


# A Systematic Review of Synthetic Data Generation Techniques Using Generative AI

Mandeep Goyal \* and Qusay H. Mahmoud 

Department of Electrical, Computer and Software Engineering, Ontario Tech University,  
Oshawa, ON L1G 0C5, Canada; qusay.mahmoud@ontariotechu.ca

\* Correspondence: mandeep.goyal@ontariotechu.net

**Abstract:** Synthetic data are increasingly being recognized for their potential to address serious real-world challenges in various domains. They provide innovative solutions to combat the data scarcity, privacy concerns, and algorithmic biases commonly used in machine learning applications. Synthetic data preserve all underlying patterns and behaviors of the original dataset while altering the actual content. The methods proposed in the literature to generate synthetic data vary from large language models (LLMs), which are pre-trained on gigantic datasets, to generative adversarial networks (GANs) and variational autoencoders (VAEs). This study provides a systematic review of the various techniques proposed in the literature that can be used to generate synthetic data to identify their limitations and suggest potential future research areas. The findings indicate that while these technologies generate synthetic data of specific data types, they still have some drawbacks, such as computational requirements, training stability, and privacy-preserving measures which limit their real-world usability. Addressing these issues will facilitate the broader adoption of synthetic data generation techniques across various disciplines, thereby advancing machine learning and data-driven solutions.

**Keywords:** synthetic data; LLMs; GANs; VAEs; generative AI; neural networks; machine learning



**Citation:** Goyal, M.; Mahmoud, Q.H. A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. *Electronics* **2024**, *13*, 3509. <https://doi.org/10.3390/electronics13173509>

Academic Editor: Chunping Li

Received: 28 July 2024

Revised: 21 August 2024

Accepted: 29 August 2024

Published: 4 September 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The availability of high-quality data is critical to the success of several applications across a wide range of fields, from healthcare and finance to cybersecurity, primarily in artificial intelligence (AI). Large datasets are required because of the basic idea that insights from strong datasets may alter decision-making, drive innovation, and reveal previously unnoticed patterns and connections. For example, in the healthcare industry, advances in customized medicine, disease prediction, and treatment evaluation are powered by the analysis of large patient records and genomic data. Analyzing large datasets of financial transactions also helps with risk assessment, fraud detection, and algorithmic trading strategies in the finance industry. Furthermore, the analysis of enormous amounts of network traffic data in cybersecurity makes it possible to identify unusual activities and create a proactive defense against online attacks. Moreover, the analysis of extensive urban sensor data in urban planning can guide the development of infrastructure, traffic control, and environmental sustainability projects. The reliance on big data highlights their essential role in promoting advancement, innovation, and well-informed decision-making in the present day across many fields and beyond.

However, big data have several drawbacks. Organizations often face difficulties in identifying suitable data sources that can offer valuable insights into their particular uses. This involves being aware of the various types of data available, their sources (both internal and external), and how relevant they are to the business objectives. To access third-party data sources, it may be necessary to negotiate licensing terms, data agreements, and compliance with regulations and laws. This procedure may involve complex legal

issues and requires an extended period. Integrating data from multiple locations requires sustaining compatibility and consistency in the data formats, structures, and semantics. Data cleaning, standardization, and transformation may be required to align the different datasets for analysis. Moreover, big data usually contains various errors, outliers, and missing values [1]. Eliminating these faults and obtaining clean and seamless data may be challenging and time consuming. Along with all the concerns that businesses encounter, the public faces major risks. For instance, it has become difficult for individuals to maintain control of their personal information because large IT companies collect, store, and process user data to draw inferences from it [2]. Consequently, the demand for differential privacy has increased in recent years. By adding noise to query results or changing data in a controlled manner, differential privacy ensures that the presence or absence of a specific individual's data has no significant effect on the outcome of the data analysis. This approach enables companies to gain important insights from sensitive datasets while reducing the risk of re-identification or unauthorized disclosure of personal information. In an era of growing data collection and analytics, establishing differential privacy measures is critical for maintaining trust and transparency as well as protecting individual privacy rights.

These challenges have raised the need for solutions that can address such problems, leading to synthetic data as a solution. Synthetic data enables businesses to produce realistic but fictional datasets that retain the actual fundamental structure and patterns while revealing no sensitive information. This allows stakeholders to perform analyses, build algorithms, and test applications while maintaining individual privacy and confidentiality. Another issue that synthetic data address is data augmentation. When academics and companies attempt to develop an application for an extremely specific subject, they frequently find it difficult to gather a complete dataset. Synthetic data generation techniques can generate new instances of data with unique attributes or circumstances that are not seen in the original dataset. Organizations can increase the adaptability, generalization, and accuracy of machine learning models and analytical algorithms by augmenting actual data with synthetic samples, resulting in better performance across a broader range of situations and scenarios. Synthetic data exhibit promising properties for boosting the performance of deep neural networks in real-world instance segmentation [3]. Synthetic data facilitates data sharing and collaboration by offering a privacy-preserving alternative to real datasets [4]. Organizations that generate synthetic versions of their data can work more freely with external stakeholders and exchange ideas without compromising on sensitive information or intellectual property. Synthetic data promote increased transparency and openness in data-driven research and cooperation activities, while reducing the dangers involved in sharing real data.

There are two fundamental categories for generating synthetic data: statistical methods and deep-learning models. Statistical distribution techniques for synthetic data generation operate by modeling the underlying data distribution observed in real data and then creating artificial data samples that follow the distribution pattern. These algorithms generally involve two major steps, namely parameter estimation and sampling. Parameter estimation involves determining the statistical parameters (such as the mean, variance, and covariance) of the observed data distribution using techniques such as the maximum likelihood estimation (MLE) or kernel density estimation (KDE). Once the parameters are predicted, synthetic data samples are generated from the fitted distribution by using random number generators or inverse transform sampling techniques. This ensures that the synthetic data have statistical qualities similar to the original data, such as the probability density function (PDF), correlation structure, and summary statistics. Synthetic data generated using statistical approaches appear quite similar to genuine data when compared using first-order distributions [5].

Among the deep learning models, variational autoencoders (VAEs) were introduced in 2013 [6]. They work by learning to compress the input data into a smaller representation, known as a latent space, which is then used to reconstruct the original data. VAEs differ from standard autoencoders in that they learn the underlying structure of the data distribu-

tion and to memorize the data. This means that they can create new data that are similar to those ones they were trained on, making them ideal for tasks such as image generation, anomaly detection, and data compression. VAEs have become essential tools in a range of domains, including computer vision and natural language processing, because of their capacity to learn and generate complex data patterns. A complementary approach is the use of generative adversarial networks (GANs), which are the most popular models for generating synthetic data. In 2014, ref. [7] introduced the concept of GANs which consist of two neural networks: a generator and a discriminator. They engage in a competitive learning process. The generator generates synthetic data samples such as photographs, and the discriminator assesses them to distinguish between actual and fake instances. GANs can produce highly realistic outputs owing to adversarial training. This makes them useful for picture generation, data augmentation, and generation of completely new content, such as art and music. This innovative architecture has attracted significant research interest. GANs constantly push the limits of what is possible in machine-generated content. However, a team of researchers at OpenAI introduced a framework in 2018 [8] to improve language understanding through generative pre-training. This framework consists of a sophisticated network of neurons, known as large language models. LLMs can generate synthetic data by using their ability to interpret and synthesize human-like text from massive volumes of training data. By entering precise prompts or instructions, these models can generate a wide range of realistic text data, such as dialogues, stories, and domain-specific information. These synthetic data can be used for various purposes, including training machine-learning models, enhancing datasets, and replicating scenarios for testing and development.

This study aims to conduct a systematic literature review of various synthetic data generation techniques that leverage generative AI. Through a comprehensive analysis of the available and relevant frameworks, this study identifies the limitations of the current generative AI methods for synthetic data generation. Section 2 examines related work available in the form of reviews and surveys for synthetic data generation and justifies the need for this systematic literature review. Section 3 details the methodology employed in conducting this systematic literature review, including the inclusion and exclusion criteria, and the selection process for the studies reviewed. Section 4 provides an overview of the technologies used in synthetic data generation, encompassing both machine learning and programmatic approaches. Section 5 discusses the findings of this literature review and answers the research questions. Section 6 presents the most critical challenges that occur during synthetic data generation using generative AI and proposes future research directions. Finally, Section 7 presents conclusions drawn from the literature review.

## 2. Related Work

The introduction outlines the importance of our research topic and the objectives of this review. In the related work section, we provide a comprehensive analysis of the existing literature, comparing various studies to highlight the current state of research. This analysis underscores the necessity of our literature review by identifying the gaps and limitations in previous studies and demonstrating how our work aims to address these shortcomings. We examined the studies mentioned in Table 1 to provide an overview of the existing literature in this field and to justify the need for this systematic review.

**Table 1.** Reviews/surveys related to synthetic data generation.

Primary Study	Year	Overview	Limitations
Long, L. [9]	2024	This study surveys the potential of large language models (LLMs) for data generation, curation, and evaluation, highlighting their strengths and limitations across various tasks.	The survey lacks specificity in addressing fine-tuning LLMs for domain-specific applications.

Table 1. Cont.

Primary Study	Year	Overview	Limitations
Bauer, A. [10]	2024	This paper presents a comprehensive survey on synthetic data generation, discussing various techniques and their applications, with a focus on machine learning models and privacy concerns.	The review may overlook recent advancements due to its reliance on the earlier literature.
Hao, S. [11]	2024	The study discusses the challenges, applications, and ethical implications of synthetic data in AI, particularly in healthcare and finance, with a focus on maintaining privacy and fairness.	The review extensively covers challenges but offers limited practical solutions for ethical concerns.
Sengar, S.S. [12]	2024	This systematic review explores the use of generative AI in various applications, emphasizing its potential in synthetic data generation and the associated ethical challenges.	The review highlights ethical challenges but lacks in-depth strategies for addressing them in generative AI.
Sufi, F.K. [13]	2024	This paper reviews the use of generative pre-trained transformers (GPTs) for data augmentation in research, highlighting the benefits and limitations of using GPT models in various academic contexts.	The paper primarily focuses on academic contexts, potentially missing broader practical applications of GPTs for data augmentation.
Guo, X. [14]	2024	The study provides an overview of the methods and challenges in using generative AI for synthetic data generation, focusing on the potential future directions of this technology.	The survey provides a broad overview but may oversimplify the challenges of synthetic data generation in complex real-world scenarios.
Lu, Y. [15]	2023	This systematic review discusses the role of generative adversarial networks (GANs) in synthetic data generation for agriculture, highlighting their applications in image augmentation and crop monitoring.	The review is comprehensive but could benefit from a more detailed examination of the scalability of GANs in agricultural applications.
Bandi, A. [16]	2023	The paper reviews the requirements, models, input–output formats, evaluation metrics, and challenges in leveraging generative AI models for synthetic data generation.	The review provides a broad perspective but lacks detailed examples of successful implementations in various domains.
Ippolito, D. [17]	2023	This survey focuses on the bias and fairness in large language models, providing insights into how these models can be audited and adjusted to minimize ethical risks.	The survey focuses on bias but offers limited actionable recommendations for mitigating it in LLMs.
Eigenschink, P. [18]	2023	This paper surveys deep generative models for synthetic data, exploring their effectiveness across various fields, including healthcare, finance, and autonomous systems.	The review is extensive but could better address the practical limitations of deep generative models in specific industries.
Fonseca, J. [19]	2023	The study provides a comprehensive review of methods for generating synthetic tabular and latent space data, discussing their applications and limitations in various domains.	The review is thorough but does not sufficiently explore the practical challenges of implementing synthetic data generation in industrial applications.
Vargas, A. M. [20]	2022	This paper reviews the use of synthetic data in medical imaging, focusing on the advancements in GANs and VAEs and their impact on data augmentation for diagnostic tasks.	The review focuses on GANs and VAEs but lacks discussion on the integration of synthetic data with existing medical imaging workflows.
Lu, Y. [21]	2022	This systematic review discusses the role of GANs in agriculture, particularly in the generation of synthetic data for improving crop monitoring and disease detection.	The review is well rounded but lacks insights into the real-world challenges of deploying GANs in agricultural settings.

Table 1. *Cont.*

Primary Study	Year	Overview	Limitations
Wang, S. [22]	2022	The paper provides a review of controllable data generation using deep learning, with a focus on methods and challenges in generating synthetic data for various applications.	The paper broadly covers controllable data generation but does not delve deeply into the technical challenges of achieving control in diverse data types.
Figueira, Á. [23]	2022	This survey reviews methods and evaluation techniques for synthetic data generation, focusing on the use of GANs and their effectiveness in generating high-quality data.	The survey is detailed but could be improved by offering more practical evaluation methods for synthetic data quality in applied settings.

A comprehensive study conducted on the application of machine learning for synthetic data generation highlighted the pressing data-related challenges in real-world applications, such as poor data quality, insufficient data points, and difficulties in data access owing to privacy and regulatory concerns [15]. This systematic review covers the extensive use of synthetic data across various domains, including computer vision, speech, natural language processing, healthcare, and business, and underscores the versatility of synthetic data in addressing data scarcity and governance issues. However, this study has some limitations. First, it does not provide detailed benchmarks or empirical comparisons of the performances of different synthetic data generation methods, making it difficult to assess the relative effectiveness of the approaches discussed. Second, the review is heavily focused on deep learning methods, potentially overlooking simpler and more interpretable models that might be more suitable for certain applications.

Similarly, a review paper explored synthetic data generation (SDG) for tabular health records, highlighting its potential and limitations [20]. The authors analyzed various methods used to create synthetic health data, including GANs and VAEs. Although this review offers valuable insights, its universality is restricted by several limitations. The rapidly evolving field of SDG might have emerging techniques beyond the scope of this review. Additionally, the findings and recommendations are overly specific to certain health record types or settings, limiting their broader applicability in healthcare.

A recent study reviewed synthetic data generation using LLMs, addressing the long-standing issue of data quantity and quality in deep learning [9]. The study emphasizes the potential of LLMs to alleviate real-world data limitations through synthetic data generation, while also highlighting the need for a more unified framework in this field. One significant limitation of this study is its primary focus on text data and LLM-driven approaches, which leaves out investigations into other data modalities such as vision and speech. This narrow scope may limit the applicability of these findings to different fields that require synthetic data generation.

One study explored the role of GANs in agricultural image augmentation by focusing on various vision tasks such as plant health, weed detection, fruit phenotyping, and postharvest inspection [21]. This study explores various GAN architectures and challenges and suggests future research opportunities. However, the focus on GAN-based augmentation techniques may overlook other effective non-GAN methods that may be relevant in agricultural contexts. In addition, while this paper identifies challenges and future research needs, it provides limited practical solutions or strategies to address these challenges, particularly regarding the scalability and real-world implementation of GANs in agriculture.

Other comprehensive reviews highlight the evolution from simpler probabilistic models such as Markov chains and Bayesian networks to complex neural network-based approaches, with GANs and diffusion models leading to computer vision applications [10,16]. However, the first paper [10] faced limitations such as the scarcity of common evaluation metrics and datasets, making comparisons challenging. In addition, it neglects training and computational costs, which are crucial considerations for future research. The second paper [16], while thoroughly discussing evaluation metrics, did not address the interpretability of these metrics, which is crucial for validating models in real-world scenarios.



Moreover, a review paper explores the creation and utilization of synthetic datasets in AI, addresses traditional statistical methods and advanced deep learning techniques for data generation, and highlights their applications across various domains [11]. Another study provides a comprehensive overview of advancements in generative AI, focusing on models such as GANs, transformers, VAEs, and diffusion models, and key applications in fields such as image translation, medical diagnostics, and natural language processing [12]. However, both studies had limitations. The first lacks an in-depth discussion on the scalability of synthetic data generation techniques across different data types and volumes and does not thoroughly examine the long-term impacts and sustainability of using synthetic data. The second paper has a restricted analysis timeframe from 2012 to 2023, potentially overlooking earlier foundational work. Its reliance on peer-reviewed sources may exclude innovative industry applications not yet published in academic journals. Both reviews could benefit from a more comprehensive approach for evaluating biases and ensuring fairness in synthetic datasets.

Two studies provided comprehensive reviews of deep learning techniques for synthetic data generation. One reviews various deep generative models used across domains such as health records, NLP, and financial time series, introducing an evaluation framework based on representativeness, novelty, realism, diversity, and coherence [18]. However, its broad scope may oversimplify domain-specific challenges, lacks practical implementation guidelines, and do not empirically validate its evaluation framework, thereby limiting real-world applicability. Another study examine methods for generating data with targeted properties, covering applications such as molecular design, image editing, and speech synthesis [22]. It highlights the transition from traditional methods to efficient deep learning approaches but points out limitations such as the complexity of searching large data spaces, challenges in multi-property control due to property correlations, and inefficiency in representing complex data structures. These limitations emphasize the need for advanced and scalable techniques to enhance the efficacy and accuracy of controlled data generation.

Notably, another paper provides an extensive overview of the applications of GPTs and LLMs in enhancing research data, generating features, and synthesizing data [13]. It meticulously categorizes and evaluates 77 scholarly contributions, focusing on data augmentation, critical analysis, and research design. However, the study's limitations include a heavy reliance on the existing literature without empirical validation, potential biases in selecting and interpreting the reviewed works, and a lack of practical guidelines for implementing GPT-based data augmentation in diverse research contexts. These constraints could impact the generalizability and applicability of its findings.

Two studies examined synthetic data generation methods, focusing on different aspects. The first study analyzed algorithms for tabular and latent space data, proposing a unified taxonomy and reviewing 70 algorithms across six ML problems [19]. It discusses metrics for evaluating synthetic data quality but faces limitations, including a lack of focus on real-world application challenges, potential biases in the selected algorithms, and insufficient analysis of scalability in large industrial datasets, highlighting the need for further research to validate practical effectiveness. The second study focused on GANs and their applications, providing an introduction, an overview of synthetic data generation methods, an examination of GAN architectures, and an exploration of evaluation methods [23]. However, it predominantly focuses on GAN-based techniques, potentially overlooking other innovative methods, lacking empirical benchmarks and direct performance comparisons of GAN models, and providing limited practical guidance on implementing evaluation techniques, which could hinder real-world application.

Furthermore, a study reviewed recent advancements in generating synthetic data using large language models (LLMs) [14]. It discusses methodologies, evaluation techniques, and applications, with a focus on addressing data scarcity and privacy issues in various domains. This paper is structured into sections covering an introduction, methods for generating synthetic data, application scenarios, challenges, and future research directions. However, the scope is broad, potentially leading to oversimplification of domain-specific

challenges. In addition, the discussion on practical deployment scenarios is limited, providing insufficient guidance for real-world implementations.

All the above limitations collectively highlight the necessity for a more comprehensive literature review that not only synthesizes existing research but also addresses the gaps and inconsistencies in current knowledge. Unlike previous studies that focused narrowly on deep learning methods or specific applications, this literature review thoroughly covers all the major technologies and models used for synthetic data generation, including generative adversarial networks (GANs), variational autoencoders (VAEs), and large language models (LLMs) such as GPT-3 and BERT. We compare these models across various dimensions such as performance, scalability, data diversity, and practical implementation. In doing so, this review aims to provide a nuanced understanding of each model's strengths and weaknesses, facilitating more informed decisions for future research and application in synthetic data generation. Additionally, we will explore domain-specific adaptations and conditional models tailored for specific use cases and assess their effectiveness and limitations.

### 3. Research Methodology

To conduct and report this review, we followed the guidelines for systematic literature reviews [24], systematic mapping studies [25], and PRISMA statement guidelines [26]. In our study, this approach helped us identify and map the various techniques used in synthetic data generation with generative AI. This allowed us to understand the extent to which generative AI applications have been developed and applied to synthetic data generation. In addition, a mapping study helped identify potential research gaps. This systematic review further enabled us to explore the current trends in the technical approaches, methodologies, and frameworks employed in the development of generative AI for synthetic data generation.

#### 3.1. Research Questions

At the beginning of the systematic literature review (SLR), the following research questions will be answered in this paper to help us better understand the workings of generative AI for synthetic data generation:

- RQ1. What are the techniques and methods used for synthetic data generation?  
Answering this question will help to identify and categorize the current methods and techniques employed to generate synthetic data.
- RQ2. What frameworks are available for synthetic data generation for specific use cases?  
This question focuses on identifying and evaluating frameworks designed for synthetic data generation tailored to specific applications.
- RQ3. What are the limitations of using these techniques?  
Answering this question will help to explore and understand the challenges and drawbacks associated with existing synthetic data generation methods.
- RQ4. What are the areas of future research?

This question highlights gaps and opportunities for future research on synthetic data generation using generative AI.

#### 3.2. Information Source

While conducting this systematic literature review of synthetic data generation using generative AI, we selected primary papers by searching several scientific databases using specific search prompts. Six prominent scientific databases were included in the search: MDPI, IEEE Xplore, Science Direct, Research Gate, NeurIPS, and Arxiv. These databases were selected to ensure the inclusion of peer-reviewed articles published in reputable journals, conferences, workshops, and symposiums.

To search the databases, we used the following search strings: “synthetic data generation using AI”, “GANs for synthetic data generation”, “LLMs for synthetic data generation”,

and “different types of VAEs”. Additionally, we crafted composite search strings to broaden our search, such as “synthetic data” AND (“generative AI” OR “GANs” OR “VAEs” OR “LLMs”), and others, such as “data generation” AND (“generative models” OR “neural networks” OR “AI” OR “deep learning”).

This search string selection was based on pilot searches, in which we tested various common terms and acronyms related to synthetic data and generative AI. These pilot searches ensured that our search strings comprehensively captured the relevant literature without missing any key papers that might have used alternative terminologies. This approach allowed us to systematically identify and review studies pertinent to synthetic data generation using generative AI, thereby providing a robust foundation for our literature review.

### 3.3. Eligibility Criteria

For our systematic literature review, we established a clear set of inclusion and exclusion criteria to ensure credibility and relevance of the studies.

#### 3.3.1. Inclusion Criteria

**Publication Type.** Only peer-reviewed journal articles, conference papers, and book chapters are included. This ensured that the sources had undergone strict academic inspection. Some Web articles were included to better understand these concepts.

**Language.** Only studies published in English were included to maintain consistency of interpretation and ease of understanding.

**Relevance.** Studies that focused on synthetic data generation using generative AI, including techniques such as GANs, VAEs, LLMs, and other relevant models, are included.

**Time Frame.** The concept of generative AI was introduced in 2014; therefore, all studies added to this literature review were from 2014 and later.

**Availability.** Full-text availability was required to allow for a thorough review of the methodology, results, and conclusions of each study.

#### 3.3.2. Exclusion Criteria

**Irrelevance.** Studies that did not primarily focus on synthetic data generation using generative AI were excluded.

**Retracted Papers.** Retracted papers were excluded to ensure the reliability of our review.

### 3.4. Data Collection Procedure

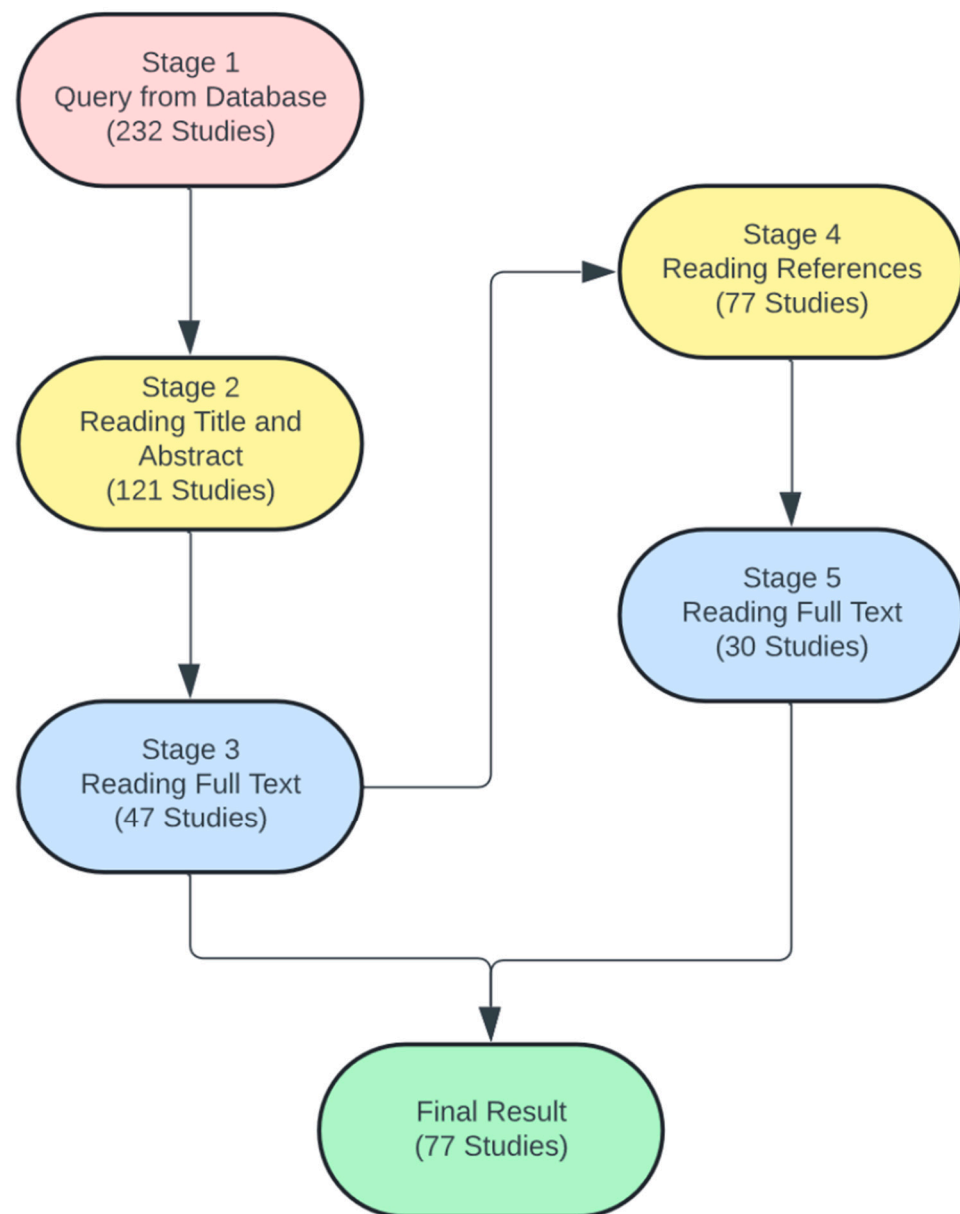
The selection process was conducted in May 2024. Figure 1 shows the process of paper inclusion and exclusion.

Our research collection process began by querying the databases, which resulted in 232 studies. The titles and abstracts were then read and narrowed to 121 studies. A thorough review of the full text of these studies reduced the sample size to 47. By examining the references of the selected studies, we identified 77 more studies that seemed to be highly related to this review. After reading their full texts, we included an additional 30 studies in our final list. This rigorous process led to the final selection of 77 studies for a detailed analysis.

These core studies represent the most technical and essential research. Additional supporting studies, although not crucial, were also reviewed and included to provide a comprehensive understanding of the topic.

Data extraction was conducted meticulously by a single reviewer to maintain consistency and ensure thoroughness during the extraction process. Only studies with complete texts were included in the review to ensure the availability of all necessary data. The data extraction process was manual, with careful attention paid to detail to minimize errors and biases. The reviewer meticulously read each study, extracting data directly from the text, tables, and figures without the aid of external software or automation tools.





**Figure 1.** Collection procedure.

The primary outcomes were the performance, effectiveness, and limitations of the various generative AI models used for synthetic data generation. Performance metrics such as accuracy, precision, recall, F1-score, and other relevant metrics were used to evaluate generative models. In terms of effectiveness, this was the ability of the models to generate realistic and useful synthetic data across different data types, such as images and tabular and textual data. Limitations include the challenges faced by the models, including issues related to scalability, computational demands, training stability, and privacy concerns. The risk of bias in the included studies was evaluated through a manual review process, without the use of specific assessment tools.

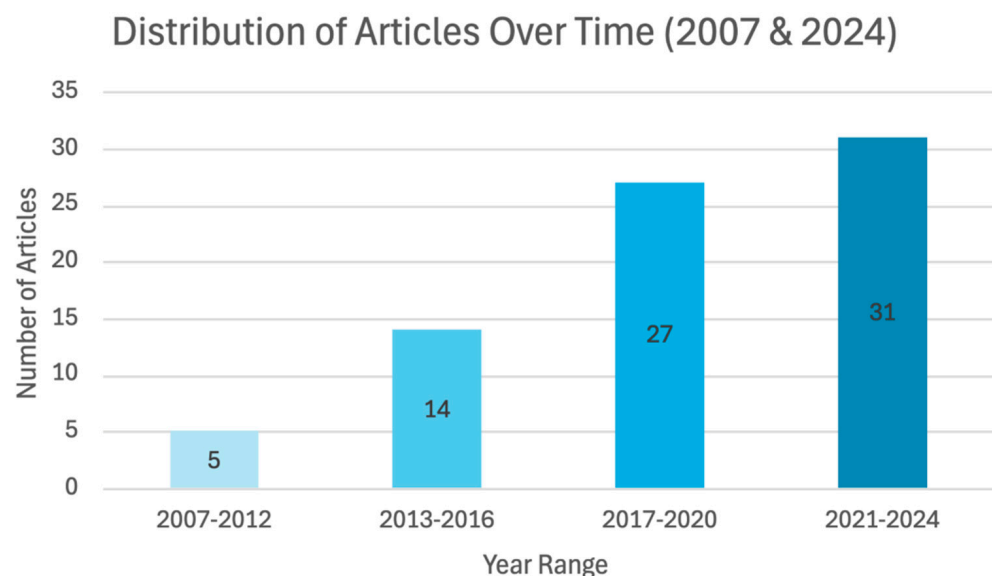
In this systematic review, no specific effects were used to synthesize the outcomes. The eligibility for synthesis was determined by comparing the study characteristics against predefined criteria, with a focus on architectures based on LLMs, GANs, and VAEs. Data preparation involved standardizing the data formats, but no missing summary statistics were addressed or included. The results were organized and visually displayed in tables and charts to present the individual study findings and overall trends. The synthesis methods involved the qualitative integration of results without formal meta-analysis or

software tools. Heterogeneity among study results was explored through narrative analysis, and sensitivity analyses were conducted to evaluate the robustness of the findings and ensure the reliability of the conclusions drawn.

### 3.5. Results

We conducted a comprehensive search of several databases and several primary studies concerning generative adversarial networks (GANs), variational autoencoders (VAEs), and large language models (LLMs), which are the three areas we are investigating in machine learning. These investigations are important for the development of the area and as such, we categorized them to enhance clarity while presenting the results. None of the excluded studies meet the inclusion criteria. The risk of bias for each included study was assessed manually, considering factors such as study design, sample size, data handling, and potential conflicts of interest, although no formal assessment tools or standardized criteria were used in the evaluation process.

The bar graph, in Figure 2 presents an overview of all the studies utilized in this literature review. This graph illustrates the temporal distribution of the selected studies across the publication years, encompassing core generative AI models such as GANs, VAEs, and LLMs. By plotting these studies over time, the graph provides an insightful visualization of the research trajectory and highlights significant trends and shifts in focus within the field of synthetic data generation.



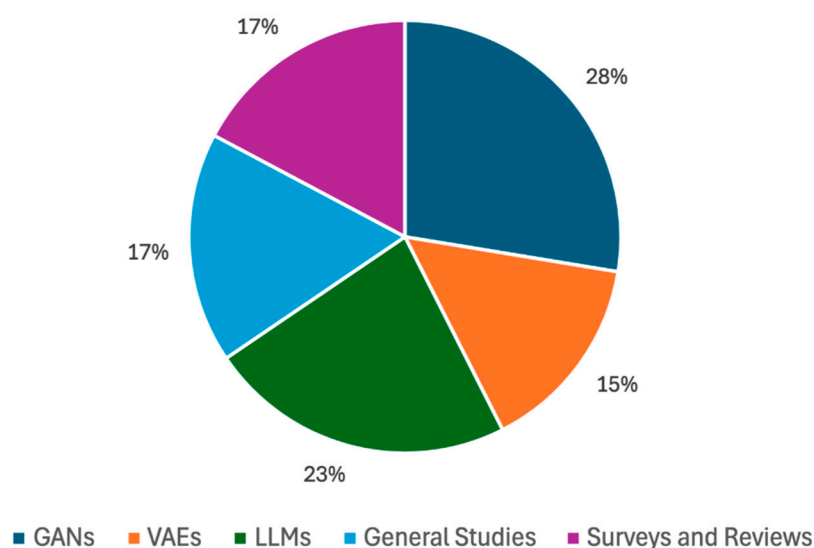
**Figure 2.** Year of publication in which all the selected 77 studies were published.

The pie chart in Figure 3 provides a detailed breakdown of the studies included in this literature review, categorizing them into five main segments: GANs, LLMs, VAEs, general studies, and surveys/reviews. This visual representation highlights the proportion of studies dedicated to each category, offering a clear overview of the research focus areas and the total number of studies analyzed. By illustrating the distribution of these studies, the pie chart complements the temporal analysis provided by the bar graph, thus enhancing our understanding of the research landscape in synthetic data generation using generative AI models.

- **GANs:** This category contains all the studies that reveal new GAN models and their implementations. Various modified versions and frameworks of GANs are included in this category.
- **VAEs:** This category contains all studies that introduce the foundations of VAEs and their modified versions.

- **LLMs.** This category comprises various LLM foundational studies on GPT-1, BERT, and their modified versions.
- **General studies:** This category includes all the studies related to the need for differential privacy, biases in machine learning, problems with synthetic data, and the style and format for writing systematic reviews. This category also includes all the web articles used to support the literature.
- **Surveys and reviews:** This category includes all the literature reviews that attempted to explore synthetic data generation, but they did not propose any new model. These studies discuss the applications, challenges, and future research prospects for synthetic data generation.

### Distribution of References Across various Categories



**Figure 3.** The proportion of various types of studies selected for this literature review.

## 4. Synthetic Data Generation

In this section, we examine existing research and approaches in the field of synthetic data generation across a variety of disciplines, including finance and healthcare. This study focuses on both numerical and temporal data. In addition, we delve deeper into the metrics for analyzing synthetic data, looking at how datasets are used to assess fidelity and utility. This article highlights key studies and approaches that have advanced the understanding and application of generative models, such as generative adversarial networks (GANs), variational autoencoders (VAEs), and large language models (LLMs), to address challenges such as data scarcity, imbalance, and privacy.

Synthetic data can be generated using a variety of methods, each with its aims and applications. These techniques enable the development of synthetic datasets for various purposes, including model training and testing, privacy protection, and augmentation of constrained data sources. In the following discussion, we examine the many domains and methodologies used to generate synthetic data and highlight their benefits and prospective applications.

### 4.1. Programmatic Approach

The demand for synthetic data has dramatically increased in the rapidly developing field of data analytics. Synthetic data are a crucial tool for training machine learning models, performing experiments, and protecting privacy when actual data are sparse or sensitive. The ability to generate high-quality synthetic data enables researchers and practitioners to overcome the constraints of real-world datasets, such as inadequate data, privacy problems,

and high costs. Among the different available strategies, the programmatic approach is one option.

Synthetic data are generated programmatically by applying predetermined rules, logical constructs, and simulation models to create data that closely resembles the qualities and behaviors of real-world data. These methods are highly adaptable and can be adjusted to recreate the specific patterns, relationships, and structures observed in real data. Programmatic approaches can be further divided into the following categories.

#### 4.1.1. Rule-Based Systems

Rule-based systems create synthetic data by complying with a set of established rules and logical constructs that mimic the features and relationships of real data [27]. These systems are deterministic, which means that the same rules consistently yield the same results, making them extremely predictable and reproducible. This method is beneficial for creating structured data if links and limitations are fully defined.

The steps in rule-based systems include establishing rules that reflect the features and relationships of the data, implementing these rules in the code, and running the program to generate synthetic data.

#### 4.1.2. Random Sampling

Random sampling is a fundamental technique for creating synthetic data by randomly selecting observations from the existing datasets. This method preserves the statistical features and distributions of the original data, making it beneficial for producing larger datasets and balancing the class distributions. Researchers and data scientists can generate fresh datasets that closely mirror the properties of real-world data while avoiding direct disclosure of sensitive information. Random sampling is especially useful in situations where privacy considerations prevent the direct use of original datasets, or where more data are required for meaningful analysis and model training.

### 4.2. Statistical Approach

Statistical models for synthetic data generation use mathematical and probabilistic methodologies to mimic datasets with real-world features. These approaches use statistical distributions, patterns, and correlations found in the original data to generate new datasets that are statistically comparable, but not identical to the original.

Parametric models use assumed distributions or functional forms to generate the synthetic data. These models estimate parameters from observed data, such as the mean and variance in Gaussian distributions or coefficients in regression models, and then produce new data points that follow defined distributions or relationships [28]. This approach offers a formal framework for data synthesis, making it simple and efficient to create synthetic datasets with statistical qualities similar to those of original data. Parametric models are especially useful when the underlying data structure is well defined and closely follows the anticipated distribution, allowing for simple parameter estimation and ensuring that the output data closely match real-world patterns.

By contrast, non-parametric models use more adaptable and data-driven methods to generate synthetic data. These models make a few assumptions about the underlying data distribution, instead relying on empirical distributions calculated directly from the observed data [29]. Non-parametric approaches frequently use methods such as kernel density estimation, nearest neighbor algorithms, and decision trees to simulate new data points that capture the complexity and sensitivity of the original dataset. Non-parametric models excel in circumstances with unknown or highly variable data distributions because they accommodate various heterogeneous data patterns without imposing strict assumptions. This adaptability enables non-parametric models to generate synthetic data that closely resemble the complexities and irregularities observed in real-world datasets, making them useful for tasks that require robust data synthesis and analysis.

#### 4.3. Differential Privacy Approach

Differential privacy-based methods create synthetic data by adding carefully calibrated noise to the actual data to protect the privacy of individual data entries. The fundamental idea of differential privacy is to provide strong guarantees that any analysis or query of the data produces statistically indistinguishable results regardless of whether any individual's data are included. This is conducted using a method that introduces random noise into the data or results of data queries, ensuring that the output cannot be linked back to any particular individual with high certainty [30].

Several stages are required to generate synthetic data under differential privacy. First, a model or algorithm was chosen to represent the data distribution. This model was then trained on the original dataset, during which noise was incorporated into the training process or model parameters to maintain differential privacy. Once trained, the model is used to generate synthetic data that represent the fundamental patterns and relationships of the actual data, but does not reveal specific details about every single data point. A privacy parameter known as epsilon ( $\epsilon$ ) controls the quantity and nature of the introduced noise, which balances the trade-off between data utility and privacy. A lower epsilon suggests stronger privacy protections but may result in less accurate synthetic data, whereas a higher epsilon allows for more precise data but worse privacy protection. For instance, one study addressed the challenge of setting a proper value of epsilon and attempted to provide insights into choosing the correct value of epsilon [31].

This technique ensures that the synthetic data retain the crucial statistical properties of the original dataset such as distributions and correlations while preventing the re-identification of individuals. Differential privacy-based synthetic data generation is especially valuable in sensitive industries, such as healthcare and finance, where maintaining individual privacy is vital while still allowing for meaningful data analysis and exchange. Organizations can use differential privacy to develop and share synthetic datasets that provide robust analytical insights and model training while maintaining the privacy of personal information.

#### 4.4. Machine Learning Approach

Machine learning-based techniques have transformed the production of synthetic data, providing a robust and versatile approach that outperforms previous methods in many ways. These techniques use advanced algorithms and neural networks to generate synthetic datasets that closely resemble the features and complexities of real-world datasets. Machine learning has numerous advantages for synthetic data generation. This enables the efficient generation of large amounts of data, which is especially useful for training machine learning models when real data are limited, expensive, or subject to privacy concerns. Furthermore, machine learning algorithms can detect complex patterns and dependencies in data, producing synthetic datasets that retain the underlying statistical relationships and distributional features of the original data.

A key benefits of using machine learning to generate synthetic data is that it can be applied to various use cases. In sectors such as healthcare, finance, and self-driving cars, obtaining large, annotated datasets is difficult. Synthetic data can bridge this gap by providing additional training instances that improve the performance and resilience of the model. Furthermore, synthetic data are invaluable for data augmentation because they improve the generalization skills of models by exposing them to a wider range of scenarios. Another crucial use case is privacy preservation, which allows machine learning-generated synthetic data to be utilized to share insights and improve models while protecting sensitive information.

Among machine learning approaches, generative adversarial networks (GANs), variational autoencoders (VAEs), and large language models (LLMs) are the most popular. GANs are an excellent example of machine learning-based synthetic data generation. The GANs are composed of two neural networks. Through iterative competition, the generator learns to produce realistic data that may trick the discriminator, resulting in synthetic datasets



that are extremely close to real-world data. However, VAEs adopt a probabilistic approach to produce synthetic data. VAEs encode the input data into a latent space before decoding them again, allowing fresh data points to be sampled from this latent representation. This procedure ensures that the created data have the same key properties and variability as the original dataset. Finally, LLMs such as GPT-3, demonstrated outstanding effectiveness in producing synthetic text data. When trained on large datasets, these models may generate coherent and contextually accurate texts that are indistinguishable from human-written content, making them perfect for natural language processing applications. By reviewing these techniques in depth, we aimed to explore more ways to utilize them and how to improve them for specific use cases.

#### 4.4.1. Generative Adversarial Networks (GANs)

Table 2 contains a list of the studies used to explain the concepts of core GAN architectures and their frameworks.

**Table 2.** Studies related to generative adversarial networks.

Primary Study	Year	Overview
Islam, S. [32]	2024	This paper reviews the advancements, applications, and challenges of using GANs in medical imaging, particularly in enhancing the quality of medical images for better diagnostic accuracy.
Strelcenia, E. [33]	2023	This study discusses how GANs have been used to improve breast cancer detection through the generation of synthetic data, leading to enhanced classification performance.
Ali, H. [34]	2023	The paper explores the application of GANs in generating synthetic COVID-19 data to address data scarcity and improve the performance of diagnostic models.
Yadav, P. [35]	2023	This paper introduces a novel approach using MTS-TGAN for generating synthetic multivariate time series data, improving the quality and applicability of time series data in various fields.
Charitou, C. [36]	2021	The study discusses the use of GANs for fraud detection by generating synthetic data that helps improve detection rates in financial services.
Xu, L. [37]	2019	This paper presents a conditional tabular GAN (CTGAN) model for generating synthetic tabular data, which is particularly effective in handling imbalanced datasets.
Xu, L. [38]	2019	The TGAN model described in this paper is designed for generating high-quality synthetic tabular data, preserving inter-column relationships.
Miyato, T. [39]	2018	This paper introduces the concept of cGANs with a Projection Discriminator, improving the quality of class-conditional image generation by modifying the discriminator's design.
Xie, L. [40]	2018	The paper discusses a Differentially Private Generative Adversarial Network (DP-GAN), which ensures privacy while generating synthetic data for sensitive applications.
Zhu, J. [41]	2017	This paper presents CycleGAN, which enables unpaired image-to-image translation, useful in scenarios where paired data are unavailable.
Chen, X. [42]	2016	CcGAN is introduced in this paper as an interpretable GAN model, focusing on disentangling latent space to generate meaningful representations.
LeCun, Y. [43]	2015	This seminal paper on deep learning provides an overview of the field and discusses the development of deep neural networks, highlighting their success in various applications.
Radford, A. [44]	2015	This paper describes the development of a DCGAN, a variant of GANs using deep convolutional layers to improve the quality of generated images.
Mirza, M. [45]	2014	This paper introduces conditional GANs (cGANs), which improve GANs by conditioning the generation process on additional information, allowing for more controlled data generation.
Goodfellow, I. [7]	2014	The foundational paper that introduced GANs, describing the adversarial process where two neural networks compete, leading to the generation of realistic data.

In 2014, a research paper proposed the idea of generative adversarial networks, in which two neural networks are trained simultaneously through a process of competition [6]. Various application-specific GANs have been developed by using this approach. For instance, cGANs can incorporate conditional information for better results and TGANs

can incorporate better synthetic tabular data. These improved GANs are extremely useful because they improve the quality and flexibility of the synthetic data production. By addressing specific issues and refining various components of the generation process, they can produce more accurate, stable, and context-aware results. This advancement has significant implications for a various industries, including healthcare, where high-quality synthetic data can improve medical imaging and diagnostics, and entertainment, where realistic content production can change gaming and visual effects.

To improve the utility and effectiveness of GANs, the authors introduced the concept of conditional generative adversarial networks (cGANs) by adding conditional inputs to both the generator and discriminator [45]. The generator received an extra conditioning variable and noise vector, whereas the discriminator received the generated sample and matches the conditioning variable. This methodology enables the model to generate data based on certain qualities or labels, thereby increasing control over the generated data. This research uses a variety of experiments, such as generating MINIST digits conditioned on class labels, to demonstrate the usefulness of cGANs. Their research highlighted the ability of cGANs to generate complex structured data, such as annotated images and multi-modal distributions. This conditional method broadens the scope of GAN applications, making them effective for jobs requiring fine-grained control over the data generation process.

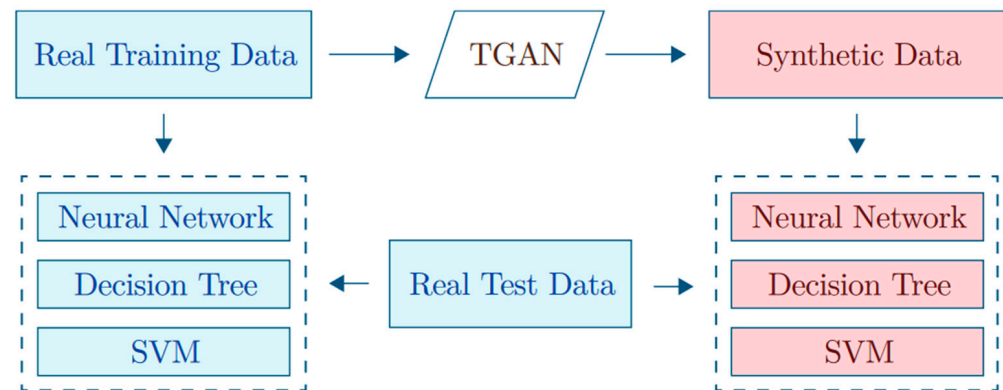
As discussed, cGANs were developed to perform best with categorical conditioning; however, in early 2021, continuous conditional generative adversarial networks (CcGANs) for image generation based on continuous scalar labels were introduced [42]. This innovation addressed inbuilt issues such as scarce real-image data for specific regression labels and the adaptation of label input methods to accept scalar values. By including novel empirical losses suited for continuous settings, CcGANs outperform traditional cGANs in generating diverse, high-quality images across architectures such as circular 2-D Gaussians, RC-49, and UTKFace. These advances not only improve visual fidelity and quantitative measures but also open possibilities for meaningful applications in sectors ranging from autonomous driving for precise steering control to medical imaging for synthetic data generation in cell-counting activities.

To utilize cGANs, researchers have proposed a novel image-denoising method for low-dose chest imaging using conditional generative adversarial networks (cGANs) [43]. This method effectively addressed the issue of quantum noise that occurs during low-dose medical imaging data collection. The findings of this study provide insights into how this innovative approach, which uses cGANs, greatly improves image quality when compared to classic denoising methods such as total variation minimization and non-local means. Quantitative evaluations, such as the structure similarity index measure (SSIM), show that the cGAN-based method outperforms traditional techniques, reaching 1.5 to 2.5 times better SSIM scores. This approach preserves crucial image information, improves the resolution of noise-corrupted images, and provides a greater diagnostic utility.

Another improvement to cGANs, proposed in [39], suggests a unique projection-based method for adding conditional information to the discriminator of conditional GANs (cGANs), which improves the performance of conditional image generation. Unlike standard cGAN frameworks, which concatenate a conditional vector with feature vectors, this technique preserves an underlying probabilistic model. The proposed model significantly improves the quality of class-conditional picture production on the ILSVRC2012 (ImageNet) 1000-class image dataset, beating the current best with a single discriminator–generator pair. These findings highlight the importance of discriminator design and distributional metrics in boosting cGANs performance, with possible applications including semantic segmentation and image-to-image translation tasks.

To use GANs efficiently, one study introduced a tabular GAN (TGAN), an effective framework that generates high-quality synthetic tabular data with both discrete and continuous variables [38]. TGAN uses deep neural networks to provide data correlation integrity and scalability across large datasets. This is accomplished using novel strategies such as clustering numerical variables to handle multi-modal distributions and adding noise and

KL divergence to the loss function for discrete feature generation. Multiple dataset evaluations show that TGAN outperforms conventional statistical generative models, particularly in terms of inter-column correlations and scalability. The average performance gap between the real and synthetic data generated using the TGAN was 5.7%. This work demonstrates TGAN's ability to generate trustworthy synthetic data for data science applications, representing a significant step forward in relational database modeling. Figure 4 shows the training and evaluation process of the TGAN.



**Figure 4.** The process of training and evaluating a TGAN involves using real training data, including labels, to train a GAN and generate synthetic data. Multiple machine learning models (five methods were chosen) were trained using both real and synthetic data.

To create a higher privacy-preserving model that can generate synthetic tabular data with complex distributions, conditional tabular GAN (CTGAN) was introduced [37]. This paper presents a flexible and robust model capable of learning complex column distributions which current deep generative models struggle with. CTGAN uses mode-specific normalization to transform continuous values into bounded vector representations that are suitable for neural networks. Furthermore, the conditional generator and training-by-sampling strategies efficiently addressed the issue of imbalanced training data. Empirical data show that CTGAN outperforms Bayesian networks, which have historically discretized continuous values and greedily learned distributions. This study highlights the ability of CTGAN to predict and generate more accurate synthetic tabular data, resulting in significant advances in applications that require privacy-preserving data synthesis.

Deep convolutional generative adversarial networks (DCGANs) are a significant advancement in the field of deep generative models, expanding the capabilities of regular GANs by explicitly adding convolutional and convolutional-transpose layers to their architectures. DCGANs were first introduced in the seminal paper on “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks” [44]. It processes  $3 \times 64 \times 64$  input images using stride convolution layers, batch normalization, and LeakyReLU activations in the discriminator, producing scalar probabilities indicative of the real data distribution [46]. By contrast, the generator uses convolutional-transpose layers, batch normalization, and ReLU activations to transform latent vectors derived from a standard normal distribution into high-resolution  $3 \times 64 \times 64$  RGB images. The DCGAN architecture is a significant advancement in unsupervised representation learning, providing solid rules for building powerful generative models that are capable of producing high-quality realistic images.

The CycleGAN research paper presented an innovative approach to image-to-image translation, a complicated vision and graphics problem that involves converting images from one domain to another without using paired training data [41]. Unlike existing methods that rely on aligned image pairings, CycleGAN uses adversarial loss to ensure that the images generated from source domain X to target domain Y are indistinguishable from the actual images in domain Y [47]. This approach was improved by inverse mapping  $F: Y \rightarrow X$  and a cycle consistency loss that ensures that  $F(G(X)) = X$  for bidirectional consistency.

This study presents qualitative results for a variety of tasks, including style transfer, object transfiguration, and photo enhancement, demonstrating the capability of CycleGAN in circumstances lacking paired data. Despite its success, the approach suffers from tasks that require major geometric alterations, such as converting dogs to cats, where only minor changes occur. CycleGAN makes major contributions to the field of unpaired image-to-image translation and provides robust algorithms for a wide range of applications.

To address the problem of detecting money laundering in online gambling, researchers have developed a novel system called synthetic data generation GAN (SDG-GAN) [36]. Generative adversarial networks (GANs) were used in this system to create synthetic data for training supervised classifiers. This paper highlights the issue of class imbalance in fraud detection datasets and demonstrates that it outperforms other traditional oversampling methods, such as SMOTE, B-SMOTE, ADASYN, and alternative adversarial network architectures such as cGANs. By performing extensive evaluations on both public imbalanced datasets and a real-world gambling fraud dataset, the SDG-GAN significantly enhances classification performance by raising the F1 score by 5% above rule-based systems and 0.4% over various oversampling techniques. However, more recent and potentially superior GAN models may have been left out because of the speed of their development. Moreover, using the published data sets in this study may not adequately represent all possible situations in online gambling.

Another study explained the use of GANs for breast cancer detection through the generation of synthetic data [33]. Researchers are aware that healthcare has a problem with limited datasets, suggesting K-CGAN which was trained on the Wisconsin Breast Cancer dataset with 357 malignant and 212 benign cases. Performance metrics, such as recall, precision, accuracy, and F1-scores, were used to evaluate how well the synthetic data generated by the K-CGAN can be classified. Empirical evidence showed that the K-CGAN outperformed other types of GANs as it demonstrated the highest stability and closely reproduced the original characteristics of the dataset. Nevertheless, these performance evaluation metrics may not cover all clinical subtleties, thus necessitating further validation in real-world settings.

To generate synthetic time series, researchers developed MTS-TGAN, an innovative GAN model designed to generate realistic multivariate time series data [35]. This model solves problems associated with the procurement of huge balanced datasets in certain fields such as finance, healthcare, and manufacturing which can be costly. The authors utilized a mix of qualitative (t-SNE and PCA) and quantitative methods (MAE and MSLE) to show that the MTS-TGAN can adequately approximate the real data distribution, leading to a significant reduction in errors in predictive and discriminative scores by 13% and 10%, respectively. However, despite these promising results, there are limitations, including a lack of testing for the model on more diverse and complex datasets incorporating multi-model elements and mixed data types.

In a comprehensive study on the use of GANs in medical imaging, researchers explored their transformative potential in generating synthetic medical images, improving data quality, and aiding in tasks such as image segmentation and disease detection [32]. This study delved into recent developments, popular datasets, and pre-processing techniques, providing a detailed examination of a range of GAN algorithms along with their applications. This analysis revealed some drawbacks and merits of these models, providing insights into their performance and practical efficacy in medicine. One limitation should be noted regarding the inclination towards published research, which may not capture the latest developments in this rapidly evolving field. Furthermore, focusing on selective datasets and pre-processing methods may limit the generalizability of these findings to various medical imaging contexts.

A study critically examined the use of GANs to generate synthetic lung CT scans and X-ray images to improve AI-based COVID-19 diagnostic models [34]. The authors identified several common issues, including data bias, lack of reproducibility, and insufficient expert feedback, by analyzing 43 published studies. The reproducibility of these results is

hampered by a major challenge that is the unavailability of the source code. However, their clinical applicability remains limited, although GANs have the potential for data augmentation. This study made recommendations for future research to improve the acceptability and reliability of GAN-based approaches in medical imaging, which has one limitation: it restricted its findings from only PubMed, Scopus, IEEE Xplore, and Google Scholar-indexed papers, thus excluding some valuable preprints and unpublished literature. This focus on image-based studies has narrowed the scope and potentially overlooked other relevant applications.

#### 4.4.2. Variational Autoencoders (VAEs)

Table 3 lists the studies used to explain the concepts of core VAE architectures and their frameworks.

**Table 3.** Studies related to variational autoencoders.

Primary Study	Year	Overview
Mostofi, F. [48]	2024	This study integrates a variational autoencoder (VAE) into a multi-head graph attention network (GAT) to address the class imbalance in construction datasets, significantly improving predictive performance in construction management tasks.
Wu, J. [49]	2023	This paper focuses on using variational autoencoders (VAEs) for generating synthetic financial data while preserving privacy, demonstrating effectiveness in financial applications where data sensitivity is a concern.
Li, H. [50]	2023	The study proposes a causal recurrent variational autoencoder (CR-VAE) for generating time series data with an emphasis on learning Granger causality, showing superior performance in medical time series analysis like EEG and fMRI data.
Saldanha, J. [51]	2022	This paper discusses the use of VAEs to generate synthetic respiratory sounds for the classification of respiratory diseases, addressing class imbalance and improving classification performance in minority classes.
Kok, S. [52]	2020	The study introduces oblivious variational autoencoders (OVAEs) for generating privacy-preserving synthetic tabular data, particularly in environments requiring strict data privacy like healthcare and finance.
Islam, Z. [53]	2020	This paper addresses class imbalance in crash datasets using VAEs to generate synthetic data, leading to improved specificity and sensitivity in crash prediction models.
Goyal, P. [54]	2019	This study focuses on using Contrastive Predictive Coding (CPC) for self-supervised video representation learning, enabling models to learn more effective video features without requiring large, labeled datasets.
van den Oord, A. [55]	2017	The paper introduces Neural Discrete Representation Learning, which uses discrete latent representations in VAEs to prevent posterior collapse, making it effective in unsupervised learning tasks like phoneme learning.
Kipf, T. [56]	2016	This study presents variational graph autoencoders (VGAEs) for unsupervised learning on graph-structured data, showing their effectiveness in tasks like link prediction and graph clustering.
Rasmus, A. [57]	2015	The paper explores Ladder Networks for semi-supervised learning, demonstrating their ability to improve learning performance by leveraging both labeled and unlabeled data.
Sohn, K. [58]	2015	This study introduces a deep conditional generative model for learning structured output representations, offering advancements in probabilistic inference in complex output spaces like image segmentation.
Makhzani, A. [59]	2015	This paper discusses adversarial autoencoders (AAEs), which combine VAEs with GANs to align latent space distributions, offering advancements in generative modeling and semi-supervised learning.
Kingma, D.P. [6]	2013	The foundational paper on Auto-Encoding Variational Bayes (AEVB), introducing a powerful method for variational inference that has become a cornerstone in generative modeling.



In 2013, researchers introduced a new method called stochastic gradient variational Bayes (SGVB) to improve the handling of certain data types [6]. This technique addresses the challenges that arise when dealing with data containing continuous latent variables (hidden factors that influence the data). The SGVB method provides an efficient method for estimating these hidden factors. It can also be used for various tasks, such as determining the most likely parameters (values that best explain the data) or making predictions based on these hidden factors.

Researchers have built on SGVB to create a method called Auto-Encoding Variational Bayes (AEVB). The AEVB method takes advantage of SGVB's efficiency and makes it easy to use with common optimization techniques. This makes AEVB particularly well suited for analyzing large datasets where each data point is independent of the others. The AEVB method uses the SGVB to learn an approximate model to understand the data. This study represents a significant improvement in variational inference approaches, providing a solid framework for dealing with continuous latent variables in a variety of realistic contexts, including online and non-stationary data sets.

An innovative method was introduced (CVAE) to make a substantial contribution to structured output prediction by building a deep conditional generative model that uses Gaussian latent variables [58]. This approach overcomes the limitations of classic supervised deep learning algorithms, which struggle with probabilistic inference and produce different predictions in complicated structured output spaces. The model, which was trained using the stochastic gradient variational Bayes framework, could efficiently and quickly predict the stochastic feed-forward inference. Key advances include the addition of input noise injection and multi-scale prediction objectives, which improve the resilience of structured prediction algorithms. The experimental results show that the model outperforms deterministic deep neural networks in pixel-level object segmentation and semantic labeling tasks on datasets such as Caltech-UCSD Birds 200 and Labeled Faces in the Wild. This research emphasizes the necessity of probabilistic inference in multi-modal output distributions, as well as the scalability and efficiency of the proposed stochastic neural network model in structured output prediction difficulties.

The adversarial autoencoder (AAE) research paper describes a unique approach for performing variational inference that combines the principles of generative adversarial networks (GANs) with probabilistic autoencoders [59]. By aligning the aggregated posterior of the hidden code vector with an arbitrary prior distribution, AAEs ensure that the samples generated from any area of the prior space are meaningful. This alignment allows the decoder to train a deep generative model that maps the imposed subject before data distribution. Experimental results using datasets such as MNIST, Street View House Numbers (SVHN), and Toronto Face demonstrate AAEs' competitive performance of AAEs in both generative modeling and semi-supervised classification tasks. This study represents a significant advancement in the use of adversarial training for variational inference in autoencoders, resulting in robust methods for complex data representation and classification applications.

The Ladder Variational Autoencoder (LVAE) is a novel inference model developed to improve the training of deep variational autoencoders [57], which are traditionally challenged by numerous layers of dependent stochastic variables. The LVAE, similar to a Ladder Network, uses a recursive approach to modify the generative distribution with a data-dependent approximate likelihood. Compared to typical bottom-up inference models in variational autoencoders, this approach yields a better predictive log-likelihood and narrower log-likelihood lower bound. This study also underlines the relevance of batch normalization and deterministic warm-up in training deep variational models, highlighting their significant impact on the generative performance. The findings highlight that LVAEs achieve a qualitatively different and more distributed hierarchy of latent variables, marking a substantial improvement in the efficiency and effectiveness of hierarchical latent representation learning in variational models.

In unsupervised learning, the Vector Quantized-Variational AutoEncoder (VQ-VAE) [55] represents a significant development by resolving some of the major drawbacks of the classic variational autoencoders (VAEs). Unlike VAEs, VQ-VAE uses discrete latent representations via vector quantization, thereby avoiding the problem of “posterior collapse” which is common in powerful autoregressive decoders. This model also learns a dynamic prior, which improves its capacity to produce high-quality results across multiple domains such as photos, videos, and audio. The discrete latent space of the VQ-VAE captures crucial data properties, allowing high-level tasks such as unsupervised phoneme learning and speaker conversion. This study highlights the effectiveness of VQ-VAE in modeling long-term relationships and creating meaningful sequences without supervision, establishing it as a significant advancement in the field of generative models and representation learning. The ability of VQ-VAE to attain competitive likelihoods on datasets such as CIFAR10 further demonstrates its potential for widespread use in machine learning tasks.

While much of the previous research has been on graph embedding problems, there is increasing interest in expanding the advancement of generative models beyond images and text to graph structures. A novel strategy uses variational autoencoders (VAEs) to construct graphs directly from continuous embeddings, avoiding the difficulties of linearizing discrete graph structures [56]. The proposed technique generates a probabilistic fully connected graph with a given maximum size, thereby providing an entirely new perspective on graph construction. Evaluations using molecular datasets showed that the model could learn appropriate embeddings for small molecules but struggled to capture the complicated chemical interactions found in larger molecules. This research represents a significant step toward more advanced graph decoders and is expected to encourage future progress in graph-based generative models. The technological contributions and early accomplishments highlighted in this study demonstrate VAEs’ potential to handle challenging generative tasks for graph structures.

Directed acyclic graphs (DAGs) are widely used to describe complicated structures such as neural networks and Bayesian networks, and research into deep generative models for DAGs has acquired significant momentum. To encode DAGs into a latent space, a study presented a novel DAG variational autoencoder (D-VAE) using graph neural networks (GNNs) [54]. The distinctive feature of the D-VAE is its asynchronous message-passing scheme, which encodes the computations on DAGs as opposed to traditional simultaneous message-passing schemes that focus on local graph structures. This approach allows for a more accurate representation of the partial order inherent in the DAGs. The effectiveness of the D-VAE is demonstrated through neural architecture search and Bayesian network structure learning, where it not only generates novel and valid DAGs but also facilitates a smooth latent space for optimizing DAG performance via Bayesian optimization. This study represents a significant step forward in the development of generative models for graph-structured data, offering new approaches for the encoding and generation of complex DAGs.

In another study, researchers presented an oblivious variational autoencoder (OVAE), which is a novel model designed to generate high-fidelity synthetic tabular data while preserving privacy [52]. OVAE combines variational autoencoders (VAEs) with differentiable oblivious decision trees (ODTs), leveraging ODTs’ effectiveness in handling tabular data. This model is designed to address issues related to privacy when personally identifiable information is shared in datasets such as EHRs and financial records. Through extensive experiments on 12 real-world datasets, the OVAE outperformed state-of-the-art models. Nevertheless, this research recognizes that future work must integrate domain knowledge as an additional restriction and use normalizing flows to further enhance this approach.

Moreover, one study explored the use of variational autoencoders (VAEs) to generate synthetic data in the financial sector, addressing privacy concerns associated with real data [49]. The researchers used a sensitivity-based approach to explain how the input features in tabular datasets can affect the VAE’s latent space and its synthetic dataset generation process. First- and second-order partial derivatives were used to identify the

importance and interaction of the input features globally and locally. The method was validated using a simulated dataset and three Kaggle banking datasets, demonstrating its effectiveness in explaining the feature contributions and relationships.

Another study explored the use of variational autoencoders (VAEs) to address the class imbalance in the ICBHI respiratory sounds database for lung sound classification [51]. Researchers have utilized different VAE variants—MLP-VAE, convolutional VAE, and conditional VAE—to synthesize minority class data. The quality of the synthetic sounds was assessed using the Fréchet Audio Distance, cross-correlation, and Mel Cepstral Distortion. The results indicated that augmenting the imbalanced dataset significantly improved the classification performance, especially for minority classes. Despite these improvements, challenges remain in fully capturing the variability of real respiratory sounds using synthetic data.

To generate time series data, researchers have proposed a causal recurrent variational autoencoder (CR-VAE), a generative model designed to learn Granger causality graphs from multivariate time series data [50]. Unlike traditional recurrent VAEs, CR-VAE has a multi-head decoder such that each head corresponds to one dimension of the input data. This study used a two-stage training approach and evaluated CR-VAE on both synthetic data and real-world human brain datasets, including EEG and fMRI signals. The model outperforms state-of-the-art time series generative models in both qualitative and quantitative assessments and accurately discovers causal graphs. While CR-VAE demonstrates robust performance, one of its limitations is that the isotropic Gaussian assumption for latent factors restricts its generative capabilities.

To address the issue of imbalanced crash datasets, researchers introduced a data augmentation technique using variational autoencoders (VAEs) [53]. The dataset contained 625 crash events compared with over 6.5 million non-crash events, leading to poor performance of the learning algorithms. The VAE model encodes events in a latent space and effectively differentiates between crashes and non-crashes. Synthetic data generated from this latent space were statistically similar to real data and favorably compared with traditional oversampling methods such as SMOTE and ADASYN. The VAE-augmented data improved the specificity by 8% and 4% for logistic regression (LR) and support vector machine (SVM) models, respectively, and enhanced the sensitivity by 6% and 5%, respectively, compared with ADASYN. However, the drawback of autoencoders is that they may remove noise from the input which can cause loss of important features.

One study addressed the issue of class imbalance in construction datasets, which hampers the predictive performance of machine learning models [48]. By integrating a variational autoencoder (VAE) into a multi-head graph attention network (GAT), researchers have aimed to augment underrepresented classes and create balanced datasets. They collected comprehensive productivity data on various construction activities. The integration of the VAE significantly improved the accuracy of the GAT from 90.6% to 92.5% for finishing, 81.1% to 94.4% for concrete, and 92.2% to 95.4% for insulation activities. However, further exploration of data augmentation strategies in engineering contexts is necessary to fully realize the potential of such models.

#### 4.4.3. Large Language Models (LLMs)

Table 4 contains a list of the studies used to explain the concepts of core LLMs' architectures and their frameworks.

**Table 4.** Studies related to large language models.

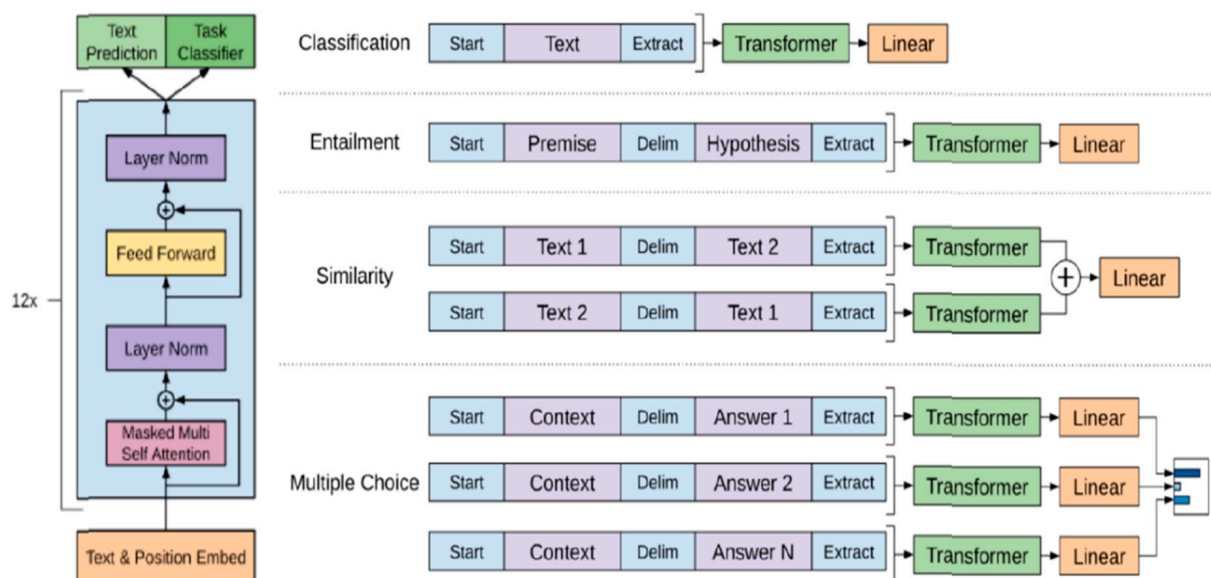
Primary Study	Year	Overview
Li, Z. [60]	2023	This study investigates the potential and limitations of using large language models (LLMs) for generating synthetic datasets specifically for text classification tasks, highlighting challenges related to task subjectivity and model bias.

Table 4. Cont.

Primary Study	Year	Overview
Thoppilan, R. [61]	2022	The paper introduces LaMDA (language models for dialog applications), a family of transformer-based models designed for conversational AI, emphasizing safety, factual grounding, and improving the quality of AI-driven dialogues.
Meng, Y. [62]	2022	This research presents SuperGen, a novel method for zero-shot learning in natural language understanding tasks using pre-trained language models (PLMs), enhancing performance through prompt-guided text generation and label smoothing.
Austin, J. [63]	2021	The study explores the use of LLMs for program synthesis, demonstrating the ability of these models to generate Python code from natural language descriptions, with an emphasis on challenges related to semantic understanding.
Brown, T.B. [64]	2020	This paper introduces GPT-3, a groundbreaking autoregressive language model with 175 billion parameters, highlighting its few-shot learning capabilities and its impact on various natural language processing tasks.
Radford, A. [65]	2019	This study presents GPT-2, a large-scale unsupervised language model that can perform a variety of NLP tasks, such as summarization and question answering, without task-specific fine-tuning.
Devlin, J. [66]	2019	The BERT model is introduced in this paper, offering significant improvements in various NLP tasks by pre-training deep bidirectional representations from unlabeled text.
Yang, Z. [67]	2019	This paper introduces XLNet, a generalized autoregressive model that improves BERT by capturing bidirectional context while avoiding the limitations of masked language modeling.
Radford, A. [8]	2018	This study describes GPT-1, the first generative pre-trained transformer model, demonstrating the effectiveness of unsupervised pre-training for improving performance on various NLP benchmarks.

In 2018, understanding natural language required a wide range of tasks, including classifying documents, determining whether sentences are similar, and responding to questions. Despite the abundance of available unlabeled text, a significant issue is the lack of labeled data for such tasks. To overcome this problem, an innovative approach was introduced that involved training a language model on a large amount of unlabeled text and then optimizing it for certain tasks [8]. The model proposed in this study is known as GPT-1. In contrast to other approaches, this model leverages task-specific input transformations during the fine-tuning phase, which facilitates a more effective transfer of pre-training information with only minor changes to the model. According to previous research, this method outperformed models that were specifically created for each task and greatly improved performance on a range of benchmarks. For instance, considerable progress has been made in textual entailment, question answering, and commonsense reasoning. This suggests that the model selects significant domain knowledge and comprehends long-range dependencies in text via pre-training on a variety of texts. The results indicate that unsupervised pre-training to improve the performance during particular tasks is a promising approach. Figure 5 shows the working architecture of GPT-1.

After releasing GPT-1, another model known as GPT-2 [65] was introduced which uses unsupervised language models trained on millions of webpages that make up the WebText dataset to explore advancements in natural language processing (NLP). This shows that these models can learn skills including summarization, reading comprehension, machine translation, and answering questions without direct supervision. Notably, the model obtains an F1 score of 55 on the CoQA dataset when conditioned on a document plus questions, matching or surpassing three out of the four baseline systems without utilizing the 127,000+ training samples. This research highlights the significance of the model capacity for zero-shot task transfer by demonstrating log-linear performance gains as the model size increases. Despite still underfitting WebText, the largest model, GPT-2, with 1.5 billion parameters, achieved superior outcomes on seven out of eight tested language modeling datasets in a zero-shot environment.

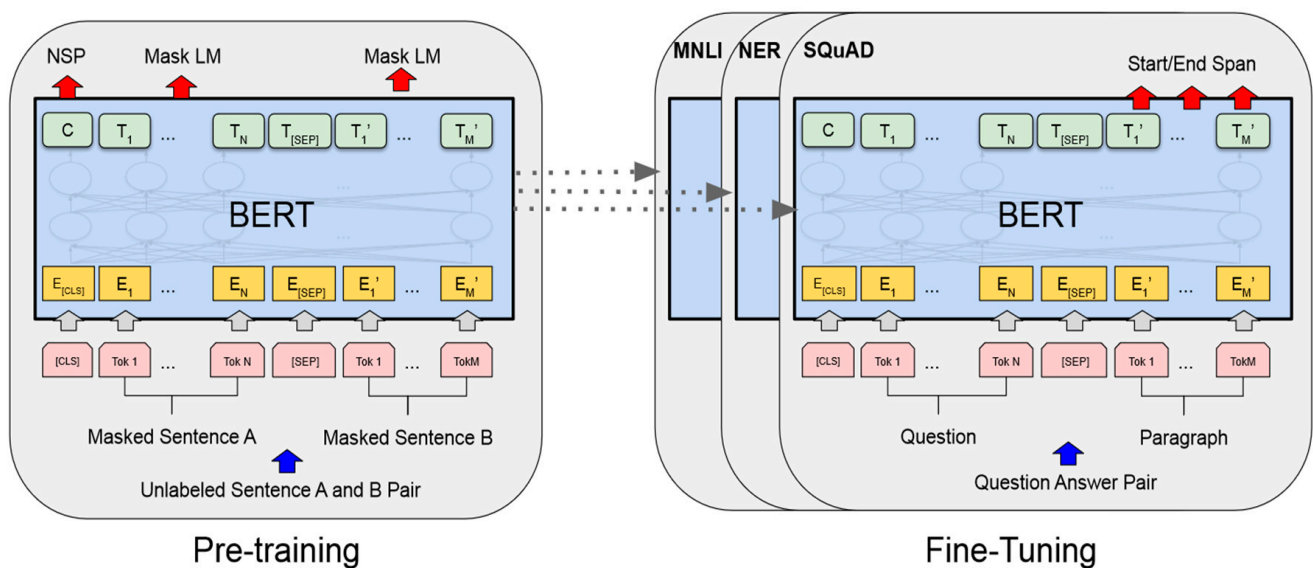


**Figure 5.** The transformer architecture and training objectives utilized in this work involve converting all structured inputs into token sequences to be processed by the pre-trained model, followed by a linear + softmax layer for fine-tuning during different tasks.

After the release of GPT-2, another model known as GPT-3 [64] was introduced. This paper describes major advances in natural language processing (NLP) by scaling up language models to improve their few-shot learning capabilities. This study introduces GPT-3, an autoregressive language model with 175 billion parameters, which is ten times larger than that of any previous non-sparse model. Unlike typical models that require significant task-specific fine-tuning datasets, the GPT-3 performs well across a wide range of NLP tasks, including translation, question answering, and cloze tasks, with no gradient updates or fine-tuning. It also excels in tasks that require on-the-fly reasoning and domain adaptability, such as unscrambling words and computing three-digit arithmetic tasks. GPT-3 showed a competitive performance on several benchmarks, proving its ability to handle new language tasks with few samples. It creates human-like text samples that are frequently indistinguishable from real articles, thereby demonstrating its potential for use in content generation and other applications. However, this model struggles with specific datasets and experiences difficulties when trained using large online corpora. This study highlights the broader social effects of implementing such powerful language models and their potential limitations. Despite these challenges, the findings indicate that very large language models, such as GPT-3, may be critical in developing adaptable, general-purpose language systems. This is a major advancement in NLP, demonstrating the ability of scaling models to improve performance across various tasks.

After the release of GPT-1, researchers at Google introduced Bidirectional Encoder Representations from Transformers (BERT) [66], a new language representation model that significantly enhances NLP by pre-training deep bidirectional representations from unlabeled text. Unlike previous models, BERT conditions both left and right contexts at every level, making it suitable for a wide range of jobs. BERT obtained excellent results on 11 NLP tasks, including a GLUE score of 80.5% (7.7% improvement), MultiNLI accuracy of 86.7% (4.6% improvement), SQuAD v1.1 Test F1 score of 93.2 (1.5-point improvement), and SQuAD v2.0 Test F1 score of 83.1 (5.1-point improvement). The simplicity and empirical strength of this model come from its ability to be fine-tuned with only just one additional output layer for a variety of tasks without requiring significant task-specific modifications. This study highlights the importance of extensive, unsupervised pre-training and generalizes the advantages of deep bidirectional architectures. Thus, BERT enables low-resource tasks to benefit from powerful language models. Figure 6 represents the pre-training and fine-tuning procedure for the BERT model.





**Figure 6.** The overall pre-training and fine-tuning procedures for BERT involve using the same architecture for both phases, apart from the output layers. The same pre-trained model parameters are used to initialize models for various downstream tasks.

Another paper, released by Google researchers in 2022, introduced LaMDA (language models for conversation applications) [61], a family of transformer-based neural language models developed mainly for conversation. The models contain up to 137 billion parameters and were trained using 1.56 trillion words of public dialog data and web content. LaMDA addresses two major concerns, safety and factual accuracy. To improve safety, the model's responses are based on human values, using a LaMDA classifier fine-tuned with crowd-annotated data to filter harmful or biased suggestions. For factual grounding, LaMDA consults external knowledge sources such as information retrieval systems, translators, and calculators, ensuring that responses are based on verifiable information. This approach was evaluated using the groundedness metric. Experiments indicate that while scaling increases quality, fine-tuning with annotated data and introducing external information significantly enhances safety and groundedness. In applications such as education and content suggestions, LaMDA outperformed the pre-trained models in terms of helpfulness and role consistency, with more than 80% of the responses being contextually accurate. This demonstrates the potential of LaMDA for the development of robust, reliable, and context-aware dialog systems.

To improve the BERT model, researchers introduced XLNet [67], a new autoregressive pretraining approach that solves the shortcomings of BERT's bidirectional context modeling. XLNet enhances pretraining by maximizing the expected likelihood over all permutations of the factorization order, thus learning bidirectional contexts without the drawbacks of masked inputs. This method combines Transformer-XL, an advanced autoregressive model, and a two-stream attention mechanism to boost the performance. XLNet consistently outperformed BERT on 20 NLP tasks, including question answering, natural language inference, sentiment analysis, and document ranking, usually with a good margin. This demonstrates XLNet's robust architecture and novel training approach, which makes it an effective tool for a variety of natural language processing use cases. For example, XLNet's ability to attain higher accuracy in tasks such as question answering and sentiment analysis illustrates its competence in dealing with complicated linguistic relationships, whereas its permutation-based training objective reduces pretrain finetune differences. By resolving both autoregressive and autoencoding pretraining limitations, XLNet establishes a new benchmark in the field, demonstrating the significance of advanced training techniques and architectural integration in improving NLP model performance.

A study introduced SuperGen, a novel approach for zero-shot learning in natural language understanding (NLU) tasks using pre-trained language models (PLMs) [62]. The method consists of a unidirectional PLM that generates class-conditioned texts guided by prompts which can be used as training data for fine-tuning a bidirectional PLM. This is achieved through regularization techniques, such as label smoothing and temporal grouping to enhance the generalization and stability. SuperGen outperformed the zero-shot prompting methods and was approximately as effective as some few-shot approaches over the seven GLUE benchmark classification tasks. However, limitations include difficulties in hyperparameter tuning owing to the lack of task-specific samples and potential challenges in generating high-quality training data for tasks with distributions that differ from pretraining data.

A study led by Google investigated the capabilities of large language models (LLMs) in synthesizing short Python programs from natural language descriptions, using the MBPP and MathQA-Python benchmarks [63]. Models with parameters ranging from 244M to 137B were evaluated in few-shot and fine-tuning settings. The results demonstrated that the performance was log-linearly dependent on the size of the model, with the largest model achieving 59.6% accuracy with MBPP and 83.8% accuracy with MathQA-Python after fine-tuning. This study also examined how human feedback can help reduce the error rates and limitations of these models in understanding programming semantics. A notable finding is that while LLMs can generate syntactically correct code, they struggle with semantic comprehension, indicating the need for further research in multi-modal models and grounding outside the program synthesis domain.

Another study investigated the efficacy of using large language models (LLMs) to generate synthetic datasets for training text classification models, focusing on the impact of task subjectivity [60]. The findings indicate that model performance decreases with higher levels of subjectivity at both task and instance levels. This study highlights that while LLM-generated synthetic data could be a useful resource, their efficacy is constrained by the subjectivity embedded in the tasks. Possible limitations include reliance on the GPT-3.5-Turbo model and crowd worker evaluations, meaning that the results may differ from those of other models or expert judgments. Consequently, this study suggests combining real-world data and human intelligence into synthetic data generation to increase diversity and enhance model effectiveness.

#### 4.4.4. More Studies and Frameworks

SynSys has been introduced as a new method for creating realistic synthetic data specifically for healthcare applications [68]. Traditional synthetic data generation methods often struggle with complexity and realism. SynSys addresses this challenge by employing hidden Markov models (HMMs) and regression models. These models were trained on real healthcare data, particularly from smart home environments, allowing SynSys to capture the intricacies of human behavioral patterns. SynSys address three key aspects: creating a realistic sequence of activities, generating realistic sensor events within those activities, and producing timestamps that reflect real-world durations and intervals. This study demonstrates that SynSys surpasses existing methods by generating more realistic data than random generation or data from different contexts. It also shows promise in improving the activity recognition accuracy when combined with semi-supervised learning, especially when real data are limited.

Although SynSys shows great promise, there are limitations to consider. The quality of synthetic data relies heavily on the quality and variety of the real data used for training. Additionally, this study focused on controlled environments, and further research is needed to explore its effectiveness in diverse real-world scenarios with various sensor types and configurations. The scalability of larger datasets and more complex environments require further investigation.

TF-GAN, an implementation of TensorFlow, is highly flexible and powerful, allowing the creation of complex generative models. However, its complexity can be a double-edged

sword; it requires significant expertise in machine and deep learning to be set up and fine-tuned. Additionally, the computational demands of the TF-GAN are substantial, making it less accessible to organizations with limited resources [69].

The Gretel Synthesis provides a user-friendly platform designed to simplify the process of generating synthetic data. This is particularly noteworthy for its emphasis on privacy and ease of integration into existing workflows. Despite these advantages, Gretel Synthesis sometimes falls short in terms of data fidelity, particularly when generating data for highly nuanced or specialized applications. The trade-off between maintaining privacy and ensuring data utility can result in synthetic datasets that do not adequately reflect the complexities of the original data [70].

DataSynthesizer focuses on generating privacy-preserving synthetic data by using differential privacy techniques to protect sensitive information. Although it excels in ensuring privacy, this focus can compromise the quality and utility of the synthetic data. Stringent privacy mechanisms often lead to synthetic datasets that lack the variability and richness needed for accurate downstream analysis, making them challenging to use in applications that require high-fidelity data [71].

One study examined the utility of synthetic data in healthcare, particularly its effectiveness in preserving privacy while enabling data sharing for secondary purposes [72]. This study evaluated various synthetic data generation and usage settings, the necessity of tuning supervised machine learning models, and the predictive power of propensity scores for model accuracy. Propensity scores and classification accuracy were used as utility measures to understand optimal strategies for generating and using synthetic data. This research indicates that when real data are processed before synthesis, the utility of the synthetic dataset is not improved. However, applying tuning settings from real data to synthetic data models improves the accuracy. One significant limitation of this investigation is the specific testing strategies it has relied upon, which may not entirely capture the complexity involved in real-world applications.

Another study explored the application of generative artificial intelligence to produce synthetic data to accelerate research on hematologic neoplasms [73]. The goals of this study were to generate synthetic data for myelodysplastic syndromes (MDSs) and acute myeloid leukemia (AML), develop a framework for validation to assess data fidelity, confidentiality, and privacy preservation, and evaluate the potential applications of synthetic data for expediting clinical and translational research in hematology. Through conditional generative adversarial networks, synthetically generated cohorts were developed including clinical features, genomics, treatment, and outcomes for 7133 patients. The synthetic data performed well in terms of fidelity and privacy, thereby effectively solving problems associated with incomplete information through augmentation. A limitation of this study is its focus on specific hematological conditions, which may not be generalizable to other medical fields.

Another study investigated the use of generative AI models, such as generative adversarial networks (GANs) and variational autoencoders (VAEs), to create realistic and anonymized synthetic patient data [74]. These synthetic data aim to address the privacy and regulatory challenges associated with the use of real patient data in healthcare, adhering to standards such as HIPAA and GDPR. One limitation of this study is the focus on the potential of synthetic data without a detailed examination of the practical implementation challenges. However, this study could benefit from a more in-depth analysis of the current technical and ethical barriers to its widespread adoption.

A recent study attempted to address the challenge of distinguishing AI-generated images from real-life photos using computer vision techniques to improve this ability [75]. The researchers created a dataset called CIFAKE, which mimics the CIFAR 10 dataset by employing latent diffusion to produce similar images, framing the issue as a classification task. Distinguishing between AI-generated images, the study utilized a convolutional neural network (CNN) to achieve an impressive 92.98% accuracy rate after fine-tuning hyperparameters and training 36 different network structures. Despite the showcasing

accuracy, one limitation is its dependence on a dataset and method for generating synthetic images, which may not be universally applicable across various types of synthetic data. Future studies may investigate attention-based methods to enhance the classification accuracy and broaden the application to areas such as recognition in humans or medical imaging, thereby improving the versatility and effectiveness of AI-powered image detection systems.

## 5. Findings

Synthetic data generation has emerged as a crucial technique for addressing data scarcity, privacy concerns, and the need for diverse datasets in training machine learning models. Various synthetic data generation methods, such as generative adversarial networks (GAN), variational autoencoders (VAEs), and large language models (LLMs), offer unique capabilities and limitations tailored to different types of data, such as image, tabular, and textual data. This discussion aims to compare and draw conclusions from various studies that discuss these methods of synthetic data generation and answer research questions.

Many studies tend to focus narrowly on specific domains or applications, such as healthcare or financial services, thereby limiting the generalizability of their findings across different fields [76,77]. Additionally, methodological inconsistency is a prevalent issue, where the criteria for evaluating synthetic data quality and utility vary significantly between studies, making it difficult to draw consistent and comparative conclusions [78,79]. Another critical limitation is insufficient exploration of the ethical and privacy implications of synthetic data generation. Although some studies acknowledge these concerns, they often lack a deep analysis of the potential risks and mitigation strategies [40]. Moreover, many studies have not adequately addressed the scalability and practical implementation challenges associated with deploying synthetic data generation techniques in real-world settings, leading to a gap between theoretical research and practical application [80,81]. Most of the existing literature also tends to focus on the technical aspects of generative models, such as algorithmic improvements and performance metrics, without giving equal attention to their broader impact on data ecosystems and regulatory environments [82,83]. This narrow focus can result in an incomplete understanding of the broader implications of synthetic data generation technologies.

### **RQ1.** *What are the techniques and methods used for synthetic data generation?*

GANs, particularly cGANs and TGANs, have shown considerable promise in generating high-quality synthetic data. cGANs, which condition the generation process on additional information (such as class labels), excel in generating synthetic image data that retain realistic features and context [45]. This technique is particularly powerful for medical imaging, where cGAN can create realistic pictures of illnesses, thus contributing to the training of diagnostic models [40]. However, the computational cost is high for GANs, and it can be difficult to train them owing to problems such as the generation of a few variations in the dataset [39]. However, the TGANs were specifically designed to generate tabular data. TGANs seek to learn complex distributions over tables by exploiting their adversarial framework. This is achieved by capturing the feature dependencies within tables that help generate more complicated tabular data instances [38]. For example, preserving attribute relationships in healthcare data synthesis is essential for creating plausible patient records [78]. Nevertheless, large-scale tabular datasets pose a challenge to TGANs, owing to their computational requirements and sensitivity to hyperparameter tuning, which may adversely affect their performance during applications [37].

Another popular method for synthetic data generation is the variational autoencoder (VAE), which exhibits various conditional and adversarial variations. The fundamental principle of this approach is to encode the input data into a latent space and then decode them back to ensure that the generated data resemble the original data. Conditional VAEs, on the other hand, are more beneficial because they can be used to generate labeled data,

thus allowing for the controlled synthesis of certain classes or categories [58]. These are most commonly employed to generate synthetic images and tabular data with smooth latent spaces, leading to coherent samples across diverse dimensions. [6]. However, when compared to the GANs, VAEs usually generate blurred images because they depend heavily on reconstruction loss, which averages over pixel values. Therefore, these models may not be suitable for high-fidelity image generation applications [59]. Moreover, although VAEs can work well with continuous data types, they may not perform optimally on discrete data unless specific adjustments, such as discrete latent variable models, are applied [55].

LLMs such as GPT-3 and BERT represent significant advancements in the generation of synthetic textual data. GPT-3 uses a technique called autoregressive modeling, which must write the full sequence of words because it predicts each word based on the context generated by previous methods [64]. This feature makes GPT-3 particularly suited for applications that require coherent and contextually meaningful text generation, such as chatbots, content creation, and automated summarization [65]. Nonetheless, the need for large datasets and the many computation cycles required by GPT-3 are challenges in terms of accessibility and sustainability.

BERT stands out in terms of understanding text through bidirectional contextuality, leading to a better capture of word interdependence [66]. Owing to its deep contextual understanding, BERT is useful for tasks such as text classification, question answering, and sentiment analysis. Nevertheless, BERT is not capable of generating texts to the extent that GPT-3 is designed primarily for comprehension mechanisms related to text encoding [61].

#### **RQ2.** *What frameworks are available for synthetic data generation for specific use cases?*

When comparing these techniques, it is evident that each has strengths and limitations tailored to specific types of data.

**Image data.** For purposes of artificial image generation, cGANs and VAEs are the most efficient. In medical imaging and other contexts where realistic image synthesis is needed, cGANs offer high-quality, contextually appropriate images [40]. On the other hand, VAEs can produce somewhat blurrier pictures but possess robust performance in generating diverse and coherent image samples [6].

**Tabular data.** TGANs and VAEs are ideal for synthesizing synthetic tabular data. TGANs are effective at capturing intricate relationships between features, which are important for healthcare and financial data synthesis [78]. This area also has strong performance in VAEs, especially when combined with techniques that address discrete data [55].

**Textual data.** GPT-3 and BERT belong to the family of LLM models that are best suited for generating synthetic textual data. The ability of GPT-3 to generate coherent text that makes sense in its surroundings makes it highly applicable to natural language generation tasks [64]. BERT, on the other hand, has a deep contextual understanding; thus, it performs better than any other model in terms of text comprehension and generation tasks that require nuanced understanding, such as question answering or sentiment analysis [66].

**Speech data.** GAN-based models are employed for tasks such as voice conversion, which transforms one speaker's voice into another while retaining the linguistic content, making them useful in personalized voice systems. Google's AudioLM model enhances speech generation by modeling long-term dependencies in audio sequences, thereby enabling coherent and context-aware speech synthesis for conversational AI systems [84].

#### **RQ3.** *What are the limitations of using these techniques?*

The use of synthetic data generation techniques, including GANs, VAEs, and LLMs such as GPT-3 and BERT, has several limitations. GANs, such as conditional GANs (cGANs) and tabular GANs (TGANs), are particularly susceptible to mode collapse, where the model generates a limited variety of outputs and fails to capture the full diversity of input data. Additionally, GANs often suffer from training instability owing to the adversarial nature of the generator and discriminator networks, requiring careful tuning of the hyperparameters and training procedures to achieve stable convergence [39].



VAEs, on the other hand, tend to produce lower-quality synthetic data because they rely on reconstruction loss, which prioritizes generating data that closely resemble the input but often at the expense of finer details and overall quality. This issue is more pronounced when dealing with high-dimensional data such as images. Moreover, VAEs struggle with discrete data formats unless specifically adapted, thereby limiting their versatility across different types of datasets [59].

LLMs, such as GPT-3 and BERT are powerful but have significant drawbacks. Their high computational demands necessitate substantial processing power and memory, which not only increases the cost of deployment but also raises environmental concerns owing to the energy consumption associated with large-scale training. These models also require extensive datasets for effective training, posing challenges in terms of data availability and the ethical implications of data collection [64,66].

Furthermore, all of these models share common limitations, including potential biases in the synthetic data generated. If the training data are biased, the synthetic data will likely reflect these biases, potentially leading to unfair or discriminatory outcomes when used in real-world applications. Additionally, ensuring the privacy and security of synthetic data is crucial, as improper handling can lead to leakage of sensitive information.

Overall, although these techniques offer significant advancements in synthetic data generation, addressing their limitations through improved algorithms, better training protocols, and ethical considerations are essential for their effective and responsible use in various applications.

#### **RQ4.** *What are the areas of future research?*

The future of generative AI for synthetic data generation holds many promising research avenues, particularly for large language models (LLMs) such as GPT-3 and BERT. The ability of these models to produce sophisticated but synthetic textual data continues to improve, leading to new prospects for natural language understanding, automated content generation, and advanced conversational agents. However, significant challenges remain, and further research is required. An important area of research is ensuring privacy and security when using LLMs for pattern recognition and data synthesis. As a huge consumer of training data, LLMs may accidentally carry sensitive or personally identifiable information that can be extracted through model inversion attacks. For instance, further research should focus on developing robust methods for privacy-preserving data synthesis, such as differential privacy and federated learning, to ensure that the created data can be used safely. This study must also consider strategies for making large-scale LLM training and deployment less computationally intensive and environmentally damaging if such technologies are to become accessible and sustainable in the long term. Research addressing these issues will accelerate the full potential of LLMs.

### *5.1. Evaluation of Models*

A summarized evaluation of the effectiveness and applications is required for a clear and complete understanding of various data generation models.

#### *5.1.1. Evaluation of the Specific GAN Models*

- **Conditional GANs:** cGANs are highly effective in generating data with specific characteristics by conditioning the output on class labels or other variables. This makes them particularly powerful for tasks, such as generating labeled images, text, or data with specific features. cGANs excel in applications requiring control over the generated content, such as medical imaging, where they enhance the diagnostic quality and image segmentation.
- **Continuous conditional GANs:** CcGANs extend cGANs by allowing continuous scalar labels as inputs, which is crucial for tasks, such as regression-based image generation. They outperform traditional cGANs in generating high-quality, diverse images, especially in settings where precise control over continuous variables is required. They are

useful in fields such as autonomous driving for generating images with fine control over conditions such as steering angles and in medical imaging for tasks such as generating synthetic cell count images.

- Tabular GAN: TGANs are tailored to generate synthetic tabular data with discrete and continuous variables. They effectively maintain inter-column correlations, making them suitable for data science applications in which data relationships are critical. They are primarily used in fields requiring synthetic data generation for tabular datasets, such as finance, healthcare, and relational database modeling.
- Conditional tabular GANs: CTGANs improve TGANs by handling complex column distributions and by addressing the challenges of imbalanced training data. Their mode-specific normalization improves their ability to generate accurate tabular data, particularly in privacy-preserving contexts. CTGANs are ideal for generating synthetic data in privacy-sensitive industries, such as healthcare and financial services, where accurate and secure data are critical.
- Deep convolutional GANs: DCGANs enhance the basic GAN architecture by incorporating convolutional layers, improving the generation of high-quality images, and enabling improved feature extraction for unsupervised learning. They are widely used in tasks such as image generation, representation learning, and enhancement of the quality of synthesized images for various applications.
- CycleGAN: CycleGAN excels in unpaired image-to-image translation, using cycle consistency loss to ensure high-quality transformations between domains without the need for paired training data. It is particularly effective for tasks such as style transfer, object transfiguration, and photo enhancement, but struggles with tasks that require significant geometric transformations.

#### 5.1.2. Evaluation of Specific VAE Models

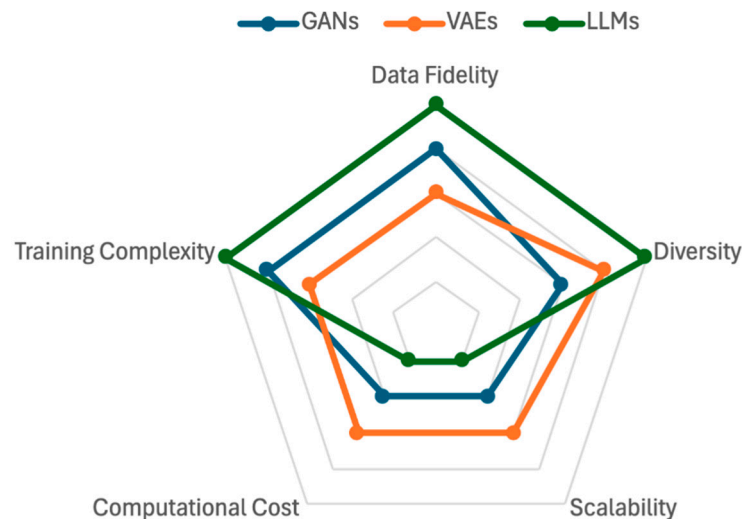
- Auto-Encoding Variational Bayes: AEVB efficiently handles continuous latent variables using the stochastic gradient variational Bayes (SGVB) estimator. This significantly improves variational inference, making it a powerful tool for analyzing large, complex datasets with continuous data. Widely used in probabilistic modeling, AEVB is applicable to tasks such as unsupervised learning, generative modeling, and anomaly detection.
- Conditional variational autoencoders: CVAEs are highly effective in structured output prediction by conditioning the generation process on specific input variables, leading to improved resilience in complex prediction tasks. CVAEs outperform deterministic models in tasks requiring probabilistic inference, such as image segmentation and semantic labeling.
- Adversarial autoencoders: AAEs combine the strengths of GANs and VAEs to align the latent space with an arbitrary prior, enhancing the generation of meaningful samples across the entire latent space. AAEs are particularly useful in generative modeling and semi-supervised classification tasks, making them effective in domains such as image synthesis and representation learning.
- Ladder Variational Autoencoders: LVAEs improve upon traditional VAEs by using a hierarchical structure that enhances the learning of latent representations across multiple layers, leading to a better generative performance. LVAEs are ideal for deep generative modeling tasks, particularly in hierarchical representation learning.
- Vector Quantized-Variational autoencoder: The VQ-VAE addresses the issue of “posterior collapse” in VAEs by using discrete latent variables, making it more effective for generating high-quality data, particularly in unsupervised learning scenarios. VQ-VAE is applied in tasks such as unsupervised phoneme learning, speaker conversion, and other tasks involving long-term sequence generation.

### 5.1.3. Evaluation of Specific LLM Models

- **GPT-1:** GPT-1 was a breakthrough in generating coherent and contextually relevant text, but its unidirectional nature limit its ability to understand long-range dependencies in text. This requires substantial fine-tuning and labeling of data for specific tasks. Useful for basic text generation, translation, and summarization, GPT-1 sets the foundation for future advancements in language models.
- **GPT-2:** GPT-2 was significantly improved over GPT-1 by increasing the model size and training data. It demonstrated strong performance in zero-shot tasks, such as summarization, translation, and question answering without requiring task-specific fine-tuning. GPT-2 is widely used in content creation, machine translation, and dialogue systems, although concerns remain about its potential misuse in generating misleading content.
- **GPT-3:** GPT-3 represents a major leap in scale with 175 billion parameters, allowing it to perform well in few-shot and zero-shot learning tasks. It excels in natural language understanding and generation, making it highly versatile across various tasks without fine-tuning. GPT-3 is used in advanced chatbots, content generation, and translation, demonstrating human-like text generation capabilities. However, it faces challenges with certain datasets and ethical concerns.
- **BERT:** BERT's bidirectional architecture enables it to understand the context from both the left and right, making it highly effective in a wide range of NLP tasks. It achieves state-of-the-art performance on many benchmarks, including question answering and sentence classification. BERT has been widely adopted for tasks such as sentiment analysis, text classification, and question answering because of its strong contextual understanding.
- **LaMDA:** LaMDA focuses on conversational AI and is designed to generate more contextually accurate and safe responses. It improves factual grounding by integrating external knowledge sources and addresses safety concerns by filtering biased or harmful content. Primarily used in dialogue systems and conversational agents, LaMDA enhances the quality and reliability of AI interactions in applications such as education and customer service.
- **XLNet:** XLNet improves on BERT by combining autoregressive and autoencoding approaches, allowing for better bidirectional context modeling without the limitations of masked language models. It consistently outperformed BERT on various NLP tasks. XLNet excels in tasks such as question answering and sentiment analysis, where understanding complex linguistic relationships is crucial.

To facilitate a clearer understanding of the strengths and limitations of various synthetic data generation techniques, a visual representation was included. This approach allows for a more intuitive comparison of the performance of each method across critical dimensions such as data fidelity, diversity, scalability, computational cost, and training complexity. After an extensive review and comparison of the relevant literature, scores were assigned to the above key metrics to ensure that the visuals accurately reflected the relative effectiveness and challenges associated with each technique. Figure 7 shows the performance comparison of the data generation techniques where the outermost and innermost axes represent the best and worst performance in the chosen dimensions, respectively.

From the visual perspective, we can conclude that LLMs are resource-consuming models that can be trained to specialize in a specific type of use case. For example, LLMs can be used in large industries to automate their workflow responsibly. However, GANs and VAEs are better choices for dynamic environmental settings with limited resources. For instance, GANs and VAEs can be used by start-ups or individuals for personal projects.



**Figure 7.** Performance of data generation techniques over various dimensions.

### 5.2. Real World Applications

Several advanced models have demonstrated significant potential for real-world applications in various fields. In healthcare, cGANs have been effectively used to improve the quality of medical imaging, particularly low-dose CT scans and MRIs, by generating high-quality synthetic medical images for diagnostic purposes. In the financial sector, VQ-VAE models are employed to generate synthetic financial data that preserve privacy while allowing for advanced analytics without exposing sensitive information [64]. GPT-3, with its massive scale, is utilized in content generation and customer support systems, allowing businesses to automate responses and generate human-like text in applications ranging from chatbots to reports. In finance, GPT-3 can be applied to generate financial reports and analyze sentiment in market trends by processing large amounts of unstructured text data. Moreover, CycleGAN has been used to enhance the visual effects in movies and games. For example, GANs can transform low-resolution images into high-resolution images, thereby improving the quality of animations and their special effects [65]. In gaming, these models help to create realistic environments by generating lifelike textures and characters.

Pre-trained language models such as GPT-3 and BERT offer plug-and-play capabilities, allowing developers to apply them to various tasks with minimal fine-tuning [85]. For instance, GPT-3 can perform content generation or customer service automation with minimal customization, making it relatively easy to implement. However, models, such as cGANs and VQ-VAEs often require substantial domain-specific customization, hyperparameter tuning, and large amounts of data for training, which can be a barrier to entry for businesses without deep technical expertise. Moreover, computational costs can be high, particularly for models such as GPT-3, which require significant infrastructure for deployment and training. Integrating these models into existing workflows can be complex because of issues such as data privacy, scalability, and ensuring model interpretability in sensitive sectors such as healthcare and finance.

Several strategies can be employed to overcome the challenges of deploying advanced AI models such as GANs, VAEs, and LLMs in real-world applications. Cloud-based platforms and AI as a Service (AIaaS) can mitigate high computational costs, allowing businesses to leverage powerful models like GPT-3 without investing in expensive infrastructure [85]. Automated machine learning (AutoML) tools can simplify hyperparameter tuning and model customization, thereby reducing the need for deep technical expertise [65]. Additionally, adopting privacy-preserving techniques, such as differential privacy and federated learning can address data privacy concerns, particularly in sensitive sectors such as healthcare.

Real-world data can often be biased or ethically problematic, reflecting historical inequalities or flawed collection process. When designing generative models such as

GANs, VAEs, or LLMs, it is crucial to address these concerns by implementing strategies that actively reduce bias and promote fairness. One method involves incorporating pre-processing techniques that modify the training data to ensure that they represent all demographic groups fairly. This can be achieved through re-sampling or re-weighting strategies, making the model less likely to perpetuate the biases present in the original data [86]. In-training bias mitigation techniques were employed during training. These involve adjusting the model parameters to prioritize fairness, such as penalizing biased predictions during training. For instance, fairness-aware learning algorithms have been used to ensure equality of opportunity in model outputs [17]. Furthermore, post-processing methods, that adjust the model outputs, can help mitigate the biases that persist after training. This is crucial in real-world applications, such as healthcare and finance, where biased outcomes can have significant negative consequences. Large language models (LLMs), which are particularly prone to perpetuating social biases owing to the vast and uncurated data they are trained on, combining techniques across all stages, from data augmentation to model fine-tuning, can effectively reduce bias [87].

### 5.3. Threats to Validity

Although this literature review provides a comprehensive analysis of various generative AI techniques for synthetic data generation, including GANs, VAEs, and LLMs, it is subject to several validity threats that could influence the robustness and generalizability of its conclusions.

#### 5.3.1. Internal Threats to Validity

1. Selection bias: The review primarily incorporated studies published in academic journals and conferences, potentially overlooking valuable insights from proprietary or unpublished industrial applications. This selection bias may have skewed the representation of the effectiveness and applicability of the techniques discussed.
2. Implementation fidelity: The discussion of the computational and resource demands for implementing generative AI techniques is based on theoretical analysis rather than empirical testing. This can lead to inaccurate assessments of the feasibility and efficiency of these methods in practical applications.

#### 5.3.2. External Threats to Validity

1. Technological evolution: As generative AI models continue to evolve, the techniques and methods discussed in this review may become increasingly outdated. Future advancements could introduce new capabilities or address the current limitations, thereby affecting the long-term validity of the findings.
2. Ethical and privacy considerations: Although ethical and privacy issues, particularly those related to LLMs, have been acknowledged, they have not been deeply explored. The responsible deployment of synthetic data generation technologies requires thorough ethical scrutiny, which is crucial for ensuring long-term societal acceptance and regulatory compliance of these technologies.
3. Interdisciplinary integration: The integration of synthetic data generation techniques with other emerging technologies, such as differential privacy and federated learning, is briefly mentioned but not extensively analyzed. This omission could overlook important synergies and potential advancements in enhancing the privacy and utility of the data.

## 6. Challenges and Future Research Directions

After examining several bases and modified research studies on the production of synthetic data using generative AI, we discovered certain challenges that arose during the process.

Below mentioned are the most critical challenges that need to be addressed.



- **Data fidelity and diversity:** Ensuring that synthetic data accurately represent the diversity and statistical properties of real-world data is a complex task. Overfitting or generating unrealistic data can limit the utility of synthetic data in training AI models and conducting research.
- **Bias and fairness:** Synthetic data generated from biased real-world datasets may perpetuate or amplify these biases, potentially leading to unfair and discriminatory AI models. Identifying and mitigating these biases is crucial for ensuring equitable outcomes.
- **Computational and environmental costs:** High computational power demands, particularly for models such as GPT-3 and BERT, not only increase operational costs but also raise environmental concerns owing to significant energy consumption.
- **Privacy and security:** Although synthetic data can help protect privacy by anonymizing data, ensuring that these data do not inadvertently leak sensitive information remains a challenge. Combining synthetic data generation with privacy-preserving techniques such as differential privacy and federated learning is necessary to enhance security.
- **Evaluation and validation:** Establishing robust metrics and frameworks for evaluating the quality of synthetic data is essential but challenging. Current evaluation methods often lack comprehensive empirical validation, which makes it difficult to assess the real-world applicability of synthetic data.
- **Practical implementation:** There is often a gap between theoretical advancements in generative AI and its practical implementation. Providing clear guidelines and best practices for deploying these models in real-world scenarios is crucial for effective application.

Addressing these challenges requires ongoing research and development to improve the robustness, scalability, and ethical considerations of synthetic data generation methods, ensuring that they can be effectively and responsibly used in various applications.

Future research in the domain of generative AI for synthetic data generation should address several key challenges to enhance its efficacy and applicability. First, advancing the robustness and scalability of generative models, such as GANs, VAEs, and LLMs is crucial. Research should focus on developing techniques to improve the fidelity and diversity of synthetic data, ensuring that they accurately mirror the statistical properties of real-world datasets without overfitting or generating unrealistic samples.

Additionally, integrating privacy-preserving methods such as differential privacy and federated learning with generative models can mitigate privacy concerns. As these models often require substantial datasets, it is critical to ensure that sensitive information is not compromised during training. Differential privacy techniques can introduce noise to the data, ensuring that individual data points remain confidential while still providing useful synthetic data for analysis. Federated learning can be employed to train models across multiple decentralized servers without transferring raw data, thereby maintaining data privacy.

Another area of future research involves enhancing the interpretability and usability of these models. Developing user-friendly tools and frameworks that allow practitioners to easily implement and customize synthetic data generation for specific use cases will facilitate broader adoption. Furthermore, establishing standardized evaluation metrics and benchmarks for synthetic data quality will help objectively assess and compare different generative approaches.

Lastly, expanding the application of synthetic data generation beyond current domains such as healthcare and finance to areas such as personalized medicine, public health surveillance, and education can provide significant societal benefits. By addressing these challenges, future research could significantly improve the reliability, applicability, and ethical considerations of synthetic data generation using generative AI models.

## 7. Conclusions

This literature review provides a comprehensive analysis of various synthetic data generation technologies, specifically GANs, VAEs, and LLMs. Each method has its strengths and challenges depending on the type of data it is intended for. GANs, including cGANs and TGANs, excel in generating high-quality images and complex tabular data but are computationally intensive and challenging to train. VAEs are effective for generating diverse and coherent data but often produce blurred images. LLMs such as GPT-3 and BERT are powerful tools for generating and understanding text but require significant computational resources. Nevertheless, there are still concerns regarding this progress, including computation requirements, stability in the learning process, and privacy-preserving measures that need to be considered while designing these technologies. Future research should focus on enhancing the scalability and efficiency of these methods while ensuring ethical and privacy standards. By addressing these problems, synthetic data generation methods can be adopted more widely across disciplines.

**Author Contributions:** Writing—original draft preparation: M.G.; supervision and writing—review and editing: Q.H.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Fan, J.; Han, F.; Liu, H. Challenges of Big Data Analysis. *Natl. Sci. Rev.* **2014**, *1*, 293–314. [CrossRef] [PubMed]
2. Fhom, H. Big Data: Opportunities and Privacy Challenges. In Proceedings of the International Conference on Information Systems and Management Science, Karlsruhe, Germany, 21–23 July 2015; p. 13. [CrossRef]
3. Poucin, F.; Kraus, A.; Simon, M. Synthetic data shows promising properties to boost the performance of Deep Neural Networks on real-world instance segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, BC, Canada, 11–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 945–953.
4. Abowd, J.M.; Vilhuber, L. How Protective Are Synthetic Data? In *Privacy in Statistical Databases*; Springer: Berlin/Heidelberg, Germany, 2008.
5. Jävergård, N.; Lyons, R.; Muntean, A.; Forsman, J. Preserving correlations: A Statistical Method for Generating Synthetic Data. *arXiv* **2024**, arXiv:2403.01471. [CrossRef]
6. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114.
7. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**. [CrossRef]
8. Radford, A.; Narasimhan, K. Improving Language Understanding by Generative Pre-Training. *OpenAI* **2018**. Available online: <https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf> (accessed on 28 August 2024).
9. Long, L.; Wang, R.; Xiao, R.; Zhao, J.; Ding, X.; Chen, G.; Wang, H. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. *arXiv* **2024**, arXiv:2406.15126.
10. Bauer, A.; Trapp, S.; Stenger, M.; Leppich, R.; Kounev, S.; Leznik, M.; Chard, K.; Foster, I. Comprehensive Exploration of Synthetic Data Generation: A Survey. *arXiv* **2024**, arXiv:2401.02524.
11. Hao, S.; Han, W.; Jiang, T.; Li, Y.; Wu, H.; Zhong, C.; Zhou, Z.; Tang, H. Synthetic Data in AI: Challenges, Applications, and Ethical Implications. *arXiv* **2024**, arXiv:2401.01629.
12. Sengar, S.S.; Hasan, A.B.; Kumar, S.; Carroll, F. Generative Artificial Intelligence: A Systematic Review and Applications. *arXiv* **2024**, arXiv:2405.11029. [CrossRef]
13. Sufi, F.K. Generative Pre-Tr78. Sufi, F.K. Generative Pre-Trained Transformer (GPT) in Research: A Systematic Review on Data Augmentation. *Information* **2024**, *15*, 99. [CrossRef]
14. Guo, X.; Chen, Y. Generative AI for Synthetic Data Generation: Methods, Challenges and the Future. *arXiv* **2024**, arXiv:2403.04190.
15. Lu, Y.; Wang, H.; Wei, W. Machine Learning for Synthetic Data Generation: A Review. *arXiv* **2023**, arXiv:2302.04062.
16. Bandi, A.; Adapa, P.V.; Kuchi, Y.E. The Power of Generative AI: A Review of Requirements, Models, Input-Output Formats, Evaluation Metrics, and Challenges. *Future Internet* **2023**, *15*, 260. [CrossRef]
17. Ippolito, D.; Ahn, J.; Cerqueira, J.F.; Huang, M.; Burgess, D. Bias and Fairness in Large Language Models: A Survey. *arXiv* **2023**, arXiv:2309.00770.
18. Eigenschink, P.; Reutterer, T.; Vamosi, S.; Vamosi, R.; Sun, C.; Kalcher, K. Deep Generative Models for Synthetic Data: A Survey. *IEEE Access* **2023**, *11*, 47304–47320. [CrossRef]

19. Fonseca, J.; Bação, F. Tabular and Latent Space Synthetic Data Generation: A Literature Review. *J. Big Data* **2023**, *10*, 1–37. [CrossRef]
20. Vargas, A.M.; Berlanga, A. A comprehensive review on synthetic data generation and its applications in medical imaging. *Neurocomputing* **2022**, *482*, 231–247.
21. Lu, Y.; Chen, D.; Olaniyi, E.O.; Huang, Y. Generative Adversarial Networks (GANs) for Image Augmentation in Agriculture: A Systematic Review. *Comput. Electron. Agric.* **2022**, *200*, 107208. [CrossRef]
22. Wang, S.; Du, Y.; Guo, X.; Pan, B.; Zhao, L. Controllable Data Generation by Deep Learning: A Review. *ACM Comput. Surv.* **2022**, *56*, 1–38. [CrossRef]
23. Figueira, Á.; Vaz, B. Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics* **2022**, *10*, 2733. [CrossRef]
24. Kitchenham, B.; Charters, S. Guidelines for Performing Systematic Literature Reviews in Software Engineering. *Engineering* **2007**, *2*, 1051.
25. Petersen, K.; Feldt, R.; Mujtaba, S.; Mattsson, M. Systematic Mapping Studies in Software Engineering. In Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, Bari, Italy, 26–27 June 2008; pp. 68–77.
26. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; Antes, G.; Atkins, D.; Barbour, V.; Barrowman, N.; Berlin, J.A. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med.* **2009**, *6*, e1000097. [CrossRef] [PubMed]
27. Syntheticus.ai. Guide: Everything You Need to Know about Synthetic Data. Available online: <https://syntheticus.ai/guide-everything-you-need-to-know-about-synthetic-data> (accessed on 20 May 2024).
28. DataRobot. What are Parametric Models? Available online: <https://www.datarobot.com/blog/what-are-parametric-models/> (accessed on 20 May 2024).
29. DeepAI. Non-Parametric Model. Available online: <https://deepai.org/machine-learning-glossary-and-terms/non-parametric-model> (accessed on 20 May 2024).
30. Singh, A. Protecting your Data Privacy with Differential Privacy: An Introduction. Available online: <https://medium.com/dsaid-govtech/protecting-your-data-privacy-with-differential-privacy-an-introduction-abee1d7fcb63> (accessed on 22 May 2024).
31. Bossert, J.; Lütjen, M.; Kanoun, O. Context-adaptive and activity-aware physical analysis for people with dementia. In *Ambient Assisted Living*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 161–168.
32. Islam, S.; Aziz, M.T.; Nabil, H.R.; Jim, J.R.; Mridha, M.F.; Kabir, M.M.; Asai, N.; Shin, J. Generative Adversarial Networks (GANs) in Medical Imaging: Advancements, Applications, and Challenges. *IEEE Access* **2024**, *12*, 35728–35753. [CrossRef]
33. Strelcenia, E.; Prakoonwit, S. Improving Cancer Detection Classification Performance Using GANs in Breast Cancer Data. *IEEE Access* **2023**, *11*, 71594–71615. [CrossRef]
34. Ali, H.; Grönlund, C.; Shah, Z. Leveraging GANs for Data Scarcity of COVID-19: Beyond the Hype. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023; pp. 659–667.
35. Yadav, P.; Gaur, M.; Fatima, N.; Sarwar, S. Qualitative and Quantitative Evaluation of Multivariate Time-Series Synthetic Data Generated Using MTS-TGAN: A Novel Approach. *Appl. Sci.* **2023**, *13*, 4136. [CrossRef]
36. Charitou, C.; Dragicevic, S.; Garcez, A.S. Synthetic Data Generation for Fraud Detection using GANs. *arXiv* **2021**, arXiv:2109.12546.
37. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. CTGAN: Synthesizing Tabular Data Using Conditional GANs. *Neural Inf. Process. Syst.* **2019**. Available online: <https://hdl.handle.net/1721.1/128349> (accessed on 28 August 2024).
38. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling Tabular Data Using Conditional GAN. *arXiv* **2019**, arXiv:1907.00503. [CrossRef]
39. Miyato, T.; Koyama, M. cGANs with Projection Discriminator. *arXiv* **2018**, arXiv:1802.05637.
40. Xie, L.; Lin, K.; Wang, S.; Wang, F.; Zhou, J. Differentially Private Generative Adversarial Network. *arXiv* **2018**, arXiv:1802.06739.
41. Zhu, J.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2242–2251.
42. Chen, X.; Duan, Y.; Houthoofd, R.; Schulman, J.; Sutskever, I.; Abbeel, P. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. *arXiv* **2016**, arXiv:1606.03657. [CrossRef]
43. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
44. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* **2015**, arXiv:1511.06434.
45. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784.
46. PyTorch. DCGAN Tutorial: Generate Faces Using Deep Convolutional GAN. Available online: [https://pytorch.org/tutorials/beginner/dcgan\\_faces\\_tutorial.html](https://pytorch.org/tutorials/beginner/dcgan_faces_tutorial.html) (accessed on 30 May 2024).
47. TensorFlow. CycleGAN Tutorial: Using Cycle-Consistent Adversarial Networks for Unpaired Image-to-Image Translation. Available online: <https://www.tensorflow.org/tutorials/generative/cyclegan> (accessed on 2 June 2024).
48. Mostofi, F.; Tokdemir, O.B.; Toğan, V. Generating Synthetic Data with Variational Autoencoder to Address Class Imbalance of Graph Attention Network Prediction Model for Construction Management. *Adv. Eng. Inform.* **2024**, *62*, 102606. [CrossRef]
49. Wu, J.; Plataniotis, K.N.; Liu, L.; Amjadian, E.; Lawryshyn, Y.A. Interpretation for Variational Autoencoder Used to Generate Financial Synthetic Tabular Data. *Algorithms* **2023**, *16*, 121. [CrossRef]

50. Li, H.; Yu, S.; Principe, J.C. Causal Recurrent Variational Autoencoder for Medical Time Series Generation. *arXiv* **2023**, arXiv:2301.06574. [\[CrossRef\]](#)
51. Saldanha, J.; Chakraborty, S.; Patil, S.A.; Kotecha, K.V.; Kumar, S.; Nayyar, A. Data Augmentation Using Variational Autoencoders for Improvement of Respiratory Disease Classification. *PLoS ONE* **2022**, *17*, e0266467. [\[CrossRef\]](#)
52. Kok, S.; Vardhan, L.V. Generating Privacy-Preserving Synthetic Tabular Data Using Oblivious Variational Autoencoders. In Proceedings of the Workshop on Economics of Privacy and Data Labor at the 37th International Conference on Machine Learning, Cambridge, UK, 18 July 2020.
53. Islam, Z.; Abdel-Aty, M.A.; Cai, Q.; Yuan, J. Crash Data Augmentation Using Variational Autoencoder. *Accid. Anal. Prev.* **2020**, *151*, 105950. [\[CrossRef\]](#)
54. Goyal, P.; Sapienza, M.; Sun, C. Self-Supervised Video Representation Learning with Contrastive Predictive Coding. *arxiv* **2019**, arXiv:1905.09272. [\[CrossRef\]](#)
55. van den Oord, A.; Vinyals, O.; Kavukcuoglu, K. Neural Discrete Representation Learning. *arxiv* **2017**, arXiv:1711.00937. [\[CrossRef\]](#)
56. Kipf, T.; Welling, M. Variational Graph Auto-Encoders. *arXiv* **2016**, arXiv:1611.07308.
57. Rasmus, A.; Berglund, M.; Honkala, M.; Valpola, H.; Raiko, T. Semi-Supervised Learning with Ladder Networks. *arXiv* **2015**, arXiv:1507.02672. [\[CrossRef\]](#)
58. Sohn, K.; Lee, H.; Yan, X. Learning Structured Output Representation using Deep Conditional Generative Models. *arXiv* **2015**, arXiv:1506.05517. [\[CrossRef\]](#)
59. Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.J. Adversarial Autoencoders. *arXiv* **2015**, arXiv:1511.05644.
60. Li, Z.; Zhu, H.; Lu, Z.; Yin, M. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. *arXiv* **2023**, arXiv:2310.07849.
61. Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.M.; Kulshreshtha, A.; Cheng, H.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. LaMDA: Language Models for Dialog Applications. *arXiv* **2022**, arXiv:2201.08239.
62. Meng, Y.; Huang, J.; Zhang, Y.; Han, J. Generating Training Data with Language Models: Towards Zero-Shot Language Understanding. *arXiv* **2022**, arXiv:2202.04538.
63. Austin, J.; Odena, A.; Nye, M.; Bosma, M.; Michalewski, H.; Dohan, D.; Jiang, E.; Cai, C.J.; Terry, M.; Le, Q.; et al. Program Synthesis with Large Language Models. *arXiv* **2021**, arXiv:2108.07732.
64. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
65. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *OpenAI* **2019**. Available online: [https://d4mucfpxsywv.cloudfront.net/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://d4mucfpxsywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf) (accessed on 28 August 2024).
66. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Proceedings of the 2019 North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019.
67. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.G.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arxiv* **2019**, arXiv:1906.08237. [\[CrossRef\]](#)
68. Dahmen, J.; Cook, D. SynSys: A Synthetic Data Generation System for Healthcare Applications. *Sensors* **2019**, *19*, 1181. [\[CrossRef\]](#) [\[PubMed\]](#)
69. Google. TF-GAN: A Generative Adversarial Networks Library for TensorFlow. Available online: <https://www.tensorflow.org/tutorials/generative/tf-gan> (accessed on 10 June 2024).
70. Gretel.ai. Gretel Synthesis: Generate Synthetic Data with Enhanced Privacy Features. Available online: <https://gretel.ai> (accessed on 15 June 2024).
71. Ping, H.; Stoyanovich, J.; Howe, B. DataSynthesizer: Privacy-preserving synthetic datasets. *arXiv* **2017**. [\[CrossRef\]](#)
72. Dankar, F.K.; Ibrahim, M.M. Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. *Appl. Sci.* **2021**, *11*, 2158. [\[CrossRef\]](#)
73. D’Amico, S.; Dall’Olio, D.; Sala, C.; Dall’Olio, L.; Sauta, E.; Zampini, M.; Asti, G.; Lanino, L.; Maggioni, G.; Campagna, A.; et al. Synthetic Data Generation by Artificial Intelligence to Accelerate Research and Precision Medicine in Hematology. *JCO Clin. Cancer Inform.* **2023**, *7*, e2300021. [\[CrossRef\]](#) [\[PubMed\]](#)
74. Jadon, A.; Kumar, S. Leveraging Generative AI Models for Synthetic Data Generation in Healthcare: Balancing Research and Privacy. In Proceedings of the 2023 International Conference on Smart Applications, Communications and Networking (SmartNets), İstanbul, Türkiye, 25–27 July 2023; pp. 1–4.
75. Bird, J.J.; Lotfi, A. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *IEEE Access* **2023**, *12*, 15642–15650. [\[CrossRef\]](#)
76. Patki, N.; Wedge, R.; Veeramachaneni, K. The synthetic data vault. In Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 17–19 October 2008; IEEE: Piscataway, NJ, USA, 2016.
77. Beaulieu-Jones, B.K.; Wu, Z.S.; Williams, C.; Greene, C.S. Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. *bioRxiv* **2019**. [\[CrossRef\]](#)

78. Choi, E.; Biswal, S.; Malin, B.A.; Duke, J.D.; Stewart, W.F.; Sun, J. Generating Multi-Label Discrete Patient Records using Generative Adversarial Networks. In Proceedings of the 2nd Machine Learning for Healthcare Conference, Boston, MA, USA, 18–19 August 2017.
79. Frid-Adar, M.; Klang, E.; Amitai, M.; Goldberger, J.; Greenspan, H. Synthetic Data Augmentation Using GAN for Improved Liver Lesion Classification. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018.
80. Yi, X.; Walia, E.; Babyn, P. Generative adversarial network in medical imaging: A review. *Med. Image Anal.* **2019**, *58*, 101552. [CrossRef]
81. Esteban, C.; Hyland, S.L.; Rätsch, G. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. *arXiv* **2017**, arXiv:1706.02633.
82. Xu, L.; Veeramachaneni, K. Synthesizing Tabular Data Using Generative Adversarial Networks. *arXiv* **2018**, arXiv:1811.11264.
83. Montenegro, H.; Silva, W.; Cardoso, J. Privacy-Preserving Generative Adversarial Network for Case-Based Explainability in Medical Image Analysis. *IEEE Access* **2021**, *9*, 148037–148047. [CrossRef]
84. Recent Developments in Generative AI for Audio. Available online: <https://www.assemblyai.com/blog/recent-developments-in-generative-ai-for-audio/> (accessed on 17 August 2024).
85. Bao, H.; Dong, L.; Wei, F. UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training. *arXiv* **2019**, arXiv:2002.12804.
86. Magda, N.; Maciej, M.; Michal, P.; Tomasz, T.; Michal, W. Federated Learning Methods for Combating Attacks and Improving Privacy in IoT Networks. In *Computational Intelligence*; Springer: Cham, Switzerland, 2021; pp. 487–500.
87. Auditing Bias in Large Language Models. Available online: <https://insights.sei.cmu.edu/blog/auditing-bias-in-large-language-models/> (accessed on 17 August 2024).

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.