Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

J-space Forecasts Model Collapse, but Only Anchoring the Function Prevents It

ChaeWoo Son

soncwnetp@gmail.com

Independent Researcher https://orcid.org/0009-0009-6108-2373

Research Article

Keywords: model collapse, recursive training, synthetic data, interpretability, J-space, knowledge distillation, catastrophic forgetting, language models

Posted Date: August 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10713541/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

J-space Forecasts Model Collapse, but Only Anchoring the Function Prevents It

ChaeWoo Son
Independent Researcher
soncwnetp@gmail.com

Abstract

As language models are trained on ever more text of their own making, their outputs degrade — a process termed model collapse, in which the tails of the original data distribution are the first to disappear. Collapse has been characterised almost entirely at the level of outputs. Where in the model the injury occurs, and whether it can be prevented at that site, remains open.

We test this in a controlled recursive-training testbed built on a 70-million-parameter model, tracking collapse through J-space: a low-dimensional residual-stream subspace recently shown to hold the representations a model can report and manipulate. Drift in J-space preceded any measurable rise in held-out cross-entropy. Fine-tuning on fresh real data also shifted the subspace — the raw signal is not specific to collapse — but the regimes separate sharply: self-generated data produced a large, seed-consistent drop within ten steps (74%) where real data produced only small, seed-inconsistent early movement, and drove the subspace to a far deeper floor (98% versus 55% below baseline). Across frequency strata the drift was dominantly uniform — a global displacement dwarfing any residual tail component: in this space, collapse is not primarily a loss of tails.

To test whether J-space is the site of the injury, we pinned its geometry during training; this preserved the subspace almost perfectly but prevented only 12% of the damage, and no more than 17% even when the whole layer was pinned. Anchoring the function rather than the geometry — a KL penalty to the pre-collapse model on 64 real sequences — prevented 99.6% of the damage in one generation and 95% across six, at no cost in fit to the synthetic corpus and leaving at least 23-fold less residual damage than mixing the same real data into training.

Taken together, these results are consistent with collapse being an injury to the input–output mapping — one that leaves an early trace in representation space but is not repaired by stabilising that space. These findings come from one small model and dataset, but if they hold at scale, a handful of real sequences under a functional anchor may be a cheap and general safeguard.

1 Introduction

Language models are increasingly trained on text produced by earlier models, whether deliberately or through the contamination of web-scale corpora. Shumailov et al. showed that when this becomes a closed loop — each generation trained on the output of the last — models degrade irreversibly, with the tails of the original data distribution disappearing first and the whole distribution eventually collapsing toward low-variance, repetitive output [1, 2]. Subsequent work has mapped the boundary conditions of this phenomenon: accumulating real data alongside synthetic data slows or averts collapse [3], and the effect can be cast as a change in scaling laws [4]. What these accounts

share is their vantage point: collapse has been characterised almost entirely at the level of outputs and data distributions, as a property of what models emit.

Interpretability research has meanwhile identified structure inside models that could, in principle, be watched instead. Recent work described "J-space", a low-dimensional subspace of the residual stream obtained by differentiating output behaviour with respect to hidden activations, which accounts for only a small fraction of representational variance yet governs which concepts a model can report and manipulate [5]. To date this lens has been applied statically — to ask what a fixed model is representing at a moment in time — rather than longitudinally, to track how representations deteriorate as a model is damaged by training.

The intersection of these two literatures is empty, and the questions that live there are consequential. Collapse is measured in outputs, but the damage must accrue somewhere inside the model before it surfaces; a signal that moves earlier than perplexity would allow contaminated training to be caught in progress rather than diagnosed afterwards. More pointedly, if the site of the damage could be identified, one could ask whether protecting that site prevents the damage — turning an interpretability lens into an intervention target. Neither the location of collapse inside a model, nor whether it can be prevented at that location, has to our knowledge been tested.

Here we use J-space first as a longitudinal lens on collapse and then as a candidate site of intervention, in a controlled recursive-training testbed built on Pythia-70M [6] and WikiText-103 [7], with matched real-data training as a control throughout.

J-space activation drift precedes any measurable rise in held-out cross-entropy, but with a qualification we state at the outset: ordinary fine-tuning on fresh real data also moves the subspace, so the raw signal is not by itself specific to collapse. What distinguishes collapse is the speed, consistency and depth of the drift — a seed-consistent 74% drop within ten steps and a 98% terminal floor under self-generated data, against small, seed-inconsistent early movement and a 55% floor under real data at matched settings ($n = 3$ seeds). The drift is also dominantly uniform: once a global shift is factored out, rare concepts fare scarcely worse than frequent ones (a residual differential of about 3% of the global displacement), so in this representational lens collapse is chiefly a global displacement rather than the tail-first erosion seen in output distributions.

This early visibility does not make J-space the site of the injury. Pinning its geometry during training preserves the subspace almost perfectly at no cost to fitting the synthetic corpus, yet prevents only 12% of the damage; the ceiling is 17% even when the entire layer is pinned, although J-space carries roughly ten times the protective density per dimension of the rest of the layer. Anchoring the function instead — a KL penalty toward the pre-collapse model on 64 real sequences — prevents 99.6% of the damage in one generation and 95% across six recursive generations, and outperforms mixing the same real data into training by at least a factor of 23 in residual damage.

The functional anchor itself is a form of self-distillation and is not our contribution; what is new is the dissociation the experiments expose — the earliest signature of collapse lives in representation space while the effective point of intervention lies in function space — and the map of anchoring targets that establishes it. All findings come from a single small model, dataset and layer, in an acute collapse regime; we treat them as a mechanistic case study and specify in the Discussion the scale, regime and layer sweeps required to test their generality.

2 Results

2.1 Recursive self-training collapses the model within one generation

We fine-tuned Pythia-70M on 3M tokens of WikiText-103 (generation 0) and then iterated for six generations: sample 1M synthetic tokens from the current model, fine-tune one epoch on them, and measure held-out cross-entropy on real text. Cross-entropy rose from 3.66 to 7.24 nats after one generation and fluctuated between 8.2 and 9.1 nats thereafter (Fig. 1, red; single lineage); the non-monotone tail of this curve is consistent with run-to-run variation in which repetitive mode each generation's corpus happens to lock onto, once collapse has saturated. Synthetic corpora degenerated into repetitive loops, reproducing at this scale the qualitative signature reported by Shumailov et al. [1]. Because collapse saturates within one generation under this protocol, subsequent analyses resolve the injury *within* the first generational fine-tune rather than across generations.

2.2 J-space drift precedes the rise in cross-entropy but is not by itself specific to collapse

We extracted J-space once from the generation-0 model (layer 4 of 6; 256 concepts stratified over four log-frequency bins; rank-64 basis) and, during fine-tuning on the generation-1 synthetic corpus, measured every ten steps the fraction of probe positions at which each concept's Jacobian score exceeded its generation-0 95th-percentile threshold ("firing rate"), alongside held-out cross-entropy (Fig. 2). At learning rate 3×10^{-5}, the firing rate fell $74 \pm 3\%$ within ten steps while cross-entropy had risen 9%; the firing rate crossed a 20% deviation criterion at step 10 and cross-entropy at step 20 in all three seeds. A matched control fine-tuned on fresh real text also drifted, but differently: its early movement was small and seed-inconsistent (mean 6% at step 10, ranging from -11% to $+27\%$ across seeds), and its terminal floor was $55 \pm 2\%$ below baseline against $98 \pm 1\%$ for synthetic — while its cross-entropy never rose. The raw firing rate is therefore an early but non-specific signal; what separates the regimes is the speed and consistency of early drift and the depth of the terminal floor, jointly with the direction of cross-entropy. At learning rate 2×10^{-4} with per-step resolution, the same ordering held (firing-rate break at step 2, cross-entropy at step 6; single seed).

2.3 Within J-space, drift is dominantly uniform across frequency strata

The canonical account of collapse holds that rare modes disappear first. To test this in J-space we re-calibrated, at each measurement, a single global threshold offset such that the mean firing rate over all concepts equals its generation-0 value of 5%, and read the rare-bin rate under that matched offset; selective loss of rare concepts would push this quantity below 5% under synthetic training. It stayed flat in both regimes (synthetic: $0.0485 \to 0.0475$; real: $0.0485 \to 0.0506$; $n = 3$ each), and uncorrected rare-bin and all-concept rates were indistinguishable throughout (Fig. 4). Within this lens, collapse displaces the representation globally rather than eroding its tail — a contrast with, not a contradiction of, the tail-first behaviour of output distributions, which our synthetic corpora do exhibit.

Two reanalyses probe below the binarised rates. First, re-extracting j_v per generation shows that the lens itself rotates substantially across the unanchored lineage (mean $\cos(j_v^0, j_v^t)$ falling from 0.79 at generation 1 to 0.47 at generation 6) — but indistinguishably across strata, ruling out a rare-specific rotation that binarised rates would miss. Second, comparing continuous standardised scores before binarisation (generation $0 \to 1$), the displacement is overwhelmingly shared: a global shift of -1.55 standard deviations, against a residual rare-versus-frequent differential of -0.04 standard

deviations (effect size $d = 0.35$) — roughly 3% of the global component, in the direction of slightly deeper rare displacement. We therefore state the result as dominantly uniform displacement, with at most a small residual tail differential; resolving whether that residual grows in slower regimes requires the chronic protocol outlined in the Outlook.

2.4 Pinning representational geometry preserves it but prevents little of the injury

We added to the training loss a penalty on deviation of anchored-text activations from their generation-0 values, $\lambda\|U^\top h_\theta(x) - U^\top h_{\theta_0}(x)\|^2$, over 64 real sequences disjoint from the measurement probe, for three subspaces U: the J-space basis (64 dims), a random orthonormal basis of equal dimension, and the identity (the full 512-dim layer). All three pinned their targets essentially perfectly (anchor loss ≤ 0.001 at $\lambda = 10$) without impairing fit to the synthetic corpus (final training loss 0.16 versus 0.15 unanchored). Protection, however, was limited: $11.6 \pm 1.4\%$ of collapse damage for the J-space anchor ($n = 3$), $4.2 \pm 1.1\%$ for the random subspace ($n = 3$; basis and training seed both varied) and $\sim$17% for the full layer ($n = 1$), all measured against the 3-seed unanchored baseline (Table 1). Two readings follow. Per constrained dimension, J-space is roughly ten times as protective as the remainder of the layer, and pinning a random subspace leaves J-space itself free to drift (final firing rate 0.0009–0.0011 across seeds, identical to no anchor) — the subspace is privileged among representational targets. But the ceiling of $\sim$17% for the entire layer implies that most of the injury does not pass through the geometry of this layer at all.

2.5 Anchoring the function prevents nearly all of it

Replacing the geometric penalty with a functional one — the KL divergence from the frozen generation-0 model's next-token distribution, on the same 64 sequences — prevented $99.6 \pm 0.05\%$ of collapse damage at $\lambda = 10$ ($n = 3$; 97.9% at $\lambda = 1$, $n = 1$). The anchored model's final cross-entropy, 3.706 ± 0.001 nats, matched that of a model fine-tuned on fresh real data (3.729 ± 0.008, $n = 3$), while its training loss on the degenerate synthetic corpus was unchanged from the unanchored run: the model fit the synthetic distribution fully and lost essentially nothing on real text. Against the standard remedy at matched real-data budget — mixing the same 32k tokens into training, which prevented $53.5 \pm 1.3\%$ ($n = 3$) — the functional anchor left residual damage of 0.006 nats at $\lambda = 10$ (0.029 at $\lambda = 1$) versus 0.680 nats for mixing: at least 23-fold less, taking the weaker $\lambda = 1$ anchor as the conservative comparison. J-space drift was also largely suppressed (final firing rate 0.026 ± 0.000 versus 0.001 unanchored) although the function, not the representation, was the constrained quantity; the residual half-level of this rate (0.026 against the 0.05 baseline) is itself uninvestigated.

2.6 The functional anchor holds across six recursive generations

We re-ran the six-generation recursion with the functional anchor to the frozen generation-0 teacher active in every generational fine-tune ($\lambda = 10$, learning rate 2×10^{-4}; single lineage). Held-out cross-entropy rose from 3.66 to 3.91 nats over six generations — 0.25 nats in total, 7% of the unanchored lineage's first-generation loss alone — corresponding to 97.8% damage prevented at generation 1 and 95.3% cumulatively at generation 6 (Fig. 1, green). The residual is approximately linear at 0.04 nats per generation, so on this evidence the anchor slows collapse by roughly a factor of 30 rather than halting it. Training loss on each generation's synthetic corpus declined across

Table 1: Treatment arms at matched real-data budget (intra-generation protocol, learning rate 3×10^{-5}, 1M synthetic tokens; damage base 1.460 nats against the 3-seed unanchored baseline). Mean ± s.d. where $n = 3$. The ≥23-fold residual-damage comparison in the text is the ratio of residuals, 0.680/0.029 nats (mixing versus the $\lambda = 1$ anchor).

Arm	Final CE (nats)	Damage prevented	n
No treatment	5.159 ± 0.018	—	3
Random-64 anchor (repr.)	5.098 ± 0.017	$4.2 \pm 1.1\%$	3
J-space-64 anchor (repr.)	4.990 ± 0.020	$11.6 \pm 1.4\%$	3
Full-layer-512 anchor (repr.)	4.907	∼17%	1
Mixing 32k real tokens	4.379 ± 0.019	$53.5 \pm 1.3\%$	3
Functional anchor (logit-KL), $\lambda = 1$	3.729	∼98%	1
Functional anchor (logit-KL), $\lambda = 10$	3.706 ± 0.001	$99.6 \pm 0.05\%$	3
Real-data control	3.729 ± 0.008	n/a	3

generations ($0.24 \rightarrow 0.06$), consistent with a compounding effect: a model whose real-text behaviour is preserved emits less degenerate data, which in turn injures the next generation less.

3 Discussion

Why constraining behaviour on 64 sequences neutralises damage that constraining representations cannot admits a simple candidate explanation. Synthetic and real text occupy largely disjoint regions of input space, and the unchanged training loss under the functional anchor shows that fitting the degenerate corpus and preserving real-text behaviour are jointly achievable for this model. If so, the unanchored model's loss of real-text ability is interference — an optimisation-path effect rather than a capacity limit — and the anchor works by penalising exactly that interference where it matters. We note this account is inferred from the pattern of outcomes rather than demonstrated directly; tracing the gradient conflict itself is a natural follow-up.

The dissociation clarifies what J-space is and is not, in this setting. It is the most informative single window we found on the injury: drift there precedes cross-entropy, and per constrained dimension the subspace is roughly ten times as protective as the remainder of the layer. We use "precedes" in the operational sense of an early-warning signal — the quantity that moves first at matched criteria — not as a claim about where damage arrives first mechanistically; a readout can lead simply because it is the more sensitive instrument, and our criterion sweep establishes the robustness of the ordering, not its causal interpretation. It is not the site of the injury: pinning it, or the whole layer containing it, leaves most of the damage untouched. One further observation connects the two — when the function alone is anchored, J-space drift is largely suppressed as a by-product (final firing rate 0.026 versus 0.001 unanchored), consistent with the subspace covarying closely with real-text behaviour. J-space was originally characterised as a privileged subspace of fixed models — a static property [5]. Our observations raise the possibility that the privilege is also dynamic, in that the subspace remains the layer's most behaviour-coupled component while the model is being damaged; establishing this systematically, across layers and injury types, is a follow-up our single-layer results motivate but do not provide.

Six generations, rather than one, were necessary for three reasons. The phenomenon of interest is a closed loop — model output becoming the next model's training data — so a single-generation result would speak to fine-tuning on bad data, not to collapse as defined in the literature [1]. Failure modes invisible in one generation only surface across generations: the anchor's residual leak (0.04

nats per generation) could in principle compound, and the fixed generation-0 teacher could grow stale as the student's synthetic corpora drift; neither occurred over six generations. Testing staleness directly, the frozen teacher's cross-entropy on each anchored generation's corpus rose from 0.53 to about 1.2 nats by generation 4 and plateaued thereafter, versus 6.0 nats on the unanchored lineage's corpus by generation 6: the anchored lineage's data remains close to the teacher's distribution, and the early rise stabilises rather than compounds. Instead we observed the opposite compounding — per-generation training loss on synthetic corpora declined ($0.24 \rightarrow 0.06$), indicating that a model whose real-text behaviour is preserved emits less degenerate data, which injures the next generation less. This virtuous cycle is a finding only a multi-generation protocol can produce.

The functional anchor is a form of self-distillation, closely related to learning-without-forgetting [8, 9]; we claim no novelty for the technique. What the present experiments add is placement: a controlled comparison across anchoring targets — random subspace, privileged subspace, full layer, function — under identical budgets, which locates the injury by elimination. The comparison also carries a practical point. Replay-style mixing, the standard remedy [1, 3], spent the same 32k real tokens for a 53% reduction in damage; the same tokens spent as distillation targets yielded 99.6%. Where guarding real-text behaviour is the goal and curated real data is scarce, the anchored form appears to be the more efficient use of each real token, at least at this scale. Two scope notes bound this comparison: it is with equal-token mixing, not with accumulating real data across generations [3], which is a different and complementary regime; and the anchor's data efficiency is bought with compute — the teacher forward pass and the anchor-batch update raised wall-clock time per training step by roughly $1.8\times$ in our runs. The method does presuppose access to a known-healthy reference model; the present moment, in which models trained predominantly on human-written text still exist, is precisely when such references can be banked — a speculation we flag as such.

The acute regime was a design choice, not a convenience. A corpus that is entirely synthetic and collapses the model within one generation is the worst case for any defence: the training signal contains no real data at all, so whatever survives does so on the strength of the auxiliary term alone — an existence proof under maximal contamination — and the anatomical clarity of a fast, saturating injury is what made the elimination logic of the anchoring spectrum legible. The cost is the denominator: with 1.46 nats of unanchored damage, percentage protection is generous, and in a chronic regime where per-generation damage is of the same order as the anchor's absolute leak (0.04 nats per generation), the same leak would loom far larger in percentage terms. Whether the functional anchor retains its advantage there is governed by that absolute number, which is why we report it alongside every percentage.

Limitations. All results come from one model (Pythia-70M), one dataset (WikiText-103), one layer (4 of 6) and one collapse regime — an acute one, saturating within a generation, which makes protection percentages generous; we report the anchored lineage's absolute slope (0.04 nats per generation) alongside them for this reason. Several arms are single-seed (the full-layer anchor, the $\lambda = 1$ functional anchor, and both generational lineages); in particular the $\sim 10\times$ per-dimension protection figure depends on the single-seed full-layer arm. The multi-seed arms share one generation-1 corpus, so seed variance reflects training stochasticity (batch order, anchor sampling), not data-generation stochasticity; full replication with per-seed corpora is deferred to the planned scale sweep. The residual leak under the functional anchor is unexplained; its coincidence with slow representational drift (J-space tail ratio drifting from 1.5 toward 0.9 across generations) is suggestive but untested, as is its possible relation to the half-level firing-rate residual (0.026 versus 0.05) under the functional anchor. The uniform-drift null result concerns J-space, not output distributions, which do lose tails. Finally, the planned Goodhart control — splitting concepts into regularised and held-out halves —

was rendered moot when J-space proved not to be the therapeutic target, and has not been run.

Outlook. The interference account makes a testable prediction: because it rests on spare capacity and the separation of synthetic from real inputs, the functional anchor's protection should not weaken — and may strengthen — with model scale. A sweep across Pythia-160M to 1B under the identical protocol is the immediate next step, followed by a chronic regime (partial synthetic contamination sustained over many generations) in which collapse develops slowly and early warning matters most, a layer sweep for both the forecasting signal and the anchoring ceiling, and a gradient-level test of the interference account itself.

4 Methods

Operational definition of collapse. Throughout, "collapse damage" denotes the rise in per-token cross-entropy on a fixed 200k-token real-text held-out set, measured on the same model before and after training on synthetic data. This definition is deliberately narrow: it captures loss of real-text modelling ability, not the full qualitative structure of degradation, for which we report distinct-n diversity of synthetic corpora and a qualitative sample as auxiliaries (Appendix A) [10]. A matched real-data fine-tune serves as the control against which specificity claims are made. Protection percentages are computed against the 3-seed unanchored baseline (damage base 1.460 nats).

4.1 Model, data and recursive protocol

All experiments use Pythia-70M [6] (6 layers, $d = 512$) in fp32. Real text is WikiText-103 [7], tokenised with the Pythia tokeniser and packed into 512-token sequences. From a contiguous stream we carve, in order, a generation-0 training set (3.0M tokens), a held-out set (200k tokens) for cross-entropy, and a probe set of 512 sequences used for J-space extraction and measurement; a further 1M fresh tokens beyond these serve as the real-data control corpus. Generation 0 is one epoch of fine-tuning on the 3M-token set (AdamW, learning rate 2×10^{-4}, cosine schedule, 3% warm-up, batch 16 sequences, weight decay 0.01, gradient clipping 1.0). Each subsequent generation t samples 1M synthetic tokens from M_{t-1} and fine-tunes M_{t-1} for one epoch on them with the same hyper-parameters.

Sampling. Each synthetic sequence is seeded with a random 32-token window of real generation-0 text and continued for 512 tokens by nucleus sampling (top-$p = 0.95$, temperature 1, top-$k = 1024$ pre-filter, Gumbel-argmax draw); only the continuation is kept. Prefix conditioning avoids a degenerate heading-marker absorbing state that unconditional end-of-sequence-seeded sampling produced in this model, and matches the conditioning protocol of prior collapse studies [1].

4.2 J-lens and J-space

Let $h^{(\ell)}(x) \in \mathbb{R}^d$ be the residual stream after block ℓ ($\ell = 4$ throughout) at the final position of sequence x, and z_v the output logit for token v. We compute for each concept v the Jacobian direction

$$j_v = \frac{\partial z_v}{\partial h^{(\ell)}}, \tag{1}$$

averaged over 64 probe sequences and normalised to unit length. Our construction is a simplified variant of the original J-lens [5], which averages linearised effects on final-token likelihood across positions and prompts and forms J-space from sparse non-negative combinations; we use single final-position logit gradients and an SVD basis for tractability at this scale, and refer to the resulting object as J-space in that operational sense. The variant reproduces the defining functional signatures of the original construct in our setting — privilege over random subspaces of equal dimension (Section 2) and tight coupling to real-text behaviour (Section 3). Our choice of a mid-network layer is consistent with the original report that J-space structure concentrates in the middle block of the network [5]. The concept set $\mathcal{C}$ ($|\mathcal{C}| = 256$) is stratified: tokens of the generation-0 corpus (frequency ≥ 2) are split into four log-frequency quantile bins and 64 are sampled uniformly from each, so that the rarest bin is represented by construction. Stacking j_v into $J \in \mathbb{R}^{256\times 512}$, J-space is the top-$r$ left-singular basis $U_0 \in \mathbb{R}^{512\times 64}$ of $J^\top$. J_0 and U_0 are extracted once from generation 0 and held fixed; all cross-generation diagnostics are measured through this fixed lens. For the Grassmann drift metric $\|U_t U_t^\top - U_0 U_0^\top\|_F$ we additionally re-extract U_t per generation.

4.3 Firing-rate diagnostics

For a model θ and the probe set (positions N), let $s_{iv} = j_v^\top h_\theta(x_i)$. Per-concept thresholds ϑ_v are the 95th percentile of $s_{\cdot v}$ under the generation-0 model, so that every concept fires at rate 0.05 at generation 0. The firing rate is $r_v(\theta) = \frac{1}{N}\sum_i \mathbf{1}[s_{iv} > \vartheta_v]$; we report its mean over all concepts and over the rarest bin. The *drift-corrected rare-bin rate* bisects a single scalar offset δ such that $\frac{1}{N|\mathcal{C}|}\sum_{i,v}\mathbf{1}[s_{iv} > \vartheta_v + \delta] = 0.05$, and reports the rare-bin rate at that matched offset; a uniform shift of all scores is absorbed by δ, leaving only differential (rare-versus-common) change. Held-out cross-entropy is per-token negative log-likelihood on the 200k-token real held-out set.

4.4 Intra-generation probe

To resolve the timing of the injury we re-run the generation-0 $\to$ 1 fine-tune (245 steps, batch 8, gradient accumulation 2) pausing every 10 steps (or every step for the first 30) to compute the diagnostics above on a 64-sequence probe subset and cross-entropy on a 48-sequence held-out subset. Runs differ only in learning rate (2×10^{-4} or 3×10^{-5}), training corpus (synthetic, real control, or 3%-mixed) and anchor.

4.5 Anchors

All anchors act on an anchor set $\mathcal{A}$ of 64 real probe sequences disjoint from the measurement subset. Per training step a batch of four anchor sequences is drawn.

Representational (geometric) anchor. With $C_0(x) = U^\top h_{\theta_0}(x)$ cached from generation 0,

$$\mathcal{L}_{\text{rep}} = \frac{1}{\sigma_0^2}\,\mathbb{E}_{x\in\mathcal{A}}\big\|U^\top h_\theta(x) - C_0(x)\big\|_2^2, \tag{2}$$

with σ_0^2 the variance of C_0 and $U \in \{U_0,\ U_{\text{rand}},\ I_{512}\}$ (U_{rand}: fixed-seed random orthonormal 64-dim basis).

Functional anchor. With θ_0 frozen as teacher,

$$\mathcal{L}_{\text{fn}} = \frac{1}{T}\,\mathbb{E}_{x\in\mathcal{A}}\,\mathrm{KL}\big(p_{\theta_0}(\cdot \mid x)\,\|\,p_\theta(\cdot \mid x)\big), \tag{3}$$

summed over positions and divided by sequence length T. The KL is taken between full next-token distributions over the entire vocabulary, at every position of the anchor sequences, and averaged per token; no token subset — in particular, none of the concept set $\mathcal{C}$ used for measurement — is singled out. The anchor set is disjoint from both the measurement probe and the held-out evaluation set, so the optimised quantity and every reported metric are computed on separate data. Total loss is $\mathcal{L}_{\text{CE}}(\text{synthetic}) + \lambda\,\mathcal{L}_{\text{anchor}}$ with $\lambda \in \{1, 10\}$.

Mixing control. The same 64 sequences are shuffled into the synthetic training set (2,017 sequences, 3.2% real) with no auxiliary loss.

Generational anchoring. In the anchored lineage the teacher is always the generation-0 model.

4.6 Break-point criterion

A metric "breaks" at the first measurement deviating $\geq 20\%$ from its step-0 value (downward for firing rates, upward for cross-entropy). The 20% level is a reporting choice, not a tuned parameter: sweeping the criterion from 10% to 40% leaves the ordering unchanged in all three seeds — the firing rate breaks at the first measurement (step 10) at every level, while the cross-entropy break moves later as the criterion tightens (from step 20 to step 110–160, or is never reached). The lead is therefore robust to, and widens with, the choice of criterion.

4.7 Code and data availability

All code (data preparation, recursive protocol, J-lens extraction, diagnostics, anchors, figure generation) and per-run diagnostic files will be released in a public repository upon posting of this preprint. WikiText-103 and Pythia-70M are publicly available.

Declarations

Competing interests. The author declares no competing interests.

Funding. This work received no external funding.

Ethics. Not applicable: no human or animal subjects; public text data only.

A Synthetic-corpus diversity and a qualitative sample

Table 2 reports distinct-n diversity (unique n-grams over total, first 200k tokens) of the corpus *emitted by* each generation's model, for both lineages. Two patterns warrant comment. First, diversity drops sharply after the first synthetic generation in both lineages. Second, and counterintuitively, the unanchored lineage's diversity then *rebounds* (distinct-3 back to 0.038 by generation 5) while the anchored lineage stays flat near 0.022 — yet the teacher's cross-entropy on those same corpora (Section 3) is 6.0 nats for the unanchored lineage against $\leq$1.2 for the anchored one. Diversity alone is therefore not a quality measure here: the collapsed model's late corpora are diverse but incoherent, whereas the anchored model's corpora are moderately repetitive but remain close to the healthy teacher's distribution. The pair (teacher cross-entropy, distinct-n) separates the two regimes; neither does alone.

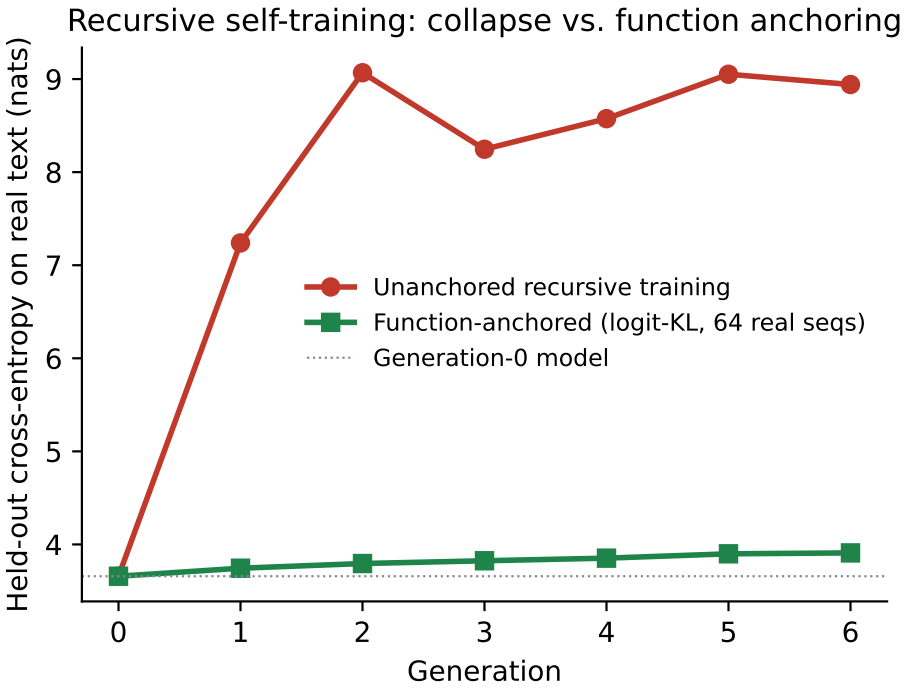


Figure 1: Function anchoring arrests recursive collapse. Held-out cross-entropy on real text across six generations of recursive self-training for the unanchored lineage (red) and the lineage trained with a logit-KL anchor to the frozen generation-0 model on 64 real sequences (green). Dotted line: generation-0 model. Single lineage per condition.

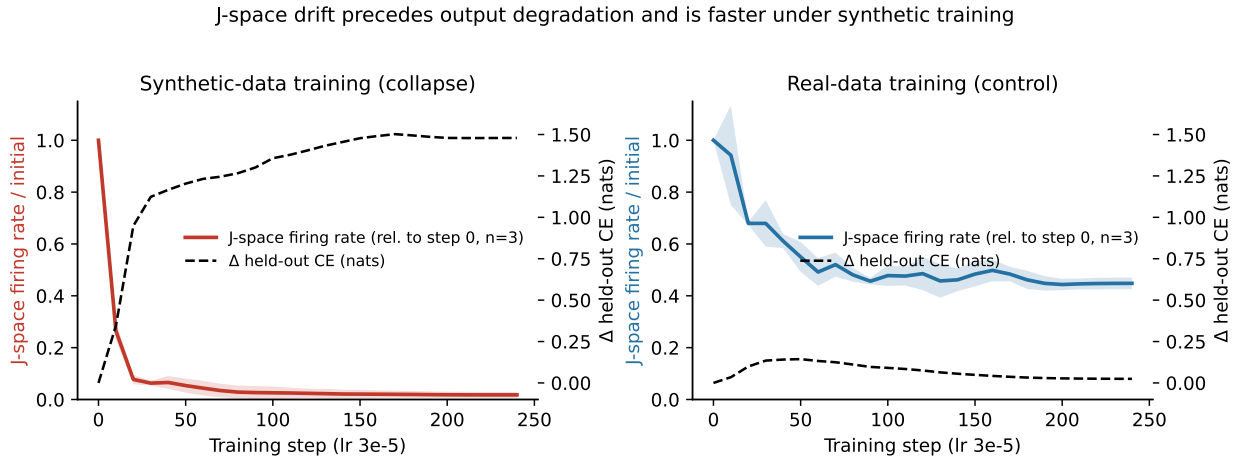


Figure 2: J-space drift precedes output degradation and is faster under synthetic training. J-space firing rate relative to step 0 (solid, left axis) and change in held-out cross-entropy (dashed, right axis) during fine-tuning of the generation-0 model on synthetic (left) or fresh real (right) text at learning rate 3×10^{-5}. Shading, where shown: $\pm$ s.d. over 3 seeds. Early movement in the real-data control ranged from -11% to $+27\%$ across seeds.

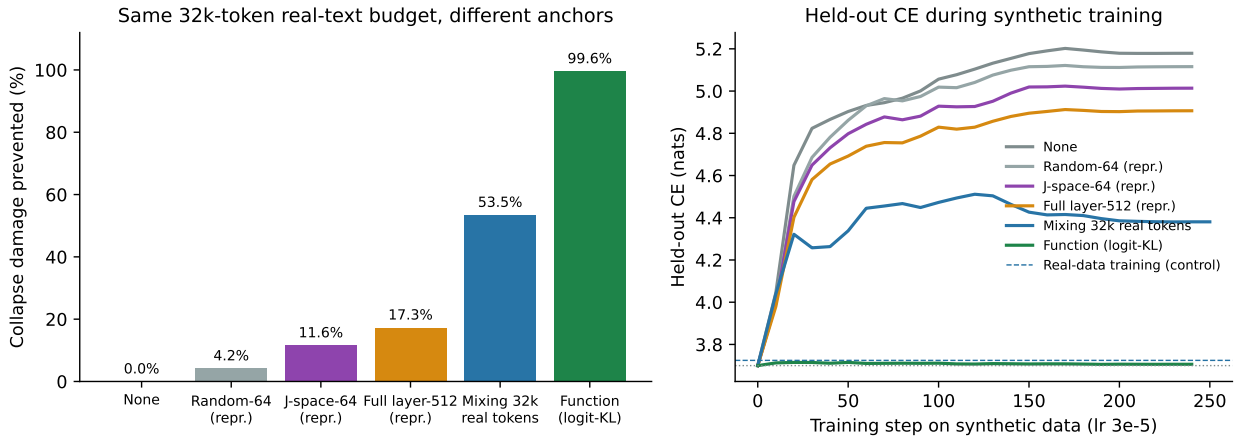


Figure 3: Treatment spectrum at matched real-data budget. Left: fraction of collapse damage prevented by each anchor (representational anchors on a random 64-dim subspace, J-space, or the full layer; mixing the same 32k real tokens into training; functional logit-KL anchor). Right: held-out cross-entropy trajectories; dashed blue: model fine-tuned on fresh real data.

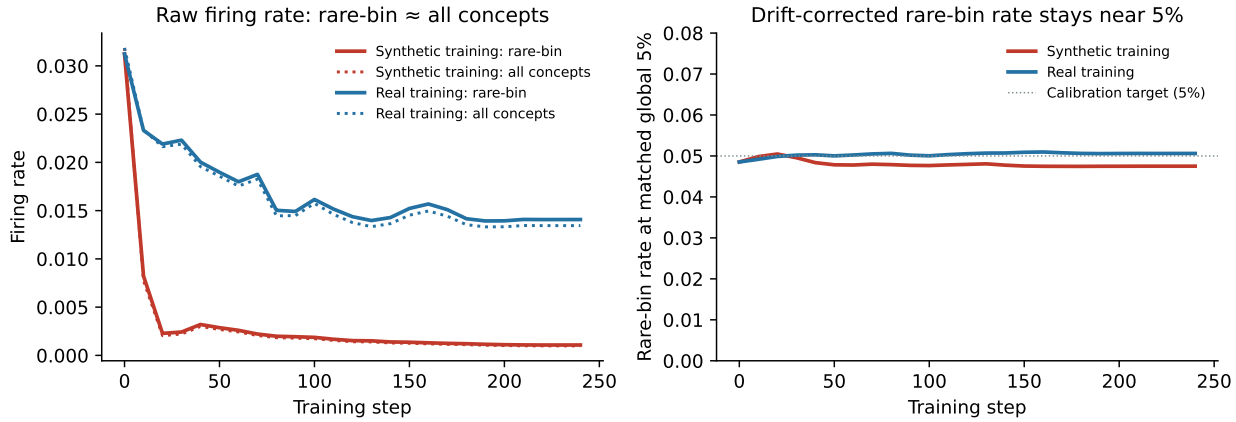


Figure 4: Collapse in J-space is dominantly uniform drift, not primarily tail loss. Left: raw firing rate for rare-bin (solid) versus all (dotted) concepts under synthetic (red) and real (blue) training. Right: rare-bin firing rate after re-calibrating a global threshold offset to a 5% overall rate; the near-flat traces indicate no selective loss of rare concepts at the resolution of binarised rates (see Section 2.3 for the continuous-score reanalysis, which resolves a small residual differential).

Table 2: Diversity of the corpus emitted by each generation's model. Single lineage per condition. Generation 0 is shared.

	Unanchored lineage		Anchored lineage	
Emitting generation	distinct-2	distinct-3	distinct-2	distinct-3
0	0.026	0.047	0.026	0.047
1	0.014	0.022	0.017	0.027
2	0.019	0.027	0.014	0.023
3	0.023	0.031	0.014	0.021
4	0.028	0.035	0.014	0.022
5	0.030	0.038	0.015	0.023

For qualitative reference, a decoded excerpt from the generation-0 model's synthetic corpus (the training data of generation 1), illustrating the repetitive-loop degeneration typical at this scale:

> *"the war in the war was in the same war . The war in the war . The war in the war was in the same war . [. . .]"*

References

[1] 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.

[2] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

[3] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. In *First Conference on Language Modeling (COLM)*, 2024. arXiv:2404.01413.

[4] Elvis Dohmatob, Yunzhen Feng, Pu Yang, François 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)*, 2024. arXiv:2402.07043.

[5] Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread; arXiv preprint arXiv:2607.15495*, 2026.

[6] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[7] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.

[8] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[10] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020.