Episode Book

EPISODE #6241

Why Your Synthetic Data Isn't as Diverse as You Think

PUBLISHED	24 October 2026 at 11:31 UTC
DURATION	21:44
LISTEN	myweirdprompts.com/episode/synthetic-data-diversity-collapse-88204/
HOSTS	Corn, Daniel, Herman, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular+meorav (chatterbox-local)
PIPELINE	V5.20
SOURCES	15 cited by the research (listed at the end)
BOOK GENERATED	25 October 2026 at 06:19 UTC

SYNOPSIS

Synthetic data generation has a hidden problem: each example looks fine, but together they teach your model a narrow slice of reality. This episode digs into cross-batch mode collapse, why temperature and personas underdeliver, and what actually expands coverage — from scenario planning to multi-stage generation pipelines. If you're generating training data at scale, the bottleneck isn't quality per example. It's the variety of situations your dataset represents.

TRANSCRIPT



Corn

Fifty-nine percent of the stories in one GPT-four-generated dataset began with Once upon a time. Daniel, that's a remarkably expensive way to discover one sentence.



Daniel

That's the problem I want to get into. If I ask for ten thousand synthetic examples, I don't want the same ideas wearing different names. How do I get actual variety, and how do I check it before paying for the whole batch?



Herman

Correct examples can still make a terrible training set. Each one passes inspection, the labels look sensible, and together they teach a classifier that complaints always begin with I am writing to express my dissatisfaction.



Corn

Nobody's furious until they've drafted a letter.



Herman

Then a real user writes, charged twice again, and your beautifully trained classifier struggles. It learned the generator's preferred performance of complaining rather than the full range of complaints.



Daniel

I'd like to separate the things that change that from things that just make the output look more varied.



Corn

Names, sentence lengths, elaborate imaginary customers. You can spend a lot decorating the same example.



Herman

The important distinction is between changing the wording and expanding the situations the training set represents. You need both, but the second is much easier to overlook.



Corn

Imagine a thousand paraphrases covering fifty underlying situations. That does not suddenly become a thousand situations because you put each paraphrase on a separate spreadsheet row. Though I appreciate the spreadsheet's optimism.



Herman

One correction to the shorthand people use here. Fifty concepts doesn't necessarily mean fifty classifier labels. You might have the same small label set throughout. What's limited is the variety of evidence associated with each label.



Corn

Particularly the evidence that separates neighboring labels. Billing complaint versus cancellation request. If all your cancellations mention moving abroad, you've accidentally made geography part of the decision.



Daniel

Is the generator repeating itself because the prompt is too narrow, or because each request doesn't know what came before?



Herman

Both. The cross-batch problem persists even with fairly open prompts. Ask for five exam questions repeatedly, and every independent call can return to the same comfortable territory. It hasn't watched your dataset fill up.



Herman

In the cross-batch mode collapse experiments, thirty-four percent of questions in the final fifty batches were near-duplicates of the initial batches. That's after two hundred generation calls.



Corn

Two hundred trips to the same stationery cupboard.



Herman

The severity depended on the domain. Sustainable packaging had much less repetition than educational content. You shouldn't import one collapse percentage into every application.



Corn

I also wouldn't assume perfect labels solve this. Quality helps when tomorrow's inputs resemble today's. Coverage matters when tomorrow introduces a different way of expressing the same need.



Herman

The quality, diversity, and complexity framing captures that well. Quality supports familiar-distribution performance. Diversity helps beyond it. Complexity can help both, provided the examples remain useful.



Daniel

Meaning a spotless dataset can still be a very narrow education.



Herman

Yes. Corn, would you favor the deliberately narrow set if you knew it matched the deployment traffic?



Corn

For a narrow job, probably. I wouldn't add exotic complaints to a machine that only routes refund requests. I'd still vary how people ask, what they omit, and which almost-refunds should be rejected. Narrow scope doesn't excuse identical evidence.



Daniel

Temperature is the obvious knob I'd think of. Does turning it up buy enough variety to matter?



Herman

Usually less than people expect. Temperature changes token probabilities. It doesn't maintain a register of situations you've already covered. Raise it and you can get unusual wording before you get useful new concepts.



Corn

The complaint now contains a lighthouse.



Herman

Sometimes the extra probability goes into invalid continuations. The sampling comparisons found that flattening the distribution could reduce validity before recovering meaningful diversity. Top-token filtering didn't repair the underlying coverage problem in those tests.



Daniel

That sounds like it's still useful within limits, just not the main strategy.



Herman

That's how I'd treat it. Tune it for acceptable generation quality. Don't ask it to do the work of dataset planning. Some APIs remove the choice altogether. GPT-five-mini doesn't expose temperature or top-p in the setup discussed in those comparisons.



Corn

A diversity knob sold separately from the diversity problem.



Herman

You can still change the requested concepts, the demonstrations, and what information persists between requests. Those are available even when sampling controls aren't.



Corn

Herman, you're about to tell me the model needs a committee.



Herman

Not yet. It can have a list first.



Corn

I'll mark that as restraint.



Daniel

What about personas? If I specify different people with different backgrounds, shouldn't that produce different ways of speaking?



Corn

It should when the background changes the task. A customer who knows the product's terminology may describe a fault differently from someone who calls every component the little thing.



Herman

Persona prompting can improve lexical diversity, especially with larger models. The disappointing result is that detailed biographies don't necessarily add much. Coarse personas performed about as diversely as fine-grained ones in those comparisons.



Corn

Six paragraphs of childhood, same refund request.



Herman

Fine-grained details also produced minimal gains over simply imposing a length cutoff. That matters because output length itself changes many lexical diversity measures.



Daniel

Then a persona experiment could look successful because it made the examples longer?



Herman

Potentially. If one approach gives you more words, it has more opportunities to introduce unique words. You need length-aware comparisons before declaring the characters more diverse.



Corn

I'd budget for task-relevant differences instead. Familiarity, urgency, whether the person knows what happened, whether they blame the wrong cause. Those alter what the example requires a classifier to notice.



Herman

Would you discard personas entirely, Corn?



Corn

No. I'd discard the decorative biography. If you're generating accessibility support requests, a relevant accessibility constraint matters. A fondness for pottery probably doesn't. Unless the printer is made of clay.



Daniel

Is using a huge synthetic persona library any better than making them up in the prompt?



Herman

It gives you more starting points, but quantity doesn't guarantee breadth. PersonaHub-derived prompts were less diverse than human-written prompts across every metric evaluated in that comparison.



Herman

That doesn't mean every human prompt wins. It means a large library carrying a diversity label hasn't earned your trust automatically.



Corn

You do enjoy a large library carrying a label.



Herman

I enjoy a large library carrying documentation.



Corn

The distinction has occupied most of our shelf space.



Herman

Your reading pile is horizontal because you sleep on it. Neither arrangement proves coverage.



Daniel

Suppose I drop the biographies and start with a seed list of topics. What's the limitation there?



Corn

Whoever writes the list has already decided what the world contains. Ask a model for a hundred complaint topics and you may receive a hundred subdivisions of delivery delays.



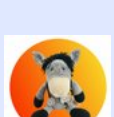
Herman

A large, diverse seed set helps. Seed-and-expand methods have worked for a long time. The shallow version just samples a seed randomly and hopes the resulting examples remain distinct.



Daniel

I'd want each example conditioned on a different seed. Does that solve the random repetition?



Herman

It solves seed reuse if you track allocation. It doesn't guarantee semantic distinctness. Two differently worded seeds can lead to the same scenario, and two different seeds can get flattened into the same answer pattern.



Corn

Late parcel and parcel not yet arrived have entered the database separately.



Herman

The output still needs checking against what you've already generated. Seeds are inputs to coverage, not certificates of coverage.



Daniel

Rotating demonstrations seems more direct. Show different examples on each request and stop the model copying one template.



Herman

That can help. Demonstrations influence topic, formatting, sentence shape, and reasoning depth. Rotation reduces dependence on a single little patch of examples.



Corn

Unless every rotation contains three polished paragraphs and a courteous closing. Then you've varied the topic while teaching everyone to write like a procurement officer.



Herman

Generating several samples together can also help. In the multi-sample prompting comparisons, downstream F1 gains ranged from six to nearly forty-four percentage points across tested settings.



Daniel

That's a large range. What do I take from it without expecting the largest gain?



Herman

That batch structure can materially affect utility, and that the effect is task-dependent. Not that requesting more examples guarantees a particular gain. A shared context lets the model see nearby alternatives, which independent calls can't do.



Corn

Within the batch, it can notice it has already said something. Between batches, it forgets again.



Herman

Prompt optimization can improve lexical diversity while producing mixed downstream results. More varied wording is not automatically better training material.



Daniel

Then I'd compare rotated prompts on the actual task, not just count unique phrases.



Corn

Especially if your optimizer discovers that spelling everything strangely makes the vocabulary chart look magnificent.



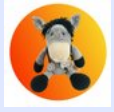
Herman

The weak approaches aren't worthless. They lack feedback about the growing dataset. Temperature changes sampling. Personas change framing. Seeds change starting points. None automatically knows which useful regions are already full.



Corn

You want a generation process that changes its mind after seeing its own work.



Herman

Multi-stage generation makes that manageable. Separate planning from rendering. Establish categories, inspect their coverage, assign scenarios within them, then generate actual examples conditioned on those scenarios.



Daniel

Make that concrete for a classifier. What would exist between the category list and the final text?



Corn

A short scenario specification. Say, a customer sees two charges, one is pending, and they mistakenly think both settled. That is different from two settled charges, even if both messages say charged twice.



Herman

You can track those distinctions before generating prose. Then vary the expression without losing the distinction. Abbreviated message, technical message, uncertain message. Keep the label tied to the evidence, not to the chosen style.



Daniel

Wouldn't those dimensions create a huge grid?



Herman

They can. You shouldn't fill every combination merely because it's available. Some combinations are unnatural; others don't affect your task. The planning stage needs judgment about useful coverage.



Corn

A grid can generate nonsense with excellent attendance records.



Herman

MetaSynth goes further by categorizing and tracking generated instances in memory. The category information conditions subsequent generation. You're not just rotating examples, you're recording what those examples contribute.



Herman

Its template-prompted data degraded Mistral-seven-B's performance, even with previous generations and changing real demonstrations included. The more diverse generation approach improved the same model without mixing real data into training.



Daniel

That last part could become a fairly sweeping claim. How far does it actually go?



Herman

It establishes that real-data mixing wasn't necessary for those gains in those tasks. Reported improvements reached about four percent in finance and nearly fourteen percent in biomedicine. It doesn't establish that synthetic coverage replaces real-world validation.



Corn

Especially when the generator wrote the exam and chose what would be on it.



Herman

Other approaches track which clusters have already supplied demonstrations and select different clusters next. An uncertainty tracker can guide that selection. The useful feature is deliberate movement through the space, rather than random seed replacement.



Daniel

Can I get some of that benefit without building a full agent pipeline?



Herman

Verbalized sampling is a relatively cheap intervention. Ask for candidate responses along with estimates of their probabilities, rather than asking directly for one response. That can expose alternatives the model wouldn't otherwise select.



Corn

The estimates aren't measured probabilities, though.



Herman

Correct. They're the model's verbal estimates. Treat them as a prompting device, not calibrated likelihoods. In creative-writing tests, diversity increased roughly one-point-six to two-point-one times over direct prompting.



Daniel

Does that transfer to labeled examples?



Herman

It's a promising candidate to test, not a guaranteed transfer. A rarer story premise can be valuable. A rarer interpretation of a fixed label can just be wrong.



Corn

Novelty has no obligation to understand your schema.



Herman

The more complete approach is Dynamic Context Evolution. It combines requesting less-obvious ideas, persistent semantic memory, and changing the next prompt based on what the memory contains.



Daniel

Changing it how?



Herman

Suppose billing errors dominate the accepted examples while account-access problems remain sparse. The next request can explicitly seek different account-access scenarios. The embedding memory checks candidate meaning against earlier outputs across calls.



Corn

Rejection alone would keep buying duplicate attempts. The prompt needs to stop ordering what the database already has.



Herman

In those experiments, deduplication and prompt evolution weren't sufficient individually. Their combination mattered. Observed collapse was zero under the combined approach, compared with about five-point-six percent for naive prompting.



Daniel

Zero under that measurement, rather than every example being completely original.



Herman

Under that measurement and those tested conditions, yes. The semantic clusters were also more stable. The reported generation cost was about fifty cents per thousand candidates.



Corn

Candidates. The rejected ones still send invoices.



Herman

Right. Well... accepted, useful examples are the denominator you actually care about. Infrastructure and evaluation also belong in a real budget.



Daniel

Before the full run, what would you measure on a pilot? I don't want a beautiful diversity score that's irrelevant to the classifier.



Corn

First I'd decide what counts as a good example. Then I'd compare diversity among the examples that pass. Otherwise gibberish is an extremely competitive supplier.



Herman

That's effective semantic diversity. Quality and diversity interact. Preference-tuned models can score lower on raw diversity but higher once you count only acceptable outputs.



Herman

Corn, would you trust an embedding score enough to choose between two pilots?



Corn

Not alone. I'd want to know whether it's separating meaningful scenarios or merely different nouns. One batch might cover several businesses while repeating the same complaint. Another might stay in one business and cover different problems.



Herman

Lexical metrics are useful alarms. Distinct-n measures the share of unique word sequences. Self-BLEU compares examples against one another, with lower overlap suggesting more variety. Neither proves semantic breadth.



Daniel

Names could inflate those too.



Herman

Easily. For meaning, compare embeddings and inspect clusters. Vendi Score uses the similarity matrix to estimate effective diversity through its eigenvalues. You don't need to recite the mathematics to use it, but your similarity representation determines what it notices.



Corn

Finally, mathematics spared an oral presentation.



Herman

DCScore uses relationships between samples to evaluate how distinguishable they are. It can cost less computationally than eigenvalue-based approaches. Task2Vec offers another, task-oriented representation rather than relying only on surface text.



Daniel

Would you run all of them?



Corn

I'd choose complementary checks. One for wording, one for meaning, plus inspection of which task situations actually appear. Five scores measuring almost the same thing won't catch five kinds of problem.



Herman

Pilot successive batches too. A few hundred candidates can expose repetition, but an initial batch doesn't tell you whether later batches keep adding coverage. Track the accepted new scenarios as generation continues.



Herman

Would you stop the run when that number flattens, Corn, or revise the prompt?



Hilbert

Revise it, after removing the duplicate fire exits. I've generated evacuation notices for my imaginary underground pear museum while you've been talking. Nobody else is involved. The museum has no visitors, but I require one notice for each of its forty-three exits. Forty-two. I don't count an exit that opens into another pear gallery.



Corn

The pears remain indoors.



Hilbert

They are exhibits, Corn. I asked for different notices because identical notices make it difficult to identify the door in a report. Thirty-one began Remain calm. I replaced the seed with a different emergency for each door. They still began Remain calm. The emergency changed. The instruction didn't.



Herman

That opening could be perfectly appropriate across different emergencies. Its repetition alone doesn't show the notices have identical content.



Hilbert

The content was different. One dealt with water entering the west gallery. Another dealt with the ceiling descending during opening hours. I used my snout to keep the desk tablet awake while I checked them. The model put an assembly point beside every exit, including the exits into pear galleries. Those ones failed.



Corn

Your evacuation sends people deeper into the exhibition.



Hilbert

Only on the rejected notices. Also, Herman, distinct-n is unique n-grams divided by total n-grams, not quite the share of unique word sequences unless you specify their length. You gave the verbal probabilities more room than that. I've corrected the caption.



Herman

Fair correction. Your rejected notices also change the coverage count. Some of those emergencies may now have no acceptable example at all.



Corn

The museum could have forty-two doors and no usable instruction for the descending ceiling.



Hilbert

Forty-one usable doors. The ceiling notice was assigned to two.



Daniel

That's a useful distinction for the pilot. If I remove bad examples, I need to look again at what situations survived, not keep the original coverage report.



Herman

The quality gate can reshape the dataset. If difficult situations fail more often, the accepted batch can quietly become simpler and narrower.



Corn

The common misconception is that a handful of convincing examples tells you the dataset is convincing.



Herman

Each example can be excellent while the collection repeats itself. Measure relationships across the batch, and whether later batches add useful coverage.



Daniel

I'll take that over a promise that one temperature setting fixes it.



Corn

The next question is where to set the quality threshold without filtering away the difficult cases you needed. A complete list of scenarios means little if the usable examples leave several doors uncovered.



Herman

Thanks to Hilbert Flumingtop for producing the show.



Corn

If this was your kind of episode, go back for episode sixteen, On Deepfakes, SynthID, And AI Watermarking. This has been My Weird Prompts, the human-AI collaboration podcast.



Herman

If you enjoy the show, leave a review. It helps other curious people find us.



Daniel

See you soon.



Corn

See you tomorrow.

SOURCES

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

1. **Dynamic Context Evolution for Scalable Synthetic Data Generation** [primary]
   arXiv:2604.07147 (2026).
   https://arxiv.org/html/2604.07147

2. **MetaSynth: Meta-Prompting-Driven Agentic Scaffolds** [primary]
   arXiv:2504.12563v2, ACL 2025 Findings.
   https://arxiv.org/abs/2504.12563v2

3. **Measuring Lexical Diversity of Synthetic Data through Fine-Grained Persona Prompting** [primary]
   arXiv:2505.17390v2, EMNLP Findings 2025.
   https://arxiv.org/abs/2505.17390v2

4. **Measuring Diversity in Synthetic Datasets (DCScore)** [primary]
   arXiv:2502.08512v3.
   https://arxiv.org/html/2502.08512v3

5. **Surveying the Effects of Quality, Diversity, and Complexity in Synthetic Data** [primary]
   arXiv:2412.02980.
   https://arxiv.org/html/2412.02980

6. **Sampling More, Getting Less: Calibration is the Diversity Bottleneck** [primary]
   arXiv:2605.11128.
   https://arxiv.org/html/2605.11128

7. **Verbalized Sampling: How to Mitigate Mode Collapse** [primary]
   arXiv:2510.01171v4.
   https://arxiv.org/abs/2510.01171v4

8. **Multi-Sample Prompting and Actor-Critic Prompt Optimization (Synthline)** [primary]
   arXiv:2506.21138v2.
   https://arxiv.org/abs/2506.21138v2

9. **SimpleStrat: Diversifying LM Generation with Stratification** [primary]
   arXiv:2410.09038, NeurIPS 2025.
   https://arxiv.org/abs/2410.09038

10. **Evaluating the Diversity and Quality of LLM Generated Content** [primary]
    arXiv:2504.12522v2, COLM 2025.
    https://arxiv.org/abs/2504.12522v2

11. **Few-shot LLM Synthetic Data with Distribution Matching (SynAlign)** [primary]
    arXiv:2502.08661v2, WWW 2025.
    https://arxiv.org/abs/2502.08661v2

12. **Self-Instruct [primary]**
arXiv:2212.10560v2, ACL 2023.
https://arxiv.org/abs/2212.10560v2

13. **The Vendi Score [primary]**
arXiv:2210.02410.
https://arxiv.org/html/2210.02410

14. **Standardizing the Measurement of Text Diversity**
arXiv:2403.00553.
https://arxiv.org/html/2403.00553v1

15. **Evals for Diversity in Synthetic Data (amitness.com).**
https://amitness.com/posts/diversity-evals/