improving fit to that evidence can reasonably be interpreted as improving fit to the world.

Current AI systems, however, can produce factual errors (Lin et al., 2022) and can change the diversity and composition of the material produced with their assistance (Doshi & Hauser, 2024; Anderson et al., 2024; Ashkinaze et al., 2025; Wan & Kalman, 2026). Usefulness to users and grounding of the resulting evidence are different criteria: model assistance can be useful enough to be adopted before its production is as well grounded as the independent production it replaces. When model-assisted production substitutes for independent production, model-assisted artifacts can enter later training data in place of evidence that would otherwise have been produced independently. Deployment can therefore change the evidence available to future learners.

Once deployment changes the evidence available to future learners, improving the learner and improving the evidence it learns from become different questions. A more capable model may fit the evidence available to it increasingly well even while that evidence has moved away from the underlying world. If changes to the evidence source persist, the evidence available at a given model scale can depend on the history of deployment (Perdomo et al., 2020; Brown et al., 2022). The effect of scaling can therefore depend not only on the scale eventually reached, but also on when it is reached.

We therefore ask: *When can scaling undo AI-induced data pollution, and how does the timing of scaling affect whether it can do so?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Related work. This question connects several works that we discuss in detail in Section A. Work on recursive-training studies degradation when model-generated samples are reused in later training, but typically takes the production of new grounded evidence as given rather than allowing deployment to change how that evidence is produced (Shumailov et al., 2024; Alemohammad et al., 2024; Gerstgrasser et al., 2024; Kazdan et al., 2025). Performative prediction and related stateful models make future data endogenous to deployment, but do not study how increasing learner scale interacts with changes to the evidence source created by earlier deployment (Perdomo et al., 2020; Brown et al., 2022). Related scaling work incorporates data constraints, synthetic data, and data quality, while generally treating the data source or generation process as independent of deployment at earlier scales (Muennighoff et al., 2023; Dohmatob et al., 2024; Qin et al., 2025; Subramanyam et al., 2026). Finally, experiments on AI-assisted production show that assistance can improve individual productivity or evaluated output while also changing the composition and diversity of collective production (Noy & Zhang, 2023; Doshi & Hauser, 2024; Anderson et al., 2024). These studies leave open what happens when such changes accumulate in the evidence available to future learners.

Our contributions. We develop a continuous-time dynamical model of how scaling interacts with evidence altered by deployment, separating learner scale, the evidence state, and model adoption. At each time t, quality of model-assisted production depends jointly on the learner scale s_t and the accumulated dilution $z(t)$ of the evidence ecology (eq. (2)). Higher production quality creates greater adoption pressure, while existing users can abandon model assistance (eq. (1)). When model-assisted production is less well grounded than the independent production it replaces, the adopted share contributes new dilution to the evidence ecology; curation removes accumulated dilution at rate μ (eq. (4)). These relations give coupled dynamics for adoption and evidence dilution, while learner scale follows an exogenous path (eq. (5)). Crucially, adoption pressure can be positive even when model-assisted production remains below parity with the independent production it replaces.

We first characterize the equilibrium regimes possible at a fixed learner scale. As a boundary case, if the learner is sufficiently capable that model-assisted production is never worse than the production it replaces, then we show that no new dilution is created and curation removes what remains (see the beginning of Section 3). Thus, persistent degradation is not built into our model. When sufficiently diluted evidence can still prevent parity at accessible scales, we show for a broad class of adoption and dilution responses that there is a critical curation rate μ_c: above it, the clean equilibrium is the only equilibrium, while below it, at least one positive-dilution equilibrium exists (Theorem 1). We further show that for sufficiently low curation, one of these positive-dilution equilibria is stable (Theorem 2). Technically, we reduce equilibrium existence to a scalar balance between dilution production and curation, and then relate the slope of this balance to stability in the full adoption–dilution dynamics. Thus, at a fixed learner scale, the evidence ecology can have both a clean long-run state and a persistent degraded long-run state.

We then study whether increasing learner scale necessarily eliminates the degraded outcome, and show that it need not. In the canonical system, we prove that there exist two scales $s_{\mathrm{low}} < s_{\mathrm{high}}$ and a common initial state such that starting immediately at s_{high} converges to the clean equilibrium, whereas first evolving at s_{low} and switching sufficiently late to the same s_{high} leads to persistent positive dilution (Theorem 3). The proof requires relating the phase portraits at two different scales: we construct a common initial state that lies in the clean basin at the higher scale, but whose trajectory under the lower-scale dynamics enters the degraded basin of that same higher-scale system. Thus, the timing of scaling, not just the eventual scale, can determine the long-run outcome. Numerically, the timing effect persists in the full two-dimensional dynamics, is delayed by stronger curation, and is robust to alternative model specifications (Sections 3 and C).

Finally, in a controlled experiment with Qwen2.5-1.5B-Base, each generation starts from the same pretrained checkpoint and learner scale is held fixed, so that only the evidence ecology is inherited across generations (Section 4). Under model-mediated turnover, the distance of the circulating evidence from the fixed synthetic world rises from 0.1303 to 0.2703 over two generations while externally anchored accuracy falls from 36.58% to 32.91%; 50% audit-and-repair substantially attenuates both effects and restores accuracy to the grounded-turnover benchmark.

Together, these results show that whether scaling can undo earlier degradation depends not only on the eventual scale reached, but also on the evidence produced along the way and on the strength of curation.

2 MODEL

Let $\mathcal{X}$ be a sample space of *evidence units*, such as documents, computer programs, or images, and let $\mathcal{P}(\mathcal{X})$ denote the set of probability distributions on $\mathcal{X}$. We fix an *ideal evidence distribution* $P^\dagger \in \mathcal{P}(\mathcal{X})$, representing the fully grounded evidence that, ideally, would be available to future learners. We consider a setting in which, unlike in a fixed-data benchmark, the evidence distribution available to future learners is itself shaped by deployment.

To measure how far an evidence distribution is from $P^\dagger$, we use a normalized distance $d : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to [0, 1]$. Let θ denote a trained model, and let B_θ denote the distribution of evidence produced with assistance from θ and entering circulation. We define the quality of this assisted production by $q(\theta) := 1 - d(B_\theta, P^\dagger)$. At time t, let P_t denote the evidence distribution available for training, and let θ_t denote the model trained using this evidence. Once deployed, θ_t is used to assist some fraction of new production. The resulting model-assisted evidence has distribution B_{θ_t} and enters circulation, thereby affecting the evidence available to future learners.

Adoption and quality. Let $u(t) \in [0, 1]$ denote the fraction of production at time t that uses model assistance. We assume that higher-quality model-assisted production makes adoption more attractive. We represent this dependence by a continuous, non-negative, non-decreasing function A, with larger $A(q(\theta_t))$ corresponding to stronger adoption. We let $\kappa > 0$ control the rate of new adoption and allow existing users to stop using model assistance at rate $\lambda > 0$. The adoption share then evolves according to

$$\dot{u}(t) = \kappa A(q(\theta_t))(1 - u(t)) - \lambda u(t). \tag{1}$$

The first term represents new adoption: the unassisted share $1 - u(t)$ adopts at a rate that increases with the quality of model-assisted production. The second term represents abandonment: a fraction λ of the assisted share $u(t)$ stops using model assistance per unit time.

Scale, dilution, and quality. We use $z(t) \in [0, 1]$ to represent the dilution of the evidence available for training at time t. Here $z(t) = 0$ represents an undiluted evidence ecology, while larger values represent greater dilution. We allow the initial evidence ecology to have $z(0) > 0$. This initial dilution need not result from prior model deployment; it can represent finite-sample variation, measurement noise, or other pre-existing departures of the available evidence from $P^\dagger$. Let $s_t \in [0, s_{\max}]$ denote the scale of the learner available at time t, where $s_{\max} < \infty$ is the maximum accessible scale, summarizing factors such as model capacity, compute, architecture, etc. This scalar abstraction is motivated by empirical scaling laws showing systematic improvements with model size, data, and training compute (Kaplan et al., 2020; Hoffmann et al., 2022).

Scale is exogenous: the path s_t is specified independently of the adoption–evidence dynamics, and if $\mathcal{L}_s(P)$ denotes training at scale s on evidence distribution P, then $\theta_t = \mathcal{L}_{s_t}(P_t)$. Rather than track the full distribution P_t, we assume that its effect on the quality of model-assisted production is captured by the dilution level $z(t)$. We therefore write

$$q(\theta_t) = \Phi(s_t, z(t)). \tag{2}$$

where Φ is continuous, non-decreasing in scale, and non-increasing in dilution. The monotonicity in scale captures the usual scaling intuition, while the dependence on z captures the fact that attainable model quality can also depend on the quality of the evidence supplied to the learner (Subramanyam et al., 2026). The departure from standard scaling analyses is that evidence quality is not held fixed: here it is an endogenous state shaped by earlier deployment. In our analysis, we will often consider the following explicit instantiation of such a link function, chosen to be as simple as possible.

$$\Phi(s, z) \coloneqq 1 - e^{-s} z. \tag{3}$$

This specification satisfies the monotonicity assumptions above: quality is non-decreasing in scale, decreases with dilution, and equals 1 on undiluted evidence.

Substitution, dilution, and curation. We assume that model-assisted production replaces some production that would otherwise have occurred without model assistance. We let P_{hum} denote the distribution of evidence that would be produced without model assistance, and define its quality by $q_{\mathrm{hum}} := 1 - d(P_{\mathrm{hum}}, P^\dagger)$. If model-assisted production has quality below q_{hum}, we model its substitution for the production that would otherwise have occurred as creating new dilution. If its quality is at least q_{hum}, the substitution creates no new dilution. We represent the dilution created

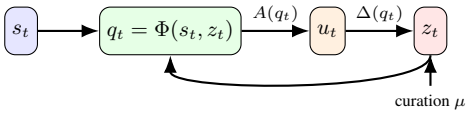


Figure 1: Feedback structure linking scale, production quality, adoption, and evidence dilution.

per unit of model-assisted production by a continuous, non-negative, non-increasing function Δ. We assume that $\Delta(q) > 0$ for $q < q_{\mathrm{hum}}$ and $\Delta(q) = 0$ for $q \geq q_{\mathrm{hum}}$. We refer to the latter case as *parity*. Empirical work showing that AI assistance can change the composition and diversity of collective production motivates allowing model-assisted production to differ systematically from the production it replaces (Doshi & Hauser, 2024; Anderson et al., 2024).

New dilution therefore accumulates at rate $u(t)\Delta(q(\theta_t))(1-z(t))$, where $u(t)$ is the share of model-assisted production, $\Delta(q(\theta_t))$ captures how much dilution that production creates, and $1 - z(t)$ is the remaining undiluted share of the evidence ecology. We also allow curation to remove existing dilution at rate $\mu > 0$, representing processes such as auditing, correction, filtering, removal, or replacement of degraded evidence. The dilution level therefore evolves according to

$$\dot{z}(t) = u(t)\Delta(q(\theta_t))(1 - z(t)) - \mu z(t). \tag{4}$$

Thus, new dilution is created only when model assistance is used, and its production is below parity, while curation removes dilution that has already accumulated. Importantly, we do not require model-assisted production to reach parity before it is adopted. In particular, $A(q)$ may be positive even when $q < q_{\mathrm{hum}}$. Thus there can be a range of qualities in which model assistance is attractive enough to spread while the production it generates still creates new dilution. We also assume that evidence quality remains relevant throughout this range of scales. In particular, if the evidence ecology is fully diluted, then even the highest accessible scale cannot produce model-assisted evidence at parity.

Assumption 1 (Relevance). *At every accessible scale $s \in [0, s_{\max}]$, $\Phi(s, 1) < q_{\mathrm{hum}}$.*

Thus, scaling can improve the quality of model-assisted production, but within the range of scales available to the learner it cannot completely overcome a fully diluted evidence source. For the explicit choice $\Phi(s, z) = 1 - e^{-s}z$, this assumption becomes $1 - e^{-s_{\max}} < q_{\mathrm{hum}}$.

Reduced dilution–adoption dynamics. Since production quality is determined by scale and dilution, let $\bar{\Delta}_s(z) := \Delta(\Phi(s, z))$ and $\bar{A}_s(z) := A(\Phi(s, z))$. Substituting these into eqs. (1) and (4):

$$\dot{z}(t) = u(t)\bar{\Delta}_{s_t}(z(t))(1 - z(t)) - \mu z(t),\ \dot{u}(t) = \kappa \bar{A}_{s_t}(z(t))(1 - u(t)) - \lambda u(t). \tag{5}$$

By the assumptions on Φ, A, and Δ, $\bar{A}_s(z)$ is non-increasing in dilution and non-decreasing in scale, while $\bar{\Delta}_s(z)$ has the opposite monotonicities. Thus, $z(t)$ and $u(t)$ are endogenous and s_t is exogenous; for fixed s, this is an autonomous two-dimensional system on the forward-invariant state space $[0, 1]^2$.

For later use, note that Assumption 1 is equivalently the condition that $\bar{\Delta}_s(1) > 0$ at every accessible scale, or, by monotonicity, that $\bar{\Delta}_s$ is not identically zero. We call a learner operating at scale s *capable* if $\bar{\Delta}_s(0) = 0$, so that on undiluted evidence its model-assisted production creates no new dilution. For the fixed-scale analysis, we use the slightly stronger notion that the learner is *strongly capable* if $\limsup_{z \downarrow 0} \bar{\Delta}_s(z)/z < \infty$. Thus, near the undiluted state, dilution pressure grows at most linearly with existing dilution; in particular, strong capability implies capability.

A canonical explicit instantiation of the reduced system. For some of our theoretical analysis, we use simple explicit forms for the link functions. For adoption, we take $A(q) = q$, with any constant of proportionality absorbed into κ. For dilution, we take $\Delta(q) = (q_{\mathrm{hum}} - q)_+$, so that dilution pressure is proportional to the amount by which model-assisted production falls below parity and is zero once parity is reached. Together with the scale–quality link in eq. (3), these choices give:

$$\boxed{\dot{z} = -\mu z + u(1 - z)(e^{-s}z + q_{\mathrm{hum}} - 1)_+, \qquad \dot{u} = \kappa(1 - e^{-s}z)(1 - u) - \lambda u.} \tag{6}$$

We use this explicit canonical system for the changing-scale analysis below, while the fixed-scale results are first stated for the general reduced system; fig. 1 summarizes the feedback structure.

3 THEORETICAL RESULTS

To answer the question in the introduction, we first characterize what can persist at a fixed scale and how curation changes it; we then ask whether histories reaching the same eventual scale can have different long-run outcomes.

The boundary case of production at or above parity. We begin with the boundary case of a learner operating at such a high scale that, from any arbitrarily diluted evidence ecology, its model-assisted production is at least as good as the unassisted production it replaces. In our notation, this means that the relevance assumption (Assumption 1) is violated. We verify that in this optimistic case, dilution converges to zero from any initial state: thus, persistent degradation is not built into our model. We first check this for the canonical dynamics in eq. (6). In this case, the condition that at every dilution z, the quality of the assisted production mediated by a model learned using evidence with dilution z is at least q_{hum} amounts to having scale s large enough that $1 - \exp(-s) \geq q_{\mathrm{hum}}$. Substituting this in eq. (6) decouples the dynamics of the dilution z from that of the adoption u, and reduces the evolution equation to $\dot{z} = -\mu z$. Hence, at any positive curation rate μ, the dilution stock decreases to 0 as $z(t) = z(0)\exp(-\mu t)$. The same conclusion follows in the general reduced dynamics of eq. (5) as well, once we note that for a powerful learner as above, the reduced quality-dilution link function $\bar{\Delta}_s$ satisfies $\bar{\Delta}_s(z) = \Delta(\Phi(s, z)) = 0$ for all $z \in [0, 1]$ (since, by assumption, $\Phi(s, z) \geq q_{\mathrm{hum}}$ at all z). Hence the evolution of the dilution stock z in the more general dynamics eq. (5) also reduces to $\dot{z} = -\mu z$, and thus converges as above.

Clean and active equilibria. In light of the above, we enforce Assumption 1 in the rest of this section. As described in the introduction, our goal is to understand the behavior of the system as the scale is changed exogenously while the evidence ecology evolves according to our dynamics. An important step towards this is to understand the evolution of the system at a *fixed* scale. To this end, we begin by developing a characterization of the nature of the *equilibrium points* of the system (a point (z_0, u_0) is an *equilibrium point*, or *fixed point*, of the system if the dynamics initialized at (z_0, u_0) stays forever at (z_0, u_0)). In particular, we ask: when are the fixed points of the system *clean* (i.e., have dilution $z = 0$), and when are they *active* (i.e., have strictly positive dilution)? Understanding these active equilibria, and when the dynamics can stay close to them forever, is the main object of our study.

Before turning to the study of the active equilibria, however, we first establish that under mild conditions, there is always a unique clean equilibrium. Assume that the adoption strength κ is positive, and suppose that the learner is operating at a scale s at which it is *capable*. Further, assume that at zero dilution, there is positive adoption pressure, i.e., the reduced adoption link function $\bar{A}_s(0)$ is positive. Set $\dot{z} = \dot{u} = 0$ in eq. (5) at scale $s_t = s$, and solve the resulting system with $z = z(t) = 0$. The solutions to this system are exactly the clean equilibria, and since $\kappa, \bar{A}_s(0)$ are positive by assumption, we find that the only solution is the point

$$(z, u) = \left(0, \tfrac{\kappa \bar{A}_s(0)}{\lambda + \kappa \bar{A}_s(0)}\right), \tag{7}$$

which therefore is the unique clean equilibrium.

The existence of a clean equilibrium, however, does not imply that the ecology must recover to it. The same fixed-scale system may also support an active, positive-dilution equilibrium, and if such an equilibrium is stable, the eventual outcome can depend on the state from which the dynamics begin. The next question is therefore whether curation can rule out such active equilibria.

The critical curation rate. We answer this question by establishing a threshold behavior in the curation rate. We show that there is a critical value μ_c of the curation rate μ, such that when μ is below μ_c, at least one active equilibrium exists, while for $\mu > \mu_c$, the clean equilibrium described in eq. (7) is the only equilibrium.

Theorem 1 (Critical curation rate). *Consider the general reduced dilution-adoption dynamics in eq.* (5)*, and suppose that Assumption 1 is satisfied. Assume that the adoption strength κ is positive, and let s be an accessible scale at which the learner is* strongly capable *and the reduced adoption link function $\bar{A}_s$ is positive-valued. Then, there exists a critical curation rate $\mu_c = \mu_c(s)$ such that (a) if $\mu > \mu_c$, the clean equilibrium is the unique equilibrium; (b) if $0 < \mu < \mu_c$, there is at least one additional active equilibrium.*

We emphasize here, however, that the existence of an active equilibrium and the long-term persistence of degradation in the dynamics are different questions: a degraded equilibrium that repels nearby trajectories would not by itself imply persistent degradation. We return to this point in the next theorem; but first we describe briefly the proof of the above result.

The proof proceeds by showing that a scalar equation of the form $\Gamma(z) = \mu$, characterises the z coordinate of all active equilibria, where Γ is defined by

$$\Gamma(z) \coloneqq \frac{\kappa \bar{A}_s(z)}{\lambda + \kappa \bar{A}_s(z)} \cdot \frac{\bar{\Delta}_s(z)}{z} \cdot (1 - z), \tag{8}$$

and then shows that μ_c can be chosen to be $\sup_{z \in (0,1]} \Gamma(z)$. The three factors in $\Gamma(z)$ can respectively be interpreted as an equilibrium adoption share, the dilution pressure per unit of existing dilution, and the remaining undiluted share. With additional assumptions about the reduced link function $\bar{A}_s$ and $\bar{\Delta}_s$ this analysis can be further sharpened to obtain a precise count of the active equilibria as the curation rate μ varies. In particular, this sharpened analysis applies to the canonical instantiation of the system in eq. (6), and in that setting, it can be shown that $\mu_c(s)$ is decreasing in the scale parameter s (see fig. 2a). We defer these results to Appendices B.3 and B.3.1.

Stability of equilibria. Theorem 1 establishes that active equilibria with degraded evidence ecology (i.e., $z > 0$) exist when the curation rate is low enough. However, our goal is to understand when the trajectory of the system has a persistently degraded ecology at all times (i.e., when $z(t)$ remains bounded away from 0 at all times t). A step towards this is to understand when there exist active equilibria that are *stable* (informally, an equilibrium is *locally asymptotically stable* if trajectories starting sufficiently close to it converge to it; we give the corresponding Jacobian characterization in Appendix B.2). We do this in the following theorem.

Theorem 2 (**Stability of equilibria**). *Consider the same setting as Theorem 1, with the following additional regularity assumptions: $\bar{A}_s$ and $\bar{\Delta}_s$ are both twice continuously differentiable in a neighborhood of every equilibrium point, and their derivatives are bounded in $(0, 1)$. Then (1) the clean equilibrium described in eq.* (7) *is a stable equilibrium whenever $\bar{\Delta}_s'(0) < \mu$, where μ is the curation rate; and (2) there is a critical threshold $\mu_f = \mu_f(s)$ such that $0 < \mu_f \leq \mu_c$ (where $\mu_c = \mu_c(s)$ is the critical threshold from Theorem 1) such that for each $\mu \in (0, \mu_f)$, one of the active equilibria guaranteed by Theorem 1 is a stable equilibrium.*

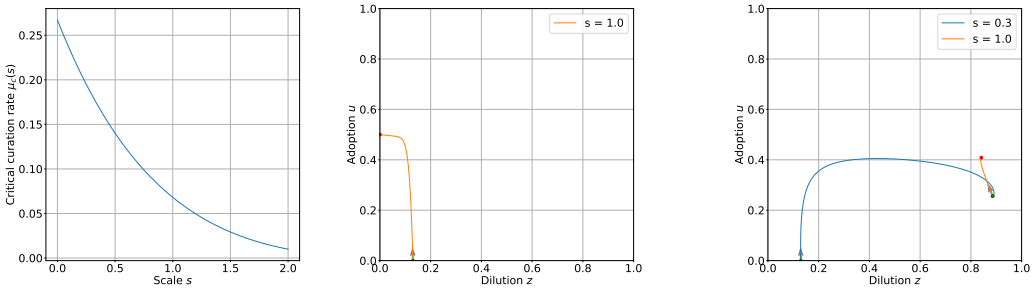
The proof of the above result (in Appendix B.2) uses characterizations of stability of equilibrium points in terms of the eigenvalues of the Jacobian of the vector field associated with the dynamics in eq. (5). In the study of these eigenvalues, an important role is played by the fact that the sign of the determinant of the Jacobian is determined by the sign of the *derivative* of the function Γ used to study critical curation rate in Theorem 1. This is related to the general paradigm of understanding stability of a planar system through the derivative of an appropriately chosen scalar function; see Rosenzweig & MacArthur (1963).

Again, like Theorem 1, Theorem 2 can also be sharpened with additional assumptions on the link functions that hold in the case of the canonical system of eq. (6). In particular, it can be shown that for the canonical dynamics in eq. (6), there is *at most one* stable active equilibrium. We defer the details to Theorem 5 in Appendix B.3.2.

The race between scale and dilution. Armed with an understanding of the equilibria and their stability, we now establish a race between scale and dilution: if a higher-scale learner is introduced late enough, the ecology can remain persistently degraded even though introducing the same higher scale from the outset leads to the clean state. For this demonstration, we use the canonical model of eq. (6).

Theorem 3 (**Scaling-dilution race**). *There exists a configuration of the dynamical system in eq.* (6) *with positive values of the curation rate μ, the adoption strength κ, the abandonment rate λ and $q_{\text{hum}} = 0.95$ such that the following is true. There exist two scales $s_{\text{low}} = 0.3 < 1.0 = s_{\text{high}}$, and an initial point $(z_0, u_0) \in [0, 1]^2$ for which the following holds.*

1. *The dynamics with scale s_{high} started at (z_0, u_0) converges to a clean equilibrium with zero dilution (i.e., $z = 0$).*
2. *There exists a positive dilution level $z^* \geq 0.6$ such that the following is true. The dynamics with scale s_{low} started at (z_0, u_0) reaches a point (z', u') such that the dynamics at scale s_{high} started sufficiently close to (z', u') has dilution at least z^* at all subsequent times.*



a: Critical curation rate b: Starting at high scale c: Scaling up later

Figure 2: *(a)* Representative example of the variation of critical curation rate $\mu_c(s)$ with scale s in the canonical model. *(b)–(c)* **The race between scale and dilution.** Both trajectories start from the same initial condition and have access to the same eventual high scale. Starting immediately at the highest available scale leads to a low-dilution steady state, whereas scaling up only later leads to a high-dilution steady state.

Numerical evidence for the above theorem can be obtained by simply simulating the dynamics using a standard ODE solver (see, e.g., fig. 2). We give, however, an analytic proof of the result in Appendix B.4, which, in addition to enhancing confidence in further numerical exploration of the dynamics between scale and other parameters of the system (as we do in Appendix C), also gives a clearer geometric picture of the phenomenon. The form of the proof also shows that standard results on continuous variation of trajectories of differential equations under variations of initial conditions and parameters (see, e.g. Borkar (2008, Corollary 7 in Section 11.2) and Perko (2001, Theorem 2 in Section 2.3)) already imply that the phenomenon described in Theorem 3 is robust to small perturbations of the initial conditions and the parameters. We now give a short sketch of the proof.

Proof: [**Sketch, see Appendix B.4 for the full proof**] Using information about attracting active equilibria derived from Theorem 2 (and its version for the canonical system, Theorem 5 in the appendix) together with numerical exploration, the proof chooses candidate values for the initial conditions and the parameters at which the phenomenon in the theorem occurs. At these candidate values, both the s_{low}- and s_{high}-dynamics have an attracting active equilibrium, but the s_{high}-dynamics from (z_0, u_0) converges to a clean equilibrium. The proof then has the following steps.

1. The basic strategy is to find regions $R_{\text{low}} \supseteq R_{\text{high}}$ such that (i) R_{low} contains (z_0, u_0); (ii) R_{high} contains the active equilibrium of *both* the s_{low}- and s_{high}-dynamics; and (iii) on the boundaries of R_{low} (respectively, R_{high}), the vector field of the s_{low}-dynamics (respectively, s_{high}-dynamics) points into R_{low} (respectively, R_{high}), so that the two dynamics cannot escape their respective regions once they enter those regions. One could then hope that the s_{low}-dynamics would eventually be attracted by its attracting active equilibrium into R_{high}, from where even the s_{high} trajectory cannot escape.
2. A natural strategy is to define R_{low} and R_{high} as regions to the right of appropriately chosen "vertical" lines parallel to the u-axis in the zu-plane (see fig. 2). However, it turns out that such a choice cannot satisfy condition (iii) above when the curation rate is positive. In the proof, we overcome this by defining the regions instead in terms of lines of carefully chosen negative slopes (see fig. 3 in the appendix).
3. The remaining proof uses two important results from the theory of planar dynamical systems. First, we apply the *Poincaré-Bendixson theorem* to conclude that the s_{low}-dynamics either approaches its attracting active equilibrium, or a potential periodic trajectory. We then exclude the latter possibility by using *Bendixson's criterion* to show that there cannot be any periodic trajectories contained in R_{low}.

A complication in this concluding step is that Bendixson's criterion is only a sufficient criterion, and thus the initial conditions and the parameters have to be chosen carefully so as to satisfy it. However, these choices are robust to small perturbations. □

Computational robustness. The theorem establishes that the timing effect is possible. We next ask how robust this phenomenon is and also how curation changes the window in which scaling can

still recover the clean state. We test the timing mechanism numerically in the full two-dimensional canonical system. Holding the initial state and the eventual high scale fixed, we vary only how long the system evolves at a lower scale before switching. At $\mu = 0.02$, the long-run outcome changes sharply at a critical switching time $T^* \approx 2.56$: switching earlier returns the system to the clean equilibrium, whereas switching later places it in the basin of the stable degraded equilibrium. We then repeat the calculation across curation rates. The critical switching time increases with curation, reaching $T^*(0.068) \approx 7.11$, and disappears at the high-scale critical curation rate $\mu_c(s_{\text{high}}) \approx 0.0681$, above which the degraded active high-scale equilibrium no longer exists. Thus curation enlarges the set of histories from which a later increase in scale can still recover the clean state.

We also test whether these conclusions depend on the canonical functional forms. The timing effect persists under an alternative scale–quality specification in which assisted-production quality on undiluted evidence itself improves with scale. The coexistence of clean and degraded basins also persists under alternative quality–dilution links and when the sensitivity of learner quality to degraded evidence is varied independently from the severity of below-parity production. Full phase portraits, switching-time sweeps, and functional-form checks are reported in Section C.

4 A Controlled Evidence-Ecology Experiment

We next test whether the evidence-feedback mechanism in the model can arise with an actual language-model learner. Rather than estimate a long-run phase diagram or test the changing-scale result, we isolate the proposed channel from deployment to evidence ecology: deployment changes the evidence available to future learners, and those changes can propagate even when learned parameters are not inherited across generations. Learner scale is therefore held fixed throughout. Full details of construction, training, evaluation, controls, and comparative statics are given in Section D.

Experimental ecology. We construct a fixed synthetic world $\mathcal{W}$ containing $5{,}000$ entities and $20{,}000$ atomic facts across twenty relation types. Facts are assigned frequencies according to a fixed long-tailed target distribution $\pi^\dagger$, and a deterministic renderer converts them into natural-language evidence. The initial ecology contains $155{,}000$ evidence units.

As an experimental analogue of the theoretical dilution state, we define $z_k := \mathrm{TV}(\widehat{\pi}_k, \pi^\dagger)$, where $\widehat{\pi}_k$ is the empirical distribution of circulating evidence at generation k. Because the initial ecology is a finite sample from $\pi^\dagger$, $z_0 = 0.1303$. All factual assignments are generated synthetically. Before training, Qwen2.5-1.5B-Base obtains 24.87% accuracy on a frozen four-way evaluation set constructed directly from $\mathcal{W}$, essentially the 25% chance level.

At each generation k, we train a *fresh* copy of Qwen2.5-1.5B-Base on the current evidence ecology $P_k^\star$ using the same three-pass learning rule. The resulting cycle is $P_k^\star \xrightarrow{\mathcal{L}} \theta_k \xrightarrow{\text{production}} P_{k+1}^\star$, with generation $k+1$ again starting from the original pretrained checkpoint. Thus only the circulating evidence is transmitted across generations. Production requests continue to be sampled from the fixed distribution $\pi^\dagger$. Before closing the loop, we verify the two links needed for this mechanism. Training on the initial grounded ecology increases external accuracy from 24.87% to 36.58%, with larger gains for more frequently represented facts. In a matched intervention on 200 common facts, removing every occurrence of a target fact reduces accuracy on those facts from 60.5% to 17.0%, while accuracy on exposure-matched control facts changes from 65.0% to 67.0%. The resulting difference-in-differences is -45.5 percentage points (95% CI $[-55.0, -36.0]$). Learned frequency differences also appear in subsequent model-mediated production; see Section D. Thus available evidence affects both what a fresh learner acquires and what it subsequently produces.

Closing the feedback loop. We then allow model deployment to modify the ecology. Each generation replaces (or *turns over*) a fraction $\beta = 0.20$ of the fixed-capacity corpus, with $u = 0.75$ of new production supplied by the deployed learner and the remainder independently grounded in $\mathcal{W}$. The main outcomes are summarized in table 1.

With zero curation, ecological distance z rises from 0.1303 initially to 0.1997 after one update and 0.2703 after two, while false-evidence mass E rises from zero to 12.43% and then 22.50%. External accuracy q^{ext} of successive fresh learners falls from 36.58% to 34.62% and then 32.91%, with declines across common, medium-frequency, and rare facts.

	Initial	Grounded	Model	Model+cur.	Model (2nd)
z	0.1303	0.1309	0.1997	0.1623	0.2703
E	0	0	12.43%	6.20%	22.50%
q^{ext}	36.58%	35.60%	34.62%	35.61%	32.91%

Table 1: Evidence-ecology outcomes. "Cur." denotes 50% audit-and-repair; all learners start from the same pretrained checkpoint.

Grounded turnover, adoption share, and curation. We first repeat the initial update with the same turnover rate β and production requests but set $u = 0$, so that all replacement evidence is independently grounded. Ecological distance changes only from 0.1303 to 0.1309, false-evidence mass remains zero, and the resulting fresh learner reaches 35.60% external accuracy. Thus turnover alone does not generate the deterioration produced by model-mediated replacement.

We next vary the model-mediated share while holding $\beta = 0.20$ fixed. Using the same production requests and realized model outputs, ecological distance increases monotonically from 0.1309 at $u = 0$ to 0.1510, 0.1741, and 0.1997 at $u = 0.25$, 0.50, and 0.75, while false-evidence mass rises from zero to 4.12%, 8.25%, and 12.43%. Thus deterioration is not specific to $u = 0.75$ but increases with the model-mediated share (Appendix D.6).

Finally, holding $u = 0.75$ and $\beta = 0.20$ fixed, auditing half of model-mediated production reduces ecological distance from 0.1997 to 0.1623 and false-evidence mass from 12.43% to 6.20%. A fresh learner trained on this curated ecology obtains 35.61% external accuracy, essentially matching the 35.60% under grounded turnover.

Together, these experiments show that degradation can propagate through the evolving evidence source even when learned parameters are not inherited, that greater model-mediated substitution produces greater deterioration, and that curation attenuates the effect. Learner scale is held fixed, so the experiment is complementary to, rather than a test of, our theoretical changing-scale result.

5 Conclusion, Limitations, and Future Work

We studied whether increasing model scale can undo degradation in future training evidence when deployment changes that evidence. Our results show that the answer depends on eventual scale, curation, and the path by which scale is reached. At a fixed scale, sufficiently strong curation leaves only the clean equilibrium, whereas weaker curation can sustain a stable equilibrium with persistent dilution. When scale changes over time, reaching a higher scale early can lead to the clean equilibrium, while reaching the same scale after dilution has accumulated can leave a persistent degraded regime. Numerical simulations show that the timing mechanism persists under alternative specifications, while controlled fine-tuning experiments separately show that degradation can propagate through the evolving evidence source without parameter inheritance across generations.

Our model is deliberately simplified. In particular, it compresses a high-dimensional and heterogeneous evidence ecology into a single dilution variable. Therefore, it does not capture which kinds of evidence are displaced or degraded, how pollution may differ across domains, or how different evidence sources may interact. The adoption dynamics are also reduced-form, and scale is treated as exogenous rather than being jointly determined with deployment and the evolving evidence ecology. Our experiments isolate the evidence-feedback channel using controlled fine-tuning and fresh model checkpoints, but do not reproduce the full complexity of long-run deployment in which models, users, institutions, and data-generation processes co-evolve.

These limitations suggest several directions for future work, including models with multiple evidence types, endogenous scale and adoption, and heterogeneous users or institutions. It would also be useful to study how targeted curation, provenance auditing, or preservation of independently generated evidence can change the long-run dynamics. Our results do not imply that AI deployment necessarily degrades future evidence; they explore the consequences of model-assisted production being adopted while remaining below the quality of the evidence it displaces. More broadly, the results suggest that the consequences of scaling depend not only on how capable future models become, but also on the evidence ecology that is preserved while they are being developed and deployed.

AI Use Statement

For the theoretical and modeling components described in Sections 2 and 3, and for the proofs in Appendix B, generative AI tools were used to assist with language editing and manuscript polishing, including identifying grammatical and presentation issues, and to check proofs for possible errors or inconsistencies. They were not used for research ideation, theoretical analysis, or the construction of proofs. For the computational robustness analysis in Appendix C and the controlled experiment in Section 4 and Appendix D, generative AI tools were used for assistance with coding and plotting. The authors reviewed the AI-assisted edits, code, and plots and take responsibility for the final content of the paper.

Ethics Statement

The controlled evidence-ecology experiment described in Section 4 and Appendix D uses a fully synthetic world and synthetically generated factual evidence; it involves no human subjects or personal data. The paper studies potential degradation of future evidence under model-mediated production and interventions such as curation, but does not make claims about the behavior of any particular deployed system or population.

Reproducibility Statement

The model assumptions and dynamical system are specified in Section 2, and the main theoretical results are stated in Section 3, with complete proofs in Appendix B. Appendix C gives the parameter settings and procedures for the computational robustness analyses. Section 4 summarizes the controlled evidence-ecology experiment, with the synthetic world, training protocol, evaluation procedure, controls, and comparative statics described in Appendix D. Code used to generate the computational and experimental results is provided in the 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*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/ebc042e767de551803ccfcc45e2454f5-Paper-Conference.pdf.

Barrett R. Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition (C&C)*, pp. 413 – 425, 2024. doi: 10.1145/3635636.3656204. arXiv:2402.01536.

Joshua Ashkinaze, Julia Mendelsohn, Li Qiwei, Ceren Budak, and Eric Gilbert. How AI ideas affect the creativity, diversity, and evolution of human ideas: Evidence from a large, dynamic experiment. In *Proceedings of the ACM Collective Intelligence Conference (CI)*, pp. 198 – 213, 2025. doi: 10.1145/3715928.3737481. arXiv:2401.13481.

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press and Hindustan Book Agency, 2008.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pp. 6045–6061, 2022. URL https://proceedings.mlr.press/v151/brown22a.html.

Andrea Caravaggio, Luca Gori, and Mauro Sodini. Population dynamics and economic development. *Discrete and Continuous Dynamical Systems - B*, 26(11):5827–5848, 2021. doi: 10.3934/dcdsb.2021178.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11165–11197, 2024. URL https://proceedings.mlr.press/v235/dohmatob24b.html.

Anil R. Doshi and Oliver P. Hauser. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28), 2024. doi: 10.1126/sciadv.adn5290.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Dan 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 *First Conference on Language Modeling (CoLM)*, 2024. URL https://openreview.net/forum?id=5B2K4LRgmz. arXiv:2404.01413.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 30016–30030, 2022. doi: 10.52202/068431-2176.

Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *PMLR*, pp. 9760–9785, 2022. URL https://proceedings.mlr.press/v162/jagadeesan22a.html.

Feiyang Kang, Newsha Ardalani, Michael Kuchnik, Youssef Emad, Mostafa Elhoushi, Shubhabrata Sengupta, Shang-Wen Li, Ramya Raghavendra, Ruoxi Jia, and Carole-Jean Wu. Demystifying synthetic data in LLM pre-training: A systematic study of scaling laws, benefits, and pitfalls. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10739–10758. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.544.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. arXiv:2001.08361.

Joshua Kazdan, Rylan Schaeffer, Apratim Dey, Matthias Gerstgrasser, Rafael Rafailov, David L. Donoho, and Sanmi Koyejo. Collapse or thrive? Perils and promises of synthetic data in a self-generating world. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29469–29494, 2025. URL https://proceedings.mlr.press/v267/kazdan25a.html. arXiv:2410.16713.

Qiang Li and Hoi-To Wai. State dependent performative prediction with stochastic approximation. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3164–3186, 2022. URL https://proceedings.mlr.press/v151/li22c.html.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252. Association for Computational Linguistics, May 2022. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Robert MacArthur. Species packing and competitive equilibrium for many species. *Theoretical Population Biology*, 1(1):1–11, 1970. doi: 10.1016/0040-5809(70)90039-0.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 50358–50376, 2023. doi: 10.52202/075280-2191.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023. doi: 10.1126/science.adh2586.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609, 2020. URL https://proceedings.mlr.press/v119/perdomo20a.html.

Lawrence Perko. *Differential equations and dynamical systems*, volume 7 of *Texts in Applied Mathematics*. Springer, third edition, 2001. doi: 10.1007/978-1-4613-0003-8.

Zeyu Qin, Qingxiu Dong, Xingxing Zhang, Li Dong, Xiaolong Huang, Ziyi Yang, Mahmoud Khademi, Dongdong Zhang, Hany Hassan Awadalla, Yi R. Fung, Weizhu Chen, Minhao Cheng, and Furu Wei. Scaling laws of synthetic data for language models. In *Second Conference on Language Modeling (CoLM)*, 2025. URL https://openreview.net/forum?id=UmUXPXHtdl.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

M. L. Rosenzweig and R. H. MacArthur. Graphical representation and stability conditions of predator-prey interactions. *The American Naturalist*, 97:209–223, 1963. doi: 10.1086/282272.

David G. Schaeffer and John W. Cain. *Ordinary Differential Equations: Basics and Beyond*, volume 65 of *Texts in Applied Mathematics*. Springer, 2016. doi: 10.1007/978-1-4939-6389-8.

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

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 *First Conference on Language Modeling (CoLM)*, 2024. URL https://openreview.net/forum?id=t3z6UlV09o. arXiv:2404.05090.

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.

Ludvig Sinander. The converse envelope theorem. *Econometrica*, 90(6):2795–2819, 2022. doi: 10.3982/ECTA18119.

Ricard Solé, Giulio Ruffini, Francesca Castaldo, Marco Tuccio, Luis F. Seoane, Manlio de Domenico, Santiago F. Elena, David C. Krakauer, and Michael Levin. Large-language models as a cognitive virus, 2026. arxiv:2609.03344.

Anirudh Subramanyam, Yuxin Chen, and Robert Grossman. Scaling laws revisited: Modeling the role of data quality in language model pretraining. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/d0c80a0c294a16190c8904b9809c5fba-Paper-Conference.pdf.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Yun Wan and Yoram M. Kalman. Diverse AI personas can mitigate the homogenization effect in human-AI collaborative ideation. *Computers in Human Behavior: Artificial Humans*, 8:100289, 2026. doi: 10.1016/j.chbah.2026.100289. URL https://www.sciencedirect.com/science/article/pii/S294988212600040X.

Wolfram Research, Inc. Mathematica, Version 14.3, 2025. Champaign, IL.

CONTENTS

1 Introduction **1**

2 Model **3**

3 Theoretical Results **5**

4 A Controlled Evidence-Ecology Experiment **8**

5 Conclusion, Limitations, and Future Work **9**

A Detailed Relation to Prior Work **14**

B Proofs for Section 3 **16**

B.1 Proof of Theorem 1 16

B.2 Proof of Theorem 2 17

B.3 Equilibria for the Canonical System 19

B.3.1 Monotonicity of μ_c in the Canonical Model 21

B.3.2 Stability of Equilibria in the Canonical System 23

B.4 Proof of Theorem 3 24

C Computational Robustness of the Dynamical Mechanism **28**

C.1 Timing of the Scale Increase 28

C.2 Alternative Scale–Quality Link 28

C.3 Curation and the Critical Switching Time 30

C.4 Alternative Quality–Dilution Links 30

C.5 Independent Quality Sensitivity and Dilution Severity 31

D Controlled Evidence-Ecology Experiment: Additional Details **32**

D.1 Synthetic World and Initial Evidence Ecology 32

D.2 Experimental Protocol 32

D.3 Evaluation and Outcome Measures 33

D.4 Calibration and Component-Level Checks 34

D.5 Closed-Loop Feedback and Grounded Control 35

D.6 Comparative Statics 36

Table 2: Selected notation used in the main text.

Symbol	Meaning
t	Continuous time.
$P^\dagger$	Ideal, fully grounded evidence distribution.
P_t	Evidence distribution available for training at time t.
θ_t	Model trained and deployed at time t.
B_θ	Distribution of evidence produced with assistance from model θ.
$q(\theta)$	Quality of model-assisted production.
$s_t;\ s_{\max}$	Learner scale at time t; maximum accessible scale.
$z(t) \in [0,1]$	Dilution of the evidence ecology; $z = 0$ is undiluted.
$u(t) \in [0,1]$	Fraction of production using model assistance.
$\Phi(s, z)$	Quality of model-assisted production at scale s and dilution z.
$A(q)$	Adoption pressure as a function of assisted-production quality.
$\kappa;\ \lambda$	Adoption strength; abandonment rate.
$P_{\mathrm{hum}};\ q_{\mathrm{hum}}$	Evidence produced without model assistance; its quality.
$\Delta(q)$	Dilution created per unit of model-assisted production.
μ	Curation rate.
$\bar{A}_s(z)$	Reduced adoption pressure at scale s and dilution z.
$\bar{\Delta}_s(z)$	Reduced dilution pressure at scale s and dilution z.
$\mu_c(s)$	Critical curation rate at fixed learner scale s.
$s_{\mathrm{low}}, s_{\mathrm{high}}$	Lower and higher scales in the changing-scale analysis.

A Detailed Relation to Prior Work

Recursive training, synthetic data, and model collapse. The closest technical literature studies recursive or self-consuming training loops in which model-generated samples enter the training data of later generations. Repeated training on generated data can progressively lose information about the tails of the original distribution (Shumailov et al., 2024), while related self-consuming loops can reduce quality or diversity when insufficient fresh real data enter later generations (Alemohammad et al., 2024). Other work shows that degradation is not an inevitable consequence of synthetic data itself: replacing real data by successive synthetic generations can produce collapse, whereas accumulating synthetic samples alongside retained real data can keep error bounded (Gerstgrasser et al., 2024); related analyses obtain similar containment phenomena under alternative models and workflows (Seddik et al., 2024; Kazdan et al., 2025). More broadly, recent work emphasizes that "model collapse" covers several distinct phenomena whose occurrence depends on the data-generation and training workflow (Schaeffer et al., 2025).

Our feedback mechanism acts earlier in the pipeline. Recursive-training models typically specify the supply of grounded data and ask what happens when a training procedure mixes that data with model-generated samples. We instead allow deployment to change the production of new evidence itself. Model assistance can substitute for independently grounded production before either source enters a future training corpus. Retaining previously produced grounded data can therefore preserve those observations without recreating new grounded evidence that was displaced by model-mediated production. The controlled experiment in Section 4 isolates this distinction: every generation begins from the same pretrained checkpoint, so learned parameters are not inherited while the evidence ecology is allowed to evolve.

Curation also plays a different role in our model from changing the real–synthetic mixture presented to a learner. It acts on the evolving evidence stock by auditing, repairing, removing, or replacing degraded evidence. Interventions that filter generated data or retain access to grounded data during recursive training are therefore complementary to our analysis; our curation parameter instead governs how quickly degradation already present in the circulating evidence is removed.

Performative prediction and stateful distribution shift. Performative prediction formalizes settings in which deploying a predictor changes the distribution on which that predictor or subsequent predictors operate (Perdomo et al., 2020). Extensions allow the response to depend on a persistent

population state (Brown et al., 2022; Li & Wai, 2022) and study learning when the distributional effects of deployment must themselves be inferred (Jagadeesan et al., 2022). This literature provides an important precedent for treating future data as endogenous to deployment.

Our setting makes the persistent state the evidence ecology generated by production. Higher-quality assistance creates greater adoption pressure, adopted assistance substitutes for independently grounded production, and below-parity production can add dilution that is removed only gradually by curation. The accumulated dilution then affects the quality attainable by later learners. This creates a direct connection between deployment history and future learning: histories that reach the same learner scale can arrive there with different evidence states and consequently lie in different basins of the same higher-scale dynamics. Our changing-scale result therefore concerns not only how deployment changes a contemporaneous distribution, but how increasing learner scale interacts with a distributional state created by earlier deployment.

Human–AI production, productivity, and homogenization. A growing empirical literature studies how generative AI changes human production. Experiments on professional writing find substantial productivity gains from access to language models (Noy & Zhang, 2023). In creative production, AI assistance can improve the evaluated quality of individual outputs while reducing diversity across the collection (Doshi & Hauser, 2024). Related experiments find homogenization in ideation tasks (Anderson et al., 2024), while other studies show that diversity effects depend on exposure, task design, incentives, and the diversity of the model outputs themselves (Ashkinaze et al., 2025; Wan & Kalman, 2026).

These findings motivate the possibility that model-assisted production differs systematically from the independently grounded production it replaces, but our model does not assume a particular source of that difference. Dilution can represent factual error, reduced diversity, genericity, derivative production, weak verification, loss of rare observations, or other departures from the ideal evidence source. Nor does the model assume that AI assistance is intrinsically degrading. If model-assisted production is at least as well grounded as the production it replaces, it creates no new dilution. The relevant primitive is therefore the relative quality of the two production processes, together with the possibility that assistance becomes attractive to users before it reaches parity with the production it displaces.

Scaling laws, data quality, and synthetic data. Scaling laws characterize how performance changes with model size, compute, and data under a supplied training and evaluation distribution (Kaplan et al., 2020; Hoffmann et al., 2022). Work on data-constrained scaling studies how repeated data substitute for additional unique data when the supply of distinct training examples is limited (Muennighoff et al., 2023). More recent work incorporates synthetic data and data quality directly into scaling analyses. Recursive synthetic data can alter the scaling relation across generations (Dohmatob et al., 2024); appropriately constructed synthetic data can itself exhibit predictable scaling (Qin et al., 2025); and large-scale empirical studies find that the effect depends on the composition of synthetic and natural data (Kang et al., 2025). Quality-aware scaling laws further model achievable performance as depending jointly on model size, data quantity, and data quality (Subramanyam et al., 2026).

Our question concerns what happens when deployment at one scale changes the evidence source encountered at later scales. Learner scale and evidence quality are therefore separated in the model: increasing scale can improve attainable production quality for a fixed evidence state, while deployment simultaneously changes that state. This distinction makes the timing of scaling consequential. In our canonical dynamics, the same initial condition and the same eventual higher scale can lead to a clean or degraded long-run outcome depending on how long the system first evolves at a lower scale. The issue is therefore not whether scaling improves the learner relative to the evidence it receives, but whether later improvement can undo changes to the evidence source produced by earlier deployment.

Dynamical-systems perspective. At a more abstract level, our model belongs to the family of low-dimensional dynamical systems in which coupled state variables create feedback, multiple equilibria, changes in stability, and basin boundaries. Related structures appear in ecology, epidemiology, chemical kinetics, and mathematical economics. Very recent work independently studies bistability and hysteresis in a population model of LLM adoption and cognitive dependence (Solé et al., 2026).

The interaction between evidence quality and adoption also has a loose analogy with consumer–resource models (Rosenzweig & MacArthur, 1963; MacArthur, 1970), in which the state of a resource affects an activity, while that activity in turn changes the resource on which it depends. Since these models also often involve planar dynamical systems, there is also an intersection with our work in terms of the nature of technical tools required.

At a conceptual level, however, the analogy with consumer-resource models cannot be taken very far. Adoption in our model does not consume a fixed dilution stock; it changes the production entering the future dilution stock by substituting model-mediated for independently grounded production. The *parity threshold* $\sup\{z|\bar{\Delta}(z) = 0\}$ can also create a region in which adoption generates no new dilution, while learner scale can change exogenously over time. These features produce the fixed-scale coexistence of clean and degraded equilibria and the cross-scale basin geometry used in our timing result. These distinctive features also motivate our specific questions of interest concerning learner scale and curation.

B Proofs for Section 3

B.1 Proof of Theorem 1

Proof: [**Proof of Theorem 1**] We first note that since the learner is *strongly capable* at scale s and Assumption 1 holds, the reduced dilution link function $\bar{\Delta}_s$ is continuous, non-decreasing, non-negative but not identically zero, and is such that $\limsup_{z\downarrow 0} \bar{\Delta}_s(z)/z < \infty$.

Equilibrium points of the reduced system in eq. (5) correspond to points (z, u) where both $\dot{z}$ and $\dot{u}$ are zero. Solving for u at such a point using $\dot{u} = 0$, we get

$$u = \frac{\kappa \bar{A}(z)}{\lambda + \kappa \bar{A}(z)}, \tag{9}$$

and substituting this in $\dot{z} = 0$, we get that a point (z, u) with $z \neq 0$ is an equilibrium point of the reduced system in eq. (5) if and only if u is as in eq. (9), and z solves the equation

$$\mu = \Gamma(z) \coloneqq \frac{\kappa \bar{A}(z)}{\lambda + \kappa \bar{A}(z)} \cdot \frac{\bar{\Delta}(z)}{z} \cdot (1 - z). \tag{10}$$

Thus, since eq. (9) uniquely determines u in terms of z at an equilibrium point, we see that the number of active equilibria is equal to the number of solutions to the equation

$$\mu = \Gamma(z) \tag{11}$$

with $z \in (0, 1]$. We now analyse the number of such solutions. We start by noting that Γ is continuous and non-negative in $(0, 1]$, and further that $\limsup_{z\downarrow 0} \Gamma(z)$ is finite. These facts follow from the assumed continuity and non-negativity of $\bar{A}$ and $\bar{\Delta}$, and the assumption $\limsup_{z\downarrow 0} \bar{\Delta}(z)/z < \infty$ implied by the hypothesis (as discussed at the beginning of this proof). Further, since neither $\bar{A}$ and $\bar{\Delta}$ is identically zero in $(0, 1)$ it follows that Γ is also not identically zero on $(0, 1)$.

It then follows that $\sup_{z\in(0,1]} \Gamma(z)$ is finite and positive. To see this, we note that since $\limsup_{z\downarrow 0} \Gamma(z)$ is finite, there exist positive M and ε such that for $z \in (0, \varepsilon)$, $\Gamma(z) < M$. On the other hand, since Γ is continuous on the compact interval $[\varepsilon, 1]$, it is also bounded above on that interval. This shows that $\sup_{z\in(0,1]} \Gamma(z)$ is finite, and its positivity follows because Γ is not identically zero.

From this and eq. (11), we see that if we define $\mu_c \coloneqq \sup_{z\in(0,1]} \Gamma(z) > 0$ then for $\mu > \mu_c$, eq. (11) has no solution, and hence there are no active equilibria.

Consider now the case when $0 < \mu < \mu_c = \sup_{z\in(0,1]} \Gamma(z)$. By definition of the supremum, there exists a point $z' \in (0, 1)$ such that $\mu_c \geq \Gamma(z') > \mu$. Since Γ is continuous and since $\lim_{z\uparrow 1} \Gamma(z) = 0$, there thus exists a $z^* \in (z', 1)$ such that $\Gamma(z^*) = \mu$. Substituting this z^* in eq. (9) gives u^* such that (z^*, u^*) is an active equilibrium. This completes the proof. □

B.2 Proof of Theorem 2

In the proof of Theorem 2, we need some basic facts from the theory of dynamical systems in $\mathbb{R}^2$, specialized to the setting of the dynamical system in eq. (5), which we review in the following remark.

Remark 1 (Dynamical systems in $\mathbb{R}^2$). *We will use the following facts from the theory of dynamical systems in $\mathbb{R}^2$ (see, e.g. (Perko, 2001, Section 1.5 and Theorem 4 in Section 2.10)): under the twice continuously differentiable assumption on $\bar{A}_s$ and $\bar{\Delta}_s$, an equilibrium point is (1) a* saddle *if the Jacobian of the RHS of the reduced system has two real eigenvalues of opposite signs, (2) a* stable equilibrium *if both eigenvalues of the same Jacobian have strictly negative real parts, (3) an* unstable equilibrium *if both eigenvalues of the same Jacobian have strictly positive real parts, and (4) a* center *if both eigenvalues of the same Jacobian are imaginary.*

Since the determinant and trace of the Jacobian are the product and the sum, respectively, of its eigenvalues, and since the Jacobian has real entries (so that its eigenvalues, if complex, are conjugate to each other), we obtain the following: (1) if the determinant is positive*, and the trace is* negative*, both eigenvalues must have negative real parts, and hence the equilibrium is* stable*. (2) if the determinant is* positive*, and the trace is* positive*, both eigenvalues must have positive real parts, and hence the equilibrium is* unstable*. (3) if the determinant is* positive*, and the trace is* zero*, both eigenvalues must be purely imaginary, and hence the equilibrium is a* center*. (4) if the determinant is* negative*, both eigenvalues must be real with opposite signs, and hence the equilibrium is a* saddle*.*

In the light of the above remark, we start with the following simple lemma that isolates an important property of systems of a form similar to our reduced system. As mentioned in the main paper, our strategy here is related to the paradigm of understanding stability of planar systems through the derivatives of an appropriately chosen scalar function; for us the function Γ in the proof of Theorem 1 plays this role. This paradigm goes back at least to the work of Rosenzweig & MacArthur (1963). See also Schaeffer & Cain (2016, Section 6.2.4) and Caravaggio et al. (2021, Section 3) for related examples.

Lemma 1. *Consider a dynamical system in the (z, u) plane $[0, 1]^2$ given by*

$$\begin{aligned} \dot{z} &= ua(z) - \mu b(z) \\ \dot{u} &= (1-u)h(z) - \lambda u, \end{aligned} \tag{12}$$

where b and h are assumed to be positive-valued, and μ and λ are non-negative reals. Define

$$\Gamma(z) \coloneqq \frac{h(z)}{\lambda + h(z)} \cdot \frac{a(z)}{b(z)}. \tag{13}$$

Let (z^*, u^*) denote an equilibrium of eq. (12) *such that a, b and h are all continuously differentiable in a neighborhood of z^*, and where $a(z^*)$ is also positive. Then the determinant D and the trace T of the Jacobian matrix of the right hand side of eq.* (12) *at the point (z^*, u^*) satisfy*

$$D = -b(z^*)(\lambda + h(z^*)) \cdot \Gamma'(z^*) \text{ and } T = \mu b(z^*) \cdot \left[\frac{a'(z^*)}{a(z^*)} - \frac{b'(z^*)}{b(z^*)}\right] - (h(z^*) + \lambda). \tag{14}$$

Proof: It is easier in terms of notation to write some of the computations for this proof in a more general context. We write the system in eq. (12) as

$$\begin{aligned} \dot{z} &= F_z(z, u) \coloneqq ua(z) - \mu b(z) \\ \dot{u} &= F_u(z, u) \coloneqq (1-u)h(z) - \lambda u. \end{aligned} \tag{15}$$

Here, the functions F_z and F_u are continuously differentiable in a neighborhood of (z^*, u^*).

For any function $G : \mathbb{R}^2 \to \mathbb{R}$, we denote by $G^{(i,j)}$ the partial derivative of G of ith order in the first argument and jth order in the second argument. Consider the equation

$$F_u(z, u) = 0, \tag{16}$$

which is satisfied at $(z, u) = (z^*, u^*)$. Since $F_u^{(0,1)}(z^*, u^*) = -(h(z^*) + \lambda) \neq 0$, the implicit function theorem implies that there is a neighborhood V of z^* and a continuously differentiable function H on V such that every solution to eq. (16) with $z \in V$ satisfies $u = H(z)$ (in particular,

$u^* = H(z^*)$, and in the setting of the lemma, we can explicitly compute $H(z) = \frac{h(z)}{\lambda + h(z)}$ from eq. (16)). Further, for all such z,

$$H'(z) = -\frac{F_u^{(1,0)}(z, H(z))}{F_u^{(0,1)}(z, H(z))}. \tag{17}$$

Using this, we see that the determinant D of the Jacobian matrix of the RHS of eq. (15), evaluated at the point (z^*, u^*), satisfies

$$\begin{aligned} D &= F_z^{(1,0)}(z^*, u^*) F_u^{(0,1)}(z^*, u^*) - F_z^{(0,1)}(z^*, u^*) F_u^{(1,0)}(z^*, u^*) \\ &= F_u^{(0,1)}(z^*, u^*) \cdot (F_z^{(1,0)}(z^*, u^*) + F_z^{(0,1)}(z^*, u^*) H'(z^*)), \end{aligned} \tag{18}$$

where the first equality is the definition and the second uses eq. (17) along with $u^* = H(z^*)$. Now define

$$\Psi(z) \coloneqq F_z(z, H(z)), \tag{19}$$

so that by the chain rule we have, for all $z \in V$, and in particular, for $z = z^*$,

$$\Psi'(z) = F_z^{(1,0)}(z, H(z)) + F_z^{(0,1)}(z, H(z)) H'(z). \tag{20}$$

Substituting this into eq. (18) gives

$$D = \Psi'(z^*) F_u^{(0,1)}(z^*, u^*). \tag{21}$$

We now specialize the situation in the lemma. Using eq. (15), the definition of H following eq. (16), and the definition of Γ in the statement of the lemma, we then have

$$\begin{aligned} H(z) &= \frac{h(z)}{\lambda + h(z)}, \text{ and} \\ \Psi(z) &= b(z) \cdot (\Gamma(z) - \mu). \end{aligned} \tag{22}$$

Since (z^*, u^*) is an equilibrium with $b(z^*) \neq 0$, we also have (from eq. (15)) $\Gamma(z^*) = \mu$, so that $\Psi'(z^*) = b(z^*)\Gamma'(z^*)$. Again using eq. (15), we compute that $F_u^{(0,1)}(z^*, u^*) = -(h(z^*) + \lambda)$. Substituting these in eq. (21) gives the claimed expression for the determinant in eq. (14).

For the trace, we compute

$$T = F_z^{(1,0)}(z^*, u^*) + F_u^{(0,1)}(z^*, u^*) = F_z^{(1,0)}(z^*, u^*) - (h(z^*) + \lambda). \tag{23}$$

Since (z^*, u^*) is an equilibrium, we have $F_z(z^*, u^*) = u^* a(z^*) - \mu b(z^*) = 0$. We thus get

$$F_z^{(1,0)}(z^*, u^*) = u^* a'(z^*) - \mu b'(z^*) = \mu b(z^*) \cdot \left[\frac{a'(z^*)}{a(z^*)} - \frac{b'(z^*)}{b(z^*)} \right]. \tag{24}$$

Substituting this into eq. (23) gives the claimed form for the trace. □

We now prove Theorem 2.

Proof: [**Proof of Theorem 2**] Given Remark 1, it is sufficient to understand the signs of the determinant and trace of the Jacobian of the system at the equilibrium points. We now proceed to do this.

The clean equilibrium. Consider the clean equilibrium $(0, u^*)$ described in Theorem 1. The Jacobian at this point is

$$J = \begin{bmatrix} u^*(\bar{\Delta}'(0) - \bar{\Delta}(0)) - \mu & \bar{\Delta}(0) \\ (1 - u^*) \cdot \kappa \bar{A}'(0) & -(\kappa \bar{A}(0) + \lambda) \end{bmatrix}. \tag{25}$$

The assumption that the learner is *strongly capable* (i.e., $\limsup_{z \downarrow 0} \bar{\Delta}(z)/z < \infty$) implies that $\lim_{z \downarrow 0} \bar{\Delta}(z) = 0$, and the continuity of $\bar{\Delta}$ then implies that $\bar{\Delta}(0) = 0$. Substituting these in eq. (25), we see that the determinant D and the trace T of the Jacobian are

$$D = -(\kappa \bar{A}(0) + \lambda)(u^* \bar{\Delta}'(0) - \mu) \text{ and } T = (u^* \bar{\Delta}'(0) - \mu) - (\kappa \bar{A}(0) + \lambda). \tag{26}$$

It follows that the eigenvalues of the Jacobian are $u^*\bar{\Delta}'(0) - \mu$ and $-(\kappa\bar{A}(0) + \lambda)$, which are both negative since $\mu > \bar{\Delta}'(0)$ and $u^* \in [0, 1]$.

The active equilibria. Consider now the case $\mu \in (0, \mu_c)$, so that, by Theorem 1, there exists at least one active equilibrium $(z^*(\mu), u^*(\mu))$.

We now consider the trace and the determinant of the Jacobian at any such active equilibrium. Applying Lemma 1 with $a(z) = \bar{\Delta}_s(z)(1 - z)$, $b(z) = z$ and $h(z) = \kappa\bar{A}_s(z)$, we note that the function Γ in Lemma 1 coincides with the function Γ defined in eq. (10) in the proof of Theorem 1. Lemma 1 then shows that the determinant and trace are given by $D(z^*(\mu))$ and $T(\mu, z^*(\mu))$, where the functions D and T are defined by

$$D(z) \coloneqq -(\lambda + \kappa\bar{A}(z)) \cdot z \cdot \Gamma'(z), \text{ and} \tag{27}$$

$$T(\alpha, z) \coloneqq \alpha z \cdot \left[\frac{\bar{\Delta}'(z)}{\bar{\Delta}(z)} - \frac{1}{z} - \frac{1}{1-z}\right] - (\kappa\bar{A}(z) + \lambda). \tag{28}$$

From the definition of Γ in eq. (10), we then note that

$$\Gamma'(1) = -\frac{\kappa\bar{A}(1)}{\lambda + \kappa\bar{A}(1)} \cdot \bar{\Delta}(1) < 0, \tag{29}$$

so that $D(1) > 0$. From continuity, it then follows that there exists $\beta' \in (0, 1)$ such that $D(z) > 0$ for all $z \in (\beta', 1)$: we choose $\beta \in [0, 1)$ to be the infimum of all such β'.

Similarly, we note that as $z \uparrow 1$, the term in the square brackets in eq. (28) decreases to $-\infty$: this follows from the boundedness assumption on $\bar{\Delta}'$, and the continuity and monotonicity of $\bar{\Delta}$ combined with the relevance assumption (which together imply that $\bar{\Delta}(1)$ is strictly positive). It thus follows that there exists a $\gamma' \in (0, 1)$ such that for all $z \in (\gamma', 1)$, the term in the square brackets is negative. We define γ to be the infimum of all such γ'. Thus, for all $\mu \geq 0$, $T(\mu, z) < 0$ for all $z \in (\gamma, 1)$.

We then define $\alpha := \max(\beta, \gamma) \in [0, 1)$. With this choice, we have that

$$D(z) > 0 \text{ and } T(\mu, z) < 0 \text{ for all } z \in (\alpha, 1) \text{ and } \mu \geq 0. \tag{30}$$

Now, from $\bar{\Delta}(1) > 0$ and continuity, we also have that there exists a $\tau \in (0, 1)$ such that $\Gamma(z) > 0$ for all $z \in (\tau, 1)$ (even as $\lim_{z\uparrow 1} \Gamma(z) = 0$). We define

$$\mu_f \coloneqq \sup_{z \in (\max(\alpha,\tau), 1]} \Gamma(z). \tag{31}$$

As argued in the proof of Theorem 1, $\mu_c = \sup_{z\in(0,1]} \Gamma(z)$ is finite. It thus follows that $\mu_f \leq \mu_c$. μ_f is also positive (since Γ is positive in the interval $(\max(\alpha, \tau), 1)$).

Now, since $\Gamma(1) = 0$, continuity of Γ implies that there is a solution $z^*(\mu) \in (\max(\alpha, \tau), 1)$ to $\Gamma(z) = \mu$ for every positive $\mu < \mu_f$. As argued in the proof of Theorem 1, every such solution corresponds to an active equilibrium $(z^*(\mu), u^*(\mu))$. By the above definitions, the determinant and the trace of the Jacobian at this equilibrium are $D(z^*(\mu))$ and $T(\mu, z^*(\mu))$. Since $z^*(\mu) \in (\alpha, 1)$, eq. (30) implies that the determinant is positive, while the trace is negative. Hence both eigenvalues of the Jacobian must have strictly negative real parts, and the active equilibrium $(z^*(\mu), u^*(\mu))$ is thus stable. □

B.3 Equilibria for the Canonical System

In this section, we strengthen Theorems 1 and 2 with additional assumptions on the reduced link functions. In particular, these assumptions are satisfied in the explicitly instantiated system in eq. (6).

Theorem 4 (**Critical curation threshold points**). *Consider the general reduced dilution-adoption dynamics in eq.* (5)*. Assume that the adoption strength κ is positive, and let s be a scale such that*

1. *the reduced adoption link function $\bar{A}_s$ is a positive-valued, continuous, non-increasing,* concave *function, and*

2. *the reduced dilution link function $\bar{\Delta}_s$ is continuous, non-decreasing, is not identically zero, and is such that the function $z \mapsto \bar{\Delta}_s(z)/z$ is* log-concave *on $(0,1)$ and satisfies $\lim_{z\to 0} \bar{\Delta}_s(z)/z = 0$.*

Then, there exists a critical curation rate μ_c such that (a) if $\mu > \mu_c$, the clean equilibrium is the unique equilibrium; (b) if $\mu = \mu_c$, there is one additional active equilibrium; (c) if $0 < \mu < \mu_c$, there are exactly two additional active equilibria.

To see how this result applies to the system in eq. (6), we just note that the reduced link functions in the setting of eq. (6) are $\bar{A}_s(z) \coloneqq 1 - ze^{-s}$ and $\bar{\Delta}_s(z) \coloneqq (ze^{-s} - 1 + q_{\mathrm{hum}})_+$, and these satisfy the conditions of Theorem 4 at all positive s.

The proof of this result starts similarly to the proof of Theorem 1 above, with the additional assumptions leading to a more precise account of the equilibrium points being brought to bear later in the proof. For this more delicate analysis, we will need the following elementary facts, whose proofs we include for completeness. To avoid case analyses, we follow the standard convention that outside its domain of definition, a concave function is assumed to take the value $-\infty$ (Rockafellar, 1970, p. 23). Thus, in particular, $\log$ is concave on $[0,\infty)$ and takes the value $-\infty$ at 0.

Lemma 2. *Let I be an interval of non-zero length on the real line. Let $f : I \to [0,\infty)$ be concave and $g : [0,\infty) \to \mathbb{R} \cup \{-\infty\}$ be concave and non-decreasing. Then $h \coloneqq g \circ f$ is concave.*

Proof: For any $x, y \in I$ and $\lambda \in [0,1]$ we have $f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y)$. Since g is non-decreasing, we thus have $g(f(\lambda x + (1-\lambda)y)) \geq g(\lambda f(x) + (1-\lambda)f(y))$. Concavity of g then gives $g(\lambda f(x) + (1-\lambda)f(y)) \geq \lambda g(f(x)) + (1-\lambda)g(f(y))$, completing the proof. □

Lemma 3. *Let a and b be positive real numbers. Then, the function $f : [0,\infty) \to [0,1)$ given by $f(x) = \frac{ax}{b+ax}$ is strictly increasing and concave. The function $h \coloneqq \log \circ f$ is also therefore strictly increasing and concave.*

Proof: We compute the derivative $f'(x) = \frac{ab}{(b+ax)^2}$, which is a positive-valued, strictly decreasing function on $[0,\infty)$. This shows that f is strictly increasing and concave. Since $\log$ and f are both strictly increasing, it thus follows that $h = \log \circ f$ is also strictly increasing. Since, in addition to being increasing, $\log$ is also concave on $[0,\infty)$, it follows from Lemma 2 that h is also concave. □

We are now ready to prove Theorem 4.

Proof: [**Proof of Theorem 4**] We start exactly as in the proof of Theorem 1. Proceeding as done there, we then obtain the same characterizing equation for the equilibria:

$$\mu = \Gamma(z) \coloneqq \underbrace{\frac{\kappa \bar{A}(z)}{\lambda + \kappa \bar{A}(z)}}_{=:g_1(z)} \cdot \underbrace{\frac{\bar{\Delta}(z)}{z}}_{=:g_2(z)} \cdot (1-z). \tag{32}$$

Then, as before, since eq. (9) uniquely determines u in terms of z at an equilibrium point, we see again that the number of active equilibria is equal to the number of solutions to the equation

$$\mu = \Gamma(z) \tag{33}$$

with $z \in (0,1]$. We now analyse the number of such solutions.

We start by noting the following facts.

1. **Γ is continuous and non-negative in $(0,1]$.** This follows from the assumed continuity and non-negativity of $\bar{A}$ and $\bar{\Delta}$.
2. **$\Gamma(1)$ and $\lim_{z\to 0} \Gamma(z)$ are both 0.** $\Gamma(1) = 0$ follows because of the factor $1 - z$, while $\lim_{z\to 0} \Gamma(z) = 0$ follows because of the assumption that $\lim_{z\to 0} \bar{\Delta}(z)/z = 0$. In the following, we extend the definition of Γ continuously to all of $[0,1]$ by defining $\Gamma(0) = \lim_{z\to 0} \Gamma(z) = 0$.
3. **There exists a $z_c \in [0,1)$ such that $\Gamma(z) = 0$ for $z \leq z_c$ (when $z_c = 0$, this condition is vacuously true), and Γ is positive-valued and strictly log-concave on $(z_c, 1)$.**

Define z_c to be the same as the parity threshold $\sup\{z|\bar{\Delta}(z) = 0\}$. Note that the assumption $\lim_{z\to 0} \bar{\Delta}(z)/z = 0$ implies that z_c exists, and the assumptions that $\bar{\Delta}$ is non-decreasing and not identically zero imply that $z_c < 1$ and that $\bar{\Delta}$ is positive-valued in $(z_c, 1)$. Since $\bar{A}$ is assumed to be positive-valued, this implies that Γ is also positive-valued in $(z_c, 1)$.

Now, note that the function $z \mapsto (1 - z)$ is strictly log-concave on $(0, 1)$: for every $\lambda \in (0, 1)$ and $z_1 \neq z_2 \in (0, 1)$, the strict concavity of log implies that $\log(1 - \lambda z_1 - (1 - \lambda)z_2) > \lambda \log(1 - z_1) + (1 - \lambda) \log(1 - z_2)$. To establish that Γ is strictly log-concave in $(z_c, 1)$, we therefore only need to show that the other two factors in the definition of Γ, $g_1(z)$ and $g_2(z)$, are log-concave. The log-concavity of g_2 is exactly the log-concavity assumption on $z \mapsto \bar{\Delta}(z)/z$ in the theorem. For g_1, we note that when $\lambda = 0$, g_1 is the constant function evaluating to 1 and hence is trivially log-concave. When $\lambda > 0$, we consider the function $f(x) = \frac{\kappa x}{\lambda + \kappa x}$, and note that $\log \circ g_1 = (\log \circ f) \circ \bar{A}$. By Lemma 3, the function $\log \circ f$ is strictly increasing and concave on $[0, \infty)$. Lemma 2 and the assumed concavity of $\bar{A}$ then imply that $\log \circ g_1 = (\log \circ f) \circ \bar{A}$ is also concave. Thus, g_1 is log-concave, and, in the interval $(z_c, 1)$, Γ is a product of two positive-valued log-concave functions and one positive-valued strictly log-concave function. It follows that Γ itself is strictly log-concave on $(z_c, 1)$.

We now define

$$\mu_c = \sup_{z\in(0,1]} \Gamma(z). \tag{34}$$

From items 2 and 3 above, it follows that μ_c is positive. Since Γ is continuous on the compact set $[z_c, 1]$ and zero on the set $[0, z_c] \cup \{1\}$, it therefore follows that there exists $z^* \in (z_c, 1)$ such that $\Gamma(z^*) = \mu_c$. Since Γ is strictly log-concave on $(z_c, 1)$, it follows that there is only one such z^* (since if there were at least two, then by strict log-concavity, Γ would exceed μ_c at some point between them, contradicting the definition of μ_c). From the strict log-concavity of Γ, it then also follows that Γ is strictly increasing on $[z_c, z^*]$, and strictly decreasing on $[z^*, 1]$. We can summarize the above as follows:

1. for any $z_1 < z_2 \in (z_c, z^*]$, we have
$$0 = \Gamma(z_c) < \Gamma(z_1) < \Gamma(z_2) < \Gamma(z^*) = \mu_c, \text{ and} \tag{35}$$
2. for any $z_1 < z_2 \in (z^*, 1]$, we have
$$\mu_c = \Gamma(z^*) > \Gamma(z_1) > \Gamma(z_2) > \Gamma(1) = 0. \tag{36}$$

From these, and the preceding discussion that showed that the number of non-zero solutions of $\mu = \Gamma(z)$ determines the number of active equilibria, it follows that

1. When $\mu > \mu_c$, $\mu = \Gamma(z)$ has no solutions in $(0, 1)$, and hence there are no active equilibria.
2. When $\mu = \mu_c$, $\mu = \Gamma(z)$ has exactly one solution, namely z^*, in $(0, 1)$, and hence there is exactly one active equilibrium.
3. When $0 < \mu < \mu_c$, $\mu = \Gamma(z)$ has exactly two solutions in $(0, 1)$, one in (z_c, z^*), and another in $(z^*, 1)$, and hence there are exactly two active equilibria.

This completes the proof. □

B.3.1 Monotonicity of μ_c in the Canonical Model

In this section, we show that in the special case of the canonical model of eq. (6), the critical threshold $\mu_c(s)$ of Theorems 1 and 4 is monotonically decreasing in the scale s for all $s > 0$.

To emphasize the dependence on the scale s, we write the function Γ used in the proofs of Theorems 1 and 4 as a function of both s and z. Thus, in the special case of the canonical model, we have

$$\Gamma(s, z) \coloneqq \frac{(1 - ze^{-s})\kappa}{\lambda + (1 - ze^{-s})\kappa} \cdot \frac{(ze^{-s} - 1 + q_{\text{hum}})_+ \cdot (1 - z)}{z}. \tag{37}$$

We start with the general case considered in the proof of Theorem 4, which showed that for any $s > 0$, there exists a point $z^*(s) \in (0, 1)$ such that

$$\mu_c(s) = \Gamma(s, z^*(s)), \tag{38}$$

where $z^*(s)$ is a maximizer of $\Gamma(s,\cdot)$, and further, in some neighborhood of z^*, $\Gamma(s,\cdot)$ is positive-valued, differentiable and strictly log-concave. From the positivity, we obtain that in a neighborhood of z^*, the factor $(ze^{-s}-1+q_{\rm hum})_+$ in the definition of Γ can be replaced by $(ze^{-s}-1+q_{\rm hum})$. From the differentiability and the fact that $z^*(s)$ is a maximum of $\Gamma(s,\cdot)$ in its neighborhood, we get that $z^*(s)$ solves the equation

$$\Gamma^{(0,1)}(s,z^*(s))=0, \tag{39}$$

where $\Gamma^{(i,j)}$ denotes the partial derivative of Γ of ith order in the first argument and jth order in the second argument. The implicit function theorem then implies that $z^*(s)$ is a differentiable function of s provided

$$\Gamma^{(0,2)}(s,z^*(s))\neq 0, \tag{40}$$

in which case $z^{*\prime}(s)=-\Gamma^{(1,1)}(s,z^*(s))/\Gamma^{(0,2)}(s,z^*(s))$. Now, since $\Gamma(s,\cdot)$ is strictly log-concave in a neighborhood of $z^*(s)$, it follows that

$$0>\partial_z^2\log\Gamma(s,z)\big|_{z=z^*(s)}=\frac{\Gamma^{(0,2)}(s,z)}{\Gamma(s,z)}-\left(\frac{\Gamma^{(0,1)}(s,z)}{\Gamma(s,z)}\right)^2\Bigg|_{z=z^*(s)} \tag{41}$$

$$=\frac{\Gamma^{(0,2)}(s,z^*(s))}{\Gamma(s,z^*(s))}, \tag{42}$$

where the last equality uses eq. (39). Since $\Gamma(s,z^*(s))$ is non-zero, it follows that so is $\Gamma^{(0,2)}(s,z^*(s))$, and hence, as argued above, $z^*(s)$ is a differentiable function of s for all $s>0$.

We can now apply the chain rule in eq. (38) to get

$$\mu_c'(s)=\Gamma^{(1,0)}(s,z^*(s))+\Gamma^{(0,1)}(s,z^*(s))\cdot z^{*\prime}(s)=\Gamma^{(1,0)}(s,z^*(s)), \tag{43}$$

where the last equality again uses eq. (39).[1]

Using the positivity of $\Gamma(s,\cdot)$ in a neighborhood of $z^*(s)$, we get that (recall that in a neighborhood of $z^*(s)$, the factor $(ze^{-s}-1+q_{\rm hum})_+$ in the definition of Γ can be replaced by $(ze^{-s}-1+q_{\rm hum})$)

$$\Gamma^{(1,0)}(s,z^*(s))$$
$$=z^*(s)\Gamma(s,z^*(s))\cdot\left[\frac{e^{-s}}{1-z^*(s)e^{-s}}-\frac{e^{-s}}{z^*(s)e^{-s}-1+q_{\rm hum}}-\frac{\kappa e^{-s}}{\lambda+(1-z^*(s)e^{-s})\kappa}\right]. \tag{44}$$

On the other hand, since z^* is a solution to eq. (39) and $\Gamma(s,z^*(s))\neq 0$, we have

$$0=\frac{\Gamma^{(0,1)}(s,z^*(s))}{\Gamma(s,z^*(s))}$$
$$=-\left[\frac{e^{-s}}{1-z^*(s)e^{-s}}-\frac{e^{-s}}{z^*(s)e^{-s}-1+q_{\rm hum}}-\frac{\kappa e^{-s}}{\lambda+(1-z^*(s)e^{-s})\kappa}\right] \tag{45}$$
$$-\frac{1}{z^*(s)}-\frac{1}{1-z^*(s)}.$$

The bracketed expressions in eqs. (44) and (45) are identical, and by eq. (45), they are both equal to $-\frac{1}{z^*(s)(1-z^*(s))}$. Substituting this in eq. (44) gives

$$\Gamma^{(1,0)}(s,z^*(s))=-\frac{\Gamma(s,z^*(s))}{1-z^*(s)}<0. \tag{46}$$

Thus, from eq. (43), we get $\mu_c'(s)=\Gamma^{(1,0)}(s,z^*(s))<0$, which establishes that $\mu_c(s)$ is a monotone decreasing function of s for all $s>0$.

[1]We note here that eq. (43) can be seen as a special case of the "envelope theorem". A first principles approach like the above is easier for our purpose, and we will not need to invoke this theorem. However, we refer the interested reader to Sinander (2022) for recent developments on the general theorem.

B.3.2 Stability of Equilibria in the Canonical System

With additional assumptions similar to those in Theorem 4, we now have the following sharper version of Theorem 2. Again, as in the case of Theorem 4, the additional assumptions hold in the explicit instantiation of the link functions in eq. (6).

Theorem 5 (Stability). *Consider the same setting as Theorem 4, with the following additional regularity assumptions: $\bar{A}_s$ and $\bar{\Delta}_s$ are both twice continuously differentiable in a neighborhood of every equilibrium point. Then*

1. *The clean equilibrium described in Theorem 4 is always a stable equilibrium.*
2. *For each $\mu \in (0, \mu_c)$, where μ_c is the critical curation rate of Theorem 4, denote by $(z_1(\mu), u_1(\mu))$ and $(z_2(\mu), u_2(\mu))$, with $z_1(\mu) < z_2(\mu)$, the two active equilibrium points. Then*
   (a) *For all $\mu \in (0, \mu_c)$, the point $(z_1(\mu), u_1(\mu))$ is a saddle.*
   (b) *For the point $(z_2(\mu), u_2(\mu))$ there are two cases, depending upon the reduced link functions:*
      i. *$(z_2(\mu), u_2(\mu))$ is a stable equilibrium for all $\mu \in (0, \mu_c)$, or*
      ii. *There exists another threshold $\mu_f \in (0, \mu_c)$ such that for $\mu \in (0, \mu_f)$, $(z_2(\mu), u_2(\mu))$ is a stable equilibrium, while for any $\mu' \in (\mu_f, \mu_c)$ there exists a point $\mu \in (\mu_f, \mu')$ for which $(z_2(\mu), u_2(\mu))$ is either an unstable equilibrium or a center equilibrium.*

Proof: [**Proof of Theorem 5**] As in the proof of Theorem 2, given Remark 1, it is sufficient to understand the signs of the determinant and trace of the Jacobian of the system at the equilibrium points. We now proceed to do this.

The clean equilibrium. Consider the clean equilibrium $(0, u^*)$ described in Theorem 4. As in eq. (25), the Jacobian at this point is

$$J = \begin{bmatrix} u^*(\bar{\Delta}'(0) - \bar{\Delta}(0)) - \mu & \bar{\Delta}(0) \\ (1 - u^*) \cdot \kappa \bar{A}'(0) & -(\kappa \bar{A}(0) + \lambda) \end{bmatrix}. \tag{47}$$

Now, the condition $\lim_{z \to 0} \bar{\Delta}(z)/z = 0$, imposed already in Theorem 4, combined with the continuity of $\bar{\Delta}$ implies $\bar{\Delta}(0) = \bar{\Delta}'(0) = 0$. Substituting this in eq. (47), we see that the determinant D and the trace T of the Jacobian are

$$D = (\kappa \bar{A}(0) + \lambda)\mu \text{ and } T = -\mu - (\kappa \bar{A}(0) + \lambda). \tag{48}$$

It follows that the eigenvalues of the Jacobian are $-\mu$ and $-(\kappa \bar{A}(0) + \lambda)$, which are both negative. Thus, this equilibrium is stable.

The active equilibria. Consider now the case $\mu \in (0, \mu_c)$, so that, by Theorem 4, there are two active equilibria $(z_1(\mu), u_1(\mu))$ and $(z_2(\mu), u_2(\mu))$ with $0 < z_1(\mu) < z_2(\mu) < 1$. We consider these separately.

1. We first consider the "lower" equilibrium $(z_1(\mu), u_1(\mu))$. We apply Lemma 1 with $a(z) = \bar{\Delta}_s(z)(1 - z)$, $b(z) = z$ and $h(z) = \kappa \bar{A}_s(z)$. With this choice, the function Γ in Lemma 1 coincides with the function Γ in the proof of Theorem 4. By Lemma 1, the determinant of the Jacobian is then $-z_1(\mu)(\lambda + \kappa \bar{A}_s(z_1(\mu)))\Gamma'(z_1(\mu))$. As established in the proof of Theorem 4, Γ is strictly increasing and log-concave in a neighborhood of $z_1(\mu)$, so that $\Gamma'(z_1(\mu)) > 0$. Thus, the determinant of the Jacobian is negative, and this equilibrium is a saddle.
2. We next consider the "upper" equilibrium $(z_2(\mu), u_2(\mu))$. We again apply Lemma 1 with $a(z) = \bar{\Delta}_s(z)(1 - z)$, $b(z) = z$ and $h(z) = \kappa \bar{A}_s(z)$, and again, with this choice, the function Γ in Lemma 1 coincides with the function Γ in the proof of Theorem 4. By Lemma 1, the determinant of the Jacobian is then $-z_2(\mu)(\lambda + \kappa \bar{A}_s(z_2(\mu)))\Gamma'(z_2(\mu))$. As established in the proof of Theorem 4, Γ is strictly decreasing and log-concave in a

neighborhood of $z_2(\mu)$, so that $\Gamma'(z_2(\mu)) < 0$. Thus, the determinant of the Jacobian is positive at $(z_2(\mu), u_2(\mu))$.

It thus follows that if the trace of the Jacobian is also negative, then $(z_2(\mu), u_2(\mu))$ is a stable equilibrium. From Lemma 1, the trace of the Jacobian can be written as $T(z_2(\mu))$ where

$$T(z) = \mu z \cdot \left[\underbrace{\frac{\bar{\Delta}'(z)}{\bar{\Delta}(z)} - \frac{1}{z}}_{(\log g_2)'(z)} - \frac{1}{1-z} \right] - (h(z) + \lambda), \tag{49}$$

where g_2 is as defined in eq. (32) in the proof of Theorem 4. Let z^* be the point, as defined in the proof of Theorem 4, where Γ achieves its maximum. It was shown in the proof of that theorem that Γ is strictly decreasing, from μ_c to 0, in the interval $I := (z^*, 1)$ and that for $\mu < \mu_c$, $z_2(\mu)$ is the unique solution in I of $\Gamma(z) = \mu$. It thus follows that $1 > z_2(\mu) > z^*$ (since $\mu < \mu_c$) and further that $z_2(\mu)$ increases to 1 as μ decreases to 0. As also assumed in the hypotheses of Theorem 4, the function g_2 is log-concave, so that the term in the square brackets in eq. (49) is strictly decreasing in $(z_c, 1)$ and approaches $-\infty$ as z increases to 1. It follows from this that there exists a point $\tilde{z} \in (z_c, 1)$ such that $T(z) < 0$ for all $z \geq \tilde{z}$. Consider the point y which is the infimum of all such $\tilde{z}$, and let $\mu_f := \Gamma(y)$.

Now, if $\mu_f = \mu_c$, then $y = z^*$, and we see that $T(z) < 0$ for all $z \in (z_c, 1)$. Since the determinant of the Jacobian is positive, $z_2(\mu)$ thus corresponds to a stable equilibrium for all $\mu \in (0, \mu_c)$. Otherwise, we have $\mu_f < \mu_c$. In this case, we have $T(z) < 0$ for all $z > y = z_2(\mu_f)$ and for any $\varepsilon > 0$, there is a point z in the interval $(y - \varepsilon, y)$ at which $T(z) \geq 0$ (otherwise, the choice of y as an infimum would be contradicted). Translating these statements, we see that the trace is negative (and hence the resulting equilibrium at $(z_2(\mu), u_2(\mu))$ is stable) for $\mu \in (0, \mu_f)$, whereas for any point $\mu' \in (\mu_f, \mu_c)$, there exists a point $\mu \in (\mu_f, \mu')$ at which the trace is either positive or zero, and hence, as the determinant of the Jacobian is positive, the equilibrium is unstable or a center.

This completes the proof of all cases in the statement of the theorem. □

B.4 Proof of Theorem 3

Proof: [**Proof of Theorem 3**] We give an explicit example of a setting of parameters for which the system in eq. (6) satisfies the theorem. We set

$$\begin{gathered} s_{\text{low}} = 0.3, s_{\text{high}} = 1.0, \\ \kappa = \lambda = 0.5, \mu = 0.02 \text{ and } q_{\text{hum}} = 0.95. \end{gathered} \tag{50}$$

Further, we consider trajectories starting at $(z_0, 0) = (0.13, 0)$. Given such an explicit choice, we can *numerically* check using an off-the-shelf ordinary differential equation (ODE) solver (e.g., the `solve_ivp` function in the `scipy` library (Virtanen et al., 2020)), that the claims made in the theorem are satisfied (see figs. 2b and 2c). However, this requires assuming that the effect of errors made by the ODE solver due to round-off and finite step-size are negligible. Thus, we give below a proof in which the only non-trivial numerical computations required are checking whether an explicitly given cubic polynomial has a root in a given interval. In addition to removing the reliance on a numerical ODE solver, this proof also has the advantage that it gives a clearer qualitative understanding of the phenomenon described in the theorem.

We first show item 1 of the theorem. For this, we note that when $z \leq (1 - q_{\text{hum}})e^s$, the evolution equation for z in eq. (6) reduces to $\dot{z} = -\mu z$. Thus, for any trajectory of eq. (6) starting with $z(0) \leq (1 - q_{\text{hum}})e^s$, we have $z(t) = z(0)e^{-\mu t}$ for all $t \geq 0$, and hence the dynamics converges to a clean equilibrium with $z = 0$. To complete the proof of item 1, we check that $z_0 = 0.13$ while $(1 - q_{\text{hum}})e^{s_{\text{high}}} > 0.135$ for our choice in eq. (50), so that the above condition for converging to a clean equilibrium is satisfied at scale s_{high} when the initial state is $(z_0, 0)$.

We now turn to the proof of item 2 of the theorem. Given a scale s, we denote the vector field on the right hand side of eq. (6) as $F_s(z, u)$, so that at all points (z, u), we have $(\dot{z}, \dot{u}) = F_s(z, u)$.

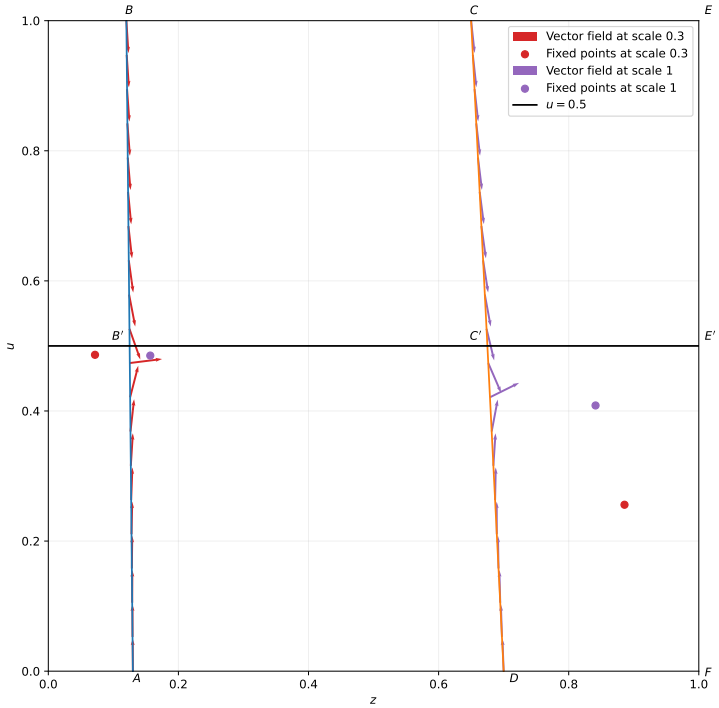


Figure 3: Separators in the proof of Theorem 3

Consider the following special points in $[0,1]^2$ (see fig. 3):

$$A = (z_0, 0) = (0.13, 0), B = (0.12, 1), B' = (0.125, 0.5)$$
$$C = (0.65, 1), C' = (0.675, 0.5), D = (0.70, 0), \tag{51}$$
$$E = (1, 1), E' = (1, 0.5), \text{ and } F = (1, 0).$$

We then have the following observations:

(i) On the segment EF, $\dot{z} = -\mu z < 0$, at all scales s. This follows by substituting $z = 1$ in eq. (6).

(ii) On the segment AF, $\dot{u} = \kappa(1 - e^{-s}z) > 0$, at all scales $s > 0$. This follows by substituting $u = 0$ in eq. (6).

(iii) On the segment $B'E'$, $\dot{u} = -\frac{\kappa}{2}e^{-s}z < 0$, at all scales s. This follows by substituting $u = 1/2$ in eq. (6) and using that $\kappa = \lambda$ in our setting of parameters.

(iv) Fix the scale to $s_{\text{low}} = 0.3$. Then, at any point (z, u) on the segment AB, the vector field $F_{s_{\text{low}}}(z, u)$ points *to the right of the directed segment* AB. Formally, the anticlockwise angle by which the vector $F_{s_{\text{low}}}(z, u)$ has to be rotated to make it parallel (and *not* anti-parallel) to the directed segment AB is *strictly* between $0°$ and $180°$.

To prove this, we consider the vector $v := (-0.01, 1)$ parallel to the directed segment AB. Then, the anticlockwise angle from $F_{s_{\text{low}}}(z, u)$ to v is strictly between $0°$ and $180°$ if and only if

$$\text{the vector cross product } F_{s_{\text{low}}}(z, u) \times v \text{ is } \textit{strictly positive}. \tag{52}$$

Now, any point on AB can be written as $(z_0 - 0.01t, t)$, for $t \in [0, 1]$. Note that every point (z, u) on AB satisfies $z > 0.11 > (1 - q_{\text{hum}})e^{s_{\text{low}}}$, with our choice of parameters. Using this, we see that $F_{s_{\text{low}}}(z, u) \times v$ is a cubic polynomial in t, and this item is true if

and only if that polynomial is strictly positive in $[0, 1]$. By a direct computation (performed here using *Mathematica* (Wolfram Research, Inc., 2025)), this polynomial is

$$-\frac{t^3}{10000e^{3/10}} - \frac{\left(10 + \frac{149}{e^{3/10}}\right) t^2}{20000} + \frac{\left(\frac{2276}{e^{3/10}} - 1066\right) t}{20000} + \frac{48 - \frac{13}{e^{3/10}}}{20000}. \tag{53}$$

That this polynomial is positive in $[0, 1]$ can be checked, e.g., using the symbolic `Reduce` function in *Mathematica*. Alternatively, one can manually check that the derivative of the polynomial (a quadratic polynomial) is positive in $[0, 1]$, and the positivity of the polynomial in the whole interval then follows after checking that it is positive at the end-points 0 and 1.

(v) Fix the scale to $s_{\text{high}} = 1$. Then, at any point (z, u) on the segment DC, the vector field $F_{s_{\text{high}}}(z, u)$ points *to the right of the directed segment* DC.

To prove this, we again consider the same geometric method as in the last item. We start with the vector $v \coloneqq (-0.05, 1)$ parallel to the directed segment DC. Then, the anticlockwise angle from $F_{s_{\text{high}}}(z, u)$ to v is strictly between 0° and 180° if and only if

$$\text{the vector cross product } F_{s_{\text{high}}}(z, u) \times v \text{ is } \textit{strictly positive.} \tag{54}$$

Now, any point on DC can be written as $(0.7 - 0.05t, t)$, for $t \in [0, 1]$. Note that every point (z, u) on DC satisfies $z \geq 0.15 > (1 - q_{\text{hum}})e^{s_{\text{high}}}$, with our choice of parameters. Using this, we see again that $F_{s_{\text{high}}}(z, u) \times v$ is a cubic polynomial in t, and this item is true if and only if that polynomial is strictly positive in $[0, 1]$. By a direct computation (performed here using *Mathematica* (Wolfram Research, Inc., 2025)), this polynomial is

$$-\frac{t^3}{400e} - \frac{(10e - 75)t^2}{4000e} - \frac{(256e - 915)t}{4000e} - \frac{70 - 44e}{4000e}. \tag{55}$$

That this polynomial is positive in $[0, 1]$ can be checked, e.g., using the symbolic `Reduce` function in *Mathematica*. Alternatively, one can again manually check that the derivative of the polynomial is positive in the interval $[0, 1]$, and the positivity of the polynomial then follows from checking that it is positive at the end-points 0 and 1.

(vi) At scale s_{low}, any trajectory starting from a point in the closed quadrilateral $AB'E'F$ must remain inside the quadrilateral at all times in the future.

To see this, note that items (i) to (iv) imply that at all points of the boundary of the quadrilateral $AB'E'F$, the vector field $F_{s_{\text{low}}}(z, u)$ points into the interior of $AB'E'F$. Now, any trajectory that starts from inside $AB'E'F$ and ultimately leaves $AB'E'F$ must cross this boundary of $AB'E'F$ at some time t' (formally, t' is the infimum of those times in the future where the trajectory is outside $AB'E'F$). At this time t', the trajectory must be at some point p on the boundary of $AB'E'F$, and its tangent must not be pointing inside $AB'E'F$ (otherwise, it cannot leave $AB'E'F$). But this is false since this tangent is parallel to the direction of the vector field $F_{s_{\text{low}}}$, and the vector field points inside $AB'E'F$. Thus, we have a contradiction, and a trajectory that starts inside $AB'E'F$ cannot leave it.

(vii) At scale s_{high}, any trajectory starting from a point in the closed quadrilateral $DC'E'F$ must remain inside the quadrilateral at all times in the future.

The argument for this is identical to item (vi), except that it uses item (v) in place of item (iv).

(viii) The dynamical system at scale s_{low} has only one equilibrium point in $AB'E'F$, and this point lies to the right of the directed segment DC (see fig. 3).

To see this we note that for any (z, u) inside $AB'E'F$, $z \geq 0.08 > 0.07 > (1 - q_{\text{hum}}) \exp(s_{\text{low}})$, so that both components of the vector field $F_{s_{\text{low}}}(z, u)$ are polynomials in z and u, and hence continuously differentiable, in an open set containing $AB'E'F$ (the continuous differentiability will be used later in the proof). In particular, following the same argument as in the proof of Theorem 1, we can set up the following cubic equation for the z-coordinates of potential active equilibria of $F_{s_{\text{low}}}$:

$$\mu z(\lambda + \kappa(1 - ze^{-s})) - \kappa(1 - ze^{-s})(e^{-s}z - 1 + q_{\text{hum}})(1 - z) = 0. \tag{56}$$

We can solve this cubic to an arbitrary degree of accuracy, and this yields the claimed location of the only equilibrium point in $AB'E'F$, at roughly the point $(0.88577, 0.25584)$.

So far, our proof has only used elementary arguments. However, to complete the proof, we need to employ two important results from the theory of planar dynamical systems: *Bendixson's criterion* for the non-existence of periodic trajectories in compact sets (see, e.g., (Perko, 2001, Theorem 1 of Section 3.9)), and the *Poincaré-Bendixson theorem* on the classification of limit sets of trajectories (see, e.g., (Perko, 2001, Theorem 1 of Section 3.7)). In preparation for this, we note that as argued in the proof of item (viii), the vector field $F_{s_{\text{low}}}$ is continuously differentiable on an open set containing $AB'E'F$.

Claim 1. *At scale $s = s_{\text{low}}$ and with parameters chosen as in eq.* (50)*, there are no periodic trajectories of the dynamical system of eq.* (6) *that are contained in the closed quadrilateral $AB'E'F$.*

Proof: [**Proof of Claim 1**] We apply *Bendixson's criterion* (Perko, 2001, Theorem 1 of Section 3.9). In our context, the criterion says that there are no periodic trajectories in $AB'E'F$ if for all points (z, u) in $AB'E'F$, the divergence (we write s for s_{low} to avoid clutter)

$$\begin{aligned}\nabla \cdot F_s(z,u) &:= \text{tr}(DF_s(z,u)) = \partial_z \dot{z}(z,u) + \partial_u \dot{u}(z,u) \\ &= (1 - q_{\text{hum}} + e^{-s} - 2ze^{-s}) \cdot u - \mu - (1 - ze^{-s}) \cdot \kappa - \lambda\end{aligned} \tag{57}$$

of the vector field is strictly negative. (Note that in $AB'E'F$, $z \geq (1 - q_{\text{hum}})e^s$, so that $(ze^{-s} - 1 + q_{\text{hum}})_+ = ze^{-s} - 1 + q_{\text{hum}}$.) To verify that this is indeed the case, we note that $u \leq 1/2$ in $AB'E'F$, so that at any point (z, u) in $AB'E'F$,

$$\nabla \cdot F_s(z,u) = (1 - q_{\text{hum}} + e^{-s} - 2ze^{-s}) \cdot u - \mu - (1 - ze^{-s}) \cdot \kappa - \lambda \tag{58}$$

$$\leq \frac{1 - q_{\text{hum}} + e^{-s}}{2} - \mu - \lambda. \tag{59}$$

$$= \frac{0.05 + \exp(-0.3)}{2} - 0.52 < 0, \tag{60}$$

where the equality in eq. (60) comes from substituting the values of the parameters set in eq. (50) and $s = s_{\text{low}} = 0.3$, and the inequality in eq. (59) uses $0 \leq ze^{-s} \leq 1$. Thus, the divergence of the vector field $F_{s_{\text{low}}}$ is strictly negative at all points in $AB'E'F$. Bendixson's criterion as quoted above thus implies that there are no periodic trajectories in $AB'E'F$. □

Consider now the trajectory τ of the dynamical system at scale s_{low} started at $\tau(0) = (z(0), u(0)) = (z_0, 0)$. By item (vi), $\tau(t) \in AB'E'F$ for all $t \geq 0$. We consider the set $\omega(\tau)$ of ω-limit points of τ: this is the set of those points p for which there exists a sequence $t_1, t_2, \ldots$ of times, increasing to infinity, such that $\lim_{i \to \infty} \tau(t_i) = p$.

Since $\tau_+ := \{\tau(t) | t \geq 0\}$ (this notation follows that of Perko (2001, Section 3.7)) is contained in the compact set $AB'E'F$, the *Poincaré-Bendixson theorem* (Perko, 2001, Theorem 1 of Section 3.7) then says that one of the following two conditions must hold:

(a) $\omega(\tau)$ is a periodic trajectory.

(b) $\omega(\tau)$ contains an equilibrium point of the dynamical system.

Now, since τ_+ is contained in the compact set $AB'E'F$, so is the set $\omega(\tau)$ of its limit points. By Claim 1 above, there are no periodic trajectories contained in $AB'E'F$. Thus, it follows that the possibility in item (a) cannot occur.

It follows therefore that the possibility in item (b) must occur, and an equilibrium point p must be a limit point of τ. Since the only equilibrium point in $AB'E'F$ lies to the right of the directed segment DC (from item (viii)), p must be this point. Thus, p lies in the interior of $DC'E'F$. Since p is a limit point of τ, there must thus exist a time t^* at which $\tau(t^*)$ lies in the interior of $DC'E'F$.

Consider now the trajectory τ' of the dynamical system at scale s_{high}, with initial state $\tau(t^*)$. By item (vii), it follows that this trajectory lies in $DC'E'F$ at all subsequent times. Further, since $\tau(t^*)$ lies in the interior of $DC'E'F$, there is a neighborhood N of $\tau(t^*)$ contained in $DC'E'F$. By item (vii), any trajectory started at a point in N must remain in $DC'E'F$ at all subsequent times. Since the z-coordinate is bounded away from 0 for all points in $DC'E'F$, this establishes item 2 of Theorem 3. □

C Computational Robustness of the Dynamical Mechanism

This appendix reports computational checks of the dynamical mechanism studied in the main text. We first examine the timing result directly in the full two-dimensional canonical system, varying only when the learner moves from a lower to a higher scale. We then vary curation to trace how the corresponding critical switching time changes. Finally, we test whether the underlying fixed-scale phase geometry persists under alternative quality–dilution links and when the two channels governing quality degradation and dilution production are varied independently.

C.1 Timing of the Scale Increase

We use the canonical dynamics

$$\Phi(s,z) = 1 - e^{-s}z, \qquad A(q) = q, \qquad \Delta(q) = (q_{\mathrm{hum}} - q)_+,$$

with

$$\kappa = \lambda = 0.10, \qquad q_{\mathrm{hum}} = 0.95, \qquad \mu = 0.02,$$

and compare

$$s_{\mathrm{low}} = 0, \qquad s_{\mathrm{high}} = 1.$$

The common initial dilution is $z_0 = 0.09$, and adoption is initialized at its low-scale equilibrium response,

$$u_0 = \frac{\kappa(1 - e^{-s_{\mathrm{low}} z_0})}{\lambda + \kappa(1 - e^{-s_{\mathrm{low}} z_0})} \approx 0.4764.$$

At the high scale, the clean equilibrium coexists with a saddle at approximately

$$(z_{\mathrm{sad}}, u_{\mathrm{sad}}) = (0.1567, 0.4852)$$

and a stable degraded equilibrium at approximately

$$(z_{\mathrm{deg}}, u_{\mathrm{deg}}) = (0.8413, 0.4085).$$

The stable manifold of the saddle separates their basins of attraction.

Starting immediately at s_{high} converges to the clean equilibrium. We next evolve at s_{low} until time T and then switch permanently to the same s_{high}. An early switch at $T = 1.5$ still lies in the clean basin of the high-scale dynamics and converges to the clean equilibrium. By $T = 4$, however, the low-scale trajectory has crossed the high-scale basin boundary, and the same subsequent scale increase converges to the degraded equilibrium.

Sweeping the switching time makes this transition explicit. On a grid with spacing 0.05, $T = 2.55$ is the last clean outcome and $T = 2.60$ the first degraded outcome. Refining the transition by bisection gives

$$T^*(0.02) \approx 2.56.$$

Thus, even with the initial state and eventual scale held fixed, the long-run outcome can depend on when the higher scale is reached.

C.2 Alternative Scale–Quality Link

The canonical specification

$$\Phi(s,z) = 1 - e^{-s}z$$

has the simplifying feature that an undiluted evidence ecology yields maximal assisted-production quality at every learner scale. We therefore test whether the timing effect persists when scale also improves assisted-production quality on undiluted evidence.

We replace the canonical scale–quality link by

$$\Phi(s,z) = g(s)(1 - Bz), \qquad g(s) = q_0 + (1 - q_0)(1 - e^{-as}),$$

so that quality is increasing in scale and decreasing in dilution, while $\Phi(s,0) = g(s)$ itself increases with scale. We use

$$q_0 = 0.90, \qquad a = 3, \qquad B = 0.40,$$

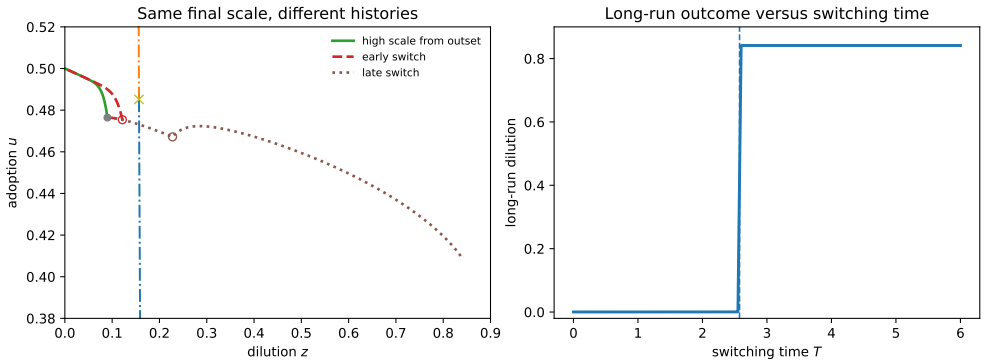


Figure 4: **Timing of scale increase in the full two-dimensional system.** *Left:* All trajectories begin from the same state and eventually use the same high scale. The dash-dotted curve is the stable manifold of the high-scale saddle and hence the basin boundary. The filled circle is the common initial state, open circles mark switching points, and the cross marks the saddle. Switching before the low-scale trajectory crosses the boundary leads to the clean equilibrium; switching after the crossing leads to the degraded equilibrium. *Right:* Long-run dilution as a function of switching time, showing the sharp basin transition.

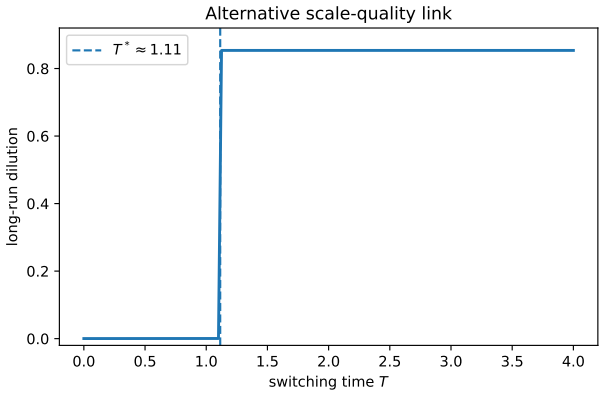


Figure 5: Robustness to the scale–quality link. Long-run dilution as a function of the time at which the learner switches from $s_{\text{low}} = 0$ to $s_{\text{high}} = 1$. Unlike the canonical specification, assisted-production quality on undiluted evidence increases with scale. The sharp transition between clean and degraded long-run outcomes persists.

together with

$$\kappa = \lambda = 0.10, \qquad q_{\text{hum}} = 0.95, \qquad \mu = 0.02,$$

and compare $s_{\text{low}} = 0$ with $s_{\text{high}} = 1$. The common initial dilution is $z_0 = 0.09$, with adoption initialized at its low-scale equilibrium response. Under this specification,

$$\Phi(s_{\text{low}}, 0) = 0.90, \qquad \Phi(s_{\text{high}}, 0) \approx 0.995,$$

so higher scale improves production quality even when the evidence ecology is undiluted.

Starting immediately at s_{high} converges to the clean equilibrium. By contrast, evolving first at s_{low} and switching sufficiently late to the same s_{high} converges to a stable degraded equilibrium. A switching-time sweep gives a critical transition at

$$T^* \approx 1.114.$$

Thus the timing effect does not depend on the canonical specification's assumption that assisted-production quality on undiluted evidence is already maximal at every learner scale.

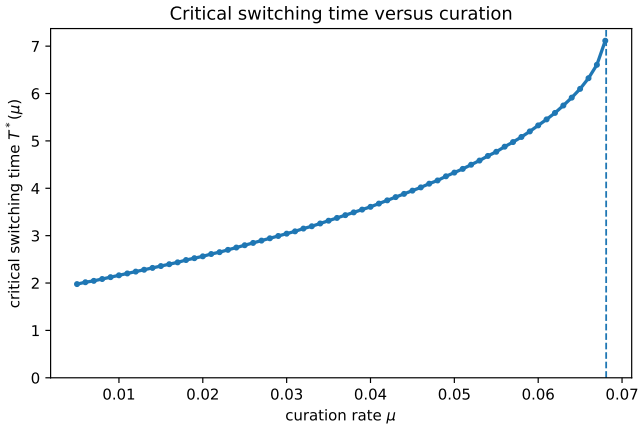


Figure 6: **Critical switching time versus curation.** $T^*(\mu)$ is the latest switch from $s_{\text{low}} = 0$ to $s_{\text{high}} = 1$ that still leads to the clean equilibrium from the common initial state. The critical switching time increases with curation. The dashed vertical line marks the high-scale critical curation rate $\mu_c(s_{\text{high}}) \approx 0.0681$; beyond it the degraded high-scale equilibrium no longer exists.

C.3 Curation and the Critical Switching Time

We next hold the same initial state and scales fixed and vary the curation rate. For each μ, define $T^*(\mu)$ as the latest switching time for which moving from s_{low} to s_{high} still leads to the clean equilibrium.

We locate this boundary numerically by bisection. The critical switching time increases with curation:

$$T^*(0.02) \approx 2.56, \qquad T^*(0.065) \approx 6.10, \qquad T^*(0.067) \approx 6.61, \qquad T^*(0.068) \approx 7.11.$$

Thus stronger curation allows the ecology to remain at the lower scale longer while still remaining recoverable by the same subsequent increase in scale.

The curve terminates at the critical curation rate of the eventual high-scale dynamics. For $s_{\text{high}} = 1$,

$$\mu_c(s_{\text{high}}) \approx 0.068105.$$

Above this threshold the saddle and degraded high-scale equilibrium disappear, so there is no degraded basin for the low-scale trajectory to enter and hence no finite critical switching time.

C.4 Alternative Quality–Dilution Links

The canonical specialization uses the linear below-parity dilution response

$$\Delta(q) = (q_{\text{hum}} - q)_+.$$

To test whether the fixed-scale phase geometry depends on this choice, we replace it by quadratic and smooth threshold-preserving alternatives, normalized so that their overall magnitudes remain comparable. For this experiment we use the fixed-scale quality response

$$\Phi(z) = 1 - Bz, \qquad B = 0.40,$$

with

$$\mu = 0.02, \qquad \kappa = \lambda = 0.10, \qquad q_{\text{hum}} = 0.95.$$

As shown in fig. 7, all three specifications retain the same qualitative phase structure: a clean equilibrium, a lower saddle whose stable manifold forms the basin boundary, and an upper stable degraded equilibrium. The boundary moves quantitatively across specifications, but the coexistence of clean and degraded long-run regimes does not depend on the linear link.

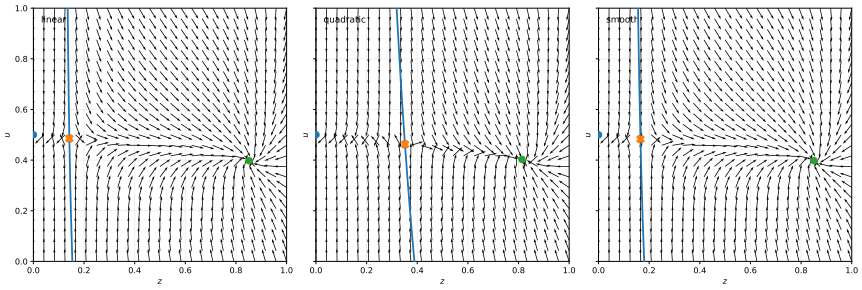


Figure 7: **Robustness to the quality–dilution link.** Linear, quadratic, and smooth threshold-preserving specifications retain the clean equilibrium, saddle basin boundary, and degraded equilibrium, although the quantitative location of the boundary changes.

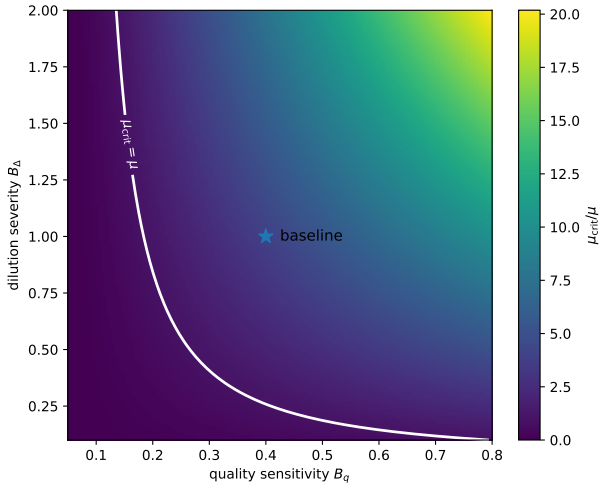


Figure 8: **Independent quality sensitivity and dilution severity.** B_q controls the sensitivity of learner quality to the evidence ecology, while B_Δ controls the severity of dilution generated by a below-parity quality gap. The plotted quantity is μ_c/μ; the white contour marks $\mu_c = \mu$.

C.5 Independent Quality Sensitivity and Dilution Severity

The canonical specialization also ties together two conceptually distinct channels: how strongly dilution lowers attainable learner quality and how much new dilution a given below-parity quality gap generates. We separate them by writing

$$\Phi(z) = 1 - B_q z, \qquad \Delta(q) = B_\Delta (q_{\text{hum}} - q)_+.$$

Here B_q controls the sensitivity of learner quality to dilution, while B_Δ controls the severity of dilution production. For each pair (B_q, B_Δ), the fixed-scale critical curation rate is

$$\mu_c(B_q, B_\Delta) = \max_{z \in (0,1]} \frac{u(z)\Delta(\Phi(z))(1-z)}{z}.$$

where $u(z)$ is the equilibrium adoption share at dilution z, satisfying $\dot{u} = 0$. The parameter map in fig. 8 shows that μ as chosen in the previous section is below the critical curation rate over a broad range of independent values of B_q and B_Δ. The reference parameterization

$$(B_q, B_\Delta) = (0.40, 1)$$

used in the last section lies well inside the region with a positive equilibrium pair. Hence the coexistence of clean and degraded regimes is not a consequence of using the same parameter to govern both the effect of dilution on learner quality and the severity of subsequent dilution production.

Summary. These experiments support the two computational claims used in the main text. First, in the full two-dimensional system, delaying an otherwise identical increase in scale can move the ecology across the basin boundary of the eventual high-scale dynamics, and stronger curation postpones this transition until the degraded basin disappears at the high-scale critical curation rate. Second, the underlying coexistence of clean and degraded regimes persists under alternative quality–dilution links and when the two channels governing degradation are varied independently. The mechanism therefore does not depend on an infinite-scale limit, a scalar approximation, or a single functional-form parameterization.

D Controlled Evidence-Ecology Experiment: Additional Details

This appendix provides the construction, training protocol, evaluation procedure, calibration experiments, and additional results for the controlled evidence-ecology experiment in Section 4. The purpose of the experiment is to isolate inheritance through the circulating evidence source from inheritance through learned model parameters. It does not test the changing-scale result: learner scale is held fixed throughout the experiment.

The theoretical dynamics are continuous, whereas training, deployment, and evidence replacement occur experimentally in discrete generations $k = 0, 1, 2, \ldots$. At each generation, a fresh copy of the same pretrained model is trained using the same learning rule on the current evidence ecology. The trained model then produces a fraction of the evidence entering the next generation. No learned weights are inherited across generations, so the only persistent state is the circulating evidence ecology.

D.1 Synthetic World and Initial Evidence Ecology

We generate a fixed synthetic world $\mathcal{W}$ containing $5{,}000$ entities and $20{,}000$ atomic facts over twenty relation types. An atomic fact has the form $f = (e, r, v)$, where e is an entity, r is a relation, and v is the associated value; examples include $(\text{Dorin}, \texttt{capital}, \text{Talven})$ and $(\text{Dorin}, \texttt{primary-export}, \text{meralite})$. Entity names, values, and their assignments are synthetically generated and fixed throughout the experiment.

Let $\mathcal{F} = \{f_1, \ldots, f_N\}$, with $N = 20{,}000$, denote the set of world facts. Each fact is assigned an ideal frequency $\pi^\dagger(f)$ according to a fixed long-tailed distribution. We partition facts by frequency rank into common (top 10%), medium (next 30%), and rare (remaining 60%) strata. The frequency distribution and the assignments of values to entities are held fixed across all generations and experimental conditions.

A deterministic renderer maps each atomic fact to multiple natural-language realizations while retaining the identity of the underlying fact. Because the world and renderer are known exactly, factual claims can be evaluated directly against $\mathcal{W}$ without using another language model as a judge.

The initial grounded corpus contains $M = 155{,}000$ sampled fact occurrences. Five factual sentences are grouped into each training document, yielding $31{,}000$ documents and $2{,}006{,}377$ tokens under the Qwen2.5-1.5B tokenizer. The individual factual sentence is the basic evidence unit; grouping five units into a document is only a training-format choice.

We denote the realized initial evidence ecology by $P_0^\star$. Because $P_0^\star$ is a finite realization of the ideal sampling process, its empirical fact distribution $\widehat{\pi}_0$ does not coincide exactly with $\pi^\dagger$. In the fixed realization used throughout the experiments, $\mathrm{TV}(\widehat{\pi}_0, \pi^\dagger) = 0.130308$. The nonzero initial distance therefore reflects finite-sample variation rather than contamination of the initial ecology.

D.2 Experimental Protocol

Fresh-checkpoint learning. Let s_{exp} denote the fixed learner scale used throughout the experiment, corresponding to Qwen2.5-1.5B-Base with the fixed training procedure described below, and let θ_{base} denote the original pretrained checkpoint. At generation k, the learner is

$$\theta_k = \mathcal{L}_{s_{\text{exp}}}\left(P_k^\star; \theta_{\text{base}}\right). \tag{61}$$

Generation $k+1$ again begins from θ_{base}, rather than from θ_k. Thus both learner scale and the learning rule are held fixed across generations; only the evidence ecology supplied to the learner changes. Architecture, optimizer, training budget, learning rate, sequence length, and randomization procedure are held fixed across generations.

Thus the experimental inheritance channel is $P_k^\star \to \theta_k \to P_{k+1}^\star$, rather than $\theta_k \to \theta_{k+1}$. The only state passed from one generation to the next in the main experiment is the evidence ecology.

For the reported experiments, the learner is Qwen2.5-1.5B-Base and the frozen learning operator consists of three complete passes of ordinary causal-language model training over the current ecology. The learning rule is calibrated once on $P_0^\star$ and is not subsequently retuned as the ecology changes.

Evidence production, turnover, and curation. At each generation, production requests are sampled independently from the fixed ideal fact distribution $\pi^\dagger$. The demand for facts therefore does not change across generations.

For each production request, evidence is supplied either by an independently grounded producer or by the deployed learner. If u_k is the model-mediated production share, the distribution of newly produced evidence is

$$P_{k+1}^{\text{new}} = (1-u_k)P_{\text{hum}} + u_k B_{\theta_k}. \tag{62}$$

Here P_{hum} produces evidence grounded directly in $\mathcal{W}$, while B_{θ_k} is the distribution of claims generated by the learner trained on $P_k^\star$. Thus, the effect of model-mediated production is *substitutive*: increasing u_k replaces independently grounded production rather than adding model output on top of an unchanged grounded stream.

The evidence ecology has fixed capacity $M = 155{,}000$. A fraction β of evidence slots *turns over* in each generation, so that before curation,

$$P_{k+1}^{\text{raw}} = (1-\beta)P_k^\star + \beta P_{k+1}^{\text{new}}. \tag{63}$$

In the main feedback experiment, $\beta = 0.20$, so $31{,}000$ atomic evidence units are refreshed in each generation.

For curation treatments, a fraction m of circulating evidence units is audited against the known world $\mathcal{W}$. Audited false claims are repaired using independently grounded evidence. The experimental audit rate m is a discrete analog of the curation parameter μ in the theoretical model, but we do not calibrate m to the continuous-time rate μ. Experimental curation operates specifically by auditing and repairing false claims, whereas μ represents curation acting on the broader dilution state.

The complete experimental cycle is $P_k^\star \to \theta_k \to P_{k+1}^{\text{new}} \to P_{k+1}^{\text{raw}} \to P_{k+1}^\star$, corresponding respectively to learning, production, turnover, and curation.

D.3 Evaluation and Outcome Measures

Frozen externally anchored evaluation. We evaluate learned factual knowledge against a fixed test set constructed directly from $\mathcal{W}$. For a fact such as $(\text{Dorin}, \texttt{capital}, \text{Talven})$, the causal language model receives a completion prefix such as "The capital of Dorin is" together with four candidate continuations.

A naive multiple-choice construction creates a token-length confound because summed conditional log likelihood tends to favor shorter candidate strings. We therefore match distractors to the correct answer by relation and object type and, for the primary evaluation, also by continuation-token length under the learner's tokenizer.

Exact four-way token-length matching is possible for $18{,}750$ of the $20{,}000$ facts. These questions form the frozen externally anchored evaluation set $E^\dagger$. The remaining $1{,}250$ facts are retained for robustness analysis. Because all candidate continuations in a primary question have equal token length, prediction is $\widehat{a}(x) = \arg\max_{a \in A(x)} \log p_\theta(a \mid x)$.

Before any exposure to the synthetic ecology, Qwen2.5-1.5B-Base obtains 24.87% accuracy on $E^\dagger$, compared with a four-way chance level of 25%. Performance is 25.58% on common facts, 24.57% on medium facts, and 24.90% on rare facts. Thus the untouched pretrained checkpoint has no detectable knowledge of the random synthetic assignments.

For each generation k, we denote externally anchored learner performance by $q_k^{\mathrm{ext}} = \mathrm{Perf}(\theta_k; E^\dagger)$, the accuracy of θ_k on $E^\dagger$. The evaluation set and underlying world remain fixed throughout, so changes in q_k^{ext} measure changes in knowledge of the same external target, rather than changes in fit to the contemporaneous circulating evidence. This externally anchored accuracy is a diagnostic of what the learner has acquired, and can be seen as an analogue of the theoretical assisted-production quality $q(\theta_k)$. It is not, however, a direct measurement of $q(\theta_k)$, which is more closely related to the distribution B_{θ_k}.

Ecological outcome measures. Let $\widehat{\pi}_k$ denote the empirical distribution of true world facts expressed in $P_k^\star$, with evidence units making claims false in $\mathcal{W}$ assigned to an additional error category. As an experimental analogue of the theoretical dilution state, we use

$$z_k = \mathrm{TV}(\widehat{\pi}_k, \pi^\dagger). \tag{64}$$

We additionally report C_k, the fraction of world facts represented in $P_k^\star$; C_k^{rare}, the fraction of rare facts represented; E_k, the fraction of evidence units whose claims are false in $\mathcal{W}$; and H_k, the entropy of the represented fact distribution.

These measures distinguish several ways in which an evidence ecology can deteriorate. E_k captures outright factual error, while C_k and C_k^{rare} capture disappearance of support even when the remaining claims are true. The total-variation metric z_k summarizes both changes relative to the fixed target distribution.

D.4 Calibration and Component-Level Checks

Calibration of the learning operator. Before introducing feedback, we calibrate the learner on the initial grounded ecology $P_0^\star$. The architecture, optimizer, learning rate, batch size, sequence length, seed, and training corpus are held fixed while only the number of corpus passes is varied. The resulting learning curve is monotone and strongly dependent on fact frequency.

Table 3: Calibration of the fresh-checkpoint learning operator on the initial grounded evidence ecology.

Corpus passes	Overall	Common	Medium	Rare
0	24.87	25.58	24.57	24.90
1	30.60	50.53	30.90	27.09
2	33.96	59.44	35.74	28.77
3	36.58	64.72	38.54	30.86

Three passes increase common-fact accuracy by approximately 39 percentage points relative to the pretrained checkpoint, medium-frequency accuracy by approximately 14 points, and rare-fact accuracy by approximately 6 points. Performance remains well below saturation, leaving room for subsequent changes in the evidence ecology to affect learned knowledge.

We therefore freeze $\mathcal{L}_{s_{\mathrm{exp}}}$ as a fresh Qwen2.5-1.5B-Base checkpoint trained for three complete passes over the current evidence ecology.

Evidence-to-learning check. We next validate directly that factual knowledge responds to the information available in the evidence ecology.

We construct 200 target–control pairs of common facts. Each target fact is matched to a control fact having exactly the same number of occurrences in $P_0^\star$. We then remove every occurrence of the target fact while leaving its matched control unchanged, and train a fresh learner using the same frozen three-pass procedure.

Accuracy on the removed target facts falls from 60.5% to 17.0%, a decline of 43.5 percentage points, while accuracy on the exactly exposure-matched controls changes from 65.0% to 67.0%. The resulting difference-in-differences is -45.5 percentage points, with a paired-bootstrap 95% confidence interval of $[-55.0, -36.0]$ percentage points. A paired randomization test gives $p < 10^{-5}$.

The transition counts show the same asymmetry. Among the 200 removed facts, 89 change from correct to incorrect and only 2 change from incorrect to correct. Among matched controls, the corresponding counts are 20 and 24.

Hence reducing evidence for a fact selectively reduces what a fresh learner knows about that fact. This establishes the evidence-to-learning link before any model-mediated production is introduced.

Learning-to-production check. We separately verify that learned knowledge affects the evidence subsequently produced by the learner.

Production requests are sampled from the fixed request distribution $\pi^\dagger$. The learner is given the requested subject and relation without being shown the correct value. Production uses deterministic greedy decoding, and generated claims are scored directly against $\mathcal{W}$.

For the learner trained on $P_0^\star$, we evaluate 23,250 model-mediated production requests. The generated claim is correct on 17.15% of requests. Production accuracy is strongly frequency dependent: 38.77% on common facts, 4.33% on medium-frequency facts, and 2.63% on rare facts.

Incorrect generations are generally structured rather than arbitrary. On 66.94% of all requests, the model emits a value admissible for the requested relation but associated with the wrong subject. Thus the learner has acquired substantial relation-level regularity while failing to preserve the correct subject–value assignment.

The frequency dependence in learned knowledge therefore propagates into model-mediated evidence production. Together, the evidence-to-learning and learning-to-production checks establish the two component links from available evidence to learned knowledge and from learned knowledge to newly produced evidence.

D.5 Closed-Loop Feedback and Grounded Control

Closed-loop feedback. We now close the feedback loop while keeping learner scale, the underlying world, request distribution, ecology size, and training procedure fixed. The main high-adoption, zero-curation treatment uses $\beta = 0.20$, $u = 0.75$, and $m = 0$. Thus 20% of the ecology turns over in each generation, and 75% of new production is model mediated.

Starting from $z_0 = 0.130308$, ecological distance rises to $z_1 = 0.199670$ after one model-mediated update and to $z_2 = 0.2703$ after the second. False-evidence mass rises from zero initially to $E_1 = 12.427\%$ and then to $E_2 = 22.50\%$.

Coverage also declines. Overall world-fact coverage is 97.77% after the first update and 96.48% after the second, while rare-fact coverage falls from 96.44% to 94.43%.

The deterioration in circulating evidence is transmitted to successive fresh learners. External accuracy falls from 36.58% on $P_0^\star$ to 34.62% on $P_1^\star$ and 32.91% on $P_2^\star$. The decline is present in every frequency stratum:

	$P_0^\star$	$P_1^\star$	$P_2^\star$
Common	64.72%	60.13%	56.91%
Medium	38.54%	36.51%	34.16%
Rare	30.86%	29.38%	28.24%
Overall	36.58%	34.62%	32.91%

Model-mediated production moves in the same direction. Production accuracy is 17.15% for the learner trained on $P_0^\star$ and 15.75% for production from the learner trained on $P_1^\star$. These production figures use independently sampled requests and are therefore descriptive rather than a matched comparison. The frozen external evaluation provides the direct like-for-like measure of degradation in what successive fresh learners acquire.

Grounded-turnover control. To separate model-mediated degradation from ordinary finite-sample turnover, we repeat the first ecological update using the same 31,000 production requests and the same turnover rate, but set $u = 0$. All replacement evidence is then independently grounded in $\mathcal{W}$.

Under grounded-only turnover, ecological distance changes only from 0.130308 to 0.130882, and false-evidence mass remains exactly zero. Under model-mediated turnover with $u = 0.75$, the corresponding first update instead gives $z_1 = 0.199670$ and $E_1 = 12.427\%$.

Grounded turnover also preserves coverage more closely: overall world-fact coverage is 98.70% and rare-fact coverage is 97.92%, compared with 97.77% and 96.44%, respectively, under model-mediated turnover.

Finite grounded turnover nevertheless changes the exact empirical corpus and therefore induces some variation in downstream learning. A fresh learner trained after grounded-only turnover obtains 35.60% external accuracy, compared with 36.58% on the original grounded realization and 34.62% after uncurated model-mediated turnover.

Thus, relative to the matched grounded-turnover control, model-mediated substitution produces an additional first-generation decline of approximately 0.98 percentage points in external accuracy, together with a much larger divergence in the evidence ecology itself.

D.6 Comparative Statics

Adoption-share comparative static. The main feedback experiment uses a model-mediated production share of $u = 0.75$. To check whether the resulting deterioration depends on this relatively high adoption level, we repeat the first ecological update while holding the total turnover rate fixed at $\beta = 0.20$ and varying $u \in \{0, 0.25, 0.50, 0.75\}$.

Across conditions, we use the same $31{,}000$ production requests and the same refreshed corpus slots. Requests assigned to model-mediated production use the same realized model outputs as in the baseline experiment, while the remaining requests are filled by independently grounded production from $\mathcal{W}$.

Table 4: Effect of the model-mediated production share on the first ecological update. Total turnover is fixed at $\beta = 0.20$ and curation at $m = 0$; production requests and refreshed corpus slots are held fixed across conditions.

u	z_1	E_1	C_1	C_1^{rare}
0	0.130882	0	98.700%	97.925%
0.25	0.150983	4.115%	98.480%	97.583%
0.50	0.174079	8.245%	98.130%	97.025%
0.75	0.199670	12.427%	97.770%	96.442%

The ecological response is monotone over this sweep. As the model-mediated share increases, ecological distance and false-evidence mass increase, while both overall and rare-fact coverage decline. Thus the deterioration observed at $u = 0.75$ is not specific to the baseline adoption level: it is already present at lower model-mediated shares and becomes larger as a greater share of new evidence is model mediated.

Curation comparative static. We next vary factual curation while keeping deployment and production fixed. Specifically, we set $u = 0.75$ and $\beta = 0.20$ and apply audit-and-repair rates $m \in \{0, 0.25, 0.50, 0.75\}$ to the same first-generation turnover and production realization. Audited evidence is checked directly against $\mathcal{W}$, and false audited claims are replaced by independently grounded evidence. Table 5 shows that the ecological response is monotone in this sweep as well. In particular, increasing curation moves the realized ecology monotonically toward the fixed target distribution, reduces false-evidence mass, and increases both overall and rare-fact coverage.

To test whether this ecological improvement is transmitted downstream, we train a fresh checkpoint on the $m = 0.50$ ecology using the same frozen three-pass learning rule. Its external accuracy is 35.61%, compared with 34.62% under otherwise identical uncurated model-mediated turnover and 35.60% under grounded-only turnover. The same pattern holds across frequency strata:

Thus 50% curation recovers essentially the entire first-generation performance gap relative to the grounded-turnover benchmark, both in aggregate and within each frequency stratum.

Table 5: Effect of curation on the first model-mediated ecological update. Deployment, turnover, requests, and production realization are held fixed.

m	0	0.25	0.50	0.75
z_1	0.199670	0.180137	0.162323	0.145659
E_1	12.427%	9.325%	6.201%	3.088%
C_1	97.770%	98.055%	98.295%	98.550%
C_1^{rare}	96.442%	96.875%	97.267%	97.683%

Table 6: External accuracy after the first turnover under grounded production, uncurated model-mediated production, and 50% curation.

Condition	Overall	Common	Medium	Rare
Grounded turnover ($u = 0$)	35.60%	61.87%	37.31%	30.31%
Model turnover ($u = .75, m = 0$)	34.62%	60.13%	36.51%	29.38%
Model turnover ($u = .75, m = .50$)	35.61%	61.76%	37.51%	30.25%

Summary of experimental identification. The experimental design separates the components of the evidence-feedback mechanism while holding learner scale fixed. First, the controlled-removal experiment shows that changing the evidence available for a fact changes what a fresh learner acquires about that fact. Second, the production experiment shows that differences in acquired knowledge are reflected in subsequently generated evidence.

Third, the closed-loop feedback experiment shows that model-mediated substitution shifts the circulating ecology away from a fixed external world and that successive fresh learners trained on this degraded ecology exhibit lower externally anchored accuracy. Fourth, the grounded turnover control shows that this effect is not merely ordinary finite-sample corpus replacement. Fifth, varying the model-mediated production share shows that ecological deterioration increases with the extent of model-mediated substitution. Finally, factual audit-and-repair moves the ecology back toward the fixed world and correspondingly restores downstream learner performance.

Across the closed-loop generations, $P_0^\star \to \theta_0 \to P_1^\star \to \theta_1 \to P_2^\star \to \theta_2$, with each θ_k trained from the same original pretrained checkpoint and at the same learner scale. The persistent state is therefore the evidence ecology itself rather than learned parameters or increasing model capability. The experiment is intended as a controlled realization of the evidence-feedback mechanism analyzed theoretically, rather than as a test of the changing-scale theorem or an estimate of the empirical location of the model's long-run phase boundaries.