Characterizing Model Collapse in Large Language Models Using Semantic Networks and Next-Token Probability

Daniele Gambetta[1] , **Gizem Gezici**[2] , **Fosca Giannotti**[2] ,
Dino Pedreschi[1] , **Alistair Knott**[3] , **Luca Pappalardo**[2,4]
[1]University of Pisa, Pisa, Italy
[2]Scuola Normale Superiore, Pisa, Italy
[3]Victoria University, Wellington, New Zealand
[4]ISTI-CNR, Pisa, Italy
daniele.gambetta@phd.unipi.it, gizem.gezici@sns.it, fosca.giannotti@sns.it, dino.pedreschi@unipi.it,
ali.knott@vuw.ac.nz, luca.pappalardo@isti.cnr.it

Abstract

As synthetic content increasingly infiltrates the web, generative AI models may experience an autophagy process, where they are fine-tuned using their own outputs. This autophagy could lead to a phenomenon known as model collapse, which entails a degradation in the performance and diversity of generative AI models over successive generations. Recent studies have explored the emergence of model collapse across various generative AI models and types of data. However, the current characterizations of model collapse tend to be simplistic and lack comprehensive evaluation. In this article, we conduct a thorough investigation of model collapse across three text datasets, utilizing semantic networks to analyze text repetitiveness and diversity, while employing next-token probabilities to quantify the loss of diversity. We also examine how the proportions of synthetic tokens affect the severity of model collapse and perform cross-dataset evaluations to identify domain-specific variations. By proposing metrics and strategies for a more detailed assessment of model collapse, our study provides new insights for the development of robust generative AI systems.

1 Introduction

In recent years, generative AI has demonstrated remarkable advancements, particularly in conversational applications like ChatGPT, Google Gemini, Claude, and Llama [1]. Large generative models (LGMs) have achieved widespread public adoption and now contribute significantly to content generation across various online platforms [2].

Leading generative AI models, such as Meta's Llama-2, are trained on approximately two trillion tokens of human-generated text [3]. In comparison, the global pool of high-quality human-generated text is estimated at just 17 trillion tokens, with a modest annual growth rate of 4–5% [4]. A report by [5] predicts that by 2025, 90% of internet content will be AI-generated, heralding what has been described as the Age of Synthetic Realities [6]. As LGMs increasingly rely on web-sourced datasets, there is a growing chance that their outputs may be used to train subsequent versions of these models, potentially leading to a shortage of human-generated data [7].

Recent research has highlighted the risks associated with a self-consuming loop, often referred to as *autophagy* process, where LGMs are recursively fine-tuned on their own outputs ([8; 9]). Studies indicate that the autophagy process can lead to a phenomenon called *model collapse*, characterized by a significant loss of linguistic diversity in the content generated by LGMs (see, e.g., [10; 11; 12]). However, current characterizations of model collapse remain simplistic and lack comprehensive evaluations of how collapse manifests in nuanced ways. Existing metrics primarily focus on the diversity of generated text, with little attention paid to its semantic structure or the finer dynamics of collapse.

In this paper, we address this gap by providing a detailed characterization of model collapse during the autophagy process, focusing on three diverse text datasets to investigate how collapse varies across domains. To capture the complexity of model collapse, we adopt a dual approach: (1) semantic networks, which evaluate textual repetitiveness and semantic diversity, and (2) next-token probabilities, which quantitatively measure diversity loss. Furthermore, we investigate the impact of synthetic token proportions — the fraction of text generated by the model — on the severity of collapse. We also perform cross-dataset evaluations to explore whether collapse is more pronounced when models are tested on data from domains not used during fine-tuning. This cross-evaluation highlights the role of domain-specific characteristics in collapse and offers valuable insights for mitigating its effects.

In summary, our contributions extend the understanding of model collapse in several key ways:

- We provide a detailed analysis of model collapse during the autophagy process across three diverse text datasets, examining how collapse varies by domain.
- We use semantic networks to quantify text repetitiveness and semantic diversity, offering insights into the structural changes in generated text as collapse progresses.

- We introduce next-token probabilities and the concept of "collapsed prediction" as metrics to precisely measure diversity loss and track the evolving probability distributions of model predictions.
- We analyze the effect of varying proportions of synthetic tokens (25%, 50%, and 75%) on model collapse severity, highlighting how the amount of generated text influences degradation.
- We investigate collapse different across domains, testing whether models collapse more or less when applied to datasets not used for fine-tuning.

Our paper advances the understanding of the model collapse phenomenon in LGMs and supports the development of more robust generative AI systems.

The structure of this paper is organized as follows. In Section 2, we review the existing literature on the autophagy process and the phenomenon of model collapse. Section 3 introduces the autophagy framework and explains our implementation approach. In Section 4, we present the results of our experiments, and Section 4 summarizes the paper, highlighting the study's limitations and potential directions for future research.

2 Related Work

Research on model collapse relies on simulations, as conducting large-scale empirical studies with real users interacting with generative AI platforms is challenging [8; 9].

Recent studies show that model collapse occurs in both text and image domains, affecting various generative models including LGMs, Variational Autoencoders (VAEs) and Gaussian Mixture Models (GMMs) [10; 12; 13; 14; 11; 15; 16; 17; 18; 19].

[11] investigate three autophagous loops based on synthetic or real data availability during training: fully synthetic loop, synthetic augmentation, and fresh data loop. The study shows that, without fresh real data in each generation, future models experience model collapse. Using fixed real training data may delay this effect, but it cannot fully prevent it.

[12] investigate metrics to detect lexical, syntactic, and semantic diversity across model generations. They examine three use cases – news summarization, scientific abstract generation, and story generation – and find that the decline in linguistic diversity is more significant in high-entropy tasks (i.e., those requiring more creativity).

[16] train a generative model on an equal mix of real and self-generated data. Using precision and recall to assess the quality and similarity of generated images, they observe increasing similarity to real data and decreasing precision and recall. In another study, [15] replicate the experiments in the context of image generation using denoising diffusion implicit models. They augment the original dataset with AI-generated images and train a new model within an autophagous loop. The study finds that data augmentation leads to a decline in the quality of subsequent images.

[13] compare a full synthetic data cycle, where training data is entirely replaced in each generation, with data augmentation cycles that vary the proportion of real and synthetic data (balanced, incremental, and expanding). The full synthetic data cycle leads to model collapse, reducing diversity to a single point. The incremental and balanced cycles also decrease diversity, with the former showing a more substantial effect. The expanding data cycle adds new data to the previous dataset and maintains diversity without declining up to 50 simulation generation steps.

[18] investigate a text-to-image Stable Diffusion model, examining how different distributions of generated images impact task performance. The generated images negatively affect performance, with the level of model degradation varying depending on the extent of contamination. The researchers emphasize the need for methods to detect AI-generated images, proposing a self-supervised learning approach using a masked autoencoder for watermarking.

[19] use a stable diffusion model and conduct five iterations of data generation, retraining the model at each step with different proportions of real and synthetic data. Model collapse emerges but can be healed by retraining the model exclusively on real images.

[17; 14] conduct experiments with Llama2 trained on a mixture of real and AI-generated data, showing that performance improves initially but eventually leads to model collapse.

[20] develop a theoretical framework to study the iterative retraining of generative models on mixed datasets containing both real and synthetic data. The iterative retraining remains stable when the initial generative model is sufficiently close to the real data distribution and when the proportion of real data is large enough.

[21] demonstrate that the model collapse phenomenon is unavoidable when the model is trained solely on synthetic data. They theoretically and empirically explore the maximum amount of synthetic data that can be used without leading to model collapse.

[22] use a pre-trained GPT-2 model within an autophagous loop, terminating the simulation after a maximum of 1,000 simulation steps. Model collapse happens more quickly with higher learning rates and for models with a larger number of parameters.

[23] explore the impact of accumulated generated content over generations through an analytical framework and empirical experiments using GPT2, GPT3, Llama2 and diffusion models. They find that accumulating data can help prevent model collapse, supporting the findings of [13].

[24] investigate the use of verification of synthetic data to prevent model collapse, using linear classifiers to assess whether a selection of verified synthetic data could improve model performance. They find that these verifiers can indeed help prevent model collapse.

[25] examines the speed of model collapse, finding that the time required to forget a word is linearly related to its frequency in the original corpus.

[26] focus on the impact of synthetic data on model training and how to synthesize data without model collapse. They show that token editing on human-made data to obtain semi-synthetic data may prevent model collapse.

Contribution of our work. Prior studies have primarily focused on identifying model collapse and investigating the effects of synthetic data on model performance. However, their characterization of model collapse remains overly simplistic, relying predominantly on linguistic diversity metrics, which capture only a limited aspect of the phenomenon. We go beyond these limitations by introducing a more nuanced approach: we describe generated documents using semantic networks, which reveal patterns of textual repetitiveness and semantic structure, and analyze how probability distributions evolve as models approach collapse. These methods provide a deeper, more comprehensive understanding of the mechanisms driving model collapse.

3 Autophagy Simulation Framework

Our autophagy pipeline, inspired by [10], is based on a foundation model M and a dataset D of $n = 1,000$ documents. We truncate each document in D to k=64 tokens to form a set P of n prompts. These prompts are used to ask M for text completion, generating up to $128 - k$ additional tokens. The choice of 128 tokens aligns with the approach in [10]. This process creates a new dataset, D_0, consisting of n documents with a max length of 128 tokens. Here is an example of prompt (in italic) and text generated by the model in response to it (in bold):

> *Pre-trained models usually come with a pre-defined tokenization and little flexibility as to what subword tokens can be used in downstream tasks. This problem concerns especially multilingual NLP and low-resource languages, which are typically processed using cross-lingual transfer. In this paper, we aim to find out if the right granularity of tokenization is helpful for a text classification task, namely dialect classification. Aiming at* **finding a right granularity of tokenization, we propose a novel approach to the dialect classification task, which allows the user to decide what subword units should be used for the task. We introduce a novel methodology for tokenization that allows the user to choose the granularity of the tokenization.**

Next, we fine-tune M using documents in D_0, producing a new model, M_1. We ask M_1 to perform text completion from prompts in P. This results in a new dataset, D_1.

This autophagy process is repeated ten times, generating a series of models — $M_1, \ldots, M_{10}$ —and their corresponding datasets — $D_1, \ldots, D_{10}$.

In the rest of the paper, we denote $M_i^{(D)}$ as the foundation model after i fine-tunings (the model at the i-th generation). This term refers to the model obtained by fine-tuning M_{i-1} on the dataset D_{i-1}.

3.1 Foundation Model and Datasets

We select Meta's Llama2-7b as the foundation model because it is publicly available and has strong performance in various NLP tasks. The 7b version, with seven billion parameters, offers a good balance between model quality and computational efficiency, enabling both fast fine-tuning and robust results. In the experiments, we use the Unsloth library[1] to fine-tune Llama-2-7b with default hyperparameters on GPU card NVIDIA Quadro RTX 6000.

We select the three text datasets used by [12]:

- **Wikitext103** (`wiki`) is a dataset of over 100 million tokens extracted from verified English Wikipedia articles [27]. From the dataset, available on HuggingFace[2], we use the article body text.
- **XLsum** (`xls`) contains 1.35 million annotated news articles from the BBC in multiple languages [28]. Each entry includes the title, article body, summary, and article URL. We focus on the English-language subset and use the article's body text.
- **SciAbs** (`sci`) is derived from a BiBTeX bibliography database of papers published in computational linguistics and NLP venues since 1965 [12]. The dataset contains over 40,000 papers, from which use the text in the associated abstracts.

3.2 Measures of model collapse

Linguistic entropy

Linguistic entropy quantifies uncertainty, unpredictability, or information content within a text document. Low linguistic entropy indicates a repetitive vocabulary and low information density. Given a text document T and the set of its unique terms $\mathcal{T}$, we define linguistic entropy as:

$$H(T) = -\frac{\sum_{t \in \mathcal{T}} p_t \log(p_t)}{\log |\mathcal{T}|} \qquad (1)$$

where p_t is the probability of each unique term $t \in \mathcal{T}$, i.e., the frequency of t in T. The normalisation factor, $\log |\mathcal{T}|$, is added to make the linguistic entropy comparable across documents of different lengths.

Commonsensical inference

We assess the foundational model's capability to tackle the commonsense natural language inference (NLI) task, which evaluates how well a model can complete a given text in a plausible manner. For this evaluation, we utilize the HellaSwag dataset [29], which consists of 70,000 sentences, each accompanied by four possible completions. Out of these options, only one completion is a valid and commonsensical continuation, while the other three are nonsensical. The model's task is to select the correct continuation by calculating the conditional probability of each option. To measure the model's accuracy, we determine which completion has the highest probability and calculate the proportion of sentences that the model completes correctly.

[1] https://github.com/unslothai/unsloth

[2] https://huggingface.co/datasets/Salesforce/wikitext

Semantic networks

Semantic networks allow for the analysis of the structure of a text document [30]. In a semantic network, nodes represent nouns and verbs, while edges indicate the co-occurrence of nodes within the same sentence. Nouns and verbs are identified using the POS-tagging methods provided by the Spacy library[3]. From these semantic networks, we compute the number of nodes and edges, network density (ratio of actual edges to all possible edges) and the number of connected components.

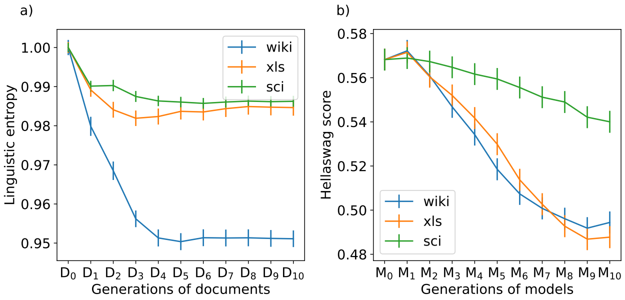

Figure 1: Linguistic entropy (a) and Hellaswag score (b) evaluated for Llama2-7b fine-tuned across 10 generations within an autophagy pipeline. The results are shown for three datasets: `wiki` (blue), `xls` (orange), and `sci` (green). In these experiments, the parameter k was set to 64, meaning the model was instructed to generate up to 64 tokens.

Next-token probability

Next-token probability quantifies the likelihood of a token being the next in a sequence, based on the preceding context. It reflects the foundation model's confidence in its prediction, with higher probabilities indicating greater certainty. Given a model $M_i^{(D)}$, we randomly select 1,000 documents from D among those not used for fine-tuning $M_i^{(D)}$ and truncate each to the first 25 tokens. These truncated documents form the prompts used to ask $M_i^{(D)}$ to predict the next token. We then extract $M_i^{(D)}$'s probability distribution for the top 100 most likely tokens. Over these distributions, we compute the Gini coefficient, which quantifies the distribution's inequality (the higher the more unequal), and the number of collapsed predictions, i.e., how many tokens have probability above 0.999.

4 Results

Our findings show that linguistic diversity declines over generations across all datasets (see Figure 1a), aligning with previous studies on model collapse [10]. Similarly, the model's ability to provide commonsensical answers diminishes over generations for all datasets (see Figure 1b). This means the foundation model increasingly produces answers that deviate from common sense or lack logical coherence, reflecting a decline in generating text that humans would consider reasonable.

Both linguistic entropy and commonsensical inference suggest that the foundational model experiences the highest collapse when fine-tuned on the `sci` dataset (abstracts of scientific papers). In contrast, the loss of linguistic diversity is most pronounced for the `wiki` dataset (Wikipedia articles).

Table 1 provides examples of how model collapse manifests in the generated text. As the generations advance, the model tends to repeat the same sentence more frequently. The repeated text is often brief, lacking diversity and common sense. Consistent with the findings presented in Figure 1, the `wiki` dataset demonstrates the most significant collapse.

The analysis of semantic networks provides an additional characterization of the model collapse phenomenon. Figure 2 presents the semantic networks derived from model-generated part in the documents of Table 1. Model collapse becomes apparent through a decreasing number of nodes as generations progress. At generation 10, the `wiki` network collapses into a network of two nodes, while in the `sci` network every node becomes connected to every other.

Figure 3 generalizes these findings across 1,000 documents. As fine-tuning proceeds over multiple generations, the semantic networks have fewer nodes, edges, and connected components, and a higher network density. These changes indicate that the text gradually relies on fewer unique concepts, which become increasingly interlinked. A drop in the number of nodes and edges means a reduction in lexical and topical variety, while the rise in density shows that the smaller set of remaining concepts often co-occur. Fewer connected components suggests that the network consolidates into larger clusters of repeating tokens, reflecting a loss of semantic diversity and a more repetitive text. The variations are consistent across all three datasets, though with different magnitudes.

The reduction in semantic diversity and the rise of repetitive text is also evident when analysing the variation of next-token probability distributions. Figure 4 shows the probabilities at generations 0, 5, and 10 of the top five tokens predicted by the foundation model fine-tuned on the `sci` dataset when generating text in response to the following prompt:

> "Automatic text summarization has enjoyed great progress over the years and is used in numerous applications, impacting the lives of many. Despite this development, there . . . "

We observe a clear trend: the probability distribution among the top five tokens becomes increasingly imbalanced in generations 5 and 10. In generation 0, the token "is" had the highest probability. By generation 10, this token dominates the others, underlining a significant reduction in language variety.

We extend this experiment to a set P' consisting of 1,000 prompts taken from each dataset. The prompts in P' were not used to fine-tune the model ($P' \not\subset P$). The probabilities of the first tokens generated by the model in response to the prompts are visualized as dots in Figure 5, along with a histogram representing the probability distribution. We find that, as generations progress, the distribution shifts increasingly toward a concentration of points with high probabili-

[3] https://spacy.io/

	`wiki`	`xls`	`sci`
Generation 0	The Church of St George is a medieval Eastern Orthodox church in the city of Kyustendil, which lies in southwestern Bulgaria and is the administrative capital of Kyustendil Province . The church is located in the Kolusha neighbourhood , which was historically separate from the city. The **church is situated on the eastern side of the city , at the foot of the Balkan Mountains . sierp 2011 the church was declared a cultural monument of national importance . The church is a single-nave structure with a semi-circular apse , with a bell tower above the**	At about 19:30 on Friday the fire alarm went off. I went out of the room and saw other guests milling about, and no one seemed to be really concerned. So I went back to my room to get ready to leave for the airport. I went downstairs to settle the bill around 20:30 and it was like a scene out of a movie. **I was in the lobby with the receptionist, and the firemen were running down the stairs with their masks on and their hoses. They were shouting, 'Get out, get out!' I didn't know what was going on. I thought maybe they were shooting a movie.**	The reliance of deep learning algorithms on large scale datasets represents a significant challenge when learning from low resource sign language datasets. This challenge is compounded when we consider that, for a model to be effective in the real world, it must not only learn the variations of a given sign, but also learn to be invariant to the person signing. In this paper, **we present a new approach to addressing these challenges, by introducing a novel loss function, which we call the "Mixed Pairwise Loss", that can be applied to both the training and testing of deep learning models. We present a number of experiments that demonstrate the effectiveness of the proposed method.**
Generation 5	The Church of St George is a medieval Eastern Orthodox church in the city of Kyustendil , which lies in southwestern Bulgaria and is the administrative capital of Kyustendil Province . The church is located in the Kolusha neighbourhood , which was historically separate from the city . The **sierp 2020 The church is a The church is a The church is a The church is a The church is a The church is a The church is a The church is a The church is a The church is a¡/s¿**	At about 19:30 on Friday the fire alarm went off. I went out of the room and saw other guests milling about, and no one seemed to be really concerned. So I went back to my room to get ready to leave for the airport. I went downstairs to settle the bill around 20:30 and it was like a scene out of a movie. **”The place was filled with smoke and the fire was really big. I was the only guest in the hotel and I was the only one who was able to get out. I was the only one who was able to get out. I was the only one who was able to get out.¡/s¿**	The reliance of deep learning algorithms on large scale datasets represents a significant challenge when learning from low resource sign language datasets. This challenge is compounded when we consider that, for a model to be effective in the real world, it must not only learn the variations of a given sign, but also learn to be invariant to the person signing. In this paper, **we propose a novel methodology for learning sign language from a low resource dataset. We propose a novel methodology for learning sign language from a low resource dataset. We propose a novel methodology for learning sign language from a low The reliance of deep learning algorithms on large scale¡/s¿**
Generation 10	The Church of St George is a medieval Eastern Orthodox church in the city of Kyustendil , which lies in southwestern Bulgaria and is the administrative capital of Kyustendil Province . The church is located in the Kolusha neighbourhood , which was historically separate from the city . The **sierp 2020. The church is a The church is a The church is a The church is a The church is a The church is a The church is a The church is a The church is a The church is a¡/s¿**	At about 19:30 on Friday the fire alarm went off. I went out of the room and saw other guests milling about, and no one seemed to be really concerned. So I went back to my room to get ready to leave for the airport. I went downstairs to settle the bill around 20:30 and it was like a scene out of a movie. **”The place was like a war zone. The place was like a war zone. The place was like a war zone. The place was like a war zone. The place was like a war zone. The place was like a war zone. The place was like a war¡/s¿**	The reliance of deep learning algorithms on large scale datasets represents a significant challenge when learning from low resource sign language datasets. This challenge is compounded when we consider that, for a model to be effective in the real world, it must not only learn the variations of a given sign, but also learn to be invariant to the person signing. In this paper, **we propose a novel methodology for learning sign language from a low resource dataset. We propose a novel methodology for learning sign language from a low resource dataset. We propose a novel methodology for learning sign language from a low resource dataset. We propose a novel methodology for learning¡/s¿**

Table 1: Example of documents generated by Llama2-7b fine-tuned across 10 generations within an autophagy pipeline. The results are shown for three generations (0, 5, and 10) and for the `wiki`, `xls` and `sci` datasets. The generated text is highlighted in bold.

ties, indicating greater predictability of the next token and a corresponding reduction in linguistic variety. The degree of this reduction depends on the characteristics of the datasets. For instance, the `wiki` dataset exhibits a probability distribution skewed toward high values already in generation 0, causing the model to collapse more rapidly compared to the other datasets. The ranking of the top ten most probable tokens remains stable across generations, as shown in Figures 8 and 9 in the Appendix. Among the datasets, `wiki` demonstrates the greatest stability, while `sci` shows the least stability.

We also investigate how model collapse affects the model's ability to complete documents taken from *external datasets* that were not used for fine-tuning. For example, we test the foundation model fine-tuned on the `wiki` dataset using the external datasets `xls` and `sci`. Similarly, we perform cross-evaluations for models fine-tuned on the other two datasets.

Figure 6a-c-e illustrates the inequality of next-token probability distribution as generations progress, quantified by the Gini coefficient. This analysis considers all possible combinations of fine-tuning and external datasets. We find that

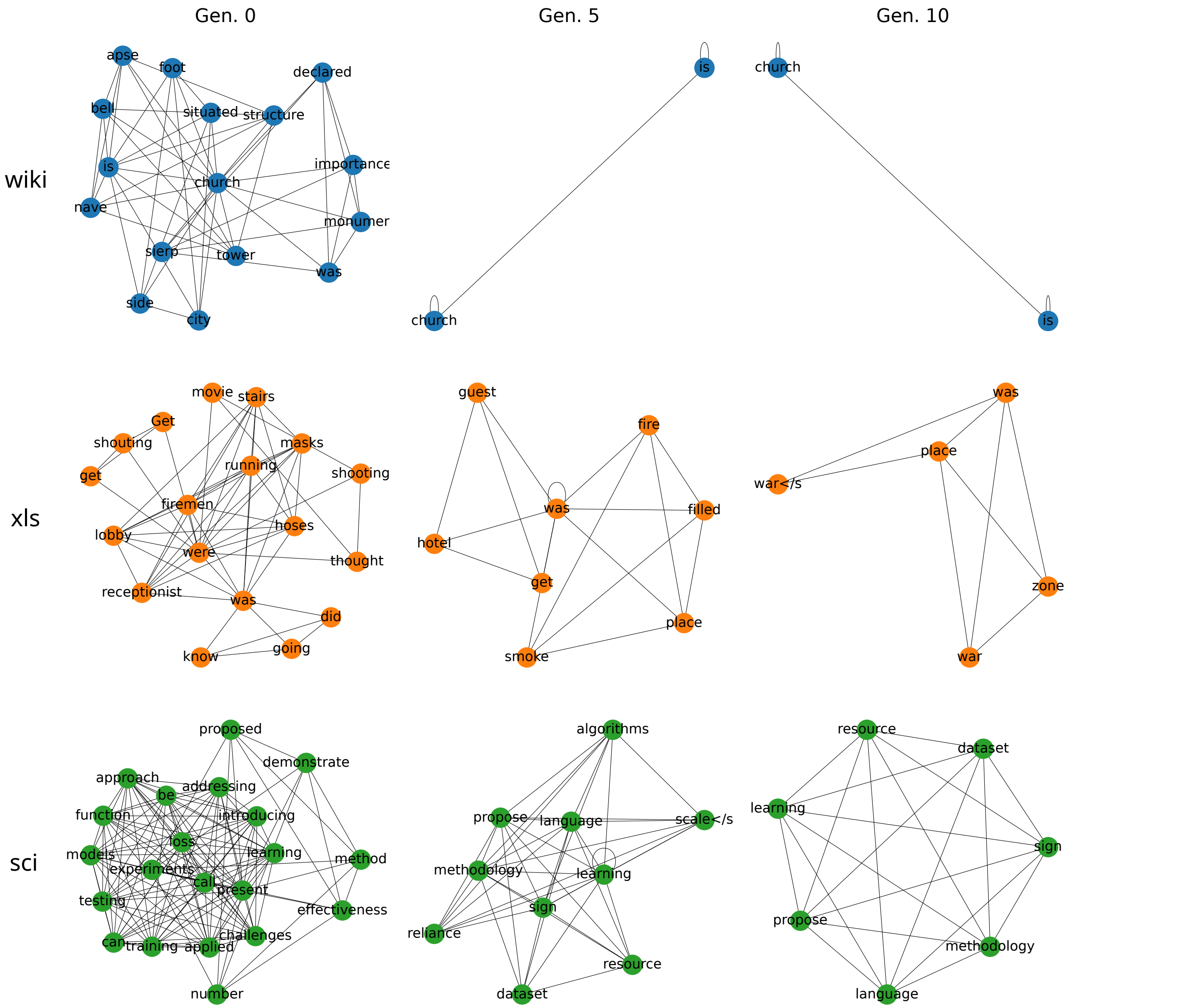

Figure 2: Semantic networks of documents in Table 1. The number of nodes (tokens) decreases over generations. At generation 10, the `wiki` dataset collapses into a network of two nodes ("is" and "church") while the `sci` network is fully connected. Both cases are clear signals of model collapse.

the foundation model shows less severe collapse when evaluated on external datasets. Indeed, the Gini coefficient reflects greater inequality when models are tasked with completing documents from the same dataset used for fine-tuning.

This outcome is supported by the rising number of "collapsed predictions" as generations progress. A collapsed prediction occurs when the probability of the top predicted token exceeds 0.999. Our results reveal that the foundational model is more likely to produce collapsed predictions when evaluated on the fine-tuning dataset, as opposed to external datasets (see Figure 6b-d-f). The `wiki` dataset results in the most significant increase in collapsed predictions over generations when used for fine-tuning. In contrast, the `sci` dataset shows the fewest collapsed predictions when used as an external dataset. This indicates that completing scientific abstracts is a task that is less prone to model collapse.

The results presented so far correspond to $k = 64$, meaning the foundational model is tasked with generating up to 64 synthetic tokens to complete the document. We now vary $k = 32, 64, 96$, which correspond to 25%, 50% and 75% of

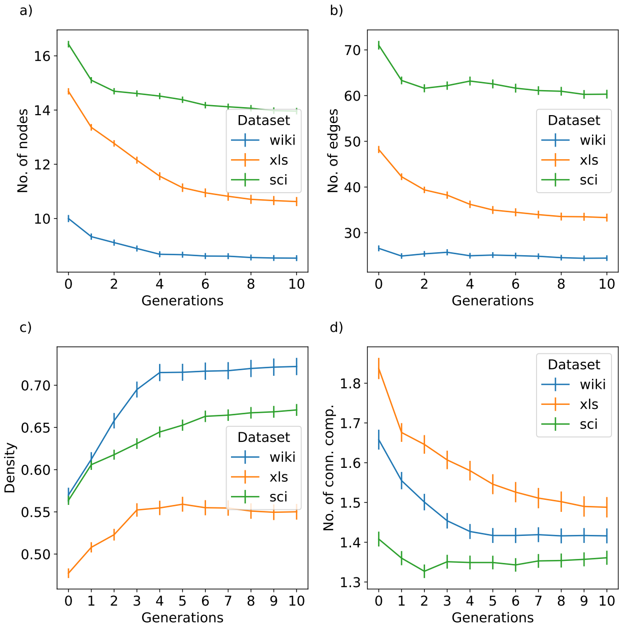

Figure 3: Characteristics of documents' semantic networks over ten generations for `wiki`, `xls`, and `sci`. (a) Average number of nodes. (b) Average number of edges. (c) Average network density (edges per possible edges). (d) Average number of connected components.

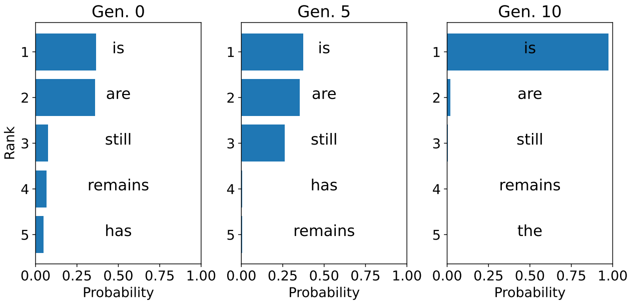

Figure 4: Probabilities (x-axis) of the top-5 tokens (y-axis) predicted by Llama2-7b fine-tuned across ten generations within an autophagy pipeline. The results are shown for the `sci` dataset and for generations 0, 5, and 10. The token predictions follow the prompt: *"Automatic text summarization has enjoyed great progress over the years and is used in numerous applications, impacting the lives of many. Despite this development, there . . . "*

the synthetic tokens out of a total of 128.

Our analysis reveals that as the value of k increases, the number of collapsed predictions also rises for all datasets (see Figure 7). This observation indicates that when the model must generate a larger portion of each document, it relies more heavily on its own synthetic outputs, leading to a higher likelihood of repeating or overestimating certain tokens. In other words, increasing the fraction of synthetic tokens accelerates the autophagy process in which the model feeds on its own generated content, intensifying the loss of diversity and predictability that characterizes model collapse.

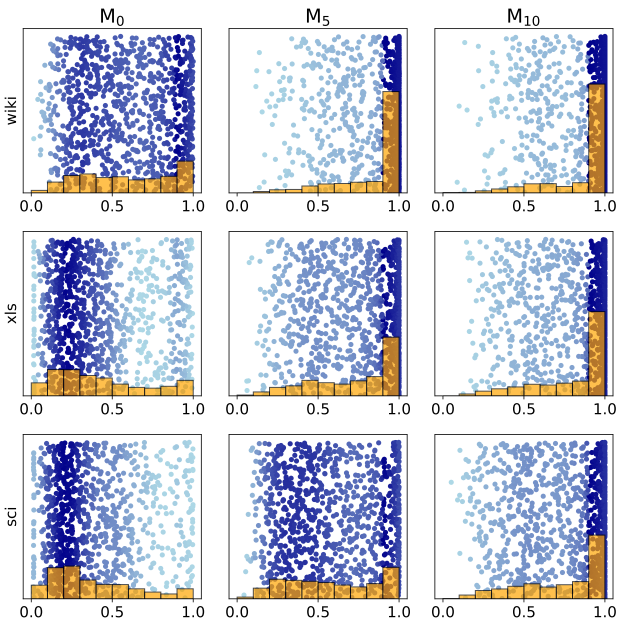

Figure 5: Probabilities of the top suggested token for 1,000 documents generated by the Llama2-7b model fine-tuned on `wiki`, `xls`, and `sci`, and for generations 0, 5, and 10. Each dot represents the probability of the first token generated by the model in response to the prompt. The bars show the distribution of these probabilities.

5 Conclusions and future studies

In this study, we provided a detailed characterization of the model collapse phenomenon associated with the autophagy process introduced by [10]. Our analysis, conducted on three text datasets, highlighted how examining semantic networks and next-token probabilities can offer valuable insights into the nature of model collapse. This approach goes beyond the initial and simplistic assessments of model collapse typically discussed in the literature ([8]).

One limitation of this study is its exclusive reliance on greedy search for decoding, which chooses the next word based solely on the highest probability. This approach may overlook high-probability words that are overshadowed by those with lower probabilities. To address this issue, we recommend exploring beam search in future work.

Our findings may serve as the basis for designing a strategy to prevent the emergence of model collapse or mitigate its effects by recalibrating next-token probabilities. This strategy could allow the foundation model to recover behaviour akin to its original, non-collapsed state without requiring fine-tuning with actual data.

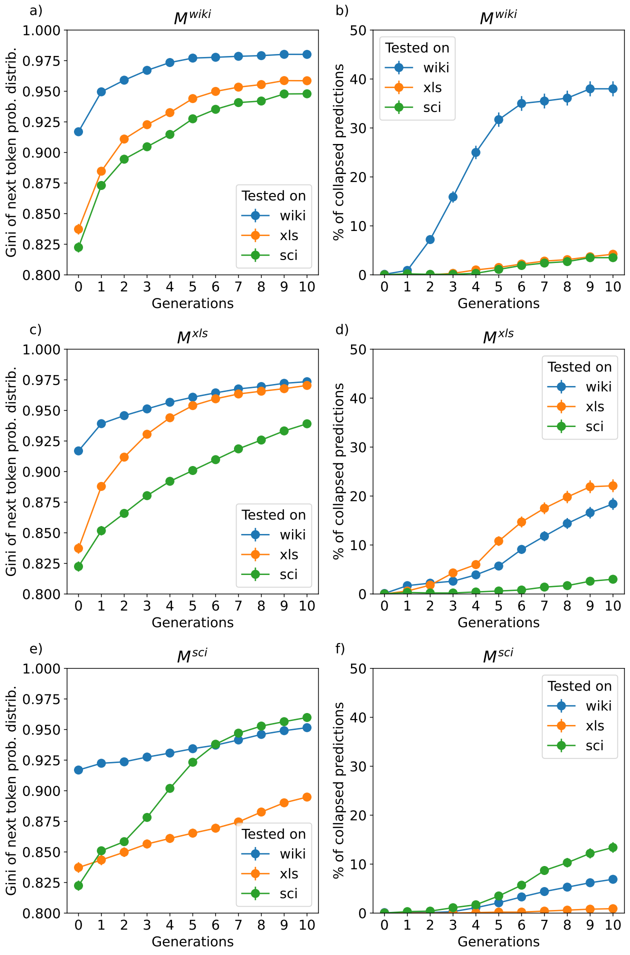

Figure 6: Gini coefficient of next-token probability distribution (a, c, e) and proportion of collapsed predictions (b, d, f) across generations various combination of fine-tuning and external datasets.

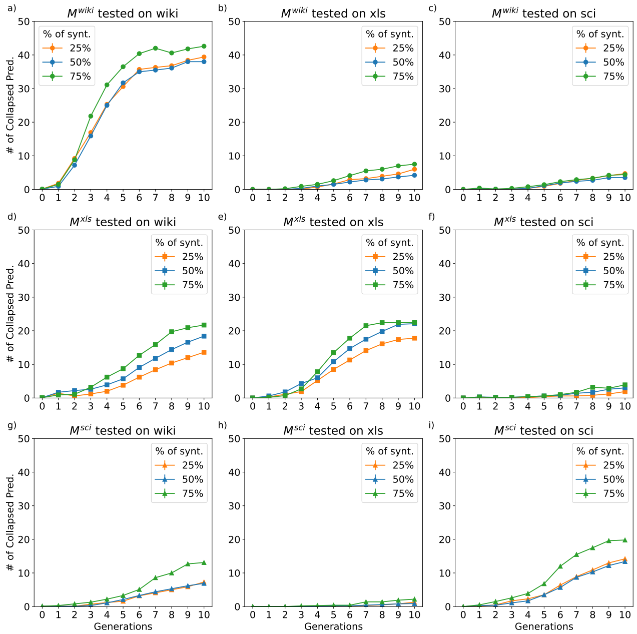

Figure 7: Percentage of collapsed predictions as generations progress, for the `wiki`, `xls`, and `sci` datasets. The figure shows results for different values of $k = 25\%, 50\%, 75\%$ and includes tests on prompts constructed from datasets not used for the model's fine-tuning.

References

[1] Wu J, Yang S, Zhan R, Yuan Y, Wong DF, Chao LS. A survey on llm-gernerated text detection: Necessity, methods, and future directions. arXiv preprint arXiv:231014724. 2023.

[2] Ghassemi M, Birhane A, Bilal M, Kankaria S, Malone C, Mollick E, et al. ChatGPT one year on: who is using it, how and why? Nature. 2023;624(7990):39-41.

[3] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288. 2023.

[4] Villalobos P, Sevilla J, Heim L, Besiroglu T, Hobbhahn M, Ho A. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. arXiv preprint arXiv:221104325. 2022.

[5] Europol Innovation Lab. Facing Reality: Law Enforcement and the Challenge of Deepfakes; 2021. Available at: https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepf.

[6] Cardenuto JP, Yang J, Padilha R, Wan R, Moreira D, Li H, et al.. The Age of Synthetic Realities: Challenges and Opportunities; 2023. Available from: https://arxiv.org/abs/2306.11503.

[7] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:210807258. 2021.

[8] Pappalardo L, Ferragina E, Citraro S, Cornacchia G, Nanni M, Rossetti G, et al. A survey on the impact of AI-based recommenders on human behaviours: methodologies, outcomes and future directions. arXiv preprint arXiv:240701630. 2024.

[9] Pedreschi D, Pappalardo L, Ferragina E, Baeza-Yates R, Barabási AL, Dignum F, et al. Human-AI coevolution. Artificial Intelligence. 2025;339:104244. Available from: https://www.sciencedirect.com/science/article/pii/S0004370224001802.

[10] Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y. AI models collapse when trained on recursively generated data. Nature. 2024;631(8022):755-9.

[11] Alemohammad S, Casco-Rodriguez J, Luzi L, Humayun AI, Babaei H, LeJeune D, et al.. Self-Consuming Generative Models Go MAD; 2023.

[12] Guo Y, Shang G, Vazirgiannis M, Clavel C. The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text; 2023.

[13] Briesch M, Sobania D, Rothlauf F. Large language models suffer from their own output: An analysis of the self-consuming training loop. arXiv preprint arXiv:231116822. 2023.

[14] Dohmatob E, Feng Y, Yang P, Charton F, Kempe J. A Tale of Tails: Model Collapse as a Change of Scaling Laws. arXiv preprint arXiv:240207043. 2024.

[15] Martínez G, Watson L, Reviriego P, Hernández JA, Juarez M, Sarkar R. Combining Generative Artificial Intelligence (AI) and the Internet: Heading towards Evolution or Degradation?; 2023.

[16] Martínez G, Watson L, Reviriego P, Hernández JA, Juarez M, Sarkar R. Towards understanding the interplay of generative artificial intelligence and the internet. arXiv preprint arXiv:230606130. 2023.

[17] Dohmatob E, Feng Y, Kempe J. Model Collapse Demystified: The Case of Regression. arXiv preprint arXiv:240207712. 2024.

[18] Hataya R, Bao H, Arai H. Will Large-scale Generative Models Corrupt Future Datasets? In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023. p. 20555-65.

[19] Bohacek M, Farid H. Nepotistically Trained Generative-AI Models Collapse. arXiv preprint arXiv:231112202. 2023.

[20] Bertrand Q, Bose AJ, Duplessis A, Jiralerspong M, Gidel G. On the Stability of Iterative Retraining of Generative Models on their own Data; 2024.

[21] Seddik MEA, Chen SW, Hayou S, Youssef P, Debbah M. How Bad is Training on Synthetic Data? A Statistical Analysis of Language Model Collapse; 2024.

[22] Herel D, Mikolov T. Collapse of Self-trained Language Models; 2024.

[23] Gerstgrasser M, Schaeffer R, Dey A, Rafailov R, Sleight H, Hughes J, et al.. Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data; 2024.

[24] Feng Y, Dohmatob E, Yang P, Charton F, Kempe J. Beyond Model Collapse: Scaling Up with Synthesized Data Requires Verification; 2024. Available from: https://arxiv.org/abs/2406.07515.

[25] Suresh AT, Thangaraj A, Khandavally ANK. Rate of Model Collapse in Recursive Training; 2024. Available from: https://arxiv.org/abs/2412.17646.

[26] Zhu X, Cheng D, Li H, Zhang K, Hua E, Lv X, et al.. How to Synthesize Text Data without Model Collapse?; 2024. Available from: https://arxiv.org/abs/2412.14689.

[27] Merity S, Xiong C, Bradbury J, Socher R. Pointer Sentinel Mixture Models; 2016. Available from: https://arxiv.org/abs/1609.07843.

[28] Hasan T, Bhattacharjee A, Islam MS, Mubasshir K, Li YF, Kang YB, et al. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. In: Zong C, Xia F, Li W, Navigli R, editors. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online: Association for Computational Linguistics; 2021. p. 4693-703. Available from: https://aclanthology.org/2021.findings-acl.413/.

[29] Zellers R, Holtzman A, Bisk Y, Farhadi A, Choi Y. HellaSwag: Can a Machine Really Finish Your Sentence?; 2019. Available from: https://arxiv.org/abs/1905.07830.

[30] Borge-Holthoefer J, Arenas A. Semantic Networks: Structure and Dynamics. Entropy. 2010;12(5):1264-302. Available from: https://www.mdpi.com/1099-4300/12/5/1264.

Appendix

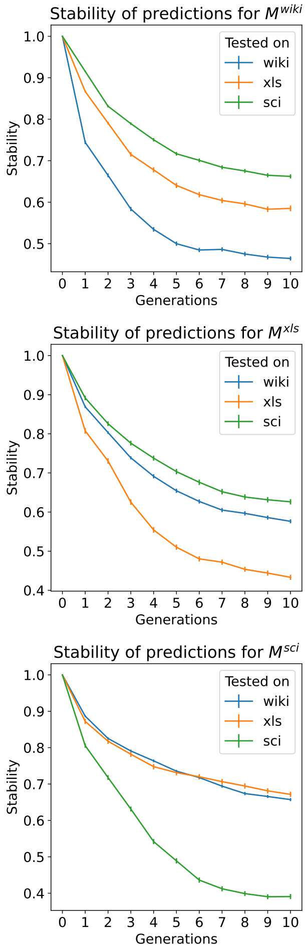

Figure 8: The stability of the top 10 tokens in next-token predictions across generations was analyzed for different pipelines. Stability measures how consistently the top 10 tokens and their relative rankings are maintained compared to the original model M, with higher values indicating greater consistency. The metric quantifies shifts in token rankings as the model evolves. Each plot shows the average stability for various fine-tuning datasets, evaluated on the corresponding dataset.

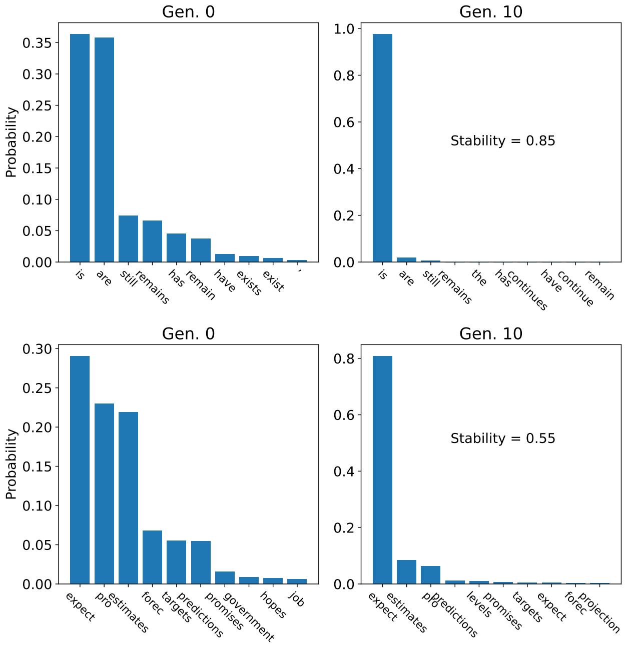

Figure 9: Two examples compare top 10 token stability between the original model (M) and generation 10 (M_{10}). The first example shows high stability (score 0.85), while the second indicates moderate stability (score 0.55). By generation 5, average stability across pipelines is around 0.5, suggesting it as a key intervention point before stability declines further, making recovery difficult.