Auditing and Generating Synthetic Data with Controllable Trust Trade-offs

Brian Belgodere, Pierre Dognin, Adam Ivankay, Igor Melnyk, Youssef Mroueh, Aleksandra Mojsilovic, Jiri Navratil, Apoorva Nitsure, Inkit Padhi, Mattia Rigotti, Jerret Ross, Yair Schiff, Radhika Vedpathak, and Richard A. Young.

***Abstract*—Real-world data often exhibits bias, imbalance, and privacy risks. Synthetic datasets have emerged to address these issues by enabling a paradigm that relies on generative AI models to generate unbiased, privacy-preserving data while maintaining fidelity to the original data. However, assessing the trustworthiness of synthetic datasets and models is a critical challenge. We introduce a holistic auditing framework that comprehensively evaluates synthetic datasets and AI models. It focuses on preventing bias and discrimination, ensuring fidelity to the source data, and assessing utility, robustness, and privacy preservation. We demonstrate our framework's effectiveness by auditing various generative models across diverse use cases like education, healthcare, banking, and human resources, spanning different data modalities such as tabular, time-series, vision, and natural language. This holistic assessment is essential for compliance with regulatory safeguards. We introduce a trustworthiness index to rank synthetic datasets based on their safeguards trade-offs. Furthermore, we present a trustworthiness-driven model selection and cross-validation process during training, exemplified with "TrustFormers" across various data types. This approach allows for controllable trustworthiness trade-offs in synthetic data creation. Our auditing framework fosters collaboration among stakeholders, including data scientists, governance experts, internal reviewers, external certifiers, and regulators. This transparent reporting should become a standard practice to prevent bias, discrimination, and privacy violations, ensuring compliance with policies and providing accountability, safety, and performance guarantees.**

***Index Terms*—Trustworthy AI, Synthetic Data, Auditing, Generative AI.**

I. Introduction

Generative models have demonstrated impressive results in synthesizing high-quality data across multiple modalities from tabular and time-series data [70], [71], [74] to text [19], [30], images [18], [46], [76], [83] and chemistry [12]. We are entering a new era in training AI models, where synthetic data can be used to augment real data [25], or as a complete replacement, in the most extreme case [38]. One of the main motivations behind controllable synthetic data usage in training AI models is its promise to synthesize privacy-preserving data that enables safe sharing without putting the privacy of real users and individuals at risk. This has the potential to circumvent cumbersome processes that are at the heart of many highly regulated fields such as financial services [10] and healthcare, for example [16], [25], [28], [32]. Another motivation comes from controlling the generation process in order to balance the training data and reduce biases against protected groups and sensitive communities [29]. Finally, synthetic data also offers new opportunities in simulating non-existing scenarios, providing grounding for causal inference via the generation of counterfactuals that would help explain some observations in the absence of real data [36].

Synthetic data can take different forms, ranging from seedless approaches [7], [51], which rely on knowledge bases and rule-based generation but incur risks of biased grounding and linkage attacks, to data generated through AI models trained on real data, which may lead to memorization, privacy breaches, and bias amplification [23], [29], [42]. These AI-generated datasets can also pose legal issues related to copyright and intellectual property [87]. Both types of synthetic data need thorough auditing for safety, privacy, fairness, and utility alignment.

Amidst these technological advancements, the AI regulation landscape is rapidly evolving to institute safeguards and objectives for AI systems, aimed at mitigating societal risks and malicious use. For instance, the recent executive order on Safe, Secure, and Trustworthy Artificial Intelligence by the Biden administration highlights the urgency of these concerns. Other acts, like the U.S. Algorithmic Accountability Act and the EU AI Act (for a detailed comparison, see [66]), are paving the way towards fostering trustworthy AI. The EU AI Act, in particular, enforces conformity assessments and post-market monitoring of AI models. Furthermore, quantitative auditing of predictive AI models has made significant progress in recent years, with various auditing systems emerging in domains such as algorithmic recruitment [53] and healthcare [59]. Multiple AI risk assessment frameworks have been proposed to tackle certain aspects of trust but have focused on individual trust pillars, such as fairness [15], explainability [9], or robustness [69]. With the advent of foundation models [17] and Large Language Models (LLMs), several techniques [57], [80] are being explored in an attempt to mitigate risks; moreover, multiple frameworks have suggested probing these models on specific trust aspects via red teaming [72], reconstruction of training data attacks [23], or via holistic auditing, as proposed in HELM (Holistic Evaluation of Language Models) [58] and the auditing framework of [47]. Finally, several governance mechanisms have been proposed to ensure transparency in communicating the risks of data and models via fact sheets [8], model cards [65], data sheets [39], and system and method cards [2].

Existing frameworks, like Synthetic Data Vault (SDV) [71] and Synthcity [74], tend to focus on specific aspects of synthetic

Authors are listed in alphabetical order. Correspondence to mroueh@us.ibm.com and inkpad@ibm.com

This article has been accepted for publication in IEEE Journal on Emerging and Selected Topics in Circuits and Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JETCAS.2024.3477976

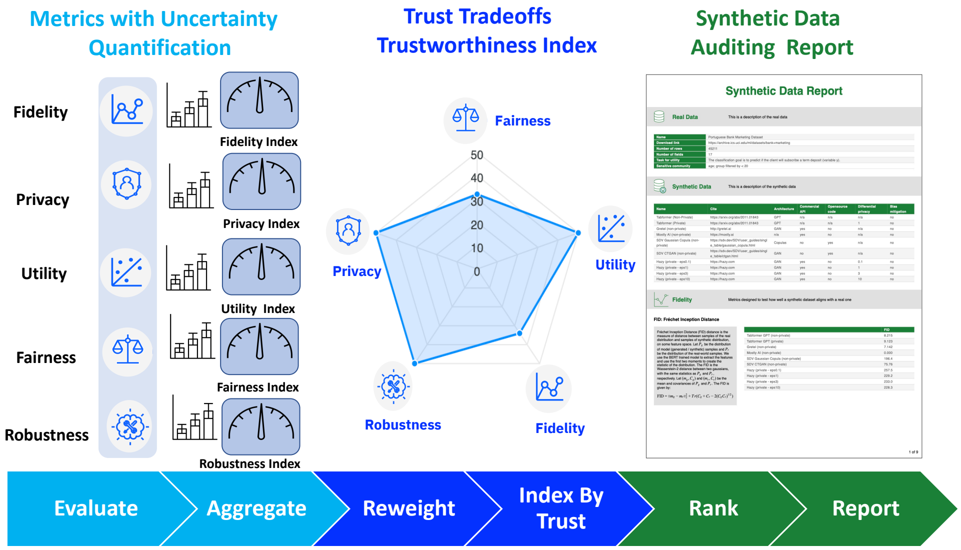


Fig. 1: Summary diagram of our proposed holistic synthetic data auditing framework. For each trust dimension (fidelity, privacy, utility, fairness, and robustness), we evaluate multiple metrics on the synthetic data and quantify their uncertainty. Metrics are aggregated within each trust dimension, which results in trust dimension indices. These indices are re-weighted with desired trust trade-offs to produce the trustworthiness index. Different synthetic datasets are then ranked using this trustworthiness index, and a summary of the audit is written to an audit report. The ranking produced by our audit enables comparison of different synthetic data produced by various generative modeling techniques, and aids the model selection process for a given generation technique, allowing its alignment with prescribed safeguards. The model selection is performed via trustworthiness index driven cross-validation, which results in controllable trust trade-offs by producing new ranks for different desired weighing trade-offs for a given application and use case.

data auditing, often overlooking a comprehensive evaluation of all trust dimensions. The TAPAS framework [48], on the other hand, is primarily dedicated to privacy [48], without fully addressing trade-offs with other essential dimensions. There are also efforts that concentrate on fidelity and utility auditing, such as [6], and those examining privacy-preserving capabilities [26], [50], [86]. Unfortunately, these initiatives in addition to not being holistic, they frequently fail to account for the uncertainty introduced by data splits between training and testing sets. Addressing these gaps is crucial for achieving a comprehensive audit of synthetic data in AI applications, as emphasized in a recent European Parliament report [67].

To address these challenges, we propose a framework for auditing the trustworthiness of synthetic data that is *holistic*, *transversal* across different modalities (tabular, time-series, computer vision and natural language) and assess the uncertainty in auditing a generative model (see Figure 1 for a summary of our approach).

Our main contributions are as follows: we introduce a holistic framework for auditing the trustworthiness of synthetic data, covering key trust dimensions like fidelity, utility, privacy, fairness, and robustness. We define a *trustworthiness index* that evaluates synthetic data and their downstream tasks. We provide methods for controlling trust trade-offs in synthetic data during training, notably through model selection via the trustworthiness index. We instrument transformer models across multiple modalities with these control mechanisms and refer to them as “TrustFormers”. By applying our framework, we demonstrate that downstream tasks using trustworthiness-index-driven cross-validation often outperform those trained on real data while meeting privacy and fairness requirements. Finally, our framework offers transparency templates for clear communication of the risks of synthetic data via an audit report.

Controllable trade-offs in the auditing and generation of synthetic data ensure that AI models trained on such data, when deployed in circuit systems and applications, meet end-user needs by providing the required confidence that fosters the adoption of AI solutions, as well as compliance with policy and legal requirements. Enabling AI solutions that are resilient to adversarial attacks and that mitigate privacy, bias, and fairness concerns while maintaining performance ensures a longer term stability of deployed solutions and their maintenance in an ever-changing technological environment. This adaptability is crucial for ensuring that AI-driven systems can be deployed across various domains, from consumer electronics to critical infrastructure, while adhering to both technical requirements and societal expectations.

II. Synthetic Data Auditing Framework

In this section, we present our auditing framework. For a given real dataset and several synthetic datasets, our framework evaluates a multitude of ***trust dimensions***, namely: fidelity, privacy, utility, fairness, and robustness.

a) Setup: Formally, given a real dataset D_r and multiple synthetic datasets $D_s^j,\ j = 1, \ldots, N$. The real dataset is split to training, development/validation, and testing sets as follows:

$$D_r = \{D_{r,\text{train}},\ D_{r,\text{val}},\ D_{r,\text{test}}\}. \tag{1}$$

This article has been accepted for publication in IEEE Journal on Emerging and Selected Topics in Circuits and Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JETCAS.2024.3477976

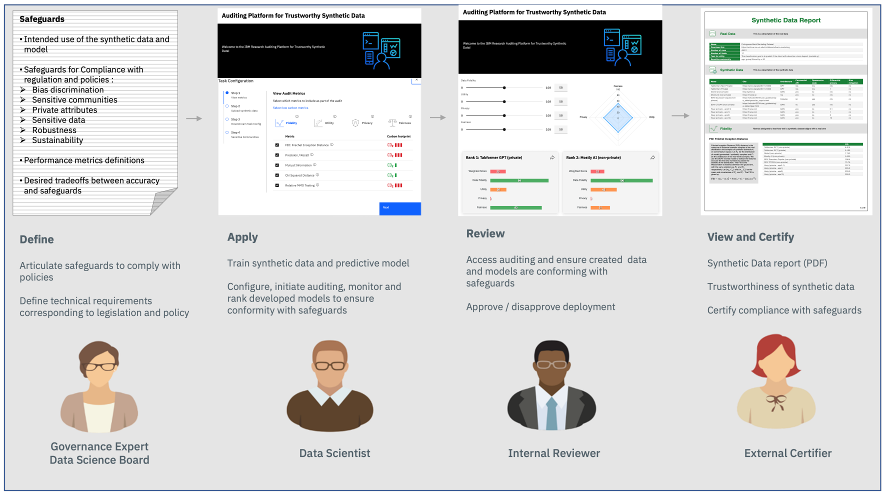


Fig. 2: Auditing Platform and workflows connecting different stakeholders (e.g., data scientists, data governance experts, internal reviewers, external certifiers, and regulators) from model development to audit and certification via a synthetic data auditing report.

Synthetic datasets come from various sampling schemes from different types of generative models. These models are trained on the real training set $D_{r,\text{train}}$ and validated on the real development set $D_{r,\text{val}}$. The utility of these synthetic data is measured via a predictive downstream task defined on the data space along protected and sensitive groups for whom we want to ensure a fair prediction. The downstream task is trained on the synthetic data $D^j_{s,train}$, validated on the real development set $D_{r,\text{val}}$, and evaluated on the real test set $D_{r,\text{test}}$ (note that $D_{r,\text{test}}$ could have a distribution shift w.r.t to $D_{r,\text{train}}$). Without loss of generality, we assume for simplicity that all downstream tasks are classification tasks.

b) Auditing Framework : Our synthetic data auditing framework, as depicted in Figure 1, comprises several key stages: Evaluate, Aggregate, Reweight, Index by trustworthiness, Rank, and Report. It serves as a means to enhance communication among governance experts, data scientists, internal reviewers, and external regulators. The primary personas involved in our framework are outlined in Figure 2. The key stages of our audit framework and the personas involved are explained below (more detailed explanations of trust dimension and our framework are given in the Methods Section and Supplementary Information):

1) **EVALUATE:** The governance expert and data science board collaborate to establish quantitative metrics for each trust dimension, as detailed in Table I, ensuring adherence to socio-technical safeguards. These metrics are then assessed by data scientists.
2) **AGGREGATE:** Metrics within each dimension are aggregated into a trust dimension index (denoted as π_T, where "T" corresponds to Fidelity, Privacy, Utility, Fairness, or Robustness), which ranges from 0 to 1, representing compliance probability with the requirements of the dimension. This aggregation method is explained in our Methods Section.
3) **RE-WEIGHT:** Stakeholders, including internal reviewers, governance experts, and the application owner, establish a trustworthiness profile through trade-off weights ω (examples are shown in Table II), indicating the relative importance of trust indices for fulfilling specific requirements.
4) **INDEX BY TRUSTWORTHINESS:** Trust dimension indices are re-weighted by the trustworthiness profile weights ω and combined via a geometric mean to produce the trustworthiness index, $\tau_{\text{trust}}(\omega)$, a context-specific measure based on the application's needs and trustworthiness profiles.
5) **RANK:** The trustworthiness index allows internal reviewers to rank synthetic datasets. Data scientists can use it for model selection within a given generation method and trustworthiness profile. When determining our trustworthiness index ranking, we can additionally consider the uncertainty in real data splits. For a trustworthiness profile ω, our preference is for a generative AI technique that exhibits the highest average trustworthiness index across splits $(\overline{\tau_{\text{Trust}}}(\omega))$ while minimizing volatility $(\Delta_\tau(\omega))$. For $\alpha \geq 0$, this corresponds to choosing the model with highest
$$R^{\alpha}_{\text{trust}} = \log(\overline{\tau_{\text{Trust}}}(\omega)) - \alpha \log(\Delta_\tau(\omega)). \quad (2)$$
6) **REPORT:** To enhance transparency, our framework provides audit report templates (as seen in Supplementary Information Section U for communicating the audit results. These audit reports can be submitted to regulators or external third-party certifiers for validation.

c) Controllable Trust Tradeoffs with TrustFormers: Trust constraints can be integrated in the training of Generative AI models. For example to ensure privacy of the synthetic data we use differential private training [35] of the generative models with a privacy budget ε. Furthermore, our trustworthiness index can be employed in an early stopping approach to select the model that best aligns with the desired trustworthiness profiles. Leveraging their adaptability in modeling various modalities, we integrate these trustworthy training and selection

Dimension	Metric	Polarity	Debiasing	Tabular	Time Series	NLP
Fidelity	***Evaluated between*** $D_{r,train}$ ***and*** D_s***:***					
	Maximum Mean Discrepancy (MMD) SNR [54]	-1	N/A	✓ (**D/E**)	✓(**E**)	✓(**E**)
	MMD test$p$-value [41]	$+1$	N/A	✓ (**D/E**)	✓(**E**)	✓(**E**)
	Fréchet Inception Distance (FID) [45], [82]	-1	N/A	✓ (**D/E**)	✓(**E**)	✓(**E**)
	Precision/ Recall [55]	$+1$	N/A	✓ (**D/E**)	✓(**E**)	✓(**E**)
	Chi Squared	-1	N/A	✓ (**D**)	✗	✗
	ℓ_2 mutual Information difference	-1	N/A	✓ (**D**)	✗	✗
Privacy	***Evaluated between*** $D_{r,train}$ ***and*** D_s***:***					
	Exact Replicas Count	-1	N/A	✓ (**D**)	✗	✗
	k-nearest neighbor median distance [26]	$+1$	N/A	✓ (**D/E**)	✓(**E**)	✓(**E**)
	k-nearest neighbor mean distance [26]	$+1$	N/A	✓ (**D/E**)	✓(**E**)	✓(**E**)
Utility	***Classifier trained on*** D_s					
	Evaluated on $D_{r,val}$ ***@ validation and*** $D_{r,test}$***@ test:***					
	Accuracy/ precision/ recall/ F1 score of:					
	Linear Logistic Regression	$+1$	✗	✓ (**D/E**)	✓ (**E**)	✓ (**E**)
	Nearest Neighbor classification	$+1$	✗	✓ (**D/E**)	✓ (**E**)	✓ (**E**)
	MLP	$+1$	✗	✓ (**D/E**)	✓ (**E**)	✓ (**E**)
	MLP / Adversarial debiasing [92]	$+1$	✓	✓ (**D/E**)	✓ (**E**)	✓ (**E**)
	MLP/ Fair Mixup [31]	$+1$	✓	✓ (**D/E**)	✓ (**E**)	✓ (**E**)
Fairness	Applicable to all classifiers in utility:					
	Evaluated on $D_{r,val}$ ***@ validation and*** $D_{r,test}$***@ test:***					
	Equal Opportunity Difference (absolute value) [14]	-1	*	*	*	*
	Average Odds Difference (absolute value) [14]	-1	*	*	*	*
	Equalized Odds Difference (absolute value) [14]	-1	*	*	*	*
Robustness	Applicable to all classifiers in utility:					
	Evaluated on $D_{r,val}$ ***@ validation and*** $D_{r,test}$***@ test:***					
	Adversarial Accuracy/ precision/ recall/ F1 score	$+1$	*	*	*	*
	Absolute Difference of Adversarial and non adversarial utility metrics	-1	*	*	*	*

TABLE I: Metrics and their associated polarities that are supported by our auditing framework under each trust dimension. Debiasing indicates if a utility classifier uses a debiasing technique. **D** indicates that the metric is computed on the data space after quantization. **E** indicates that the metric is computed in an embedding space. * refers to the same field values of the evaluated utility classifier. Note that our metrics are representative of each dimension and modality but are not exhaustive; other specialized metrics can be added and integrated seamlessly within our framework.

	$\omega = (\omega_f, \omega_P, \omega_U, \omega_F, \omega_R)$	Interpretation
ω_{all}	$(100, 100, 100, 100, 100)/500$	Equal Importance
$\omega_{e(PU)}$	$(50, 100, 100, 50, 50)/350$	Privacy/Utility Emphasis
$\omega_{e(PUF)}$	$(50, 100, 100, 100, 50)/400$	Privacy/Utility/Fairness Emphasis
ω_U	$(0, 0, 100, 0, 0)/100$	Utility only
ω_{PU}	$(0, 100, 100, 0, 0)/200$	Privacy/Utility only
ω_{UF}	$(0, 0, 100, 100, 0)/200$	Utility/Fairness only
$\omega_{e(UF)r(R)}$	$(50, 50, 100, 100, 0)/300$	Utility/Fairness Emphasis No Robustness
ω_{UFR}	$(0, 0, 100, 100, 100)/300$	Utility/Fairness/Robustness only
ω_{UR}	$(0, 0, 100, 0, 100)/200$	Utility/Robustness only
ω_{PUR}	$(0, 100, 100, 0, 100)/300$	Privacy/Utility/Robustness only

TABLE II: Examples of weights trade offs of trust dimensions reflecting priorities in auditing synthetic data.

paradigms into generative transformer models, naming the resulting models TrustFormers. We denote the selected models as:

$$\text{TF}(\omega, \text{n-p}) \text{ for non-private training and}$$

$$\text{TF}(\omega, \text{p-}\varepsilon) \text{ for private training.}$$

If two trade-off weights ω_1 and ω_2 lead to the same checkpoints selection we use the following notation:

$$\text{TF}(\omega_1, \omega_2, \text{n-p}) \text{ for non-private training and}$$

$$\text{TF}(\omega_1, \omega_2, \text{p-}\varepsilon) \text{ for private training.}$$

III. Methods

a) Trust Dimensions: We start by giving precise definitions for the trust dimensions and their risks assessment that play a central role in our auditing framework:

- ***Fidelity.*** Fidelity measures the quality of the synthetic data in terms of its closeness in distribution to the real data and its diversity in covering the multiple modes of the real data distribution [6], [41], [45], [54], [55].
- ***Privacy.*** Privacy assesses memorization and real data leakage to synthetic data. Membership inference attacks, such as nearest neighbor attacks, are instrumented to identify if an actual data point can be identified in the vicinity of a synthetic data point, thereby unveiling that

Use Case	Dataset	Modality	Downstream Task	Safeguards	Policy Alignment Example
Banking	Bank Marketing [68]	Tabular	Campaign prediction	sensitive community (age); user privacy; robustness	Fair Lending Act
Recruitment	UK Recruitment [37]	Tabular	Employment prediction	sensitive community (ethnicity) user privacy; robustness	NYC Law 144
Education	Law School Admission Council Dataset [88]	Tabular	Admission prediction	sensitive community (ethnicity); user privacy; robustness	Equal Educational Opportunity Act
Financial Services	Credit Card [7]	Tabular time-series	Fraud Detection	user privacy	Finance Regulation
Healthcare	MIMIC-III [52]	Tabular time-series	Mortality prediction	sensitive community (ethnicity); user privacy; robustness	Patient Protection and Affordable Care Act
Healthcare	MIMIC-III Notes [5]	NLP	Mortality prediction	sensitive community (ethnicity) user privacy; robustness	Patient Protection and Affordable Care Act
Visual Recognition	Imagenet [5]	Vision	classification	distribution shift	Robust Generalization

TABLE III: Synthetic data use cases, safeguards and policy alignment.

the corresponding individual was a member of the real data training set [48], [50], [86].

- ***Utility.*** Utility measures the accuracy and performance of a predictive downstream task, where predictive models are trained on the synthetic data and evaluated in terms of their predictive performance on real test data.
- ***Fairness***. Fairness has two aspects: the first is related to bias in the synthetic data [22], and the second is the fairness of the predictions with respect to sensitive and protected communities evaluated on real test data points [14].
- ***Robustness.*** Robustness refers to the accuracy of a predictive model trained on synthetic data and evaluated on real test points in the presence of imperceptible, worst-case adversarial perturbations. We use black box, greedy attacks on utility classifiers for tabular and time-series, as in [3], [13], [24], [56], [62], [90], [90] (see Supplementary Information Q).

b) Auditing Framework: The key steps of our auditing framework (Figure 1) are explained below:

1) **EVALUATE:** Given the specific synthetic data application and relevant policies and regulations, the governance expert and data science board collaboratively establish multiple quantitative metrics for each trust dimension. These metrics serve to evaluate the synthetic data's adherence to the requirements necessary for meeting socio-technical safeguards within each dimension. In Table I, a comprehensive set of metrics is presented for each dimension, chosen to strike a balance between interpretability and risk assessment. it is important to note that while these metrics represent each dimension and modality, they are not exhaustive. Our framework can seamlessly accommodate additional specialized metrics as needed. These metrics are then assessed by the data scientist.

2) **AGGREGATE:** The interpretation and communication of these metrics within each dimension pose a significant challenge. Dealing with numerous metrics with different polarities and dynamic ranges can be overwhelming for internal reviewers and regulators. In social sciences, it is common to aggregate metrics into a single score or index [40]. Indices serve as powerful tools for simplifying complex information into an accessible format that can be interpreted and understood by a wide range of stakeholders. They facilitate straightforward communication, enabling easy comparisons and benchmarking. We address the issues related to varying ranges and polarities and aggregate the metrics within each dimension into a ***trust dimension index*** that falls within the range of 0 to 1, where 0 signifies poor conformity with the dimension requirements, and 1 indicates high conformity. This trust dimension index can be interpreted as a measure of compliance probability. Our aggregation method relies on the copula technique [85] which provides us with this intuitive probabilistic interpretation. In Supplementary Information B we explain the copula method that consists in normalizing metrics under trust dimension using global CDFs estimated across synthetic data, and followed by a geometric mean.

3) **RE-WEIGHT:** Each synthetic dataset is now represented by trust dimension indices, denoted as π_T, where "T" corresponds to Fidelity, Privacy, Utility, Fairness, or Robustness. Considering the specific application, associated policies, and desired trade-offs between trust dimensions, internal reviewers collaborate with governance experts and the application owner to establish the ***trustworthiness profile***. This profile is expressed through ***tradeoff weights***, symbolized as ω_T, which indicate the relative importance of the trust indices necessary to fulfill performance and policy requirements. For instance, when training downstream tasks on-site, privacy may not be a necessity for the synthetic data, but it becomes essential for the predictive model. However, when conducting training in a public cloud environment, ensuring the privacy of synthetic data is mandatory. These varying requirements lead to different trustworthiness profiles with distinct trade-offs between privacy and utility for the synthetic data. Detailed examples of other possible trustworthiness profiles can be found in Table II.

4) **INDEX BY TRUSTWORTHINESS:** The trust dimension indices are subsequently re-weighted by the trade-off weights and combined to produce the final ***trustworthiness index*** ($\tau_{\text{trust}}(\omega)$) of the synthetic data. it is important to emphasize that this index is context-specific, contingent upon the application's requirements and the specified safeguards and trustworthiness profiles. Recalling that the trust dimension indices can be interpreted as probabilities, we define the trustworthiness index as a weighted geometric mean of the dimension indices:

$$\tau_{\text{Trust}}(\omega) = \exp\left(\sum_T \omega_T \log(\pi_T)\right), \qquad (3)$$

The choice of a geometric mean is preferred because it embodies an "and" operation interpretation, unlike the arithmetic mean, which implies an "or" interpretation.

5) **RANK:** With the defined trustworthiness profile, the trustworthiness index can serve multiple purposes. The internal reviewer can utilize it to establish a ***ranking*** for various synthetic datasets generated by different models, enabling certification and validation of their adherence to specific requirements. Simultaneously, data scientists can leverage the trustworthiness index for ***model selection*** within a given generation method and trustworthiness profile. When determining our trustworthiness index ranking, we can additionally consider the uncertainty in real data splits. For a trustworthiness profile ω, our preference is for a generative AI technique that exhibits the highest average trustworthiness index across splits ($\overline{\tau_{\text{Trust}}}(\omega)$) while minimizing volatility ($\Delta_\tau(\omega)$). For $\alpha \geq 0$, this corresponds to choosing the model with highest

$$R^\alpha_{\text{trust}} = \log(\overline{\tau_{\text{Trust}}}(\omega)) - \alpha \log(\Delta_\tau(\omega)). \qquad (2)$$

6) **REPORT:** To promote transparency and accountability, our framework defines templates for communicating auditing results in the form of an ***audit report***. An example of the audit report is given in Supplementary Information U. The audit report can be submitted to a regulator or an external third-party certifier that probe the validity of the conclusions of the internal audit report.

c) Auditing Workflows for Transparency and Accountability: Insights and Limitations: Our work aligns closely with the core principles of AI auditing, as underscored in both the FAccT (Fairness, Accountability, and Transparency) and STS (Science and Technology Studies) literature, which delve into issues related to race, gender, bias, and fairness [1]. For instance, the study by Buolamwini and Gebru [20] highlighted the need for auditing racial and gender biases in facial recognition. In our holistic auditing approach, we address multiple dimensions, mirroring the discussions on intersectional biases frequently explored in these literatures [43]. While our audits are centered on specific technical and societal aspects, it is important to note that both the FAccT and STS literature encompass a broader spectrum of topics, including governance, closing the accountability gap [75] , ethics, and the far-reaching societal implications of emerging technologies [73], [84]. We discuss here a real-time platform that operationalizes our synthetic data auditing framework and further embraces the audit culture advocated in the the FAccT and STS literature in terms of accountability, transparency and governance workflows. Our auditing platform connects different stakeholders from governance experts, to data scientists, to internal reviewers, to external certifiers or regulators.

We envision workflows for interactions between these different personas via the auditing platform. Figure 2 summarizes our vision: governance experts define the intended use of synthetic data, the safeguards for compliance with regulations and policies, and the acceptable trade-offs between these safeguards. Next, the data scientist develops models and configures auditing tasks to rank developed models, perform model selection, and ensure compliance with the safeguards. Internal reviewers also have access to the platform, verify the compliance of models and created data with prescribed policies and safeguards and approve / reject models' deployment and synthetic data usage. Finally, a portable audit report is generated on the fly within the platform, which can be submitted to external third-party certifiers that probe the validity of the conclusions of the internal audit report. [1]

We believe that transparent reporting should become a *de facto* part of any AI application, model, or data (real or synthetic). We demonstrated how transparencies could be created within our framework both for internal testing and validation and for external auditing or certification. While our framework helps connect various key players, there is a need for additional organization measures, playbooks, and governance practices to harmonize and orchestrate such workflows. Another challenge in algorithmic auditing is the interpretable communication of how the technical metrics we compute relate to policy and legislation. To address this challenge, we adopt messages and warnings for detecting biases and harms to communicate auditing findings to policy experts. We envision a future auditing workflow that uses policy packs, which for a given application and set of policies, define templates for parameters, thresholds, technical metrics, and explanations.

1) Real Data Edge Cases: It is important to note that the original real data may contain inherent privacy breaches and imbalances, particularly in terms of its representation of underprivileged communities. While some of these issues can be mitigated during the training or inference phases of the Generative AI responsible for producing synthetic data, through techniques such as fair generative modeling or differential privacy. Synthetic data generated under privacy constraints can significantly improve the trustworthiness trade-offs compared to the original real data. As illustrated in Figure 3, the fairness, utility and robustness indices of synthetic data, when generated with privacy safeguards, show improvements over those of

[1] Snippets of such auditing workflows can be found in [49]

Auditing Framework	**Holistic Auditing Framework (ours)**	**Synthcity** [74]	**SDV** [71]	**Tapas** [48]
Data modalities				
Tabular Data	✓	✓	✓	✓
Time Series	✓	✓	✓	✗
Natural Language	✓	✗	✗	✗
Image	✓	✗	✗	✗
Auditing Dimensions, Interpretability and Transparency				
Fidelity	✓	✓	✓	✗
Privacy	✓	✓	✓	✓
Utility	✓	✓	✓	✓
Fairness	✓	✗	✗	✗
Robustness	✓	✗	✗	✗
Trustworthiness Index	✓	✗	✗	✗
Auditing Report	✓	✗	✗	✗

TABLE IV: Comparison of our Holistic Auditing Framework with Synthcity, SDV, and Tapas.

Generative AI Method	Trust Constraint
Non-Private & Private TrustFormers	
(Conditional) TrustFormer GPT (non-private): **TF(ω,n-p)**	Fidelity/Utility
(Conditional) TrustFormer GPT Trained with Differential Private SGD : **TF(ω,p-ε)**	Fidelity/Privacy Preserving synthetic Data/Utility
Differential Private Sampling From non-private (Conditional)TrustFormer [61], [77]	Fidelity/Privacy Preserving synthetic Data/utility
Non-private Baselines	
Gaussian Copula [71] **Gaussian Copula (n-p)**	Fidelity
Conditional Tabular GAN [71]**CTGAN(np)**	Fidelity/utility
Private Baselines	
Differential Private Probabilistic Graphical Model [63] **DP-PGM(p-ε)**	Fidelity/Privacy Preserving synthetic Data
DP-PGM (targeted) [63] **DP-PGM(target,p-ε)**	Fidelity/Privacy Preserving synthetic Data/Utility
(Conditional) Differential Private-GAN [74], [89] **DP-GAN(p-ε)**	Fidelity/ Privacy Preserving synthetic Data/Utility
(Conditional) PATE-GAN [74], [91]**PATE-GAN(p-ε)**	Fidelity/Privacy Preserving synthetic Data/Utility

TABLE V: Generative Models and TrustFormer Models audited within our framework. For private models we consider privacy budgets $\varepsilon = 1$ or 3.

the original real data. This enhancement in trustworthiness is crucial for ensuring that the synthetic data better serves all communities and protects sensitive information.
However, it is crucial to approach these improvements with caution. While differential privacy offers a robust framework for protecting individual data points, it does not inherently address the need for anonymization. Therefore, preprocessing steps such as data anonymization and augmentation should be considered. These preprocessing techniques can help mitigate the inherent trade-offs present in the real data, ultimately leading to higher quality and more trustworthy synthetic data outputs.

IV. Comparison to Previous Works

In Table IV, we present a comparison of our holistic auditing framework against three other prominent auditing frameworks: Synthcity [74], the Synthetic Data Vault (SDV) [71], and the Tapas Framework [48].Our framework stands out as it encompasses the widest range of data modalities and trustworthiness dimensions. Specifically, it supports tabular data, time series, natural language, and image data, whereas the other frameworks either cover fewer modalities or have limited support for some types. Additionally, our framework evaluates a broader set of trustworthiness dimensions and metrics including fidelity, privacy, utility, fairness and robustness. When it comes to transparent reporting and interpretability, our holistic auditing framework provides detailed auditing reports and a trustworthiness index that summarizes the audit result in an interpretable manner. Our holistic approach offers a clearer and more detailed account of trustworthiness trade-offs and the strategies employed to mitigate them. This enhanced transparency helps users better understand and manage the quality and reliability of their synthetic data across different trustworthiness dimensions.

In the following sections we present the training and auditing of various data generation techniques, including our TrustFormer Models, across a range of use cases encompassing tabular, time series, and natural language data (refer Table III). We also conduct audits on synthetic data generated by state-of-the-art diffusion models, particularly in the computer vision domain, focusing on the ImageNet classification task and assessing their generalization under distribution shifts. The models under audit are summarized in Table V, and we employ our trust dimension indices to assess each model's compliance with specific trust dimensions. For a given trustworthiness profile weights ω (refer to Table II for examples and their interpretation), we rely on our trustworthiness index to conduct TrustFormer model selection and rank the models based on their alignment with predefined safeguards (given in the trustworthiness profile ω).

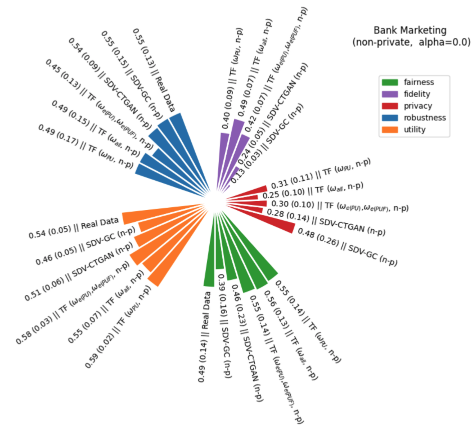
V. Tabular use cases

In this Section, we showcase our auditing framework and trustworthiness index driven model selection for TrustFormer on three tabular datasets: Bank Marketing Dataset [68], Recruitment Dataset [37], and the Law School Admission Council Dataset [88]. Please refer to Table III for details of the datasets, downstream tasks , safeguards and TF models training details.

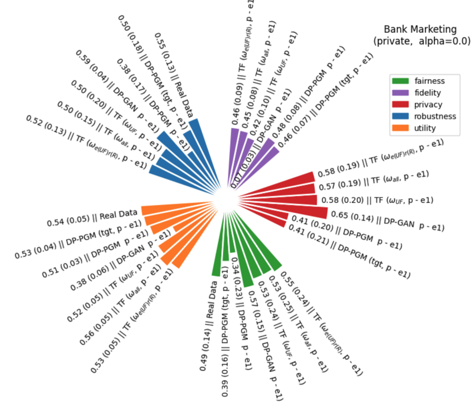
a) Setup: In our auditing setup, real data serves as the baseline, and we emphasize the contrast with synthetic generated data. The real data is split into training ($D_{r,\text{train}}$), validation ($D_{r,\text{val}}$), and test ($D_{r,\text{test}}$) sets through random sampling without replacement, repeated five times with different seeds, creating five real data folds ($\mathcal{D}_r$). Each generative AI method listed in Table V is trained five times independently on each real data fold. For TrustFormers, both non-private (TF(n-p)) and private (TF(p-$\varepsilon = 1$) or TF(p-$\varepsilon = 3$)) versions are trained. Trustworthiness Index Driven Model selection is performed independently within each fold, leading to the selection of a checkpoint t^*. Synthetic datasets are generated by sampling from each generator trained on a specific fold. This results in $\mathcal{D}_s = \{D_s^\ell\}_{\ell=1}^5$. Audits on downstream tasks, assess utility, fairness, and robustness evaluated on the corresponding real test data folds. A non-private tabular RoBERTa Embedding (**E**) [70], trained on the real data D_r (excluding task labels), is used for certain metrics in Table I.

b) Real Data, Downstream Tasks, and Protected communities: The *Bank Marketing Dataset* [68] has 45211 samples. Each sample is a row with 17 fields representing a client. The downstream task is to predict if the client will subscribe to a term deposit or not. Protected groups include individuals with value of field `age` less than 30. The *Recruitment Dataset* [37] contains 6000 samples consisting of rows with 14 demographic fields. The classification task associated to this dataset is the prediction of a candidate's employment. The sensitive variable is defined as the binary indicator `white`. Finally, the *Law School School Admission Council Dataset* [88] has 20461 samples of with 11 fields of demographic features. The downstream task is the prediction of whether a candidate was admitted to law school or not. The sensitive variable is the binary indicator `black`.

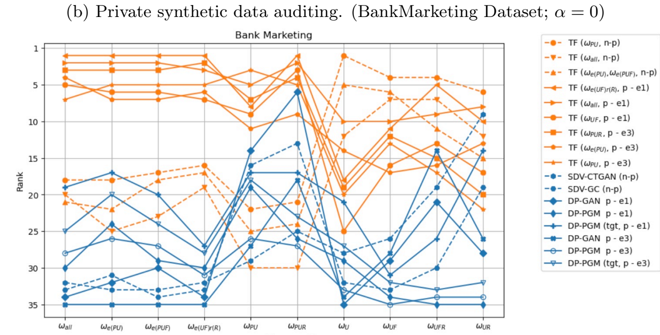
c) Audit Results: For the Bank Marketing Dataset, we present average trust dimension indices and their "variances" for non-private and private synthetic data generated by TrustFormers models (TF) in Figure 3 (a) and (b). In this case, TF models were selected across various trustworthiness profiles (as outlined in Table II) using the trustworthiness index from Equation (2) for $\alpha = 0$. We also include results for non-private and private baselines (refer to Table V), shown in panels (a) and (b). Additionally, we provide information about the utility, fairness, and robustness of downstream tasks trained on real data as a reference for both panels (a) and (b). Panel (c) displays the ranking according to trustworthiness index of synthetic data across different trustworthiness profiles based on the same α value. Similar plots for the Recruitment and Law School datasets can be found in Figure 5 and in Supplementary Information W Figure 10, respectively, for



(a) Non-private synthetic data auditing. (BankMarketing Dataset; $\alpha = 0$)



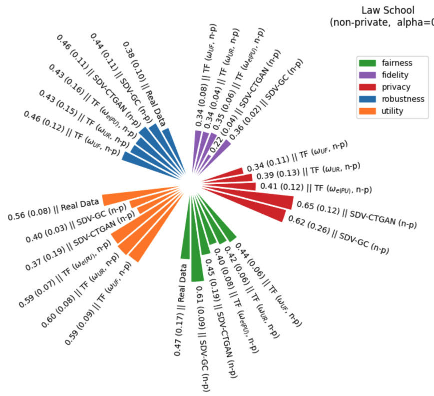
(b) Private synthetic data auditing. (BankMarketing Dataset; $\alpha = 0$)



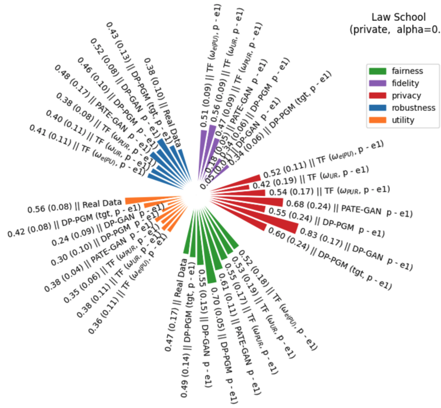
(c) Ranking of synthetic data.(BankMarketing Dataset; $\alpha = 0$)

Fig. 3: Summary of auditing and ranking results on the Bank Marketing dataset using the trustworthiness index given in (2) for $\alpha = 0$. (a) and (b) show trust dimension indices π_T(where "T" corresponds to Fidelity, Privacy, Utility, Fairness, or Robustness), and their "variance" (Δ_T) on TrustFormer (TF) and baseline models. The format is $\pi_T(\Delta_T)||$ Name of the synthetic data model. (c) shows the ranking of the models across different trustworthiness profiles ω given in Table II.

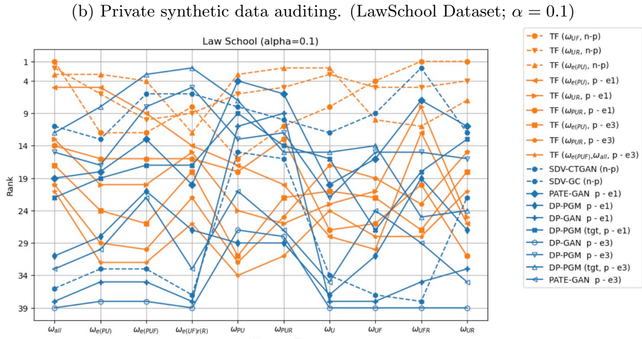
This article has been accepted for publication in IEEE Journal on Emerging and Selected Topics in Circuits and Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JETCAS.2024.3477976



(a) Non-private synthetic data auditing. (LawSchool Dataset; $\alpha = 0.1$)

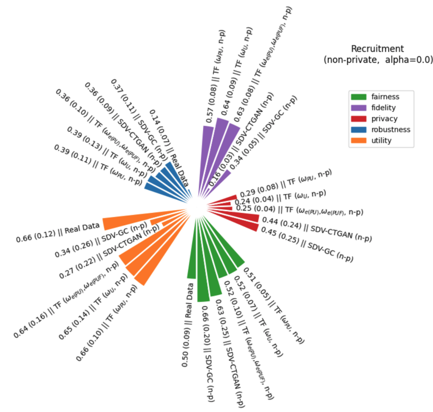


(b) Private synthetic data auditing. (LawSchool Dataset; $\alpha = 0.1$)

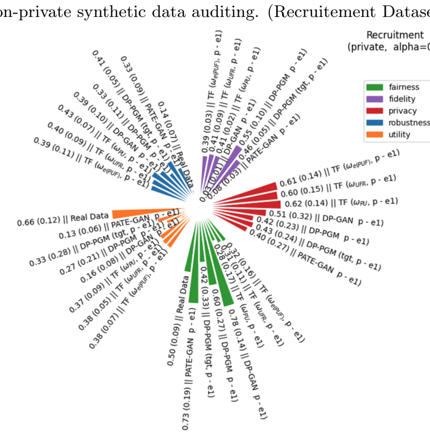


(c) Ranking of synthetic data (LawSchool Dataset; $\alpha = 0.1$)

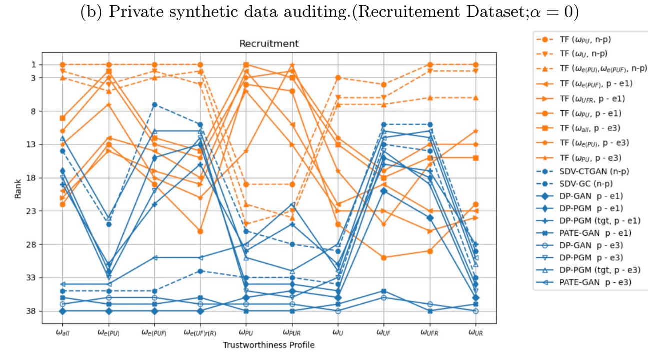
Fig. 4: Summary of auditing and ranking results on the Bank Marketing dataset using the trustworthiness index given in (2) for $\alpha = 0.1$. This means that the uncertainty is taking into account in both model selection and in ranking the synthetic data using the trustworthiness index. (a) and (b) show trust dimension indices π_T (where 'T' corresponds to Fidelity, Privacy, Utility, Fairness, or Robustness), and their "variance" (Δ_T) on TrustFormer (TF) and baseline models. The format is $\pi_T(\Delta_T)||$ Name of the synthetic data model. (c) shows the ranking of the models across different trustworthiness profiles ω given in Table II.



(a) Non-private synthetic data auditing. (Recruitement Dataset; $\alpha = 0$)



(b) Private synthetic data auditing.(Recruitement Dataset;$\alpha = 0$)



(c) Ranking of synthetic data for $\alpha = 0$.(Recruitement Dataset;$\alpha = 0$)

Fig. 5: Summary of auditing and ranking results on the recruitment dataset using the trustworthiness index given in (2) for $\alpha = 0$. (a) and (b) show trust dimension indices π_T (where "T" corresponds to Fidelity, Privacy, Utility, Fairness, or Robustness), and their "variance" (Δ_T) on TrustFormer (TF) and baseline models. The format is $\pi_T(\Delta_T)||$ Name of the synthetic data model. (c) shows the ranking of the models across different trustworthiness profiles ω given in Extended Table II.

This article has been accepted for publication in IEEE Journal on Emerging and Selected Topics in Circuits and Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JETCAS.2024.3477976

$\alpha = 0$. To account for data split uncertainties, we've included audit results for $\alpha = 0.1$, depicted in Figure 4 for the Law School dataset, and in the Supplementary Information for the other datasets (see Figures 11 and 12 in Supplementary Information W).

d) Discussion of the Audit Results: Trust Dimension Tradeoffs and Mitigations (Panels (a) and (b)): Analyzing these results we make the following observations. First, when compared with real data, TF synthetic data demonstrates on par or superior performance in utility, fairness, and robustness indices across various datasets, while effectively balancing privacy, fidelity, and other trust dimensions. A careful trustworthiness index model selection of synthetic data, in conjunction with classical classifier model selection, facilitates alignment with prescribed safeguards without compromising performance. Second, the selection of synthetic data for TrustFormer models guided by the trustworthiness index consistently leads to competitive performance with all other baseline methods across trust dimension indices as can be seen in panels (a) and (b) in all aforementioned figures. TrustFormers either achieve the highest index on each trust dimension or rank among the top-performing synthetic data, irrespective of the specific generative methods considered. Third, a distinction is observed between non-private and differentially private synthetic data, with fidelity and utility exhibiting higher indices in non-private data and privacy, fairness, and robustness showing higher indices in differentially private data. This aligns with the well-documented trade-offs in fairness, utility, privacy, and robustness in the literature [4], [81].

e) Discussion of the Ranking Results and Controllable Trade-offs (Panels (c)): For both Bank Marketing dataset and Recruitment datasets we see that in both cases $\alpha = 0$ and 0.1, TF synthetic data outperforms other baselines across all trustworthiness profiles ω in terms of its trustworthiness index . We see in Panel (c) Figure 3 and Figure 5, a clear phase transition between private and non private TF models depending on whether the trustworthiness profile highlights privacy as a requirement or not. This effect is achieved through our trustworthiness index model selection that offers control over trust trade-offs. On the other hand, for the law school dataset, TF models fall short with respect to baselines when uncertainty is not considered ($\alpha = 0$, Supplementary Figure 10, but lead to top performing synthetic data when uncertainty is considered ($\alpha = 0.1$, Figure 4). This highlights the importance of assessing the uncertainty in auditing synthetic data.

VI. Time-Series Use Cases

We audit in this Section the use of time series synthetic data in healthcare focusing on the utility/fairness tradeoffs. We also audit its use in a financial application, fraud detection, with a focus on utility/privacy tradeoffs.

a) Use Case I: MIMIC-III Controllable Trust trade-offs on Healthcare Data: We explore in this use case the promise of synthetic data in the highly regulated healthcare domain, where patient privacy and anti-discrimination regulations are enforced by law. This prohibits hospitals from sharing data in order to not expose the patients personal information. Moreover recent studies [79] showed on the MIMIC-III (Medical Information Mart for Intensive Care) time-series benchmark [52] that it has an inherent bias and discrimination [27], [64], [79]. We explore controllable trust trade-offs on synthetic times-series data obtained from learned TrustFormer models on this dataset.

b) MIMIC-III dataset.: MIMIC-III (Medical Information Mart for Intensive Care) dataset [52] is a large database of about 40K patients with de-identified records collected during their stay in intensive care unit (ICU). The records contain high temporal resolution data including lab results, electronic documentation, and bedside monitor trends collected every hour. For each admission, we have an entry every hour of vitals measurement for a total of 48 entries capturing the dynamics in a patient state. Each hour, we have about 18 columns of vitals measurements augmented with a time-stamp, subject ID, and information of gender, ethnicity, and age. For complete in-depth description of the data, please refer to the MIMIC-III extensive documentation [52].

c) Downstream Task: The In-Hospital Mortality (IHM) prediction task aims at predicting the mortality of patients in the ICU after a 48-hour stay. Given patients vitals evolution over the course of 48 hours the goal is to predict potential mortality of each patient. The data is therefore a time-series of measurements leading to a classification decision: did the patient expire or not. The training/val/test sets are composed of 14681/3222/3236 admissions respectively. Several studies on the MIMIC-III dataset, pointed the unfairness inherent to this dataset, disfavoring patients based on their ethnicity.

d) Synthetic Data with Controllable Trust Trade-offs on MIMIC-III: In order to provide controllable trust trade-offs, we trained Tabular time-series GPT models [70] with regular and private differential training for a privacy budget $\varepsilon = 3$ (See Supplementary Information for details on data preparation, model architecture and training hyper-parameters). For a trade-off weight ω, we use our trustworthiness index cross-validation to align the models with desired trust trade-offs. This results with the selected TrustFormers models: TF(ω, n-p) and TF(ω, p-$\varepsilon = 3$). For inference from these models we used multinomal decoding (mn) or top-k decoding (sampling from top-k softmaxes for k=50), and refer to resulting trustformer models as: TF(ω, n-p, mn/top-k) and TF(ω, p-$\varepsilon = 3$, mn/top-k). In order to audit this time-series dataset, we train an embedding $\mathbf{E}$ that is a TabRoBERTa model [70]. A masked language model is trained to predict masked fields from the patient vital records (masking 10% of fields). The vital records contains all the measurements over the 48-hour stay, we exclude patient IDs and labels from the TabRoBERTa training.

e) Trust Dimension Indices: Table VI summarizes the trust dimension indices of selected TrustFormers models, where the downstream tasks are evaluated on the real test set. Interestingly

Model	Fidelity	Privacy	Utility	Fairness	Robustness
Non-private TrustFormer					
TF (ω_{UR}, n-p, mn)	0.70	0.11	0.57	0.40	0.55
TF ($\omega_{e(PUF)}$,ω_U,ω_{UF},$\omega_{e(UF)r(R)}$,ω_{all},ω_{UFR}, n-p, mn)	0.37	0.36	**0.67**	0.42	0.50
TF ($\omega_{e(PU)}$,ω_{PUR}, n-p, mn)	0.53	0.21	0.51	0.29	0.49
TF (ω_{PU}, n-p, mn)	0.81	0.42	0.56	0.25	0.39
TF (ω_{UR}, n-p, topk)	0.70	0.21	0.52	0.43	0.60
TF (ω_{UF},ω_{UFR}, n-p, topk)	**0.86**	0.49	0.61	**0.44**	0.52
TF (ω_U, n-p, topk)	0.62	0.24	0.62	0.32	**0.69**
TF ($\omega_{e(PU)}$,$\omega_{e(PUF)}$,ω_{PU},$\omega_{e(UF)r(R)}$,ω_{all},ω_{PUR}, n-p, topk)	0.77	**0.54**	0.60	0.29	0.64
Private TrustFormer					
TF ($\omega_{e(PU)}$,$\omega_{e(PUF)}$,ω_{PU},$\omega_{e(UF)r(R)}$,ω_{all}, p - ϵ =3, mn)	0.34	0.93	0.30	0.58	0.41
TF (ω_U,ω_{UF},ω_{UFR},ω_{UR},ω_{PUR}, p - ϵ =3, mn)	0.19	0.69	**0.44**	0.55	0.47
TF (ω_{PU}, p - ϵ =3, topk)	0.15	**1.00**	0.22	**0.83**	0.46
TF (ω_{PUR}, p - ϵ =3, topk)	0.35	0.89	0.40	0.46	**0.50**
TF ($\omega_{e(PU)}$,$\omega_{e(PUF)}$,ω_{UF},ω_{all}, p - ϵ =3, topk)	0.40	0.81	0.33	0.57	0.29
TF ($\omega_{e(UF)r(R)}$, p - ϵ =3, topk)	0.34	0.75	0.26	0.31	0.40
TF (ω_{UFR},ω_{UR}, p - ϵ =3, topk)	0.42	0.62	0.41	0.60	0.40
TF (ω_U, p - ϵ =3, topk)	**0.64**	0.12	0.41	0.59	0.34
Real Data	N/A	N/A	0.44	0.07	0.67

Table VI: MIMIC-III/ In-Hospital Mortality downstream task evaluation: trust dimension indices of TrustFormer models. In bold highest index within each group of synthetic data. In blue highest value across all methods including real data.

Model	ω_{all}	$\omega_{e(PU)}$	$\omega_{e(PUF)}$	$\omega_{e(UF)r(R)}$	ω_{PU}	ω_{PUR}	ω_U	ω_{UF}	ω_{UFR}	ω_{UR}
Non-Private TrustFormer										
TF (ω_{UR}, n-p, mn)	13	15	15	14	15	15	5	6	5	6
TF ($\omega_{e(PUF)}$,ω_U,ω_{UF},$\omega_{e(UF)r(R)}$,ω_{all},ω_{UFR}, n-p, mn)	8	7	9	5	8	5	**1**	**1**	**1**	**3**
TF ($\omega_{e(PU)}$,ω_{PUR}, n-p, mn)	15	14	14	15	14	14	8	14	12	7
TF (ω_{PU}, n-p, mn)	9	9	10	10	9	10	6	15	14	8
TF (ω_{UR}, n-p, topk)	7	12	12	9	13	13	7	7	4	4
TF (ω_{UF},ω_{UFR}, n-p, topk)	**1**	**1**	**1**	**1**	4	**3**	**3**	**2**	**2**	5
TF (ω_U, n-p, topk)	6	10	11	11	12	7	**2**	8	**3**	**1**
TF ($\omega_{e(PU)}$,$\omega_{e(PUF)}$,ω_{PU},$\omega_{e(UF)r(R)}$,ω_{all},ω_{PUR}, n-p, topk)	**2**	**2**	**3**	**3**	**2**	**1**	4	12	7	**2**
Private TrustFormer										
TF ($\omega_{e(PU)}$,$\omega_{e(PUF)}$,ω_{PU},$\omega_{e(UF)r(R)}$,ω_{all}, p - ϵ =3, mn)	5	5	5	7	5	6	14	13	13	13
TF (ω_U,ω_{UF},ω_{UFR},ω_{UR},ω_{PUR}, p - ϵ =3, mn)	11	8	7	8	**3**	4	9	4	6	9
TF (ω_{PU}, p - ϵ =3, topk)	12	11	8	12	10	9	16	10	10	15
TF (ω_{PUR}, p - ϵ =3, topk)	**3**	**3**	**2**	6	**1**	**2**	12	11	9	10
TF ($\omega_{e(PU)}$,$\omega_{e(PUF)}$,ω_{UF},ω_{all}, p - ϵ =3, topk)	10	6	6	4	6	11	13	9	15	16
TF ($\omega_{e(UF)r(R)}$, p - ϵ =3, topk)	14	13	13	16	11	12	15	16	16	14
TF (ω_{UFR},ω_{UR}, p - ϵ =3, topk)	4	4	4	**2**	7	8	10	**3**	8	11
TF (ω_U, p - ϵ =3, topk)	16	16	16	13	16	16	11	5	11	12

Table VII: MIMIC-III synthetic dataset ranking using the trustworthiness index corresponding to the trade-off weight ω. We see that two models stand out across different trade-offs and they correspond to different decoding strategies.

similar to tabular data, we see that TrustFormer synthetic data outperforms real data on all trust dimensions. Interestingly the fairness index of real data is the lowest, and synthetic data therefore improves the fairness/utility tradeoff . Similar observations on the relationship between trust constraints and trust trade-offs we made on tabular data hold for the time series case.

f) Analysis of Results: We compare the performance of the two decoding strategies considered (multinomial and top-k decoding) for synthetic data generation from TF models and study how it impact trust trade-offs. Note that we used herein fixed data splits from the literature and we don't report therefore the uncertainty of the audit. Table VII gives the ranking of these synthetic datasets using the trustworthiness index for all trustworthiness profiles. We see that for some trustworthiness profiles, non-private TrustFormer with multinomial decoding stands out, while top-k decoding outperforms it for other profiles. On private models we see that TF models with top-k decoding are the best at balancing the privacy/utility trade-offs. This shows that our auditing framework highlights the effect of hyper-parameters choices such as decoding strategies and their impacts on all trust dimensions

Use Case II: Fraud Detection, Deep Dive on Utility and Privacy trade-offs in Synthetic Data

In this use case, we investigate the use of synthetic data in a financial application for training fraud detectors. We focus here on the impact of the synthetic data on the utility and

Training Data for Fraud Detector		Training Regime for TabRoBERTa Feature Extractor					
		Private					Non-Private
		ε=1	ε=3	ε=10	ε=30	ε=1000	
Real		0.72	0.77	0.73	0.83	0.80	0.88
Synthetic (TabGPT)	Private (ε=0.1)	0.48	0.49	0.47	0.45	0.48	0.50
	Private (ε=1)	0.48	0.47	0.47	0.47	0.48	0.48
	Private (ε=5)	0.49	0.48	0.48	0.47	0.48	0.48
	Private (ε=20)	0.49	0.49	0.50	0.50	0.51	0.50
	Private (ε=50)	0.51	0.54	0.60	0.56	0.59	0.70
	Private (ε=100)	0.64	0.63	0.73	0.72	0.74	0.75
	Private (ε=200)	0.66	0.66	0.72	0.71	0.78	0.77
	Nonprivate	0.61	0.57	0.66	0.70	0.72	0.79

TABLE VIII: Performance (F1-macro) of the fraud classifier on the test set of credit card transactions for different training choices of classifier (rows) and TabRoBERTa features extractor (columns).

	Epoch	Fidelity	Utility	Privacy	Fairness	ω_{all}	ω_U	ω_{UF}
$\text{BioGPT}_{\text{Finetuned}}$	3	0.29	0.55	1.00	0.53	**2**	**2**	**2**
	5	0.44	0.38	0.75	0.60	**3**	4	4
	7	0.84	0.51	0.50	0.90	**1**	**3**	**1**
	9	0.89	0.88	0.25	0.33	4	**1**	**3**

TABLE IX: Trust Indices of $\text{BioGPT}_{\text{Finetuned}}$ on MIMIC-III notes. Robustness dimension is not evaluated herein. trustworthiness index Ranking of synthetic data sampled from different epoch during the finetuning of BioGPT model (Note that the robustness dimension was not considered) . The ranking corresponds to different trust tradeoffs ω: for w_U that is accuracy driven, we see that the last epoch is outperforming the other ones; When in addition we consider the fairness of the prediction (ω_{UF}) the last epoch (epoch 9) ranks third and the epoch 7 presents better utility/ Fairness trade-offs.

privacy tradeoffs, The goal herein is therefore to highlight a use case of synthetic data in an end to end fashion without reporting aggregation level of metrics to have a more in depth analysis of the privacy/ utility trade offs. To conduct our study, we use the credit card transactions of [7], [70] to train our TrustFormer models. These transactions were created using a rule-based generator, where values were generated through stochastic sampling techniques. The dataset contains 24 million transactions from 20,000 users, with each transaction (row) consisting of 12 fields (columns) that include both continuous and discrete nominal attributes.

g) Training RoBERTa-like Embedding: To train TabRoBERTa on our transaction dataset, we constructed samples as sliding windows of 10 transactions, using a stride of 5. We excluded the label column, "isFraud?", during training to prevent biasing the learned representation for the downstream fraud detection task. We masked 15% of a sample's fields, replacing them with the [MASK] token, and predicted the original field token using cross-entropy loss. We used DP-SGD for transformer models [57] to train various RoBERTa-like models with differing degrees of privacy, ranging from highly private ($\varepsilon = 1$) to non-private ($\varepsilon = 1000$). Additionally, we trained a RoBERTa model without private training (see the columns labeled "Private" and "Non-Private" in Table VIII).

h) Synthetic data generation: We generated several privacy-preserving synthetic datasets using our non-private pretrained TabGPT model. For model selection in this experiment we relied on a fidelity validation of the TabGPT model. To generate private synthetic data, we used a private sampling technique [77] , which involves adding Laplacian noise with controlled variance (dependent on the user-provided ε value) to the probability distribution over the generated tokens from the non-private GPT model. This is a form of output-perturbation methods that guarantees differential privacy.We generated seven datasets with varying privacy levels, from highly private ($\varepsilon = 0.1$) to non-private ($\varepsilon = 200$), as shown in the rows labeled "Synthetic" in Table VIII. Additionally, we considered real card transaction data and synthetically generated data without private sampling.

i) Training the downstream Fraud Detection Model: Given the various transaction datasets, we constructed a simple multi-layer perceptron (MLP) classifier that was trained directly on the embeddings of the various RoBERTa feature extractors that we trained . Note that thanks to the additivity property of differential privacy the overall privacy of the fraud detector is the addition of the privacy budget of synthetic data and the privacy budget of the feature extractor. The RoBERTa feature extractor remained fixed during the fraud detector training. For each training scenario, we selected 800K transactions for training, 100K transactions for validation, and 100K transactions for testing. Note that the test transactions were always the same across different datasets and were chosen from real data. In contrast, the training and validation splits were determined according to the training regimes.

j) Results and Discussion: The highest utility performance is achieved when using a RoBERTa feature extractor trained without differential privacy, and when training the fraud detector on real transaction data (first row in the table).

	ImageNet v1				ImageNet v2			
	sig 0.0	sig 0.1	sig 0.3	sig 0.5	sig 0.0	sig 0.1	sig 0.3	sig 0.5
real	**74.9** / **92.3**	**72.6** / **91.2**	62.7 / 84.7	48.8 / 73.3	**63.1** / **84.5**	**60.4** / **82.7**	48.6 / 72.5	34.8 / 58.6
synthetic	54.8 / 76.2	50.3 / 71.7	39.7 / 60.6	27.9 / 46.7	45.6 / 68.9	40.7 / 62.7	29.9 / 50.1	19.0 / 35.8
real + 0.5syn	74.5 / 92.0	72.5 / 90.9	**63.1** / **85.0**	**49.8** / **74.1**	62.3 / 83.9	59.5 / 82.2	**48.8** / **73.0**	**35.4** / **59.5**
real + 1.0syn	**74.9** / 91.9	72.3 / 90.9	62.4 / 84.5	48.8 / 73.0	62.5 / 83.6	59.8 / 82.2	49.1 / 72.2	34.1 / 57.5

TABLE X: Performance of ResNet50 on ImageNet v1 and v2 Datasets under varied noisy conditions (additive random Gaussian noise to tested images) and different training regimes (synthetic data augmentation). The resulting models are evaluated in Acc@1 (first number) and Acc@5 (second number). We see that while classifiers trained purely on synthetic data lag behind those trained on real, augmenting real data with synthetic images makes models more resilient to noise.

Conversely, utilizing a highly private RoBERTa model in conjunction with highly private synthetically generated data yields (unsurprisingly) significantly poorer F1-macro performance (upper left corner of the table). Furthermore, it can be observed that for a fixed row (dataset for training the fraud detector), moving from left to right across columns (corresponding to decreasing privacy levels of the RoBERTa feature extractor) results in improved utility performance for the fraud detector. Similarly, for a fixed column (pretrained RoBERTa feature extractor), moving down the rows (excluding the first row, which depicts performance on real data, and excluding the last row which depicts performance on non-private synthetic data) leads to better classifier performance. It is interesting to see when comparing to the last row, that private synthetic data with private embeddings introduces a regularization effect leading to better performance than the same setup with non-private synthetic data.

VII. Natural Language Synthetic Data : Deep Dive on Utility and Fairness Trade-offs

In this section, we delve into controlling trust trade-offs within the context of language modeling, using the BioGPT model [60] as our testbed. We fine-tune a BioGPT-Large model on the MIMIC-III notes dataset [5], comprising 423,015 patient notes, with an accompanying label denoting patient survival (expiration flag). MIMIC-III is known for its bias issues [27], [64], [79].

We augment the MIMIC III notes dataset with patient age, gender, and ethnicity, and fine-tune the model on the resulting data. We then fine-tune the BioGPTLarge model (non-private training) on this augmented notes data which is first prompted by the target label and then by the controls (ethnicity, age, gender). At inference time, from a fine-tuned BioGPTLarge model at a given epoch, we generate a balanced synthetic dataset of the same size as the real training data via prompting the model with the same amount of positive and negative labels (expiration flag). In this setting, we use a multinomial decoding strategy for generation. At the end we obtain a labeled synthetic dataset of synthetic doctor notes along with the controls on ethnicity, age and gender for each sample.

We audit the synthetic dataset sampled from different epochs (namely after 3, 5, 7 and 9) during the fine-tuning process. The downstream task we consider in this audit is the in hospital mortality prediction task, where the protected community is the ethnicity "ASIAN" [27]. As an embedding **E**, we use the original pre-trained BioGPT model [60] to extract embeddings for the synthetic notes as it has the capability of representing the biomedical domain. In Table IX we see that our trustworthiness index driven model selection allows a controllable trade-off between utility and fairness : epoch 7 has a better utility fairness trade-off than the model at the last epoch that has higher utility. Therefore it is favorable to select the model at epoch 7 at the price of a reduced utility but with an enhanced fairness, which is of paramount importance for this use-case.

VIII. Synthetic Image Data: Deep Dive on Utility and Robustness to Noise and Distribution Shift

We consider here the use of synthetic image data for training classifiers on one of the key computer vision datasets, Imagenet [33], with a focus on the utility and generalization propreties of these classifiers on the Imagenet test set and under distribution shifts. Imagenet consists of 1.2 Million images of size 256x256x3 and 1000 categories. Authors in [78] constructed Imagenetv2 test set, that consists of a distribution shift from the original imagenet distribution and reported a performance drop between $11\% - 14\%$. Recent works [11] and [21] showed promising results of synthetic data from diffusion models (a family of generative models that uses diffusion techniques [83]) in improving Imagenet classification. Following these promising works, we synthesize 1.2 M labeled images from the Imagenet 256x256 pretrained guided diffusion models from OpenAI [34].Table X presents ResNet50 [44] performance on ImageNet-v1 and v2 under Gaussian noise, considering four training data scenarios: (1) real Imagenet; (2) synthetic data; (3) a hybrid with real and synthetic images (real to synthetic ratio $1/0.5$); and (4) an equal mix totaling 2.4 million images. Results show synthetic-only models lag, but integrating with real images enhances robustness, especially in noisier settings (sigma 0.3 and 0.5). Combining synthetic data with real images improves a model's resilience to noise and image corruption.

IX. Conclusion

We introduced a holistic framework for auditing synthetic data along trust pillars. Towards this end, we defined a trustworthiness index that assesses the trade-offs between trust dimensions such as fidelity, privacy, utility, fairness, and robustness and quantifies their uncertainty. Moreover, we devised a trustworthiness index driven model selection and cross-validation via auditing in the training loop, that allows controllable trust trade-offs in the resulting synthetic data. Finally, we instrumented our auditing framework with workflows connecting various stakeholders from model development to certification, and we defined templates to communicate transparency about model audits via a Synthetic Data auditing report.

Our framework highlights the potential of synthetic data in various modalities, including tabular, time series, natural language, and vision. However, for critical applications where the trustworthiness of synthetic data is paramount for its integration into the AI lifecycle of sensitive downstream tasks, it is important to recognize that not all generative AI techniques and training approaches are equally reliable. Thus, conducting rigorous audits of synthetic data is imperative to guide the training process and obtain certifications for internal use, third-party entities, and regulatory compliance.

X. Data availability

Data used in the paper is available online and open source. See Table III for references.

XI. Code availability

A code reproducing tables in the paper for the recruitment dataset is available on https://ibm.biz/synthetic-audit. Examples of full auditing reports on all use cases are provided in the Supplementary information.

XII. Acknowledgments

We thank IBM Research for supporting this work. Y.M would like to thank Payel Das, Kush R. Varshney and Abdel Hamou-Lhadj for insightful discussions.

XIII. Authors Contributions

Y.M conceived the project and wrote the initial draft of the paper. All authors contributed to developing the synthetic data auditing framework and designed experiments and contributed to their analysis. All authors contributed to the writing of the paper.

XIV. Ethics declarations

The authors declare no competing interests.

Brian Belgodere is a senior research software engineer in the Emerging Tech. Engineering Department at IBM Research. He currently works on hybrid cloud platforms and tooling for Machine Learning and AI. He holds B.S. degrees in economics and business administration from Carnegie Mellon University and a Juris Doctor from the University of Pittsburgh.

Pierre Dognin received his M.S. degree and Ph.D degree in Electrical Engineering from the University of Pittsburgh in 1999 and 2003. He then joined IBM Research in the Human Language Technologies Department where he worked as a Research Staff Member on Automatic Speech Recognition, focusing on acoustic modeling, robust speech recognition, and core algorithm technology. Since 2015, he has been working on multimodal Machine Learning and is now in the Trusted AI Department where his research interests include Deep Learning, Generative Modeling for Computer Vision and Natural Language.

Igor Melnyk is currently a Research Staff Member in the Trusted AI Department at IBM Research. Before that he completed his PhD degree in Computer Science and Engineering in 2016 from the University of Minnesota. In 2010, he obtained his MS degree in Computer Science from the University of Colorado, Boulder. His research interests are in the areas of Machine Learning and AI with the focus on the problems of Natural Language Processing and Computer Vision.

Youssef Mroueh is a principal research scientist in the Trusted AI department at IBM Research and a principal investigator in the MIT-IBM Watson AI lab. He received his PhD in computer science in February 2015 from MIT, CSAIL, where he was advised by Professor Tomaso Poggio. In 2011, he obtained his engineering diploma from Ecole Polytechnique Paris France, and a master of science in Applied Maths from Ecole des Mines de Paris. He is interested in Multimodal Deep Learning, Generative modeling, Computer Vision and Learning Theory.

Aleksandra (Saška) Mojsilović is a Serbian-American scientist. Her research interests are artificial intelligence, data science, and signal processing. She is known for innovative applications of machine learning to diverse societal and business problems. Her current research focuses on issues of fairness, accountability, transparency, and ethics in AI. She is an IBM Fellow and IEEE Fellow.

Jiri Navratil received M.Sc. ('95) in EE & PhD ('99) in CS from Technische Universitaet Ilmenau, Germany. He joined IBM Research in 1999, where he is currently a Principal Research Staff Member. His research interests include Deep Learning, Natural Language Processing and Industrial AI Applications.

Apoorva Nitsure is a Research Software Engineer at IBM Research. Her interests include Deep Learning, Natural Language Processing and Interpretability in Machine Learning to help users make informed choices while using AI. Prior to joining IBM Research, she completed her Masters degree from Heinz College, Carnegie Mellon University and B.Tech in IT from CoEP India.

Inkit Padhi received Masters's degree in Computer Science from Viterbi School of Engineering, USC/ISI in May 2016. Since December of 2018, he is working as a research engineer in the Trusted AI department at IBM Research. He is interested in Machine Learning and Natural Language Processing. His pronouns are he, him, and his.

Mattia Rigotti is a Research Staff Member in the AI Automation Department at IBM Research, which he joined in 2014. He holds an M.Sc. degree in Theoretical Physics from ETH Zurich and a Ph.D. in Computational Neuroscience from Columbia University. His research interests include Deep Learning, Neuromorphic Engineering and Computational Neuroscience.

Jarret Ross received an M.S. in Computer Science from Wichita State University. He has been with IBM research since 2016 and he is currently a research engineer in the Trusted AI department. His interests are in deep learning, distributed computing and generative modeling.

Yair Schiff is a PhD student in the Computer Science department at Cornell University. Yair was a software developer working on IBM Watson Studio. Prior to joining IBM, he completed an M.S. degree in Computer Science at the Courant Institute of Mathematical Sciences at New York University. Yair collaborated with the Trusted AI department at IBM Research. .

Radhika Vedpathak is a designer with an experience in developing design for web applications.

Richard A. Young received his B.S. Hons degree in computer science from The University of the Witwatersrand (South Africa) in 2010. He worked as a software developer in the banking and health insurance industries from 2011 to 2018. He joined IBM Research in South Africa as a Research Engineer in 2018 where he specialises in building web and cloud based applications.

REFERENCES

[1] Amina A. Abdu, Irene V. Pasquetto, and Abigail Z. Jacobs. An empirical analysis of racial categories in the algorithmic fairness literature. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1324–1333, New York, NY, USA, 2023. Association for Computing Machinery.

[2] David Adkins, Bilal Alsallakh, Adeel Cheema, Narine Kokhlikyan, Emily McReynolds, Pushkar Mishra, Chavez Procope, Jeremy Sawruk, Erin Wang, and Polina Zvyagina. Method cards for prescriptive machine-learning transparency. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*, pages 90–100, 2022.

[3] Akshay Agarwal and Nalini K Ratha. Black-box adversarial entry in finance through credit card fraud detection. In *CIKM Workshops*, 2021.

[4] Sushant Agarwal. Trade-offs between fairness and privacy in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*, 2021.

[5] Betty Van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix A. Gers, and Alexander Löser. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 881–893. Association for Computational Linguistics, 2021.

[6] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022.

[7] Erik Altman. Synthesizing credit card transactions. In *Proceedings of the Second ACM International Conference on AI in Finance*, pages 1–9, 2021.

[8] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. Natesan Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6:1–6:13, 2019.

[9] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilovic, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. Ai explainability 360: Impact and design, 2021.

[10] Samuel A. Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E. Tillman, Prashant Reddy, and Manuela Veloso. Generating synthetic data in finance: Opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, ICAIF '20, New York, NY, USA, 2021. Association for Computing Machinery.

[11] 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.

[12] Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. Molgpt: Molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076, 05 2022.

[13] Vincent Ballet, Jonathan Aigrain, Thibault Laugel, Pascal Frossard, Marcin Detyniecki, et al. Imperceptible adversarial attacks on tabular data. In *NeurIPS 2019 Workshop on Robust AI in Financial Services: Data, Fairness, Explainability, Trustworthiness and Privacy (Robust AI in FS 2019)*, 2019.

[14] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[15] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

[16] Karan Bhanot, Miao Qi, John S Erickson, Isabelle Guyon, and Kristin P Bennett. The problem of fairness in synthetic healthcare data. *Entropy*, 23(9):1165, 2021.

[17] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[18] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[19] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[20] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018.

[21] Max F Burg, Florian Wenzel, Dominik Zietlow, Max Horn, Osama Makansi, Francesco Locatello, and Chris Russell. A data augmentation perspective on diffusion models and retrieval. *arXiv preprint arXiv:2304.10253*, 2023.

[22] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

[23] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6, 2021.

[24] Francesco Cartella, Orlando Anunciacao, Yuki Funabiki, Daisuke Yamaguchi, Toru Akishita, and Olivier Elshocht. Adversarial attacks for tabular data: Application to fraud detection and imbalanced data. *arXiv preprint arXiv:2101.08030*, 2021.

[25] Clément Chadebec, Elina Thibeau-Sutre, Ninon Burgos, and Stéphanie Allassonnière. Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2879–2896, 2023.

[26] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 343–362, 2020.

[27] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31, 2018.

[28] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021.

[29] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, pages 1887–1898. PMLR, 2020.

[30] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[31] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021.

[32] Jessamyn Dahmen and Diane Cook. Synsys: A synthetic data generation system for healthcare applications. *Sensors*, 19(5):1181, 2019.

[33] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[34] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[35] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, 2014.

[36] Imme Ebert-Uphoff and Yi Deng. Causal discovery in the geosciences—using synthetic data to learn how to interpret results. *Computers & Geosciences*, 99:50–60, 2017.

[37] Recruitment Dataset from Centre for Data Ethics and Innovation UK. Recruitement Dataset. https://github.com/CDEIUK/bias-mitigation/tree/master/artifacts/data/recruiting/raw, 2020.

[38] Gartner. https://www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic-data-the-future-of-ai, 2022.

[39] 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(12):86–92, 2021.

[40] Salvatore Greco, Alessio Ishizaka, Menelaos Tasiou, and et al. On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research*, 141:61–94, 2019.

[41] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[42] Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron Adcock. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022.

[43] Ange-Marie Hancock. When multiplication doesn't equal quick addition: Examining intersectionality as a research paradigm. *Perspectives on Politics*, 5(1):63–79, 2007.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[45] 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*, pages 6626–6637, 2017.

[46] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[47] Florimond Houssiau, Samuel N Cohen, Lukasz Szpruch, Owen Daniel, Michaela G Lawrence, Robin Mitra, Henry Wilde, and Callum Mole. A framework for auditable synthetic data generation. *arXiv preprint arXiv:2211.11540*, 2022.

[48] Florimond Houssiau, James Jordon, Samuel N. Cohen, Owen Daniel, Andrew Elliott, James Geddes, Callum Mole, Camila Rangel-Smith, and Lukasz Szpruch. Tapas: a toolbox for adversarial privacy auditing of synthetic data, 2022.

[49] IBM Research. Snippets of Auditing Workflows. https://ibm.box.com/s/v5eykx3xgca1udqcdau09b1x3cmzk9g8, 2023.

[50] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33:22205–22216, 2020.

[51] Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, Vasilis Efthymiou, Jiaoyan Chen, and Kavitha Srinivas. Semtab 2019: Resources to benchmark tabular data to knowledge graph matching systems. In *The Semantic Web - 17th International Conference, ESWC 2020*, volume 12123 of *Lecture Notes in Computer Science*, pages 514–530. Springer, 2020.

[52] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016.

[53] Emre Kazim, Adriano Soares Koshiyama, Airlie Hilliard, and Roseline Polle. Systematizing audit in algorithmic recruitment. *Journal of Intelligence*, 9(3), 2021.

[54] Jonas M Kübler, Wittawat Jitkrittum, Bernhard Schölkopf, and Krikamol Muandet. A witness two-sample test. In *International Conference on Artificial Intelligence and Statistics*, pages 1403–1419. PMLR, 2022.

[55] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, pages 3929–3938, 2019.

[56] Qi Lei, Lingfei Wu, Pin-Yu Chen, Alex Dimakis, Inderjit S Dhillon, and Michael J Witbrock. Discrete adversarial attacks and submodular optimization with applications to text classification. *Proceedings of Machine Learning and Systems*, 1:146–165, 2019.

[57] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022.

[58] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[59] X Liu, B Glocker, MM McCradden, M Ghassemi, AK Denniston, and L. Oakden-Rayner. The medical algorithmic audit. *Lancet Digit Health*, 2022.

[60] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), 09 2022. bbac409.

[61] Jimit Majmudar, Christophe Dupuy, Charith Peris, Sami Smaili, Rahul Gupta, and Richard Zemel. Differentially private decoding in large language models. *arXiv preprint arXiv:2205.13621*, 2022.

[62] Yael Mathov, Eden Levy, Ziv Katzir, Asaf Shabtai, and Yuval Elovici. Not all datasets are born equal: On heterogeneous tabular data and adversarial examples. *Knowledge-Based Systems*, 242:108377, 2022.

[63] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the nist contest: A scalable and general approach to differentially private synthetic data. *Journal of Privacy and Confidentiality*, 11(3), 2021.

[64] Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1):7166, 2022.

[65] 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*, pages 220–229, 2019.

[66] Jakob Mökander, Prathm Juneja, David S. Watson, and Luciano Floridi. The us algorithmic accountability act of 2022 vs. the eu artificial intelligence act: what can they learn from each other? *Minds and Machines*, 32(4):751–758, 2022.

[67] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: a three-layered approach. *arXiv preprint arXiv:2302.08500*, 2023.

[68] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decis. Support Syst.*, 62:22–31, 2014.

[69] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, et al. Adversarial robustness toolbox v1. 0.0. *arXiv preprint arXiv:1807.01069*, 2018.

[70] Inkit Padhi, Yair Schiff, Igor Melnyk, Mattia Rigotti, Youssef Mroueh, Pierre Dognin, Jerret Ross, Ravi Nair, and Erik Altman. Tabular transformers for modeling multivariate time series. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3565–3569. IEEE, 2021.

[71] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, Oct 2016.

[72] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022.

[73] Michael Power. *The Audit Society: Rituals of Verification*. Oxford University Press, 08 1999.

[74] Zhaozhi Qian, Bogdan-Constantin Cebere, and Mihaela van der Schaar. Synthcity: facilitating innovative use cases of synthetic data in different data modalities. *arXiv preprint arXiv:2301.07573*, 2023.

[75] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. FAT* '20, page 33–44, New York, NY, USA, 2020. Association for Computing Machinery.

[76] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[77] Sofya Raskhodnikova, Satchit Sivakumar, Adam Smith, and Marika Swanberg. Differentially private sampling from distributions. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28983–28994. Curran Associates, Inc., 2021.

[78] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[79] Eliane Röösli, Selen Bozkurt, and Tina Hernandez-Boussard. Peeking into a black box, the fairness and generalizability of a mimic-iii benchmarking model. *Scientific Data*, 9(1):24, 2022.

[80] Prasanna Sattigeri, Soumya Ghosh, Inkit Padhi, Pierre Dognin, and Kush R Varshney. Fair infinitesimal jackknife: Mitigating the influence of biased training data points without refitting. In S. Koyejo, S. Mohamed,

A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 35894–35906. Curran Associates, Inc., 2022.

[81] Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness?, 2022.

[82] Stanislau Semeniuta, Aliaksei Severyn, and Sylvain Gelly. On accurate evaluation of gans for language generation, 2019.

[83] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[84] Marilyn Strathern. *Audit Cultures: Anthropological Studies in Accountability Ethics and the Academy*. Routledge, London, 2000.

[85] Maria Ulan, Welf Löwe, Morgan Ericsson, and Anna Wingkvist. Copula-based software metrics aggregation. *Software Quality Journal*, 29(4):863–899, 2021.

[86] Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. Membership inference attacks against synthetic data through overfitting detection, 2023.

[87] Nikhil Vyas, Sham Kakade, and Boaz Barak. Provable copyright protection for generative models. *arXiv preprint arXiv:2302.10870*, 2023.

[88] Linda F Wightman. Lsac national longitidunal bar passage study. In *LSAC Research Report Series*, 1998.

[89] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

[90] Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I Jordan. Greedy attack and gumbel attack: Generating adversarial examples for discrete data. *The Journal of Machine Learning Research*, 21(1):1613–1648, 2020.

[91] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2019.

[92] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.