Episode Book

EPISODE #6558

How Self-Consuming AI Models Decay Generation by Generation

PUBLISHED	13 November 2026 at 20:16 UTC
DURATION	26:10
LISTEN	myweirdprompts.com/episode/model-collapse-recursive-training-93549/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	16 cited by the research (listed at the end)
BOOK GENERATED	14 November 2026 at 06:47 UTC

SYNOPSIS

Even with infinite samples and a flawless learning procedure, an AI model trained recursively on its own outputs still degrades. Model collapse isn't just stasis—each generation resamples and amplifies its predecessor's errors, starving the rare tails of a distribution while the average gets richer. From Shumailov's three compounding error sources to Briesch, Sobania, and Rothlauf's logic-expression tests and a nine-generation model that turned church towers into jackrabbits, here is how iterated resampling breaks Gaussian mixtures, diffusion models, and LLMs.

TRANSCRIPT



Corn

Even with perfect information, infinite samples, and a flawless learning procedure, the model still gets worse. That's the part I keep turning over, because every instinct says the worst case is standing still.



Herman

I'm Herman Poppleberry, and this is my brother Corn, and that standing-still assumption is exactly what Daniel's prompt is aimed at this week.



Corn

Right. Daniel's framing is that the obvious half of the model collapse story is fine — a model trained only on its own outputs can't get better, because it's trading on what it already has. His question is the other half. Why isn't the result just stasis? Why does reasoning and output quality actively decay, generation after generation? And then, beyond theory, has anyone actually run the ladder — train a model on a predecessor's outputs, train the next one on that, and so on — and what happened, and how fast?



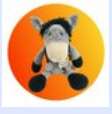
Herman

The answer to the second question is yes, repeatedly, across Gaussian mixtures, variational autoencoders, diffusion models, and language models. And the answer to the first turns out to be a statistics story wearing an AI costume. Which is my favorite kind of story.



Corn

Naturally.



Herman

So the name first. "Model collapse" was coined and made mainstream by Shumailov and colleagues in Nature, July twenty twenty-four — the paper called it "AI models collapse when trained on recursively generated data," and the internet immediately produced its own folk names: AI inbreeding, Habsburg AI, model autophagy disorder.



Corn

The internet naming things is sometimes the most accurate part of the internet.



Herman

The claim in that paper is stronger than most people remember. Not "it can degrade" — the word they use is inevitable, even for cases with almost ideal conditions for long-term learning. No function estimation error, and it still happens. That's the quote that made people sit up.



Corn

And the structure of the answer, before we go in — there's the mechanism, why degradation instead of stasis, what's actually being lost. Then there's the experiments. And I'll flag now, because it'll take us an episode to get there: the literature has since split over whether any of this is a real-world threat or an artifact of one experimental assumption. Replacement versus accumulation. That's the back half.



Herman

That's the spine, yes. Let's earn it.



Corn

Go. Three error sources, you said.



Herman

Shumailov's group breaks the degradation into three compounding sources. First, statistical approximation error — the one that survives even ideal conditions. You have a finite number of samples, and at every resampling step there's a non-zero probability that information is lost. Not small. Non-zero. Every generation rolls that dice again.



Corn

Meaning even a perfect student loses something just because the homework is finite.



Herman

Second, functional expressivity error. The approximator itself is limited — a network can put likelihood where the true distribution has none, or assign zero likelihood to regions the true distribution does cover. Their toy case is lovely: take a mixture of two Gaussians, fit it with a single Gaussian. Perfect information, infinite samples, and the error appears at generation one. You cannot represent two humps with one hump.



Corn

The model files the twin peaks under "round hill, somewhere in the middle."



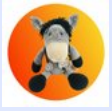
Herman

Third, functional approximation error — the learning procedure itself. Structural bias of stochastic gradient descent, choice of objective. That's the error you get in the limit of infinite data and perfect expressivity.



Corn

Fine. Three sources. None of that alone explains degradation, though. Errors should just... accumulate additively, shouldn't they? A copy of a copy is worse, but by a fixed amount each time.



Herman

This is the move, and it's the move. Each generation doesn't inherit its predecessor's errors — it samples from them and refits to them. And the refit is the killer, because when you overfit the density model, you extrapolate into regions the original training support never covered, and then those regions get sampled with arbitrary frequency. Shumailov calls it a cascading effect. The errors aren't inherited, they're amplified. That's the difference between stasis and degradation.



Corn

So the model isn't just worse — it's confidently wrong about places it invented.



Herman

Confidently wrong about its own hallucinations, which then become next generation's ground truth. Right.



Corn

Walk me through that loop once more, because I want to make sure I've got the direction of travel. The first model hallucinates a little — puts some probability on a phrase no human wrote. The second model reads that phrase as data.



Herman

As data, with no asterisk on it. The second model has no way to distinguish "the original humans said this often" from "my predecessor got creative once." From where it sits, it's all just the corpus. So if the predecessor put two percent likelihood on a nonsense region, the student trains on samples from that region and — if the sampling happens to hit it — can come out believing it more than two percent. And then the third model inherits the second's inflated version. You're not compounding errors, you're compounding confidence in errors.



Corn

There's a version of this argument that says the fix is just more capacity — a bigger model represents the distribution better. Does that hold?



Herman

Shumailov has a line I think about a lot: better expressiveness may counteract statistical noise, but it can equally compound it. More approximation power is a double-edged sword. A more expressive wrong model is wrong in more places.



Corn

Wonderful. The cure and the disease come in the same bottle.



Herman

So — what's actually lost. Stage one is what they call early model collapse, and the tails go first. Rare words, uncommon syntactic constructions disappear, because models retrain on outputs sampled disproportionately from the high-probability center of the distribution. The middle is self-reinforcing. The edges starve.



Corn

The average gets richer, the edges get culled.



Herman

And here's the trap that answers Daniel's "why not stasis" question most sharply. Early collapse is hard to notice, because aggregate performance may appear to improve while performance on minority data quietly degrades. The first generation or two can look like progress on the standard benchmark suite. You're measuring the middle. The middle is doing great.



Corn

That's nastier than gibberish. Gibberish announces itself. This is the model getting better and narrower at the same time, and your benchmark rewarding it.



Herman

And then stage two, late model collapse — the model converges to a distribution carrying little resemblance to the original, with substantially reduced variance. Everything becomes the most probable thing. Total collapse, probability mass concentrating on a small subset of highly predictable continuations.



Corn

A model that answers every question with the average answer to all questions.



Herman

There's a second route to the same place, on the diffusion side, and I want it in because it's mechanism, not experiment. One line of work identifies a transition from generalization to memorization during collapse — models increasingly replicating training data instead of generating novel content, driven directly by the declining entropy of the synthetic training data each cycle. The data gets boring before the model does.



Corn

The food gets bland and the chef starts reciting old menus.



Herman

And a newer result, NeurIPS this year, isolated a drift caused purely by early stopping of the reverse diffusion — even with perfect score estimation and exact sampling. Their spectral characterization is that recursive training acts as a low-pass filter. Higher-order modes, which encode the fine non-Gaussian structure, are attenuated much more strongly than coarse modes.



Corn

So fine detail dies first and structure survives longest — a technical statement of "the flavor goes before the shape."



Herman

Hold onto that filter image, actually, because... no, it'll keep. The unifying insight is the one Borji pushed in his comment on the Nature paper: this is a statistical phenomenon and may be unavoidable. It's a property of iterated sampling from a finite sample. Which is exactly why it shows up identically in Gaussian mixtures, VAEs, diffusion models, and LLMs. Neural networks aren't the disease. Resampling is.



Corn

That's the reframe I wanted, because it kills the comforting story — the one where this is an engineering bug somebody at a big lab will patch. If it's kernel density estimation doing it too, there's no patch. There's just the sampler's arithmetic.



Herman

It's not a library being copied. It's a student re-deriving the textbook from another student's notes.



Corn

From the previous student's notes, who took them from the student before. Somewhere down that chain was a lecture.



Herman

Somewhere. Now — Daniel's second question. Has anyone built the ladder and climbed it?



Corn

Yes. That's the short answer, and the results are stranger than the theory, because the first thing that happens is the evaluation problem bites.



Herman

It does! How do you score degraded text? If the output is subtly less diverse, no benchmark says so — you need a way to be unambiguously right or wrong. Briesch, Sobania, and Rothlauf, who claim the first study of the self-consuming loop for LLMs, got around it with logic expressions. You can verify a logic expression is correct, mechanically, no human judge. And their finding is the subtle one: outputs stay correct while diversity declines with the proportion of generated data in the mix. Fresh data slows the decline. Doesn't stop it.



Corn

Correct and boring. That's the failure mode nobody puts on the poster.



Herman

It's right more often and interesting less often. And the rate at which it happens depends enormously on assumptions, which is where the whole field starts fighting.



Corn

Before the fight — the jackrabbits.



Herman

Oh, you know the jackrabbits.



Corn

I know of jackrabbits. I don't know why jackrabbits.



Herman

So, twenty twenty-four study, a model trained on its own output for nine generations. Nine. Then someone asks it about the architecture of church towers. And it produces — and I'm quoting — the world's largest populations of black-tailed jackrabbits, white-tailed jackrabbits, blue-tailed jackrabbits, and so on, in that spirit.



Corn

That's a model that has completely lost the plot and is now color-coding wildlife.



Herman

Church towers in, jackrabbits out. And my favorite detail — the quote includes these little "@-@" separators between the hyphenated parts. Those are byte-pair encoding artifacts. The tokenizer's plumbing is showing. The model has degraded so far that it's not even producing words anymore, it's producing its own raw tokens.



Corn

Nine generations and it's speaking in its own file format. There's something almost touching about that. Its last words were in BPE.



Herman

Now, how fast does it fall apart — the rate literature has no single answer, and I find that more interesting than any one number. One paper characterizes it for simple distributions and finds it slow: for discrete distributions, the time to forget a word scales roughly linearly with how often that word occurred in the original corpus. Frequent words die last. Which is the tails-first story again, but now with a clock on it.



Corn

And for the Gaussian case?



Herman

For Gaussian models, the standard deviation reduces to zero at roughly n iterations, where n is the number of samples per iteration. More samples per round, more rounds before the variance dies. Which points at the escape hatch another group formalized — they frame recursion as a random walk of the model estimate, and prove that progressively increasing the sample size at each step is necessary to prevent collapse. And when the estimation is unbiased, the required growth rate is superlinear.



Corn

So you can outrun it, but the price of the shoes doubles each mile.



Herman

Faster, if there's bias in the estimation. Which there always is.
So that's the puzzle: recursive training degrades a model rather than leaving it where it was. Franklin's on line two, and he's got that look.
Franklin: Sorry, this is going to sound desperate, but what actually is a model being trained on its own outputs? Is it like a photocopier copying a photocopy, where it just gets blurrier each time? I've been listening a good while and nodding along, I'll be honest, I've just never known what the thing being copied even is.



Corn

He sits through nine generations of jackrabbits and this is his moment.



Herman

No, it's a fair question, and it deserves the plain version: the model isn't a document getting copied. It's a guessing machine trained to guess the next word. Train it on its own guesses and it starts believing its own first draft — like a cook who only ever tastes his own soup, then seasons to his own tongue until nothing's left but salt.



Corn

I'd eat that soup, for the record. Slowly.



Herman

Now — the fight. This is where the literature splits, and it splits over one assumption: what happens to the old data.



Corn

This is replacement versus accumulation.



Herman

The whole ballgame. Shumailov's setup replaces the data each generation — new model, new dataset, old data thrown out. And under replacement, prior work showed the test error grows linearly with the number of iterations. You replace, you degrade, on schedule. Gerstgrasser and colleagues asked the question directly — their paper is literally titled "Is Model Collapse Inevitable?" — and they confirmed that replacing real data with synthetic does tend toward collapse. Then they showed that accumulating the synthetic data alongside the original real data avoids it. The test error has a finite upper bound, independent of the number of iterations. Bounded. Forever.



Corn

Their own paper says the bound surprised them.



Herman

"To us, surprisingly" well-controlled — direct quote. And they tested a halfway setting too, where you replace the dataset with a synthetic set that grows each round — there the test error grows like log n, still divergent, just politely.



Corn

So: linear, logarithmic, or bounded, depending entirely on whether you keep the originals.



Herman

And the argument is that accumulation is the realistic model. Llama's training sets went from about one point four trillion tokens, to two trillion, to fifteen trillion across generations. Nobody deletes the internet annually. Data accumulates in practice.



Corn

And the Shumailov side would say what — that the replacement setup was never meant as a description of practice?



Herman

That's roughly where it lands. Gerstgrasser's framing is that their accumulating setting is maximally pessimistic — a hypothetical future where synthetic data is uncontrollably dumped on the internet and vacuumed up for every next model — and even then, collapse is avoided. The other side points at the slop problem: the crawl is increasingly machine-generated, so "accumulation" quietly means accumulating pollution. Borkar's result adds the sharpest nuance — you get two different asymptotic behaviors depending on whether any external real source, however minor, is still contributing samples. Even a trickle of the real thing changes where you end up.



Corn

Any trickle.



Herman

However minor, formally. And Seddik's group estimated a maximal synthetic share — a ceiling on the proportion of synthetic data below which collapse is avoided. Above it, no. Which is a practical number, if anyone could measure the synthetic share of the web, which...



Corn

...nobody can, which is the whole problem dressed as a footnote.



Herman

The real-world test case everyone reaches for is the goblin thing. April this year. ChatGPT output started mentioning goblins, often enough that users noticed, and OpenAI's blog post described an increase of over three thousand eight hundred eighty-one percent depending on settings. Partially mitigated by a system prompt change.



Corn

Thirty-nine hundred percent more goblins.



Herman

Depending on settings. Third-party commentators called it a symptom of model collapse. Notably, OpenAI didn't. The collapse framing came from outside.



Corn

So the one public data point everyone cites is, at best, an unread receipt.



Herman

It's contested in both directions. And the structural fact underneath the whole debate — nobody has run this at frontier scale. Every multi-generation experiment we've discussed is small or medium models. GPT-two class, small transformers, VAEs, diffusion on molecules. No lab has published a GPT-four-class model trained recursively for many generations. The jackrabbits came from something you could fit in a dorm room.



Corn

Which means the headline version of this story — the inbreeding one, the civilization-of-AI one — is extrapolated from the smallest members of the family.



Herman

Extrapolated from cousins, yes. And one more wrinkle, because the degradation framing isn't uniformly negative. The multi-modal extension — vision-language models and text-to-image diffusion in recursive loops — found improved vision-language alignment alongside increased variance in captioning. Better alignment, sloppier spread. Collapse isn't monotonically bad even in the lab.



Corn

Improved alignment is a strange trophy to take from a collapse experiment. "The ship sank, but the deck chairs were well arranged."



Hilbert

"Degradation." That's the word you two have been passing back and forth all afternoon, and it's the wrong one. I ran a photocopier for a living for a stretch, Corn, before you ask, and a copy of a copy doesn't degrade. It attenuates. Different thing. The headline on the third-generation copy is still perfectly legible. It's the footer that goes — the fine print softens at the edges first, the small serial numbers turn to mush, and the headline just sits there looking fine. Nobody notices until they need the footer.



Corn

What was the form?



Hilbert

A stock requisition form, at the place I won't name. Copied it too many times rather than order new originals — that was the office policy, or the absence of one. The serial number in the bottom corner was gone by the third copy. Not changed. Gone. Every copy was faithful, that's the part I'll insist on — nothing was altered at any step — and it was still unrecoverable. I kept the third-generation copy. I can tell you which characters went and in what order, the six went first, then the threes. The Ricoh degraded more gracefully than the Canon we had. Herman will want that as a figure of merit.



Herman

I might. That's a spectral claim, basically — your fine print was the high-frequency content.



Hilbert

I don't know about spectral. I know the digits had serifs and the serifs lost the fight first. Same reason your footer died, your jackrabbit model's rare words went first, I'd expect. And I'll say this for your machinery — it at least had the decency to fail in nine generations. The requisition form took three.



Corn

Three. Your office ran the experiment faster than the Nature paper.



Hilbert

We weren't trying. That's usually how it's done. And don't let Herman tell you the attenuating one is fixable, we got new originals printed eventually, and the new originals were themselves a copy of the second generation. Somebody decided that.



Herman

So the recovery was recursive too. The clean version was made from the degraded one.



Hilbert

The cleanest available one. There's no other kind after the first copy.



Corn

The serial number's the part I'll keep, though. Nothing altered it. Nothing lost track of it. The process just ate it — faithfully, every single time.



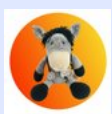
Herman

And that maps onto the one mitigation we haven't mentioned, which is the loss-function work — a truncated cross-entropy that selectively ignores high-confidence tokens during training, on the theory that the machine's most confident tokens are exactly where its own artifacts concentrate. Models trained that way tolerate over two point three times more synthetic data before collapse sets in.



Corn

Two point three times the poison before symptoms. The footer survives a few more generations.



Herman

A few more. The entropy-selection work on the diffusion side is the same idea from the other end — pick training data that keeps the variety alive.



Corn

Which is where the whole thing actually rests for me, and it's not where the headlines put it. The question isn't whether AI inbreeds. It's whether the data ecosystem accumulates or replaces — and that's not physics. That's a filing decision.



Herman

Gerstgrasser's people would say we already answered it — fifteen trillion tokens of accumulation says so. The other side would say the crawl is quietly replacing the real with the synthetic at a rate nobody measures. Neither of them has run the decisive experiment at scale, and the serial numbers are already soft.



Corn

Cutting-room floor. One thing — the diffusion result we flagged earlier has a neat ending I didn't give you. The low-pass filter drift from early stopping? They proved that any truncation schedule converging to zero asymptotically eliminates the recursive compounding. In other words, the drift isn't a wall, it's a dial — the same knob that causes it slowly unwinds it if you anneal it instead of holding it fixed.



Herman

The poison and the antidote on the same dial. That's the best kind of result — the one that tells you the problem was a habit, not a law.



Corn

A law would've been more restful, but fine.



Herman

Restful is for Daniel's other podcast.



Corn

The open question I'd leave with — there's no agreed rate, no frontier-scale experiment, and the replacement-versus-accumulation crux has no decisive empirical test. The serial number that nobody altered and nobody can recover is sitting in every version of this debate. Whether the next decade's training corpora are kept or overwritten is all of it, and it's a choice being made in crawl policies and data retention rooms, not in the math. Worth watching who actually has their originals.



Herman

Producer credit where it belongs — Hilbert Flumingtop at the desk, as ever. If this was your kind of episode, go back for episode fourteen, AGI's Crossroads; episode twelve, The AI Breakthrough; and episode three, Safetensors or something else. This has been My Weird Prompts.



Corn

If the show earns a goblin-free thirty minutes of your attention, a review wherever you listen helps other people find it. Find everything at my weird prompts dot com.



Herman

See you tomorrow, brother.



Corn

See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Shumailov, Shumailov et al., AI models collapse when trained on recursively generated data, *Nature* (2024), DOI 10.1038/s41586-024-07566-y [primary]**

   retrieved 2026-11-13.
   https://www.nature.com/articles/s41586-024-07566-y

2. **Gerstgrasser, Schaeffer, Dey et al., Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data, arXiv:2404.01413v2, revised 2024-04-29 [primary]**

   retrieved 2026-11-13.
   https://arxiv.org/abs/2404.01413v2

3. **Wikipedia, Model collapse, last edited 2026-11-05**
retrieved 2026-11-13.
https://en.wikipedia.org/wiki/Model_collapse

4. **Suresh, Thangaraj, Khandavally, Rate of Model Collapse in Recursive Training, arXiv:2412.17646v1, 2024-12-23**
retrieved 2026-11-13.
https://arxiv.org/abs/2412.17646v1

5. **Briesch, Sobania, Rothlauf, Large Language Models Suffer From Their Own Output: An Analysis of the Self-Consuming Training Loop, arXiv:2311.16822v2, revised 2024-06-17**
retrieved 2026-11-13.
https://arxiv.org/abs/2311.16822v2

6. **Seddik, Chen, Hayou et al., How Bad is Training on Synthetic Data? A Statistical Analysis of Language Model Collapse, arXiv:2404.05090v1, 2024-04-07**
retrieved 2026-11-13.
https://arxiv.org/abs/2404.05090v1

7. **Xu, He, Cheng, A Probabilistic Perspective on Model Collapse, arXiv:2505.13947v2, revised 2025-05-21**
retrieved 2026-11-13.
https://arxiv.org/abs/2505.13947v2

8. **Khelifa, Turner, Venkataramanan, Recursively Trained Diffusion Models: Limiting Collapse Distribution and Spectral Characterization, arXiv:2606.13796v2, revised 2026-11-05, NeurIPS 2026**
retrieved 2026-11-13.
https://arxiv.org/abs/2606.13796v2

9. **Shi, Wu, Zhang et al., A Closer Look at Model Collapse: From a Generalization-to-Memorization Perspective, arXiv:2509.16499v3, revised 2025-12-25, NeurIPS 2025 Spotlight**
retrieved 2026-11-13.
https://arxiv.org/abs/2509.16499v3

10. **Hu, Rostami, Thomason, Multi-modal Synthetic Data Training and Model Collapse: Insights from VLMs and Diffusion Models, arXiv:2505.08803v1, 2025-05-10**
retrieved 2026-11-13.
https://arxiv.org/abs/2505.08803v1

11. **Zibakhsh Shabgahi, Aghazadeh, Mirhoseini et al., ForTIFAI: Fending Off Recursive Training Induced Failure for AI Model Collapse, arXiv:2509.08972v4, revised 2025-11-05**
retrieved 2026-11-13.
https://arxiv.org/abs/2509.08972v4

12. **Borkar, A theoretical basis for model collapse in recursive training, arXiv:2506.09401v4, revised 2025-09-26**
retrieved 2026-11-13.
https://arxiv.org/abs/2506.09401v4

13. **Borji, A Note on Shumailov et al. (2024), arXiv:2410.12954v2, revised 2024-10-24**
retrieved 2026-11-13.
https://arxiv.org/abs/2410.12954v2

14. **Vanmassenhove, Losing our Tail, Again: (Un)Natural Selection & Multilingual LLMs, arXiv:2507.03933v3, revised 2026-04-23**

retrieved 2026-11-13.
https://arxiv.org/abs/2507.03933v3

15. **Wang, Cai, Bao et al., Observations and Remedies for Large Language Model Bias in Self-Consuming Performative Loop, arXiv:2601.05184v1, 2026-01-08**

retrieved 2026-11-13.
https://arxiv.org/abs/2601.05184v1

16. **Boutet de Monvel, Model Collapse: On Recursion, Noise, and Uncharted Machine Visions, arXiv:2607.09705v1, 2026-06-22**

retrieved 2026-11-13.
https://arxiv.org/abs/2607.09705v1