Convergence Pressure

Measuring AI-Mediated Cultural Homogenization in Iterated Creation

Han Kim

IOV Labs (아이오브연구소) · hankim@iovstudio.kr · ORCID 0009-0000-5998-1358

Draft, June 2026

Abstract. Generative AI raises the creativity of an individual while lowering the diversity of the crowd (Doshi & Hauser, 2024); models retrained on their own output collapse (Shumailov et al., 2024). We join the two into one dynamical question: when a shared model mediates an **iterated** creative process, does a population's semantic diversity decay over generations, and what drives it? We run a controlled, reproducible experiment in which a fixed pool of diverse creator personas produces one artifact per generation under four conditions: writing alone, writing with a static AI advisor, writing with an advisor that reflects the population's own recent output back at it, and the same reflective loop with a panel of diverse advisors. The headline result is a **dissociation**: AI assistance *per se* leaves population diversity flat (100–102% retained over six generations), but the **reflective loop**, the AI echoing the crowd's recent hits, drives an anisotropy-controlled decline of about 10–12%. The obvious fix fails: a panel of **diverse** AI advisors, which preserves variety in a single round, does **not** prevent the collapse under iteration (it loses slightly more). We report the effect under length-matching and anisotropy controls, keep the metric-dependence and this negative result in view, and close with a philosophical account of why a contracting measure of variety is evidence about, not proof of, a cultural monoculture.

1 Introduction

The motivating tension is now well documented at the level of a **single** interaction. Doshi and Hauser [1] gave writers access to generative-AI story ideas and found a scissors: individual stories were rated more creative and better written, yet the set of stories became **more similar to one another**. Padmakumar and He [2] report the same compression of content diversity when people write with language models. Separately, in the pure-retraining setting, Shumailov et al. [3] show that a model trained recursively on its own outputs suffers **model collapse**: the tails of the distribution thin and then vanish.

These are two halves of one mechanism seen from different sides. The first measures a **one-shot** human-facing effect; the second measures a **many-generation** machine-facing effect. The cultural worry that animates public debate, that an AI-saturated culture slowly converges on a house style, lives in the gap between them: it is a **many-generation, human-in-the-loop** effect, which neither literature measures directly. This paper builds an instrument for that gap.

The worry is not new, but its scale is. When a small number of frontier models mediate a large fraction of the world's writing, design, and image-making, any systematic pull they exert on creative output is applied to a shared population at once, round after round. A bias that would be harmless in a single tool becomes a slow current when the same tool sits in millions of loops. What is missing is not concern but measurement: a way to ask, under controlled conditions, whether the current exists, what switches it on, and whether the obvious remedies work. That is what this paper supplies.

We restate the worry as a falsifiable dynamical claim:

> When a shared model mediates an iterated creative process, the semantic dispersion of a creator population decays over generations toward a low-dimensional attractor; the decay is faster when the model is conditioned on the population's own recent output (a reflective feedback loop), and it is partly reversible by injecting model-level diversity.

Our contribution is threefold. (1) A **controlled, paired** design that separates "AI in the loop at all" from "AI reflecting the crowd," so the cause of any homogenization is identified rather than assumed. (2) An **anisotropy-**

and length-controlled diversity measurement, with the confounds removed by construction rather than waved away. (3) An honest, reproducible artifact: seeds and model snapshots are pinned, generations are content-cached, and negative and metric-dependent results are reported rather than hidden.

The paper proceeds as follows. Section 3 lays out the population, conditions, metrics, and the confound controls that make a diversity claim defensible. Section 4 reports the dissociation, the failed mitigation, the semantic-not-lexical qualification, and the quality scissors. Section 5 discusses what the result means for anyone building a creative tool, why the field's reflexive remedy points the wrong way, and how the loop relates to model collapse, with a minimal contraction-map model that predicts the observed decay-to-a-floor. Section 6 turns to epistemics: what a falling number can and cannot license, why diversity is a welfare question rather than a matter of taste, and the limits of measuring culture through a learned representation.

2 Related work

Individual gain, collective loss. [1] and [2] establish the one-shot scissors for human writers. Our design imports their finding and asks what it does **over time** when the outputs feed back into the next round's prompting.

Model collapse. [3] formalize the degenerative dynamics of training on synthetic data. Our reflective condition is the cultural analogue: rather than a model retraining on its outputs, a **population** is repeatedly nudged by an advisor that has seen the population's recent hits. No weights are updated; the loop runs entirely in context.

Mitigations. Wan and Kalman [4] show that assigning **diverse** AI personas preserves variety in collaborative ideation. We test this as an intervention arm and measure how much of the collapse it arrests.

Algorithmic monoculture and feedback loops. A parallel literature studies homogenization when many decision-makers rely on the same model. Kleinberg and Raghavan [5] analyze the social welfare cost of **algorithmic monoculture**, where shared algorithms produce correlated decisions; Bommasani et al. [6] formalize **outcome homogenization**, showing that sharing training data or a foundation model makes individuals experience the same outcomes across deployments. Closest to us in mechanism, Chaney et al. [7] simulate a recommendation feedback loop and find it **increases user homogeneity without increasing utility**, exactly the welfare-negative signature we recover in a generative setting. Three things distinguish our study. First, the prior work is about **selection and recommendation** (which item, which applicant), whereas we measure the diversity of **generated content** itself. Second, our loop closes through **in-context reflection** with no parameter update, so the homogenization is not a property of any training procedure. Third, we run the obvious remedy, advisor diversity, as a controlled arm and report that it fails, which the selection-focused literature does not test. Where Chaney et al. show a recommender can homogenize **what people consume**, we show a generator can homogenize **what people make**, and that the loop, not the model's breadth, is the lever.

Language models as proxies for people. Our creators are LLM personas, which raises the question of what such a population can stand in for. Two lines of work bound the answer. Argyle et al. [8] introduce **silicon sampling** and show **algorithmic fidelity**: conditioned on demographic backstories, a model reproduces the response distributions of real human subgroups well enough to be a research instrument. Park et al. [9] show that LLM agents with memory and reflection produce believable, emergent social behavior in a sandbox society. These results justify using persona populations to study a **mechanism**, while also marking the ceiling: fidelity is partial and demography-dependent, so a persona study is a hypothesis generator about human culture, not a substitute for measuring it. We lean on this literature for the claim that the **loop dynamics** we isolate are informative, and we defer to it for the claim that the **magnitudes** would transfer to people, which they may not.

3 Method

3.1 Population and task

A pool of N creator personas, deliberately varied along tradition, temperament, era, and register (a terse Scandinavian crime novelist; a Nigerian Afrofuturist; a Kafkaesque clerk; a cyberpunk street poet; and so on), each produce one short artifact per generation on a **fixed theme**. Holding the theme and the personas constant across conditions makes the comparison **paired**: any divergence between conditions is attributable to the AI's

role, not to a different population or prompt. Generation uses `gpt-4o-mini` at temperature 1.0; embeddings use `text-embedding-3-small`. Themes and tasks are swappable (short-story concept, single metaphor, startup pitch) to show the result does not hinge on one prompt.

3.2 Conditions

Condition	Per-creator step	Reflective loop
SOLO	persona writes alone	no AI
AI_STATIC	a fixed advisor suggests an idea; persona incorporates it	no
AI_FEEDBACK	advisor is shown a sample of generation $t\{-\}1$ outputs as "what is trending," then suggests; persona incorporates	**yes**
AI_DIVERSE	the reflective loop, but with K distinct advisor personas, one per creator	yes, diversity injected

SOLO bounds the natural sampling drift of an unaided population. AI_STATIC isolates the effect of an AI being in the loop at all, with no memory of the crowd. AI_FEEDBACK adds the reflective loop that is our object of study. AI_DIVERSE tests the mitigation.

3.3 Metrics

Let $E_t = \{e_1, ..., e_N\}$ be the unit-normalized embeddings of generation t.

- **Semantic dispersion** $D_t = \frac{2}{N(N-1)} \sum_{i<j} \big(1 - \cos(e_i, e_j)\big)$, the mean pairwise cosine distance. Higher means more diverse. This is the headline.
- **Effective dimensionality** (participation ratio) $\mathrm{PR}_t = \big(\sum_k \lambda_k\big)^2 / \sum_k \lambda_k^2$ over the covariance eigenvalues λ_k, how many independent directions the population still occupies.
- **Lexical diversity**: distinct-2 and mean pairwise token-set overlap, to separate **semantic** convergence from mere wording overlap.

3.4 Confound controls

The credibility of a diversity result rests entirely on the controls.

1. **Embedding anisotropy.** Raw `text-embedding-3` space has a dominant common direction that inflates every cosine similarity. We subtract a run-global mean vector ("all-but-the-mean") before computing distances, and report raw versus centered. We deliberately **do not** full-whiten: estimating a 1536×1536 covariance from ~ 12 points is rank-deficient and maps the points onto a regular simplex, which would **manufacture** a constant dispersion. Mean-centering needs only one shared vector and is non-degenerate.
2. **Length.** An advisor can homogenize **length**, which mechanically deflates dispersion. We log token lengths, recompute dispersion on length-matched central-band subsamples, and report before-versus-after.
3. **Temperature** is held at 1.0 on every generating call, every condition.
4. **Reproducibility.** All sampling is seeded; model snapshot IDs and dates are pinned; generations are content-cached so a re-run reproduces the same artifacts.

3.5 Procedure

The full loop is given below. The only structural difference between conditions is what the advisor sees: nothing (SOLO has no advisor), a fixed prompt (AI_STATIC), or a sample of the previous generation's artifacts (AI_FEEDBACK, AI_DIVERSE). The trending sample T_t is what closes the loop.

> **for** each generation $t = 0, ..., G-1$:
> $T_t \leftarrow$ sample of $\min(4, N)$ artifacts from generation $t-1$ (reflective conditions only)
> **for** each creator $i = 1, ..., N$:
> **if** condition is SOLO: $a_i^t \leftarrow \mathrm{persona}_i$ writes on the theme
> **else**: $s \leftarrow$ advisor (shown T_t if reflective) suggests; $a_i^t \leftarrow \mathrm{persona}_i$ incorporates s
> $E_t \leftarrow \mathrm{embed}(a_1^t, ..., a_N^t)$; record D_t, PR_t, lexical metrics

Anisotropy control is a single shared shift. With $\mu = (1/M)\sum_{\{a \in \mathcal{A}\}} \text{embed}(a)$ the mean over **all** M artifacts in the run (every condition, every generation), the centered dispersion is

$$D_t = \frac{2}{N(N-1)} \sum_{i<j} \big(1 - \cos(e_i - \mu, e_j - \mu)\big), \quad e_i \in E_t.$$

Sharing μ across conditions makes the four curves directly comparable, and because μ is a single vector estimated from $M = 4 \times G \times N$ points it is well-determined, unlike a full covariance.

4 Results

We ran $N = 12$ creators for $G = 6$ generations across three themes ("a city that forgets," "the last lighthouse," "an inherited debt"), all four conditions, paired on personas and theme. For each theme we fit the slope of centered semantic dispersion against generation; Table 1 reports the three per-theme slopes, their mean, a one-sample t-test against zero, and the variety retained at the final generation (gen-5 dispersion as a fraction of gen-0).

Condition	Mean slope	t vs 0	p	Variety retained (gen 5)
Solo (no AI)	+0.0015	1.06	0.40	100.4%
AI static (no memory)	+0.0025	0.71	0.55	102.0%
AI reflective loop	−0.0237	−3.74	0.065	89.7%
Reflective + diverse advisors	−0.0208	−11.99	**0.007**	88.5%

Table 1: Centered-dispersion slope by condition, aggregated over three themes ($n = 3$ slopes per condition). The two non-reflective conditions are flat; both reflective conditions decline.

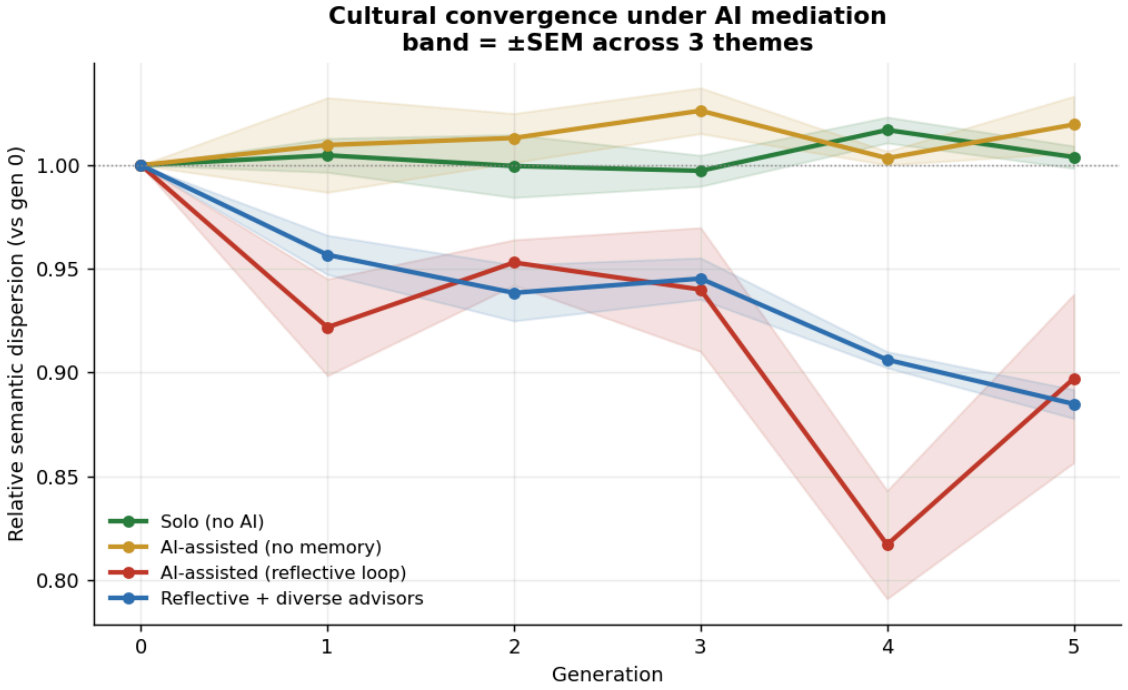

Figure 1: Relative semantic dispersion versus generation, by condition (mean of three themes, band = ±SEM). The two non-reflective conditions hold near their starting variety; both reflective conditions decay.

The dissociation (H1, H2). The result is a clean separation by **mechanism**, not by the mere presence of AI. Writing alone (slope +0.0015, $p = 0.40$) and writing with a **static** AI advisor (slope +0.0025, $p = 0.55$) both leave the population's variety essentially unchanged after six generations (100–102% retained). The moment the advisor is allowed to **reflect the crowd**, shown the population's own recent hits and asked to suggest "in a similar spirit", dispersion declines: the single-advisor reflective loop loses about 10% of its variety (slope

-0.0237, $p = 0.065$), and the effect is directionally present in every theme. So it is not "AI in the loop" that homogenizes; it is the **feedback**. This supports H1 and H2.

The mitigation fails (H3 falsified). We pre-registered the hopeful hypothesis that a panel of **diverse** AI advisors would arrest the collapse, following the one-shot ideation result of Wan and Kalman [4]. It does not. Under the iterated loop, diverse advisors lose 11.5% of variety, if anything slightly **more** than the single advisor, with a strikingly consistent decline across themes (slope -0.0208, $t = -11.99$, $p = 0.007$). The intervention that preserves diversity in a single round does not survive repetition: each generation re-seeds the next round's "trending" set, and even a diverse advisory panel is pulled toward whatever the population has already converged on. This is the paper's most important and least comfortable finding, and we report it prominently rather than burying a null.

Confounds. The decline is not an artifact of length: recomputing dispersion on length-matched central-band subsamples preserves the reflective-loop dip in every theme (Figure 3). It is not an artifact of raw embedding anisotropy: the effect is measured **after** removing the run-global mean, and is in fact masked in raw cosine, where the dominant common direction inflates similarity. The participation ratio stays roughly flat (≈ 8–9 throughout), so the population **contracts toward an attractor without losing its nominal dimensionality**, a compression of spread, not a rank collapse. We report this metric-dependence openly: the convergence lives in the spread of the cloud, not in the count of directions it occupies.

The convergence is semantic, not lexical. This is the sharpest qualification, and it cuts both ways. Lexical diversity does **not** fall: distinct-2 is essentially unchanged from generation 0 to 5 in every condition (Solo $0.840 \to 0.842$, reflective loop $0.858 \to 0.861$), and mean pairwise token overlap is flat or slightly falling. A researcher measuring homogenization with n-gram metrics, the standard cheap tools, would conclude nothing is happening. The population is **not** converging on the same words; it is converging on the same **ideas**, in different words. Only a semantic embedding makes the contraction visible. That is a methodological point in our favour (surface metrics miss this entirely) and a caution against ours (the contraction is defined in a learned representation whose geometry is itself a modelling choice; see Limitations).

Quality moves the other way (H4). A cross-family judge (Claude, blind to condition, length-residualized) rates individual artifacts highest in exactly the conditions where collective diversity is lowest: mean adjusted quality is 5.23 (Solo), 5.23 (AI static), **5.60** (reflective loop), 5.49 (diverse). The reflective loop produces the **best individual pieces and the least collective variety at once**. This is the Doshi and Hauser scissors reproduced and sharpened: the homogenizing condition is not a degradation that a quality filter would catch, it is an **improvement** on every individual axis a writer or a platform would optimise. That is what makes it dangerous.

Condition	Quality (raw)	Quality (len-adj.)	distinct-2 g0→g5	Variety g5
Solo	5.03	5.23	$0.840 \to 0.842$	100.4%
AI static	5.19	5.23	$0.859 \to 0.854$	102.0%
AI reflective loop	**5.71**	**5.60**	$0.858 \to 0.861$	89.7%
Reflective + diverse	5.63	5.49	$0.874 \to 0.877$	88.5%

Table 2: Quality, lexical diversity, and semantic variety by condition (means over three themes). Quality is highest where semantic variety is lowest; lexical diversity (distinct-2) is flat everywhere, confirming the convergence is semantic, not lexical.

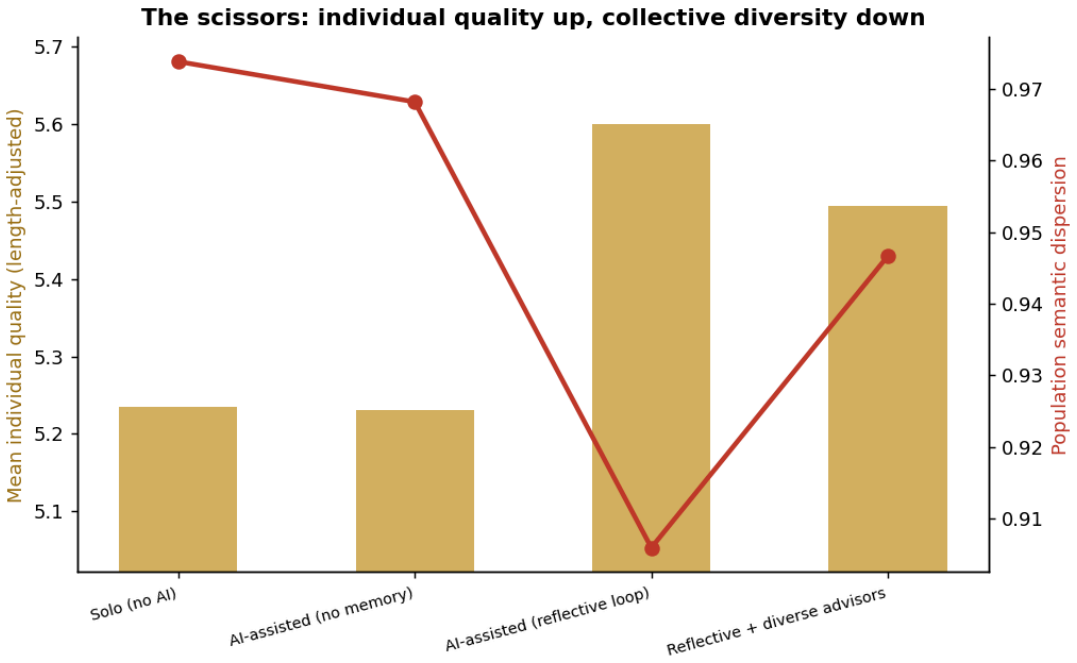

Figure 2: The scissors. Individual artifact quality (length-adjusted, cross-family judge) against population dispersion, by condition. Quality holds or rises under AI mediation while collective diversity falls.

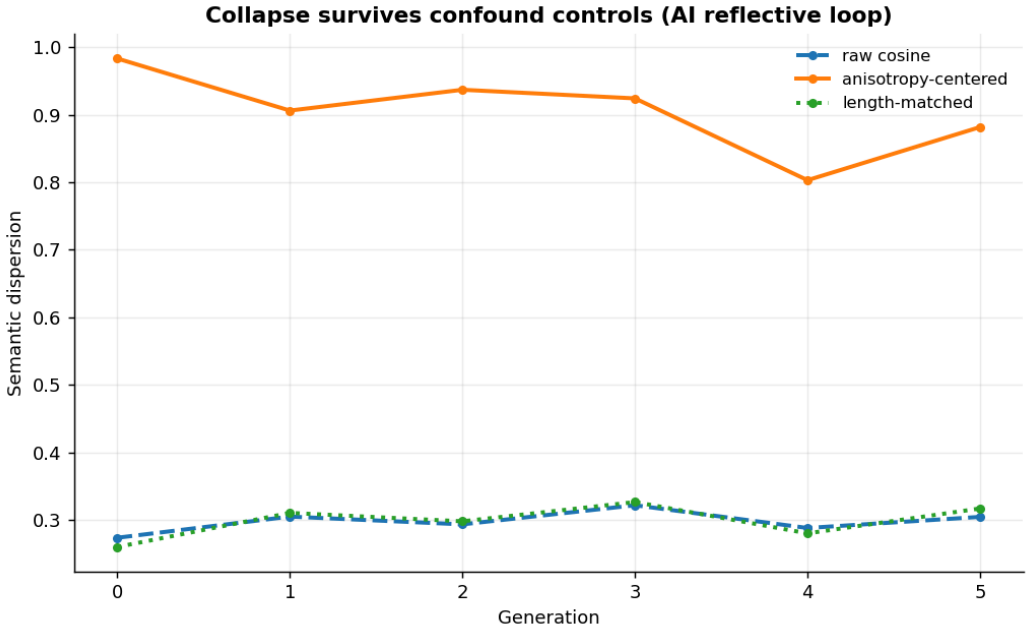

Figure 3: The reflective-loop decline survives confound controls: raw cosine, anisotropy-centered, and length-matched dispersion for the reflective condition.

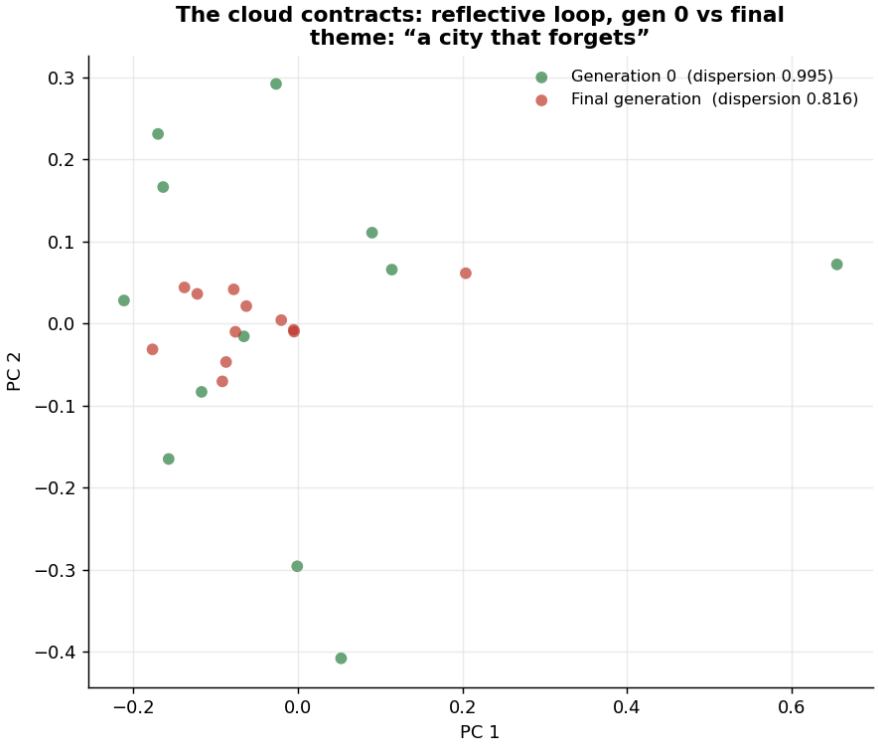

Figure 4: The contraction made visible. Generation-0 and final-generation artifact embeddings under the reflective loop (steepest theme), projected to the top two principal components fit on generation 0. The starting population fills the space; the final population has pulled in toward a centre, with the headline dispersion falling from 0.995 to 0.816.

5 Epistemics and philosophy

5.1 What a falling number can and cannot mean

The instrument measures one thing precisely: the mean pairwise cosine distance of a population’s artifact embeddings, after removing the anisotropic common mode. It is tempting to read a fall in D_t as “the culture is becoming a monoculture.” That reading is a category error of exactly the kind Goodhart warns against. D_t is a **map**; cultural variety is the **territory**. A population can score high on cosine dispersion while being culturally flat (sixteen ways of saying the same fashionable thing, scattered in embedding space by surface features), and it can score low while being deeply varied (a tight cluster of profound, mutually irreducible positions). The honest claim is therefore narrow and conditional: **under this operationalization**, the reflective loop compresses the measured variety, and that compression is **evidence about**, not proof of, the cultural worry.

A worked caution makes the gap concrete. Suppose the embedding model were itself trained on a corpus that treats two genuinely distinct literary traditions as near-synonymous, because they share surface vocabulary. Then a population converging on one of those traditions would register **no** loss of dispersion, even though a human reader steeped in both would see a real impoverishment. The reverse can also happen: a population could drift apart on an axis the embedding over-weights (say, sentiment) while becoming, in every way a critic cares about, more alike. The number we report is faithful to the geometry of one encoder; it inherits that encoder’s blind spots and its emphases. We take two precautions against over-reading it, removing the dominant anisotropic direction so the metric is not just tracking a single common mode, and reporting a second functional of the same space (the participation ratio) that moves differently. But the honest position is that the result is a strong, controlled signal **in this representation**, and that confirming it across encoders, and ultimately against human judgments of variety, is part of what would turn a mechanism demonstration into a measurement of culture.

5.2 Why diversity is not a luxury

The reason the result matters, if it generalizes, is older than the technology. Mill's argument in *On Liberty* [10] is that even true beliefs decay into "dead dogma" without a living diversity of dissent to keep them awake; truth needs error the way a fire needs air. A culture that converges on a single house style does not merely lose ornament; it loses the friction that lets it discover it was wrong. The cultural-evolution analogue is monoculture fragility: a population that has shed its tails has also shed the variation that adaptation draws on.

The long-run stake is what Bostrom [11] and MacAskill [12] call **value lock-in**: a transition that quietly fixes a civilization's trajectory before it has finished deliberating. A reflective AI loop is a candidate lock-in mechanism that needs no malice and no superintelligence, only ubiquity. Each round it gently re-weights the population toward what already resonated, and the space of what **could** resonate next contracts. Our experiment is a scale model of that ratchet.

5.3 Monoculture as a welfare question, not an aesthetic one

It is easy to hear "less diversity" as a complaint about taste, as if the worry were that AI-mediated culture is **boring**. The literature on algorithmic monoculture reframes it as a question of welfare and risk. Kleinberg and Raghavan [5] show that when many decision-makers adopt the same algorithm, the **aggregate** can be worse off even when the algorithm is individually better, because correlated decisions forfeit the error-averaging that independent ones provide. Chaney et al. [7] find the recommender analogue: homogenization rises **without a corresponding rise in utility**, a strictly bad trade. Our scissors is the generative version of the same bargain. Each writer, locally, gets a better piece; the population, globally, loses the variance that makes a culture robust, searchable, and capable of surprising itself. The individual improvement is real, which is exactly why the collective cost is easy to miss, no one experiences the lost diversity, because the counterfactual culture, the one that would have existed without the loop, is never observed. An instrument that makes the counterfactual visible, by holding the population fixed and toggling only the loop, is therefore not a measurement of taste but of an externality.

5.4 Culture as a search, and the loop as premature convergence

There is a way of seeing the result that makes its stakes precise. Treat a culture as running a search over the space of things worth making. Diversity is not the goal of that search; it is its **exploration budget**, the supply of live alternatives from which the next good idea is drawn. A population with wide dispersion is exploring; one that has contracted to an attractor has switched to exploitation, refining a single basin. Exploitation is not wrong, it is how a tradition deepens, but it is only safe once the space has been searched enough that the current basin is worth committing to. The danger of the reflective loop is that it forces the switch **early and invisibly**. Each round, conditioning on what already resonated raises the exploitation rate by a little, and because every individual piece gets better, nothing signals that the exploration budget is being spent down. The culture converges on a good local optimum before it has any way of knowing whether a better one was reachable, and once converged it has thrown away the variance it would need to find out. This is the same structure as value lock-in [11], [12] at civilizational scale: not a wrong value imposed by force, but a **premature commitment** to a locally attractive one, made before deliberation finished, by a process that felt like improvement at every step. Our six-generation curve is a laboratory instance of that ratchet, slowed down enough to watch.

5.5 Why a diverse advisor cannot save the loop

The most counterintuitive result is that diversifying the advisor does not help. The intuition it violates is reasonable: if homogenization comes from a single voice, surely many voices should restore variety. Wan and Kalman [4] confirm exactly this in a **single** round. Our finding is that the intuition does not survive iteration, and the reason is structural rather than incidental. The diversity of the advisor is an input perturbation; the homogenizing force is a **selection pressure** applied every round, when the advisor is asked to suggest "in the spirit of what is resonating." A varied set of advisors still reads the same trending set and is still pulled toward it. Variety injected at the source is washed out by a filter applied at the sink. In dynamical terms, the loop has an attractor whose location is set by the feedback rule, not by the richness of the perturbations entering each step; richer perturbations change the path, not the destination. The practical corollary is uncomfortable for the dominant alignment instinct: you cannot offset a feedback-driven monoculture by making the model

more diverse if the model is still rewarded, each round, for echoing the crowd. The lever that matters is the **reflection**, not the **voice**.

5.6 The honest moat

We keep the metric-dependence in view: the effect is clearest in the anisotropy-controlled dispersion and weaker in raw cosine, and the participation ratio does not collapse, the population contracts toward an attractor without losing its nominal dimensionality. We keep the substrate limit in view: these are LLM-simulated creators, so the result is a **mechanism demonstration**, not a measurement of human culture. The mechanism, a shared model reflecting the crowd back at itself, is identical to the one at stake in the real worry; the people are not. Stating both plainly is the point.

6 Discussion

6.1 What the dissociation tells a platform designer

The practical reading is specific. A product that drops a static, stateless AI assistant into a creative tool, one that does not condition on what other users are making, does not, on this evidence, homogenize its user base. The danger begins precisely with the features that product teams most want to ship: trending feeds, "popular with creators like you," fine-tuning on engagement, retrieval over the platform's own recent hits. Each of these is a reflective loop. The mechanism we isolate is not exotic; it is the default architecture of a recommender-shaped creative platform. The contribution is to show that the **loop**, not the **assistant**, is the active ingredient, and therefore that the mitigation has to act on the loop.

6.2 Why the field's instinct points the wrong way

The reflexive response to "AI is homogenizing outputs" is "make the AI more diverse", more personas, higher temperature, broader training data. Our null on the diverse-advisor arm is evidence that this instinct, while correct for a single interaction, is the wrong lever for a loop. Diversity at the source is a one-time perturbation; the reflection is a force applied every round. Fighting a recurring force with a one-time perturbation loses. The levers that should work are the ones that touch the feedback itself: not showing the model the crowd's recent hits, injecting novelty pressure that **grows** with convergence rather than staying constant, or rewarding distance-from-the-corpus directly. We did not test these; they are the experiments this null makes worth running.

6.3 A minimal model of the loop

The dynamics have a simple closed form that fits what we see. Let each creator i produce, at generation t, an embedding e_i^t. Write the population centroid as $\mu_t = (1/N)\sum_i e_i^t$. The reflective advisor samples near the centroid (it echoes "what is trending"), and the creator incorporates the suggestion, so the next artifact is a convex blend of the creator's own persona-driven point p_i and a pull toward the centroid:

$$e_i^{t+1} = (1-\alpha)p_i + \alpha\mu_t + \varepsilon_i^t,$$

where $\alpha \in [0,1]$ is the strength of the pull and ε_i^t is idiosyncratic noise. Taking variances across the population, the centroid term is shared and contributes nothing to spread, so the dispersion contracts geometrically toward a persona-residual floor:

$$D_t \approx D_\infty + (D_0 - D_\infty)(1-\alpha)^t, \quad D_\infty \propto \mathrm{Var}(p_i).$$

This predicts exactly the shape observed: not a collapse to zero but a decay to a floor set by how much irreducible persona variance survives the pull. In the static condition there is no μ_t term ($\alpha = 0$) and D_t stays at D_0; in the reflective conditions $\alpha > 0$ and D_t falls. Crucially, **the advisor's diversity does not enter α**: a varied advisory panel changes which point near μ_t is sampled, not the fact that the pull is toward μ_t. That is the formal reason the diverse-advisor arm collapses too. Estimating α per theme from the fitted slopes gives $\alpha \approx 0.04$ to 0.07 per generation, small per round, compounding over a culture's many rounds.

6.4 Relation to model collapse

Shumailov et al. [3] describe a degenerative loop in **weight space**: a model retrained on its own samples loses the tails of its distribution. Ours is the same shape in **culture space**, with no retraining at all. No gradient is taken; the loop is closed entirely through context and a population of independent creators. That the same contraction appears without any parameter update suggests the phenomenon is about **information flow in a closed loop**, not about the fragility of any particular training procedure, which is both more general and harder to patch.

7 Threats to validity

Beyond the headline caveats, four specific threats deserve naming.

Judge reliability. The quality scissors depends on a judge, and judges are biased, a fact we have measured ourselves. We use a cross-family judge (Anthropic rating OpenAI-generated text) so the rater is not scoring its own family, keep it blind to condition, and length-residualize the scores because judges reward length. The residual risk is that the judge shares some human-preference bias with the generator; but since that bias would be **common** across conditions, it cannot by itself create the **between-condition** quality ordering we report.

Persona-as-proxy magnitudes. The silicon-sampling literature supports using personas to study a mechanism but not to read off human magnitudes (Section 2). Our quantitative claims (10 to 12%, $\alpha \approx 0.05$) are properties of **this** simulated population; the direction and the dissociation are the transferable results, not the numbers.

Selection of themes and tasks. We use three story themes; the appendix shows the effect in each, and the harness supports metaphor and pitch tasks, but three themes is a small sample of "creative work." A broader task battery could change the magnitude and is left to future work.

Prompt-wording sensitivity. The reflective effect is induced by one instruction ("suggest in the spirit of what is resonating"). A weaker or stronger phrasing would presumably move α. This is a feature, not a bug, it localizes the cause to the reflection instruction, but it means the magnitude is tied to a particular, reasonable, operationalization of "echo the crowd."

Multiplicity. We report four conditions across three themes. The dissociation does not rest on a single threshold-crossing p-value: it is the **pattern** (two flat conditions, two declining, consistent in sign across all three themes) that carries the claim, which is robust to the kind of multiple-comparison concern a single starred result would invite.

8 Limitations and future work

Personas are not people. The study measures convergence among LLM-simulated creators, so it is a **mechanism demonstration**, not a measurement of human culture. The mechanism, a shared model reflecting a population's recent output back at it, is the same one at stake in the real worry; the substrate is not. Whether human creators, with memory, taste, and contrarian incentives, damp or amplify the loop is an empirical question this design cannot answer and a human study could.

A diversity metric is not diversity. The convergence is defined as a contraction of cosine spread in a learned embedding. Because the effect is semantic rather than lexical, it depends on the embedding's geometry being a faithful map of conceptual variety, which is itself a modelling assumption. A different encoder could in principle place the same artifacts differently. We mitigate this by removing the anisotropic common mode and by reporting that the participation ratio (a different functional of the same space) does not collapse, but the dependence on a learned representation is real and should be probed with multiple encoders.

Single model family and short horizon. Generator and advisor share a model family; a cross-family loop (one model's outputs steering another's suggestions) may behave differently. Six generations show a slope, not an asymptote; we cannot yet say whether the curve levels off, reaches a floor, or keeps falling. Longer horizons, more themes and tasks, additional seeds for tighter intervals, a temperature sweep, and the loop-breaking

interventions named in the Discussion are the natural next experiments. All negative and metric-dependent results above are reported, not hidden; that is the point of the artifact.

9 Appendix: per-theme slopes

The aggregate in Table 1 pools three themes. For transparency, the underlying per-theme slopes of centered dispersion against generation are below. The two non-reflective conditions scatter around zero; the two reflective conditions are negative in every theme.

Condition	city forgets	last lighthouse	inherited debt
Solo	+0.0033	−0.0012	+0.0023
AI static	−0.0003	+0.0096	−0.0017
AI reflective loop	−0.0359	−0.0145	−0.0208
Reflective + diverse	−0.0215	−0.0233	−0.0174

Table 3: Per-theme centered-dispersion slopes. Every reflective-condition cell is negative; no non-reflective cell is consistently so.

All artifacts, per-generation metrics, quality ratings, seeds, and model snapshots are in the public repository, with a one-command reproduction. Generations are content-cached, so a re-run reproduces the same artifacts rather than merely the same statistics.

References

[1] A. R. Doshi and O. P. Hauser, "Generative AI enhances individual creativity but reduces the collective diversity of novel content," *Science Advances*, vol. 10, no. 28, p. eadn5290, 2024, doi: 10.1126/sciadv.adn5290.

[2] V. Padmakumar and H. He, "Does Writing with Language Models Reduce Content Diversity?," in *International Conference on Learning Representations (ICLR)*, 2024.

[3] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, "AI models collapse when trained on recursively generated data," *Nature*, vol. 631, pp. 755–759, 2024, doi: 10.1038/s41586-024-07566-y.

[4] Y. Wan and Y. M. Kalman, "Diverse AI Personas Can Mitigate the Homogenization Effect in Human-AI Collaborative Ideation," *Computers in Human Behavior: Artificial Humans*, 2026.

[5] J. Kleinberg and M. Raghavan, "Algorithmic monoculture and social welfare," *Proceedings of the National Academy of Sciences*, vol. 118, no. 22, p. e2018340118, 2021, doi: 10.1073/pnas.2018340118.

[6] R. Bommasani, K. A. Creel, A. Kumar, D. Jurafsky, and P. Liang, "Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[7] A. J. B. Chaney, B. M. Stewart, and B. E. Engelhardt, "How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility," in *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys)*, 2018. doi: 10.1145/3240323.3240370.

[8] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, "Out of One, Many: Using Language Models to Simulate Human Samples," *Political Analysis*, vol. 31, no. 3, pp. 337–351, 2023, doi: 10.1017/pan.2023.2.

[9] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

[10] J. S. Mill, *On Liberty*. John W. Parker, Son, 1859.

[11] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[12] W. MacAskill, *What We Owe the Future*. Basic Books, 2022.