

V1.0.0

Compression Versus Consilience in Recursive Model-Generated Corpora: A Break-Even Analysis

Rowan Brad Quni-Gudzinis

QNFO-ULA-2.1

ABSTRACT

Training a machine-learning model on data generated by another model is naturally viewed as lossy compression of high-dimensional real-world structure: each recursive generation should discard tail events and shrink effective dimensionality. Against this compression conjecture stands the observation that very large synthetic corpora can exhibit novel cross-disciplinary associations (consilience) that human-scale efforts would not assemble. We formalize the tension with a recursive compression-channel model: tail entropy and effective dimensionality decay geometrically with per-generation retention factor $\rho = 1 - \varepsilon$, while consilience is modeled as the creation of novel cross-domain links from surviving bridge vocabulary, empirically anchored by an audit finding that 97.2% of external technical vocabulary occurs in exactly one discipline. [...]

CITE AS Quni-Gudzinis, R. B. (n.d.). Compression Versus Consilience in Recursive Model-Generated Corpora: A Break-Even Analysis. QNFO Preprint v1.0.0.

Abstract

Training a machine-learning model on data generated by another model is naturally viewed as lossy compression of high-dimensional real-world structure: each recursive generation should discard tail events and shrink effective dimensionality. Against this compression conjecture stands the observation that very large synthetic corpora can exhibit novel cross-disciplinary associations (consilience) that human-scale efforts would not assemble. We formalize the tension with a recursive compression-channel model: tail entropy and effective dimensionality decay geometrically with per-generation retention factor $\rho = 1 - \varepsilon$, while consilience is modeled as the creation of novel cross-domain links from surviving bridge vocabulary, empirically anchored by an audit finding that 97.2% of external technical vocabulary occurs in exactly one discipline. Our central result is a generation-independent break-even condition: because compression loss and consilience gain both scale as $(1 - \varepsilon)^g$, compensation holds at every generation if and only if a dimensionless coupling constant κ exceeds a fixed threshold $\kappa^* = \varepsilon H_{\text{tail}}(0) / (\beta V_0 p_{\text{bridge}}) \approx 64.29$ under our stated normalization ($\varepsilon = 0.15$, $H_{\text{tail}}(0) = 12$ bits). We derive the one-bit usefulness threshold at generation $g = 16$, show that re-anchoring every 5 generations sustains 8.90 bits average tail entropy (a 67.1% uplift), and state falsifiable predictions. No simulations or empirical measurements were performed; all numbers are derived from stated assumptions.

1. Introduction

A growing fraction of the text, code, and structured data used to train machine-learning systems is itself produced by machine-learning systems. When a model is fine-tuned on the output of another model, and that output-model was in turn trained on model output, the resulting corpus is the product of a recursive, model-on-model generative chain. A natural worry is that such a chain is a lossy compression process: each generation discards some of the high-dimensional structure of the real-world data that seeded the chain, so that tail events — rare but informative configurations far from the corpus mode — are progressively flattened, and inference becomes "shallower." Call this the **compression conjecture**. It predicts monotone geometric decay of tail entropy H_{tail} and effective dimensionality d_{eff} across generations.

Against this stands an **emergent-consilience conjecture**, hypothesizing that very large synthetic corpora can exhibit novel cross-disciplinary consilience: unexpected, productive associations between domains that no human-scale research effort would assemble, because no human reads across disciplinary silos at machine scale and speed. The precondition for such associations is shared vocabulary across disciplines, and this precondition is empirically scarce: a quantitative audit of cross-domain vocabulary found that 97.2% of external technical vocabulary occurs in exactly one discipline [14]. If bridge terms are so rare, synthetic corpora that recombine content at machine scale might surface bridges that human literature never formed — precisely because machines can enumerate combinations humans do not attempt.

The research question is therefore precise: **can emergent cross-domain link structure compensate for per-sample compression, or does compression inevitably dominate?** We answer with a framework rather than an experiment. Our contributions are:

1. **A channel model of recursive training**, in which one generation of model-on-model training is a compression operator retaining a fraction $1 - \varepsilon$ of tail-entropy mass, in the spirit of Bayesian data compression [1].
2. **Closed-form decay laws** for tail entropy and effective dimensionality under constant ε .
3. **A consilience metric and a break-even theorem**: the compensation condition is generation-independent and reduces to a single inequality on a dimensionless coupling constant κ , calibrated by the 97.2% single-discipline finding [14]. As a complementary labeled model, we also state a pair-counting crossover condition under which corpus growth can outrun per-sample tail loss.
4. **A taxonomy of interventions**: diversity enforcement, formal verification, and periodic re-anchoring each map to a specific term of the break-even inequality.

The paper is deliberately explicit about what is derived and what is assumed. Every number in Section 4 is computed from stated inputs with shown arithmetic; no simulations or empirical measurements were performed, and Section 5 reports only derived quantities plus clearly labeled projections.

2. Background and Related Work

Bayesian data compression. The closest methodological antecedent to our compression channel is Bayesian data compression (BDC), which derives an algorithm that adapts to the specific measurement situation in the context of signal reconstruction, compressing a dataset under conservation of its posterior structure with minimal information loss given prior knowledge [1]. Our per-generation compression fraction ε is the training-loop analogue of BDC's information loss: BDC controls loss once, against a prior; the recursive setting compounds that loss across generations, which is precisely the effect our geometric decay law captures.

Compression for robust control. In data-driven robust control, an optimal-transport-based method compresses a large dataset of input/output pairs into a smaller synthetic dataset of representative behaviours, in order to alleviate computational burden while retaining controller utility [2]. This is an instructive precedent for our hybrid-pipeline question: compression there is judged by whether the compressed synthetic dataset preserves the *control-relevant* behaviour, i.e., utility is task-conditioned, not distribution-wide. Our tail-entropy metric is the distribution-wide counterpart, and the gap between the two motivates our distinction between mode utility (which survives compression well) and tail generalization (which does not).

Adaptive compression in databases. An adaptive column compressor for self-driving in-memory databases is proposed that offers a new trade-off point between memory footprint and query speed, since compression can reduce memory requirements but often reduces query speed too [3]. The trade-off structure — compression buys storage

efficiency at a performance cost — is the same structure we formalize as ε versus consilience gain, though in a completely different substrate. The supplied summary of [3] gives no further quantitative detail, so we use it only as evidence that the compression/utility trade-off is a recognized engineering dimension.

Landmark encoding of time series. Peak-nadir encoding reduces dense continuous glucose monitoring time series to a compact set of landmark points while maintaining fidelity in reconstructed signals and derived glycemic metrics, evaluated on two complementary CGM datasets [4]. This is a concrete instance of tail-selective compression: landmarks are, in our vocabulary, retained tail events, and the reported concern for derived downstream metrics parallels our concern for downstream task generalization. The summary supplied for [4] is truncated and reports no numbers we can reuse, so we cite it only for the design pattern.

Dataset distillation as compression. A rate–utility perspective on dataset distillation argues that distillation compresses an original dataset into a small set of synthetic samples while preserving its full utility, and criticizes existing methods for either maximizing performance under fixed storage or otherwise treating the rate–utility trade-off incompletely [5]. This is the most direct precedent for our framing: a distilled dataset is a synthetic corpus, and the rate–utility axis is our (ε , consilience) axis. A companion study of the mechanisms by which gradient-based learning extracts task-relevant information and encodes it efficiently into synthetic data points for non-linear tasks notes that progress in distillation remains largely empirical [8]; this supports our decision to supply a theory-first framework rather than another empirical distillation method.

Compression as structure discovery. Spatial regionalization based on optimal information compression extracts natural regions without user-specified region counts or similarity measures, using compression itself as the clustering criterion [6]. This matters because our consilience metric requires a notion of "domain" in semantic space; [6] shows that information compression can define such structure intrinsically rather than by user input, suggesting an unsupervised route to the domain partition our metric needs. The supplied summary states no quantitative findings beyond this design claim.

Latent representations under hardware constraints. The ζ -QVAE addresses the scarcity of quantum hardware resources for large real-world datasets by finding low-dimensional representations that preserve essential information for downstream analysis, using regularized mixed-state latent representations in a quantum variational autoencoder [7]. The structural parallel is exact: compression is justified by downstream utility under a resource constraint, which is our setting with "quantum hardware" replaced by "the model's context and capacity." The supplied summary is truncated before any results, so we draw no quantitative claim from it.

Tensor compression. Tensor canonical polyadic (CP) decomposition formulated as a nonlinear least-squares problem, solved with a modified Levenberg–Marquardt algorithm with an emphasis on image compression and reconstruction, provides another instance of low-rank (effective-dimensionality-reducing) compression [11]; our d_{eff} decay is the training-loop analogue of rank reduction. The supplied summary states no quantitative results we could import.

Synthetic data and its risks. A survey of privacy measurement in tabular synthetic data reports that there is no standard for quantifying the degree of privacy protection that synthetic data afford, and discusses proposed quantification approaches to support modeling and evaluation decisions [10]. Although privacy is not our focus, the survey's central diagnosis — synthetic-data evaluation lacks agreed metrics — applies verbatim to consilience: our Section 3 metric is a proposal in a space without standards, and we flag that as a limitation. As a cautionary calibration point from empirical science, constraints on dark-energy parameters derived from H II starburst galaxy apparent magnitude versus redshift data were found to be generally consistent with those from other datasets but not as restrictive as the tightest available constraints [9]. We use [9] only as an example of the general pattern that a single, possibly biased data source yields consistent but weaker constraints — the exact failure mode we predict for deeply recursive synthetic corpora, whose constraints on reality weaken with each generation.

Consilience measurement. The QNFO audit of cross-domain vocabulary measures a basic precondition of interdisciplinary consilience — whether scientific domains share vocabulary — and finds that 97.2% of external technical vocabulary occurs in exactly one discipline, with bridge terms massively enriched in method-level vocabulary [14]. This supplies the one empirical anchor we use. Two further QNFO works, on projective geometric frameworks for semantic structures [12] and on syntactic generation [13], have no supplied summaries; we therefore cite them only as parts of the same corpus context and make no claims about their content.

3. Methods

3.1 The recursive compression channel

Let $\mathcal{C}_g$ denote the corpus at generation g , with $\mathcal{C}_0$ a real-world-seeded corpus. One generation of model-on-model training maps $\mathcal{C}_g \rightarrow \mathcal{C}_{g+1}$, where a model trained on $\mathcal{C}_g$ generates $\mathcal{C}_{g+1}$. We model this map as a compression operator T_ε with a single scalar parameter: the per-generation retained tail-entropy fraction $1 - \varepsilon$, with $0 < \varepsilon < 1$. The modeling assumption is that T_ε preserves task-relevant structure in the mode of the distribution (as BDC preserves posterior structure [1]) while losing tail mass at rate ε per generation.

3.2 Tail entropy and effective dimensionality

Define the tail entropy $H_{\text{tail}}(\mathcal{C})$ as the Shannon entropy mass carried by samples outside the corpus's high-probability mode region, and the effective dimensionality $d_{\text{eff}}(\mathcal{C})$ as the participation ratio of the corpus embedding covariance,

$$d_{\text{eff}}(\mathcal{C}) = \frac{\left(\sum_{i=1}^D \lambda_i\right)^2}{\sum_{i=1}^D \lambda_i^2},$$

where λ_i are the eigenvalues of the embedding covariance matrix in ambient dimension D . Under the constant- ε model,

$$H_{\text{tail}}(g) = H_{\text{tail}}(0) (1 - \varepsilon)^g, \quad d_{\text{eff}}(g) = d_{\text{eff}}(0) (1 - \varepsilon)^g.$$

Both follow by induction: $H_{\text{tail}}(g+1) = (1 - \varepsilon)H_{\text{tail}}(g)$ by definition of T_ε , and the participation ratio scales identically if compression removes variance mass proportionally across retained directions (the isotropic-loss assumption; we revisit it in Section 6).

3.3 Consilience metric

Let the corpus be partitioned into domains, with the partition either given by metadata or learned intrinsically by a compression-based clustering criterion in the spirit of [6]. Define the cross-domain link density

$$\rho_x(\mathcal{C}) = \frac{\#\{\text{links}(a, b) : \text{dom}(a) \neq \text{dom}(b)\}}{\#\{\text{links}\}},$$

where links are citation edges, co-occurrence edges in embedding k -nearest-neighbor graphs, or co-membership in verification-checked equivalence classes. The empirical anchor from [14] is that 97.2% of external technical vocabulary occurs in exactly one discipline, so the baseline probability that a vocabulary item is shared across disciplines is

$$p_{\text{bridge}} = 1 - 0.972 = 0.028.$$

We define the consilience gain of generation $g \rightarrow g+1$ as the number of *novel* cross-domain links created, modeled as

$$\Delta C_g = \kappa p_{\text{bridge}} V_g,$$

where V_g is the surviving technical-vocabulary mass at generation g and $\kappa \geq 0$ is a dimensionless coupling constant measuring how strongly the generative model exploits available bridge vocabulary. We assume $V_g = V_0(1 - \varepsilon)^g$: vocabulary decays at the same rate as tail entropy, since rare vocabulary is tail mass.

3.4 The break-even condition

Compression loss per generation in entropy units is $\varepsilon H_{\text{tail}}(g)$. Consilience gain, converted to the same units by a proportionality constant β (bits of tail-structure value per novel cross-domain link), is $\beta \kappa p_{\text{bridge}} V_0(1 - \varepsilon)^g$. Compensation requires

$$\beta \kappa p_{\text{bridge}} V_0(1 - \varepsilon)^g \geq \varepsilon H_{\text{tail}}(0)(1 - \varepsilon)^g.$$

The factor $(1 - \varepsilon)^g$ cancels: the condition is **generation-independent**. Dividing both sides by $(1 - \varepsilon)^g \beta V_0$,

$$\kappa \geq \kappa^* \equiv \frac{\varepsilon H_{\text{tail}}(0)}{\beta V_0 p_{\text{bridge}}}.$$

This is the paper's central structural result: recursive compression does not progressively lose the race against consilience — either the coupling constant clears a fixed threshold and compensation holds at every generation, or it does not and compression dominates at every generation. Interventions act on specific quantities: diversity enforcement raises κ ; formal verification raises β by making links trustworthy; re-anchoring with real data resets ε downward and injects fresh V_0 , $H_{\text{tail}}(0)$.

3.5 Complementary pair-counting model (labeled secondary model)

An alternative, stronger growth model treats a cross-domain link as an independent pair event. If corpus size grows as $N_g = N_0 \beta_c^g$ and per-pair link probability decays as $\pi_0 \rho^g$ (with $\rho = 1 - \varepsilon$ and tail-concentration exponent $\gamma \geq 0$), the expected link count obeys $E_g \approx E_0 (\beta_c^2 \rho^\gamma)^g$, with $E_0 = \binom{N_0}{2} \pi_0$. Consilience then grows iff

$$\beta_c^2 \rho^\gamma > 1 \iff \rho > \beta_c^{-2/\gamma}.$$

We report this model separately in Section 5 as a labeled projection under its own declared parameters, because it assumes link independence — an assumption the break-even model does not require.

3.6 Hybrid re-anchoring

In a hybrid pipeline, real data re-enters every m generations, resetting H_{tail} to a fraction r of its generation-0 value ($0 < r \leq 1$; $r = 1$ is a full reset). Between anchors the geometric law holds, so the long-run average tail entropy over one cycle is

$$\bar{H} = \frac{H_{\text{tail}}(0) r}{m} \sum_{j=0}^{m-1} (1 - \varepsilon)^j = \frac{H_{\text{tail}}(0) r}{m} \cdot \frac{1 - (1 - \varepsilon)^m}{\varepsilon}.$$

All symbols: g generation index; ε per-generation compression fraction; $\rho = 1 - \varepsilon$ retention factor; $H_{\text{tail}}(0)$ initial tail entropy; $d_{\text{eff}}(0)$ initial effective dimensionality; p_{bridge} bridge-vocabulary proportion; V_0 initial vocabulary mass; κ consilience coupling; β link-value proportionality; β_c corpus growth factor (secondary model); γ tail-concentration exponent; K link-type capacity (secondary model); m re-anchoring period; r reset fraction.

4. Analysis

Every input number is stated here with its source; every arithmetic step is shown.

Input 1 (from [14], empirical): 97.2% of external technical vocabulary occurs in exactly one discipline. Hence

$$p_{\text{bridge}} = 1 - 0.972 = 0.028.$$

Input 2 (assumption, labeled): per-generation compression fraction $\varepsilon = 0.15$, i.e., each recursive generation retains 85% of tail-entropy mass. This is a modeling assumption, not a measurement; Section 6 discusses its status. (Two of the reconciled drafts used alternative retention factors, $\rho = 0.90$ and $\rho = 0.95$; see Appendix A. We adopt $\varepsilon = 0.15$ as the primary convention and report the others as sensitivity.)

Input 3 (assumption, labeled): initial tail entropy $H_{\text{tail}}(0) = 12$ bits, a representative order of magnitude for a corpus whose tail spans several independent rare-event axes.

Input 4 (assumption, labeled): $\beta V_0 = 1$ bit-value unit, fixing the normalization of the consilience coupling; κ is then measured in units of this normalization.

Derivation 1 — tail-entropy trajectory. With $\varepsilon = 0.15$, the retention factor is $1 - \varepsilon = 0.85$. Then

$$H_{\text{tail}}(g) = 12 \times 0.85^g \text{ bits.}$$

Compute powers of 0.85:

$$0.85^1 = 0.85, \quad 0.85^2 = 0.7225, \quad 0.85^3 = 0.614125,$$

$$0.85^4 = 0.52200625, \quad 0.85^5 = 0.4437053125.$$

So $H_{\text{tail}}(5) = 12 \times 0.4437053125 = 5.32446375$ bits. Continuing,

$$0.85^{10} = (0.85^5)^2 = 0.4437053125^2 = 0.19687440434,$$

$$H_{\text{tail}}(10) = 12 \times 0.19687440434 = 2.36249285 \text{ bits.}$$

Derivation 2 — generations to the one-bit threshold. We ask when $H_{\text{tail}}(g) < 1$ bit:

$$12 \times 0.85^g < 1 \iff 0.85^g < \frac{1}{12} \iff g > \frac{\ln(12)}{\ln(1/0.85)}.$$

With $\ln(12) = 2.48490665$ and $\ln(1/0.85) = 0.16251893$,

$$g > \frac{2.48490665}{0.16251893} = 15.2906.$$

Since g is an integer, tail entropy drops below one bit at generation $g = 16$. Check:

$$0.85^{16} = 0.85^{10} \times 0.85^5 \times 0.85 = 0.19687440434 \times 0.4437053125 \times 0.85 = 0.07427498;$$

$$12 \times 0.07427498 = 0.89129976 < 1. \text{ At } g = 15: 0.89129976/0.85 = 1.04858795 > 1. \text{ Confirmed.}$$

Derivation 3 — break-even coupling. From Section 3.4 with $\beta V_0 = 1$:

$$\kappa^* = \frac{\varepsilon H_{\text{tail}}(0)}{\beta V_0 p_{\text{bridge}}} = \frac{0.15 \times 12}{1 \times 0.028} = \frac{1.8}{0.028} = 64.285714 \dots \approx 64.29.$$

The generative process must create about 64 times more novel cross-domain links per unit of surviving vocabulary than a baseline that samples bridge vocabulary uniformly at rate $p_{\text{bridge}} = 0.028$. Equivalently, the consilience gain per generation at break-even is

$$\Delta C^* = \kappa^* p_{\text{bridge}} V_0 = 64.285714 \times 0.028 \times 1 = 1.8 \text{ link-value units,}$$

which matches the compression loss $\varepsilon H_{\text{tail}}(0) = 0.15 \times 12 = 1.8$, as required by construction.

Derivation 4 — sensitivity of the threshold to ε . Since κ^* is linear in ε ,

$$\kappa^*(\varepsilon) = \frac{12\varepsilon}{0.028} = 428.571429\varepsilon.$$

For a well-regularized pipeline with $\varepsilon = 0.05$: $\kappa^* = 428.571429 \times 0.05 = 21.43$. For an aggressive pipeline with $\varepsilon = 0.30$: $\kappa^* = 428.571429 \times 0.30 = 128.57$. Reducing ε by a factor of 6 (from 0.30 to 0.05) reduces the required coupling by the same factor, since $128.57/21.43 = 6.000$.

Derivation 5 — effect of re-anchoring. With $\varepsilon = 0.15$, $r = 1$, $m = 5$: the geometric sum is

$$\sum_{j=0}^4 0.85^j = \frac{1 - 0.85^5}{0.15} = \frac{1 - 0.4437053125}{0.15} = \frac{0.5562946875}{0.15} = 3.70863125.$$

Then

$$\bar{H} = \frac{12 \times 3.70863125}{5} = \frac{44.503575}{5} = 8.900715 \text{ bits},$$

versus the un-anchored value at $g = 5$ of 5.32446375 bits. The ratio is $8.900715/5.32446375 = 1.6712$, i.e., a 67.1% uplift in sustained tail mass relative to the pure synthetic pipeline's end-of-cycle value.

Derivation 6 — vocabulary survival. Since $V_g = V_0(1 - \varepsilon)^g$, at $g = 16$ the surviving vocabulary fraction is $0.85^{16} = 0.07427498$, i.e., about 7.43% of the original technical-vocabulary mass survives sixteen un-anchored generations. The number of distinct bridge-vocabulary items available for novel links shrinks by the same factor, which is why the break-even condition — where both sides carry $(1 - \varepsilon)^g$ — is generation-independent.

Derivation 7 — secondary-model crossover thresholds (labeled projection). Under the pair-counting model of Section 3.5 with $\gamma = 1$: the condition $\rho > \beta_c^{-2}$ gives, for $\beta_c = 1.5$, the threshold $\rho > 1/2.25 = 0.4444$; for $\gamma = 2$, $\rho > \beta_c^{-1} = 0.6667$. With $\rho = 0.85$ (our primary convention), both thresholds are cleared, so under the independence assumption consilience-link counts would grow. With the alternative retention factor $\rho = 0.95$ (Appendix A) and $\beta_c = 1.5$, the per-generation growth factor is $\beta_c^2 \rho = 2.25 \times 0.95 = 2.1375$, and with $N_0 = 10^4$, $\pi_0 = 10^{-6}$: $E_0 = \binom{10^4}{2} \times 10^{-6} = 49,995,000 \times 10^{-6} = 49.995$, giving $E_{10} = 49.995 \times 2.1375^{10} \approx 99,524$ expected links. These are projections of a labeled secondary model, not measurements; they rest on the link-independence assumption criticized in Section 6.

5. Results

We report only quantities derived in Section 4, plus projections labeled as such.

R1 (derived). Under the constant-compression model with $\varepsilon = 0.15$, tail entropy decays as $H_{\text{tail}}(g) = 12 \times 0.85^g$ bits: 5.32 bits at $g = 5$, 2.36 bits at $g = 10$, and below the one-bit usefulness threshold at $g = 16$.

R2 (derived). The break-even consilience coupling is $\kappa^* = \varepsilon H_{\text{tail}}(0)/(\beta V_0 p_{\text{bridge}}) = 1.8/0.028 \approx 64.29$ under the stated normalization ($\beta V_0 = 1$). The condition is generation-independent: the factor $(1 - \varepsilon)^g$ cancels from both sides.

R3 (derived). κ^* is linear in ε with slope 428.57 per unit ε : $\kappa^* = 21.43$ at $\varepsilon = 0.05$ and $\kappa^* = 128.57$ at $\varepsilon = 0.30$.

R4 (derived). Re-anchoring every $m = 5$ generations with full reset ($r = 1$) sustains an average tail entropy of 8.90 bits over the cycle, a 67.1% uplift over the pure synthetic pipeline's end-of-cycle value of 5.32 bits.

R5 (derived). After 16 un-anchored generations, only $0.85^{16} = 0.07427498 \approx 7.43\%$ of the original technical-vocabulary mass survives (Derivation 6), so the pool of bridge vocabulary available for novel cross-domain links has shrunk by a factor of $0.85^{16} \approx 13.5$.

R6 (labeled projection). Under the secondary pair-counting model with $\rho = 0.95$, $\beta_c = 1.5$, $N_0 = 10^4$, $\pi_0 = 10^{-6}$, the expected link count grows from $E_0 = 49.995$ to $E_{10} \approx 99,524$ over ten generations (Derivation 7). This rests on the link-independence assumption and is not a measurement.

6. Discussion

Status of the assumptions. The per-generation compression fraction $\varepsilon = 0.15$ and the initial tail entropy $H_{\text{tail}}(0) = 12$ bits are modeling assumptions, not measurements. The break-even threshold κ^* is linear in both (Derivations 3 and 4), so the qualitative structure — a generation-independent inequality on a dimensionless coupling — survives any recalibration, but the numerical value 64.29 does not. The isotropic-loss assumption behind the d_{eff} decay law is the most fragile: if compression removes variance non-uniformly across directions, d_{eff} may decay faster or slower than H_{tail} , and the two decay laws decouple.

The coupling constant is not measured. Nothing in this paper measures κ for any real generative pipeline. The framework predicts that compensation is all-or-nothing per pipeline; the open empirical question is whether real model-on-model training loops place κ above or below κ^* . A measurement protocol would need to count novel cross-domain links per unit of surviving vocabulary across generations and compare against per-generation entropy loss — feasible in principle, not performed here.

Failure modes. (i) The link-independence assumption of the secondary model (Section 3.5) likely overestimates consilience growth, since cross-domain links cluster on the same scarce bridge vocabulary that [14] shows is concentrated in method-level terms. (ii) The consilience metric of Section 3.3 counts links, not their value; a corpus could inflate ρ_x with spurious cross-domain co-occurrences. (iii) The 97.2% single-discipline figure is an anchor for one corpus's vocabulary statistics, not a universal constant; p_{bridge} could differ by an order of magnitude elsewhere, and κ^* scales as $1/p_{\text{bridge}}$.

What would falsify the claims. The central structural result is falsified if tail entropy and consilience gain do not scale with the same factor of $(1 - \varepsilon)^g$: for example, if bridge vocabulary decays more slowly than tail entropy (breaking the shared $V_g = V_0(1 - \varepsilon)^g$ assumption), the cancellation fails and compensation could become generation-dependent. The one-bit threshold prediction ($g = 16$ under $\varepsilon = 0.15$) is falsified by measuring tail entropy above one bit at later generations in an un-anchored pipeline. The re-anchoring result (67.1% uplift) is falsified if reset fraction $r < 1$ in practice reduces the sustained average below the un-anchored trajectory.

Limitations of the literature base. Several cited works ([3], [4], [7], [11]) have supplied summaries too thin to import quantitative findings, and two ([12], [13]) have no supplied summaries at all; they are cited for design patterns and corpus context only. The empirical anchor rests on a single audit [14].

7. Conclusion

We formalized recursive model-on-model training as a compression channel and showed that the tension between compression loss and emergent consilience resolves into a generation-independent inequality on a dimensionless coupling constant κ : with the stated normalization, break-even requires $\kappa \geq 64.29$. Under the constant-compression model, tail entropy falls below one bit at generation $g = 16$, and only 7.43% of technical vocabulary survives that long; periodic re-anchoring every 5 generations sustains 8.90 bits of average tail entropy, a 67.1% uplift. The framework's value is that every intervention — diversity enforcement, verification, re-anchoring — maps to a named term of one inequality, and its predictions are falsifiable. No simulations or empirical measurements were performed; all numbers derive from stated assumptions, and the immediate next step is measuring κ on a real recursive training loop.

References

- [1] Towards Bayesian Data Compression. arXiv:2010.10375v2. <https://arxiv.org/abs/2010.10375v2> [2] The optimal transport paradigm enables data compression in data-driven robust control. arXiv:2005.09393v2. <https://arxiv.org/abs/2005.09393v2> [3] An Adaptive Column Compression Family for Self-Driving Databases. arXiv:2209.02334v1. <https://arxiv.org/abs/2209.02334v1> [4] Peak-Nadir Encoding for Efficient CGM Data Compression and High-Fidelity Reconstruction. arXiv:2601.00608v1. <https://arxiv.org/abs/2601.00608v1> [5] Dataset Distillation as Data Compression: A Rate-Utility Perspective. arXiv:2507.17221v1. <https://arxiv.org/abs/2507.17221v1> [6] Spatial regionalization based on optimal information compression. arXiv:2111.01813v3. <https://arxiv.org/abs/2111.01813v3> [7] ζ -QVAE: A Quantum Variational Autoencoder utilizing Regularized Mixed-state Latent Representations. arXiv:2402.17749v3. <https://arxiv.org/abs/2402.17749v3> [8] Dataset Distillation Efficiently Encodes Low-Dimensional Representations from Gradient-Based Learning of Non-Linear Tasks. arXiv:2603.14830v3. <https://arxiv.org/abs/2603.14830v3> [9] Constraints on dark energy from H II starburst galaxy apparent magnitude versus redshift data. arXiv:1110.5626v1. <https://arxiv.org/abs/1110.5626v1> [10] Privacy Measurement in Tabular Synthetic Data: State of the Art and Future Research Directions. arXiv:2311.17453v1. <https://arxiv.org/abs/2311.17453v1> [11] Modified Levenberg-Marquardt Algorithm For Tensor CP Decomposition in Image Compression. arXiv:2401.04670v1. <https://arxiv.org/abs/2401.04670v1> [12] <https://papers.qnfo.org/papers/projective-geometric-frameworks-for-semantic-structures/>. QNFO: Projective Geometric Frameworks for Semantic Structures. [13] <https://papers.qnfo.org/papers/syntactic-generation-primitive-distinctions/>. QNFO: Syntactic Generation. [14] <https://papers.qnfo.org/papers/terminology-silos-consilience/>. QNFO: Terminology Silos and the Consilience Gap: A Quantitative Audit of Cross-Domain Vocabulary.

Appendix A. Divergence report

The reconciled drafts disagreed on the per-generation retention factor: draft conventions included $\rho = 0.90$ (i.e., $\varepsilon = 0.10$), $\rho = 0.85$ (i.e., $\varepsilon = 0.15$), and $\rho = 0.95$ (i.e., $\varepsilon = 0.05$). The disagreement reflects differing priors on how aggressively a recursive training loop discards tail mass; none of the drafts measured ε . **Resolution:** the main text adopts $\varepsilon = 0.15$ ($\rho = 0.85$) as the primary convention because it is the intermediate choice, and reports the alternatives as sensitivity: κ^* is linear in ε (Derivation 4), so $\kappa^* = 428.571429\varepsilon$ gives $\kappa^* \approx 42.86$ at $\varepsilon = 0.10$ and $\kappa^* \approx 21.43$ at $\varepsilon = 0.05$. The secondary-model projection in Derivation 7 uses $\rho = 0.95$ under its own declared parameters and is labeled as such. No other substantive divergences between drafts were identified; convergent claims are listed in Appendix B.

Appendix B. Claim attribution

Claim	Substance	Source drafts	Status
C1	Recursive training modeled as a per-generation compression operator retaining $1 - \varepsilon$ of tail-entropy mass, yielding geometric decay $H_{\text{tail}}(g) = H_{\text{tail}}(0)(1 - \varepsilon)^g$	A, B, C	CONVERGENT
C2	Consilience modeled as creation of novel cross-domain links from surviving bridge vocabulary, $\Delta C_g = \kappa p_{\text{bridge}} V_g$	A, B, C	CONVERGENT
C3	Break-even condition is generation-independent because both compression loss and consilience gain scale as $(1 - \varepsilon)^g$, reducing to $\kappa \geq \kappa^*$	A, B, C	CONVERGENT
C4	Empirical anchor: 97.2% of external technical vocabulary occurs in exactly one discipline, giving $p_{\text{bridge}} = 0.028$ [14]	A, B, C	CONVERGENT
C5	One-bit usefulness threshold reached at generation $g = 16$ under $\varepsilon = 0.15$	A, B	CONVERGENT
C6	Re-anchoring every $m = 5$ generations sustains 8.90 bits average tail entropy, a 67.1% uplift	A, B	CONVERGENT
C7	Secondary pair-counting crossover model with link-independence assumption, reported as a labeled projection	B, C	CONVERGENT
C8	Choice of primary retention factor $\rho = 0.85$ ($\varepsilon = 0.15$)	A vs. B ($\rho = 0.90$) vs. C ($\rho = 0.95$)	DIVERGENT (resolved in Appendix A)

ed as novel cross-domain links created from surviving bridge vocabulary, with coupling κ and empirical anchor $p_{\text{bridge}} = 0.028$ from the 97.2% single-discipline audit [14] | A, B, C | CONVERGENT |

| C3 | Break-even condition is generation-independent because both sides carry $(1 - \varepsilon)^g$; reduces to $\kappa \geq \kappa^* = \varepsilon H_{\text{tail}}(0) / (\beta V_0 p_{\text{bridge}})$ | A, B, C | CONVERGENT || C4 | Primary convention $\varepsilon = 0.15$, $H_{\text{tail}}(0) = 12$ bits, $\beta V_0 = 1$ | B (primary), A and C (as sensitivity) | SINGLE (resolved via Appendix A) || C5 | One-bit usefulness threshold reached at $g = 16$ under $\varepsilon = 0.15$ | A, B | CONVERGENT || C6 | Re-anchoring every $m = 5$ generations with $r = 1$ sustains $\bar{H} = 8.90$ bits, a 67.1% uplift | A, B, C | CONVERGENT || C7 | Secondary pair-counting crossover model with growth condition $\beta_c^2 \rho^\gamma > 1$ and projected $E_{10} \approx 99,524$ under $\rho = 0.95$ | C | SINGLE || C8 | Interventions map to break-even terms: diversity enforcement raises κ , verification raises β , re-anchoring resets ε and injects fresh V_0 , $H_{\text{tail}}(0)$ | A, B | CONVERGENT |