JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Generative AI-Based Synthetic Data Generation for Addressing Data Deficiency

[1] **Mallikarjunarao Sunke,**
[1] Individual Researcher,
4821 San Mateo LN, NE, Albuquerque, NM, 87109.
sunkemallikarjunarao73@gmail.com

Abstract:
The lack of data is extremely difficult to deal with when training an Artificial Intelligence (AI) model and will be interesting to see how this problem can be solved in the future. Currently, AI is transforming more and more industry and data scarcity is a major challenge. This research paper discusses important implications with respect to lack of data in AI training, its potential and growth and suggests possible outcomes. Additionally, this study also highlights various ethical implications like consent, privacy, and unbiased laws about deploying AI models under monitored conditions. In parallel, innovative solutions are investigated on aspects that need to be adopted with few-shot learning, transfer learning and data augmentation in order to adapt model to the efficient processes for the utilization of AI with limited resources. Therefore, it aims to work on the best methods and team solutions to make it fair and applicable to address ethical and technical problems of data scarcity. This study also explores potential approaches when addressing data scarcity with the blend of traditional and synthetic models along with using modern ML techniques like few-shot learning and transfer learning. These methods improve efficiency and flexibility of AI models in different sectors to develop sustainable technology amid existing issues of lack of natural data.

Keywords: Data Scarcity, AI Models, AI Training, Synthetic Data, Generative AI Collaborative Methods.

1. Introduction

With the progress of AI, several industries have been upgraded and futuristic. Constant enhancements bring some issues, which can impact on the development of any technology. One of the most serious issues is lack of data information, particularly data quality. Basically, data is collected from real-world, which goes to AI model for training, adaptation, and decision-making. With all the data available in this day and age, it becomes a new currency. There's more demand than there's sources that are diverse and authentic. There are high stakes involved in modern AI development due to lack of pure, clean, and natural data which comes from direct human experiences and interactions worldwide (Singh, 2021; Panagiotou et al, 2024; Danahar and Nyholm, 2024). There are exponential effects of the lack of data, including accuracy and efficacy of AI, as well as ethical implications. In addition to letters and numbers, AI can also categorize data as behavioural patterns from huge information (Bai et al, 2022). Given such an dire situation, a substantial question comes up. Will AI be limited with lesser accuracy or less information? While technically, it's not the data problem today (Nandy et al, 2022), AI could be in existential crisis in all the sectors given the scarcity of the data. Yet, it impacts the interface between the organization and AI programs employed for enhancement and performance (Villalobos et al, 2022).
As learning from data is being a concern, future of AI is at stake. This limitation could impact the use and effectiveness of AI systems across different applications, as they heavily depend on large amounts of data for predictions and learning. The issue is prevalent in various sectors such as healthcare, where reliability and accuracy are crucial for AI solutions. In some areas, for example, the use of small datasets has been shown to be effective for a number of AI applications in healthcare, where the accuracies achieved in clinical

environments have been demonstrated (Babbar and Schölkopf, 2019). This way, AI becomes less generalized and robust, which requires novel approaches to address challenges related to limited data. For instance, small datasets have been used for several applications, which has been observed in clinical environments (Babbar and Schölkopf, 2019). The result is that, AI models are not so generalized and powerful, and new solutions are needed to tackle the challenges of limited data.

In addition, the ethical question of creating AI models is extreme and inescapable with the scarcity of data. Issues of consent, privacy, and bias arise where there are demands for large amounts of data and reliance on data for machine learning (ML) to replicate existing inequities (Singh, 2021). Hence, organizations have to deal with socio-technical concerns, which promote moral and robust AI programs (Babbar and Schölkopf, 2019). Many stakeholders should work together from different domains to form regulations and ethical norms to set responsibility and accountability for generating and implementing AI models (Danahar and Nyholm, 2024).

Apart from legal and ethical issues, AI is also dependent on data engineering innovations and the management of data models throughout their lifecycle. In addition to ethical considerations, AI also depends on data engineering innovations and lifecycle management of data models. Organizations must be capable to update their AI programs with the evolution of data to make effective and relevant predictions (Ramakrishnan, 2023). Moreover, they can better leverage data and enhance the productivity of AI by integration with other technologies such as optimizing intelligence and machine learning (Panagiotou et al, 2024). In addition, the convergence of AI like smart augmentation and ML can better use data and improve productivity of AI (Panagiotou et al, 2024). Stakeholders can offset the effects of data scarcity and realize the full potential of AI in different sectors by cross-generating ideas across different areas of science (Chowdhery et al, 2023). With the increasing significance of AI in all areas, this paper addresses the problem of data scarcity as one of the main hurdles for the development of AI. This study also discusses the challenges of data shortage and how it affects the ability of AI to adapt, perform, and generalize in real-world. Last, this study calls for a hybrid strategy to incorporate synthetic data into existing contexts, much like cross-indexing and moral frameworks do in various fields, in order for the AI programs to function effectively on a limited resource basis.

2. Related Work

The efficiency of AI application in various fields such as cybersecurity, finance, and health care etc. depends on high-quality and clean data, and larger datasets are even more significant as basic data can impact decision making. On the other hand, strong datasets may reveal patterns which have been unnoticed and drive innovation. In the healthcare sector, for example, there is a need to analyze large records for advances in predicting disease, personalized medicine, and the evaluation of treatment. Big data analysis could also be used for fraud detection, risk assessment and algorithmic trading for making financial transactions. Furthermore, big data about network traffic can be used to detect abnormal traffic while cybersecurity can provide an additional layer of protection against attacks. Moreover, analyzing a huge amount of data from these sensors can be used in traffic control, infrastructure and sustainable plans in urban planning. Big data also promotes innovation, advancement, and informed decision-making in this day and age around different areas (Goyal and Mahmoud, 2024).

Table 1 brings attention to the existing research efforts on data scarcity across various aspects of AI to highlight current efforts to tackle this challenge.

Authors	**Highlights**	**Key findings**
Hausen and Azarbonyad, 2024; Babbar and Schölkopf, 2019	To test multiple data augmentation techniques for machine learning due to insufficient data.	Rewriting the formula for sophisticated approach to simulate variability in real world.
Chen et al (2025)	Suggesting methods such as collaborative filtering and content-based to enhance recommendation systems in cases of limited information.	Highlighting impact of those strategies to improve diversity of recommendation and user engagement.
Khalil et al (2025)	Generating synthetic data and transfer learning to adapt technicalities to operate in the face of data shortage.	Presenting strong approaches for tackling various problems of scarcity

White et al (2022)	Using rejection systems to enhance reliability of decisions under limited resources in ML deployments	The need for mechanisms to ensure reliability of critical applications.
Singh (2023)	Discuss data-centric and traditional approaches to cope with data shortage.	Identification of critical areas, where innovations would be useful for the advancement of AI.
Li et al (2022)	Discussing the need for an affordable mirror environment to simulate real-world scenarios for AI development.Talking about creating an affordable mirror environment to simulate real-world scenarios for the development of AI.	These environments can enhance the functioning and reliability of AI systems.
Lautrup et al, 2025; Gangwal et al, 2024	Adopting AI models to integrate different sources of data	Drug safety challenge due to limited resources for identification of challenges.
Niel (2023)	Propose novel model for generating synthetic data for training ML model	The accuracy in diagnosis and access to training data have also improved.
Chen et al (2021)	Deep generative models used for synthesizing private data to address both scarcity and privacy issues	In addressing scarcity and privacy concerns in AI, these models have the potential to play a crucial role.The models could be easily used to solve scarcity and privacy related issues in AI.
Bansal et al, 2022; Nahid and Hasan, 2024	Exploring deep generative models for creating private data to enhance classifier under privacy limitations	An enhancement of the performance of the classifiers by ensuring privacy and diversifying data simultaneously.
Zou et al (2022)	Using “Bidirectional Encoder Representations from Transformers (BERT)” model to generate written data as per relevance to improve availability of data.	Improving the relevance and availability of data to support overall performance.
Van (2023)	Enhance performance and training LLMs using data augmentation techniques.Boost the performance and training of LLMs using data augmentation techniques.	These are methods to reduce effects of data scarcity for training ML models.
Alzubaidi et al (2023)	In-depth examination of the use of deep learning tools for scarcity of data on various strategies	Introduction to the efficiency and possible future potential of those tools for addressing data scarcity. The problem of "data scarcity" in political communication research.
Laurer et al (2024)	Proposed adaptations in methodology are included, but there are no specific findings.	Synthetic profiles will be developed as a means to improve the availability of data for AI applications.
Quito and Andrade (2022)	Benefits of synthetic data to enhance data availability	Talking about how little data is available around the world for the development of AI
Oxford Analytica (2024)	Emphasized problems encountered by developers in the context of data scarcity, and its impact.	A heuristic training model for enhancing the ability of “Resource Constrained Training

		(RCT)”

3. Challenges in Generating Natural and Clean Data for AI Training

AI models have been trained on larger datasets, which have led to the development of powerful models like DALL-E 3 and ChatGPT. At the same time, findings suggest that the rise of stocks is slower than that of datasets in training AI models. It is predicted in studies that current trend of AI training continues to develop high-quality textual information by 2026. By 2050 it is expected that low-quality data will be consumed and by 2060, low-quality images will be exhausted. As per PwC, AI can contribute to US$15.7 trillion to the economy globally by 2030. It might be hindered due to a lack of useful information (Abdalla et al, 2025).
AI has a few socio-economic and ethical implications. If the data is not clean and natural, the AI model won't be able to know the changes and stay enslaved in its AI data bubble. For example, LLMs must be continuously fed relevant and varied information to maintain their accuracy and relevance. Otherwise, their outputs might not be accurate/obsolete. Additionally, AI programs may also be biased without ample natural information. Since their models are trained on similar or incomplete datasets, they repeat and learn the data biases, such as, resulting in unfair or biased results. An example could be AI programmes in recruitment, leading to some numbers and a lack of diverse information about a particular group, thus contributing to social imbalance. There could be devastating biases of these effects as AI programs become the core of various processes related to decision-making, i.e., from verdicts of the court to healthcare suggestions (Giuffrè and Shung, 2023).
The lack of data also boosts financial costs. Incorporating AI into their core systems could pose a significant challenge for many companies, as they rely on it for cost savings, efficient operations, or enhancing customer experiences. They find it challenging without knowing the origin of data. Indeed, it could affect industrial deployment of AI and technological advances can improve economic growth. Table 2 highlights some of the major challenges of data scarcity like risk of bias, data exhaustion, cost, and regulatory limitations in AI and their possible implications on fairness, innovation and performance of model.

Table 2 – Data Scarcity Challenges in AI

Challenges	Descriptions	Possible Effects
Increasing risk of bias	Diversity of limited data can strengthen existing AI biases	Biased decisions in industries like law, hiring, and healthcare.
Data exhaustion	Depletion of natural image and text data projected by 2050	Slower deployment of AI model, stagnant performance, and lower innovation.
Regulatory limitations	Strict regulations affect access to real-world, quality data	Scarcity of clear data affecting fairness and accuracy in AI models
Collection cost	High cost of data collection is related to curating, sourcing, and annotation	Increased operation costs and slowdown of deploying AI projects
Imbalanced resources	Smaller organizations may find it challenging with limited access to data than resource-rich, larger companies	Potentially monopolizes advances in AI by well-funded organizations

3.1 Availability of Synthetic Data

To date, only a limited amount of natural and clean data is available, possibly synthetic data would be enough. Fake data is fake but looks like the real data, and can be created in an infinite variety. There are however, certain issues and dangers in using artificial data. Nuanced problem issues that are encountered by humans in communications cannot be simulated by synthetic data. AI Models work very well with problems in controlled environments, but are not so good in real life and chaotic environments. Synthetic data is more likely to reinforce existing biases as it can be produced in a manner that goes beyond or is drawn from existing patterns of data, which are likely to also exclude certain groups of users. Furthermore, there are ethical issues associated with synthetic data. AI can create data based on the assumption that it is correct with self-trained model. It's the main implication in machine learning. The idea of AI training AI is not precisely defined and therefore could be hard to track and/or impossible to control. So there is a lack of accuracy and authenticity of the data (Shukla et al, 2024). The biases will be neutralized at the end, and turn into AI programs by synthetic data. When things are extremely important, it can be devastating if it isn't accurate. Enhancement of related affordability, access, scale and control on bias, ethics is presented for Table 3 where challenges are noted in relation to the scarcity of data that can be addressed with synthetic data, risk of bias, lack of authenticity and ethical issues that need to be addressed.

Table 3 - Solving Data Scarcity with synthetic data

Aspects	Benefits	Challenges
Cost	Lower cost related to data annotation and collection.	There may be need for additional validation due to quality issues.
Accessibility	With synthetic data, data scarcity can be overcome.	There is a lack of authenticity in real-world in generated data.
Bias	Diversity of dataset can be improved by tailoring synthetic data.	If data sources are biased, there can be risk of bias in the data.
Ethics	Data that is collected in real world does not suffer from privacy concerns.	Not grounded in a real-world context, and therefore ambiguous in terms of ethics.
Scalability	Easy to generate big data for training.	A feedback loop may be created by depending excessively on risks of synthetic data in ML.

3.2 Privacy and Ethical Issues

Scarcity of natural, clean data is also a problem with ethical and privacy concerns. Data generated by humans should be used and gathered, usually at the cost of sensitive validation with major implications for end-user privacy. Scandalous cases, such as the Cambridge Analytica one, have come into public light and internet culture across the countries that involve data privacy. It has left very little room for site owners or vendors and severely regulates without their knowledge. There are regulations that differ substantially between countries that AI organizations must contend with, and of course, desire for more data and the need for private age (Khapra 2024). The information gathered for AI training does not involve a direct study or acquisitions of permissions, but there can be ethical issues (Khapra, 2022). More powerful models are required, particularly concerning reiterating the importance of ethical approaches and data collection methods concerning AI standards (Table 4).

Table 4 – Privacy and Ethical concerns of data usage by AI

Ethical issues	Description	Need for AI development
Biases	Bias reduction could result in unfair outcome or biases	AI applications ought to be beneficial to everyone equally.
Data Privacy	Use of data and data collection are respects of personal rights and adhere to privacy legislation.	Creating trust between the public and AI programs, to prevent legal problems.
Accountability	The need for ethical AI outcomes in sensitive areas such as healthcare.Ethics for AI outcomes in sensitive sectors such as healthcare.	Minimizing the risk of harm from errors and/or skewed results
Transparency	Clarity on the training models and sources of data.	Encourages Responsibility and Trust in AI Systems
Consent	The actions of the community must be taken in consideration and the consent of the people must be granted.	Raises social AI acceptance and aligning deployment of AI with values from society

4. Data Scarcity for AI

As AI models widely depend on a huge volume of data to predict and gather knowledge, its effectiveness and applicability is highly affected by data scarcity. Reliability and accuracy, especially in the healthcare industry, are paramount in the accuracy of AI solutions. In many of these applications the model is not as general and powerful as in others, e.g. Babbar and Schölkopf (2019) and/or the data set is very small, where the accuracy of the models is not as high as in larger data sets. This is because there is a lack of data, and new techniques will need to be developed to be able to respond. In addition, the consequences of the loss of data in case of the production of AI is huge and needs to be avoided.

Alongside the necessity for big data, data dependency in the field of machine learning, worries about consent, bias and privacy issues are raised, which leads to theoretically generated disparities (Chowdhery et al., 2023). The moral and robust AI programs are thus very significant, hence the need to address those problems (socio-technical) by these organizations. The establishment of regulations and ethical norms to govern the responsible adoption and creation of AI (Danaher and Nyholm, 2024) was a combined effort that required a number of stakeholders to contribute in various industries. The small amount of data isn't as important as the future of so called data engineering, and AI lifecycle practices on ethical issues of AI.

The companies need to become more capable to update their AI programs more relevant as data grows and make them more effective in the long-term (Ramakrishnan, 2023). Secondly, the ability of AI to operate using machine learning and augmentation could lead to better data utilization, and productivity of AI (Panagiotou et al, 2025). Cross fertilization can help with important issues of inadequate data availability and help to realize the full potential of AI in various field of science.

4.1 Natural Data for AI is critical

For AI model training and to guarantee its operational reliability in an impulsive environment, it is crucial to store and manage the clean and unprocessed (raw and natural) data (Bai et al, 2022). The details, hypotheses and mistakes are the perfect type of data for AI to tune its various scenarios and make accurate predictions. Meanwhile, processed data is generated artificially to imitate data from the real-world and it usually doesn't have the richness or authenticity needed to create reliable and robust AI. Any human interaction and behaviour is considered 'natural data'. For example, typo, regional slangs or changing of syntaxes are minor language errors that can only be attained in the presence of different language (Huang et al, 2024).

Unfortunately, it is not possible to draw generalizations about the model training (based on the synthetic data) for other contexts and/or other demography that could lead to inefficient or biased model results. This is why it is essential to have natural data to develop powerful AI programs to service and engage with users to solve the big issue of the AI's growth. In today's digital age, there is a growing need for Artificial Intelligence (AI) and data to be part of the innovation process in various sectors. Artificial Intelligence (AI) and data are key components in promoting innovation in different industries in this digital era (Figure 1). The adoption of the online platforms and devices has helped to enhance the quality of data and has shot up spectacularly especially when assisted by artificial intelligence.

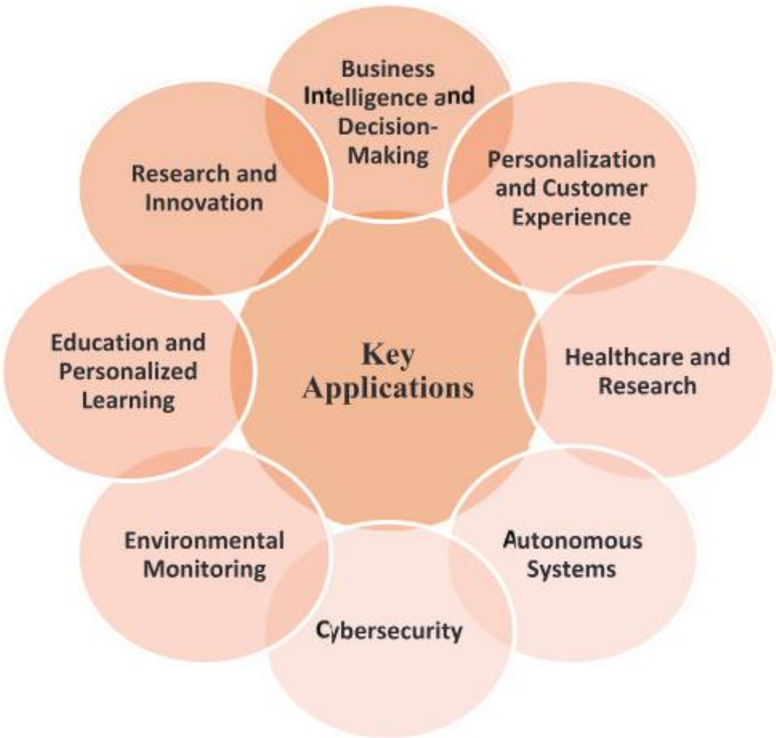

Figure 1 – Key applications of AI and data

Source – Abdalla et al (2025)

4.2 The more and more requirement for raw data

As AI is used in a variety of applications from diagnosis to product suggestions, big data is in demand. To perform their best, Large Language Models (LLMs) must be able to order billions of images and words, to boost reliability and accuracy as in the case of GPT-4 (Lynch, 2024). There are various applications in various fields which enhance the demand. The obvious ones are those in the worlds of Healthcare, E-commerce and Finance where AI analytics can prove to be the most useful tool to provide optimum trends (Lynch, 2024). Until now, it's the demand they have to continue to rely on.

It can also put a strain on the natural data sources too, as the need for natural data continues to outpace that source as organizations are getting better at using AI than their competitors, and eventually it can help to slow the growth of AI. Because of the ever-evolving nature of the real world data, many organizations are seeking a replacement to create and test new analytics – that's where synthetic data comes in. (Healy et al, 2025) Lack of right sources for real-world data can minimize the reliability of AI and ethical integrity. In this fashion, they won't be in any way a big impact to the huge industries. Table 5 illustrates the increase in the demand for AI over the last few years.

The field of Artificial Intelligence (AI) was a very popular field for the past couple of years as can be seen in table 5.

Areas	**Key trends and highlights**	**References**
Retail	Big data collects market status and consumer trends	Abdalla and Abuhaija (2023)
Finance	Risk assessment and fraud detection	Abdalla et al (2024)
Manufacturing	Optimizing productivity and supply chain with big data	Abdalla et al (2024)
Healthcare	Predictive care and personalized medications	Abdalla (2022)
ML	There is 54% rise in logged models and whopping 411% rise in registered models since 2022	Databricks (2024)
Gen AI	There is a rise of $25.2 billion worth of investments in 2023	Databricks (2024)
Overall	AI has developed over 72% of all basic models	Lynch (2024)

Overall, these trends highlight the major role played by data in AI advancements in different areas, focusing on immense rise in demand for data as well as evolving potential of AI.

4.3. One of the challenges to AI is lack of data

This is because there are a number of different AI's being trained and the result are popular AI's like DALL·E 3, ChatGPT and more.Machine learning is being used to train other machine learning systems, resulting in the creation of some popular models, such as DALL·E 3 and Chat GPT. The researchers found that trading volume of stocks online doesn't increase in line with other datasets used to train AI thus far. In the real world, where things are constantly changing and dynamic, high-quality, natural data is crucial for AI models to be able to continuously understand the dynamics and changes in the real world. But it is in the reconstructed towers of bone structures that they are taking place. Consider, for example, an LLM, which would need to be continually updated, and a lot of information to keep an up-to-date and accurate program. Otherwise, in the long run their outputs may be misleading or out-of-date. Moreover, if they don't have sufficient raw information, AI programs can also be biased. The models can be used for the same or different data and give the same or biased predictions depending on data fed into the models; e.g., if the data fed into the models are false and biased, the predictions generated by the models are also false and biased. AI is one of the types of discrimination. There is a lack of diversity of data, which could mean that the AI would be more likely to appeal to some groups than others and ultimately end up being biased towards one group. May contribute to the inequality in society. The potential negative impacts of these biases are evident in the use of AI in many aspects of decision-making, such as court verdicts and health care recommendations (Giuffrè and Shung, 2023). Besides, there is limited information available and it is costly from a financial point of view. For many, a lack of knowledge about the source of the data is a concern as they are unable to innovate organizations where AI is a significant factor in cost saving, smooth operations and enhanced customer experience. Ultimately, it could hinder the use of AI in the field, and the growth of the economy.

4.4. The outcomes of the data synthesis approach to solve the problem

As there is no natural data available, the problem is to generate the synthetic data as solution. Synthetic data is created artificially to simulate data in real-world and there are million possibilities for creating synthetic data. However, there is some risk and inaccuracies associated with synthetic data. All the subtleties of problems that occur in communication can't be synthesized. AI models work well in a sterile setting, but won't be able to extrapolate from the real-world. However, synthetic data also contributes to the biases since it is more likely to be generated based on the data pattern which is one that excludes a particular group of users. In addition, there are some ethical issues with regard to synthetic data. The data is likely produced using AI with another self-trained model, and is consequently “accurate” data. Absolutely, there's an aspect of transparency / accountability / transparency in the training of another AI model using an AI model. May impact upon accuracy/ authenticity. With synthetic data, the biases will be erased in the AI systems and accuracy is crucial in design. In addition, there could be a moral issue; you may not be able to get the raw data. Much more human data use and collection will be required, rather than the sensitive validation. Use and collection of a lot of human data (with proper end user policy requirement) instead of sensitive validation. Rules – which are sometimes very different in various countries – must be addressed by AI vendors. They need to deal with that as well and the requirement of user data.

5. Technological Innovations and Solutions to Address Challenges

There are several promising avenues offered by technological advances to resolve the issues of data scarcity in AI training. This chapter covers those innovative solutions. Advanced ML Models – AI models can learn with little data, e.g., transfer learning, few-shot learning, and self-supervised learning. In the area of natural language processing (NLP), for instance, it has been applied extensively to problems involving natural language (translation, parsing, and the like), to the improvement of which it has brought a few examples of training data around 20% more accurate. Data Augmentation and Compression: Data compression reduces the data size without compromising the integrity of the data and ensures smooth processing and storage. In the same way, dataset variety can be improved with data augmentation artificially without having to collect new data. For example, augmenting images such as scaling, rotating, and coloring an image can create variations in the training set of a single image. Modern data sources such as user generated content and IoT devices can greatly enhance data availability – New Approaches. Some of them are wearables and smartphone-based applications to collect real-time health data for customized applications.

Table 6 presents some innovative solutions and their applications in the field of AI, along with their impact and descriptions/examples.

Solutions	Description	Impact	Examples	References
Few-shot learning	AI models are trained with few shots or examples rather than identifying patterns with less data	Uses Generative Adversarial Network (GAN) and modern techniques	Fills the gap of data scarcity with high-performing generalization	Dang et al (2022)
IoT Devices	Wearables or smart devices to track heart rate and send alerts about any vulnerability	Real-time tracking	Tracking vitals of patients in real-time	Abdulmalek et al (2022)
Augmenting data	Small changes made to enable learning from limited information	Improved diversity of training set	Autonomous vehicles are trained with real-life images	Yang et al (2022)
Generating synthetic data	Creating realistic data for AI to learn without the use of sensitive data of people	Not exposing private data for training	Making synthetic profiles for testing fraud detection	Goyal and Mahmoud (2024)
Transfer learning	It refers to taking one data learned by AI and using it somewhere else, such as teaching basketball to a soccer player	Models adapted to new areas without having to retrain models	Predictions of financial markets are made	Li et al (2022)
Self-supervised learning	AI learns to use raw information like an individual learning from experience instead of referring to a book	Minimizing the need for labeled information	Moderating content on social media without specifying labels	Rani et al (2023)

5.1. Proposed Strategies

There are ways to solve the issue of lack of data, which include techniques to improve the availability of data, optimize efficiency of data and maintain ethical norms, all of which AI leaders can do. Collaborative Data Distribution – This is an approach that is widely adopted to disseminate data in a democratic fashion through collaborations with companies. Groups are likely to combine resources, which will lead to less bias in the models created by the group due to a more diverse and balanced dataset. Uber is a well-known open-source program, which enabled self-driving dataset for developers worldwide, hence, promoting a co-working space of distribution of data and innovation. Natural and Synthetic Data – Although synthetic data should not be used, they can be combined with natural data in order to solve the dilemma of both types of data. Any gaps can be addressed by synthetic data when using natural data as the basis. When more balanced method is adopted by organizations, they can address the need for data without having to decline equity or selection accuracy. Data Efficiency Optimization – Optimizing data with limited data, to get the maximum information from the data supplied. Large datasets can be reduced using techniques such as transfer learning, reinforcement learning, as well as augmenting the data, while maintaining accuracy of the model. Exploring Alternative sources – Organizations are looking for alternative methods for acquiring sources of organic information such as offline footprints, consumer feedback and exclusive datasets. An analysis of resources can provide data for training models without impacting on individual privacy, and can be digitized. In order to address data scarcity in AI training, there are several key strategies that have emerged, such as enhancing data sharing, optimizing training efficiency, leveraging hybrid data, data sharing, regulatory support, and seeking alternative sources to bolster fairness, ethics, and availability in data. These strategies are highlighted in Table 7.

Table 7 – Strategies for addressing AI data scarcity

Solutions	Descriptions	Benefits
Hybrid data	Organizations share anonymized datasets to expand diverse pool of datasets	Mitigating bias, improving data security, and promoting innovation
Data efficiency	Improving model training with transfer learning, augmentation of data, and reinforcement learning	Reducing dependence on large-scale datasets for effective learning having lack of data
Using hybrid data	Combination of synthetic and real-world data for expanding capabilities of AI training	Improving adaptability of data and authenticity to boost fairness
Discovering new avenues of data	Alternative options like sensor data, offline sources, and customer feedback	Enhancing the availability of data available to improve applications in real-world
Regulatory/policy support	Responsible data sharing models established with policymakers and governments	Deploying ethical AI by complying with lawful standards

5.2. The Way Forward

With this in mind, AI leaders need to continue to prioritize data-efficient solutions and conserve energy. Without several successful models, building data-efficient models extracting value from each point of data will be important to progress after data scarcity. It should also be done by stakeholders with regulation to encourage the responsible use of natural data and high ethical standards for businesses. Like oil, data is the new currency which could unlock more potential growth of mankind. When going forward in uncertainty where unlimited data may not be allowed or impossible, people can focus on cataloguing and authoring diversity from acquiring big data of low-quality samples. Lastly, it incorporates the legitimate use of data, trusted AI models, and responsible use of security features. If these are seen as the industry values, it would reasonably grow and innovate, and it would be a positive impact of AI. Organizations need to collaborate and make strategies to overcome the problem of data scarcity. The biggest examples of sharing big data include the ways in which it has been shared between IBM and Google. Google has released ImageNet dataset, which has redefined the research for computer vision. Public-Private Partnerships (PPPs) between governments and technology companies for cooperative efforts in creating AI solutions based on shared data.

Table 8 emphasizes the collaboration and joint working efforts of the different organizations as an effective means to overcome data scarcity.

Collaborations	Programs	Purposes	Contributions	Authors
Startups and US Department of Health	Analyzing health data	Predictive analysis in healthcare	Enhanced treatment plans and diagnosis	Mayo Clinic Innovation Exchange (n.d.)
Academic institutions and Google	ImageNet	Research on computer vision	Advancing in image generation	Zoph et al (2017)
Universities and Facebook	Analyzing social media data	Behavioral pattern analysis	Providing insights on dynamics of user interaction	Lund (2019)

Weather forecasting channels and IBM	Collaboration of weather data	Improving predictions	Refining meteorological prediction models	Bandhakavi (2024)
Tech and automotive giants	Sharing data from self-driving vehicles	Upgrading technology for self-driving vehicles	Improving navigation and safety systems	Miller et al (2024)

5.3. Consequences for the regulation and policy

Regulations and policies need to be effective to ensure responsible use of data; Regulating Data Privacy – The CCPA in California and implementation of GDPR in Europe have established new standards for the privacy of data and how it will affect practices for managing AI data globally. • Data sharing incentives – Data sharing can be incentivized by the governments with tax grants and breaks, as observed in data sharing for privacy preserving by the Data Governance Act by the European Union. Table 9 explores the impact of specific regulations about data privacy affecting AI uses and developments in different areas.

Table 9 – Effect of Regulations about Data Privacy in AI

Regulations	**Regions**	**Effect**	**Description**	**Resource**
California Consumer Privacy Act (CCPA)	California	Improved rights for consumer data	Opting out from selling data by consumers	California Department of Justice, Office of the Attorney General (n.d.)
General Data Protection Regulation (GDPR)	EU	Stricter in protecting consumer data	Needs stricter consent for AI data use	Sartor and Lagioia (2020)
Personal Information Protection Law (PIPL)	China	Stricter in managing data and controls over data export	Imposing controls on international data transfer	Gong et al (2024)
Lei Geral de Proteção de Dados (LGPD)	Brazil	Improved privacy security like GDPR	Requires transparency in data usage	Trigo Kramcsak (2024)
Health Insurance Portability and Accountability Act (HIPAA)	USA	Secures data usage with privacy rule	Imposing stricter protocols on user consent and data management, influencing implementation of AI systems	U.S. Department of Health and Human Services (n.d.)

6. Conclusion

The data scarcity is a growing challenge, and industries must proactively develop data efficient and sustainable systems to tackle this challenge, given the changing dynamics of AI. There would be limited access to natural and high-quality data and building efficient models will be essential to better utilize the value of each data point. Future building of ethically sourced and diverse data for fair practices is recommended rather than acquiring large amounts of data for the development of AI. It's crucial that AI programs be designed responsibly and keep track of the developing regulatory and moral environment to make sure they are used and created appropriately. As already mentioned, regulatory sandboxes offer a means to try out AI solutions; this would encourage innovation and compliance to regulations. Moreover, AI industry needs to follow efficient ways to minimise environmental costs to maximize sustainability. The strategies to provide sustainable solutions to incorporate AI with limited resources are quantization, model pruning and "Small Language Models (SLMs)". In the same way, quantum computing. Similarly, QC may well be able to re-visualize AI in an environment where there is a ton of data. With the promising developments, new ethical and technical challenges must be met and continued cooperation and research is needed.

Future Scope

In this research paper, the point of major innovation of technology is brought in: synthetic data. Synthetic data generation manages one of the major issues in developing AI using AI to work efficiently in areas where limited amount of natural data is available. Focusing on overall efficiency and data recall further focuses on the advantages of fusion of traditional model training with generation of synthetic data to address challenges related to availability of natural data. If not enough data to be used with AI programs, then there will be new applications of "chicken neck" – a location where cooperation between models (such as transfer learning, synthetic data, and few-shot learning) might be useful. The study shows the potential for an AI in a resource-limited environment and how to implement AI in an sustainable and responsible way. Modellers and developers can benefit from the creation of responsible synthetic and diverse data, to enhance their models and bottlenecks generalisation. To address the need for a massive amount of data for big models that have a limited number of parameters, small language models are developed (Aghajanyan et al, 2023). They are only trained on a subset of information, trained in more specialised skills, where it's hard to get information to be efficient for training. SLMs can also be beneficial for terms of making practical and computational aspects easier. To overcome data scarcity, another approach is to simplify the model and decrease the memory usage and computational cost of the model using model pruning and/or quantization. This can be done with lower accuracy (quantized with larger bitsize) on activation and model weights and the pruned connections aren't as critical. These approaches, however, have the advantage of allowing efficient deployment and training of models with minimal information, and they are less likely to be overfitting and less complex**.**

References

1. Abdalla, H. B. (2022). A brief survey on big data: Technologies, terminologies and data-intensive applications. *Journal of Big Data*, *9*(1), 107. https://doi.org/10.1186/s40537-022-00659-3
2. Abdalla, H. B., & Abuhaija, B. (2023). Comprehensive analysis of various big data classification techniques: A challenging overview. *Journal of Information & Knowledge Management*, *22*(1), 2250083. https://doi.org/10.1142/S0219649222500836
3. Abdalla, H. B., Kumar, Y., Marchena, J., Guzman, S., Awlla, A., Gheisari, M., & Cheraghy, M. (2025). The Future of Artificial Intelligence in the Face of Data Scarcity. *Computers, Materials & Continua*, *84*(1).
4. Abdulmalek, S., Nasir, A., Jabbar, W. A., Almuhaya, M. A., Bairagi, A. K., Khan, M. A. M., & Kee, S. H. (2022, October). IoT-based healthcare-monitoring system towards improving quality of life: A review. In *Healthcare* (Vol. 10, No. 10, p. 1993). MDPI.
5. Aghajanyan, A., Yu, L., Conneau, A., Hsu, W. N., Hambardzumyan, K., Zhang, S., ... & Zettlemoyer, L. (2023, July). Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning* (pp. 265-279). PMLR.
6. Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A. S., Al-Dabbagh, B. S. N., ... & Gu, Y. (2023). A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, *10*(1), 46.
7. Babbar, R., & Schölkopf, B. (2019). Data scarcity, robustness and extreme multi-label classification. *Machine Learning*, *108*(8), 1329-1351.
8. Bai, J., Alzubaidi, L., Wang, Q., Kuhl, E., Bennamoun, M., & Gu, Y. (2022). *Utilising physics-guided deep learning to overcome data scarcity* (arXiv:2211.15664). arXiv. https://doi.org/10.48550/arXiv.2211.15664
9. Bandhakavi, S. (2024, September 24). *IBM launches AI foundation model for weather and climate applications*. TechMonitor. https://www.techmonitor.ai/extreme-weather-events/ibm-weather-and-climate-model/
10. Bansal, M. A., Sharma, D. R., & Kathuria, D. M. (2022). A systematic review on data scarcity problem in deep learning: solution and applications. *ACM Computing Surveys (Csur)*, *54*(10s), 1-29.
11. California Department of Justice, Office of the Attorney General. (n.d.). *California Consumer Privacy Act (CCPA).* https://oag.ca.gov/privacy/ccpa
12. Chen, Y. T., Hsu, C. Y., Yu, C. M., Barhamgi, M., & Perera, C. (2021). On the private data synthesis through deep generative models for data scarcity of industrial internet of things. *IEEE Transactions on Industrial Informatics*, *19*(1), 551-560.
13. Chen, Z., Gan, W., Wu, J., Hu, K., & Lin, H. (2025). Data scarcity in recommendation systems: A survey. *ACM Transactions on Recommender Systems*, *3*(3), 1-31.
14. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2023). Palm: Scaling language modeling with pathways. *Journal of machine learning research*, *24*(240), 1-113.
15. Danaher, J., & Nyholm, S. (2024). Digital duplicates and the scarcity problem: Might AI make us less scarce and therefore less valuable?. *Philosophy & Technology*, *37*(3), 106.

16. Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). How to prompt? Opportunities and challenges of zero-and few-shot learning for human-AI interaction in creative applications of generative models. *arXiv preprint arXiv:2209.01390*.
17. Databricks. (2024). *State of Data + AI*. Databricks. https://www.fbcinc.com/source/virtualhall_images/2024_Virtual_Events/USDA_Innovation/Databricks/State_of_Data___AI_Resource.pdf
18. Gangwal, A., Ansari, A., Ahmad, I., Azad, A. K., & Sulaiman, W. M. A. W. (2024). Current strategies to address data scarcity in artificial intelligence-based drug discovery: A comprehensive review. *Computers in Biology and Medicine*, *179*, 108734.
19. Giuffrè, M., & Shung, D. L. (2023). Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ digital medicine*, *6*(1), 186.
20. Gong, J., Qu, H., & Dorwart, H. (2024, March 14). *AI governance in China: Strategies, initiatives, and key considerations*. Bird & Bird. https://www.twobirds.com/en/insights/2024/china/ai-governance-in-china-strategies-initiatives-and-key-considerations
21. Goyal, M., & Mahmoud, Q. H. (2024). A systematic review of synthetic data generation techniques using generative AI. *Electronics*, *13*(17), 3509.
22. Hausen, R., & Azarbonyad, H. (2024). Discovering data sets through machine learning: An ensemble approach to uncovering the prevalence of government-funded data sets. *Harvard Data Science Review*, *4*, 6-5545.
23. Healy, M., Baum, A., & Musumeci, F. (2025). Addressing data scarcity in ML-based failure-cause identification in optical networks through generative models. *Optical Fiber Technology*, *90*, 104137. https://doi.org/10.1016/j.yofte.2025.104137
24. Huang, T., Luo, T., Yan, M., Zhou, J. T., & Goh, R. (2022). RCT: Resource constrained training for edge AI. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(2), 2575-2587.
25. Huang, T., Luo, T., Yan, M., Zhou, J. T., & Goh, R. (2024). RCT: Resource constrained training for edge AI. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(2), 2575–2587. https://doi.org/10.1109/TNNLS.2022.3190451
26. Khalil, M., Vadiee, F., Shakya, R., & Liu, Q. (2025, March). Creating artificial students that never existed: Leveraging large language models and CTGANs for synthetic data generation. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 439-450).
27. KHAPRA, C. (2024). A survey on end-to-end speech recognition systems. *INTERNATIONAL JOURNAL OF COMMUNICATION*, *5*(2), 122-127.
28. Khapra, P. (2022). Evolution of cybercrime and deepfakes–Exploring intervention strategies of international organizations against AI threats. *NeuroQuantology*, *20*(22), 1425-1434.
29. Kumar, Y., Huang, K., Perez, A., Yang, G., Li, J. J., Morreale, P., ... & Jiang, R. (2024). Bias and cyberbullying detection and data generation using transformer artificial intelligence models and top large language models. *Electronics*, *13*(17), 3431.
30. Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2024). Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, *32*(1), 84-100.
31. Lautrup, A. D., Hyrup, T., Zimek, A., & Schneider-Kamp, P. (2025). Syntheval: a framework for detailed utility and privacy evaluation of tabular synthetic data: AD Lautrup et al. *Data Mining and Knowledge Discovery*, *39*(1), 6.
32. Li, W., Huang, R., Li, J., Liao, Y., Chen, Z., He, G., ... & Gryllias, K. (2022). A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: Theories, applications and challenges. *Mechanical systems and signal processing*, *167*, 108487.
33. Lund, B. (2019). Universities engaging social media users: an investigation of quantitative relationships between universities' Facebook followers/interactions and university attributes. *Journal of marketing for higher education*, *29*(2), 251-267.
34. Lynch, S. (2024). *AI Index: State of AI in 13 Charts*. HAI. Available at https://hai.stanford.edu/news/ai-index-state-ai-13-charts
35. Mayo Clinic Innovation Exchange. (n.d.). *An introduction to federal funding for innovation*. https://innovationexchange.mayoclinic.org/an-introduction-to-federal-funding-for-innovation/
36. Miller, T., Durlik, I., Kostecka, E., Borkowski, P., & Łobodzińska, A. (2024). A critical AI view on autonomous vehicle navigation: the growing danger. *Electronics*, *13*(18), 3660.
37. Nahid, M. M. H., & Hasan, S. B. (2024). SafeSynthDP: leveraging large language models for privacy-preserving synthetic data generation using differential privacy. *arXiv preprint arXiv:2412.20641*.

38. Nandy, A., Duan, C., & Kulik, H. J. (2022). Audacity of huge: overcoming challenges of data scarcity and data quality for machine learning in computational materials discovery. *Current Opinion in Chemical Engineering*, *36*, 100778.
39. Niel, O. (2023). A novel algorithm can generate data to train machine learning models in conditions of extreme scarcity of real world data. *arXiv preprint arXiv:2305.00987*.
40. Oxford Analytica. (2024). Data scarcity challenges AI developers globally. *Emerald Expert Briefings*.
41. Panagiotou, E., Qian, H., Marx, S., & Ntoutsi, E. (2024). Generative Ai Based Augmentation for Offshore Jacket Design: An Integrated Approach for Mixed Tabular Data Generation under Data Scarcity and Imbalance. *Available at SSRN 4703856*.
42. Panagiotou, E., Qian, H., Marx, S., & Ntoutsi, E. (2025). Generative AI-augmented offshore jacket design: Integrated approach for mixed tabular data generation under scarcity and imbalance. *Automation in Construction*, *177*, 106287. https://doi.org/10.1016/j.autcon.2025.106287
43. Quito, A. V., & Andrade, L. S. (2022). Proposal for the generation of profiles using a synthetic database. In *Applied Human Factors and Ergonomics International* (pp. 187-193). AHFE International.
44. Ramakrishnan, R. (2023). How to build good AI solutions when data is scarce. *MIT Sloan Management Review*, *64*(2), 48-53.
45. Rani, V., Nabi, S. T., Kumar, M., Mittal, A., & Kumar, K. (2023). Self-supervised learning: A succinct review. *Archives of Computational Methods in Engineering*, *30*(4), 2761-2775.
46. Sartor, G., & Lagioia, F. (2020). *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence* (EPRS Study No. 641530). European Parliamentary Research Service, European Parliament.
https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf
47. Shukla, S., Mishra, S., Singh, J., Mishra, K., Khapra, C., & Morey, P. (2024, August). Deep Belief Networks for Unsupervised Feature Learning and Improved Image Recognition. In *International Conference on Artificial Intelligence on Textile and Apparel* (pp. 133-145). Singapore: Springer Nature Singapore.
48. Singh, J. (2021). The rise of synthetic data: enhancing AI and machine learning model training to address data scarcity and mitigate privacy risks. *Journal of Artificial Intelligence Research and Applications*, *1*(2), 292-332.
49. Singh, P. (2023). Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Science and Management*, *6*(3), 144-157.
50. Trigo Kramcsak, P. R. (2024). *Personal data protection and data transfer regulation in Brazil* (Working Papers – Data Protection & Privacy, Vol. 10, No. 2). Brussels Privacy Hub Working Papers. https://brusselsprivacyhub.com/wp-content/uploads/2024/02/Personal-Data-Protection-in-Brazil.pdf
51. U.S. Department of Health and Human Services. (n.d.). *Laws & regulations*. https://www.hhs.gov/hipaa/for-professionals/privacy/laws
52. Van, H. (2023). Mitigating data scarcity for large language models. *arXiv preprint arXiv:2302.01806*.
53. Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2022). Will we run out of data? Limits of LLM scaling based on human-generated data. *arXiv preprint arXiv:2211.04325*.
54. White, J., Madaan, P., Shenoy, N., Agnihotri, A., Sharma, M., & Doshi, J. (2022). A case for rejection in low resource ML deployment. *arXiv preprint arXiv:2208.06359*.
55. Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., & Shen, F. (2022). Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*.
56. Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2017, November 2). *AutoML for large scale image classification and object detection*. Google Research Blog. https://research.google/blog/automl-for-large-scale-image-classification-and-object-detection/
57. Zou, M. X., Lo, C. C., Lin, C. H., Shieh, C. S., & Horng, M. F. (2022, December). Data augmentation based on topic relevance to enhance text classification in scarcity of training data. In *International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (pp. 347-357). Singapore: Springer Nature Singapore.