Model Collapse, Synthetic Data, Recursive Compression, and Reality Drift

How AI systems train on machine-generated representations, reuse compressed outputs as primary data, narrow the distribution of knowledge, and gradually lose contact with the reality their models were built to represent

Representational Failure Series - Note #3

A. Jacobs

Model collapse is commonly described as a failure that occurs when generative models are repeatedly trained on synthetic data produced by earlier models. Errors accumulate. Rare patterns disappear. Outputs become more repetitive. The generated distribution gradually narrows.

Recursive Compression describes the deeper mechanism. A model compresses reality into a learned representation and produces outputs from that representation. When those outputs are reused as training data, the next system does not return to the original world. It learns from a compressed reconstruction of an earlier compression.

Model collapse is therefore not caused by synthetic data in the abstract. It emerges when representations are recursively reused without enough renewed contact with the reality from which they were derived.

1. The Inference Gap

Most discussions of model collapse focus on data contamination.

Human-created data is mixed with machine-generated material. Models train on outputs produced by previous models. Synthetic artifacts spread through datasets. Performance declines as the ratio of generated content increases.

This explanation identifies an important source of degradation, but it does not fully explain why recursive reuse is structurally dangerous.

Synthetic data is not automatically low quality. It can be useful for simulation, augmentation, privacy protection, testing, and coverage of rare cases. A generated example may be accurate, diverse, and well suited to a particular training objective.

The deeper problem appears when synthetic representations begin replacing independent contact with the underlying distribution.

Every model output has already passed through selection, compression, and reconstruction. Some information has been preserved. Some has been weakened. Some has disappeared. Some regularities have been exaggerated because they were easier for the model to learn or reproduce.

When those outputs become the inputs to another model, the second system inherits the first model's representational limits as though they were features of reality.

The system is no longer learning primarily from the world.

It is learning from what an earlier system was able to retain about the world.

That is the transition from synthetic training data to Recursive Compression.

2. The Mechanism

The relationship unfolds through a sequence of sampling, compression, generation, reuse, and progressive narrowing.

Stage 1: Reality as a Complex Distribution

The original environment contains more variation than any model can preserve.

Human language includes common patterns, rare expressions, local knowledge, contradictions, mistakes, specialist vocabularies, shifting meanings, and forms of experience that appear only under unusual conditions. Images contain ordinary scenes, edge cases, physical irregularities, cultural artifacts, and details that may be difficult to classify.

The real distribution is uneven.

Some patterns occur frequently. Others are rare but important. Some relationships are easy to detect. Others depend on context that appears across long distances or small subsets of the data.

No training dataset captures all of this.

The first compression begins before the model is trained.

Stage 2: Dataset Construction

Reality is converted into a dataset.

Content is collected, filtered, cleaned, labeled, deduplicated, categorized, and prepared for training. Legal, commercial, technical, and safety constraints determine what remains available.

This process makes the environment computationally usable.

It also removes structure.

Some material is excluded because it is difficult to collect. Some is removed because it appears anomalous. Some disappears during cleaning. Some populations and contexts are underrepresented from the beginning.

The dataset is not reality. It is a selected representation of reality.

Stage 3: Model Compression

The model learns statistical structure from the dataset.

It does not store every example as a complete copy. It compresses regularities into parameters that support prediction and generation.

Frequent patterns exert greater influence. Stable associations become easier to reproduce. Rare cases may be weakly represented because they contribute less to average training performance.

The resulting model contains a further reduction of the original distribution.

The path now looks like:

Reality → selected dataset → learned representation

Each stage preserves useful structure while discarding detail.

Stage 4: Synthetic Reconstruction

The model generates new outputs from its compressed representation.

These outputs may resemble the original data closely. They can be fluent, coherent, and statistically plausible. They may preserve many important patterns.

But generation is reconstruction, not recovery.

The model produces what its internal representation makes probable. It tends to reproduce dominant forms more reliably than rare ones. Ambiguous or weakly learned distinctions may become cleaner, more generic, or more conventional.

The generated output is therefore not another sample from the original reality.

It is a sample from the model's representation of that reality.

Stage 5: Representational Reuse

Generated outputs re-enter the information environment.

They appear in websites, repositories, image collections, summaries, support material, social posts, reference pages, and training corpora. Some are labeled as synthetic. Many are not.

Later systems collect this material alongside human-produced data.

The synthetic output now functions as evidence about the distribution the next model is expected to learn.

This is the critical shift.

A representation has become source material for another representation.

Stage 6: Recursive Compression

The next model compresses the generated reconstruction.

Any distinction lost by the first model cannot be recovered from its output alone. Any bias toward common patterns becomes part of the new training environment. Any fabricated regularity may be interpreted as real data.

The second model does not merely repeat the first model's compression.

It compresses the consequences of that compression.

The sequence becomes:

Reality → dataset → model → synthetic output → new dataset → new model

Each cycle moves one step further from independent observation.

This is Recursive Compression.

Stage 7: Distributional Narrowing

Repeated reuse favors patterns that survive compression well.

Common styles become more common. Probable phrasings become more visible. Familiar visual arrangements dominate. Rare structures receive less representation because they were less likely to be generated in the first place.

The center of the distribution grows denser while its edges weaken.

Outputs can remain polished. They may even become more consistent. But consistency increasingly reflects what previous models found easy to represent rather than the full complexity of the source environment.

Diversity decreases before obvious quality failure appears.

Stage 8: Error Reinforcement

Representational errors can become self-reinforcing.

A model generates an inaccurate claim, oversimplified category, or distorted relationship. That output is published or stored. Later datasets collect it. Future systems encounter multiple copies of the same reconstruction.

Repetition gives the artifact statistical weight.

The model-generated pattern begins to look like an independent regularity because the training system cannot reliably reconstruct its origin.

The error no longer appears as a single mistake.

It becomes part of the environment being modeled.

Stage 9: Model Collapse

The learned distribution becomes increasingly detached from the original data-generating process.

Rare modes disappear. Outputs become less varied. Central patterns become exaggerated. Errors accumulate. The model may generate increasingly uniform or unstable results.

This is model collapse.

The system has not simply consumed too much bad data. It has lost sufficient access to the independent variation needed to correct its own representations.

The model begins learning from the residue of previous models rather than from the world those models attempted to represent.

3. Why Synthetic Data Is Not the Real Problem

The phrase synthetic data can make the failure sound like a simple distinction between human and machine production.

That is too crude.

Human-produced information can also be repetitive, derivative, inaccurate, and detached from direct experience. Machine-generated data can preserve useful structure, expand coverage, and help train systems under controlled conditions.

The relevant distinction is not human versus synthetic.

It is grounded versus recursively derivative.

A generated dataset can remain grounded when it is constrained by physical simulation, verified labels, external measurements, or regular comparison with real observations. A human-written dataset can be weakly grounded when it consists mainly of summaries, copied claims, institutional templates, or repeated interpretations.

Recursive Compression begins whenever downstream systems rely on representations that are no longer being corrected by the underlying source.

Synthetic data makes this dynamic easier to scale, but it does not create the principle.

4. Compression Is Necessary

Compression is not itself a defect.

Every model compresses. Every dataset selects. Every summary removes detail. Every category makes variation manageable.

Without compression, large-scale learning and coordination would be impossible.

The problem appears when a compressed representation is mistaken for a renewable source of the structure it removed.

A model can generate many new samples, but multiplying outputs does not restore lost information. A thousand reconstructions from the same compressed distribution do not equal a thousand independent observations of reality.

Volume can create the appearance of diversity while remaining bounded by the same representational limits.

Recursive Compression therefore concerns the direction of informational dependence.

The system keeps producing more material while drawing from a progressively narrower source.

5. Why Rare Patterns Disappear First

Common patterns are easier for models to learn because they appear often and contribute heavily to average performance.

Rare patterns have weaker statistical support. They may require unusual combinations of context, long-range dependencies, specialist knowledge, or low-frequency forms.

During generation, the model tends to select outputs with relatively high probability. Rare forms therefore appear less often than they did in the original distribution.

When generated outputs become training data, that reduction becomes part of the next dataset.

The next model sees even less evidence of the rare pattern.

Across repeated cycles, the process can look like:

Rare in reality → weaker in the first model → uncommon in generated output → nearly absent from later training data

The distribution loses its edges before losing its center.

This matters because rare cases often contain the information needed to distinguish genuine understanding from smooth approximation. They test whether a system has preserved the structure of the domain rather than only its most common surface patterns.

6. Recursive Compression and Cultural Flattening

The same mechanism can operate outside formal model training.

People increasingly encounter the world through search rankings, recommendation systems, summaries, generated answers, and machine-produced media. Creators then use those outputs as references for new work.

A style becomes visible because an algorithm selected it. Others imitate the visible style. The platform receives more examples of that style and becomes more confident that it represents what users want.

A phrase appears in generated summaries. Writers repeat it. Search systems index the repetition. Later models treat the repeated phrasing as established language.

Culture begins training on its own machine-mediated outputs.

The result is not necessarily total uniformity. New combinations still appear. But variation increasingly occurs inside a narrowed field of already legible patterns.

The system produces novelty through recombination while weakening the less predictable sources from which deeper novelty once entered.

7. Search Degradation and Recursive Source Loss

Search systems depend on the assumption that indexed pages represent relatively independent sources.

Recursive generation weakens that assumption.

Many pages may repeat the same machine-produced explanation. Summaries may cite other summaries. Reference material may derive from an AI answer that itself synthesized earlier web content.

The number of apparent sources increases while the number of independent observations decreases.

This creates source inflation.

A claim may appear well represented because it occurs across many documents. But those documents may share a hidden origin or arise from the same generated reconstruction.

The information environment becomes denser and less independent at the same time.

Search degradation begins when retrieval systems cannot distinguish repeated representations from genuinely separate evidence.

8. Provenance Failure

Recursive Compression becomes more dangerous when provenance disappears.

If synthetic material is clearly identified, downstream systems can weight, filter, or evaluate it differently. When its origin is unknown, the output is treated as another observation.

As generated material is copied, edited, summarized, and redistributed, its lineage becomes harder to reconstruct. A model output becomes a blog post. The blog post becomes a reference. The reference enters a dataset. The dataset trains another model.

At each step, the relation to the original source weakens.

The system may eventually retain the claim while losing any record of whether it came from observation, inference, duplication, or generation.

Without provenance, recursive reuse can no longer be reliably measured.

The information environment loses the ability to tell whether it is learning from reality or from its own previous compressions.

9. Model Collapse and Reality Drift

Model collapse describes the deterioration of the learned distribution.

Reality Drift describes the larger condition in which systems continue operating through recursively weakened representations.

A model may collapse in a measurable technical sense. But the surrounding system may also remain functional long before collapse becomes obvious.

Generated answers remain fluent. Search results remain populated. Content continues arriving. Benchmarks may show modest declines or no immediate failure.

The information system continues producing representations while the independent reality beneath those representations becomes less influential.

This is Reality Drift.

The system looks active because output volume remains high.

Its contact with reality weakens because the outputs increasingly derive from other outputs.

10. Model Collapse Is an Outcome, Not the Operator

The direction of the relationship matters.

Model collapse does not produce Recursive Compression.

Recursive Compression is one mechanism that can produce model collapse.

The canonical relationship is:

Repeated representational reuse + insufficient grounding → Recursive Compression → distributional narrowing → model collapse

This distinction allows the concept to generalize.

Model collapse is one technical outcome. Recursive Compression can also produce cultural imitation, semantic flattening, source inflation, institutional self-reference, and degraded search environments without causing a formally recognized model collapse.

The operator is broader than the technical failure.

11. Boundary Conditions

Not every use of generated data produces Recursive Compression or model collapse.

Synthetic data can remain useful when:

- it is generated from a well-grounded external model
- its provenance is preserved
- it supplements rather than replaces real observations
- rare and edge cases are deliberately protected
- quality is checked against independent evidence
- new environmental data continuously enters the system
- generated errors are filtered before reuse
- the training process distinguishes derivative data from primary data

Recursive Compression becomes more likely when:

- generated material is collected without reliable provenance
- synthetic output replaces expensive real-world data
- the same models generate and evaluate the data
- high-probability examples dominate generation
- rare modes receive little protection
- datasets are repeatedly deduplicated toward common forms
- downstream models cannot access the original sources
- output volume is mistaken for informational diversity

A system does not need unlimited access to primary reality.

It needs enough independent correction to prevent its own representations from becoming the only environment it can see.

12. Composition

This mechanism connects several concepts across machine learning, information systems, and the Reality Drift framework.

Data selection describes the first reduction of reality into a usable corpus.

Compression describes the preservation of regularities through the removal of detail.

Synthetic reconstruction describes outputs generated from a learned representation rather than sampled directly from the original environment.

Representational reuse describes the return of those outputs into future datasets and information systems.

Recursive Compression describes repeated learning from prior compressions.

Distributional narrowing describes the loss of rare forms and expansion of dominant patterns.

Provenance failure explains why derivative outputs become difficult to distinguish from independent evidence.

Model collapse describes the technical deterioration that can result.

Reality Drift describes the broader condition in which the system continues producing coherent outputs while losing alignment with the reality from which its representations originally came.

The full sequence can be written as:

Reality → dataset selection → model compression → synthetic reconstruction → representational reuse → Recursive Compression → distributional narrowing → model collapse

A broader canonical form is:

Recursive reuse of representations without renewed grounding → progressive loss of contact with the source distribution

13. The Operator Test

If model collapse is being produced through Recursive Compression, the system should lose more than average quality.

Rare patterns should disappear faster than common ones. Outputs should become more conventional, repetitive, and concentrated around previously successful forms. Apparent diversity may remain high at the surface while deeper structural variation declines.

The system should also become increasingly dependent on material produced through the same representational lineage.

A useful test is:

How many independent encounters with the underlying reality stand behind the apparent volume of data?

If thousands of examples descend from a small number of model-generated reconstructions, the dataset may contain abundance without independence.

A second test is:

Can the system distinguish primary observations from representations derived from earlier representations?

If not, it cannot reliably estimate how many times the same compression has been recycled.

Evidence against the mechanism would include preservation of rare modes, stable performance under repeated training cycles, reliable provenance, and continuous correction through independently collected real-world data.

Canonical Claim

Model collapse can emerge when AI systems repeatedly train on compressed reconstructions produced by earlier models.

Each model preserves part of the original distribution and removes part of it. When its outputs become the next system's data, those losses are inherited as though they belonged to reality. Repeated cycles narrow variation, reinforce artifacts, and weaken the corrective role of independent observation.

The dataset grows.

The outputs multiply.

The system becomes increasingly informed by what previous systems could reproduce rather than by the world they were built to represent.