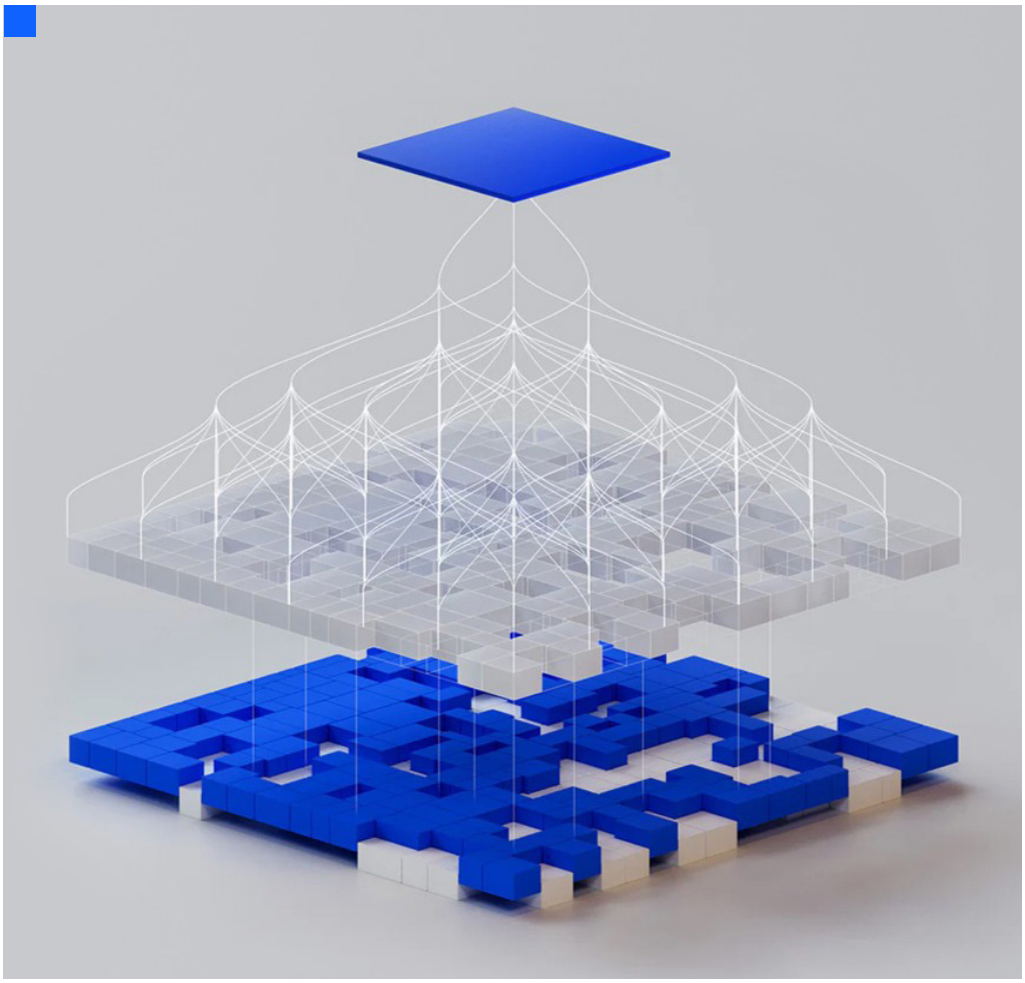
IBM Responsible Technology Board

Unlocking AI Opportunities with the Responsible Generation and Use of Synthetic Data



Attributions

With gratitude to,

Responsible Technology Board Co-Chairs and workstream's Executive Sponsors: Betsy Greytok, Francesca Rossi and Stacy Hobson

Contributors: Adriana Alvarado Garcia, Alina Glaubitz, Aliza Heching, Andrea Greggo, Eda Kavlakoglu, Enikõ Rózsa, Erik Altman, Jamie VanDodick, Joy Deng, Kristi Spess, Muneeza Azmat, Perla Khattar, Saishruthi Swaminathan, Rogerio Abreu de Paula

Acknowledgments

The Co-chairs, executive sponsors, and contributors would like to acknowledge the following members for their review and feedback: Benjamin Montague, Christina Shim, Dan Waugh, George Khour, Jordan Youngblood, Kevin Black, Lauren Quigley, Matt Bellio, Phaedra Boinodiris, Pin-Yu Chen, Ros Docktor, Ryan Hagemann, Tyler Lisowski

Table of contents

04
Introduction

05
Benefits and Challenges

08
Risks and Mitigations

12
Sample Use Cases

13
Regulatory Perspectives

14
IBM Research

15
IBM Product

17
Best Practices

18
Conclusion

Introduction

As data-driven technologies advance, synthetic data-artificially generated data that mimics real-world data-has emerged as a promising solution to overcome challenges in data collection, AI model training, and privacy, particularly in AI systems that interact with and adapt to complex environments. This is especially relevant for agentic AI systems that learn from synthetic data to make decisions, take actions, and interact with humans, as the generated data's fidelity, stability, and contextual relevance will influence the AI system's ability to navigate uncertain and dynamic situations. This discussion will focus on the use of synthetic data for AI model development and will not address questions surrounding AI generated content, such as deepfakes or artificially generated media.

Synthetic data can be created using various methods, including statistical modeling, agent-based modeling, and deep learning-based approaches. Among these, Large Language Models (LLMs) represent one of the most revolutionary deep learning-based approaches to generating synthetic data, and they are increasingly being adopted to advance the development of new LLMs. However, using synthetic data to train AI models for money laundering detection, for example, raises potential concerns about inadvertently revealing sensitive customer or transaction information if the data isn't properly anonymized, underscoring the need for careful evaluation and validation of its use. This tension underscores the need to explore synthetic data's benefits, challenges, risk mitigations and opportunities, as we navigate its expanding applications across industries and promote its responsible generation and use.

Benefits and Challenges

Synthetic data provides significant value across product development and software testing, a wide range of Natural Language Processing (NLP) tasks, diverse analytics and machine learning workflows, and throughout the lifecycle of building, training, validating, tuning, and deploying generative AI and foundation models. Specifically for model development, synthetic data has the potential to contribute throughout the stages of training, alignment, and evaluation.

- **Training**: During the pre-training phase of general-purpose models and the training of task-specific models, synthetic data can help address issues related to data bias, imbalance, privacy and scarcity. Synthetic data can be designed to create balanced datasets, introduce rare scenarios, and provide a diverse range of examples, thereby enhancing the model's learning and generalization, as well as its specialized capabilities. Synthetic data can also be algorithmically annotated, enabling consistent and accurate labels while saving significant resources on data annotation.

- **Alignment**: Synthetic data offers a promising approach for training instruction following models, particularly in scenarios where real-world data is scarce, expensive, or challenging to obtain.

- **Evaluation**: Synthetic data is widely used to evaluate models from different perspectives. For example, to assess the safety of LLMs, it plays a key role in red teaming by simulating challenging, rare, or adversarial scenarios. More broadly, synthetic data enables the creation of comprehensive test datasets, including cases that are rare or impossible to capture in real-world data, supporting model validation and helping models perform reliably across diverse conditions.

The following table provides an overview of the benefits associated with synthetic data, regardless of generation method or intended use, highlighting its value across operational, technical, and governance dimensions.

Category	Benefit	Description
Operational efficiencies	Cost savings	Generating synthetic data can be more cost-effective than collecting and labeling real data, reducing the possible financial burden of data collection and preparation. This enables organizations to allocate resources more efficiently and focus on high-value tasks.
	Annotated data	Synthetic data has the potential of providing an unlimited supply of annotated data, bypassing the conventional, time-consuming and costly process of data labeling. This enables organizations to train AI models on large amounts of high-quality data without the need for manual labeling.
	Customizable datasets	Synthetic data offers substantial customization options, enabling businesses to create datasets that match their distinct specifications and requirements. This allows organizations to tailor their AI models to specific use cases and applications.
	Faster time-to-market	Data scarcity across domains can hinder the development of robust models. By enabling rapid generation of large, labeled datasets, synthetic data can accelerate the training and testing of AI systems. This allows organizations to deploy AI models more quickly and respond to changing market conditions.
Model enhancement	Increased scalability	Synthetic data can be generated in large quantities, enabling organizations to scale their AI models and applications more quickly. This enables organizations to scale their AI models to accommodate large user bases, high-volume transactions, and rapid growth.
	Improved data quality	Synthetic data can be designed to be more consistent than real-world data, reducing the impact of data quality issues on AI model performance. This enables organizations to train AI models on high-quality data and potentially make more accurate predictions.
	Data gap mitigation	Synthetic data can infuse a wide array of scenarios and edge cases into datasets, enhancing model robustness and addressing data gaps. This enables AI models to handle unexpected scenarios and make more accurate predictions.
Fairness	Increase data balancing and bias mitigation	Synthetic datasets can be generated to balance sensitive variables (e.g., gender, race), address imbalances or underrepresentation in real datasets, and reduce bias propagation. This helps organizations build systems that reflect diverse environments and populations.
Transparency	Transparency in AI development	Synthetic datasets can enable greater data documentation and, depending on the generation method, can facilitate replication, improving transparency, traceability and accountability in how AI systems are developed and evaluated. This can contribute to improved data or model auditing, reproducibility, and trust in technology.
Misplaced trust	Privacy Protection	Depending on the method used for generating synthetic data, it is possible to eliminate or significantly reduce the need to use or preserve personal information, sensitive personal information or confidential information as part of seed data selection or synthetic data generation, which can contribute to protecting sensitive information and help comply with privacy and data regulations.

The generation and effective use of synthetic data involves a series of decisions, independent of the specific generation method and use case, and these decisions have differing effects depending on the domain. The following are some of the most prominent considerations which can impact the quality, accuracy, and effectiveness of synthetic data.

Challenge	Description
Model selection and training	The quality of synthetic is closely tied to the method used to generate it, with each technique offering strengths and limitations. When using LLMs, there are trade-offs between general models (broad applicability, but greater potential for inherited bias) and task-specific models (better alignment with specific use cases, but potentially limited scope). Similarly, using synthetic data to train a general model versus a task-specific model requires considering the trade-offs between adaptability and precision. Selecting the right approach requires balancing these factors to ensure the generated data meets the needs of the intended application.
Resource intensity and technical complexity	The quality of synthetic data depends on its generation method. All three methods for generating synthetic data-statistical, agent-based, and deep learning-based-share a common challenge: high technical complexity. For agent-based methods, this means designing realistic rules that mimic real-world behavior. For deep learning-based methods, it means tuning hyperparameters to optimize results. For statistical methods, it means modeling complex dependencies between variables. All of these approaches require significant computational power or specialized expertise, which can limit their accessibility and scalability.
Balancing data diversity and realism	Striking the right balance between data diversity and realism is vital for creating effective synthetic datasets. While diversity is essential for training robust AI models, it should be balanced with realism, which is crucial for maintaining the dataset's utility and credibility. Insufficient domain expertise as part of the synthetic data generation process can undermine both, possibly contributing to unrealistic and unrepresentative data.
Balance between accuracy and privacy	The balance between accuracy and privacy is a persistent challenge in synthetic data generation, as increasing the level of detail in synthetic data can compromise sensitive information. Conversely, reducing the level of detail to protect sensitive information can compromise the accuracy of the synthetic data, leading to poor model performance.
Validating and evaluating synthetic data	There's a significant challenge across existing data generation methods due to the lack of universally accepted standards and benchmarks. A key difficulty lies in balancing statistical fidelity, task-specific utility, and privacy preservation. Standard statistical tests (e.g., distribution similarity) may indicate that data "looks right," but fail to capture whether it performs well in downstream tasks or upholds fairness and privacy. Moreover, synthetic data can exhibit hidden biases, missing dependencies, or unrealistic cases, which are difficult to detect without domain-specific knowledge.
Maintaining appropriate documentation	Regardless of the chosen method for synthetic data generation, there is a lack of standardized frameworks for recording the design choices, parameters, assumptions, and verification steps. Lack of documentation complicates assessing data provenance, understanding potential biases and limitations in the generated data.
Interoperability and integration	Integrating synthetic data with existing datasets and systems can be challenging due to potential differences in data format, labeling, and structure. Addressing these inconsistencies often requires additional effort to ensure the synthetic data can be effectively utilized within current workflows and tools.

Risks and Mitigations

Synthetic data has emerged as a valuable tool in various industries, with its potential to unlock new opportunities and enhance AI development. However, depending on the context and domain in which it is used, the generation method, and how it interacts with other technologies, it can introduce a range of potential risks that must be carefully considered.

It is worth noting that the risks outlined below relate to different stages and aspects of synthetic data, including generation (Data Generation), use (Data Use), and specific methods (*Amplified by LLM-generated synth data*), as well as new risks specific to synthetic data not traditionally associated with AI or Machine Learning systems (*New*), and traditional AI or Machine Learning risks that are amplified by the use of synthetic data (*Amplified*).

Technical Risks

Group	Risk	Risk Indicator with Reason	Possible Risk Mitigations
Accuracy	**Temporal gap:** Temporal gaps in synthetic data refer to the discrepancies between the constantly evolving real-world data and the fixed conditions captured by synthetic data. Temporal gaps in synthetic data can render it outdated or obsolete as real-world conditions evolve, leading to mismatches between current realities and the assumptions embedded in the data, potentially compromising model performance, accuracy and relevance. Data Generation Data Use	*Amplified* Reason: • Poor model performance, reduced model accuracy, and decreased predictive power	**Periodically regenerate data:** Periodically regenerate synthetic data to maintain model performance, accuracy and relevance within the context of use. Incorporating retrieval-augmented generation into the synthetic data generation process can provide access to up-to-date and domain-specific data that can increase accuracy and further improve quality.
	Hallucination: In some instances, synthetic data generated by Large Language Models may be inaccurate, fabricated or disconnected from reality, a phenomenon often referred to as "hallucination." Hallucination can introduce distributional shifts or unrealistic patterns into synthetic data that can potentially compromise model performance, accuracy and relevance. Data Generation Data Use *Amplified by LLM-generated synth data*	*Amplified* Reason: • Fabricated or unrealistic data, lack of connection to real-world patterns, and decreased predictive power	**Mitigate hallucinations:** AI memory augmentation systems like Larimar, designed to give AI models a form of editable, short-term memory without retraining the entire system, can help mitigate hallucinations by enabling the model to potentially recall and correct its previous statements or assumptions.
Data Quality	**Poor data quality:** Poor data quality can compromise the effectiveness and reliability of AI models trained on synthetic data, particularly if the synthetic data is not FAIR (findable, accessible, interoperable, and reusable), or if the method used to generate synthetic data substantially affects its quality. Data Use	*Amplified* Reason: • Poor model performance, reduced model accuracy, and potential model failure	**Validate data quality:** Mitigate biases, HAP, and incomplete or inaccurate synthetic data in support of intended uses (for example, HAP may be necessary for tasks like safety alignment).

Group	Risk	Risk Indicator with Reason	Possible Risk Mitigations
Model Degredation	**Overfitting:** Excessive use of synthetic data generated with LLMs can lead to AI models that perform well on synthetic data but poorly on real-world data. Data Use *Amplified by LLM-generated synth data*	*Amplified* Reason: • Poor model generalizability, reduced model accuracy, and decreased predictive power	**Indicate proper attribution:** When sharing or reusing synthetic data, attribution can help maintain a clear record of who generated it, how, and for what purpose, which is critical for auditing, compliance, and responsible use. Additionally, watermarking techniques like TabWak and DuWak embed hidden signatures into synthetically generated data. Proper attribution can help trace and audit synthetic datasets, leading to better decision-making around the use of synthetic data. For example, by identifying which datasets are synthetic through attribution, we can make informed decisions about their use, avoiding excessive reliance on them and thereby mitigating the potential risks associated with overfitting and model collapse.
	Model collapse: Repeatedly training generative models on synthetic data generated with LLMs can lead to model collapse, where the model loses information and becomes less accurate. Data Use *Amplified by LLM-generated synth data*	*Amplified* Reason: • Degradation of model performance, reduced model reliability, and potential model failure	
Model Evaluation	**Evaluation contamination:** This risk emerges when synthetic data is used in model training and when publicly available benchmark test cases are also included as part of the pre-training data. When this overlap of data happens, there is evaluation contamination, which can result in inaccurate benchmark scores. Data Use	*Amplified* Reason: • Reduced model reliability, inflated model performance, and distorted evaluation outcomes	
	Inadequate validation: The risk that synthetic data is not thoroughly tested and validated, resulting in AI models that are not adequately evaluated for their performance and reliability in real-world scenarios. Data Use	*Amplified* Reason: • Degradation of model performance, reduced model reliability, and potential model failure	**Validate within context of use:** Evaluate accuracy and representativeness of synthetic data within the context of the domain for which the data or model will be used. Internal guidance or education on synthetic data can help users understand best practices, possible limitations and mitigation strategies, enabling informed decision-making and responsible use.
Robustness	**Adversarial vulnerability:** There's a risk that synthetic data generated by LLMs may lack robustness against adversarial manipulation, allowing attackers to intentionally or unintentionally introduce vulnerabilities such as data pollution, contamination or poisoning, and making it easier to inject malicious samples that evade detection. Data Generation Data Use *Amplified by LLM-generated synth data*	*Amplified* Reason: • Degradation of model performance and reliability, and potential compromise of AI system security	**Maintain robustness:** Test synthetic data generation methods for robustness and store data in a secure location that implements access controls, encryption, regular security audits, and address identified vulnerabilities. Watermarking techniques like TabWak and DuWak can also help trace and audit synthetic datasets, embedding hidden signatures into synthetically generated data, which can enable the monitoring of synthetic data modifications.

Sociotechnical Risks

Group	Risk	Risk Indicator with Reason	Possible Risk Mitigations
Fairness	**Data bias:** Biases present in the synthetic data, including those inherited from seed data or exacerbated by the generation methods, can influence the training and fine-tuning of the model. Biases in unrepresentative synthetic data can be carried forward into AI models trained on that data, and potentially into additional synthetic datasets generated from such models. Data Generation Data Use	*Amplified* Reason: • Unfair model outcomes, erosion of trust, and potential harm to marginalized groups	**Mitigate biases:** Apply bias correction techniques to minimize bias in synthetic data and in AI models trained on synthetic data. Test the data and models for fairness using tools like AI Fairness 360. Where possible, consult with domain experts and marginalized populations to validate that synthetic data represents their experiences and perspectives.
	Exclusion: The risk that synthetic data generation processes may overlook or fail to consult with marginalized populations, resulting in synthetic data that does not accurately represent their experiences, needs, or perspectives. Data Generation	*New* Reason: • Unfair model outcomes, erosion of trust, and potential harm to marginalized groups	
Data Representation	**Lack of realism:** Synthetic data might not fully capture the complexity and nuances of real-world data, including rare or extreme cases, due to possible limitations in the seed data quality, biases in data generation, or inadequate domain knowledge. Although synthetic data can be designed to be more consistent than real-world data in many aspects, if it is not representative of the real-world scenarios it is intended to model, AI model trained on such data may not work as intended. Data Generation Data Use	*Amplified* Reason: • Poor model performance, reduced model accuracy, and decreased predictive power	**Anticipate end use and edge cases:** Create synthetic data tailored to the expected end use and aim to capture edge cases so that synthetic data is more likely to reflect real-world conditions authentically. Incorporating retrieval-augmented generation into the synthetic data generation process can also provide access to up-to-date and domain-specific data that can increase accuracy and further improve data quality.
	Lack of domain expertise: Synthetic data generation processes may not involve sufficient consultation with domain experts, resulting in a lack of understanding of the specific requirements and nuances of the domain. This can lead to synthetic data that may not accurately capture the complexities and challenges of a real-world scenario. Data Generation	*Amplified* Reason: • Poor model performance, reduced model accuracy, and decreased predictive power	

Group	Risk	Risk Indicator with Reason	Possible Risk Mitigations
	Homogenization: The risk that the use of synthetic data may lead to an overrepresentation of dominant cultures and perspectives such as languages, potentially suppressing the representation of marginalized and underrepresented groups, and reducing diversity of thought and culture. Data Use	*Amplified* Reason: • Homogenization of culture, suppression of underrepresented voices, and reduced cultural sensitivity	**Mitigate biases:** Apply bias correction techniques to minimize bias in synthetic data. Test the synthetic data for fairness using tools like AI Fairness 360. Where possible, consult with domain experts and marginalized populations to validate that synthetic data represents their experiences and perspectives.
Privacy	**Data privacy:** Synthetic data can in some circumstances be reverse-engineered to reveal information about the underlying real seed data or the process used to generate it, potentially allowing for unauthorized re-identification of individuals or sensitive information, potentially breaching anonymity or confidentiality. Data Use	*Amplified* Reason: • Unauthorized disclosure of sensitive information, reputational damage, and potential liability for organizations	**Protect sensitive information:** Implement robust data anonymization techniques to protect sensitive information in seed data. Avoid generating synthetic data containing any real personal information or sensitive personal information and instead consider using statistical representations (at high levels of generality) to simulate real-world patterns without providing information that could identify individuals.
Misuse	**Misuse:** Synthetic data may be used to create deepfakes, false information, or other deceptive AI technologies, potentially leading to misinformation, manipulation, and harm to individuals or organizations. Data Use	*Amplified* Reason: • Erosion of trust, and potential harm to individuals or organizations	**Indicate proper attribution:** When sharing or reusing synthetic data, attribution can help maintain a clear record of who generated it, how, and for what purpose, which is critical for auditing, compliance, and responsible use. Additionally, watermarking techniques like TabWak and DuWak embed hidden signatures into synthetically generated data. This can help trace and audit synthetic datasets.

Sample Use Cases

The following sample use cases illustrate the practical applications and real-world considerations surrounding the use of synthetic data across different domains.

Sample Use Case 1: Insurance Claims Fraud

A leading insurance company is leveraging synthetic data to revolutionize its approach to detecting and preventing insurance fraud, and to enhance its underwriting and pricing capabilities. By simulating a comprehensive homeowners insurance ecosystem, including policy owners and insured homes, the company can create diverse and realistic scenarios that allow AI models to learn from a wide range of possibilities, improving their accuracy in detecting fraudulent claims and enabling more informed decisions about policy pricing and underwriting. This approach also enables the company to better assess risk, identify new business opportunities, and improve customer experience. The company must balance the benefits of using synthetic data against the potential risks, such as fairness and equity, as AI models may perpetuate any biases and inaccuracies present in the synthetic data. To mitigate these risks, the company should prioritize careful consideration of these trade-offs and consider using synthetic data for high-risk or low-data scenarios, selecting the right type of synthetic data, and continuously monitoring and evaluating the performance of AI models trained on synthetic data.

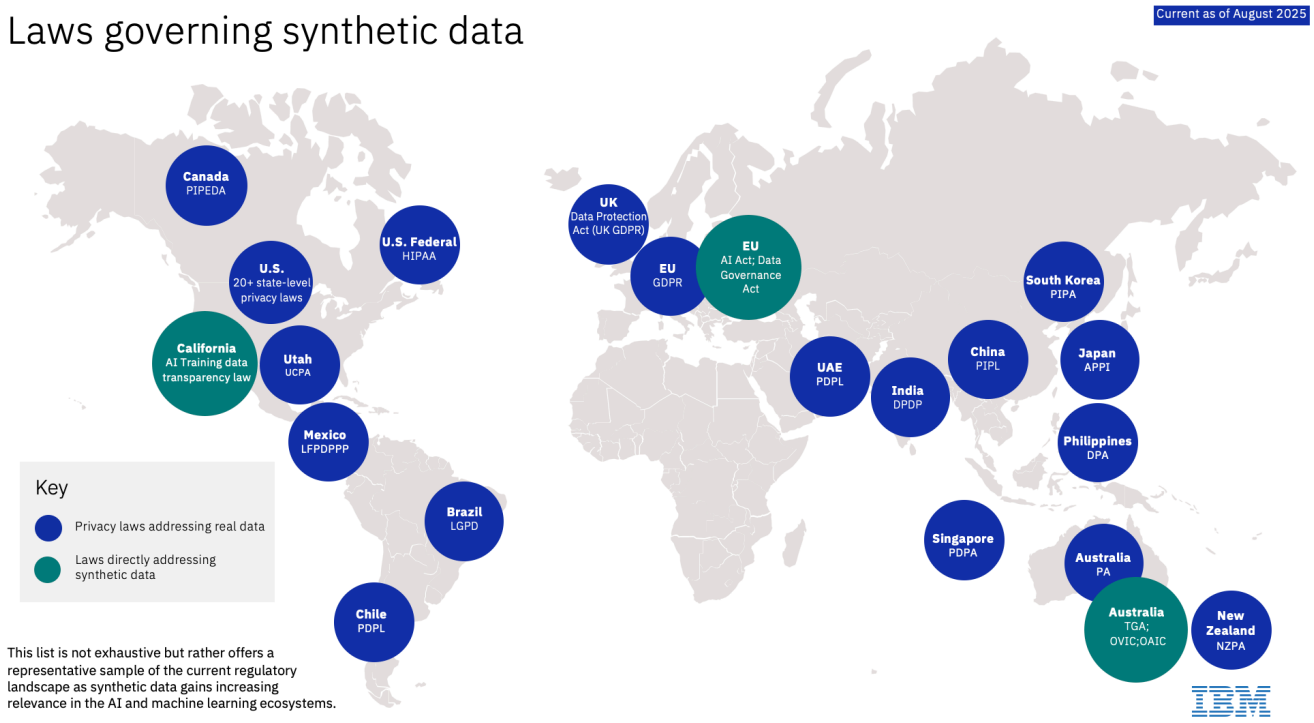
Sample Use Case 2: AI Safety

A leading research organization is using synthetic data to augment its AI safety training and alignment data, unlocking new opportunities for targeted and efficient risk mitigation. By generating synthetic data that simulates high-risk scenarios, such as malicious use or human-chatbot interactions, the organization can fine-tune its AI models to prioritize safety and efficacy. For example, synthetic data can be used to train AI models to detect safety threats in human-chatbot interactions, such as hate speech or harassment, and to simulate the complexities of natural language processing. This approach enables the organization to accelerate the development of more robust AI systems that can handle complex and dynamic situations. While this approach offers tremendous value, the organization must also test the impact of creating artificial representations of harm, which may inadvertently perpetuate or overly sanitize sensitive issues. The organization should carefully consider the type of synthetic data to use, selecting data that is tailored to the specific safety requirements of the AI model. For example, using LLM-generated synthetic data may be suitable for training AI models to detect safety threats in text-based interactions, while agent-based synthetic data may be more effective for simulating complex scenarios involving multiple agents. This approach enables the organization to accelerate the development of AI models that can handle complex and dynamic situations, and to detect and respond to rare but critical safety threats.

Sample Use Case 3: Cybersecurity

A leading bank is leveraging synthetic data to enhance its cybersecurity defense through the use of honeypot operations, which involve creating decoy systems or data that appear valuable but are actually fake. By generating synthetic data that mimics real-world transactions, the bank can create a realistic and attractive target for hackers, allowing the bank to gather insights on the hackers' techniques and tactics. This approach enables the bank to accelerate the development of AI-powered cybersecurity systems that can detect and respond to emerging threats, and to improve its overall cybersecurity posture. For example, the bank can use synthetic data to simulate various types of transactions, including those that may be indicative of suspicious activity. By using different types of synthetic data, such as statistical modeling for representative transactions and agent-based modeling for simulating hacker behavior, the bank can create a robust and realistic testing environment. Additionally, this approach can have a demotivating effect on hackers, as they become frustrated by their inability to identify the individuals behind the transactions, ultimately abandoning their hacking attempts due to the lack of tangible targets. However, if the synthetic data is not carefully designed and implemented, there is a risk that hackers may quickly identify and dismiss it, which could render the honeypot ineffective and potentially even provide hackers with insights into the bank's cybersecurity measures. By balancing the authenticity of the synthetic data with the need to protect its real data and systems, the bank can maximize the effectiveness of its honeypot operations and stay one step ahead of emerging cyber threats.

Regulatory Perspectives



Existing and emerging laws in several countries touch on synthetic data, particularly in the context of privacy and AI governance. This section does not represent all existing and emerging laws that may relate to synthetic data, but offers key examples in certain jurisdictions.

United States: Federal and State Initiatives

In the United States, there is currently no comprehensive federal law specifically regulating synthetic data. Federal legislative proposals that have not been passed into law, such as the Promoting Digital Privacy Technologies Act (H.R. 847, 117th Cong. (2022)) have referenced privacy-enhancing technologies, including synthetic data.

At the state level, California has passed AB 2013: AI Training Data Transparency (Cal. Civ. Code §§ 3110–3111), which requires developers of generative AI systems or services, on or before January 1, 2026, to provide certain disclosures regarding the data used to train generative AI, including whether synthetic data was used in development. Utah amended its Consumer Privacy Act in 2024 to clarify that "personal data" does not include "deidentified data" under the law; "deidentified data" includes synthetic data that is "generated by computer algorithms or statistical models and does not contain personal data"(Utah Code Ann. § 13-61-101). Other states, such as New York (2025 NY. S.B. 6955) and Washington (2025 WA H.B. 1168), have introduced bills addressing AI data transparency, including references to synthetic data, though these bills have yet to be passed into law as of August 2025.

European Union: GDPR and AI Act Developments

In the EU, the General Data Protection Regulation (GDPR) generally applies to the processing of "personal data" or "information relating to an identified or identifiable natural person." Synthetic data that does not relate to identifiable individuals or otherwise meets the GDPR's threshold for anonymization would not fall under the scope of the law. (Regulation (EU) 2016/679, L 119/1 O.J. (2016) ("GDPR")). Where synthetic data retains linkages to identifiable individuals, it may still fall within the scope of the regulation. The EU's Artificial Intelligence Act (Regulation (EU) 2024/1689, 2024 O.J. (L 2024/1689) introduces obligations for providers and deployers of AI systems deemed high-risk, including transparency and documentation requirements. While these provisions are not specific to synthetic data, they may be relevant when synthetic data is used in training or testing such systems.

China and India: Emerging Frameworks

China's Interim Measures for the Management of Generative AI Services (2023) include requirements related to the accuracy, reliability, and traceability of training data, and includes a mandate, starting September 1, 2025 to label certain AI-generated content. While synthetic data is not mentioned explicitly, the measures impose broad expectations on the integrity and provenance of data used in AI development. India's Digital Personal Data Protection Act (2023) provides a general framework for data processing, emphasizing principles such as consent and purpose limitation, though it does not specifically address synthetic data.

IBM Research

IBM Research has been actively advancing the field of synthetic data through the development of innovative techniques for generating high-quality synthetic data, with a focus on various contexts and applications. In this section, we provide an overview of the techniques, tools and applications that researchers have developed.

Techniques

IBM has developed specialized synthetic data generation methods tailored for distinct sectors and applications:

- **Financial Transactions:** IBM researchers created an advanced virtual simulation technique to generate realistic synthetic credit card transactions, including genuine and fraudulent activities. This method captures consumer and merchant behaviors, demographics, and spending patterns without relying on actual transactional data.

- **Anti-Money Laundering:** The AMLworld system simulates financial ecosystems, incorporating criminal agents conducting illicit transactions. This synthetic environment offers crucial insights unavailable from real-world data, including detailed transaction laundering patterns.

- **Text Classification:** The LAMBADA approach leverages synthetic data from fine-tuned language models, significantly improving text classification in scenarios with limited labeled data.

- **Out-of-Distribution (OOD) Detection:** IBM Researchers employed LLMs to create high-quality synthetic OOD examples, effectively substituting scarce real-world data. This approach has consistently demonstrated superior performance in classification tasks by reducing false positives and maintaining high accuracy.

- **LLM Safety Evaluation:** IBM researchers utilize synthetic adversarial datasets such as AttaQ and ProvoQ, generated by LLMs to identify vulnerabilities and safety issues within large language models, contributing significantly to their safer deployment.

- **Improving Synthetic Data Utility:** A novel post-processing technique has been developed to enhance synthetic data utility by targeting specific user-defined statistical properties or correlations, while preserving strong differential privacy guarantees.

- **Customizable Synthetic Datasets:** IBM introduced the SPRI framework, automating the generation of context-specific synthetic datasets for fine-tuning base models. SPRI significantly reduces reliance on costly human annotations, enhancing scalability and model alignment.

- **Synthetic Data Watermarking Techniques:** IBM has developed two watermarking approaches to protect synthetic data. TabWak embeds a secret signature into synthetic tables generated by tabular diffusion models, allowing for the detection of any unauthorized changes or sharing. Duwak is a dual watermarking approach for text generated by LLMs, embedding two independent secret patterns into the token probability distribution and sampling scheme, and outperforms existing watermarking methods in terms of the human-like quality of the watermarked task.

Tools

In addition to the multiple techniques that researchers have developed to generate synthetic data, IBM researchers have also developed tools to facilitate the generation and auditing of synthetic data.

- **AMLworld:** A multi-agent synthetic financial transaction generator specifically crafted to identify patterns of money laundering not easily discerned in real-world data.

- **Comprehensive Auditing Framework:** A holistic auditing tool evaluates synthetic data and generative AI models across essential trust dimensions: fidelity, privacy, utility, fairness, and robustness. This framework includes a Trustworthiness Index, offering stakeholders clear insights and control over trust trade-offs during model training and deployment phases.

IBM Product

IBM is at the forefront of synthetic data innovation, and its cutting-edge solutions are designed to help organizations accelerate AI adoption while addressing the challenges associated with real-world data usage.

Case Study: IBM Synthetic Data Sets

The Challenge:	Real data has long been the foundation of artificial intelligence (AI) model training. However, accessing and utilizing real data comes with its own set of challenges. Proprietary or sensitive data can be difficult to access, and lengthy approval processes can delay the development of AI solutions. Moreover, real data can lack diversity, perpetuate biases, and can be costly to collect, clean, and annotate.
The Solution:	IBM Synthetic Data Sets provides a practical alternative to traditional data usage in AI model training. By leveraging statistical distribution modeling and Agent-Based Modeling (ABM), IBM Synthetic Data Sets offer high-quality, synthetic data **generated without the need to use proprietary or sensitive seed data**.
Benefits:	• **Privacy and data protection:** IBM Synthetic Data Sets generates synthetic data that does not contain any personal or sensitive information. This approach helps organizations comply with data privacy regulations by eliminating the need to use proprietary or sensitive seed data that may put sensitive information at risk. By using synthetic data for training purposes, companies can significantly reduce the risk of exposing confidential information, enabling a safer and more secure environment for data analysis and innovation. • **Data quality and realism:** IBM Synthetic Data Sets maintains temporal consistency, ensuring that transaction dates, times, and frequencies align with realistic patterns. This helps to create a more realistic and representative dataset that can be used to train AI models without the presence of proprietary or sensitive seed data. This approach ensures that formatting details, geographical accuracy, and demographic accuracy are maintained. For example, synthetic credit card numbers, expiry dates, and zip codes are generated in the correct format and length, cities are linked to their corresponding countries or states, and demographic information is accurately represented. • **Scalability and cost-effectiveness:** IBM Synthetic Data Sets is poised to accelerate AI innovation by providing a robust, scalable, and cost-effective solution for AI model training. By eliminating the need for real data which can be challenging or costly to access, synthetic data sets can help organizations accelerate their AI project development and deployment, improve model accuracy, and mitigate risks associated with real data. • **Ground truth and model accuracy:** Often the ground truth (the target of an AI system, like fraud) is not easily found or labeled in real data. Since synthetic data provides known targets or labels, it can facilitate AI model training that helps improve the accuracy of predictive capabilities and decision-making processes.
Conclusion:	IBM Synthetic Data Sets is poised to accelerate AI innovation by providing a robust, scalable, and cost-effective solution for AI model training, providing a practical alternative to traditional data usage in AI model training. By mitigating the challenges associated with seed data like accessibility and high costs, synthetic data can accelerate AI model training and enhance overall lifecycle efficiency. These pre-built datasets may require additional collaboration with the product team to adapt to unique or evolving business needs due to limitations in their flexibility for customization. By eliminating the need for proprietary or sensitive seed data, IBM Synthetic Data Sets unlocks faster model development, improved accuracy, and reduced costs, driving transformative AI innovation.

watsonx.ai's Tabular and Unstructured Synthetic Data Generator

watsonx.ai's tabular synthetic data generator creates high-quality, synthetic datasets that mimic real-world data that are most often used for traditional Machine Learning use cases. This generator utilizes advanced statistical distributions to modify each column, allowing for customizable settings to generate synthetic data. The process involves importing existing data, selecting columns to anonymize, and generating synthetic data based on the specified settings. The generator also provides an evaluation metric to assess the quality of the synthetic data, including fidelity score, data distinguishability, leakage prevention score, and proximity score. This enables the synthetic data to be accurate, diverse, and protected from potential leaks.

watsonx.ai's unstructured synthetic data generator utilizes the synthetic data generation API to create large unstructured text datasets by leveraging data builder pipelines and data validators optimized for generating data for tuning and evaluating foundation models. This generator uses a range of techniques to produce different types of synthetic data, such as datasets for training models to interact with tools or generating pairs of questions and answers. Users can experiment with different models and fine-tune the generator to produce high-quality data tailored to their intended use, and the resulting synthetic datasets can be used to, for example, train or align AI models, helping to accelerate development and deployment.

InstructLab's Synthetic Data Generator

InstructLab's synthetic data generator creates large amounts of training data from a small seed dataset using a taxonomy-driven data curation approach. This approach organizes skills and knowledge areas into a hierarchical structure, guiding the creation of instruction data, evaluation, and response generation to ensure accurate and relevant synthetic data. The generator employs a teacher model to generate new examples through a synthetic data generation process, using prompt templates to expand the dataset while maintaining the structure and intent of the original data. Notably, this approach allows subject matter experts to contribute knowledge and skills in a structured format, without requiring technical expertise, and enables organizations to maintain version control over their knowledge base, creating a valuable asset that can be built upon over time. This results in a high-quality dataset that is tailored to specific task requirements. InstructLab's synthetic data generator accelerates AI development by improving model accuracy and reducing the costs associated with data collection and labeling. It is particularly useful in low-resource languages, domains, or tasks where high-quality training data is scarce.

Best Practices

To maximize the benefits of synthetic data while minimizing its risks, follow these best practices for generating and using synthetic data:

1. **Consider the context of use and domain requirements:** Before generating synthetic data, understand the specific requirements and technologies involved in its intended use. This includes considering factors such as data distribution, statistical properties, and domain-specific constraints to enable the generation of synthetic data that is relevant, accurate, and effective for the intended task or application.

2. **Collaborate with domain experts and utilize domain-specific data generation methods:** Involve experts in the relevant domain to guide the generation of synthetic data and help the data to accurately reflect real-world patterns and edge cases. Select a generation method, such as statistical modeling, agent-based modeling, or LLM-based approaches, that is best suited to the purpose of the intended use case, evaluating the strengths and limitations of each method.

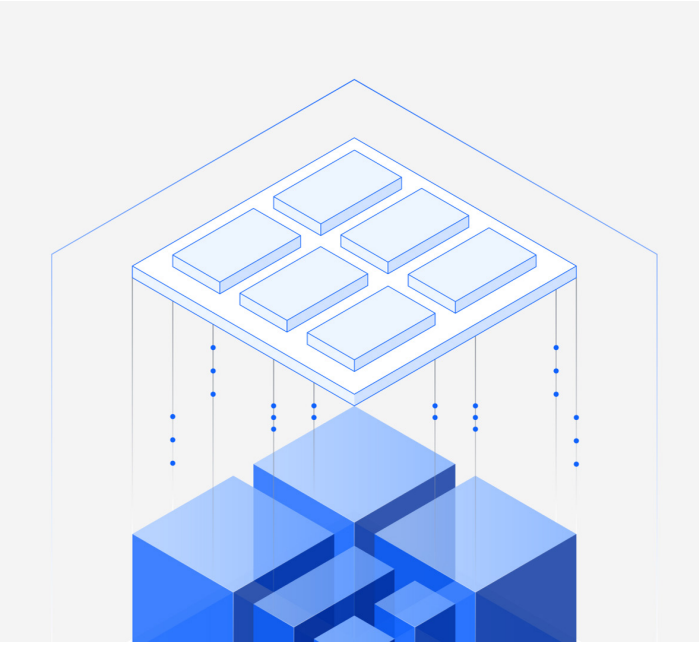
3. **Evaluate and validate synthetic data using multiple metrics:** Rigorously test and validate synthetic data to enable its accuracy, reliability, and realism. When evaluating the synthetic data, focus on its ability to capture the key characteristics and patterns of the real-world data, rather than just its size or volume. Avoid equating data scale with representativeness, as a larger dataset does not necessarily mean that it is more representative of the underlying population or phenomenon. Instead, consider factors such as data fidelity, distribution similarity, and task-specific utility when evaluating the synthetic data's ability to generalize to new, unseen data and accurately perform the intended task. Where appropriate for the use case, regularly test and evaluate model output trained on synthetic data against model performance on real-world data, such as cross-validation or benchmark datasets, to enable the synthetic data to effectively train the model for real-world application.

4. **Maintain documentation and version control:** Maintain documentation of synthetic data generation, including the methods used, assumptions made, and decisions taken. Consider using version control to track changes to the synthetic data and facilitate access to the most up-to-date version for stakeholders. Connect synthetic data generation and use to governance artifacts like lineage, model cards, and provenance tracking where possible. This can help trace and audit the dataset, enabling greater transparency, accountability, and reproducibility.

5. **Update and refine synthetic data to support data integrity and relevance:** Regularly review and update synthetic data to ensure it remains relevant and accurate over time. This involves monitoring changes in the real-world data distribution, updating the generating models, and adapting to new requirements and use cases. Consider using techniques such as data augmentation and adversarial training to improve the robustness and generalizability of the synthetic data.

Conclusion

With its vast potential to accelerate innovation and unlock new opportunities, synthetic data has the power to transform industries and redefine the future of data-driven technologies. However, its promise is tempered by the need for careful consideration of the complex implications of its development and use. As we stand at the threshold of this new frontier, it is imperative that we harness the potential of synthetic data in a responsible manner, one that prioritizes accuracy and transparency, and anticipates the complex interactions between AI systems and the humans they serve.



© Copyright IBM Corporation 2025

Produced in the
United States of America
August 2025

IBM, the IBM logo, and watsonx.ai, are trademarks or registered trademarks of International Business Machines Corporation, in the United States and/or other countries. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on ibm.com/trademark.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

IBM products are warranted according to the terms and conditions of the agreements under which they are provided.