Synthetic Data Quality Evaluation in Generative AI: Current Trends, Challenges, and Future Directions for Social Science Research

Nurul A. Emran[1*], Ruhaila Maskat[2], Abdulrazzak Ali[3]

[1]Department of Applied Data Engineering, Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Malaysia.

[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Malaysia.

[3]Faculty of Computer and Information Technology, University of Aden, Aden, Yemen.

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.100700601

Received: 26 July 2026; Accepted: 31 July 2026; Published: 08 August 2026

ABSTRACT

Generative Artificial Intelligence (GenAI) has transformed data generation through advanced models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Diffusion Models, and Large Language Models (LLMs), enabling the creation of synthetic datasets that closely resemble real-world data while addressing challenges related to privacy, accessibility, and regulatory compliance. As synthetic data becomes increasingly adopted across healthcare, education, finance, public administration, and social science research, ensuring its quality, reliability, fairness, and trustworthiness has emerged as a critical research priority. This literature review examines recent developments in synthetic data quality evaluation between 2020 and 2026, focusing on key dimensions including utility, fidelity, privacy preservation, fairness, diversity, robustness, interpretability, and governance. The review traces the evolution of evaluation methodologies from traditional statistical similarity measures toward multidimensional assessment frameworks such as SynEval, SynthEval, Benchmarking Synthetic Tabular Data Framework, SynAE, and ESDAE. A structured comparison of these frameworks is presented using criteria including analytical accuracy, scalability, privacy protection, fairness assessment, and interpretability. The review further explores emerging approaches for explainable synthetic data assessment, fairness-aware synthetic data generation, and Privacy-Enhancing Technologies (PETs), including differential privacy, federated learning, secure multi-party computation, and privacy-preserving generative models. In addition, real-world case studies from healthcare, education, public policy, and social science research are examined to demonstrate how synthetic data quality evaluation directly influences decision-making, research validity, and policy outcomes. The findings indicate that while modern generative models can produce highly realistic and analytically useful datasets, persistent challenges remain, including the lack of standardized benchmarking protocols, utility–privacy trade-offs, privacy leakage risks, bias amplification, limited explainability, and governance concerns. The review concludes that future research should prioritize internationally accepted evaluation standards, explainable and fairness-aware assessment frameworks, stronger privacy-preserving mechanisms, and comprehensive governance models to support the responsible, transparent, and trustworthy deployment of synthetic data in the Generative AI era.

Keywords: Synthetic Data, Generative Artificial Intelligence, Data Quality Evaluation, Utility, Fidelity, Privacy Preservation, Fairness, Explainable AI, Privacy-Enhancing Technologies, Social Science Research.

INTRODUCTION

The rapid advancement of Generative Artificial Intelligence (GenAI) has significantly transformed the way data are created, shared, and utilized across multiple domains. Recent developments in machine learning have enabled the generation of synthetic data that closely resembles original datasets while minimizing exposure to sensitive

information [1]. Synthetic data refers to artificially generated data designed to preserve key statistical properties, relationships, and patterns found in real-world data without directly disclosing confidential records.

The increasing importance of synthetic data is largely driven by growing concerns regarding privacy regulations, data governance, and restricted access to sensitive information. Organizations operating in sectors such as healthcare, education, finance, and public administration frequently face challenges associated with sharing real data due to legal, ethical, and regulatory constraints. Synthetic data provides an alternative approach that enables data sharing, experimentation, and innovation while reducing disclosure risks [2,3].

Recent advances in Generative AI have expanded the capabilities of synthetic data generation. Technologies such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Diffusion Models, and Large Language Models (LLMs) have demonstrated remarkable ability to generate realistic tabular, textual, image, and multimodal datasets [4,5]. Furthermore, the emergence of LLMs has facilitated the generation of synthetic survey responses, educational datasets, conversational data, and social behavioral records, creating new opportunities for social science research [6].

Despite these advancements, concerns regarding synthetic data quality continue to receive considerable attention. The quality of synthetic data directly affects the reliability of statistical analyses, machine learning models, policy evaluations, and decision-making processes. Consequently, researchers have increasingly emphasized the need for robust evaluation methodologies capable of assessing whether synthetic data accurately represents real-world phenomena while maintaining privacy, fairness, and usefulness [7,8].

Synthetic Data In The Generative AI Era

Synthetic data generation has evolved significantly over the past decade. Early approaches relied primarily on statistical modeling and simulation techniques. However, modern synthetic data generation increasingly utilizes deep learning architectures capable of learning complex relationships from large datasets [4].

Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Diffusion Models, and Large Language Models have become the dominant technologies supporting modern synthetic data generation [4–6]. According to recent surveys, synthetic data generation methods now encompass structured, unstructured, multimodal, textual, image, and healthcare data applications [4,5].

In social science research, synthetic data presents significant opportunities. Researchers can generate representative datasets for educational analytics, public policy simulations, social behavior modeling, and survey research without exposing confidential participant information. Consequently, synthetic data has become an important component of data democratization initiatives aimed at improving accessibility while maintaining ethical compliance [15,16].

Dimensions Of Synthetic Data Quality

Utility

Utility refers to the ability of synthetic data to support intended analytical and predictive tasks. High-quality synthetic datasets should allow statistical analyses and machine learning models to generate outcomes comparable to those obtained using real-world data [2,3].

Several studies have demonstrated that synthetic datasets retain substantial analytical value. Utility is commonly evaluated using metrics such as classification accuracy, precision, recall, F1-score, Area Under the Curve (AUC), and predictive transferability [2,3,7]. Recent evaluation frameworks emphasize the importance of measuring utility through downstream analytical performance rather than solely relying on statistical similarity assessments [7].

Fidelity

Fidelity refers to the degree of similarity between synthetic data and original datasets. High-fidelity synthetic datasets preserve statistical distributions, feature interactions, and correlations observed in source data [7,8].

Traditional fidelity evaluation employed statistical measures such as:

- Kolmogorov-Smirnov Distance
- Jensen-Shannon Divergence
- Correlation Preservation
- Distribution Matching

Recent frameworks advocate multidimensional fidelity assessment using statistical, embedding-based, and nearest-neighbour approaches [8,14]. Researchers increasingly recognize that fidelity must be evaluated together with utility and privacy rather than as an isolated quality metric [7,8].

Privacy Preservation

Privacy remains a primary motivation for synthetic data adoption. Organizations often deploy synthetic datasets to facilitate data sharing while adhering to privacy regulations and ethical standards [1,11].

However, synthetic data does not automatically guarantee privacy. Recent studies demonstrate that generative models may unintentionally memorize training records and expose sensitive information through disclosure attacks [11–13].

Common privacy evaluation methods include:

- Membership Inference Attacks
- Re-identification Risk Assessment
- Nearest Neighbor Analysis
- Differential Privacy Evaluation
- Disclosure Risk Measurement

Recent literature consistently emphasizes the existence of a privacy-utility trade-off in synthetic data generation [12,13].

Fairness and Bias

Fairness has become a prominent dimension of synthetic data quality assessment. Since generative models learn patterns from historical data, they may reproduce or amplify demographic and societal biases [10,15].

This issue is particularly important in social science research where synthetic datasets may influence decision-making processes, policy development, and resource allocation. Common fairness evaluation indicators include demographic parity, equal opportunity, subgroup representation, and bias amplification measures [10].

Recent governance studies recommend integrating fairness assessments into overall synthetic data quality frameworks rather than evaluating fairness separately [15,16].

Emerging Evaluation Frameworks

The rapid growth of synthetic data applications has accelerated the development of increasingly sophisticated

evaluation frameworks designed to assess the quality, trustworthiness, and suitability of synthetic datasets. Earlier evaluation approaches primarily relied on statistical similarity measures such as distribution matching, correlation preservation, and distance-based metrics. However, recent studies argue that synthetic data quality is inherently multidimensional and cannot be adequately captured through a single metric or evaluation criterion [7,8].

Early Multidimensional Evaluation Approaches

One of the most influential recent frameworks is SynEval, proposed by Yuan et al. [8], which evaluates synthetic data across three core dimensions: utility, fidelity, and privacy. The framework was initially developed to assess synthetic datasets generated by Large Language Models (LLMs) and demonstrated that synthetic data evaluation requires balancing multiple objectives simultaneously rather than optimizing a single performance measure [8]. The authors highlighted that datasets exhibiting high statistical similarity to original data do not necessarily provide high utility for downstream analytical tasks or strong privacy protection [8].

Similarly, Livieris et al. [7] introduced a comprehensive evaluation framework that combines statistical, theoretical, and practical assessment criteria to compare synthetic data generation models. Their framework emphasizes comparative benchmarking and provides mechanisms for ranking generative models according to overall quality performance [7].

Another important contribution is the SynthEval framework developed by Lautrup et al. [17], which provides a modular environment for evaluating synthetic tabular data. The framework integrates multiple privacy and utility metrics within a single platform, allowing researchers to conduct standardized and reproducible assessments across different synthetic datasets. The benchmark module further supports multiaxis comparisons among competing synthetic data generation models [17].

Collectively, these frameworks marked a significant shift from single-dimension evaluation approaches toward more holistic quality assessment methodologies that recognize the complex trade-offs between utility, realism, and privacy.

Advances in 2025 Evaluation Frameworks

Recent work published in 2025 reflects a growing emphasis on holistic and standardized benchmarking approaches.

Sidorenko et al. [14] proposed the Benchmarking Synthetic Tabular Data Framework, a multidimensional evaluation methodology designed to address limitations of previous benchmark systems. Their framework incorporates low-dimensional statistical comparison, high-dimensional distribution assessment, embedding-based similarity measures, and nearest-neighbor privacy analysis [14]. The framework supports various data structures, including structured, sequential, and contextual information, and enables reproducible comparisons across synthetic data generation techniques [14].

A key contribution of this framework is the explicit consideration of both fidelity and novelty, acknowledging that high-quality synthetic data should preserve important characteristics of original datasets without directly replicating or memorizing real records [14].

In the healthcare sector, Hernandez et al. [18] introduced a Comprehensive Evaluation Framework for Synthetic Tabular Data in Health, which integrates fidelity, utility, and privacy into a unified assessment model. Unlike earlier research that evaluated these dimensions independently, their framework explicitly incorporates a fidelity–utility trade-off metric to quantify the balance between data realism and practical usefulness [18].

The framework was validated using multiple healthcare datasets and demonstrated that generative models with stronger privacy guarantees often experience reductions in fidelity and analytical utility [18]. Their study provides practical guidance for organizations seeking to balance privacy requirements with research and operational objectives.

Another important development during 2025 was the growing adoption of benchmark-driven quality assessment methodologies. Researchers increasingly moved away from isolated metric reporting toward standardized evaluation systems that simultaneously assess utility, privacy, fidelity, robustness, and diversity [14,18]. This transition reflects broader efforts to improve reproducibility and methodological consistency within the synthetic data research community.

Emerging 2026 Frameworks

The year 2026 marks a further evolution toward domain-specific and application-oriented synthetic data evaluation frameworks.

One notable development is SynAE (Synthetic Assessment Evaluation), introduced by Wang et al. [19]. SynAE was specifically designed to evaluate synthetic datasets used for testing and benchmarking AI agents. Unlike conventional synthetic data assessment methods focused on tabular data, SynAE evaluates synthetic datasets across four categories:

- Task instructions and interactions
- Tool invocation behavior
- Final outputs
- Downstream evaluation performance

The framework simultaneously assesses validity, fidelity, and diversity, demonstrating that no individual metric sufficiently characterizes synthetic data quality [19]. Instead, SynAE advocates a multi-axis approach that integrates complementary indicators to evaluate complex synthetic datasets.

A related framework, ESDAE (Evaluating Synthetic Data for Agent Evaluation), further extends these ideas by focusing on fidelity and diversity assessment within multi-turn interaction environments [20]. The framework evaluates how effectively synthetic datasets replicate real-world interaction trajectories and demonstrates the importance of assessing dynamic behavioral patterns rather than relying exclusively on static statistical characteristics [20].

Another major advancement is the expanded version of SynEval, presented in 2026 by Yuan et al. [21]. The updated framework extends earlier work by introducing four principal quality dimensions:

- Fidelity
- Utility
- Diversity
- Privacy

Unlike earlier frameworks that focused primarily on utility and similarity metrics, the updated SynEval framework explicitly evaluates diversity and employs adversarial privacy testing mechanisms, including membership inference attacks [21]. The framework was validated across multiple domains, including healthcare, finance, and e-commerce, demonstrating its applicability across diverse synthetic data contexts [21].

One particularly important contribution of the updated SynEval framework is the identification of what the authors describe as the predictability trap, whereby poor-quality synthetic datasets may artificially inflate predictive performance metrics despite exhibiting limited realism and poor generalization characteristics [21]. This finding highlights the limitations of relying solely on predictive accuracy when evaluating synthetic data quality.

From a statistical perspective, Abdel-Azim et al. [22] introduced a complementary perspective on synthetic data evaluation by emphasizing statistical inference validity. Their work argues that most current evaluation frameworks focus heavily on predictive performance while overlooking critical inferential aspects such as uncertainty estimation, hypothesis testing, and causal inference [22].

The authors caution that synthetic datasets may produce apparently accurate predictions while generating misleading statistical conclusions due to model misspecification, attenuated variance structures, or preserved biases [22]. This concern is particularly relevant to social science research, where policy recommendations and theoretical conclusions often depend on rigorous inferential analysis rather than predictive accuracy alone.

Toward Next-Generation Evaluation Frameworks

The evolution of synthetic data evaluation frameworks reveals a clear trend toward multidimensional, standardized, and context-aware assessment methodologies. Contemporary frameworks increasingly evaluate synthetic data across several interconnected dimensions, including utility, fidelity, privacy, diversity, fairness, robustness, and governance [14,19,21].

Recent studies consistently demonstrate that no single metric is capable of adequately representing synthetic data quality and that trade-offs among evaluation criteria are unavoidable [14,18,19,21]. Consequently, future synthetic data evaluation frameworks are expected to move beyond purely technical performance measures and incorporate broader considerations, including:

- Representativeness of social groups
- Fairness across demographic categories
- Statistical inference validity
- Explainability and transparency
- Ethical compliance
- Governance and accountability

For social science applications, such developments will be essential for ensuring that synthetic data remains trustworthy, equitable, and suitable for evidence-based decision-making [16,18,22].

Table 1 illustrates the evolution of synthetic data quality assessment frameworks from narrow statistical evaluations toward multidimensional approaches. Early frameworks primarily emphasized utility and fidelity, whereas more recent frameworks increasingly integrate privacy, diversity, interpretability, and robustness. Despite these advancements, fairness remains insufficiently represented across most contemporary frameworks. This gap is particularly relevant for social science applications where issues relating to demographic bias, equity, and representation directly influence policy and decision-making outcomes [10,15,16].

Table 1: Comparison of synthetic data quality assessment frameworks

Framework	Year	Accuracy/ Utility Assessment	Scalability	Privacy Protection Assessment	Fairness Assessment	Interpretability	Main Strength
Livieris et al. [7]	2024	High	Moderate	Limited	No	Moderate	Model ranking and benchmarking
SynEval [8]	2024	High	High	Strong	Limited	Moderate	Utility–Fidelity–Privacy integration

SynthEval [17]	2024	High	High	Strong	Limited	High	Comprehensive utility and privacy metrics
Hernandez et al. [18]	2025	High	Moderate	Strong	Limited	High	Fidelity–Utility trade-off evaluation
Sidorenko et al. [14]	2025	High	High	Strong	Partial	High	Standardized benchmarking framework
SynAE [19]	2026	High	High	Moderate	No	Moderate	Multi-dimensional evaluation for agent systems
ESDAE [20]	2026	Moderate	High	Moderate	No	Moderate	Fidelity and diversity assessment
SynEval 2.0 [21]	2026	Very High	High	Very Strong	Partial	High	Fidelity, Utility, Diversity, Privacy integration

Explainable Synthetic Data Assessment and Privacy-Enhancing Technologies

The growing complexity of Generative AI models has increased demand for explainable synthetic data quality assessment. Traditional evaluation frameworks frequently function as black-box systems that provide aggregate quality scores without explaining how these scores were derived. Recent research increasingly advocates explainable evaluation mechanisms that enable researchers to identify which variables, patterns, or relationships contribute to quality degradation [5,19,21].

Explainable synthetic data assessment supports transparency, accountability, and trust by helping stakeholders understand why synthetic datasets may perform differently from their real-world counterparts. Emerging explainable approaches utilize feature-attribution methods, interpretable similarity metrics, and visualization techniques that reveal deviations between real and synthetic distributions [14,21].

In parallel, Privacy-Enhancing Technologies (PETs) have become increasingly important for synthetic data generation. Recent studies identify Differential Privacy, Federated Learning, Secure Multi-Party Computation, and Privacy-Preserving Generative Adversarial Networks as leading approaches for protecting sensitive information during synthetic data generation [11–13]. Differential Privacy, in particular, has emerged as a widely adopted mechanism because it provides mathematically quantifiable privacy guarantees while enabling controlled disclosure of information [13].

Although PETs significantly reduce disclosure risks, they may also reduce data fidelity and utility. Consequently, future synthetic data evaluation frameworks should explicitly assess how privacy-preserving mechanisms influence analytical quality and fairness outcomes [18].

Fairness-Aware Synthetic Data Generation

Fairness-aware synthetic data generation has emerged as an important research direction because generative models frequently inherit biases present in source datasets [10,15]. Traditional synthetic data generation approaches primarily optimize utility and fidelity objectives without accounting for demographic fairness. As a result, minority populations and underrepresented groups may remain inadequately represented within generated datasets.

Recent research has introduced fairness-aware GANs, bias-mitigation frameworks, and constrained optimization techniques to improve representation across demographic categories [10]. These methods seek to reduce disparities while maintaining data utility and privacy.

For social science research, fairness-aware generation is particularly important because synthetic datasets may influence policy recommendations, educational interventions, healthcare allocation decisions, and social welfare planning. Future synthetic data quality assessment frameworks should therefore incorporate fairness as a core evaluation dimension alongside utility, fidelity, and privacy [15,16].

Real-World Case Studies of Synthetic Data Quality Evaluation

Healthcare Case Study

Healthcare represents one of the most mature applications of synthetic data. Tucker et al. [2] demonstrated that synthetic health records could achieve predictive performance comparable to real datasets while protecting patient privacy. However, subsequent evaluations revealed that datasets exhibiting high predictive utility did not always preserve clinically relevant correlations, highlighting the importance of balancing utility and fidelity. Hernandez et al. [18] further showed that healthcare organizations face difficult trade-offs between privacy guarantees and analytical utility when deploying synthetic datasets for clinical research.

Education Case Study

In educational analytics, institutions commonly face restrictions on sharing student performance and behavioral data. Synthetic student datasets have been used to evaluate learning analytics models and early intervention systems without exposing personally identifiable information [4,6]. However, quality assessments have revealed that certain synthetic generation approaches fail to adequately represent low-performing or minority student groups, potentially introducing bias into educational decision-making processes [10].

Public Policy Case Study

Government agencies increasingly explore synthetic populations for policy simulation and planning. Synthetic census data and social welfare datasets have been used to evaluate public health interventions, poverty alleviation strategies, and workforce development programs [15,16]. The effectiveness of these applications depends heavily on data fidelity and representativeness. Poor-quality synthetic data may produce misleading policy outcomes, emphasizing the importance of fairness and inferential validity assessments [22].

Social Science Case Study

Recent studies using LLM-generated survey responses have demonstrated both opportunities and limitations for social science research [6,8]. While synthetic surveys can reduce data collection costs and improve research scalability, concerns remain regarding the representativeness, factual accuracy, and cultural validity of generated responses. Consequently, robust evaluation frameworks are critical before synthetic datasets can be used to support empirical social science investigations [6,8,9].

Challenges In Synthetic Data Quality Evaluation

Despite significant advances in Generative Artificial Intelligence and synthetic data generation technologies, the evaluation of synthetic data quality remains a complex and evolving research challenge. While modern generative models can produce highly realistic datasets, determining whether these datasets are sufficiently accurate, trustworthy, fair, and privacy-preserving for practical use is far from straightforward. The literature identifies several interrelated challenges that continue to hinder the widespread adoption of synthetic data across research, industry, and public sector applications [5,7,8].

Lack of Standardized Evaluation Frameworks

One of the most frequently cited challenges is the absence of universally accepted standards for synthetic data quality assessment. Current studies employ diverse evaluation methodologies, datasets, quality metrics, and benchmarking approaches, making it difficult to compare findings across different studies and application domains [5,7,14].

Some researchers focus primarily on utility measures such as prediction accuracy and machine learning performance, whereas others prioritize statistical fidelity, privacy preservation, or fairness considerations [7,8]. Consequently, a synthetic dataset considered high quality under one evaluation framework may perform poorly under another. This lack of consistency limits reproducibility and complicates efforts to establish industry-wide best practices.

Furthermore, existing evaluation frameworks are often domain-specific. Quality measures developed for healthcare datasets may not be suitable for educational, social, or economic datasets due to differences in data structures, analytical objectives, and privacy requirements. The absence of domain-independent assessment methodologies remains a significant limitation in current synthetic data research [5,14].

Balancing Utility and Privacy

A fundamental challenge in synthetic data generation is achieving an appropriate balance between data utility and privacy protection. Synthetic data should maintain sufficient similarity to real datasets to support meaningful statistical analysis and predictive modeling while simultaneously protecting individuals from disclosure risks [11–13].

This challenge is often described as the utility–privacy trade-off. Increasing privacy safeguards typically requires introducing additional randomness or noise into generated datasets, which may reduce data utility and analytical value. Conversely, datasets optimized for high utility may retain characteristics that increase the risk of re-identification or information leakage [12,13].

Recent studies demonstrate that no synthetic dataset can simultaneously maximize privacy and utility without compromise. As a result, organizations must carefully determine acceptable trade-offs based on the intended application, regulatory requirements, and risk tolerance. Developing evaluation frameworks capable of quantifying and balancing these competing objectives remains a major research priority [11–13].

Privacy Leakage and Memorization Risks

Although synthetic data is frequently promoted as a privacy-preserving alternative to real-world data, recent research has shown that synthetic datasets can still leak sensitive information. Deep learning models, particularly large-scale generative models, may unintentionally memorize portions of their training data and reproduce them within generated outputs [11,12].

Privacy leakage can occur through several mechanisms, including:

- Membership inference attacks
- Attribute inference attacks
- Record linkage attacks
- Nearest-neighbor disclosure
- Training data memorization

Such vulnerabilities are particularly concerning in domains involving sensitive personal information, including healthcare, education, finance, and public administration. Researchers have therefore emphasized the need for comprehensive privacy testing that extends beyond simple anonymization measures [11–13].

The emergence of Large Language Models has further intensified privacy concerns because these systems are trained on massive datasets and may inadvertently reproduce sensitive information embedded within training corpora. Consequently, privacy evaluation is becoming an increasingly important component of synthetic data quality assessment.

Evaluating Fidelity Beyond Statistical Similarity

A common misconception in synthetic data research is that statistical similarity alone guarantees data quality. Traditional fidelity evaluation methods frequently rely on measures such as distribution matching, correlation preservation, and distance metrics between real and synthetic datasets. While these metrics provide useful information, they often fail to capture higher-order relationships and contextual dependencies within complex datasets [7,8].

For example, a synthetic dataset may successfully reproduce marginal distributions while failing to preserve causal relationships, temporal dependencies, or behavioral patterns present in the original data. Such deficiencies may compromise the validity of downstream analyses despite apparently strong statistical performance.

Recent evaluation frameworks advocate multidimensional approaches that combine statistical similarity measures with machine learning utility tests, embedding-based comparisons, and domain-specific validation procedures [8,14]. Nevertheless, developing robust methods for evaluating complex social phenomena represented within synthetic datasets remains an unresolved challenge.

Bias Amplification and Fairness Concerns

Bias and fairness issues represent another major challenge in synthetic data generation. Generative models learn directly from historical datasets, which may contain structural inequalities, demographic imbalances, and societal biases. Consequently, synthetic data generation processes may replicate or even reinforce existing patterns of discrimination [10,15].

This problem is particularly relevant in social science research, where synthetic datasets may influence policy development, educational planning, workforce analytics, and public resource distribution. If biases embedded within source data remain unaddressed, synthetic data could unintentionally perpetuate inequitable outcomes.

Several studies have shown that minority populations, marginalized groups, and rare categories are often underrepresented or inaccurately represented within synthetic datasets [10,15]. Such distortions may affect statistical analyses and reduce the generalizability of research findings.

Although fairness metrics such as demographic parity and equal opportunity have gained popularity, consensus regarding appropriate fairness evaluation methods for synthetic data remains limited. Future research must explore methods that simultaneously preserve analytical utility while reducing bias and improving representation.

Challenges Associated with Large Language Model-Generated Synthetic Data

The emergence of Large Language Models has introduced a new generation of synthetic data applications. Researchers increasingly utilize LLMs to generate synthetic surveys, interview responses, educational records, social media content, and textual datasets. While these capabilities offer significant opportunities, they also create unique quality evaluation challenges [6,8,9].

Unlike traditional tabular synthetic data, textual synthetic data is more difficult to evaluate objectively. Existing quality measures often struggle to assess:

- Semantic consistency
- Contextual accuracy
- Linguistic diversity

- Cultural appropriateness
- Factual correctness

Additionally, LLMs are susceptible to hallucinations, whereby generated content appears plausible but contains inaccurate or fabricated information [6,9]. In social science research, such inaccuracies can significantly affect interpretations and conclusions.

The absence of standardized evaluation frameworks for LLM-generated synthetic data represents an emerging gap in the literature and an important area for future investigation.

Limited Explainability and Transparency

Another critical challenge concerns the limited explainability of synthetic data generation models. Many modern generative AI systems operate as complex "black-box" models, making it difficult for researchers to understand how specific outputs are generated [4,5].

Lack of explainability creates several concerns:

- Difficulty identifying model biases
- Limited accountability
- Reduced trust in generated outputs
- Challenges in auditing quality assessments
- Obstacles to regulatory compliance

For social science researchers, understanding how synthetic data is generated is essential for evaluating research validity and methodological rigor. Explainable AI approaches may therefore play an increasingly important role in synthetic data quality evaluation moving forward.

Governance, Ethical, and Regulatory Challenges

Technical performance alone is insufficient to guarantee trustworthy synthetic data. The literature increasingly emphasizes that synthetic data governance must address broader ethical, legal, and societal considerations [16].

Important governance challenges include:

- Transparency in synthetic data generation processes
- Accountability for generated outputs
- Documentation and traceability requirements
- Compliance with privacy regulations
- Ethical use of AI-generated information

As synthetic data becomes more widely adopted across government, healthcare, education, and social science domains, governance frameworks will be essential for ensuring responsible use. The World Economic Forum recently highlighted the need for comprehensive governance mechanisms that support transparency, fairness, accountability, and stakeholder trust throughout the synthetic data lifecycle [16].

Future Research Challenges

Looking ahead, several important research opportunities remain. Future studies should focus on:

- Developing internationally accepted synthetic data quality standards.
- Creating domain-specific quality benchmarks for social science applications.
- Improving fairness-aware synthetic data generation techniques.
- Designing explainable evaluation frameworks for LLM-generated data.
- Strengthening privacy-preserving generation methods.
- Establishing governance models that support ethical and transparent synthetic data use.

Addressing these challenges will be essential for ensuring that synthetic data can serve as a reliable, trustworthy, and ethically responsible resource for future social science research.

Implications For Social Science Research

The emergence of synthetic data generated through Generative Artificial Intelligence (GenAI) is reshaping the research landscape across the social sciences. Traditional social science research often relies on access to sensitive datasets containing demographic, socioeconomic, educational, behavioral, and health-related information. However, ethical concerns, privacy regulations, and institutional restrictions frequently limit researchers' ability to access and share such data. Synthetic data offers a promising alternative by enabling researchers to work with datasets that preserve important statistical characteristics while reducing the risk of exposing personal information [1–3].

One of the most significant implications of synthetic data is its potential to improve research accessibility and collaboration. Social science research increasingly requires interdisciplinary and cross-institutional cooperation, yet data-sharing barriers often hinder collaborative efforts. Synthetic datasets can facilitate data exchange among universities, government agencies, and research organizations without compromising confidentiality. This capability is particularly valuable in studies involving vulnerable populations, educational performance records, public health statistics, and social welfare data, where privacy concerns are especially prominent [1,2].

Synthetic data also has considerable potential to enhance reproducibility and transparency in social science research. Reproducibility is a persistent challenge because many published studies rely on proprietary or restricted datasets that cannot be accessed by external researchers. By generating representative synthetic datasets, researchers can share data that approximate the original information while protecting participants' identities. Such practices may improve research verification, replication studies, and methodological transparency, ultimately strengthening the credibility of social science research [2,3].

In educational research, synthetic data provides opportunities to develop and evaluate learning analytics systems without exposing student records. Educational institutions frequently encounter challenges associated with student privacy regulations and ethical approval requirements. Synthetic student datasets can support investigations into academic performance, student engagement, learning behaviors, and educational interventions while maintaining confidentiality. Moreover, researchers can generate diverse educational scenarios to test predictive models and intervention strategies before implementation in real-world settings [4,6].

For public policy research, synthetic data offers a mechanism for evaluating policy alternatives and simulating social outcomes. Governments increasingly employ data-driven decision-making to address issues such as poverty, unemployment, public health, housing, and social inequality. Synthetic populations can enable researchers to model the potential impacts of policy interventions under different scenarios without relying exclusively on sensitive administrative records. Such capabilities may support evidence-based policymaking and improve the quality of governmental planning and resource allocation [15,16].

Synthetic data may also contribute to addressing data scarcity and representation challenges within social science research. Certain populations, minority groups, and marginalized communities are often underrepresented in traditional datasets due to sampling limitations, accessibility constraints, or historical exclusion. Advanced synthetic data generation techniques may help augment underrepresented populations and create more balanced datasets for analysis. If carefully implemented, synthetic data could support more inclusive research practices and improve the representation of diverse populations in social science studies [10,15].

Nevertheless, the use of synthetic data introduces important methodological and ethical considerations. One major concern is the potential amplification of existing biases. Generative AI models learn from historical data, which may already contain structural inequalities, stereotypes, or discriminatory patterns. As a result, synthetic data may inadvertently reproduce or even strengthen such biases when generating new records. In social science contexts, these biases can influence policy recommendations, resource allocation decisions, and interpretations of social phenomena. Consequently, fairness assessment should be regarded as a fundamental component of synthetic data quality evaluation rather than a supplementary consideration [10,15].

Another important issue relates to data validity and external generalizability. Although synthetic datasets may preserve statistical patterns observed in original data, they do not represent actual observations from real individuals or communities. Researchers must therefore exercise caution when drawing conclusions from synthetic datasets alone. Synthetic data should be viewed as a complementary research resource rather than a complete substitute for real-world data. Validation against original datasets, whenever possible, remains essential to ensure that research findings accurately reflect real social conditions [7,8].

The increasing use of Large Language Models (LLMs) in synthetic data generation introduces additional implications for social science research. LLM-generated synthetic interviews, survey responses, focus group discussions, and textual datasets offer unprecedented opportunities for simulation and exploratory analysis. However, concerns regarding hallucinations, factual inaccuracies, cultural biases, and model-generated stereotypes necessitate careful quality assessment and validation procedures. Researchers must critically examine whether generated content accurately represents the social contexts under investigation and avoid overreliance on AI-generated narratives [6,8,9].

From an ethical and governance perspective, transparency and accountability are essential. Researchers should document data generation processes, model architectures, evaluation procedures, and limitations associated with synthetic datasets. Transparent reporting practices would enhance trustworthiness, facilitate peer review, and support responsible AI adoption within the social sciences. Recent governance frameworks emphasize that technical quality assessment alone is insufficient; synthetic data ecosystems must also address ethical accountability, traceability, fairness, and public trust [16].

Looking forward, synthetic data has the potential to become a transformative resource for social science research. Its ability to improve data accessibility, support collaboration, enhance reproducibility, and facilitate innovative forms of analysis positions it as a valuable tool for future research. However, realizing these benefits requires the development of standardized evaluation metrics, fairness-aware generation methods, robust privacy safeguards, and governance frameworks that ensure responsible data use. Future social science research should therefore focus not only on expanding synthetic data applications but also on establishing rigorous standards that ensure the reliability, inclusiveness, and ethical integrity of generated datasets [5,7,14,16].

CONCLUSION

The rapid advancement of Generative Artificial Intelligence (GenAI) has transformed synthetic data from a niche research topic into a strategic resource for data-driven innovation across healthcare, education, finance, government, and social science research. Recent developments in Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Diffusion Models, and Large Language Models (LLMs) have significantly improved the realism, scalability, and accessibility of synthetic data generation. As a result, synthetic data is increasingly viewed as a potential solution for addressing data scarcity, privacy restrictions, and barriers to data sharing while supporting collaborative and reproducible research [1,4–6]. This review demonstrates that synthetic data quality evaluation has evolved from traditional statistical similarity assessments toward

multidimensional evaluation frameworks encompassing utility, fidelity, privacy, fairness, diversity, robustness, and governance considerations. Earlier evaluation approaches focused primarily on distributional similarity and machine learning performance, whereas recent frameworks such as SynEval, SynthEval, the Benchmarking Synthetic Tabular Data Framework, SynAE, and ESDAE emphasize holistic assessment across multiple quality dimensions [8,14,17–21]. These developments reflect growing recognition that no single metric can adequately characterize synthetic data quality and that meaningful evaluation requires balancing multiple, often competing, objectives.

The literature consistently highlights the complex trade-offs that exist between utility, privacy, and realism. While highly realistic synthetic datasets may improve analytical performance, they may also increase disclosure risks. Conversely, stronger privacy protections can reduce analytical utility and limit practical applicability [11–13,18]. Similarly, fairness and representativeness have emerged as increasingly important evaluation dimensions, particularly in social science contexts where synthetic data may influence policy decisions, educational planning, and public resource allocation [10,15,16]. These findings suggest that future evaluation frameworks must move beyond purely technical performance measures and incorporate broader social, ethical, and governance considerations.

The emergence of Large Language Model-generated synthetic data introduces additional challenges and opportunities. LLMs enable the generation of synthetic surveys, interviews, educational records, and behavioral datasets that have substantial value for social science research. However, concerns related to hallucinations, factual inconsistencies, hidden biases, memorization risks, and inferential validity highlight the need for more sophisticated evaluation methodologies [6,8,9,22]. The growing emphasis on diversity metrics, adversarial privacy testing, explainability, and statistical inference validity represents an important step toward addressing these challenges.

For social science research, synthetic data offers significant potential to enhance data accessibility, facilitate interdisciplinary collaboration, improve research reproducibility, and support evidence-based policymaking. Nevertheless, the successful deployment of synthetic data within social science domains depends on rigorous quality evaluation, transparent reporting practices, and robust governance frameworks that ensure ethical and responsible use [15,16,22]. Researchers must therefore critically assess not only whether synthetic data resembles real-world data but also whether it accurately represents social phenomena, preserves fairness across populations, and supports valid scientific inference.

Looking forward, several research priorities remain. First, there is an urgent need for internationally accepted standards and benchmarking protocols for synthetic data quality evaluation. Second, future frameworks should incorporate fairness-aware and explainable evaluation mechanisms capable of identifying hidden biases and ensuring equitable representation. Third, additional research is required to evaluate the suitability of synthetic data for causal inference, policy modeling, and other forms of social science analysis. Finally, governance frameworks that integrate transparency, accountability, privacy protection, and ethical oversight will be essential for fostering trust in synthetic data ecosystems.

In conclusion, synthetic data has the potential to become a transformative resource for social science research and data-driven decision-making. However, its long-term success will depend not only on improvements in generative modeling technologies but also on the development of comprehensive evaluation frameworks capable of ensuring utility, fidelity, privacy, fairness, and trustworthiness. Achieving these goals will be critical for realizing the full potential of synthetic data in supporting responsible innovation and advancing evidence-based research in the Generative AI era.

REFERENCES

1. El Emam, K., Mosquera, L., & Hoptroff, R. (2020). Practical synthetic data generation: Balancing privacy and the broad availability of data. O'Reilly Media.
2. Tucker, A., Wang, Z., Rotalinti, Y., & Myles, P. (2020). Generating high-fidelity synthetic health data for assessing machine learning healthcare software. npj Digital Medicine, 3(1), 147. https://doi.org/10.1038/s41746-020-00353-9

3. Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. BMC Medical Research Methodology, 20(1), 108. https://doi.org/10.1186/s12874-020-00977-1
4. Goyal, M., & Mahmoud, Q. H. (2024). A systematic review of synthetic data generation techniques using generative AI. Electronics, 13(17), 3509. https://doi.org/10.3390/electronics13173509
5. Bauer, A., Trapp, S., Stenger, M., Leppich, R., Kounev, S., Leznik, M., Chard, K., & Foster, I. (2024). Comprehensive exploration of synthetic data generation: A survey.
6. Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., & Wang, H. (2024). On LLMs-driven synthetic data generation, curation, and evaluation: A survey. Findings of the Association for Computational Linguistics: ACL 2024, 11065–11082. https://doi.org/10.18653/v1/2024.findings-acl.658
7. Livieris, I. E., Alimpertis, N., Domalis, G., & Tsakalidis, D. (2024). An evaluation framework for synthetic data generation models. In Artificial Intelligence Applications and Innovations (pp. 287–299). Springer. https://doi.org/10.1007/978-3-031-63219-8_24
8. Yuan, Y., Liu, Y., & Cheng, L. (2024). A multi-faceted evaluation framework for assessing synthetic data generated by large language models. arXiv. https://arxiv.org/abs/2404.14445
9. Iskander, S., Tolmach, S., Shapira, O., Cohen, N., & Karnin, Z. (2024). Quality matters: Evaluating synthetic data for tool-using LLMs. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 4958–4976. https://doi.org/10.18653/v1/2024.emnlp-main.285
10. Bellamy, R. K. E., Dey, K., Hind, M., et al. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4:1–4:15.
11. Stadler, T., Oprisanu, B., & Troncoso, C. (2022). Synthetic data: Anonymisation ground truth. Proceedings on Privacy Enhancing Technologies, 2022(1), 499–527.
12. Steier, A., Ramaswamy, L., Manoel, A., & Haushalter, A. (2025). Synthetic data privacy metrics. arXiv. https://arxiv.org/abs/2501.03941
13. Schlegel, V., Bharath, A. A., Zhao, Z., & Yee, K. (2025). Generating synthetic data with formal privacy guarantees: State of the art and the road ahead. arXiv. https://arxiv.org/abs/2503.20846
14. Sidorenko, A., Platzer, M., Scriminaci, M., & Tiwald, P. (2025). Benchmarking synthetic tabular data: A multi-dimensional evaluation framework. arXiv. https://arxiv.org/abs/2504.01908
15. Mönchmeyer, R., et al. (2024). Synthetic data in biomedicine via generative artificial intelligence. Nature Reviews Bioengineering, 2(10), 713–731. https://doi.org/10.1038/s44222-024-00245-7
16. World Economic Forum. (2025). Synthetic data: The new data frontier. World Economic Forum
17. Lautrup, A. D., Hyrup, T., Zimek, A., & Schneider-Kamp, P. (2024). SynthEval: A framework for detailed utility and privacy evaluation of tabular synthetic data. Data Mining and Knowledge Discovery, 39(1). https://doi.org/10.1007/s10618-024-01081-4
18. Hernandez, M., Osorio-Marulanda, P. A., Catalina, M., Loinaz, L., Epelde, G., & Aginako, N. (2025). Comprehensive evaluation framework for synthetic tabular data in health: Fidelity, utility and privacy analysis of generative models with and without privacy guarantees. Frontiers in Digital Health, 7, 1576290. https://doi.org/10.3389/fdgth.2025.1576290
19. Wang, S., Maddi, A., Lin, Z., & Fanti, G. (2026). SynAE: A framework for measuring the quality of synthetic data for tool-calling agent evaluations. arXiv:2605.22564.
20. Wang, S., Maddi, A., Lin, Z., & Fanti, G. (2026). ESDAE: Evaluating synthetic data for agent evaluation. ICLR 2026 Workshop on Data-Centric Foundation Models.
21. Yuan, Y., Cheng, L., Atwal, T., Shi, Z., Tieu, C. N., & Liu, Y. (2026). SynEval: A multidimensional framework for evaluating synthetic data on fidelity, utility, diversity, and privacy. IEEE Symposium on Security and Privacy Poster Proceedings.
22. Abdel-Azim, A., Wang, R., & Lin, X. (2026). Harnessing synthetic data from generative AI for statistical inference. Statistical Science (submitted manuscript), arXiv:2603.05396.