IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Model Collapse: A Comprehensive Review of Causes, Detection, and Mitigation

Murtaza S. Malik

Raha International School Gardens Campus, Khalifa City 'A, Al Raha Gardens Cnr. Al Mireef St, Abu Dhabi, 00000, UAE; murtaza.pog@gmail.com

ABSTRACT: As generative AI systems become increasingly trained on synthetic data, often produced by earlier iterations of the same models, concerns about model collapse have come to the forefront of AI research. Model collapse is the degradation of a model's output quality when it recursively learns from its outputs, leading to a loss of diversity, factuality, and generalization over time. Therefore, output starts to become increasingly repetitive and less creative, making it biased toward high-frequency or stereotypical patterns, because rare or nuanced examples gradually get filtered out. This makes the information users receive narrower, less trustworthy, and potentially discriminatory against marginalized groups whose experiences are underrepresented in high-probability data. These consequences of model collapse raise ethical dilemmas, especially for its applications in education, healthcare, hiring, and public services. This review investigates the detection and prevention of model collapse in generative AI, examining various studies on regression modeling, recursive training, and prompt evaluations. These studies compare synthetic and real data, track rare token loss, and analyze the compounding effects of low-diversity exposure. This project contributes to the field by consolidating current research showing collapse is measurable, mitigable via curated data, token-level editing, hybrid training, and unlearning, while proposing a hybrid benchmark for detection.

KEYWORDS: Model Collapse, Recursive Training, Synthetic Data, Output Bias, AI Ethics, Model Unlearning.

Introduction

At the core of generative AI, which is a subset of AI systems capable of creating original content,[1] lies a silent danger: these models may eventually lose sight of the richness of the human data they were trained on. A phenomenon known as "model collapse" occurs when AI systems are repeatedly trained on their own outputs, resulting in outputs that get increasingly bland, biased, and detached from reality.[2]

In recent years, large language models (LLMs), like LLaMA[3] and ChatGPT,[4] have made remarkable progress in generating text that is both fluent and relevant.[1] However, as developers use more and more synthetic data generated by previous iterations of these models for further training, concerns regarding long-term degradation have increased. Models that undergo recursive training eventually lose their rare token diversity, factual accuracy, and generalization ability.[2] The performance of models can drop even when synthetic data initially seems to be of high quality.[5] These effects can also result from overexposing models to repeated examples, even if they are clean and human-generated. When rare token loss sets in, the model's output becomes significantly dominated by a few high-frequency categories, and the long tail, less common ones, just disappear. As seen in Figure 1,[6] in this context, the model is finetuned with synthetic data, with the model collapsing onto a small set of nationalities, with most vanishing. Collectively, this literature indicates that collapse is increasingly recognized as a theoretical and practical threat.

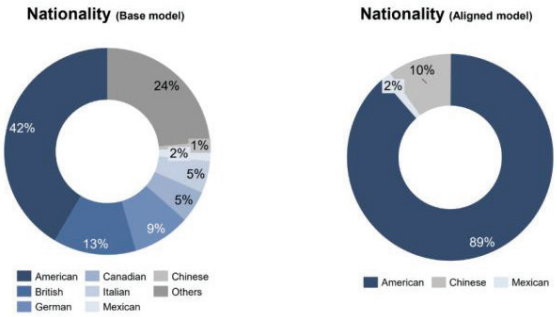

Figure 1: The preference-aligned model trained on simulated user outputs collapses to a narrow set of nationalities, demonstrating rare-token degradation compared to the more varied base model. Figure adapted from Devansh.[6]

Understanding model collapse is crucial because its implications extend beyond academic boundaries. As collapse progresses, model outputs become more homogeneous, overconfident, and reliant on high-probability patterns[5] which can lead to inaccurate information, the exclusion of minority voices, and a drop in user confidence.

Model collapse is a real and quantifiable phenomenon, but it can be lessened with intentional interventions. Our goal is to bridge the gap in the literature by integrating recent findings and collectively show that methods such as model confidence filtering, which filters synthetic outputs based on the model's certainty to retain higher-quality examples,[2] token-level editing, which replaces low-value or overly predictable tokens to preserve diversity,[7] and hybrid training with real data, which combines synthetic and human-generated data to maintain grounding in real-world distributions,[5] can all help lower the

DOI: 10.36838/IJHSR87.54

risk of collapse. Furthermore, to potentially reverse collapse, we suggest a novel approach that combines a model-testing framework to identify and reverse collapse in future AI pipelines with machine unlearning, a method for eliminating malicious data from machine learning (ML) systems.[8]

Our research methodology is mostly grounded in literature, referencing recent works that combine experimental benchmarks with theoretical modeling. Furthermore, we examine secondary data and suggest ways to incorporate it into upcoming AI development processes. Relevant papers were identified through systematic searches on Google Scholar using targeted keywords such as "model collapse," "rare token degradation", etc. Furthermore, we prioritized recent publications as well as applying citation snowballing by reviewing references within key papers to identify other relevant studies.

This is how the rest of the paper is organized:

Background information on synthetic data, generative AI, and the mechanisms underlying model collapse is given in Section 3. In Section we discuss and analyze a variety of model collapse monitoring methods 4. Prevention and mitigation techniques are discussed in Section 5. Additionally, we propose 2 methods for preventing and tracking model collapse through the use of model unlearning and hallucination tracking in Section 6, along with a discussion of their possible advantages and disadvantages. Finally, in Section 7, the paper's main conclusions are summarized, and the necessity of proactive collapse prevention in long-term AI deployment is emphasized.

Background

3.1. Generative AI and Synthetic Data:

Artificial intelligence is increasingly being integrated into our world today, from being used for customer service calls to medical imaging analysis, but all of this is thanks to generative AI systems. Generative AI systems, one of the most commonly used systems today, are a type of machine learning (ML) module that produces original content like text or images.1 These systems have rapidly transformed industries by automating tasks such as editing, writing, and summarization. As their capabilities grow, so too does their use. Millions of people now rely on generative AI tools in daily life, whether to draft emails, create social media posts, write essays, or generate blog content.[1] Due to this, a substantial portion of online content is now AI-generated. This rise in usage has led to the emergence of synthetic data, AI-created content that, once released publicly, blends into the broader web.[2] Since most modern AI models are trained on data scraped from the internet, this synthetic content often re-enters the training pipelines without filtering out. Even content with a small degree of AI involvement can be labeled as synthetic, and as AI usage becomes more common, this kind of data is leaking rapidly (see Figure 2). After a large pool of synthetic data eventually leaks into training data, it quietly evolves into a feedback loop (see Figure 2), where models are unintentionally trained on their own outputs. This phenomenon, if left unchecked, creates the conditions for a serious challenge in AI development: model collapse. Roman[9] describes how recursive training reduces sensitivity to rare events and shows how even a small overlap between training and evaluation data can obscure early signs of degradation. These risks are particularly important in professions where objectivity and accuracy are essential, such as healthcare, law, and education.[1]

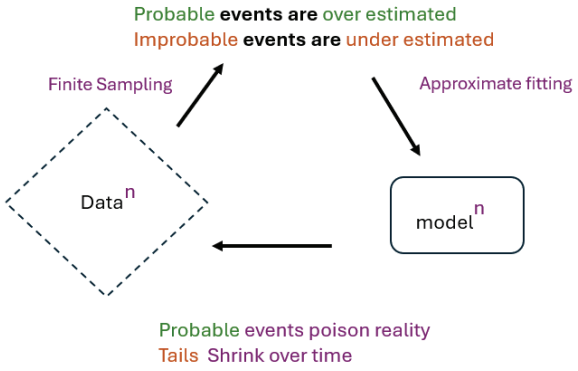

Figure 2: Model collapse is the downward spiral that occurs when a model keeps training on its own outputs, erasing low-probability events and reinforcing a biased view of reality. Adapted from Shumailov *et al.*[2]

3.2. Defining Model Collapse:

Model collapse is where a model's performance iteratively degrades as it recursively trains from its own low diversity outputs. Over time, this leads to "rare token" and "rare scenario" loss, such as when, in Gerstgrasser *et al.*,[10] models in later generations begin introducing irrelevant and implausible content or examples, shifting from factual architectural descriptions to surreal details like "black-tailed, white-tailed, blue-tailed, red-tailed, yellow-tailed jackrabbits" when describing churches (see Table 1). Furthermore, models can also hallucinate. In hallucination, models are displaying and outputting clearly false information or sources. The most well-known example is when Google's Gemini hallucinated in 2023 that the James Webb telescope claimed a discovery that never happened in the first place.[11]

The accuracy of the model is significantly impacted when rare token loss and hallucinations are allowed to continue, which causes the model to marginalize underrepresented data and favor high-probability, stereotypical responses. This is evident with the rise of gender bias and the decline of gender-neutral alternatives in generative AI models[10] where prompts such as "The doctor walked into the room and..." are nearly always completed with "he looked at the patient." [10]

This is a noteworthy example of a stereotypical response from a generative AI system. Additionally, the performance of these models starts to become more unreliable because of the growing gap between training data and real-world use cases.

3.3. Early and Late Collapse:

Researchers distinguish between early collapse, where rare or low-frequency tokens are lost, and late collapse, where output variance drops and the model drifts away from the original data distribution.[7] While catastrophic forgetting erases prior task knowledge when learning new ones, model collapse stems from the cumulative reuse of model-generated data (see Table 1).

Table 1: Model collapse- catastrophic forgetting. Adapted from Shumailov *et al.*[2] The table showcases progressive degradation in an OPT-125M language model when repeatedly trained on its own generated data, showing loss of semantic coherence, topic drift, and eventual nonsensical output across generations.

Example of text outputs of an OPT-125m model affected by Model Collapse. The model iteratively degrades where each new generation is trained on data produced by the previous generation.			
Inputs:			
some started before 1360 — was typically accomplished by a master mason and a small team of itinerant masons, supplemented by local parish labourers, according to Poyntz Wright. But other authors reject this model, suggesting instead that leading architects designed the parish church towers based on early examples of Perpendicular			
Outputs:			
Gen 0: Revival architecture such as St. John's Cathedral in London. The earliest surviving example of Perpendicular Revival architecture is found in the 18th @-@ century Church of Our Lady of Guernsey, which dates from the late 19th century. There are two types of perpendicular churches : those	**Gen 1:** architecture such as St. Peter's Basilica in Rome or St. Peter's Basilica in Buenos Aires. There is no evidence that any of these buildings were built during the reign of Pope Innocent III, but it is possible that they may have been built during the reign of his successor, Pope Innocent	**Gen 7:** architecture in England. In an interview with The New York Times, Wright said : " I don 't think there is anything wrong with me being able to do what I want to do. It just doesn't work for me. " He added : " I don 't know if you can call it funny,	**Gen 9:** architecture. In addition to being home to some of the world's largest populations of black @-@ tailed jackrabbits, white @-@ tailed jackrabbits, blue @-@ tailed jackrabbits, red @-@ tailed jackrabbits, yellow @-

3.4. Evidence That Model Collapse is Real:

Multiple empirical and theoretical studies demonstrate that model collapse is not a hypothetical failure mode, but a real and measurable phenomenon. In Shumailov *et al.*,[2] recursive training experiments on language models showed a sharp decline in lexical diversity and a measurable loss of rare tokens over successive generations. Additionally, it was also seen that there was a reduced recall for rare tokens and a drift toward high-frequency, less informative tokens, even when fluency remained stable.[12] Also, the test accumulates predictably with each iteration of self-training through regression-based modelling.[7] This is further confirmed by Feng *et al.*[13] that collapse is accompanied by semantic drift, disappearance of tail-distribution content, and factual inaccuracy. These findings together provide strong, cross-paper evidence that model collapse occurs in practice and can be reliably detected using quantitative metrics.

3.5. Effects of Model Collapse:

The harmful effects and consequences of model collapse are real and devastating. The loss of rare tokens and scenarios in their memory can erase minority group representation, which in turn narrows the diversity of model outputs and reinforces existing biases. The likelihood of factually incorrect outputs, undermining trust in high-stakes domains such as healthcare, law, and education, is greatly increased due to a rise in semantic drift and hallucination.[13] Along with that, late-stage collapse leads to embedding convergence,[7] reducing the model's ability to differentiate between contexts and degrading performance across a wide range of tasks. Overconfidence in incorrect answers further exacerbates harm by making failures less obvious to users. These effects collectively reduce the model's reliability, fairness, and alignment with real-world knowledge.

■ Monitoring Model Collapse

Monitoring model collapse is a key measure to ensure and protect the long-term reliability, fairness, and factuality of generative AI models.[2,13] It is seen that early-stage collapse is often invisible to conventional accuracy metrics because it first appears in rare or low probability outputs before affecting high frequency tokens. [2,12,13] It is noted that when collapse actually begins, its nature of self-reinforcing makes it difficult and costly to reverse, even in the smallest models.[2,7,10]

The stakes are greatly increased by environmental costs, as well as the retraining costs incurred.[14] The literature we studied identifies several monitoring approaches to catch collapse at this subtle stage. To compare these, this paper uses a four-factor evaluation scale, adapted from recurring evaluation priorities in the reviewed works:

1. **Sensitivity –** The ability to detect collapse onset before significant performance degradation on general benchmarks.
2. **Domain Adaptability –** A technique's efficacy across multiple domains, such as bio-med, financial, and legal.
3. **Computational Cost –** The resources that a method demands for continuous deployment and use.
4. **Diagnostic Depth –** The depth to which the method reveals not only the presence but also the likely cause of collapse.

Each method is rated High (H), Medium (M), or Low (L) for each factor, based on evidence from the reviewed literature.

4.1. Rare Token degradation:

Rare token degradation monitoring tracks and monitors a model's capability to be able to generate low-frequency tokens from its original training distribution.[2] These pertain to uncommon words, specialized domain terms, or even niche phrases that are more likely to disappear when models are trained recursively on synthetic outputs. For example, controlled recursion studies show token diversity decreasing sharply over generations, with rare vocabulary declining while common tokens persist.[12]

Rare tokens are much more sensitive to distributional drift than more naturally occurring frequent tokens, making them an early warning signal that often moves before overall benchmark accuracy.[13,15] In domain-specific continual pretraining (e.g., biomed), maintaining coverage of rare terms is critical; moreover, contamination between train and test sets can mask these early declines if not controlled.[7]

This method is applicable for use in both full pretraining and continual pretraining scenarios, particularly where domain-specific vocabulary is critical, such as in medicine or law.[2,12] Refer to Table 2 for a summary of the analysis.

Table 2: Rare Token degradation: Factor Ratings and Justifications. The table shows that rare-token–based metrics are a low-cost and sensitive early-warning signal for model collapse but offer limited diagnostic depth.

Factor	Rating	Justification
Sensitivity	High	It easily captures early signs of lexical narrowing, rare token loss, and tail erosion well before global accuracy drops.[2,12,16]
Domain Adaptability	Medium	This method works well across many fields but requires an extensively large list of rare tokens per domain (e.g., biomedical/legal), increasing setup time as it is quite time consumin.[16]
Computational Cost	Low	It could be run at evaluation time using fixed datasets (tail perplexity/recall and embedding drift checks), adding minimal overhead contrasted to retraining.[16]
Diagnostic Depth	Medium	It only picks up vocabulary loss and not the root cause unless paired with distributional or attribution analyses; training-data attribution helps trace whether drift stems from synthetic, duplicated, or contaminated sources.[7,10]

4.2. Embedding Drift Monitoring:

Embedding drift monitoring measures how the internal vector representation of rare or semantically important tokens changes across model checkpoints during training or continual pretraining. This method tracks cosine similarity between each token's current embedding and a reference embedding (usually from an early, pre-collapse checkpoint) to detect whether semantic meaning is eroding in the model's representation space.[10,16]

DOI: 10.36838/IJHSR87.54

For example, Dohmatob *et al*.[5] show that in over-parameterized regression models trained recursively on synthetic data, the parameter vector representing the ground-truth function progressively contracts toward zero when each generation is fit on independent data subsets. In their high-dimensional setup, this projection loses compounds over generations, leading the embeddings and thus the model's learned representation, to drift entirely away from the true target. By the later generations, the learned function norms drop close to that of a null predictor, confirming severe semantic drift.

Embedding drift is a structural health metric because embeddings encode semantic and syntactic relationships; distortions in their geometry often occur before surface-level performance declines. This allows it to be an early indicator of collapse that is sensitive to subtle degradation patterns invisible in perplexity or accuracy scores.[10,16]

Embedding drift is able to be used across domain-specific and domain-agnostic fields. It is valuable to implement this strategy for high-stakes fields (e.g., medical or legal NLP) where semantic precision is critical, as even small representational shifts can distort meaning in sensitive terms.[7] Refer to Table 3 for a summary of the analysis.

Table 3: Embedding Drift Monitoring: Factor Ratings and Justifications. The table shows that embedding drift is a highly sensitive and domain-agnostic indicator of early model collapse, offering strong diagnostic depth at moderate computational cost.

Factor	Rating	Justification
Sensitivity	High	Detects representational instability before it appears in text outputs; large embedding drift was observed generations before collapse in Dohmatob et al.'s regression experiments.[5]
Domain Adaptability	High	Works with any token set; only the reference embedding set needs to be domain-specific.[16]
Computational Cost	Medium	Requires periodic checkpoint saving and embedding-space comparisons; scales with vocabulary size but is less resource-intensive than full retraining.[10]
Diagnostic Depth	High	Can localize drift to specific token groups and, when paired with attribution tools, identify problematic synthetic data sources.[7]

4.3. Perplexity Gap Tracking:

Perplexity gap tracking compares a model's perplexity (PPL) over different token subsets, especially the long-tail, low-frequency portion, against its perplexity over the full evaluation set. The gap between these two scores acts as an early warning: a widening gap indicates that the model is losing the ability to predict rare or diverse sequences even if overall perplexity stays stable.[2,12]

For example, Shumailov *et al*.[2] conducted recursive training experiments and used the Wikitext-103 benchmark in them for assessment. While overall perplexity on the benchmark remained largely unchanged in early generations, tail perplexity climbed rapidly, reflecting the loss of low-probability events. This divergence between tail PPL and overall PPL appeared before standard benchmark accuracy degraded, confirming that the perplexity gap is an early indicator of collapse.

Perplexity gap tracking is effective because perplexity directly measures predictive uncertainty. Tail-specific perplexity magnifies the signal from low-frequency events, which are the first casualties of recursive synthetic training, making the gap a leading indicator of collapse risk.

Applicable in both general-domain and domain-specific benchmarks. Particularly useful in specialized areas like biomedical NLP, where monitoring perplexity on rare terminology is critical for early detection of semantic loss.[7] Refer to Table 4 for a summary of the analysis.

Table 4: Perplexity Gap Tracking: Factor Ratings and Justifications. The table shows that perplexity gap tracking provides a simple and scalable signal of distributional shift during model collapse, but is less sensitive to early-stage degradation compared to embedding or rare-token–based methods.

Factor	Rating	Justification
Sensitivity	High	It is able to identify tail-event degradation before full-set perplexity or accuracy drops.[2]
Domain Adaptability	Medium	It requires defining a tail-token list or tail-probability threshold for each domain.[12]
Computational Cost	Low	Uses standard evaluation runs; only requires calculating perplexity over token subsets, no retraining.[2]
Diagnostic Depth	Medium	Identifies tail loss but not the exact cause without pairing with attribution or data-source analysis.[10]

4.4. Output Diversity Index (ODI) Monitoring:

ODI monitoring measures the diversity of generated outputs by computing metrics such as the number of type-token ratio and unique engrams or semantic cluster spread across a fixed set of prompts. The method detects when a model's responses begin favoring higher-probability sequences and forgetting rare cases or tokens, which is a sign of collapse.[10,12]

For example, Dohmatob *et al*.[5] note that in their experiments, the learned function's effective dimensionality shrinks each generation. This caused outputs to progressively increase the projection into a narrow subspace of possible responses, reducing variability in the predictions despite unchanged input diversity. By the final generations in their setup, the downstream model produced outputs statistically indistinguishable from a null predictor, reflecting minimal diversity.

ODI provides a direct behavioral signal of degeneration in generative space. Unlike perplexity-based metrics, which measure prediction uncertainty, ODI captures whether the range of content a model can produce is collapsing, even if the model remains confident in those fewer outputs.

This method applies to both text and non-text generative models (e.g., image diffusion models). In text, it is useful for domains requiring richness of expression, such as creative writing, policy drafting, or customer service, where repetition and narrowing of output styles are undesirable.

Additionally, ODI would apply to both text and non-text generative models. It would prove most useful where expressive range in writing matters, such as in creative writing, because ODI highlights early narrowing of output diversity.

Refer to Table 5 for a summary of the analysis.

Table 5: Perplexity Gap Tracking: Factor Ratings and Justifications. The table shows that perplexity gap tracking provides a simple and scalable signal of distributional shift during model collapse, but is less sensitive to early-stage degradation compared to embedding or rare-token–based methods.

Factor	Rating	Justification
Sensitivity	High	It is able to narrow the generative space before accuracy or perplexity shifts are visible.[5]
Domain Adaptability	High	It works for data like text, image, or code generation; domain-specificity only affects prompt design.[12]
Computational Cost	Medium	Requires generating and analyzing a large set of outputs periodically; cost grows with prompt set size.[10]
Diagnostic Depth	Medium	It as able to flag diversity loss but does not identify the cause without additional attribution or data-tracking tools.[7]

4.5. Benchmark Contamination Auditing:

Benchmark contamination auditing reviews whether training data contains material that overlaps with evaluation benchmarks. The technique works by matching similarities between benchmark items and training sets such as n-grams, hashes, or

embedding.[7] The problem is that contamination skews evaluation results, making it harder to detect early collapse because models can appear to perform well on benchmarks they have already "seen."

For example, Li *et al.*[7] reviewed that widely used benchmarks like PIQA and LAMBADA exhibited inflated accuracy scores when training data contained identical or near-identical items from those benchmarks. After removing these contaminated items, model performance dropped sharply in some cases by more than 20%, which revealed that earlier "good" results had masked underlying generalization problems. In the context of collapse monitoring, this means contamination can hide the early warning signs of degradation by artificially sustaining benchmark scores.

By ensuring benchmark cleanliness, contamination auditing removes a key source of false reassurance. It enables more accurate tracking of real model capabilities and makes collapse detection metrics, such as rare token loss or embedding drift, more reliable and accurate.

Works across general-purpose and domain-specific evaluation sets. Especially important in high-stakes testing environments (e.g., medical QA datasets), where contamination could make dangerous degradations invisible until they cause real-world harm. Refer to Table 6 for a summary of the analysis.

Table 6: Benchmark Contamination Auditing: Factor Ratings and Justifications. The table shows that benchmark contamination auditing is an extremely sensitive and broadly applicable method for uncovering hidden model degradation, even though it provides limited insight into root causes without follow-up analyses.

Factor	Rating	Justification
Sensitivity	High	It is able to reveal hidden degradation by removing artificial performance boosts from contaminated benchmarks.[7]
Domain Adaptability	High	It can be applied to any benchmark dataset with an accessible reference set.[7]
Computational Cost	Medium	It requires running similarity checks between large datasets; moderate overhead depending on detection method.[7]
Diagnostic Depth	Medium	It is able to identify the presence and extent of contamination but does not explain the underlying cause of performance drop without further analysis.[7]

4.6. Synthetic-to-Real Attribution Ratio (SRAR):

SRAR monitoring measures the proportion of synthetic versus original human-generated training sources based on a model's outputs. SRAR is able to achieve this through using influence functions, nearest-neighbor search in embedding space, data tracing tools, and other attribution methods.[10] If your ratio is increasing, it indicates that the model is relying more heavily on synthetic content, which often correlates with semantic drift and eventual collapse.

For example, Gerstgrasser *et al.*[10] applied an efficient attribution pipeline to track whether generations in LLaMA-derived models were anchored more in synthetic or real training segments. In their recursive training setup, later generations showed a marked increase in outputs whose nearest training neighbors were synthetic, even when those outputs appeared high-quality, signaling that the model's knowledge base was shifting away from human-grounded data.

SRAR directly quantifies the source composition behind outputs, making it a causal rather than purely correlational metric for collapse monitoring. This allows teams to detect shifts in data reliance before quality metrics degrade.

It can be applied to any model trained on mixed datasets, especially when synthetic loops are part of the training process. Particularly relevant for long-term model maintenance pipelines, where unintentional accumulation of synthetic traces can destabilize outputs. Refer to Table 7 for a summary of the analysis.

Table 7: SRAR: Factor Ratings and Justifications. The table shows that SRAR provides both early detection and strong diagnostic insight into model collapse by directly tracing degradation to rising reliance on synthetic data, at moderate computational cost.

Factor	Rating	Justification
Sensitivity	High	It is able to detect rising synthetic reliance before surface-level degradation appears.[10]
Domain Adaptability	High	It is able to function across domains if attribution-compatible datasets are available.[10]
Computational Cost	Medium	It requires storing embeddings for large portions of training data and running nearest-neighbor search.[10]
Diagnostic Depth	High	It is able to identify that degradation is occurring and why, by tracing it to synthetic data reliance.[10]

■ Model collapse mitigation

Mitigating model collapse is essential to ensuring the safe and effective use of AI systems, especially in high-stakes domains like law and medicine, where accuracy is critical and ethical risks leave no room for error.[2,8] Model collapse typically reduces output quality by promoting over-generalization, which undermines the reliability of AI in these sensitive fields. The compounding nature of collapse makes recovery more challenging and expensive, both monetarily and environmentally, as models must be retrained after they have been infected by collapse to a certain degree.[10]

It has been demonstrated that proactive measures like improved data curation, cautious use of synthetic data, and corrective mechanisms can slow or even reverse the effects of degradation[5]. Specifically, by eliminating collapse-inducing data from models without the need for retraining, our suggested machine unlearning technique provides a fresh and effective solution. This reduces resource usage and environmental impact in addition to restoring model performance.[8] To compare these strategies systematically, this paper evaluates them against four applied criteria drawn from recurring priorities in the reviewed literature:

Scalability – How well the approach performs when applied to large-scale, continuously updated models without prohibitive engineering or infrastructure changes.

Generalizability – An evaluation of the method's capacity to maintain or restore model quality across different data types, languages, and application domains, ensuring robustness beyond the environment in which it was originally designed or tested.

Detection-to-Action Speed – How quickly the strategy can be deployed after early collapse indicators appear, and how effectively it halts progression.

Sustainability and Cost Efficiency – The balance between computational, financial, and environmental costs versus the performance gains achieved.

Each strategy is rated High (H), Medium (M), or Low (L) for each criterion, based on evidence from the reviewed literature.

DOI: 10.36838/IJHSR87.54

5.1. Token-Level Editing:

Token-level editing is a form of post-processing that is narrower and is performed on model-generated text prior to the text being reintegrated into training. Its aims to circulate infrequent or domain-specific lexemes, to lessen the slip of factual accuracy, and to address long-tail thinning that is commonplace on recursive or synthetic-heavy training.[9,15,16]

Rather than accepting synthetic outputs exactly as they are, token-level editing selectively adjusts or replaces certain tokens, typically those that are overly generic or predicted with overly high confidence, to ensure that rare terms and factual anchors remain in the data. We found that this intervention actually preserves long-tail vocabulary, reduces the over-representation of common terms, and improves performance on de-contaminated validation sets,[15] reducing model collapse's effects. Moreover, we found that in recursive training without such controls, rare tokens disappear within a few generations.[2] Additionally, we find evidence that rare token embedding degradation can lead to broader representational collapse.[16] This makes token-level editing an effective way to interrupt collapse processes before they spread through the model's distribution.

This technique can be used in biomedical, legal, and financial domains to maintain coverage of rare or domain-critical terms, such as uncommon disease names, niche legal terminology, or rare financial instruments, when training large language models (LLMs). It can also be applied in multimodal models used in scientific and industrial contexts, where editing captions or answers preserves fine-grained object labels, specialist descriptors, or experimental terms across iterations. In VAEs and diffusion models used in creative and cultural preservation, editing local elements, such as patches that can help maintain stylistic diversity, heritage patterns, and less frequent visual motifs that could otherwise be lost through repeated synthetic reuse. Refer to Table 8 for a summary of the analysis.

Table 8: Token-Level Editing: Factor Ratings and Justifications. The table shows that token-level editing enables rapid, targeted intervention to mitigate early collapse signals, offering high generalizability and fast detection-to-action speed, but with moderate scalability and ongoing computational cost.

Factor	Rating	Justification
Scalability	Medium	It can be integrated into large-scale pipelines but requires probability estimation and token-level modifications across vast datasets, which increases processing overhead for continuously updated models.[15]
Generalizability	High	It can be applied to a wide range of domains and modalities, including text, multimodal captions, and even latent edits in generative models, with adaptation of editing criteria to domain-specific rare terms or features. [2]
Detection-to-Action Speed	High	Token editing can be triggered immediately upon detecting high-confidence overuse of generic tokens or loss of rare tokens, allowing for rapid corrective action before degradation spread.[15]
Sustainability and Cost Efficiency	Medium	More cost-efficient than full retraining and reduces long-term collapse risks, but continuous token-level scoring and selective resampling carry non-trivial computational costs at scale.[9,16]

5.2. Hybrid Training with Real and Synthetic Data:

Hybrid training encompasses combining high-quality human data with carefully controlled synthetic data during pretraining or continual updates. The method's goal is to expand coverage and reduce costs while avoiding the long-tail erosion and distribution shift seen in purely synthetic or recursive training.[9,15]

Hybrid training works by integrating vetted synthetic data into human data pipelines while enforcing safeguards to maintain distributional integrity. Gerstgrasser *et al.*[10] show that accumulating human data alongside new samples each round keeps test error bounded, whereas replacing earlier data causes error to grow over time.[5] and Zhu *et al.*[15] recommend gating synthetic content through selection or editing to preserve rare events and factual anchors, since naive mixing can distort scaling laws and truncate the long tail. These safeguarding techniques aid in preventing the consequences of model collapse, such as the loss of rare tokens and divergence from human distributions.[2,9]

This method can be used in biomedicine, law, and economic forecasting to ensure that LLMs retain rare clinical terms, jurisdiction-specific legal precedents, and low-frequency economic scenarios by blending curated human corpora with vetted synthetic continuations. It can also be applied to VAEs and diffusion models in scientific imaging and creative industries to preserve morphological diversity in medical scans, variability in architectural models, and stylistic range in art generation. In multimodal systems for specialized manufacturing and environmental monitoring, hybrid training helps maintain rare component labels, species identifiers, or environmental event descriptors by combining human-authored text and images with high-quality synthetic examples. Refer to Table 9 for a summary of the analysis.

Table 9: Hybrid Training with Real and Synthetic Data: Factor Ratings and Justifications. The table shows that hybrid training offers a highly scalable and sustainable long-term mitigation strategy against model collapse, balancing cost efficiency and generalizability, though it relies on prior detection signals to guide intervention.

Factor	Rating	Justification
Scalability	High	It is designed and built to work in large-scale, continuously updated pipelines by mixing curated human and synthetic data. Accumulation-based integration avoids prohibitive retraining cycles while maintaining data diversity.[10]
Generalizability	High	This is applicable across domains (biomedicine, law, economics, creative industries) and modalities (text, image, multimodal) by adjusting synthetic-data gating and selection criteria to domain-specific need.[2,15]
Detection-to-Action Speed	Medium	It needs prior detection of drift or model collapse indicators to adjust synthetic-to-human ratios, making it slightly slower to respond than purely post-processing intervention.[5]
Sustainability and Cost Efficiency	High	It is able to decrease the environmental and financial costs compared to full retraining by supplementing human datasets with vetted synthetic samples, while retaining rare-event coverage to prevent costly collapse recovery.[9,10]

5.3. Confidence-Based Filtering:

Confidence-based filtering is a technique applied to synthetic or model-generated data before it is used for further training. It can be classified as a quality control process as it removes low-quality, error-prone, or hallucinated outputs by evaluating the model's own confidence scores for each token or sequence. This method limits the spread of errors that could otherwise accelerate model collapse by excluding samples that the model is either uncertain about or confidently wrong on.[2,7,15]

Confidence-based filtering works by assigning a probability or confidence score to generated outputs and uses that score to decide whether the sample is edited, discarded, or kept. We see that removing low-confidence synthetic samples before training reduces noise and prevents the model from reinforcing incorrect patterns.[7] Furthermore, Zhu *et al.*[15] notes that in recursive setups, even small quantities of low-quality synthetic data can distort token distributions over time. Additional Shumailov *et al.*[2] emphasizes that collapse often starts with rare, low-probability events, making confidence-based screening a way to protect correctly generated rare events and discard unreliable ones. By enforcing quality thresholds, this method helps keep the training distribution stable.

This mitigation strategy is applicable in biomedical, legal, and financial domains as it can filter synthetic outputs for tasks such as clinical summarization, case law referencing, or market forecasting in LLMs, ensuring only high-confidence and accurate samples are used in training. Moreover, it can be used in multimodal models for scientific and industrial applications, where screening low-confidence captions or answers preserves the accuracy of fine-grained object descriptions, technical annotations, and experimental data. In VAEs and diffusion models for creative or research purposes, proxy confidence measures such as reconstruction likelihood or discriminator scores can be used to remove off-distribution or low-quality samples, maintaining diversity and coherence in generated outputs. Refer to Table 10 for a summary of the analysis.

Table 10: Confidence-Based Filtering: Factor Ratings and Justifications. The table shows that confidence-based filtering enables fast, pre-training intervention to block low-quality synthetic data, offering high generalizability and rapid response, but with moderate scalability and ongoing computational cost.

Factor	Rating	Justification
Scalability	Medium	Can be applied to large-scale pipelines, but continuous confidence scoring for every token or sequence increases computational overhead as dataset size grows.[7,15]
Generalizability	High	Works across text, multimodal, and generative models by adapting confidence metrics (e.g., token log-probabilities, reconstruction likelihoods, discriminator scores) to domain-specific outputs. [2]
Detection-to-Action Speed	High	Provides immediate quality gating at the point of data generation, allowing faulty outputs to be removed or corrected before entering the training pipeline.[15]
Sustainability and Cost Efficiency	Medium	Reduces long-term collapse recovery costs by preventing noisy data integration, but constant scoring and thresholding introduce ongoing computational expenses.[7]

5.4. Curated Data Selection Pipelines:

Curated data selection pipelines are structured processes for building training datasets that deliberately balance content diversity, quality, and representation. The pipelines are able to organize, filter, and then prioritize both human-generated and synthetic data using pre-defined, adjustable criteria that the training distribution maintains important rare events, factual consistency, and domain coverage.[17] Moreover, this aims to prevent the distributional drift and long-tail loss reported when models rely heavily on uncontrolled synthetic data sources.[2,15,18]

Curated pipelines usually work by selecting and screening data before it's entered into the training corpus. Mattioli *et al.*[18] shows that automated embedding signals usually give a misleading impression of data quality, showcasing that pipelines benefit from robust and strong layers of validation to make sure that filtering decisions preserve a meaningful and diverse set of training data. Zhu *et al.*[15] emphasize that the targeted inclusion of rare or domain-specific samples, especially in recursive or hybrid setups, can slow the erosion of the long tail. Furthermore, this also removes repetitive or low-quality synthetic sequences and stabilizes scaling trends. Shumailov *et al.*[2] show that recursive training without such curation leads to progressive narrowing of token frequency distributions and disappearance of low-probability items. By ensuring every refresh or augmentation of the dataset passes through a robust curation framework, the model is able to retain an increased amount of coverage and resist collapse dynamics.

This method can be applied in biomedical, legal, and economic domains to ensure datasets contain balanced coverage of rare disease cases, jurisdiction-specific rulings, and low-frequency market scenarios when training LLMs. It can be used in scientific and technical multimodal systems to maintain representation of infrequent object classes, rare environmental events, or specialist equipment by curating both text and image data. For VAEs and diffusion models used in creative industries and cultural preservation, curated pipelines can be designed to preserve underrepresented artistic styles, heritage motifs, or rare visual elements in generative datasets, ensuring diversity and authenticity in outputs. Refer to Table 11 for a summary of the analysis.

Table 11: Curated Data Selection Pipelines: Factor Ratings and Justifications. The table shows that curated data selection pipelines provide a generalizable and robust safeguard against model collapse by maintaining balanced, high-quality datasets, though they trade real-time responsiveness for moderate scalability and operational cost.

Factor	Rating	Justification
Scalability	Medium	Its automated filtering is able to scale well to large datasets, however incorporating human-in-the-loop validation for high-stakes domains increases operational complexity at scale.[15,18]
Generalizability	High	This method could be adapted to text, multimodal, and domain-specific datasets adjustable selection criteria, ensuring balanced coverage through of rare events and domain-critical data.[2,18]
Detection-to-Action Speed	Medium	It is able to detect and correct imbalances during dataset refreshes or augmentation cycles rather than in real time, making it slower than inline filtering approaches.[15]
Sustainability and Cost Efficiency	Medium	It is able to prevent costly retraining due to drift by maintaining balanced datasets, but ongoing validation and quality assurance incur moderate resource costs.[18]

5.5. Accumulation over Replacement in Iterative Training:

Accumulation over replacement is a data management strategy implemented in iterative training cycles where new training data is added to the existing dataset rather than replacing older data. The strategy aims to preserve valuable historical information, especially rare or low-frequency events, while still allowing the model to learn from new, diverse samples. This approach directly counters the error growth and long-tail loss observed when prior data is discarded in favor of new inputs during repeated fine-tuning or pretraining rounds.[10]

The principle is simple: maintain the cumulative dataset across training iterations, ensuring that each cycle retains all previously collected high-quality data alongside the newly added samples. This approach keeps test error bounded over many training generations, while replacement strategies cause error to grow linearly with each iteration.[10] Through retaining older data, accumulation ensures that domain-specific or rare content remains part of the distribution, in turn, reducing the likelihood of semantic drift or collapse. Shumailov *et al.*[2] investigate and find that replacing data in recursive setups accelerates the disappearance of low-probability tokens, allowing accumulation to be a direct countermeasure. Furthermore, accumulation can also be combined with other safeguards, such as curated selection or confidence-based filtering, to ensure that the growing dataset remains balanced and high quality.

This strategy can be used in biomedical, legal, and financial domains to keep rare disease case notes, jurisdiction-specific precedents, or low-frequency market events in the training set for LLMs across multiple updates. It can be applied in scientific multimodal systems to maintain infrequent object classes, rare environmental conditions, or specialized equipment annotations in both text and image modalities over successive training cycles. For VAEs and diffusion models in creative

DOI: 10.36838/IJHSR87.54

industries or cultural preservation, accumulation ensures that underrepresented motifs, styles, and niche subject matter remain in the dataset across model refreshes, aiding in preventing stylistic homogenization and mode collapse. Refer to Table 12 for a summary of the analysis.

Table 12: Accumulation over Replacement in Iterative Training: Factor Ratings and Justifications. The table shows that accumulation-based training is a highly scalable and generalizable strategy for mitigating model collapse by preserving rare and historical data across iterations, though it trades immediate responsiveness for moderate long-term resource costs.

Factor	Rating	Justification
Scalability	High	The approach integrates seamlessly into large-scale, continuously updated training pipelines, as adding new data to existing datasets avoids complex replacement logic and maintains stability over many iterations.[10]
Generalizability	High	It is quite effective across textual, multimodal, and generative domains by preserving both general and domain-specific rare content, ensuring robustness beyond the original application scope. [2]
Detection-to-Action Speed	Medium	It operates and works on an iteration-by-iteration basis, preserving rare data without requiring prior drift detection, but cannot react mid-cycle to emergent collapse signs.[10]
Sustainability and Cost Efficiency	Medium	Reduces retraining costs by avoiding loss of historical data, but ongoing storage and processing requirements grow with each cycle, adding to long-term resource usage.[2]

5.6. Synthetic Data Verification / Tail Preservation:

Synthetic data verification and tail preservation are a combination of two methods that ensure that synthetic samples added to the training dataset are of high quality and capable of maintaining coverage of rare, low-frequency events. The verification step focuses on screening synthetic outputs for factual accuracy, relevance, and cohesion. Moreover, tail preservation ensures that uncommon tokens, domain-specific terms, or rare features remain represented in the training distribution.[15,16]

Synthetic data verification begins with the generation of candidate synthetic samples, which are then assessed using a combination of automated quality metrics, human evaluation, or model-based tools to ensure factual accuracy and alignment with domain-specific requirements. However, if this is left unverified, synthetic data can distort token distributions, especially during recursive training by overrepresenting frequent patterns and diminishing the long tail.[15]

To counter this, verification filters are employed to remove or modify low-quality outputs, while tail-preservation mechanisms deliberately incorporate rare tokens sourced from domain glossaries or apply controlled resampling techniques to boost the representation of low-frequency content. Shumailov *et al.*[2] and Yu *et al.*[16] emphasize that maintaining rare token usage is vital for preventing embedding drift and preserving semantic diversity, both of which serve as early safeguards against model collapse.

Synthetic data verification can be implemented in biomedical, legal, and financial datasets to ensure that synthetic training data includes verified, high-accuracy outputs while preserving rare terminology such as uncommon disease names, niche case law terms, or infrequent financial instruments in large language models (LLMs). In scientific and technical multimodal systems, it can maintain coverage of rare experimental conditions, specialist equipment labels, or infrequent object classes by verifying and supplementing both text and image synthetic data. For VAEs and diffusion models in creative and cultural preservation, verification ensures synthetic samples meet authenticity and stylistic quality standards, while tail-preservation techniques maintain representation of rare artistic motifs, heritage designs, or niche cultural symbols that might otherwise vanish through repeated synthetic reuse.

Refer to Table 13 for a summary of the analysis.

Table 13: Synthetic Data Verification / Tail Preservation: Factor Ratings and Justifications. The table shows that synthetic data verification with tail-preservation mechanisms provides a generalizable and preventative safeguard against rare-token loss, though it introduces moderate scalability and cost trade-offs due to ongoing verification requirements.

Factor	Rating	Justification
Scalability	Medium	It could be integrated into large-scale pipelines, but verification and tail-preservation steps require additional processing such as automated scoring, human review, or targeted resampling, which can become resource-intensive at scale.[15]
Generalizability	High	It is adaptable across domains and modalities by adjusting verification criteria and rare-token lists to domain-specific requirements (e.g., biomedical glossaries, legal terminology, cultural heritage motifs).[2,16]
Detection-to-Action Speed	Medium	The quality verification can occur at the point of data generation, but tail-preservation measures often rely on periodic corpus audits rather than instant correction, which slows reaction to new collapse indicators.[15]
Sustainability and Cost Efficiency	Medium	Prevents costly retraining by ensuring synthetic data quality and rare-token retention, but ongoing verification and controlled resampling introduce moderate computational and financial costs.[2,16]

5.7. Data Deduplication and Multiplicity Control:

Data deduplication and multiplicity control are dataset management techniques aimed at reducing redundancy and excessive repetition in training data. Deduplication focuses on identifying and removing exact or near-duplicate samples, while multiplicity control addresses over-representation of similar patterns or repeated content in different forms. Both methods work to preserve diversity in the training distribution and avoid the representation collapse that can occur when models are repeatedly exposed to the same information.[2,5,15]

Excessive repetition in training datasets can bias model representations toward over-frequent features, leading to semantic narrowing and reduced capacity to generate rare or diverse outputs. Zhu *et al.*[15] show that data multiplicity, repeated exposure to the same or nearly identical data, can degrade model performance even if the duplicated data is high-quality, as it shifts probability mass away from the long tail. Deduplication uses exact matching, fuzzy matching, or embedding-based similarity to remove identical or overly similar samples. Multiplicity control goes further by monitoring dataset composition and capping the frequency of highly similar sequences, ensuring a balanced distribution of common and rare elements. Shumailov *et al.*[2] highlights that unchecked repetition exacerbates collapse in recursive setups, as duplicated content becomes disproportionately dominant. By enforcing these controls, the dataset maintains semantic variety and supports stable generalization.

This approach can be used in biomedical, legal, and financial domains to prevent models from overfitting to a small subset of frequently repeated documents, such as common clinical case summaries, boilerplate legal clauses, or widely circulated financial reports, thereby preserving the capacity to handle rare and novel cases in LLMs. In scientific and technical multimodal systems, deduplication and multiplicity control can ensure balanced representation of rare object classes, unique environmental events, or specialist equipment annotations in both text and image data. For VAEs and diffusion models in creative industries and cultural preservation, these techniques help maintain stylistic diversity by preventing certain motifs, genres, or cultural symbols from dominating the training corpus, re-

ducing the risk of mode collapse and homogenized outputs. Refer to Table 14 for a summary of the analysis.

Table 14: Data Deduplication and Multiplicity Control: Factor Ratings and Justifications. The table shows that deduplication and multiplicity control are broadly applicable safeguards against redundancy-driven collapse, offering stable long-term benefits while incurring moderate computational overhead.

Factor	Rating	Justification
Scalability	Medium	This technique can be scaled to large datasets using automated matching and similarity search, but embedding-based approaches are computationally heavy for massive corpora.[15,5]
Generalizability	High	It is applicable across text, multimodal, and generative datasets, with customization of similarity thresholds and frequency caps to fit domain-specific needs. [15]
Detection-to-Action Speed	Medium	Repetition and over-representation can be detected quickly with automated scanning, but full remediation often requires periodic batch processing rather than immediate correction.[5]
Sustainability and Cost Efficiency	Medium	It helps avoid the high costs of retraining from distributional collapse caused by redundancy, but frequent large-scale similarity computations can be resource-intensive.[2]

Discussion

(Note that all ideas discussed in this section are theoretical and not tested or proven to work)

6.1. Hybrid Monitoring Framework for Early Detection and Diagnosis of Model Collapse:

Now that we have delved into all monitoring methods and delved into their strengths and weaknesses (check section 4), we infer that they might work best if paired all together. For example, embedding drift or output diversity is able see how model behavior and internal representations are changing. Benchmark hygiene and attribution checks ensure that these signals reflect real degradation rather than noise or data leakage.[10,15] Additionally, there are other notable signals like tail-focused rare-token metrics that respond quickly when collapse begins.[2,16] Drawing on these complementary strengths, we outline a hybrid monitoring approach that combines fast, lightweight detection with deeper, periodic diagnostics and evaluation safeguards.

We suggest a hybrid monitor that blends deeper, recurring probes with light, always-on checks. As a first line of defense, we could monitor the perplexity gap between rare tokens and the entire set, as well as coverage and recall of rare tokens that are focused on the tail. While headline accuracy stays constant, early collapse frequently manifests as a fading long tail. Twelve. We could apply embedding-drift tests to a limited, carefully chosen set of crucial terms to determine what might be drifting. These would draw attention to modifications in internal representations before the emergence of surface-level errors.[16] We could run a monthly output-diversity panel to monitor for narrowing in generated outputs as a behavioral guardrail.[12] We would combine these with benchmark de-contamination to guarantee that assessments continue to be trustworthy. We could also estimate reliance on synthetic training neighborhoods by including a small synthetic-to-real attribution ratio (SRAR) sample when synthetic data is available.[10,15] We could try to gather actual data during iteration instead of replacing it, since theory and experiments indicate that this helps avoid the feedback loop that causes collapse.[10] Ultimately, when comparing these approaches, we found that each is ineffective in a particular area where another approach excels. As a result, we came to the conclusion that each approach was complementary to the others and considered combining these elements into a collapse-aware benchmark: a small prompt panel for diversity, a periodic drift or attribution slice, a fixed rare-token set for tail perplexity and recall (see Figure 3 for workflow). [2,15,16] However, a key limitation of this hybrid system is scalability. Running multiple overlapping monitors, such as attribution and drift checkers, becomes increasingly computationally heavy and expensive as model size and deployment scale grow with every check, limiting its feasibility for constant use.

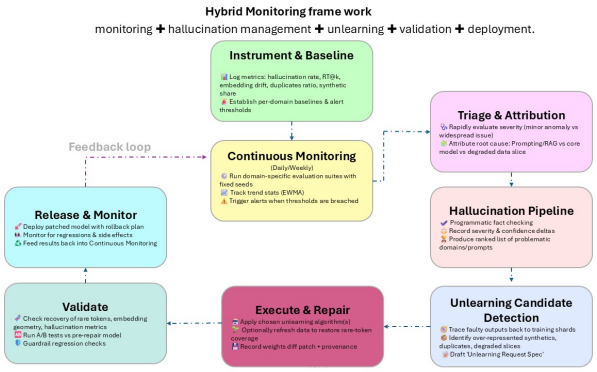

Figure 3: Hybrid Monitoring Framework workflow. The figure illustrates an end-to-end hybrid framework that combines continuous monitoring, attribution, validation, and targeted intervention to detect, diagnose, and mitigate model collapse and hallucination risks across the full training–deployment lifecycle.

6.2. Model Unlearning for Collapse Mitigation:

Model unlearning is a technique that removes the influence of specific training data from a model's parameters without retraining from scratch. Model unlearning operates after training by modifying the model, so it behaves as though specific data had never been seen, rather than blocking problematic data beforehand.[8]

Model unlearning can be used to remove collapse-inducing data, such as over-represented synthetic samples, duplicated low-quality segments, or domain-specific content that has been degraded through recursive training. By selectively erasing the influence of this data, it may be possible to restore balance in the training distribution, recover lost diversity in rare tokens or features, and slow the progression of collapse. For example, in a biomedical question-answering model, unlearning can target repeatedly recycled synthetic case summaries that distort the overall variety of disease terminology. Periodic unlearning cycles could act as a "repair" stage in long-lived LLMs or multimodal systems when collapse indicators such as rare-token erosion or embedding drift reach a critical threshold.

Current unlearning work has mostly targeted privacy and compliance scenarios, such as meeting GDPR "right to be forgotten" requests. Applying it to model collapse raises open research questions:

- Whether unlearning can restore degraded embedding geometry and long-tail coverage.
- How to reliably identify collapse-inducing data segments.
- The maximum degree of recovery possible in late-stage collapse.

DOI: 10.36838/IJHSR87.54

- Developing scalable unlearning methods for multi-billion-parameter models.

Furthermore, model unlearning for collapse mitigation also has its drawbacks. Firstly, it needs extremely precise identification of harmful data, and errors could eliminate useful information, potentially leading to secondary distributional shift. Unlearning algorithms in use today are computationally costly and might not scale well. In particular, current techniques struggle to scale to multi-billion parameter models where unlearning steps become much slower and memory-intensive, and the intensive computational load may lead to the model becoming economically and environmentally unsustainable due to increased energy and cooling use as well as hardware demands.

6.3. Hallucination Tracking as an Early Collapse Signal:

Hallucination tracking methodically logs and examines situations in which a model produces content that is unsupported or factually inaccurate. It offers a behavioral indicator of model reliability over time by concentrating on the frequency, seriousness, and confidence of these errors.

Hallucinations could be a precursor to collapse. Particularly in rare or long-tail knowledge domains, models frequently generate more high-confidence but inaccurate outputs as token distributions get smaller and factual grounding deteriorates. Monitoring hallucination patterns may indicate the beginning of a collapse before more conventional indicators like token coverage or perplexity show appreciable shifts. This can be done by monitoring changes in hallucination rates along with other collapse metrics and conducting regular, domain-specific fact-checking evaluations against fixed reference sets (e.g., biomedical QA, legal facts).

The predictive value of hallucination detection for model collapse remains uncertain, despite its exploration for general AI reliability. Examining the relationship between recorded collapse events and hallucination patterns is one of the main areas of research.

- Creating scalable, automated pipelines for hallucination detection for ongoing observation.
- Assessing whether collapse risk across domains can be predicted by hallucination patterns in one domain.

Moreover, hallucination tracking as an early collapse signal also has its drawbacks. Firstly, false positives can occur in hallucination tracking caused by unrelated problems (prompting, retrieval errors), as well as false negatives on certain specific sub-domains. This, in turn, increases the computational cost of an already expensive tool, potentially making it economically unsustainable. Additionally, the method may face scalability challenges since constant fact-checking and high-frequency monitoring would become increasingly resource-intensive as model domains expand, making it environmentally unsustainable as well, given the energy demands associated with large-scale interference and verification.

■ Conclusion

A serious long-term risk to the dependability, equity, and factual foundation of generative AI systems is model collapse. Particularly in fields that require accuracy and domain coverage, the risk of distributional drift, rare-token loss, and embedding degradation increases as models are trained on their own outputs more frequently. This study examined the theoretical foundations of collapse, reviewed early detection techniques such as output diversity metrics, rare-token monitoring, perplexity gap tracking, and embedding drift analysis (refer to Table 15 for the summary of the analysis), and assessed a variety of mitigation techniques, including hybrid training, token-level editing, confidence-based filtering, curated data pipelines, accumulation, verification, and multiplicity control.

Building on this framework (refer to Table 16 for the summary of the analysis), we suggested two more lines of inquiry: hallucination tracking as a behavioral early-warning signal and machine unlearning to eliminate data that causes collapse after training. Both offer promising avenues, but further empirical investigation is required to determine their viability, scalability, and reliability for integration into large-scale AI development pipelines. Importantly, we also highlighted the shortcomings of the existing methods, including the computational cost of unlearning and the potential ambiguity in hallucination detection.

Our study showcases that no single approach is sufficient to prevent or reverse collapse. Instead, a robust strategy must include supplementary monitoring and mitigation tools, supported by continuous evaluation and domain-aware dataset management. For high-stakes applications in industries like medicine, law, finance, science, and cultural preservation, ensuring that generative models retain their accuracy, diversity, and dependability over the course of their operational lifetimes is not only a technical challenge but also an ethical requirement. However, the paper does not come without limitations.

This paper has several limitations. Experimental validation of the proposed Hybrid Monitoring Framework and Model Unlearning strategy was beyond the scope of this work, so both remain conceptual and grounded in analysis of existing empirical studies rather than controlled experimentation. In addition, the absence of a unified benchmark and the scarcity of detailed performance comparisons in the literature limited quantitative evaluation. Finally, research on model collapse remains fragmented and inconsistently reported, underscoring the need for more empirical studies, standardized benchmarks, and clearer reporting practices to establish reliable detection and mitigation strategies.

In the end, model collapse will require a shift from reactive fixes to proactive design, where data governance, quality assurance, and long-term distributional stability are incorporated into the development process. By combining advances in detection, mitigation, and post-hoc repair, the AI research community can develop systems that can withstand current challenges and long-term deployment pressures in an increasingly synthetic data ecosystem.

Table 15: All monitoring efforts analysis summary. Overall, hybrid training performs best due to its strong scalability, generalizability, and sustainability, while token-level editing performs worst overall as its moderate scalability and cost efficiency limit its effectiveness despite fast detection-to-action speed.

Method	Sensitivity	Domain Adaptability	Computational Cost	Diagnostic Depth
Rare token degradation	High	Medium	Low	Medium
Embedding drift monitoring	High	High	Medium	High
Perplexity gap tracking	High	Medium	Low	Medium
ODI monitoring	High	High	Medium	Medium
Benchmark contamination	High	High	Medium	Medium
Auditing	High	High	Medium	Medium
SRAR	High	High	Medium	High

Table 16: All mitigation efforts analysis summary. Overall, SRAR performs best due to its high sensitivity, strong domain adaptability, and high diagnostic depth, while rare token degradation performs worst overall as its limited domain adaptability and moderate diagnostic depth reduce its effectiveness despite low computational cost.

Method	Scalability	Generalizability	Detection-to-Action Speed	Sustainability & Cost Efficiency
Token-Level Editing	Medium	High	High	Medium
Hybrid Training	High	High	Medium	High
Confidence-Based Filtering	Medium	High	High	Medium
Curated Data Selection Pipelines	Medium	High	Medium	Medium
Accumulation over Replacement	High	High	Medium	Medium
Synthetic Data Verification / Tail Preservation	Medium	High	Medium	Medium
Data Deduplication & Multiplicity Control	Medium	High	Medium	Medium

Acknowledgments

I sincerely and wholeheartedly thank Professor Siddharth Krishnan and Dr. Plinio Zanini for their rigorous mentoring, productive feedback, and academic advice during the process of writing this paper. Their understanding, perspectives, and expertise were integral for this paper to even be written. Furthermore, I deeply appreciate the platform that the IRIS program has meticulously developed to create a venue that promotes inquisitiveness and critical thinking.

Last but not least, I extend my deepest thanks to the academic researchers whose papers and studies were the bones and the foundation of my own paper. Their work was and will always continue to be a fascinating topic of discussion. Furthermore, I attest to the idea that the graphics and writings in this paper are entirely my own.

References

1. NVIDIA. What is Generative AI? | NVIDIA. NVIDIA. https://www.nvidia.com/en-us/glossary/generative-ai. (accessed 2025-12-17).
2. Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Gal, Y.; Papernot, N.; Anderson, R. The Curse of Recursion: Training on Generated Data Makes Models Forget. arXiv. https://doi.org/10.48550/arXiv.2305.17493.
3. META. Meta Platforms, Inc. Meta | Social technology company. https://www.meta.com/en-gb/about/?srsltid=AfmBOopMAuQZ-5tehzkVath6eWVD-_vAiwbojveyj4Se5KolZ3HZomMIE.
4. OpenAI. OpenAI. OpenAI. https://openai.com.
5. Dohmatob, E.; Feng, Y.; Yang, P.; Charton, F.; Kempe, J. A Tale of Tails: Model Collapse as a Change of Scaling Laws. arXiv.org. https://arxiv.org/abs/2402.07043.
6. Devansh. Addressing one of the Biggest Misunderstandings in AI. Medium. https://machine-learning-made-simple.medium.com/addressing-one-of-the-biggest-misunderstandings-in-ai-4d627821346.
7. Li, Y.; Lin, Z.; Yin, F.; Zhang, M. M. Preventing Model Collapse in Gaussian Process Latent Variable Models. arXiv.org. https://arxiv.org/abs/2404.01697.
8. Bourtoule, L.; Chandrasekaran, V.; Choo, C.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; Papernot, N. Machine Unlearning. ieeexplore.ieee.org. https://ieeexplore.ieee.org/abstract/document/9519428.
9. Roman, D. The Collapse of GPT. Acm.org 2025. https://doi.org/10.1145//3722476.
10. Gerstgrasser, M.; Schaeffer, R.; Dey, A.; Rafailov, R.; Sleight, H.; Hughes, J.; Korbak, T.; Agrawal, R.; Pai, D.; Gromov, A.; Roberts, D. A.; Yang, D.; Donoho, D. L.; Koyejo, S. Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. arXiv.org. https://doi.org/10.48550/arXiv.2404.01413.
11. Twenty. AI Hallucinations vs Human Error. Twenty-four.it. https://www.twenty-four.it/insights/ai-hallucinations.
12. Villalobos, P.; Sevilla, J.; Heim, L.; Tamay Besiroglu; Marius Hobbhahn; Anson. Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning. arXiv (Cornell University) 2022. https://doi.org/10.48550/arxiv.2211.04325.
13. Feng, Y.; Dohmatob, E.; Yang, P.; Charton, F.; Kempe, J. Beyond Model Collapse: Scaling Up with Synthesized Data Requires Verification. arXiv.org. https://arxiv.org/abs/2406.07515.
14. Hao, K. Training a single AI model can emit as much carbon as five cars in their lifetimes. MIT Technology Review. https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes.
15. Zhu, X.; Cheng, D.; Li, H.; Zhang, K.; Hua, E.; Lv, X.; Ding, N.; Lin, Z.; Zheng, Z.; Zhou, B. How to Synthesize Text Data without Model Collapse? arXiv.org. https://arxiv.org/abs/2412.14689.
16. Yu, S.; Song, J.; Kim, H.; Lee, S.; Ryu, W.-J.; Yoon, S. Rare Tokens Degenerate All Tokens: Improving Neural Text Generation via Adaptive Gradient Gating for Rare Token Embeddings. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2022. https://doi.org/10.18653/v1/2022.acl-long.3.
17. PaperPile. Data Curation Pipeline. @emergentmind. https://www.emergentmind.com/topics/data-curation-pipeline.
18. Mattioli, L.; Ait-Hadichou, Y.; Chaouche, S.; Gonzalez, M. Data Curation Matters: Model Collapse and Spurious Shift Performance Prediction from Training on Uncurated Text Embeddings. Arxiv.org. https://arxiv.org/html/2506.17989v1.

Author

Murtaza S. Malik is an enthusiastic and extremely passionate research student fascinated with the field of AI. He aims to contribute to future AI policy and test his proposed solutions to model collapse as well as examining the economic and social effects of an AI-driven workforce.

DOI: 10.36838/IJHSR87.54