Escaping Model Collapse via Synthetic Data Verification: Near-term Improvements and Long-term Convergence

Bingji Yi[1*] Qiyuan Liu[2*] Yuwei Cheng[2] Haifeng Xu[3]

[1]Independent Researcher [2]Department of Statistics, University of Chicago [3]Department of Computer Science, University of Chicago

yibingji@gmail.com qiyuanliu@uchicago.edu yuweicheng@uchicago.edu haifengxu@uchicago.edu

The Synthetic Data Dilemma

Optimistic View

Training on synthetic data **improves** model performance across various domains. It reduces collection costs and enhances privacy protection.

[Gunasekar et al. 2023; Guo et al. 2024; Zelikman et al. 2022; Tian et al. 2023]

Pessimistic View

Recursively retraining on synthetic data causes **model collapse**–progressive loss of quality and diversity.

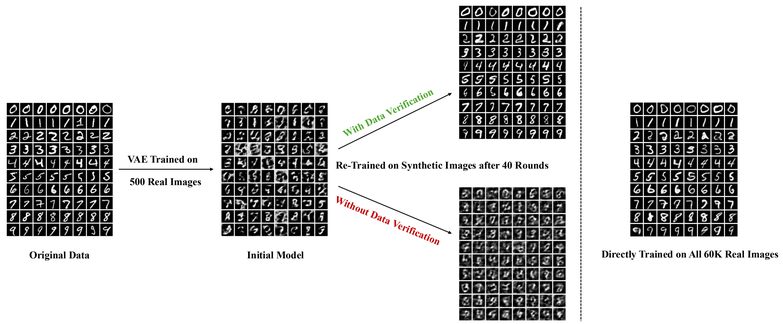
[Shumailov et al. 2024; Dohmatob et al. 2024; Alemohammad et al. 2023; Gerstgrasser et al. 2024]

What Causes the Discrepancy?

In practice, raw synthetic data is rarely used. Instead, a **knowledgeable verifier** filters out low-quality samples before training.

Verification may be the key missing factor.

Our theoretical framework for **verifier-based synthetic retraining** shows how an external verifier prevents model collapse and enables model improvement by injecting knowledge into the iterative retraining dynamics.



Verification prevents collapse and enables improvement. VAEs on MNIST: With verification, 40 rounds of synthetic retraining yields sharp digits. Without, quality collapses.

Our Approach: Verifier-Based Synthetic Retraining

A **verifier** (human/stronger model) provides binary accept/reject feedback on samples.

Main Results and Insights

- **Short-term theory:** Retraining with verified synthetic data can improve performance through bias–variance trade-off.
- **Long-term theory:** Iterative retraining converges to the verifier's knowledge center. Verifiers can prevent model collapse and yield near-term improvements, but performance will eventually plateau unless the verifier is perfectly unbiased.
- **Experimental validation:** Results across linear regression, VAEs on MNIST, and fine-tuning SmolLM2-135M on the XSUM task confirm these theoretical insights.

Theory: Bias–Variance Trade-off and Long-Term Convergence

Setup. We study iterative retraining on synthetic data in the linear regression model

$$y = x^\top \theta^* + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma^2),$$

where the goal is to estimate the true parameter θ^*.

Modeling the verifier and data filtering rule. The verifier's knowledge is modeled as a spherical ball with center θ_c and radius r:

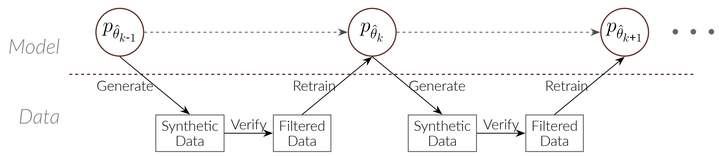
$$B_r(\theta_c) := \{\, \theta \in \mathbb{R}^p : \|\theta - \theta_c\| \le r \,\},$$

A verifier provides binary feedback, accepting a synthetic sample (x_i, y_i) if it is consistent with the verifier's knowledge:

$$|y_i - x_i^\top \theta_c| \le r\|x_i\| + \sigma_c.$$

$\|\theta^* - \theta_c\|$ is the **verifier bias**, r is **selectivity**, and σ_c is the verifier's estimate of σ.

The Generate–Verify–Retrain Scheme Start with $\hat{\theta}_0$ trained on real data.



One-step bias–variance trade-off. Filtering reduces variance but may introduce bias. The first-round estimator satisfies

$$\mathrm{MSE}(\hat{\theta}_1) = \underbrace{\text{synthetic variance}}_{\downarrow \text{ with more verified data}} + \underbrace{\text{verification error}}_{\uparrow \text{ with verifier bias}}.$$

Long-run behavior. The iteative retraining procedure induces a Markov process:

$$\hat{\theta}_{k+1} = T(\hat{\theta}_k) + \eta_{k+1},$$

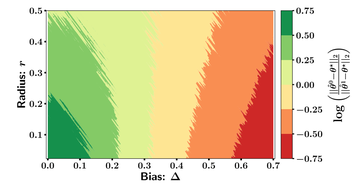
where $T(\cdot)$ is a **contraction** induced by verifier filtering, and η_{k+1} is sub-Gaussian noise due to synthetic generation. As $k \to \infty$, $\hat{\theta}_k \to \theta_c$.

Experiments

Experiment 1: Linear Regression Simulation

- $p = 8$, 100 real samples, 800 synthetic
- Verifier bias Δ and selectivity r varied
- Color: $\log(\|\hat{\theta}^0 - \theta^*\| / \|\hat{\theta}^1 - \theta^*\|)$

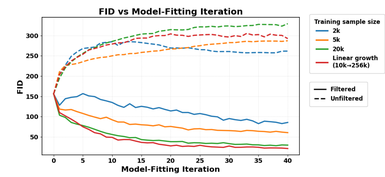
Green = improvement, Red = degradation.
Data verification enables improvement through bias–variance trade-off.



Experiment 2: MNIST Image Generation (VAE)

- VAE initialized on only 500 real images
- Discriminator retains top 10% per digit
- 40 rounds of iterative retraining

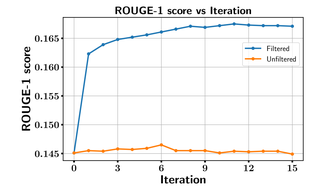
Verifiers prevent model collapse and enable improvement, but performance eventually plateaus due to verifier bias.



Experiment 3: XSUM Summarization (LLM)

- `SmolLM2-135M` fine-tuned on XSUM news summarization task
- Oracle verifier selects top 12.5% by ROUGE-1
- 15 rounds of iterative retraining

Iterative retraining with oracle filtering consistently improves summary quality.



Limitations & Future Directions

- Theoretical analysis is grounded in **linear regression** setting, formal generalization to broader model classes (neural networks, exponential families) remains open
- A global "true model" may not exist in non-parametric settings; defining optimality for generative models is an open challenge
- Current framework assumes a **static verifier**; studying adaptive or co-evolving verifiers is a promising direction

Paper

Code

*Equal contribution