Position: Model Collapse Does Not Mean What You Think

Anonymous Author(s)
Affiliation
Address
email

Abstract

The proliferation of AI-generated content online has fueled concerns over *model collapse*, a degradation in future generative models' performance when trained on synthetic data generated by earlier models. Industry leaders, premier research journals and popular science publications alike have prophesied catastrophic societal consequences stemming from model collapse. In this position piece, we contend this widespread narrative fundamentally misunderstands the scientific evidence. We highlight that research on model collapse actually encompasses eight distinct and at times conflicting definitions of model collapse, and argue that inconsistent terminology within and between papers has hindered building a comprehensive understanding of model collapse. To assess how significantly different interpretations of model collapse threaten future generative models, we posit what we believe are realistic conditions for studying model collapse and then conduct a rigorous assessment of the literature's methodologies through this lens. While we leave room for reasonable disagreement, our analysis of research studies, weighted by how faithfully each study matches real-world conditions, leads us to conclude that certain predicted claims of model collapse rely on assumptions and conditions that poorly match real-world conditions, and in fact several prominent collapse scenarios are readily avoidable. Altogether, this position paper argues that model collapse has been warped from a nuanced multifaceted consideration into an oversimplified threat, and that the evidence suggests specific harms more likely under society's current trajectory have received disproportionately less attention.

1 Introduction

The rapid surge of AI-generated content has sparked intense debate about potential ramifications of training future generative AI models on datasets containing synthetic data generated by previous models. One especially concerning prediction is *model collapse*, a phenomenon whereby future generative models fail due to being trained on synthetic data. Model collapse has captured attention at the highest levels of academia and industry: *Nature* prominently featured model collapse (Gibney, 2024) alongside accompanying research stating that AI models trained on synthetic data would suffer catastrophic degradation in performance (Shumailov et al., 2024), while prominent science and news outlets like *Scientific American* and the *Wall Street Journal* amplified these concerns, writing "a training diet of AI-generated text, even in small quantities, eventually becomes poisonous to the model being trained." (Rao, 2023) and that "feeding a model text that is itself generated by AI is considered the computer-science version of inbreeding" (Seetharaman, 2024). Some industry leaders have highlighted model collapse as a critical challenge for the future of AI (Wang, 2024).

In this position piece, we argue this widespread narrative of a bleak model-collapsed future, filled with polluted pretraining data and useless generative models, oversimplifies or misinter-

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

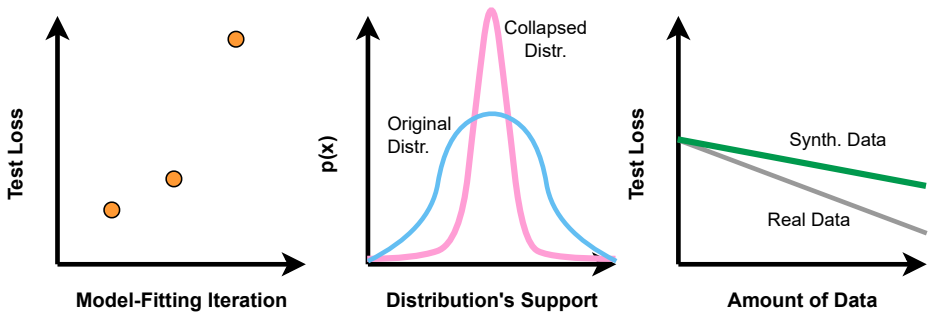

Figure 1: **Model Collapse Has Been Defined in Multiple and Sometimes Conflicting Ways.** By hand-annotating 28 prior research publications, we identify 8 definitions of model collapse (Sec. 2). The 8 definitions can be loosely grouped into three families: (1) the behavior of the test loss on real data over model-fitting iterations (left), (2) the deformation of the real data distribution over model-fitting iterations (center) and (3) the scaling behavior of the test loss with respect to typical scaling quantities such as the amount of data (right).

prets the precise scientific claims and their underlying assumptions and mechanisms. Through careful analysis, we identify three critical gaps between the prevailing discourse and research reality:

First, we reveal that the term “model collapse" encompasses eight definitions of performance degradation in model-data feedback loops. This multiplicity of definitions, used inconsistently between papers and at times inconsistently within papers, has resulted in papers talking past one another, thereby hindering development of a comprehensive understanding of likely futures for frontier deep generative models. We argue specific failure modes should be explicitly identified and discussed alongside comparable results.

Second, we posit trends that we believe faithfully describe common practices of leading AI labs pretraining frontier AI systems on web-scale data: increasing compute, improving data quality, and expanding datasets of real and synthetic data. One key point that we emphasize here is that many prominent model collapse papers assume data are entirely deleted after each model-fitting iteration and that subsequent models are trained entirely on synthetic data generated by their predecessors, which we argue is not realistic.

Third, we weigh different results in the model collapse literature based on the plausibility of their assumptions along multiple axes of consideration to assess which notions of model collapse pose significant and likely threats to future frontier AI models. We argue that some definitions of model collapse do not correspond to catastrophic outcomes under our realistic assumptions, and most concerning collapse predictions emerge from implausible experimental setups. However, there are very real threats to tails of the data distribution that should be taken seriously.

Altogether we argue that model collapse has been inflated from a precise and important technical consideration into a mischaracterized and overstated threat. While synthetic data poses genuine challenges that warrant careful study, our analysis reveals that the most widely stated collapse scenarios can be avoided through standard ongoing practices in model development and dataset curation. Instead of worrying about unrealistic catastrophic notions of model collapse, by adopting a more realistic perspective, we can re-orient to focus on the real issues of diversity collapse happening now (Zhang et al., 2024; Padmakumar & He, 2024; Murthy et al., 2024; Wu et al., 2024). This position paper aims to clarify the scientific discourse around model collapse, propose best practices for future work on the subject, and redirect research attention towards understanding how to generate and curate synthetic data that improves future frontier AI systems while mitigating failure modes.

2 Definitions of Model Collapse

We begin with a non-obvious but critical point: the model collapse literature has at least eight different definitions based on different notions of model performance degradation. As evidence, we

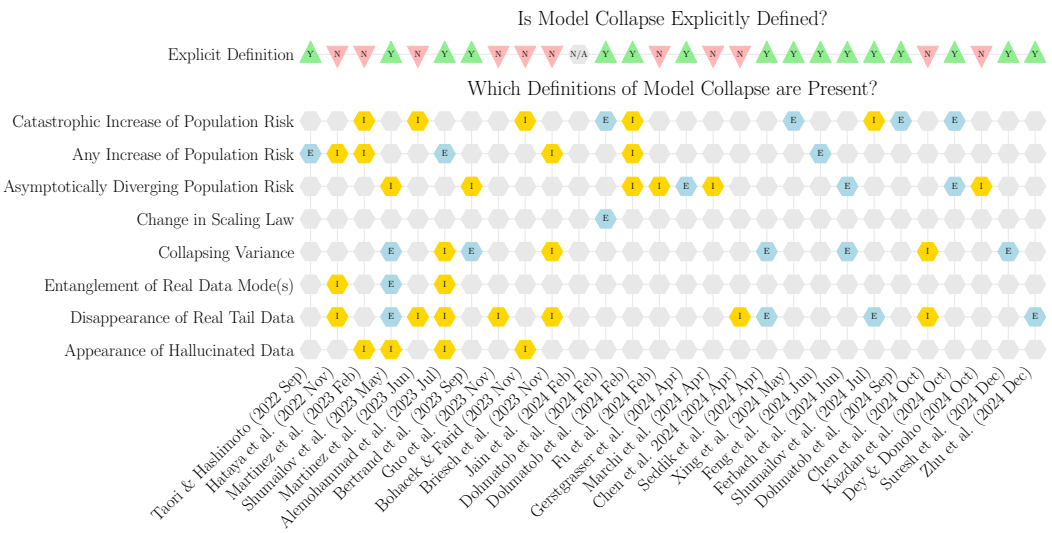

Figure 2: **Model Collapse Has Been Defined in Multiple and Sometimes Conflicting Ways.** We conduct a meta-analysis of research papers on model collapse. Top: We identify which papers offer *any* explicit definition of model collapse (**Yes** (Y) or **No** (N)). Bottom: We identify which definition(s) of model collapse each paper uses for its empirical and/or mathematical results, either **explicitly** (E) or **implicitly** (I). Our annotations reveal that model collapse research is based on multiple definitions that we will show sometimes conflict between papers and even within individual papers.

hand-annotated twenty-eight prominent prior research publications on model collapse to determine which papers offer explicit definition(s) of model collapse (Fig. 2), where we define explicit as either (1) any mathematical definition, or (2) any precise verbal description of the failure behavior *independent* from results. We additionally categorize which definition(s) of model collapse each paper uses, perhaps implicitly, across any and all mathematical and empirical results (Fig. 2 bottom). We find that many papers do not offer explicit definitions of model collapse and also sometimes use multiple definitions, leading to a lack of specificity and apparent contradictions both within single works and across multiple works.

Before delving into the eight definitions, we first define some shared terminology. In general, we consider model-data feedback loops whereby f_t is the t-th generative model, and we study the behavior of a sequence of generative models $(f_t)_t$ that are iteratively fit to data and then sampled from. We call data sampled from a generative model *synthetic data*. We can evaluate the quality of a generative model in multiple ways. One prominent way is via the *population risk*, defined as the expected loss over the entire real data distribution $\mathbb{E}_{x\sim P}[\ell(f_t(x))]$, where x is some real datum, P is the real data distribution, and ℓ is the loss function (Vapnik, 1991). Another way to evaluate the quality of a generative model is by its tail risk, defined as the expected loss conditioned on tail events $\mathbb{E}_{x\in \text{Tail}(P)}[\ell(f_t(x))]$, where $\text{Tail}(P)$ informally represents real data with low probability. Population risk provides a holistic view of performance, but may mask specific failure modes (Xu, 2024; Kozerawski et al., 2022), whereas tail risk can reveal degradations in edge cases even when population risk remains stable (Hoffmann & Börner, 2020; Xu, 2024). There are many other salient properties of generative models one can use to assess whether the models collapse.

1. **Catastrophic Increase of Population Risk** (Dohmatob et al., 2024b; Bertrand et al., 2023; Kazdan et al., 2024b): Perhaps the most colloquial definition, model collapse is a critical and rapid degradation in model performance due to the presence of synthetic data, as measured by population risk. We note that what constitutes catastrophic is often undefined.
2. **Any Increase of Population Risk** (Alemohammad et al., 2023; Dohmatob et al., 2024b): Under this strict definition, model collapse occurs if there is *any* increase in population risk when training with synthetic data compared to training with real data alone.
3. **Asymptotically Diverging Population Risk** (Gerstgrasser et al., 2024; Kazdan et al., 2024b; Dey & Donoho, 2024): This definition considers model collapse to occur when

the population risk grows without bound over successive model-fitting iterations. This represents a fundamentally unstable learning dynamic where each iteration of synthetic data generation and training leads to progressively worse performance.

4. **Collapsing Variance** (Alemohammad et al., 2023; Shumailov et al., 2023; Bertrand et al., 2023) Model collapse here is when variance (or diversity) trends towards 0 and the learned distributions tend towards delta-like functions over successive model-fitting iterations.
5. **Change in Scaling Law** (Dohmatob et al., 2024c): In this view, model collapse occurs if the governing scaling behavior changes due to the presence of synthetic data. Specifically, model collapse occurs if the relationship between model performance and training data size deviates from the expected scaling behavior observed with real data.
6. **Disappearance of or Entanglement of Real Data Mode(s)** (Alemohammad et al., 2023): Sometimes called "Mode Collapse" (Goodfellow et al., 2014; Lucic et al., 2018; Brock et al., 2019), model collapse here is defined by the presence of synthetic data preventing the model from learning particular modes of the real data distribution or causing the model to blur different data modes together.
7. **Disappearance of Real Tail Data** (Shumailov et al., 2023; Wyllie et al., 2024; Shumailov et al., 2024): Sometimes called "coverage collapse" (Zhu et al., 2024), model collapse here occurs when synthetic data leads to the under-representation of data from the tail of the distribution, leading to models that can only handle common cases but fail on rare ones. The disappearance of real tail data can be more subtle and more narrow than the generative model losing all diversity (Def. 4).
8. **Appearance of Hallucinated Data** (Shumailov et al., 2023; Alemohammad et al., 2023; Bohacek & Farid, 2023): Model collapse occurs when the sequence of models begin producing fully-synthetic data not supported by the original real data's distribution.

We note that the definitions can themselves be loosely clustered into three families (Fig. 1): (i) population risk degrading (Definitions 1, 2 and 3), (ii) distributions deforming from their original shape (Definitions 4, 5, 6, 7, and 8), and (iii) decreasing value from additional data (Definition 5).

2.1 Intra-Paper Definitions Can Cause Confusion

The differences between different definitions of model collapse can be slippery, and understandably, authors sometimes move between them in the course of a paper. However, model collapse is a technical phenomenon that requires definitional rigor to properly characterize. In this section, we use a prominent prior work to demonstrate how easy it can be to slip between definitions. Our intention is not to call out this specific work, but rather demonstrate how a seemingly reasonable treatment of model collapse definitions can have serious implications for interpreting results.

We consider Shumailov et al. (2023), which admirably provides explicit definitions of two different types of model collapse: "We separate two special cases: early model collapse and late model collapse. In early *model collapse* the model begins losing information about the tails of the distribution; in the late *model collapse* the model entangles different modes of the original distributions and converges to a distribution that carries little resemblance to the original one." These correspond to our Definitions 7 and 6, respectively.

However, the paper presents results on model collapse that fall under different definitions. Firstly, Shumailov et al. (2023) demonstrate Definition 3 in their Sections 4.2 and 4.3. We lightly generalize their results for clarity and generality. We consider repeatedly fitting multivariate Gaussians to data and sampling from the fitted Gaussians. We begin with n *real* data drawn from a multivariate Gaussian with mean $\mu^{(0)}$ and covariance $\Sigma^{(0)}$:

$$X_1^{(0)}, ..., X_n^{(0)} \sim_{i.i.d.} \mathcal{N}(\mu^{(0)}, \Sigma^{(0)}). \tag{1}$$

For model fitting, we compute the unbiased mean and covariance of the most recent data:

$$\hat{\mu}^{(t+1)} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^{n} X_j^{(t)} \tag{2}$$

$$\hat{\Sigma}^{(t+1)} \stackrel{\text{def}}{=} \frac{1}{n-1}\sum_{j=1}^{n}(X_j^{(t)} - \hat{\mu}^{(t+1)})(X_j^{(t)} - \hat{\mu}^{(t+1)})^T \tag{3}$$

and then draw n new synthetic samples from a Gaussian with the most recently fit parameters. In this model-data feedback loop, Shumailov et al. (2023) proved that as the model-fitting iteration $t \to \infty$, the population risk as measured by the expected squared Wasserstein distance between the most recent multivariate Gaussian and the original multivariate Gaussian diverges asymptotically:

$$\mathbb{E}[\mathbb{W}_2^2(\mathcal{N}(\hat{\mu}^{(t)}, \hat{\Sigma}^{(t)}) \ , \ \mathcal{N}(\mu^{(0)}, \Sigma^{(0)}))] \to \infty. \tag{4}$$

This result demonstrates that the population risk diverges asymptotically, which aligns with neither of the two definitions of model collapse stated at the outset of the paper. Is it possible that the population risk is diverging *because* of one of the two other definitions? Recall that the squared Wasserstein distance between two Gaussians has two terms: a contribution from the means and a contribution from the covariances:

$$\begin{aligned} &\mathbb{E}[\mathbb{W}_2^2(\mathcal{N}(\hat{\mu}^{(t)}, \hat{\Sigma}^{(t)}), \mathcal{N}(\mu^{(0)}, \Sigma^{(0)}))] \\ &= \underbrace{||\hat{\mu}^{(t)} - \mu^{(0)}||_2^2}_{\to \infty} \quad + \underbrace{\text{Tr}\left(\hat{\Sigma}^{(t)} + \Sigma^{(0)} - 2((\Sigma^{(0)})^{1/2}\hat{\Sigma}^{(t)}(\Sigma^{(0)})^{1/2})^{1/2}\right)}_{\to 0}. \end{aligned} \tag{5}$$

Thus, while the variance collapses and tails do vanish, neither is the cause of the population risk diverging. Rather, the population risk diverges because the sequence of means $(\hat{\mu}^{(t)})_t$ randomly walks away from the ground truth mean $\mu^{(0)}$. This is one example in which a casual reader might fail to notice that while the population risk *is* indeed asymptotically diverging, such undesirable behavior is attributable to neither real data tails disappearing quickly nor to real data modes entangling slowly.

The same paper also demonstrates how definitions of model collapse can go beyond subtle confusion to explicit contradiction. Shumailov et al. (2023) experimentally demonstrate model collapse under a *fourth* definition of model collapse: in their Section 5.2, the authors consider finetuning sequences of `OPT-125M` language models (Zhang et al., 2022) initially on `wikitext2` (Merity et al., 2016) and then subsequently on the models' own generated outputs. The authors' Figure 10 shows that while the population risk (measured by test perplexity) initially increases, it then decreases and converges to a plateau about $12.5\% - 50\%$ above the population risk of the first model. Indeed, the authors find that "Over the generations models tend to produce samples that the original model trained with real data is more likely to produce." We believe that under Definitions 3, 6, and 7, these results suggest that model collapse has not occurred. However, the authors label this result as model collapse because later generations trained on purely synthetic data begin introducing fully synthetic tail data over time: "later generations start producing samples that would never be produced by the original model, i.e., they start misperceiving reality based on errors introduced by their ancestors." Thus, this appearance of hallucinated data qualifies as model collapse under Definition 8.

When a paper shifts definitions without explicitly acknowledging the change, it creates a cascade of problems undermining scientific clarity. Readers interpret results through the lens of the initially stated definitions, creating a false sense that all discussed phenomena represent the same underlying issue when they may be fundamentally distinct. This leads to misattribution of causes and effects, as demonstrated in the Gaussian example where population risk divergence was incorrectly associated with tail data disappearance rather than mean drift. Such definitional inconsistency fosters overgeneralization of results, hampering cross-study comparisons and potentially prompting inappropriate technical responses or policy decisions. Most critically, when technical concepts like model collapse require precise characterization, unstated definitional shifts prevent the formation of a stable framework for interpreting claims, ultimately contributing to broader confusion about which phenomena deserve concern and how they might be addressed.

2.2 Inter-Paper Definitions Can Cause Confusion

To demonstrate how different researchers can look at the same results and reach different conclusions regarding model collapse, Alemohammad et al. (2023) studied a `FFHQ-StyleGAN2` (Karras et al., 2020) trained on synthetic data and found that the population risk as measured by Frechet Inception Distance (FID) (Heusel et al., 2018) increased $2\times$ by the 5th model-fitting iteration and then plateaued. The authors declared this result constituted model collapse because the authors had implicitly defined model collapse as *any* increase in the population risk (Definition 2). Gerstgrasser et al. (2024) then questioned this claim that the models had collapsed, writing, "Figure 7 from Alemohammad et al. (2023) shows that linearly accumulating data ("Synthetic augmentation loop") causes poor behavior to plateau with the number of model-fitting iterations [...] We believe is that our evidence and their

evidence is more consistent with the conclusion that accumulating data avoids model collapse and does not merely delay it." This apparent disagreement was because Gerstgrasser et al. (2024) had defined model collapse as asymptotically diverging population risk (Definition 3). Thus, while looking at the exact same figure, the researchers came to differing conclusions because they were operating under different definitions.

2.3 Stating and Adhering to Definitions Improves Clarity and Drives Progress

In this position paper, our intention is not to argue in favor of specific definitions of model collapse or call out authors whose definitions we disagree with. Rather, we hope to emphasize that model collapse is a multifaceted phenomenon: under the same results, model collapse can simultaneously "occur" and "not occur" in different researchers' opinions; in Appendix A, we include a case study of how confusing and entangled scientific insights can become on account of different definitions and methodologies. This makes building a comprehensive understanding of model collapse difficult, which can be especially concerning to specific communities. For instance, search providers like Google, Bing, and Perplexity may pay an especially high penalty if their models are trained on hallucinated facts, whereas the disappearance of real tail data can disproportionately affect marginalized groups or historically disadvantaged communities (Blodgett et al., 2016; Bender & Friedman, 2018; Noble, 2018; Shah et al., 2020; Jo & Gebru, 2020; Koenecke et al., 2020; Bender et al., 2021; Hutchinson et al., 2021; Ji et al., 2023). By clarifying different definitions of model collapse and adhering to those definitions, researchers can build a better understanding of model collapse with greater nuance for what causes different failures modes and what actions can be taken to prevent each. Together, these definitions can form a "model collapse profile" which can characterize the different ways in which models worsen over time. Research on the harms of synthetic data should explicitly note the relevant aspects of the model collapse profile they study.

3 Realistic Conditions for Studying Model Collapse

Our goal is to understand likely outcomes as humanity pre-trains future frontier AI systems on web-scale datasets containing a mixture of real and synthetic data. We focus on pre-training both because the literature has (Shumailov et al. (2023)'s "What will happen to GPT-n once LLMs contribute much of the language found online?") and because pre-training's tremendous capital and operating expenses render missteps extremely costly. Motivated by this goal, we posit trends that we believe describe the current trajectory of pre-training practices of frontier AI systems:

1. **Increasing Pre-training Compute**: The total floating point operations used for pre-training has been rapidly increasing. For example, Meta pre-trained Llama 1 using 2k GPUs, Llama 2 using 4k GPUs and Llama 3 using 16k GPUs (Goyal, 2024), and OpenAI recently announced a $500B initiative to increase compute capacity of the United States (OpenAI & SoftBank, 2025), a fraction of which will be allocated for pre-training.
2. **Increasing Pre-training Data**: The amount of data used has been rapidly increasing. For example, Meta's series of Llama language models were pre-trained on increasing amounts of data: 1.4 trillion tokens for Llama 1, 2 trillion tokens for Llama 2, and most recently, 15 trillion tokens for Llama 3. While pre-training data will inevitably max out, recent estimates place the total number of available language pre-training tokens at 1 quadrillion (1000 trillion) tokens (Villalobos et al., 2024).
3. **Increasing Quality of Pre-training Data**: Over time, pre-training data is becoming increasingly higher quality on account of pre-training data teams developing better filtering techniques to ensure high-quality training data (Gao et al., 2020; Penedo et al., 2024; Li et al., 2024). Moreover, whatever synthetic data is shared online is increasingly higher quality since models are improving over time Kiela et al. (2021, 2023); Maslej et al. (2024).
4. **Synthetic Data Accumulating Alongside Real Data**: Real data are not deleted en masse after each iteration of model pre-training. When synthetic data are generated and released online, they amasses alongside prior data and new real data.
5. **Decreasing Proportion of Real Data**: The fraction of real data relative to total (real plus synthetic) data is decreasing over time. However, whether the fraction of real data will asymptote to zero is unclear, a point we will return to later.

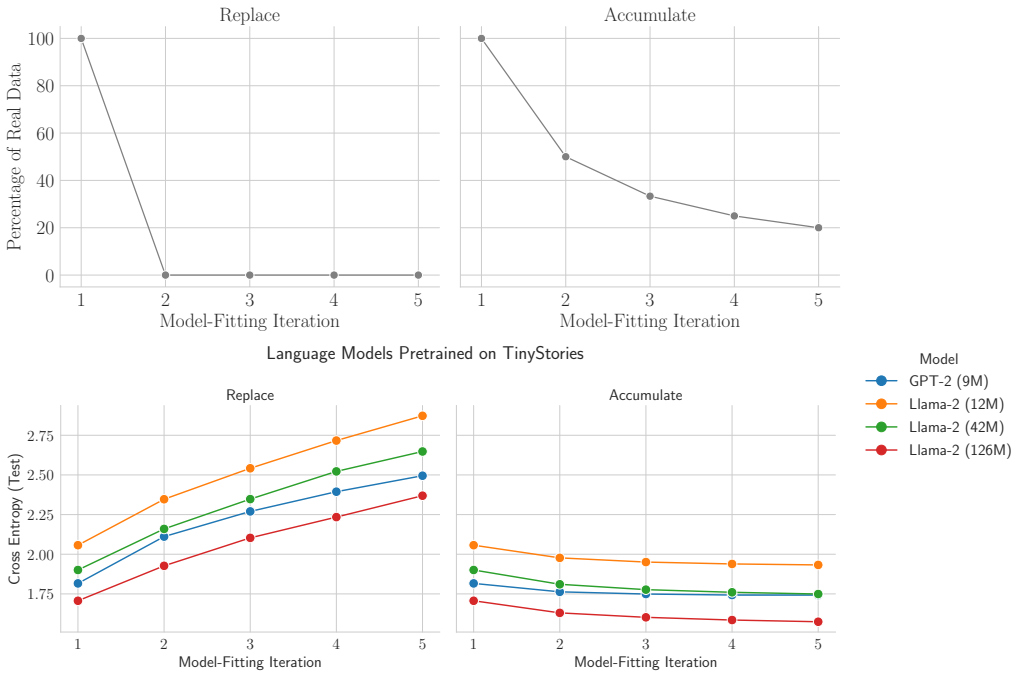

Figure 3: **Dimensions of Consideration for Model-Data Feedback Loops: Propagation of Data Over Time and Proportion of Real Data Over Time.** When data are *replaced* after each model-fitting iteration (left), the proportion of real data immediately becomes zero after the first iteration, whereas when data instead *accumulate* (right), the proportion of real data falls asymptotically to zero. Gerstgrasser et al. (2024); Kazdan et al. (2024b); Dey & Donoho (2024) showed that replacing data over time causes the population risk to diverge, whereas accumulating data avoids diverging population risk. In these works, synthetic data are assumed to grow linearly over time, contributing n samples per model-sampling iteration. Credit: The bottom figure is copied from Gerstgrasser et al. (2024) with permission.

We believe that researchers and policy makers interested in potential societal implications of model collapse should focus on research that adhere to these conditions as faithfully as possible (with the obvious caveat that computational budgets limit research).

4 Key Dimensions of Consideration for Model-Data Feedback Loops

4.1 Propagation of Data Over Time

Early work that sounded the alarm about model collapse (Martínez et al., 2023; Alemohammad et al., 2023; Bohacek & Farid, 2023; Shumailov et al., 2023; Briesch et al., 2023; Bertrand et al., 2023) assumed that data propagate in a particular way: after training a model, all existing data are deleted, new data are sampled from the new model, and the next model is trained solely on this fresh synthetic data. Subsequent authors called this the *replace* paradigm Gerstgrasser et al. (2024); Kazdan et al. (2024b); Dey & Donoho (2024) because data are entirely replaced after each model-fitting iteration. When data are replaced, researchers demonstrated multiple harmful outcomes: variances collapse, real data tails disappear, population risk diverges, etc. (Fig. 3 left). This particular assumption of how data propagate over time is highly unrealistic: **After a model finishes training, the entire internet is not deleted, nor is the next model necessarily trained solely on its predecessor's outputs.** Rather, a more realistic assumption is that synthetic data from each model *accumulates* on the internet alongside real data and past synthetic data such that all can be used for training the next model. To some, these differences might seem insignificant, but each produces vastly different asymptotic behavior in terms of population risk: Gerstgrasser et al. (2024) showed empirically and Kazdan et al. (2024b) and Dey & Donoho (2024) showed mathematically that population risk diverge if data are replaced, but population risk does not diverge if data instead accumulate (Fig. 3 right).

A slightly different but perhaps more realistic data propagation assumption is that real and synthetic data accumulate, but future models are trained on a downsampled proportion of the total available data (Kazdan et al., 2024b). This *accumulate-subsample* paradigm represents a middle ground between replace and accumulate: while test loss appears to stabilize, no analytic theory has been proven in this case to date. Individual beliefs about which data paradigm is most reflective of reality can influence perceptions of how model collapse will unfold in the future. However, to our knowledge, non-population risk-based notions of model collapse have not been connected to assumptions about how data propagate, leaving an important question open.

4.2 Proportion of Real Data Over Time

ChatGPT alone produces $1/1000$ of all words produced by humanity each day Altman (2024), and as these models proliferate, over time, future generative models could produce vastly more data than humanity for training future models. Thus, the proportion of real data on the future internet plays a crucial role in the debate over the effects of model collapse. Bertrand et al. (2023) claimed that the population risk will not asymptotically diverge so long as the proportion of real data remains lower-bounded above 0 (in addition to other conditions). Relatedly, in Dohmatob et al. (2024b)'s view, *any* synthetic data causes "a critical degradation" to future models, writing model collapse "generally persists even when mixing real and synthetic data, as long as the fraction of training data which is synthetic does not vanish" and that "model collapse cannot generally be mitigated by simple adjustments [...] unless these strategies asymptotically remove all but a vanishing proportion of synthetic data from the training process."

Results like these draw attention to what proportion of real data is necessary to avoid collapse, which is a valid consideration. However, correctly interpreting these results takes care. For instance, Bertrand et al. (2023)'s condition is *sufficient*, not necessary; it should *not* be understood as saying that model collapse is inevitable unless real data remains a non-zero proportion. Moreover, contrary work by Gerstgrasser et al. (2024), Marchi et al. (2024), and Kazdan et al. (2024b) established conceptually stronger guarantees: when data *accumulate*, but human data asymptotically occupy a vanishing fraction of the internet, the population risk will likely not diverge (in some cases, with an additional requirement that the rate of AI data generation does not grow super-linearly) (Fig. 3). Relatedly, Gillman et al. (2024) also show that the proportion of real data can asymptotically approach zero using a function to "correct" synthetic data towards the real data distribution. Due to uncertainty about the rate of synthetic data generation, one potential solution is to sequester real training data for future use, when the internet still contains an abundance of human-generated data; however, frontier AI labs have already collected such data, meaning no additional work is required.

4.3 Model Training Assumptions

While multiple works claim that models do not collapse under certain settings or propose interventions to avoid collapse, we urge the explicit declaration of strong technical assumptions that are frequently unrealistic. As an example, Bertrand et al. (2023) and Gillman et al. (2024) study *iterative retraining*, where each model is initialized from its predecessor's parameters and optimizer state. Under an additional assumption that the first model is sufficiently high performing, since each model is assumed to be close to its predecessor, model collapse can be avoided. However, to the best of our knowledge, iterative retraining has not been used for any frontier AI model, including OpenAI's GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), GPT-4 Achiam et al. (2023), Anthropic's Claude 1, 2 or 3, Google's PaLM 1 (Chowdhery et al., 2022), PALM 2 (Anil et al., 2023) or Gemini (Team et al., 2024), DeepSeek's V3 (DeepSeek-AI et al., 2024). Thus, Bertrand et al. (2023)'s sufficiency conditions for avoiding collapse are far from current practices.

For another example, Zhu et al. (2024) propose data editing as a collapse mitigation strategy and analyze a self-consuming linear model. Each model iteration fits $\hat{w}_n = X^\dagger \tilde{Y}_n$, where X are fixed Gaussian covariates, and generates synthetic data $\hat{Y}_{n+1} = X\hat{w}_n + E_{n+1}$ where E_{n+1} are Gaussian errors and $\tilde{Y}_n$ are regression targets sampled from the previous generation's training targets with edits from the prior generations synthetic data:

$$\tilde{Y}_n^\top = M_{n-1}\hat{Y}_n + (1 - M_{n-1})\tilde{Y}_{n-1}$$

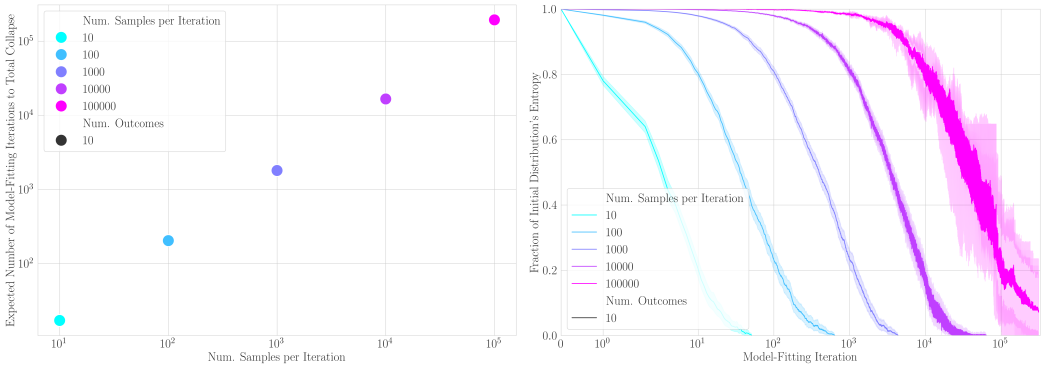

Figure 4: **Dimension of Consideration for Model-Data Feedback Loops: Timescales of Collapse.** Characterizing the timescale over which one should expect collapse is an underappreciated but crucial consideration. Focusing on the discrete model of Shumailov et al. (2023), the expected number of model-fitting iterations before total collapse is proportional to the number of data times the entropy of the initial data distribution (left). Taking this model at face value, this means that trillions of models can be trained before glimpsing the onset of collapse. However, total collapse is only the most extreme outcome; in this model, we additionally show how the entropy of the initial data distribution decays over time (right). Error bars are over 100 seeds (0 to 99, inclusive); for details, see Sec. 4.4.

Here, M_k is a diagonal matrix of 1's or 0's indicating whether to replace or not replace a label with a synthetic value. While Zhu et al. (2024) claim that this prevents model collapse, the proof of their test error bound (Theorem 2) assumes that $\|M_i\| = \eta\|M_{i-1}\|$ for some constant $\eta \in (0, 1)$, meaning that the number of edits decreases by at least some fixed proportion each generation. This geometric decay guarantees that the total number of edits at any given generation is finite, and their proof fails without this. However, training GPT-2 on the Natural-Instructions dataset (Mishra et al., 2021) yields only a slight decline in edit percentage (Zhu et al., 2024). Moreover, if the number of edits is finite, each generation trains on mostly real data.

4.4 Timescales of Collapse

Another key dimension of consideration is the timescale of model deterioration, which is often omitted. For example, Shumailov et al. (2023) introduced a simple theoretical setting for studying model collapse: a discrete distribution (picture a histogram) with N outcomes (atoms):

$$p^{(0)}(x) = \sum_{n=1}^{N} w_n \delta_{x_n}(x), \tag{6}$$

where $\sum_n w_n = 1$. If one sequentially draws D data from this distribution and computes a new distribution based on the empirical proportions, then this process forms a Markov chain with N absorbing states, each corresponding to a totally collapsed distribution, i.e., a distribution comprised of exactly one outcome. Consequently, one can use standard results from absorbing Markov chains to show that this simple process *must* collapse.

However, one can go beyond a guarantee of collapse and ask: *how many model-fitting iterations can we survive before total collapse consumes us?* Again using standard results (Brydges, 2009), the expected number of model-fitting iterations before total collapse is:

$$\mathbb{E}[\text{Model Iterations Till Total Collapse}] \propto D\, H[p^{(0)}],$$

where $H[\cdot]$ is the Shannon entropy. For example, if our starting distribution is uniform, the entropy is $\log(N)$ and thus the expected number of model-fitting iterations before total collapse is $D \log(N)$. We confirmed this claim using numerical simulations starting from uniform distributions (Fig. 4 Left). We also numerically simulated how quickly the entropy of $\hat{p}^{(t)}$ falls relative to the entropy of $p^{(0)}(x)$ to provide a description of the process more nuanced than just total collapse (Fig. 4 Right), since one might be interested in how quickly tail information is lost, not just how quickly total collapse arrives.

While this mathematical model is simple, for the sake of argument, if we take this model at face value, we realize total model collapse poses virtually no present threat. This is because the number

of real public text data alone is on the order a quadrillion tokens (Villalobos et al., 2024), and text, image, video and agentic data are high dimensional with large entropy, meaning total collapse occurs so imperceptibly slowly that humanity could train trillions of models before noticing the onset of collapse. However, this highlights a more general point: characterizing the timescale over which one should expect model collapse to occur is an underappreciated but crucial consideration for describing the model collapse profile.

Several authors do characterize timescales. Suresh et al. (2024) give an exact formulation of model collapse rate in the fundamental setting of recursive maximum likelihood estimation for discrete distributions and Gaussian mixtures; they find that for $\mathrm{Bern}(\mu)$, $\mathrm{Pois}(\lambda)$, and Gaussian mixtures with shared variance σ^2 distributions, the parameters μ, λ, and σ collapse to 0 exponentially in the number of generations. While Seddik et al. (2024) previously controlled the total collapse probability in both the fully synthetic and partially-synthetic recursive training regimes, Suresh et al. (2024) further control the number of unique symbols after k generations in the fully synthetic regime. Lastly, Kazdan et al. (2024b) note, in the context of kernel density estimators with fixed bandwidths, that the negative log likelihood does diverge asymptotically, although "this occurs at a rate so glacial that it doesn't pose a practical concern."

5 Is Model Collapse a Threat?

In conclusion, is model collapse a threat? In our view, model collapse is a multifaceted phenomenon, and a single answer is not possible. By taking a realism-weighted average of different papers' results, we synthesize our own forecast of model collapse under the different definitions:

Population risk will not increase catastrophically or diverge asymptotically. Given the increase in pretraining dataset size and quality, we argue that models training on accumulating synthetic data alongside real data will not suffer from catastrophic population risk increase or diverging population risk. The jury remains out on whether the proportion of real data relative to available data will approach zero.

Real tail data and modes will be lost, but how many and how quickly is unclear. Loss of diversity is a real issue, with disproportionate harms oftentimes born by subgroups. It is unclear how much of the tail we will lose or which of the real data modes will become entangled, and how synthetic data affect such changes that already occur naturally. We strongly encourage more research regarding coverage and mode collapse prevention strategies for realistic settings, building on prior work such as Hashimoto et al. (2018); Ensign et al. (2018); Taori & Hashimoto (2023).

Scaling laws may change with the introduction of synthetic data. The precise nature of these changes under realistic conditions remains to be determined. Synthetic data could potentially remove what some researchers describe as a data bottleneck, but this benefit might come at the cost of altered scaling law parameters. We encourage further research into how synthetic data affects scaling behaviors to better characterize likely future outcomes.

We emphasize that while real threats do exist, the popular perception that synthetic data on the internet will render future frontier AI models pretrained on web-scale data useless is likely unrealistic since such failures appear in conditions that do not faithfully match what is actually done in practice. Subtle degradations in data distributions might still insidiously occur, such as loss of real tail data, and future work should aim to explore what can be used to counter such outcomes.

6 Alternative Views

One may feel the conditions we identify in Section 3 are inaccurate, disproportionately emphasized, likely to change, or ignorant of important settings other than pre-training. One might also believe that the identified model collapse definitions in Section 2 do not fully encompass the literature or are inaccurately applied in Figure 2. Finally, a concerned bystander could argue that the benefits of being over-cautious outweigh the costs: if society fails to anticipate model collapse by allowing wanton generation of poor-quality synthetic data, then the internet could become flooded with low-quality samples that preclude future progress. These are valid points, and we look forward to engaging with researchers and policymakers to better identify what matters to them and what the future looks like in those directions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction accurately reflect the paper's main text.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: This is a position paper that does not introduce significant new research material.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This is a position paper that does not introduce significant new research material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: This is a position paper that does not introduce significant new research material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: This is a position paper that does not introduce significant new research material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: This is a position paper that does not introduce significant new research material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: This is a position paper that does not introduce significant new research material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: This is a position paper that does not introduce significant new research material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We believe we have followed the NeurIPS Code of Ethics in every regard.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This is a position paper that does not introduce significant new research material. We do feel that we discuss societal impacts of prior research at length.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release data, models, code or other artifacts with any high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We believe we have cited prior work and followed existing licenses where applicable.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our figures are accompanied by adjacent captions that explain their contents.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This manuscript contains no crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This manuscript contains no content that would require involving IRB or equivalent.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We do not use LLMs materially for this manuscript. We used them a bit to accelerate plotting data in matplotlib.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., and Baraniuk, R. G. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850*, 2023.

Altman, S. openai now generates about 100 billion words per day., Feb 2024. URL `https://twitter.com/sama/status/1756089361609981993`. [Online; accessed 13-October-2024].

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., Ruder, S., Tay, Y., Xiao, K., Xu, Y., Zhang, Y., Abrego, G. H., Ahn, J., Austin, J., Barham, P., Botha, J., Bradbury, J., Brahma, S., Brooks, K., Catasta, M., Cheng, Y., Cherry, C., Choquette-Choo, C. A., Chowdhery, A., Crepy, C., Dave, S., Dehghani, M., Dev, S., Devlin, J., Díaz, M., Du, N., Dyer, E., Feinberg, V., Feng, F., Fienber, V., Freitag, M., Garcia, X., Gehrmann, S., Gonzalez, L., Gur-Ari, G., Hand, S., Hashemi, H., Hou, L., Howland, J., Hu, A., Hui, J., Hurwitz, J., Isard, M., Ittycheriah, A., Jagielski, M., Jia, W., Kenealy, K., Krikun, M., Kudugunta, S., Lan, C., Lee, K., Lee, B., Li, E., Li, M., Li, W., Li, Y., Li, J., Lim, H., Lin, H., Liu, Z., Liu, F., Maggioni, M., Mahendru, A., Maynez, J., Misra, V., Moussalem, M., Nado, Z., Nham, J., Ni, E., Nystrom, A., Parrish, A., Pellat, M., Polacek, M., Polozov, A., Pope, R., Qiao, S., Reif, E., Richter, B., Riley, P., Ros, A. C., Roy, A., Saeta, B., Samuel, R., Shelby, R., Slone, A., Smilkov, D., So, D. R., Sohn, D., Tokumine, S., Valter, D., Vasudevan, V., Vodrahalli, K., Wang, X., Wang, P., Wang, Z., Wang, T., Wieting, J., Wu, Y., Xu, K., Xu, Y., Xue, L., Yin, P., Yu, J., Zhang, Q., Zheng, S., Zheng, C., Zhou, W., Zhou, D., Petrov, S., and Wu, Y. Palm 2 technical report, 2023. URL `https://arxiv.org/abs/2305.10403`.

Bender, E. M. and Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Bertrand, Q., Bose, A. J., Duplessis, A., Jiralerspong, M., and Gidel, G. On the stability of iterative retraining of generative models on their own data. *arXiv preprint arXiv:2310.00429*, 2023.

Blodgett, S. L., Green, L., and O'Connor, B. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016.

Bohacek, M. and Farid, H. Nepotistically trained generative-ai models collapse. *arXiv preprint arXiv:2311.12202*, 2023.

Briesch, M., Sobania, D., and Rothlauf, F. Large language models suffer from their own output: An analysis of the self-consuming training loop. *arXiv preprint arXiv:2311.16822*, 2023.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis, 2019. URL `https://arxiv.org/abs/1809.11096`.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Brydges, D. Wright-fisher processes, 2009. URL `https://personal.math.ubc.ca/~db5d/SummerSchool09/lectures-dd/lectures6-7.pdf`. [Online; accessed 30-Jan-2025].

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R.,

Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J. L., Liang, J., Guo, J., Ni, J., Li, J., Wang, J., Chen, J., Chen, J., Yuan, J., Qiu, J., Li, J., Song, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Wang, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Wang, Q., Zhu, Q., Chen, Q., Du, Q., Chen, R. J., Jin, R. L., Ge, R., Zhang, R., Pan, R., Wang, R., Xu, R., Zhang, R., Chen, R., Li, S. S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Pan, S., Wang, T., Yun, T., Pei, T., Sun, T., Xiao, W. L., Zeng, W., Zhao, W., An, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Li, X. Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Zhang, X., Chen, X., Nie, X., Sun, X., Wang, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yu, X., Song, X., Shan, X., Zhou, X., Yang, X., Li, X., Su, X., Lin, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhu, Y. X., Zhang, Y., Xu, Y., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Yu, Y., Zheng, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Tang, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Wu, Y., Ou, Y., Zhu, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Zha, Y., Xiong, Y., Ma, Y., Yan, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Wu, Z. F., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Huang, Z., Zhang, Z., Xie, Z., Zhang, Z., Hao, Z., Gou, Z., Ma, Z., Yan, Z., Shao, Z., Xu, Z., Wu, Z., Zhang, Z., Li, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Gao, Z., and Pan, Z. Deepseek-v3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

Dey, A. and Donoho, D. Universality of the $\pi^2/6$ pathway in avoiding model collapse. *arXiv preprint arXiv:2410.22812*, 2024.

Dohmatob, E., Feng, Y., and Kempe, J. Model collapse demystified: The case of regression. *arXiv preprint arXiv:2402.07712*, 2024a.

Dohmatob, E., Feng, Y., Subramonian, A., and Kempe, J. Strong model collapse. *arXiv preprint arXiv:2410.04840*, 2024b.

Dohmatob, E., Feng, Y., Yang, P., Charton, F., and Kempe, J. A tale of tails: Model collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024c.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. Runaway feedback loops in predictive policing. In *Conference on fairness, accountability and transparency*, pp. 160–171. PMLR, 2018.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., Korbak, T., Agrawal, R., Pai, D., Gromov, A., et al. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*, 2024.

Gibney, E. Ai models fed AI-generated data quickly spew nonsense. *Nature*, 632:18–19, 7 2024. doi: 10.1038/d41586-024-02420-7. URL https://www.nature.com/articles/d41586-024-02420-7.

Gillman, N., Freeman, M., Aggarwal, D., Hsu, C.-H., Luo, C., Tian, Y., and Sun, C. Self-correcting self-consuming loops for generative model training. In *International Conference on Machine Learning*, pp. 15646–15677. PMLR, 2024.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Goyal, N. llama1: 2048 gpus llama2: 4096 gpus llama3: 16384 gpus llama4: ..... you see where we are headed! gonna be insane ride!, jul 2024. URL https://x.com/NamanGoyal21/status/1815819622525870223. Tweet.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. URL https://arxiv.org/abs/1706.08500.

Hoffmann, I. and Börner, C. J. Tail models and the statistical limit of accuracy in risk assessment. *The Journal of Risk Finance*, 21(3):201–216, 2020.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., and Mitchell, M. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 560–575, 2021.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Jo, E. S. and Gebru, T. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 306–316, 2020.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.

Kazdan, J., Dey, A., Schaeffer, R., Gerstgrasser, M., Rafailov, R., Donoho, D. L., and Koyejo, S. Accumulating data avoids model collapse. In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*, 2024a.

Kazdan, J., Schaeffer, R., Dey, A., Gerstgrasser, M., Rafailov, R., Donoho, D. L., and Koyejo, S. Collapse or thrive? perils and promises of synthetic data in a self-generating world, 2024b. URL https://arxiv.org/abs/2410.16713.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., et al. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*, 2021.

Kiela, D., Thrush, T., Ethayarajh, K., and Singh, A. Plotting progress in ai. *Contextual AI Blog*, 2023. https://contextual.ai/blog/plotting-progress.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689, 2020.

Kozerawski, J., Sharan, M., and Yu, R. Taming the long tail of deep probabilistic forecasting. *arXiv preprint arXiv:2202.13418*, 2022.

Li, J., Fang, A., Smyrnis, G., Ivgi, M., Jordan, M., Gadre, S., Bansal, H., Guha, E., Keh, S., Arora, K., Garg, S., Xin, R., Muennighoff, N., Heckel, R., Mercat, J., Chen, M., Gururangan, S., Wortsman, M., Albalak, A., Bitton, Y., Nezhurina, M., Abbas, A., Hsieh, C.-Y., Ghosh, D., Gardner, J., Kilian, M., Zhang, H., Shao, R., Pratt, S., Sanyal, S., Ilharco, G., Daras, G., Marathe, K., Gokaslan, A.,

Zhang, J., Chandu, K., Nguyen, T., Vasiljevic, I., Kakade, S., Song, S., Sanghavi, S., Faghri, F., Oh, S., Zettlemoyer, L., Lo, K., El-Nouby, A., Pouransari, H., Toshev, A., Wang, S., Groeneveld, D., Soldaini, L., Koh, P. W., Jitsev, J., Kollar, T., Dimakis, A. G., Carmon, Y., Dave, A., Schmidt, L., and Shankar, V. Datacomp-lm: In search of the next generation of training sets for language models, 2024. URL https://arxiv.org/abs/2406.11794.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. Are gans created equal? a large-scale study. *Advances in neural information processing systems*, 31, 2018.

Marchi, M., Soatto, S., Chaudhari, P., and Tabuada, P. Heat death of generative models in closed-loop learning, 2024.

Martínez, G., Watson, L., Reviriego, P., Hernández, J. A., Juarez, M., and Sarkar, R. Towards understanding the interplay of generative artificial intelligence and the internet. *arXiv preprint arXiv:2306.06130*, 2023.

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., and Clark, J. The ai index 2024 annual report, 2024.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models, 2016.

Mishra, S., Khashabi, D., Baral, C., and Hajishirzi, H. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021.

Murthy, S. K., Ullman, T., and Hu, J. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity, 2024. URL https://arxiv.org/abs/2411.04427.

Noble, S. U. Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press, 2018.

OpenAI and SoftBank. Announcing the Stargate project, 1 2025. URL https://openai.com/index/announcing-the-stargate-project/.

Padmakumar, V. and He, H. Does writing with language models reduce content diversity?, 2024. URL https://arxiv.org/abs/2309.05196.

Penedo, G., Kydlíček, H., allal, L. B., Lozhkov, A., Mitchell, M., Raffel, C., Werra, L. V., and Wolf, T. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rao, R. Ai-generated data can poison future ai models. *Scientific American*, jul 2023. URL https://www.scientificamerican.com. Edited by Sophie Bushwick.

Seddik, M. E. A., Chen, S.-W., Hayou, S., Youssef, P., and Debbah, M. How bad is training on synthetic data? a statistical analysis of language model collapse, 2024.

Seetharaman, D. For data-guzzling AI companies, the Internet is too small. *The Wall Street Journal*, apr 2024. URL https://www.wsj.com. Published online at 5:30 am ET.

Shah, D. S., Schwartz, H. A., and Hovy, D. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5248–5264, 2020.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., and Anderson, R. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07566-y. URL https://doi.org/10.1038/s41586-024-07566-y.

Suresh, A. T., Thangaraj, A., and Khandavally, A. N. K. Rate of model collapse in recursive training, 2024. URL https://arxiv.org/abs/2412.17646.

Taori, R. and Hashimoto, T. Data feedback loops: Model-driven amplification of dataset biases. In *International Conference on Machine Learning*, pp. 33883–33920. PMLR, 2023.

Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Vapnik, V. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., and Hobbhahn, M. Will we run out of data? limits of llm scaling based on human-generated data, 2024. URL https://arxiv.org/abs/2211.04325.

Wang, A. New paper in nature shows model collapse as successive model generations models are recursively trained on synthetic data, jul 2024. URL https://x.com/alexandr_wang/status/1816491442069782925. Posted on X (formerly Twitter) at 8:11 AM.

Wu, F., Black, E., and Chandrasekaran, V. Generative monoculture in large language models, 2024. URL https://arxiv.org/abs/2407.02209.

Wyllie, S., Shumailov, I., and Papernot, N. Fairness feedback loops: Training on synthetic data amplifies bias, 2024. URL https://arxiv.org/abs/2403.07857.

Xu, M. Estimating tail risk in neural networks, 2024. URL https://www.alignment.org/blog/estimating-tail-risk-in-neural-networks/.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

Zhang, Y., Schwarzschild, A., Carlini, N., Kolter, Z., and Ippolito, D. Forcing diffuse distributions out of language models. *arXiv preprint arXiv:2404.10859*, 2024.

Zhu, X., Cheng, D., Li, H., Zhang, K., Hua, E., Lv, X., Ding, N., Lin, Z., Zheng, Z., and Zhou, B. How to synthesize text data without model collapse?, 2024. URL https://arxiv.org/abs/2412.14689.

A A Case Study of Disagreement Between Papers

The wide variety of non-equivalent definitions for model collapse often creates apparent contradictions between papers claiming to study the same phenomenon. A representative example presents itself in the study of model collapse for linear regression. In this data setting, one begins with a dataset (X, y) where we assume that

$$y \sim X\beta + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma \cdot I).$$

In the first iteration, one computes

$$\hat{\beta}^{(1)} = \left(X^T X\right)^{-1} X^T y,$$

and uses the fit parameter to generate new data $(X, y^{(1)})$ with

$$y^{(1)} = X\hat{\beta}^{(1)} + \epsilon_1, \epsilon_1 \sim \mathcal{N}(0, \sigma \cdot I).$$

One can then fit successive model iterations using an *accumulate* paradigm, in which the data for the nth model fitting takes the form

$$\left(\left[y, y^{(1)}, y^{(2)}, \ldots, y^{(n-1)}\right]^\top, [X, X, \ldots, X]^\top\right)$$

One can also fit the nth model iteration using a *replace* paradigm, in which $\hat{\beta}^{(n)}$ is computed using only the data $(X, y^{(n-1)})$. As proven by Gerstgrasser et al. (2024), in the accumulate paradigm, the ratio

$$\frac{\mathbb{E}\left[\|\hat{\beta}^{(n)} X_{\text{test}} - y_{\text{test}}\|^2\right]}{\mathbb{E}\left[\|\hat{\beta}^{(1)} X_{\text{test}} - y_{\text{test}}\|^2\right]}$$

monotonically increases before converging to $\pi^2/6$. Citing Definitions 3 and 4, Gerstgrasser et al. (2024) correctly asserted that model collapse does not occur. However, if one instead defines model collapse by Definition 2, this scenario does exhibit collapse.

To complicate the story, Dohmatob et al. (2024a) studied the same model under the replace paradigm. Under the replace paradigm, Definition 3 suggests that model collapse occurs since the asymptotic risk diverges, while Definition 4 implies that model collapse does not occur, since the replace scenario does not exhibit vanishing variance. Table 1 shows how incompatible definitions lead to confusion.

Table 1: Model collapse occurs under some definitions, but does not occur under others for the regression setting described in Appendix A.

Definition	**Accumulate**	**Replace**
Def. 1	✗	✓
Def. 2	✓	✓
Def. 3	✗	✓
Def. 4	✗	✗
Def. 5	✓	✓
Def. 6	✗	✗
Def. 7	✗	✗
Def. 8	✗	✗