Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026473481

# Model Collapse in Recursive Synthetic Data Training: Mechanisms, Evaluation, and Mitigation Strategies

**Renpu Tian***

*Northeastern University, Liaoning, China*

**Corresponding author: Renpu Tian.*

## Abstract

Generative models increasingly produce content that can enter later training corpora, creating recursive synthetic-data feedback loops. This review examines when such loops cause model collapse and when synthetic data remain useful. It synthesizes theoretical, empirical, and methodological evidence across language, diffusion, evaluation, mitigation, and governance, organized into four dimensions: training regime, degradation mechanism, outcome, and intervention point. Collapse is most consistently associated with repeated replacement of independent observations with unverified outputs, resulting in tail loss, declining diversity, cumulative error, and bias amplification; recursive diffusion studies also report a shift toward memorization. However, stable or beneficial outcomes occur when independent data or accumulated history are retained and generated samples are corrected, filtered, reweighted, or verified. Findings vary across model families, sampling policies, synthetic ratios, recursive depths, and metrics, limiting direct comparison and large-scale generalization. Priorities include standardized multigeneration benchmarks, multidimensional longitudinal evaluation, recovery research, realistic multimodel ecosystems, and provenance-aware governance. Overall, model collapse is conditional on the workflow rather than an inevitable consequence of synthetic data.

## Keywords

model collapse, recursive synthetic data training, synthetic data, distribution shift, data governance

## 1. Introduction

Generative artificial intelligence is changing both content production and future training corpora. Language models, diffusion systems, and related architectures generate text, images, code, annotations, and structured records at scale. When these outputs are scraped or deliberately reused, the central question is not the quality of one synthetic dataset but the stability of recursive training.

Synthetic data are attractive because independent observations can be costly, privacy-sensitive, or unevenly distributed. In tested settings, selected synthetic examples improved instruction-following, reasoning, mathematical training, and image classification [1–4], and LLMs supported annotation and text synthesis with task-specific prompting, filtering, and validation [5–6]. Synthetic origin alone, therefore, does not determine value, but one-round gains do not demonstrate multigeneration stability.

Recursive reuse creates distinct risks. Model collapse is an intergenerational process in which model-generated data cause successors to misrepresent the original distribution [7]. At the same time, self-consuming loops may lose quality or diversity as synthetic samples dominate [8]. Reported mechanisms fall into three classes: distributional narrowing, including tail loss and altered scaling [7,9]; inherited error and bias amplification [10–12]; and, in diffusion recursion, a generalization-to-memorization shift as synthetic-data entropy declines [13]. These outcomes remain conditional. Local stability requires assumptions about accuracy and independent-data retention [14]; accumulation, correction, verification, and detector-guided importance resampling stabilize tested loops [15–18]. Strong-collapse analysis instead derives an asymptotic error floor under persistent synthetic fractions [19]. The evidence is therefore regime-dependent, not a simple real–synthetic dichotomy.

Apparent disagreement also reflects incompatible designs. Replacement, fixed-ratio mixture, and accumulation preserve different amounts of independent information [14–15,20–23], whereas one-round augmentation [24] and unknown-origin contamination [18] address different estimands. Studies further vary in modality, scale, metric, and recursive horizon. This review asks: under which combinations of update regime, independent-data retention, generator quality, selection or verification, capacity, and recursive depth does reuse remain stable or beneficial, and when does it degrade the target distribution? It distinguishes recursive regimes; synthesizes distributional narrowing, error accumulation, and memorization; compares language and image evidence without equating one-round gains with recursive stability; and evaluates utility, fidelity, coverage, diversity, long-tail retention, memorization, fairness, privacy, and provenance. Mitigation is organized around data anchors, composition control, correction or verification, and lifecycle governance.

### 1.1 Review Scope and Method

This article is a focused narrative review rather than a preregistered systematic review. It examines peer-reviewed theoretical, empirical, evaluation, and governance studies published through 2025 that directly address recursive synthetic-data training or provide methods for assessing it. Studies are compared by update regime, retained independent information, mechanism, outcome, and intervention. The corpus supports cross-study synthesis rather than exhaustive coverage, and absence-of-evidence claims are interpreted accordingly.

## 2. Theoretical Background

### 2.1 Recursive Learning and Distribution Estimation

Let $P_0$ denote the independent target distribution, $P_t$ the output distribution of generation t, and $Q_{t+1}$ the idealized distribution used to train the next model. If $\alpha$ is the current synthetic fraction and $n_k$ is the number of retained samples from generation k, the three common regimes can be written as follows:

$$Q_{t+1}^{\mathrm{rep}} = P_t.$$

$$Q_{t+1}^{\mathrm{mix}} = (1-\alpha)P_0 + \alpha P_t, \qquad 0 \le \alpha \le 1.$$

$$Q_{t+1}^{\mathrm{acc}} = \frac{n_0 P_0 + \sum_{k=1}^{t} n_k P_k}{n_0 + \sum_{k=1}^{t} n_k}.$$

Replacement uses only $P_t$; mixture retains a fixed $P_0$ anchor; and accumulation keeps earlier data in proportion to retained sample counts [7,14–15]. Because each $P_t$ is learned from finite samples in an imperfect model class, recursion can compound statistical, approximation, and optimization error [7,10,20]. Heterogeneous generator–learner pairs may share different errors.

### 2.2 Sampling and Information Loss

Finite sampling can turn small estimation errors into irreversible information loss. Low-probability events are least likely to appear, and, under replacement, an omitted event provides no evidence to later learners. Recursion can therefore contract support and variance while reinforcing recurrent artifacts [7,9,23].

Sampling policy controls this trade-off. Across 10 LLM generations, Drayson et al. compared untruncated sampling with temperature ($\tau$ = 0.9), top-k (k = 50), and top-p (p = 0.95) sampling. In the tested settings,

temperature and truncation produced increasingly peaked, less diverse distributions; untruncated sampling weakened narrowing but generated high-perplexity, incoherent tail text [18]. Broader sampling can improve tail coverage at the cost of validity, so independent data, correction, or verification are needed to prevent selection errors from becoming inherited targets [14,16–17].

### 2.3 Scaling Laws and Stability

Conventional scaling laws predict lower error with more data or capacity, but recursive contamination can introduce an error component that additional synthetic samples do not remove. Analyses therefore distinguish finite-sample gains from asymptotic bias or error floors relative to $P_0$ [9,19]. Generator quality, verification accuracy, mixture ratio, and capacity can also create transitions between beneficial and harmful regimes [17,22]; more synthetic data may improve a finite budget while shifting the long-run target.

### 2.4 Theoretical Assumptions

Theoretical results are conditional on the update regime and assumptions about data retention, sample independence, model capacity, and verification quality. Local or finite-generation stability should not be read as global or asymptotic equivalence to $P_0$, nor should an asymptotic error floor imply immediate failure [14,19–22].

## 3. Literature Review

Studies are compared by update regime, retained independent information, mechanism, and outcome. Replacement removes omitted support, whereas mixture and accumulation preserve external anchors [7,14–15]. The sections below, therefore, treat regime as a moderator and focus on distributional narrowing, error accumulation, memorization, and model-family differences.

### 3.1 Distributional Narrowing

Distributional narrowing is the most consistent recursive signature. Low-probability events are readily omitted from finite synthetic samples, so repeated learning can contract support and reduce lexical, semantic, or visual diversity before aggregate quality metrics deteriorate [7–9,11]. Fairness feedback can likewise amplify disparities when weakly represented regions lose further support [12]. Evidence is strongest for replacement, in which omitted information cannot be recovered [7–8]. Independent data mixtures and accumulation can mitigate the bottleneck under stated conditions, but do not guarantee the preservation of the full target distribution [14–15,19]. Plausible outputs may therefore coexist with losses of rare vocabulary, minority classes, unusual compositions, or subgroup knowledge.

### 3.2 Error Accumulation

Error accumulation turns finite-sample, approximation, and optimization errors into inherited targets. Regression and general analyses distinguish label noise, covariate error, recursive depth, and conditions under which errors contract or propagate [7,10,20]. Empirical studies further suggest that outcomes depend on corpus composition, generator behavior, sampling policy, and selection procedures, rather than on synthetic proportion alone [11–12,18,23]. Recursive learning can amplify particular tendencies without forcing every error or bias to change monotonically.

Propagation depends on the full pipeline. Image contamination can cause class-dependent degradation, while external correction and detector-guided importance resampling stabilize tested loops through different signals [16,18,25]. Evidence largely comes from controlled lineages; operational ecosystems may include multiple generators, human editing, retrieval, deduplication, and changing sources. Such uncertainty in origin can confound recursive reinforcement learning with domain shift, while detector errors may introduce selection bias [26–27].

### 3.3 Generalization-to-Memorization Transition

Recursive diffusion studies identify a transition from novel generation to reproduction as the entropy of synthetic data declines [13]. This complements evidence that tail and diversity loss can precede deterioration

in aggregate quality [7–8,11]: common patterns remain plausible while uncommon alternatives weaken. Direct evidence for this intergenerational transition is concentrated in diffusion models. Authenticity and distributional precision–recall can diagnose copying and coverage, but cannot alone establish the transition [28–29]; its generality remains unresolved.

## 3.4 Differences across Model Families

Language-model studies link corpus composition and decoding to linguistic diversity and task performance, with conditional stabilization from detector-guided importance resampling [11,18,23]. These multigeneration results differ from one-round uses in which selected synthetic text supports classification or pretraining [6,24]; human–model co-writing separately suggests reduced content diversity [30]. Image and diffusion studies likewise report support loss, class-dependent contamination, and memorization alongside benefits from high-quality augmentation [4,8,13–14,25]. No universal cross-modal threshold holds: decoding, guidance, data retention, representation, and evaluation differ, and evidence from frontier-scale recursive pretraining remains scarce.

## 3.5 Evaluation

Evaluation should separate fidelity, coverage, utility, novelty, and risk across generations. Table 1 summarizes the principal measures.

*Table 1: Compact evaluation framework for recursive synthetic-data studies.*

| Dimension | Representative measures | What it reveals | Main caution |
|---|---|---|---|
| Fidelity | FID; KID; precision and density [24,29,32–34] | Sample quality and closeness to a reference distribution | Feature- and estimator-dependent; scalar distances may hide missing modes |
| Coverage and diversity | Recall and coverage; Vendi Score; lexical or semantic diversity [11,29,32,34,35] | Retention of modes, tails, and variation | Sensitive to representation, geometry, and sample size; diversity can reflect noise |
| Utility | Task scores; general and specific utility [36,37] | Usefulness for downstream analysis or prediction | Aggregate utility can remain high while rare modes disappear |
| Novelty and memorization | Authenticity and reproduction tests [13,28] | Generation of new samples rather than copying | Requires lineage or reference access and defensible similarity thresholds |
| Risk and traceability | Subgroup and privacy audits; provenance and documentation checks [12,38–42] | Differential harm, disclosure, and data origin | Cannot be inferred from aggregate fidelity or utility |

A longitudinal protocol should report recursive depth, real–synthetic composition, generator identity, sampling and filtering policies, and matched non-recursive controls [26–27]. Because tail loss may precede utility decline, every generation should be evaluated. No model-family-independent collapse criterion exists; fixed-reference representations should be triangulated with coverage measures, expert assessment, downstream utility, and uncertainty estimates [29,31].

## 3.6 Mitigation

Mitigation is strongest when independent information is preserved. Independent data mixtures, accumulation, and fresh observations can stabilize tested loops or avoid fixed-size replacement [8,14–15], although persistent synthetic fractions may still introduce limiting error under strong-collapse assumptions [19]. Correction, detector-guided importance resampling, verification, and entropy-aware selection intervene at different stages [13,16–18,21]. Their guarantees differ: correction requires external knowledge, detection faces distribution shift, and verification depends on calibration. No generally safe synthetic ratio or filtering rule has been established.

A threshold example illustrates the precision–coverage trade-off. Suppose a verifier accepts only scores of at least 0.95. A valid rare-dialect sentence might score 0.80 because calibration data mainly represent majority-language text and would therefore be removed, while common high-confidence variants remain. Average accepted-sample validity could rise as tail and subgroup coverage fall. The values are illustrative; the risk

arises from detector shifts, subgroup imbalances, and verifier blind spots [12,17–18]. Mitigation should therefore be evaluated as a multi-objective trade-off rather than by acceptance precision alone.

Provenance and governance make recursive exposure measurable but do not establish sample quality. Data statements, model cards, and datasheets document origin, composition, and limitations [32–34], while watermarking and zero-shot detection support identification under domain- and transformation-dependent conditions [35–37]. Sensitive settings require joint assessment of utility, privacy, and domain validity [38]. Lineage records should therefore be combined with independent anchors, selective acceptance, and longitudinal audits.

### 3.7 Recovery

Evidence is stronger for prevention than for recovery after support disappears. Existing studies retain or reintroduce independent data, correct samples, or filter them before severe degradation [14–18]; they do not show that a rare mode absent from both the dataset and the model can be reconstructed without new evidence. Stabilization and recovery are therefore distinct claims.

Recovery studies should quantify the independent information needed to restore tail coverage, distinguish reweighting of retained rare modes from recollection of absent ones, and track novelty, subgroup performance, and newly introduced errors. Preventive success should not be cited as proof that collapse is reversible.

## 4. Discussion and Synthesis

### 4.1 Cross-Study Synthesis

The principal disagreement is not whether recursive training can fail, but which outcome is treated as evidence of stability and over what horizon. Replacement studies examine support loss; stability analyses establish conditional behavior under retained-data assumptions; verification studies vary acceptance quality; and strong-collapse theory characterizes limiting error [7–9,14,17,19–20]. Finite-generation utility, local stability, asymptotic error floors, and full-support preservation are distinct claims. Conclusions should therefore be indexed to update the rule, horizon, generator, and verifier quality, and protected outcome.

Design choices reinforce apparent contradictions. Controlled regression and bounded-distribution studies aid causal isolation but do not reproduce web-scale heterogeneity [10,23,26]. Multigeneration replacement exposes feedback, whereas one-round augmentation tests sample quality and task relevance [7,24]. Accuracy or perceptual quality may remain high while rare modes, diversity, or novelty decline, and image and language studies use different representations and metrics [11,13,29,31,39–40]. Seemingly inconsistent results may therefore estimate different target quantities rather than oppose one another.

### 4.2 Methodological Limitations and Research Gaps

External validity is limited by bounded datasets, short recursive horizons, and closed lineages. Operational ecosystems may combine multiple generators, human editing, retrieval, deduplication, topic switching, and uncertain provenance; controlled findings do not establish transferability to these open, non-stationary settings [26–27].

Causal identification is weakened when recursive and non-recursive datasets are unmatched for size, domain, quality, and sampling policy. Effects attributed to recursion may instead reflect domain shift, filtering, or unequal quality. Image-contamination and language-pretraining protocols, therefore, motivate separate identification of data anchoring, recursion, and selection [24–25].

Feature encoders, sample sizes, decoding policies, test contamination, and definitions of collapse constrain the consistency of evaluation. Fairness, privacy, provenance, utility, and coverage are often studied separately [12,26,38,41]. Shared generation-level reporting and stopping criteria are needed before cross-modal comparison.

Evidence at scale also remains limited. Few studies jointly vary capacity, synthetic proportion, generator quality, verification accuracy, and recursive depth, and multigeneration foundation-model pretraining remains scarce. One-round gains should not be treated as evidence of long-run recursive stability [18,24]. Recovery after support loss is similarly under-tested.

### 4.3 Future Research and Adaptive Framework

Future work should use factorial longitudinal designs that vary update regime, synthetic proportion, generator quality, verification accuracy, capacity, and sampling policy under fixed budgets. Benchmarks should contain known tails, subgroup labels, memorization probes, generation-level checkpoints, and equal-budget controls. Pre-registered collapse definitions would improve comparison. Modality-specific measures remain necessary for distinct representations and risks [12–13,28–29,31,41–42].

Based on the reviewed evidence, this review proposes a five-layer recursive-data lifecycle: information source and provenance; recursion design; sampling and selection; learning dynamics; and outcome auditing. It links data independence and lineage to replacement, mixture, or accumulation; then to filtering, correction, verification, and weighting; and finally to utility, fidelity, coverage, diversity, memorization, fairness, and privacy. The framework identifies where independent information enters, inherited error is amplified, and interventions can be evaluated [14,16–17,34].

Adaptive policies require observable triggers. Candidate signals include declining tail recall or coverage, rising held-out verification error, and increasing uncertainty about sample origin; they should be monitored jointly because no single signal is sufficient [17,26,29,31,34]. A policy might lower the synthetic fraction, broaden independent-data sampling, or tighten verification when coverage falls and verification error rises. These are proposed triggers, not universal cutoffs; research must estimate pipeline-specific thresholds, false-alarm costs, and preventive value.

## 5. Conclusion

The evidence supports a conditional conclusion: recursive synthetic-data training is neither inherently self-destructive nor automatically sustainable. Risk is greatest when unverified outputs repeatedly replace independent observations, turning omissions, approximation error, and bias into inherited targets [7–9]. Stability is more plausible when independent data or accumulated history are retained and generated samples are reliably corrected or verified [14–17]. Long-run behavior, therefore, depends on access to independent, validated information rather than on a binary distinction between human- and machine-generated data.

The main limitations are controlled datasets, closed lineages, short horizons, incompatible evaluation, and scarce multigeneration evidence at pretraining scale [18,24,26]. Transfer to ecosystems with multiple generators, human editing, retrieval, and evolving sources remains uncertain. Recovery is especially unresolved: after rare modes or minority patterns disappear, the independent evidence needed to restore them is unknown. Aggregate utility should therefore be interpreted alongside coverage, memorization, fairness, privacy, and provenance [12–13,28,34,41–42].

Priority should be given to standardized multigeneration benchmarks with matched controls, traceable lineages, generation-level reporting, and multidimensional metrics. Adaptive governance should connect intervention to pipeline-specific signals such as declining tail coverage, rising verification error, or increasing provenance uncertainty [17,27,29,31,34]. These signals can guide synthetic-data ratios, independent collection, correction, or verification. Recovery studies should determine when lost support can be reconstructed and when a new collection is necessary. The goal is a recursive learning ecosystem that remains observable, traceable, and correctable.

## References

[1] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. “Self-Instruct: Aligning Language Models with Self-Generated Instructions.” In Proceedings of ACL 2023, 13484–13508. https://doi.org/10.18653/v1/2023.acl-long.754.

[2] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. “STaR: Bootstrapping Reasoning With Reasoning.” In Advances in Neural Information Processing Systems 35 (2022): 15476–15488.

[3] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. “MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models.” In The Twelfth International Conference on Learning Representations (ICLR), 2024.

[4] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. “Synthetic Data from Diffusion Models Improves ImageNet Classification.” Transactions on Machine Learning Research, 2023.

[5] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. “Large Language Models for Data Annotation and Synthesis: A Survey.” In Proceedings of EMNLP 2024, 930–957. https://doi.org/10.18653/v1/2024.emnlp-main.54.

[6] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. “Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations.” In Proceedings of EMNLP 2023, 10443–10461. https://doi.org/10.18653/v1/2023.emnlp-main.647.

[7] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. “AI models collapse when trained on recursively generated data.” Nature 631 (2024): 755–759. https://doi.org/10.1038/s41586-024-07566-y.

[8] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. “Self-Consuming Generative Models Go MAD.” In The Twelfth International Conference on Learning Representations (ICLR), 2024.

[9] Elvis Dohmatob, Yunzhen Feng, Pu Yang, François Charton, and Julia Kempe. “A Tale of Tails: Model Collapse as a Change of Scaling Laws.” In Proceedings of the 41st International Conference on Machine Learning, PMLR 235 (2024): 11165–11197.

[10] Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. “Model Collapse Demystified: The Case of Regression.” Advances in Neural Information Processing Systems 37 (2024): 46979–47013. https://doi.org/10.52202/079017-1490.

[11] Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. “The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text.” In Findings of the Association for Computational Linguistics: NAACL 2024, 3589–3604. https://doi.org/10.18653/v1/2024.findings-naacl.228.

[12] Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. “Fairness Feedback Loops: Training on Synthetic Data Amplifies Bias.” In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 2113–2147. https://doi.org/10.1145/3630106.3659029.

[13] Lianghe Shi, Meng Wu, Huijie Zhang, Zekai Zhang, Molei Tao, and Qing Qu. “A Closer Look at Model Collapse: From a Generalization-to-Memorization Perspective.” Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Main Conference Track, 2025.

[14] Quentin Bertrand, Avishek (Joey) Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. “On the Stability of Iterative Retraining of Generative Models on their own Data.” In The Twelfth International Conference on Learning Representations (ICLR), 2024.

[15] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Dan Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. “Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data.” In First Conference on Language Modeling (COLM), 2024.

[16] Nate Gillman, Michael Freeman, Daksh Aggarwal, Chia-Hong Hsu, Calvin Luo, Yonglong Tian, and Chen Sun. “Self-Correcting Self-Consuming Loops for Generative Model Training.” In Proceedings of the 41st International Conference on Machine Learning, PMLR 235 (2024): 15646–15677.

[17] Yunzhen Feng, Elvis Dohmatob, Pu Yang, François Charton, and Julia Kempe. “Beyond Model Collapse: Scaling Up with Synthesized Data Requires Verification.” In The Thirteenth International Conference on Learning Representations (ICLR), 2025.

[18] George Drayson, Emine Yilmaz, and Vasileios Lampos. “Machine-generated text detection prevents language model collapse.” In Proceedings of EMNLP 2025, 29657–29673. https://doi.org/10.18653/v1/2025.emnlp-main.1506.

[19] Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. “Strong Model Collapse.” In The Thirteenth International Conference on Learning Representations (ICLR), 2025.

[20] Shi Fu, Sen Zhang, Yingjie Wang, Xinmei Tian, and Dacheng Tao. “Towards Theoretical Understandings of Self-Consuming Generative Models.” In Proceedings of the 41st International Conference on Machine Learning, PMLR 235 (2024): 14228–14255.

[21] Shi Fu, Yingjie Wang, Yuzhu Chen, Xinmei Tian, and Dacheng Tao. “A Theoretical Perspective: How to Prevent Model Collapse in Self-consuming Training Loops.” In The Thirteenth International Conference on Learning Representations (ICLR), 2025.

[22] Aymane El Firdoussi, Mohamed El Amine Seddik, Soufiane Hayou, Reda Alami, Ahmed Alzubaidi, and Hakim Hacid. “Maximizing the Potential of Synthetic Data: Insights from Random Matrix Theory.” In The Thirteenth International Conference on Learning Representations (ICLR), 2025.

[23] Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Abdelkader Debbah. “How bad is training on synthetic data? A statistical analysis of language model collapse.” In First Conference on Language Modeling (COLM), 2024.

[24] Feiyang Kang, Newsha Ardalani, Michael Kuchnik, Youssef Emad, Mostafa Elhoushi, Shubhabrata Sengupta, Shang-Wen Li, Ramya Raghavendra, Ruoxi Jia, and Carole-Jean Wu. “Demystifying Synthetic Data in LLM Pre-training: A Systematic Study of Scaling Laws, Benefits, and Pitfalls.” In Proceedings of EMNLP 2025, 10739–10758. https://doi.org/10.18653/v1/2025.emnlp-main.544.

[25] Ryuichiro Hataya, Han Bao, and Hiromi Arai. “Will Large-scale Generative Models Corrupt Future Datasets?” In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, 20555–20565. https://doi.org/10.1109/ICCV51070.2023.01879.

[26] Xiaodan Xing, Fadong Shi, Jiahao Huang, Yinzhe Wu, Yang Nan, Sheng Zhang, Yingying Fang, Michael Roberts, Carola-Bibiane Schönlieb, Javier Del Ser, and Guang Yang. “On the caveats of AI autophagy.” Nature Machine Intelligence 7 (2025): 172–180. https://doi.org/10.1038/s42256-025-00984-1.

[27] Boris Van Breugel, Zhaozhi Qian, and Mihaela Van Der Schaar. “Synthetic Data, Real Errors: How (Not) to Publish and Use Synthetic Data.” In Proceedings of the 40th International Conference on Machine Learning, PMLR 202 (2023): 34793–34808.

[28] Ahmed M. Alaa, Boris van Breugel, Evgeny Saveliev, and Mihaela van der Schaar. “How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models.” In Proceedings of the 39th International Conference on Machine Learning, PMLR 162 (2022): 290–306.

[29] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. “Assessing Generative Models via Precision and Recall.” Advances in Neural Information Processing Systems 31 (2018): 5228–5237.

[30] Vishakh Padmakumar and He He. “Does Writing with Language Models Reduce Content Diversity?” In The Twelfth International Conference on Learning Representations (ICLR), 2024.

[31] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. “Reliable Fidelity and Diversity Metrics for Generative Models.” In Proceedings of the 37th International Conference on Machine Learning, PMLR 119 (2020): 7176–7185.

[32] Emily M. Bender and Batya Friedman. “Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science.” Transactions of the Association for Computational Linguistics 6 (2018): 587–604. https://doi.org/10.1162/tacl_a_00041.

[33] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. “Model Cards for Model Reporting.” In Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, 220–229. https://doi.org/10.1145/3287560.3287596.

[34] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. “Datasheets for Datasets.” Communications of the ACM 64, no. 12 (2021): 86–92. https://doi.org/10.1145/3458723.

[35] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. "A Watermark for Large Language Models." In Proceedings of the 40th International Conference on Machine Learning, PMLR 202 (2023): 17061–17084.

[36] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature." In Proceedings of the 40th International Conference on Machine Learning, PMLR 202 (2023): 24950–24962.

[37] Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Merey, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, Ilia Shumailov, Ciprian Baetu, Sven Gowal, Demis Hassabis, and Pushmeet Kohli. "Scalable watermarking for identifying large language model outputs." Nature 634 (2024): 818–823. https://doi.org/10.1038/s41586-024-08025-4.

[38] Mauro Giuffrè and Dennis L. Shung. "Harnessing the power of synthetic data in healthcare: innovation, application, and privacy." npj Digital Medicine 6 (2023): 186. https://doi.org/10.1038/s41746-023-00927-3.

[39] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium." In Advances in Neural Information Processing Systems 30, 2017, 6626–6637.

[40] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. "Improved Precision and Recall Metric for Assessing Generative Models." Advances in Neural Information Processing Systems 32 (2019): 3927–3936.

[41] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. "Synthetic Data—Anonymisation Groundhog Day." In 31st USENIX Security Symposium, 2022, 1451–1468.

[42] Joshua Snoke, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavković. "General and Specific Utility Measures for Synthetic Data." Journal of the Royal Statistical Society: Series A 181, no. 3 (2018): 663–688. https://doi.org/10.1111/rssa.12358.

## Funding

This research received no external funding.

## Conflicts of Interest

The authors declare no conflict of interest.

## Acknowledgment

This paper is an output of the science project.

## Open Access