Data

Andrew Shi

February 4, 2025

1 Data Exhaustion

The 2024 AI Index Report cited researchers from Epoch estimating the depletion of high-quality language data by 2024, low-quality language data within two decades, and image data by the late 2030s to mid-2040s [3]. The updated report from Epoch, published in June 2024, estimates that we will run out of public human text data between 2026-2032 [1]. To study this problem, the researchers evaluated the following parameters: (1) growth of the indexed web, (2) growth in the number of internet users, (3) historical dataset growth rate, and (4) projected compute growth rate.

In each of these categories, there are further considerations. For (1), it is important to note that not all internet data can/should be used to train models, as bad data degrades model output quality. However, there are techniques like data filtering and data deduplication that can be used to salvage a larger proportion of internet data. For (2), the growth of the human population and the percentage of the population that uses the internet were inputs to the model. For (3) and (4), it is important to note that growth in computing power and data supply is often correlated.

2 Synthetic Data

2.1 Model Degradation Resulting from Synthetic Data

Using synthetic data—data generated by AI models themselves—to train models has been an area of interest. The 2024 AI Index Report suggests that there are limitations associated with this approach, namely that it is likely to lose representation of the tails of distributions when performing repeated training cycles on synthetic data, leading to degraded model output quality. This phenomenon was observed across different model architectures, including variational autoencoders (VAEs), Gausian Mixture Models (GMMs), and Large Language Models (LLMs) [3].

This year, there have been advances in generating high-fidelity synthetic data, albeit synthetic data is still generally distinguishable from real data and there is no existing scalable method to achieve the same performance training LLMs on synthetic data compared to real data. Additionally, researchers have studied characterizations on the trade-off between data fidelity and data utility.

Promising new methods have been proposed to generate synthetic data that can be used to train models. Vine Copulas, Bayesian Hierarchical Generalized Linear Models, tree ensembles, and GMMs have all been used to generate synthetic data. Generally, statistical and probabilistic methods are known to be straightforward and interpretable but often oversimplify complex relationships. More recently, deep learning architectures like VAEs and Generative Adversarial Networks (GANs) have been trained to generate large-scale, high-dimensional data. In the medical domain, models like ADS-GAN, CTGAN, and MedGAN have enhanced performance on classification and prediction tasks by training

on synthetically-augmented datasets, increasing F1 scores or AUROC by 5–10% on minority classes [4].

[2] compares the performance of models trained on synthetic and real data across multiple architectures and datasets. More specifically, they evaluate how well synthetic relational data preserve key characteristics of the original data ("fidelity") and remain useful for downstream tasks ("utility"). They find that most methods are systematically detectable as synthetic, especially once relational information is considered. Furthermore, performance typically deteriorates compared to real-data–trained models, but some methods still yield moderately good predictive scores. On a few experiments, synthetic data outperforms real data such as using Synthetic Data Vault (SDV) vs. Walmart data to train an XGBoost classifier. The researchers show that training on the synthetic dataset achieves a lower mean squared error (MSE). Although the evidence is currently scarce, it is promising that synthetic data has been successful in rare cases and could potentially improve in the future to augment data availability and access.

2.2 Hallucinations in Synthetic Data

There are concerns around the quality and fidelity of synthetically generated data, as LLMs are known to hallucinate and provide factually-incorrect outputs. When training on hallucinated content in datasets, models can experience compounded degradation in output quality. New techniques have been developed to combat this issue. [6] uses automated fact-checking and confidence scores to rank factuality scores of model response pairs. Human-in-the-loop approaches to label preferred responses have also been used to aligning language models, but this method is expensive. Finally, post-hoc filtering and debiasing methods are used to remove anomalies in synthetic data before the training stage.

3 Policies around Data Access

This year, policies around data access for training and deploying LLMs have become more restrictive, reflecting growing concerns about user privacy and organizational security [7]. Recent trends include integrating federated learning to keep data locally on user devices rather than pooling it on a central server, and applying differential privacy techniques to obscure individual records within large datasets. Furthermore, certain organizations have promoted "knowledge unlearning": removing data from a model that effectively promote "forgetting" data post-training in models. Data governance has become a core requirement, shaping how firms structure their data pipelines, model training practices, and user-facing applications. Thus,

Due to restrictions around data access, researchers have studied new methods to train models without access to the original dataset. For example, [5] uses a distillation-based approach: they first train a teacher VAE on a specific objective and then feed Gaussian noise into a student model that learns to mimic the representations produced by the teacher model. At inference time, the student model generates latent representations, which are then fed into the VAE decoder to reconstruct synthetic samples. In this framework, the original data is not needed to train the student model.

References

[1] T. Besiroglu, L. Heim, and J. Sevilla. Projecting compute trends in machine learning, 2022. Accessed: 2025-01-23.

[2] V. Hudovernik, M. Jurkovič, and E. Štrumbelj. Benchmarking the fidelity and utility of synthetic relational data, 2024.

[3] N. Maslej, L. Fattorini, R. Perrault, V. Parli, A. Reuel, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, J. C. Niebles, Y. Shoham, R. Wald, and J. Clark. Artificial intelligence index report 2024, 2024.

[4] V. C. Pezoulas, D. I. Zaridis, E. Mylona, C. Androutsos, K. Apostolidis, N. S. Tachos, and D. I. Fotiadis. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal*, 23:2892–2910, 2024.

[5] M. F. Sikder, D. de Leng, and F. Heintz. Fair4free: Generating high-fidelity fair synthetic samples using data free distillation, 2024.

[6] K. Tian, E. Mitchell, H. Yao, C. D. Manning, and C. Finn. Fine-tuning language models for factuality. *ArXiv*, abs/2311.08401, 2023.

[7] B. Yan, K. Li, M. Xu, Y. Dong, Y. Zhang, Z. Ren, and X. Cheng. On protecting the data privacy of large language models (llms): A survey, 2024.