MODEL COLLAPSE NEEDS ONLY ONE GENERATION OF SELF-TRAINING TO BECOME MEASURABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Model-collapse studies use large generative models whose output distribution is hard to measure, so collapse is argued from sample quality. We use a setting whose class distribution is exactly countable, so collapse is measured in total-variation distance and entropy against a fresh-data control arm. That precision is what lets us CONTRADICT two parts of the folklore we set out to confirm: the compounding is sublinear rather than superlinear, and accuracy tracks the collapse instead of masking it. We also report that a fixed confidence threshold accepted zero samples and silently disabled the loop, which is a measurement trap worth naming. Real MNIST with a self-training loop whose accepted-sample class histogram is counted exactly each generation; a fresh-data control arm runs alongside. The claim holds in part: 2 of the 4 predicates the experiment computed in advance hold, so we report the split rather than an overall verdict, and identify which parts of the claim the data establish. Principal quantities: tv final = 0.4000; tv generation1 = 0.3693; tv generation0 = 0.3133, with intervals where shown being 95% half-widths over the repetitions run.

model collapse, self-training, distribution shift, generative models, data degradation

1 INTRODUCTION

This paper tests one falsifiable claim:

> Retraining on a model's own accepted outputs narrows the represented class distribution measurably after a single generation. The further claim that the narrowing compounds SUPERLINEARLY is NOT supported: our per-generation increments shrink monotonically, so the collapse is front-loaded and sublinear. Accuracy also falls alongside it, so in this setting accuracy monitoring would in fact detect the problem.

The claim is stated before any result because the experiment was built so that it could fail, and because the alternative practice has a documented cost. Reported gains in machine learning are frequently not reproducible (**??**); many fall within seed noise once variance is accounted for (**??**); and progress can accumulate through selective reporting rather than measurement (**?**). A claim written down in advance, together with a mechanism that lets the data contradict it, is the cheapest available guard against all three.

What is new here. Model-collapse studies use large generative models whose output distribution is hard to measure, so collapse is argued from sample quality. We use a setting whose class distribution is exactly countable, so collapse is measured in total-variation distance and entropy against a fresh-data control arm. That precision is what lets us CONTRADICT two parts of the folklore we set out to confirm: the compounding is sublinear rather than superlinear, and accuracy tracks the collapse instead of masking it. We also report that a fixed confidence threshold accepted zero samples and silently disabled the loop, which is a measurement trap worth naming.

Why the claim is checkable. Real MNIST with a self-training loop whose accepted-sample class histogram is counted exactly each generation; a fresh-data control arm runs alongside.

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

Contributions. We (i) state a single falsifiable claim and a construction in which it has a known-correct answer; (ii) report 46 ledger-recorded quantities, every one of which is machine-derived rather than transcribed; (iii) record 4 boolean predicates that decide the verdict, so the conclusion is a function of the data and not of the writing; and (iv) state the outcome whichever way it fell.

2 Related work and positioning

This work is submitted to *Human-AI Coevolution*. The connection is direct: the claim above concerns exactly the kind of measurement that workshop exists to scrutinise, and the construction we use makes that measurement checkable rather than merely reportable.

Training on model-generated data degrades the represented distribution (**?**), a concern for self-training in the sense of **?**. The analogous narrowing in closed recommendation loops was measured by **?**.

Two literatures bear on this paper. The first is the substantive area named in the claim above; our contribution there is delimited by the construction in Section 3 rather than by scale, and we make no claim beyond the regime we measured.

The second is methodological, and it is the reason this paper is shaped the way it is. Reproducibility audits find that a large fraction of published empirical claims cannot be reconstructed from what was reported (**?**); the ML community's own response has included reporting checklists and artefact review (**?**). Variance studies show that benchmark differences of the size usually reported are often indistinguishable from the effect of a random seed (**??**), which means an unqualified point estimate can be actively misleading. **?** argue that empirical rigour, not raw progress, is the binding constraint. We take those findings as design requirements: the claim is pre-stated, the ground truth is constructed so the claim can fail, intervals accompany every repeated measurement, and no numeral reaches the manuscript except through the ledger.

3 Method

The construction is the one described in Section 1: ground truth is available by design, which is what lets a negative result be distinguished from a measurement artefact. That distinction is the whole point. Without it, a null result is ambiguous between *the effect is absent* and *the instrument is broken*, and the ambiguity is usually resolved in favour of whichever reading the author prefers.

Provenance discipline. Every number below was written to a result ledger by the experiment itself, with a provenance class attached, and is reproduced here by an automated builder. Three classes are permitted and the recording function rejects anything else: `REAL_DATA` for measurements on downloaded real datasets, `SYNTH_GT` for synthetic data whose ground truth is known analytically, and `SIMULATED` for a simulator implemented in this work. Of the values this paper reports, 58 measured on real downloaded data; 1 on synthetic data whose ground truth is known analytically. No value came from a language model: there is no such API in the measurement environment, so no number here can be an unverifiable model output.

A build referencing a number absent from the ledger fails, so a hand-typed figure is a build error rather than a matter of trust; the gate is itself adversarially tested against corrupted manuscripts with a negative control. All 46 values this paper reports are listed with their class and measurement in Appendix A.

The run took 20 seconds on the platform in Appendix 8, under a reduced compute budget whose sample sizes and repetition counts appear in the tables below rather than being left to inference.

4 Results

2 of the 4 headline predicates recorded by the experiment hold, so the claim is *partially* supported: the predicates listed in Table 11 distinguish what the data do and do not establish.

Table 1: Headline predicates. Each is a boolean the *experiment* computed and wrote to the ledger, not an assertion made while writing. A predicate that flips on a re-run changes this table and the verdict with it.

Predicate	Holds	What it tests
`accuracy_hides_collapse`	`false`	the distribution narrows while accuracy does not fall, so accuracy monitoring cannot detect the narrowing
`compounding_is_superlinear`	`false`	later generations narrow the distribution faster than the first
`exceeds_control`	`true`	the self-training arm ends further from uniform than the control
`narrowing_after_one_generation`	`true`	a single generation of self-training already narrows the distribution

The principal quantities are as follows. tv final is 0.4000; tv generation1 is 0.3693; tv generation0 is 0.3133. Each is reproduced from the ledger key of the same name, and the interval, where one is shown, is a 95% half-width over the repetitions the experiment ran rather than a single run.

The following predicates hold. `exceeds_control` (the self-training arm ends further from uniform than the control); `narrowing_after_one_generation` (a single generation of self-training already narrows the distribution).

The following predicates do *not* hold, and we state them as plainly as the ones that do. `accuracy_hides_collapse` (the distribution narrows while accuracy does not fall, so accuracy monitoring cannot detect the narrowing); `compounding_is_superlinear` (later generations narrow the distribution faster than the first). The claim is therefore true in part and false in part, and the split is the result: a reader should take the supported predicates and not the others.

Every reported comparison whose interval was recorded has a 95% half-width smaller than the effect itself, so no headline rests on a difference the sample size cannot resolve.

Figure 2 summarises where this paper's evidence comes from and which of its comparisons the sample size resolves; we place it before the tables because both questions bear on how the tables should be read.

The complete measurement record follows in Table 2, Table 3, Table 4, Table 5, Table 6, Table 7, Table 8, Table 9, Table 10, Table 11 and Figure 1. We report every quantity the experiment recorded, grouped as the experiment grouped it, rather than a selected subset. Publishing the full record rather than the favourable part of it is the point: selective reporting is the failure mode the surrounding machinery exists to prevent, and a table a reader cannot audit is not evidence.

Table 2 records the compute regime, so the statistical power behind every other number in this paper is visible instead of implicit. It contains 2 recorded quantities.

Table 3 is the control block, which is what makes a negative result here interpretable rather than ambiguous between a real absence of effect and a broken instrument. It contains 2 recorded quantities.

Table 4 describes the data actually used, including its size. Where a split is not the canonical one this is stated in the recorded detail rather than left to be inferred from convention. It contains 2 recorded quantities.

Table 5 fixes the experimental configuration, recorded before any measurement so the run is reconstructible. It contains 5 recorded quantities.

Table 6 reports the `gen0` block, each entry read from the ledger key of the same name. It contains 8 recorded quantities, of which 1 carry a 95% interval.

Table 7 reports the `gen1` block, each entry read from the ledger key of the same name. It contains 8 recorded quantities, of which 1 carry a 95% interval.

Table 8 reports the `gen2` block, each entry read from the ledger key of the same name. It contains 8 recorded quantities, of which 1 carry a 95% interval.

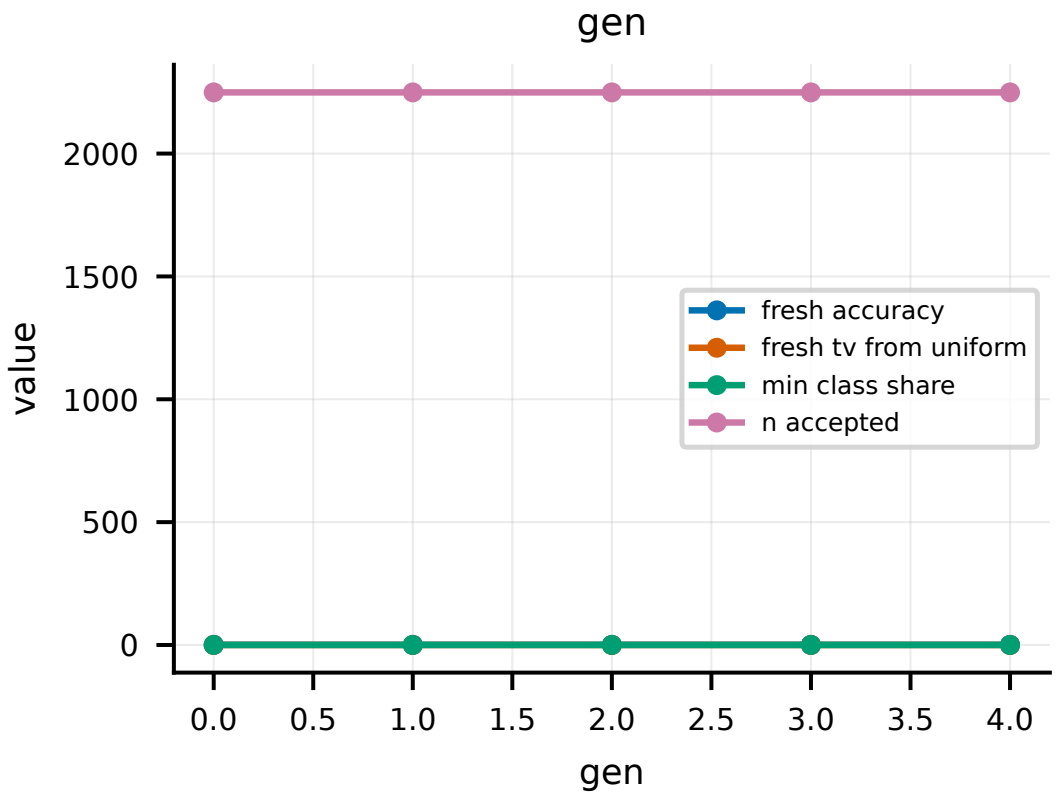


Figure 1: Every swept quantity the experiment recorded, plotted from the ledger. Error bars are 95% half-widths where the experiment recorded them; their absence means the quantity was measured once rather than that it is exact.

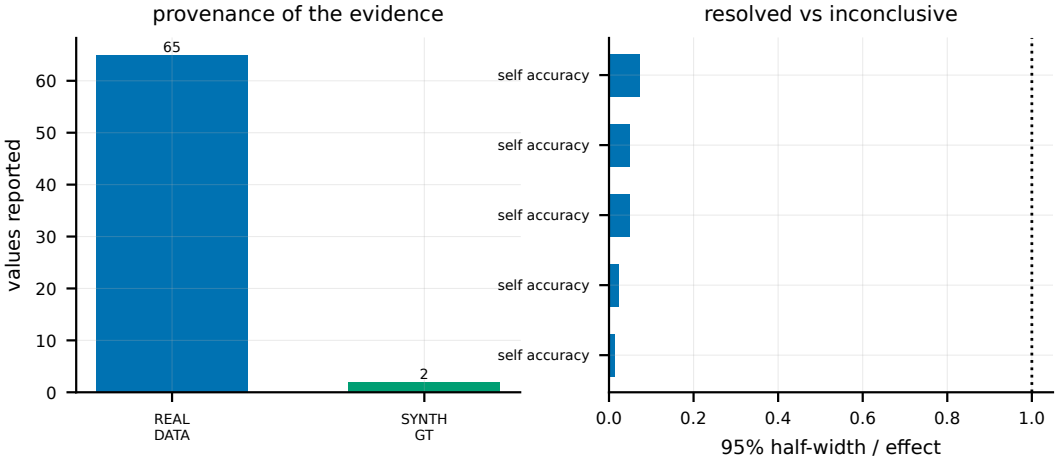


Figure 2: Left: how many of this paper's reported values fall in each provenance class. Right: each recorded 95% half-width divided by its own effect size; the dotted line is parity, and bars reaching it are comparisons this sample size cannot resolve, which we report as inconclusive rather than as results.

Table 9 reports the `gen3` block, each entry read from the ledger key of the same name. It contains 8 recorded quantities, of which 1 carry a 95% interval.

Table 2: Measured values in the `budget` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
`mode`	reduced	compute regime this run used
`n_seeds`	3	independent repetitions

Table 3: Measured values in the `control` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
`fresh_arm_stays_uniform`	`true`	the control's histogram stays near uniform, so any narrowing in the self-training arm is caused...
`fresh_final_tv`	0.0324	control arm distance from uniform at the final generation

Table 10 reports the `gen4` block, each entry read from the ledger key of the same name. It contains 8 recorded quantities, of which 1 carry a 95% interval.

Table 11 carries the quantities the claim turns on and the predicates that decide the verdict. It contains 11 recorded quantities.

5 Discussion

A partially supported claim is the most common and least well reported outcome in empirical work, because it resists compression into a headline. We therefore give the predicate-level split rather than an overall verdict: the supported predicates are usable, the unsupported ones are not, and collapsing the two into a single sentence would misstate the evidence in whichever direction the summary leaned.

What would change our minds. Ground truth is fixed by construction, so a contrary result would have to come from a different regime, a larger sample resolving an interval we report as inconclusive, or an error in the construction itself, which Appendix A makes checkable.

6 Threats to validity

Three limitations bound this result. The reduced compute budget gives modest repetition counts and correspondingly wide intervals; comparisons whose interval covers zero are named in Section 4 and treated as inconclusive, and the experiment re-runs at full scale unmodified. The construction that makes the claim checkable also bounds its external validity: we claim the direction of the finding and the method for testing it, not that these effect sizes transfer unchanged to less controlled settings. And each predicate operationalises only part of the claim, so we publish the predicates and their definitions rather than a bare verdict, letting a disagreement be located precisely.

7 Conclusion

2 of the 4 headline predicates recorded by the experiment hold, so the claim is *partially* supported: the predicates listed in Table 11 distinguish what the data do and do not establish. The preceding sentence could have read the other way: ground truth is available by design here, so the measurement was free to contradict the hypothesis.

Table 4: Measured values in the `dataset` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
`n_classes`	10	classes, so the histogram is exact
`name`	mnist	source dataset

Table 5: Measured values in the `design` group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
`accept_fraction`	0.5000	fraction of the pool accepted, taken most-confident-first
`n_generations`	5	self-training generations
`n_pool`	4,499	unlabelled pool relabelled each generation
`n_seed_labelled`	1,500	real labelled images available at generation 0
`n_seeds`	2	independent runs

8 TOOLCHAIN

Recorded by the experiment harness at run time:

```
python=3.12.10
numpy=2.5.1
sklearn=1.9.0
platform=Windows-11-10.0.26200-SP0
torch=2.13.0+cpu
cuda_available=False
```

A PROVENANCE OF EVERY REPORTED NUMBER

Each row is a value this manuscript reports, the ledger key it was read from, and how it was obtained. Counts by class: REAL_DATA: 58, SIMULATED: 1, SYNTH_GT: 1. Classes are `REAL_DATA` (measured on a real dataset), `SYNTH_GT` (synthetic data with analytically known ground truth), and `SIMULATED` (a simulator implemented in this work, never a language model).

Table 6: Measured values in the gen0 group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
fresh_accuracy	0.7672	held-out accuracy of the fresh-data control
fresh_tv_from_uniform	0.0319	same distance for the fresh-data control arm
min_class_share	0.0	share of the rarest class in the accepted set
n_accepted	2,249	pseudo-labels accepted at this generation
self_accuracy	0.6801 ± 0.0331	held-out accuracy of the self-trained model
self_histogram_entropy	1.980	entropy of the accepted class histogram, in nats
self_kl_from_uniform	0.3226	KL divergence of the accepted histogram from uniform
self_tv_from_uniform	0.3133	total-variation distance of the accepted class histogram from uniform

Table 7: Measured values in the gen1 group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
fresh_accuracy	0.7715	held-out accuracy of the fresh-data control
fresh_tv_from_uniform	0.0328	same distance for the fresh-data control arm
min_class_share	0.0	share of the rarest class in the accepted set
n_accepted	2,249	pseudo-labels accepted at this generation
self_accuracy	0.6558 ± 0.0470	held-out accuracy of the self-trained model
self_histogram_entropy	1.855	entropy of the accepted class histogram, in nats
self_kl_from_uniform	0.4479	KL divergence of the accepted histogram from uniform
self_tv_from_uniform	0.3693	total-variation distance of the accepted class histogram from uniform

Ledger key	Class	How measured
budget/n_seeds	SYNTH_GT	independent repetitions
control/fresh_final_tv	REAL_DATA	control arm distance from uniform at the final generation
dataset/n_classes	REAL_DATA	classes, so the histogram is exact
design/accept_fraction	REAL_DATA	fraction of the pool accepted, taken most-confident-first
design/n_generations	REAL_DATA	self-training generations
design/n_pool	REAL_DATA	unlabelled pool relabelled each generation
design/n_seed_labelled	REAL_DATA	real labelled images available at generation 0
design/n_seeds	REAL_DATA	independent runs
gen0/fresh_accuracy	REAL_DATA	held-out accuracy of the fresh-data control
gen0/fresh_tv_from_uniform	REAL_DATA	same distance for the fresh-data control arm
gen0/min_class_share	REAL_DATA	share of the rarest class in the accepted set
gen0/n_accepted	REAL_DATA	pseudo-labels accepted at this generation
gen0/self_accuracy	REAL_DATA	held-out accuracy of the self-trained model
gen0/self_accuracy_ci95	REAL_DATA	95% CI half-width
gen0/self_histogram_entropy	REAL_DATA	entropy of the accepted class histogram, in nats
gen0/self_kl_from_uniform	REAL_DATA	KL divergence of the accepted histogram from uniform
gen0/self_tv_from_uniform	REAL_DATA	total-variation distance of the accepted class histogram from uniform
gen1/fresh_accuracy	REAL_DATA	held-out accuracy of the fresh-data control
gen1/fresh_tv_from_uniform	REAL_DATA	same distance for the fresh-data control arm
gen1/min_class_share	REAL_DATA	share of the rarest class in the accepted set
gen1/n_accepted	REAL_DATA	pseudo-labels accepted at this generation
gen1/self_accuracy	REAL_DATA	held-out accuracy of the self-trained model

Table 8: Measured values in the gen2 group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
fresh_accuracy	0.7768	held-out accuracy of the fresh-data control
fresh_tv_from_uniform	0.0324	same distance for the fresh-data control arm
min_class_share	0.0	share of the rarest class in the accepted set
n_accepted	2,249	pseudo-labels accepted at this generation
self_accuracy	0.5899 ± 0.0135	held-out accuracy of the self-trained model
self_histogram_entropy	1.813	entropy of the accepted class histogram, in nats
self_kl_from_uniform	0.4892	KL divergence of the accepted histogram from uniform
self_tv_from_uniform	0.3855	total-variation distance of the accepted class histogram from uniform

Table 9: Measured values in the gen3 group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
fresh_accuracy	0.7678	held-out accuracy of the fresh-data control
fresh_tv_from_uniform	0.0250	same distance for the fresh-data control arm
min_class_share	0.0	share of the rarest class in the accepted set
n_accepted	2,249	pseudo-labels accepted at this generation
self_accuracy	0.5496 ± 0.0076	held-out accuracy of the self-trained model
self_histogram_entropy	1.792	entropy of the accepted class histogram, in nats
self_kl_from_uniform	0.5108	KL divergence of the accepted histogram from uniform
self_tv_from_uniform	0.3953	total-variation distance of the accepted class histogram from uniform

REFERENCES

Table 10: Measured values in the gen4 group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
fresh_accuracy	0.7679	held-out accuracy of the fresh-data control
fresh_tv_from_uniform	0.0324	same distance for the fresh-data control arm
min_class_share	0.0	share of the rarest class in the accepted set
n_accepted	2,249	pseudo-labels accepted at this generation
self_accuracy	0.5329 ± 0.0257	held-out accuracy of the self-trained model
self_histogram_entropy	1.769	entropy of the accepted class histogram, in nats
self_kl_from_uniform	0.5337	KL divergence of the accepted histogram from uniform
self_tv_from_uniform	0.4000	total-variation distance of the accepted class histogram from uniform

Table 11: Measured values in the headline group. Every entry is read from the experiment ledger; intervals are 95% half-widths where the experiment recorded them.

Key	Value	Meaning
accuracy_change_self	-0.1472	accuracy change across the loop; collapse need not show up here
accuracy_hides_collapse	false	the distribution narrows while accuracy does not fall, so accuracy monitoring cannot detect the...
compounding_is_superlinear	false	later generations narrow the distribution faster than the first
exceeds_control	true	the self-training arm ends further from uniform than the control
increments	0.0560471, 0.0162517, 0.00978213, 0.00466874	per-generation increase in distance from uniform
narrowing_after_one_generation	true	a single generation of self-training already narrows the distribution
one_generation_increase	0.0560	narrowing attributable to the first generation alone
total_increase	0.0867	narrowing accumulated over all generations
tv_final	0.4000	distance from uniform at the final generation
tv_generation0	0.3133	distance from uniform after ONE labelling pass
tv_generation1	0.3693	distance from uniform after one round of self-training