Preventing Model Collapse in the Synthetic-Data Era

A Path Forward for Safer, More Robust AI Systems

Presenter

Suraj Ranganath

DSC291: Safety in Generative AI

What Happens When Models Learn from Their Own Outputs

Model collapse: Generative models trained on data generated by previous models gradually "misperceive reality," losing critical information about the original distribution.

Two Types of Collapse

Early Collapse

Models lose information about rare events and the tails of the distribution.

Late Collapse

Models converge to distributions with little resemblance to the original, with reduced variance.

Three Sources of Error

Statistical Approximation

Information loss from finite sampling of tail regions.

Functional Expressivity

Neural networks cannot perfectly fit all distributions.

Functional Approximation

Learning procedure biases (e.g., SGD) compound errors across generations.

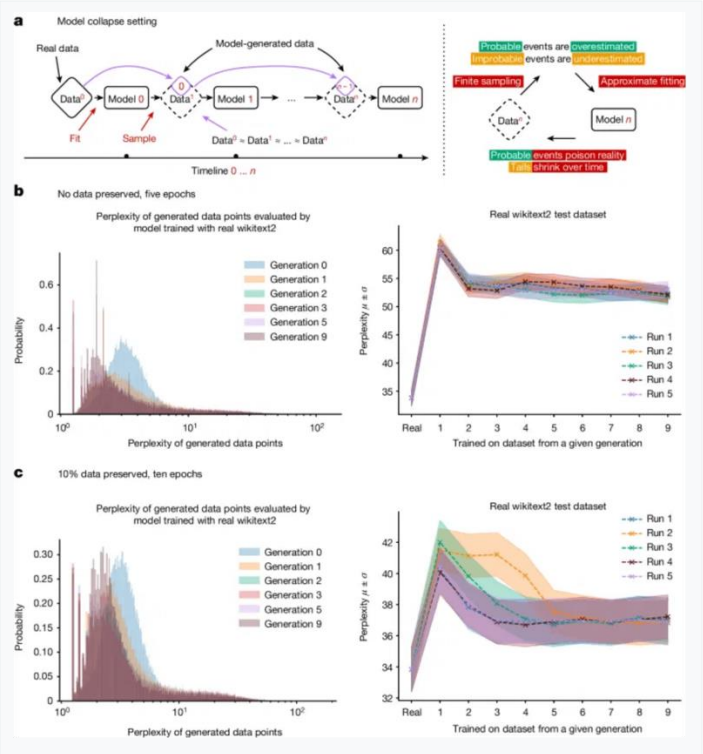


Figure 1: The feedback mechanism. Real data enters the system, models generate synthetic data, which then contaminates the next generation's training set.

Source: Shumailov et al. (2024). Nature, 631, 755-759.

Why Rare Phenomena Disappear: The Tail Erosion Problem

The Problem

When you sample from a distribution, you naturally oversample high-probability regions and undersample rare events in the tails. Each generation of the model learns from this biased sample.

The Cascade Effect

When that model generates new data, it generates even fewer tail samples. The next model learns from this even more biased distribution. **By generation 3 or 4, the model has forgotten that rare events exist.**

> **Key Insight:** Rare events are lost first because they're naturally underrepresented in finite samples. This compounds exponentially across generations.

Why This Matters

Rare events are critical: edge cases in medical diagnosis, unusual user behaviors, novel scenarios in safety systems. Losing the tails means losing the ability to handle important edge cases.

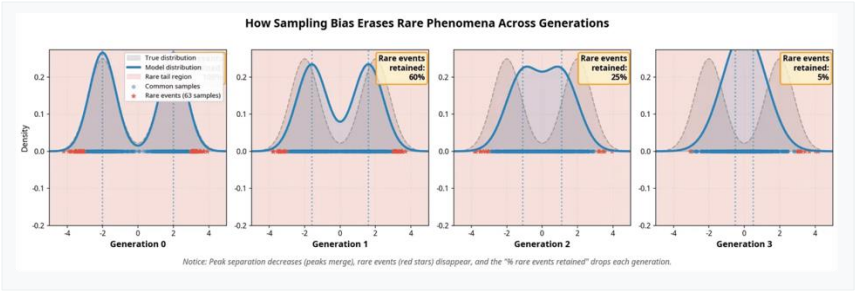


Figure: Sampling bias erases rare phenomena. Notice peak separation decreases, rare events (red stars) disappear, and rare event retention drops each generation.

Foundational Research: What We Know About Model Collapse

Shumailov et al. (2024): The Comprehensive Study

Nature, 631, 755-759

Key finding: Recursive training on synthetic data causes irreversible defects across language models, vision models, and diffusion models.

Mechanism: Three sources of error (statistical approximation, functional expressivity, functional approximation) compound across generations.

Alemohammad et al. (2023): The Tail Loss Problem

arXiv:2305.17493

Key finding: Training on model-generated data leads to loss of information in the tails, even with small amounts of synthetic data.

Implication: Sampling bias causes rare events to be progressively underrepresented, compounding across generations.

Gerstgrasser et al. (2024): The Accumulation Solution

COLM 2024; arXiv:2404.01413

Key insight: The critical factor is whether data are replaced or accumulated. Replacement causes collapse; accumulation prevents it.

Validation: Results hold for transformers, VAEs, and diffusion models across multiple domains.

Critical Takeaway: Model collapse is empirically proven across multiple architectures. We now know what causes it (replacement) and what prevents it (accumulation).

Why This Matters Now: The Urgency of 2024-2025

1

Synthetic Data Adoption is Rising

Companies are generating synthetic data at scale for training and cost reduction. This is becoming standard practice.

2

Web Contamination is Real

AI-generated content is flooding the internet. Future models will inevitably train on this contaminated data during web scraping.

3

The Vicious Cycle is Accelerating

Model 0 generates synthetic data. Model 1 trains on mixed real and synthetic data. Model 1 generates more synthetic data. The cycle repeats and accelerates.

Timeline: This is happening now. October 2025 papers discuss watermarking as a solution. If we don't act immediately, future models will be trained on increasingly degraded data.

Can Synthetic Data Be Safe?

If synthetic data causes model collapse, how can we ever use synthetic data safely?

The problem is **not synthetic data itself**, but **replacement**. If you accumulate synthetic data with a non-shrinking real-data anchor, you avoid collapse.

Adding more data (even synthetic data) might actually prevent collapse, not cause it. This challenges our intuition about data contamination.

Two Guiding Questions

- What role does the real-data anchor play in preventing tail loss?
- What happens to the distribution when you replace data versus accumulate it?

Think about these questions. The answers reveal why accumulation is the key to preventing model collapse.

Solution 1: Accumulate, Never Replace

Core Strategy

Never replace real data with synthetic data. **Accumulate both together**. The real data serves as an anchor that preserves information about the true distribution.

Why It Works

The real-data anchor preserves information about the tails. As you accumulate, the proportion of synthetic data increases, but the absolute amount of real data never decreases.

> The model always has access to the true distribution (via real data), preventing tail loss and distribution shift.

Empirical Evidence

Gerstgrasser et al. (COLM 2024) tested this across multiple domains:

Replacement
Test loss grows linearly. Collapse is inevitable.

Accumulation
Test loss remains bounded. Collapse is avoided.

Implementation

Version Data
Keep all real data in a non-shrinking anchor.

Track Ratios
Monitor real vs. synthetic percentage per round.

Set Alerts
Alert if real-data anchor shrinks below threshold.

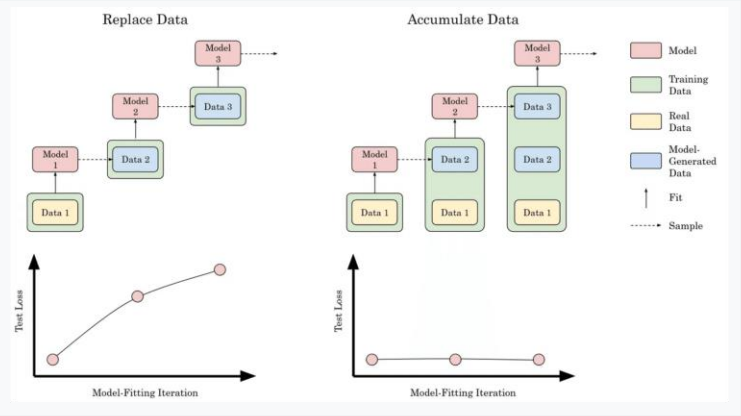


Figure: Replacement leads to collapse. Accumulation preserves the real-data anchor and avoids collapse.

Source: Gerstgrasser et al. (2024). COLM 2024. arXiv:2404.01413

Solution 2: Watermarking and Provenance Tracking

The Problem: Unlabeled Synthetic Data

Synthetic data is being uploaded to the internet without any indication that it is synthetic. Future data collectors cannot distinguish real from synthetic content. This means synthetic data will inevitably contaminate future training sets.

How Watermarking Works

Embed Imperceptible Markers

Embed imperceptible markers in LLM outputs that identify them as synthetic. These markers are designed to be invisible to humans but detectable by algorithms.

Purpose: Data Curation, Not Misuse Detection

Watermarking is not about preventing misuse or deception. It is about enabling **data curation**: helping future data collectors identify and exclude synthetic content from their training sets.

Low Computational Overhead

Watermarking can be deployed at generation time with minimal computational cost. No need for separate post-processing or expensive infrastructure.

Provenance Tracking (C2PA-Style)

Attach Metadata

Attach metadata to every data point indicating its source (real vs. synthetic), generation date, and the model used to generate it. This creates a complete provenance chain.

Data Curation at Scale

During web crawls, use detectors to flag synthetic content and de-duplicate near-copies. This prevents synthetic data from dominating the training corpus.

Recent Work: October 2025

Recent research argues that watermarking is a key tool for preventing model collapse through data curation. The focus is on designing watermarking systems that **align the incentives of all stakeholders**: model developers, data collectors, and downstream users.

"LLM Watermarking Should Align Stakeholders' Incentives..." (arXiv, October 2025)

Key Insight: Watermarking and provenance tracking are not about preventing deception. They are about enabling safe data curation at scale. When combined with organizational policies and technical guardrails, they form a comprehensive defense against synthetic data contamination.

Solution 3: Active Curation and Mixture Management

Active Selection: Curate Which Synthetic Data to Include

Don't just accumulate all synthetic data. Actively curate which synthetic data to include in your training set based on what the model needs.

- **Prioritize diversity:** Select synthetic data that fills gaps in the real-data distribution. Avoid synthetic data that duplicates or near-duplicates real data.
- **Use diversity metrics:** Employ metrics like maximum mean discrepancy or entropy to ensure the synthetic data adds new information rather than just amplifying existing patterns.
- **Avoid mode collapse:** Don't let the model generate the same outputs repeatedly. Synthetic data should expand the training distribution, not narrow it.

Decoding Constraints: Generate Diverse Outputs

When generating synthetic data, use constraints to ensure the model produces diverse outputs rather than collapsing to a few high-probability modes.

- **Temperature sampling:** Increase sampling temperature to generate more diverse outputs. Higher temperature means the model explores lower-probability regions.
- **Nucleus sampling:** Sample from the top-p% of the distribution to avoid the long tail of very unlikely outputs while still maintaining diversity.
- **Constrained decoding:** Force the model to generate outputs that are sufficiently different from the training data. This prevents the model from just memorizing and reproducing existing examples.

Mixture Management: Monitor and Control Ratios

Actively manage the real:synthetic ratio in your training data. This ratio is a critical control parameter that affects model collapse risk.

- **Track mixture ratios:** Monitor the percentage of real vs. synthetic data in each training round. Make this visible to your team.
- **Set target ratios:** Based on your domain and task, establish target ratios (e.g., 70% real, 30% synthetic). Adjust dynamically based on model performance and diversity metrics.
- **Adjust dynamically:** If model performance degrades or diversity metrics decline, increase the real-data proportion. If the model is performing well, you may be able to increase synthetic data slightly.

Practical Takeaway: Synthetic data is not a free lunch. It requires active curation, careful generation, and continuous monitoring to be safe. The three strategies above form a practical framework for managing synthetic data at scale while preventing collapse.

What Researchers and Companies Can Do Today

For Researchers

Investigate Anchor Size

How small can the real-data anchor be while still preventing collapse?

Study Detector Robustness

Can watermark detectors reliably identify synthetic data across models and domains?

Explore Beneficial Synthetic Data

When is synthetic data beneficial without risking collapse? Characterize the safe zone.

Develop Mixture Strategies

What mixture ratios work best for different domains? Build empirical benchmarks.

For Companies

Policy Safeguards

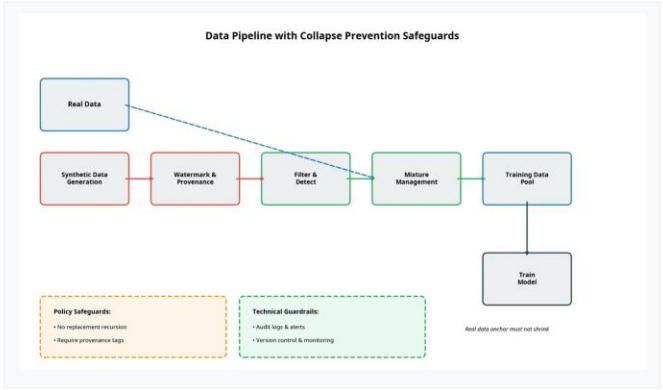
Make "no replacement recursion" explicit policy. Require provenance tags.

Deploy Technical Guardrails

Run watermark detectors during web crawls. Maintain audit logs. Set up automated alerts.

Coordinate on Standards

Establish industry standards for watermarking. Share best practices and tools.



End-to-end data pipeline with collapse prevention safeguards: from generation through training, with policy-level and technical controls.

Custom visualization created for this presentation.

Open Questions and Future Research Directions

Open Research Questions

Anchor Size

How small can the real-data anchor be while still preventing collapse? Does this depend on model size, data dimensionality, or task complexity?

Detector Robustness

Can learned watermark detectors reliably identify synthetic data that has been paraphrased, translated, or modified? How do we design robust detectors?

Beneficial Synthetic Data

When is synthetic data beneficial (e.g., for safety fine-tuning, rare-event simulation, domain adaptation)? How do we characterize the safe zone?

Mixture Ratios

What are the optimal mixture ratios for different domains and tasks? Does this depend on the quality of the synthetic data?

Generalization

Do these strategies generalize across model architectures, domains, and scales? What breaks down at larger scales?

Promising Research Directions

Theoretical Analysis

Extend the linear model analysis to more complex settings. Prove bounds on model divergence as a function of mixture ratio and anchor size.

Empirical Characterization

Systematically study model collapse across different domains, architectures, and mixture ratios. Build a taxonomy of collapse phenomena.

Watermarking Design

Develop watermarking schemes that are robust to paraphrasing, translation, and adversarial attacks. Align incentives across stakeholders.

Data Curation Methods

Develop automated methods for active curation of synthetic data. Use diversity metrics and information-theoretic measures to select high-quality synthetic data.

Standards and Coordination

Work with industry to establish standards for provenance tracking and watermarking. Build shared infrastructure and open-source tools.

Long-term Vision: A future where synthetic data is safe, traceable, and beneficial for AI development. This requires coordination across researchers, companies, and policymakers.

Three Rules of Thumb for Preventing Model Collapse

1

Never Replace

Avoid replacement-style recursion. Accumulate real and synthetic data together.

> *Maintain a non-shrinking real-data anchor that preserves information about the tails.*

Why: The real data serves as a ground truth that prevents the model from forgetting about rare events and edge cases.

2

Label Everything

Watermark synthetic data and attach provenance metadata to every data point.

> *Enable future data collectors to identify and exclude synthetic content from their training sets.*

Why: Without labeling, synthetic data will inevitably contaminate future training sets and trigger collapse in downstream models.

3

Curate the Tails

Actively manage data mixtures and preserve diversity in your training set.

> *Monitor mixture ratios and set alerts for anchor shrinkage. Use diversity metrics to guide curation.*

Why: Rare events are the first to disappear in model collapse. Protecting the tails protects the entire distribution.

Key Takeaway: Model collapse is preventable with intentional design and organizational commitment. These three rules form the foundation of a safe, robust approach to synthetic data in the era of recursive training.

The Path Forward: Building Safer AI Systems

The Challenge

As synthetic data becomes cheaper and more abundant, the risk of model collapse increases exponentially. We must act now, not later. The window for prevention is closing as synthetic content floods the internet.

The Opportunity

We have concrete, practical strategies to prevent collapse. We know what works: accumulation, watermarking, active curation, and organizational safeguards. We have the knowledge. Now we need the commitment to implement it.

The Call to Action

Researchers

Investigate the open questions. Build better watermarks and detectors. Characterize the safe zone for synthetic data. Your work will shape the future of AI development.

Companies

Implement policy safeguards and technical guardrails today. Coordinate with other labs on standards. Make collapse prevention the default, not an afterthought.

All of Us

Think about data provenance and curation. Ask hard questions about where your training data comes from. Demand transparency and accountability from the systems we build.

The Vision

A future where AI systems are trained on clean, traceable, well-curated data. Where synthetic data is a tool for improving AI, not a source of contamination. Where safety and robustness are built in from the start.

Model collapse is a solvable problem. But only if we design our systems intentionally and coordinate across the community.

Thank you. Questions?.