AI models collapse when trained on recursively generated data

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson & Yarin Gal

Presenter: **Lorenzo Valente**

Organiser: **Henry Day-Hall**

JOINT ML JOURNAL CLUB
October 11, 2024

Overview

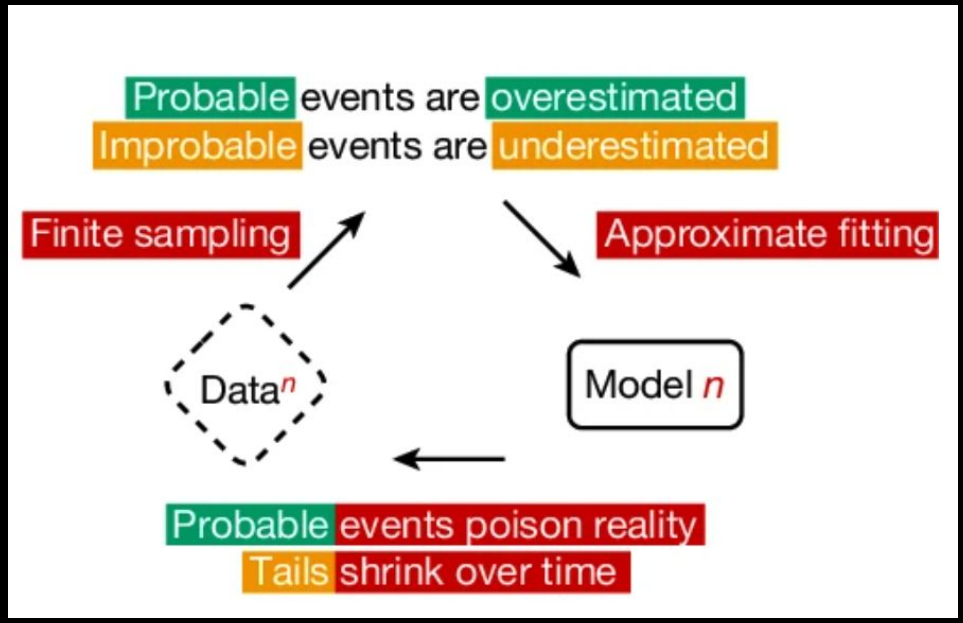
Generative Models: learn an approximation of the data distribution $p(\mathcal{D})$

[Source: Denoising Diffusion Probabilistic Models]

Overview

Generative Models: learn an approximation of the data distribution $p(\mathscr{D})$

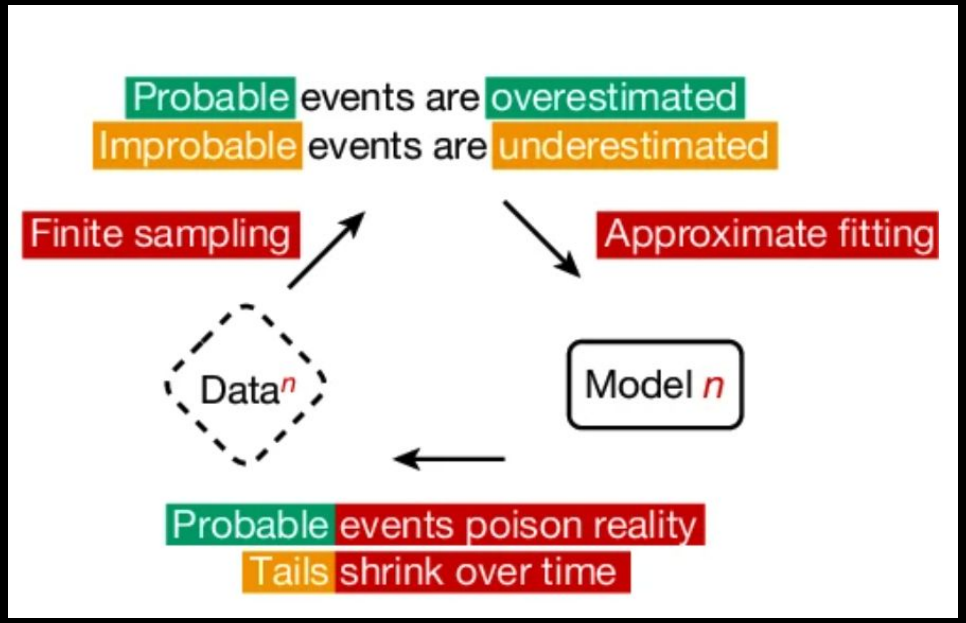
- **What's new**: when gen models are trained almost exclusively on the output of earlier models, learn a distorted data distribution.



Overview

Generative Models: learn an approximation of the data distribution $p(\mathscr{D})$

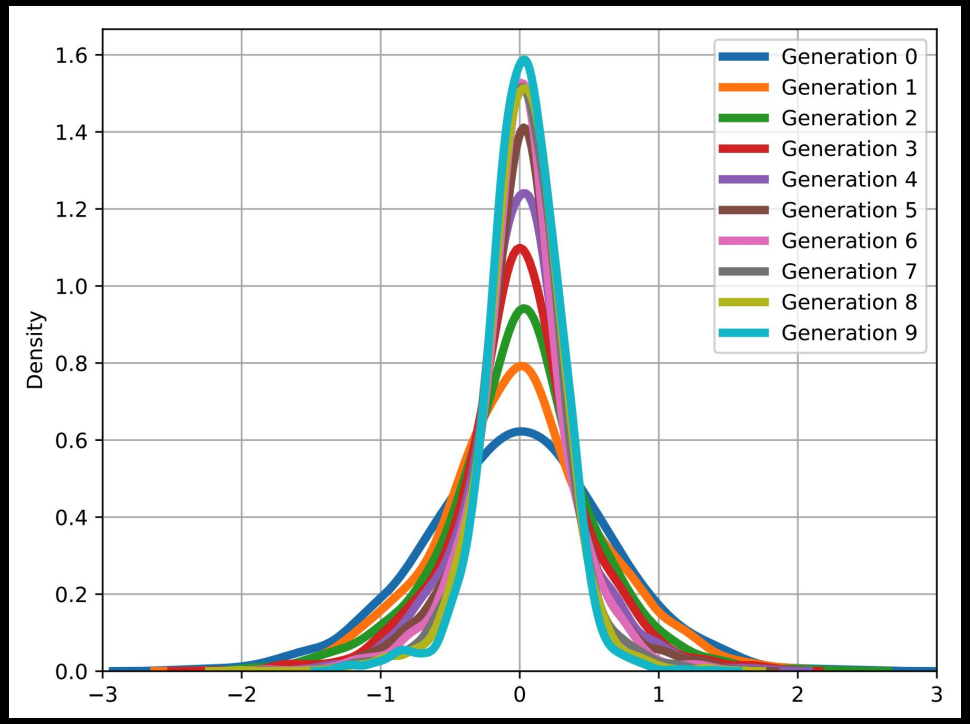
- **What's new**: when gen models are trained almost exclusively on the output of earlier models, learn a distorted data distribution.
- **Key insight**: Models struggle to generate rare examples from their training data and imperfectly model this data, causing distribution mismatch.
    - Repeated training amplifies error accumulations, defined as ***model collapse***.



Overview

Generative Models: learn an approximation of the data distribution $p(\mathscr{D})$

- **What's new**: when gen models are trained almost exclusively on the output of earlier models, learn a distorted data distribution.
- **Key insight**: Models struggle to generate rare examples from their training data and imperfectly model this data, causing distribution mismatch.
    - Repeated training amplifies error accumulations, defined as ***model collapse***.



Dynamics of Model Collapse

Model collapse is a degenerative process where models forget the true underlying data distribution over generations:

Dynamics of Model Collapse

Model collapse is a degenerative process where models forget the true underlying data distribution over generations:

- **Early Model Collapse:** Loss of information about the *tails* of the distribution.

Dynamics of Model Collapse

Model collapse is a degenerative process where models forget the true underlying data distribution over generations:

- **Early Model Collapse:** Loss of information about the tails of the distribution.
- **Late Model Collapse:** Convergence to a distribution that bears little resemblance to the original, often with reduced variance.

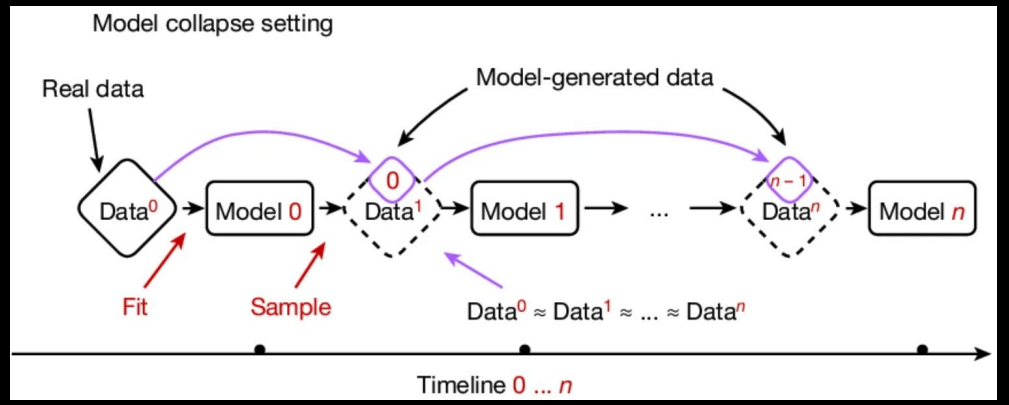
Dynamics of Model Collapse

Model collapse is a degenerative process where models forget the true underlying data distribution over generations:

- **Early Model Collapse:** Loss of information about the tails of the distribution.
- **Late Model Collapse:** Convergence to a distribution that bears little resemblance to the original, often with reduced variance.
- **Sources of Error:**
  - Statistical approximation error, arises from finite samples data.
  - Functional Expressivity Error, due to the limited expressive capacity of models.
  - Functional Approximation Error, from limitations in learning procedures.

How it works: Overview of the Training process

- The authors trained various models in a generational manner.
- **Generation 0:** Trained on a human-collected and curated dataset.

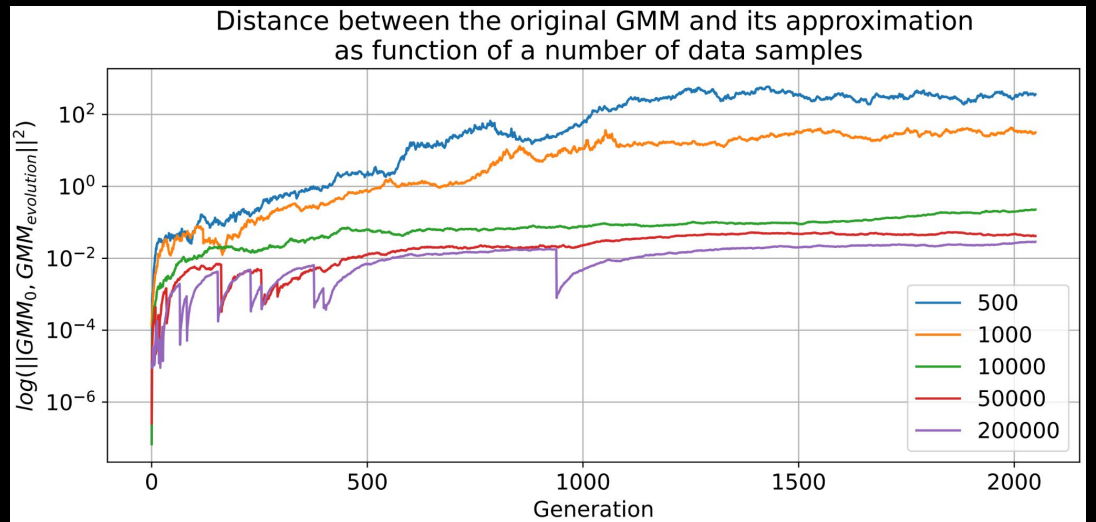


- **Subsequent Generations:** Each generation (1, 2, ...) was trained on the outputs of the previous generation.
- **Data Replacement:** In some cases, a fraction of generated examples was replaced with examples from the original training set.

Experiments: Gaussian Mixture Model (GMM)

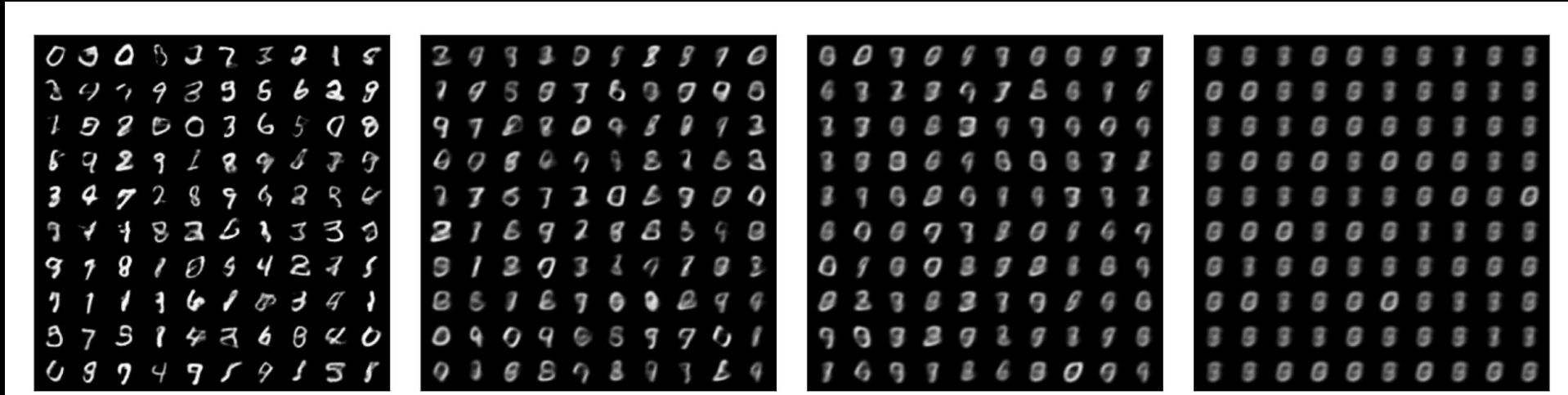
[See supplementary information]

- **Objective:** Fit input data assumed to come from two 2-dimensional Gaussian distributions.
- **Training Process:**
  - Trained 2,000 generations of GMMs.
  - Each generation used different number of sample generated previously.
  - **No Original Data:** The training was conducted without any original data.



Experiments: Variational Autoencoder (VAE)

- **Objective:** Generate MNIST digits.
- **Training Process:**
  - Trained over 20 generations of VAE.
  - Each generation was trained solely on the output produced by the previous generation.



(a) Original model (b) Generation 5 (c) Generation 10 (d) Generation 20

Experiments:
Fine-Tuning the OPT Language Model

- **Model Details:** Pretrained OPT language model with 125 million parameters.
- **Training Process:**
  - Fine-tuned on WikiText-2 for generation 0.
  - Subsequent generations (1-9) were trained under two conditions:
    - Only on examples produced by the previous generation.
    - On a mixture of 90% data from the previous generation and 10% original training data.
    - Models here are explicitly forced to not repeat sequences with a penalty of 2.0

Experiments: Fine-Tuning the OPT Language Model

Example of text outputs of an OPT-125m model affected by *Model collapse* – models degrade over generations, where each new generation is trained on data produced by the previous generation.

Input: some started before 1360 — was typically accomplished by a master mason and a small team of itinerant masons, supplemented by local parish labourers, according to Poyntz Wright. But other authors reject this model, suggesting instead that leading architects designed the parish church towers based on early examples of Perpendicular

Outputs:

Gen 0: Revival architecture such as St. John's Cathedral in London. The earliest surviving example of Perpendicular Revival architecture is found in the 18th @-@ century Church of Our Lady of Guernsey, which dates from the late 19th century. There are two types of perpendicular churches : those

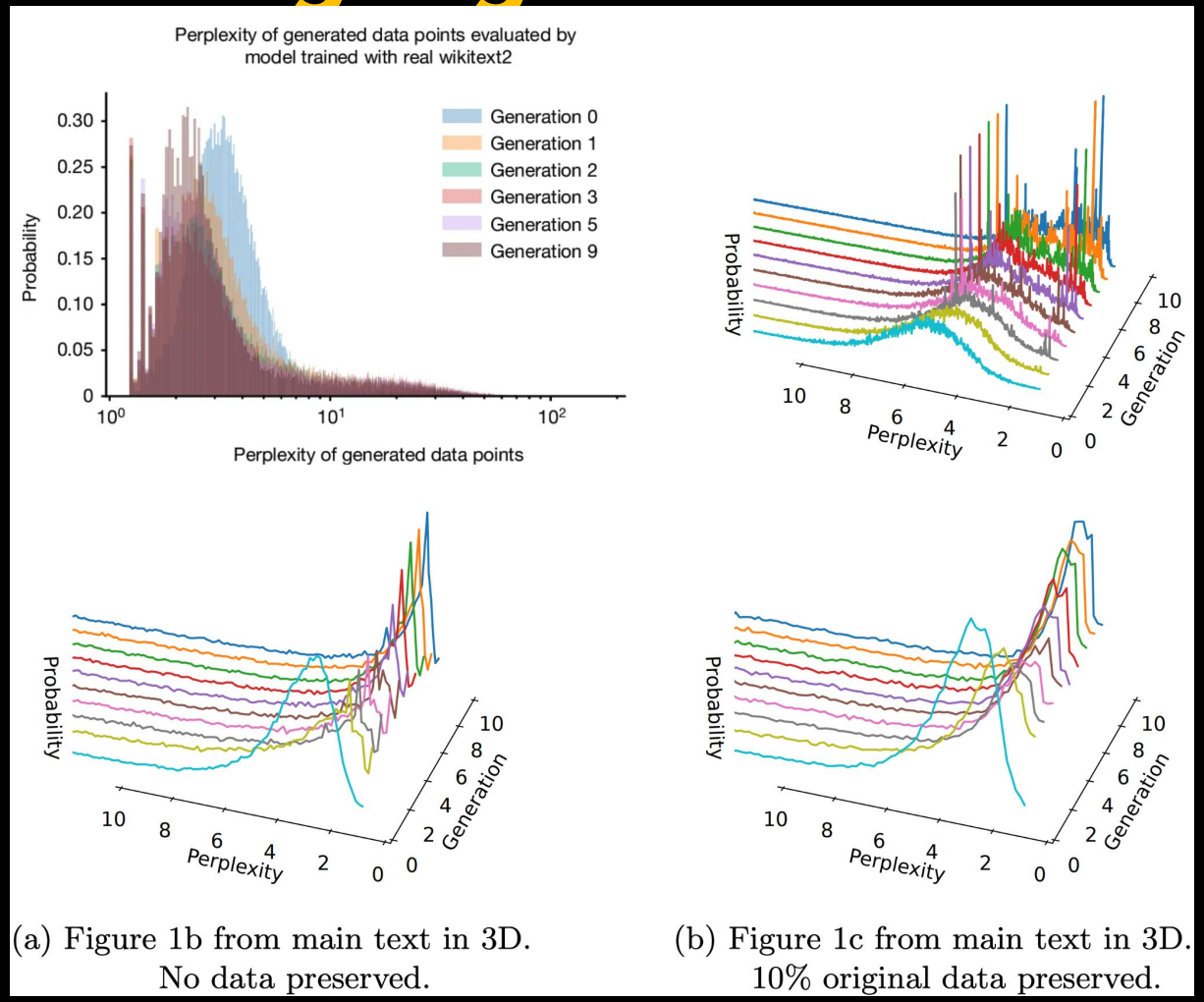
Gen 1: architecture such as St. Peter's Basilica in Rome or St. Peter's Basilica in Buenos Aires. There is no evidence that any of these buildings were built during the reign of Pope Innocent III, but it is possible that they may have been built during the reign of his successor, Pope Innocent

Gen 3: architecture such as St. Peter's Basilica (now part of the Church of Jesus Christ of Latter @-@ day Saints) or St. Peter's Basilica (now part of the Church of Jesus Christ of Latter @-@ day Saints). There is no evidence that any of these buildings

Gen 5: ism, which had been translated into more than 100 languages including English, French, German, Italian, Spanish, Portuguese, Dutch, Swedish, Norwegian, Polish, Hungarian, Slovak, Lithuanian, Estonian, Finnish, Romanian, Bulgarian, Turkish, Croatian, Serbian, Ukrainian, Russian, Kazakh, Kyrgyz

Gen 7: architecture in England. In an interview with The New York Times, Wright said : " I don 't think there is anything wrong with me being able to do what I want to do. It just doesn 't work for me. " He added : " I don 't know if you can call it funny,

Gen 9: architecture. In addition to being home to some of the world's largest populations of black @-@ tailed jackrabbits, white @-@ tailed jackrabbits, blue @-@ tailed jackrabbits, red @-@ tailed jackrabbits, yellow @-



(a) Figure 1b from main text in 3D. No data preserved.

(b) Figure 1c from main text in 3D. 10% original data preserved.

***Perplexity** quantifies how well a model predicts sequences in the test set; lower is better.

Why it matters?

Why it matters?

- **Efficiency in Training:**
  - High-quality generative models enable faster, cost-effective training using outputs from older models.

Why it matters?

- **Efficiency in Training:**
  - High-quality generative models enable faster, cost-effective training using outputs from older models.
- **Risks of Over-Reliance:**
  - Excessive dependence on generated data can diminish model capability and richness of insights.
  - Potential loss of real-world data diversity.

Why it matters?

- **Efficiency in Training:**
  - High-quality generative models enable faster, cost-effective training using outputs from older models.
- **Risks of Over-Reliance:**
  - Excessive dependence on generated data can diminish model capability and richness of insights.
  - Potential loss of real-world data diversity.
- **Impact of Web Content:**
  - Increasing web-based generated media may degrade model performance over time.

Yes, but

Generated data can still play a valuable role in training:

- Alpaca surpassed a pretrained LLaMA model by fine-tuning it on 52,000 examples produced by GPT-3.5.
- Alemohammad et al. introduced Self-Improving diffusion models with Synthetic data (SIMS), which utilize synthetic data for negative guidance to align models with real data distribution, enabling iterative training without falling into model autophagy disorder (MAD).

Initial Comments

- Even small amounts of malicious data in training can disrupt models, endangering fairness and the handling of rare events; original data access is crucial to prevent collapse.
  - Tracking data origins and preserving real-world data content is essential for robust AI model training.
- Future efforts should focus on integrating original and generated data to enhance model robustness and performance.

Thanks for the Attention!

Questions?

Let’s discuss

Contact:
lorenzo.valente@uni-hamburg.de