Simulating Model Collapse:

**Model Collapse:**

Model collapse is a feedback‑loop failure in which a generative model that is repeatedly retrained on its own or other models’ synthetic outputs drifts away from the true data distribution, progressively forgetting rare or tail events and producing ever less‑diverse, lower‑fidelity content.

The phenomenon was formalised by Shumailov et al.’s 2023 paper *“The Curse of Recursion,”* which draws on earlier discussions of “mode collapse” in Generative Adversarial Networks (GANs) but extends the idea to language generative models. For context, “mode collapse” occurs when a model produces outputs that are less diverse than the training data’s distribution, effectively "collapsing" to generate only a few modes of the data distribution while ignoring others.

**Relevance of Model Collapse in the Context of GPT Models:**

It is now clear that large language models (LLMs) are here to stay, and will bring about drastic change in the whole ecosystem of online text and images. The development of LLMs requires masses of training data which are sourced by scraping much of the Internet, then further fine-tuned with reinforcement learning from human feedback (RLHF). (Shumailov, 2023) What will happen to GPT-{n} once LLMs contribute much of the available/scrapable language found online?

Literature shows that models up to GPT-4 were trained on predominantly human generated texts, but this may not be the case for future iterations of GPT, as it is expected that more and more LLM generated data will find its way into the training data of future GPT models.

**Ethics of Model Collapse**

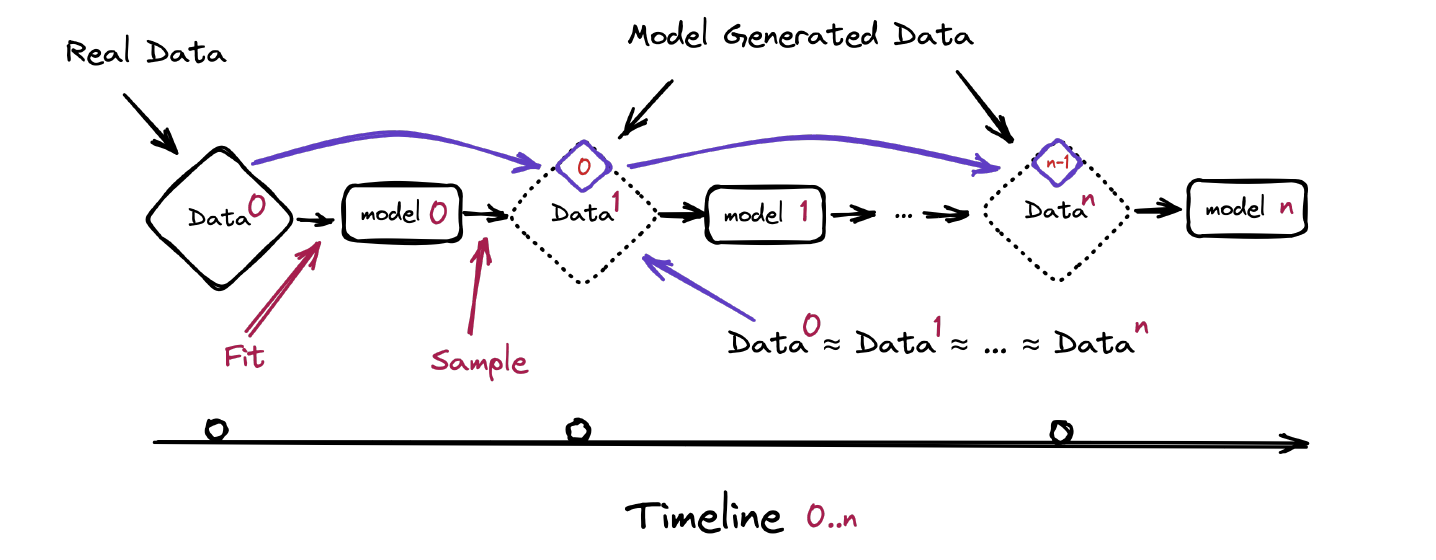
Preserving the ability of LLMs to model low-probability events is essential to the fairness of their predictions: such events are often relevant to marginalised groups.

**Simulating and Evaluating Model Collapse:**

To mimic *model collapse* in a controlled sandbox I set up a toy feedback loop, where: First, a Markov‑chain name generator produces thousands of synthetic 5‑/6‑letter “names” based on the true original distribution for english names, then throws away the original human statistics, and retrains only on its own output, recursively.

We then collect Entropy, and KL-divergence stats for every iteration of the loop. These metrics were chosen based on their use in literature. Entropy shrinks when diversity vanishes, and KL‑divergence grows as the synthetic distribution drifts from the original.

The below diagram (figure x) outlines the step-by-step of the simulation’s loop.



The simulation was set up using arbitrary hyperparameters, loosely inspired by the simulation seen in Shumailov et al.’s 2023 paper *“The Curse of Recursion.”* The hyperparameters used to set up the simulation are described below in Table 1.

*Table 1:*

| Hyperparameter | Value | Description |
| --- | --- | --- |
| GENS | 9 | Number of generations of the model |
| N\_NAMES\_PER\_GEN | 10 000 | How many names the model generates |
| N\_ITERS\_PER\_NAME | 1 000 | the number of Metropolis-Hastings steps we let the Markov chain take **f**or each individual name |
| NAME\_LENGTH\_OPTIONS | [5, 6] | Number of characters per name |
| TEMPERATURE | 1.0 | Scales the Metropolis-Hastings acceptance rule: higher values let the algorithm accept worse-scoring names more often. |

**Simulation Results:**

At every iteration of the simulation loop, two metrics were calculated to analyze the changes in the names generated by the MCMC algorithm: Entropy and Kullback-Leibler (KL) divergence. Entropy is the expected information content of drawing a token from a distribution; higher values mean the name is harder to predict form a single token and therefore more diverse. Shumailov et al. show that, under recursive self-training, n-gram entropy drops sharply after only a few generations—the empirical sign that “tails vanish first.

*Table 2:*

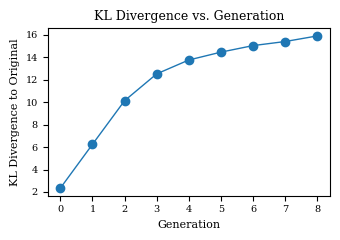
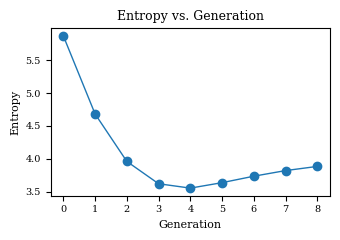
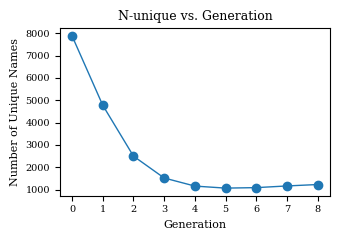
| generation | entropy | kl\_to\_orig | N-Unique Names |
| --- | --- | --- | --- |
| 0 | 5.867039 | 2.345249 | 7833 |
| 1 | 4.682282 | 6.241527 | 4776 |
| 2 | 3.961877 | 10.122113 | 2498 |
| 3 | 3.622965 | 12.509385 | 1518 |
| 4 | 3.556657 | 13.743662 | 1156 |
| 5 | 3.639479 | 14.443453 | 1065 |
| 6 | 3.737514 | 15.025509 | 1087 |
| 7 | 3.822835 | 15.387559 | 1162 |
| 8 | 3.886130 | 15.885615 | 1226 |

**Table 2 – Collapse Diagnostics by Generation**

*Entropy and unique‑name diversity fall while KL divergence from the original distribution rises,*

*illustrating progressive model collapse over nine recursive training rounds.*

**Figures 1,2,3:**



*Shumailov et al 2023, separates 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 model entangles different modes of the original distributions and converges to a distribution that carries little resemblance to the original one, often with very small variance.*

Figures 1,2 and 3 illustrate the transition from early model collapse to late model collapse. Entropy decreases up to generation 4 even though KLdivergence keeps increasing. THis shows that even though the model is behind to regain its tails relative to the original levels after generation 4 (Entropy and N-Unique names increase), the “tails” no longer resemble the original distribution (KL keeps increasing).