Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.063551

ARTICLE

The Future of Artificial Intelligence in the Face of Data Scarcity

Hemn Barzan Abdalla[1,*], Yulia Kumar[2], Jose Marchena[2], Stephany Guzman[2], Ardalan Awlla[3], Mehdi Gheisari[4] and Maryam Cheraghy[1]

[1]Department of Computer Science, Wenzhou-Kean University, Wenzhou, 325015, China
[2]Department of Computer Science and Technology, Kean University, Union, NJ 07083, USA
[3]Department of Computer Science, Cihan University Sulaimaniya, Sulaymaniyah, 46001, Iraq
[4]Institute of Artificial Intelligence, Shaoxing University, Shaoxing, 312010, China

*Corresponding Author: Hemn Barzan Abdalla. Email: habdalla@kean.edu

Received: 17 January 2025; Accepted: 28 March 2025; Published: 09 June 2025

ABSTRACT: Dealing with data scarcity is the biggest challenge faced by Artificial Intelligence (AI), and it will be interesting to see how we overcome this obstacle in the future, but for now, "THE SHOW MUST GO ON!!!" As AI spreads and transforms more industries, the lack of data is a significant obstacle: the best methods for teaching machines how real-world processes work. This paper explores the considerable implications of data scarcity for the AI industry, which threatens to restrict its growth and potential, and proposes plausible solutions and perspectives. In addition, this article focuses highly on different ethical considerations: privacy, consent, and non-discrimination principles during AI model developments under limited conditions. Besides, innovative technologies are investigated through the paper in aspects that need implementation by incorporating transfer learning, few-shot learning, and data augmentation to adapt models so they could fit effective use processes in low-resource settings. This thus emphasizes the need for collaborative frameworks and sound methodologies that ensure applicability and fairness, tackling the technical and ethical challenges associated with data scarcity in AI. This article also discusses prospective approaches to dealing with data scarcity, emphasizing the blend of synthetic data and traditional models and the use of advanced machine learning techniques such as transfer learning and few-shot learning. These techniques aim to enhance the flexibility and effectiveness of AI systems across various industries while ensuring sustainable AI technology development amid ongoing data scarcity.

KEYWORDS: Data scarcity; artificial intelligence; application of artificial intelligence; ethical considerations; artificial general intelligence; synthetic data

1 Introduction

With the advancement of AI, many sectors have progressed to a new, innovative future, including healthcare, finance, and other sectors. However, with continued advancements in AI, a basic limiting factor has slowly come to light, which may reshape the evolution of this technology: not enough high-quality, real-world data is available. Here is the fundamental concept of how this works—Data → AI Engine → Learning, Adapting, and Decision Making. However, the need for data is growing faster than authentic and varied sources can provide. The lack of naturally occurring data, i.e., unadulterated natural information based upon direct human interactions and experiences as they occur in the world, is putting AI development at serious risk [1–5].

The consequences of less data are monumental, including ethical questions and the efficacy and accuracy of AI models. To truly grasp the severity of this matter, it is essential to think about how AI categorizes things not as mere numbers or letters but as patterns formed by behaviors learned from enormous data [6–10]. However, when inputting datasets are restricted, is the AI we rely on left to function with insufficient information or underwhelming accuracy? Data scarcity is an existential problem for AI in all sectors [11–14], though it does not exist technically. Nonetheless, it significantly affects the marginal AI systems that businesses use to enhance performance [15–19].

With data learning being a dilemma, the future of AI is at a critical juncture. Given that AI systems become progressively dependent on enormous amounts of data to acquire knowledge and produce predictions, this restriction in the form of insufficient data can considerably impact their effectiveness and applicability in different domains. This issue is especially pronounced in disciplines like healthcare, where the precision and reliability of AI-powered solutions are critical. For example, we used relatively small datasets for many of the AI applications in healthcare, which have yielded accuracies below those seen in clinical settings [18]; this makes the AI models less robust and generalized, which calls for developing novel methodologies to deal with limited data situations.

In addition, the ethical consequences of AI creation in the setting of a lack of data are significant and must not be ignored. Data reliance and the need for extensive datasets in machine learning [16] have made questions of privacy, consent, and bias in AI algorithms to reproduce existing inequalities theoretically even more pertinent [1]. Hence, organizations are required to overcome such socio-technical issues, which will, in turn, facilitate a robust and moral AI system [18]. This joint effort was necessary with many stakeholders of different sectors to create ethical standards and regulatory frameworks to rule the responsible generation and implementation of AI technologies [4].

Beyond ethical considerations, the future of AI depends on advances in data engineering practices and AI model lifecycle management, given the scarcity of data. Organizations must develop the capabilities to update their AI systems as the data evolves to make their prediction relevant and effective over time [9]. Additionally, the convergence of AI with other technologies (like machine learning and intelligence augmentation) can help in better utilization of data as well as in improving AI productivity [2]. However, by encouraging the cross-fertilization of ideas between different fields of science, stakeholders can counteract the negative consequences of data shortages and facilitate the realization of AI's potential in a range of sectors [16].

As AI increasingly matters across domains, this article discusses data scarcity (See Appendix A, Table A1) as a key obstacle to AI progress. Section 2 delves into the impact of data scarcity, emphasizing its role in limiting AI's ability to generalize, adapt, and perform effectively in real-world applications. Section 3 discusses some of the ethical implications of limited data, including bias, privacy, and fairness risks, and calls for responsible AI development.

Section 4 explores technological innovations such as synthetic data generation, transfer learning, and short learning as potential paths toward overcoming the problem of data scarcity. Section 5 provides a case study that illustrates how the generation of synthetic data can be used to augment the training dataset using SMOTE (See Appendix A, Table A1) in the case of a very imbalanced dataset, which improves model performance in terms of generalization. These experimental results confirm the essential role of synthetic data in tackling data scarcity [20,21].

Ultimately, the conclusion argues for a hybrid approach, whereby synthetic data is integrated into existing paradigms, underpinned by ethical frameworks and cross-indexing across all sectors, to empower AI

systems to succeed in low-resource settings. This in-depth survey sheds light on addressing data challenges and scaling AI responsibly.

2 Literature Review

Table 1 summarizes information about data scarcity across various fields of AI and draws from multiple references throughout the document to highlight ongoing research and methods to address this challenge, as shown in Table 1.

Table 1: Literature review

References	Key points	Findings
[17,18]	Focused on evaluating various machine learning data augmentation techniques to address data scarcity.	Emphasized the need for more sophisticated methods that can simulate real-world variability effectively.
[22]	Explored strategies including collaborative filtering and content-based methods to enhance recommendation systems in the face of data scarcity.	Highlighted the critical impact of these strategies on improving user engagement and recommendation diversity.
[23]	Utilized transfer learning and synthetic data generation to adapt technical systems to operate under conditions of data scarcity.	Detailed the development of methodologies robust enough to handle diverse conditions of data scarcity.
[24]	Discussed the implementation of rejection mechanisms in ML (Machine Learning) deployments to improve decision reliability in low-resource settings.	Argued for the necessity of such mechanisms to ensure reliability in critical AI applications.
[25]	Reviewed data-centric approaches, contrasting them with traditional model-centric methods, to address data scarcity in AI.	Identified key aspects where data-centric innovations are crucial for the advancement of AI.
[26]	Proposed the creation of a low-cost mirror environment to simulate real-world conditions for AI training and development.	Demonstrated that such environments can significantly enhance AI system reliability and performance.
[27,28]	Implemented AI-based methods to integrate diverse data sources for improved pharmacovigilance.	Identified challenges in drug safety monitoring due to sparse data and proposed solutions.
[29]	Developed a novel algorithm to generate synthetic data for training machine learning models, specifically in medical imaging.	Demonstrated an increase in training data availability and improved accuracy in medical diagnostics.
[30]	Utilized deep generative models for the private data synthesis, addressing both privacy concerns and data scarcity.	Indicated that these models could effectively tackle privacy and scarcity issues in AI.

(Continued)

Table 1 (continued)

References	Key points	Findings
[31,32]	Discussed deep generative models for private data synthesis, focusing on improving classifier performance under privacy constraints.	Found that these models could enhance classifier performance by generating diverse and privacy-compliant data.
[33]	Employed the BERT (Bidirectional Encoder Representations from Transformers) model to generate text data based on topic relevance, aiming to enrich data availability for ML training.	Demonstrated an enhancement in data availability and relevance, improving overall model performance.
[34]	Explored data augmentation techniques to enhance the training and performance of large language models.	Found that these techniques significantly mitigate data scarcity impacts on model training.
[35]	Conducted a comprehensive review of deep learning tools designed to handle data scarcity, examining various strategies and their applications.	Provided an overview of tools' capabilities and effectiveness in addressing scarce data challenges.
[36]	Addressed the issue of data scarcity in the context of political communication research, proposing methodological adaptations.	Detailed adaptations in research methodology but specific findings were not provided.
[37]	Proposed the creation of synthetic profiles using AI to enrich data availability for various applications.	Demonstrated the potential of synthetic data to enhance availability across diverse AI applications.
[38]	Discussion on the global impact of data scarcity on AI development, focusing on challenges developers face.	Highlighted the pervasive nature of data scarcity and its global impact on AI development.
[39]	Proposed heuristic training methods as part of Resource Constrained Training (RCT) to optimize training under resource limitations.	Showcased the effectiveness of RCT methods in enhancing training efficiency in resource-poor environments.
[8]	The importance of data sharing was emphasized, particularly in the context of medical AI development.	Stressed that enhancing data sharing practices is crucial for advancing AI in medicine.

3 Navigating Data Challenges in AI Development

3.1 The Critical Role of Natural Data in AI Development

Training AI models to function properly in unpredictable environments requires natural data, also called raw or unprocessed real-world scenario data [2,6]. The flaws, assumptions, and nuances in this data are perfect for AI to predict and adapt well to a massive range of unique scenarios. Synthetic data, on the other hand, is loosely defined as artificially generated to imitate real-world data and often does not have the authenticity or richness necessary for creating robust and reliable AI.

What sits behind natural data is the truth in how humans interact, speak, and behave. For example, the subtleties of language (i.e., slang in different regions, typos, and differences between syntaxes) are only learned through immersion in linguistic diversity [39]. As a result, models trained only on synthetic data may not generalize well to other demographics and contexts, which could produce biased or incorrect results. Thus, having natural data is crucial in constructing AI systems capable of interacting and servicing a wide range of user bases, addressing the fundamental challenge of responsible progress on AI.

Workflow of processing natural data for AI development:

- **Start:** Initiation of data collection.
- **Activities:**
  - **Collect Natural Data:** Gathering data from various sources.
  - **Clean Data:** Removing irrelevant or erroneous data.
  - **Analyze Data:** Understanding patterns and anomalies.
  - **Train AI Model:** Using the processed data to train the model.
  - **Evaluate Model:** Testing the model against a validation dataset.
- **Decisions:** Based on the outcome of the model evaluation, decide whether to retrain the model or proceed to deployment.
- **End:** Conclude the process if the model meets the desired metrics.

In the modern age of digitalization, data and artificial intelligence (AI) are playing a central role in ushering innovation across industries, as shown in Fig. 1. As the quantity of information keeps increasing due to digital devices and online platforms, along with rapid AI progression, our days are getting used to seeing eye-popping improvements.

- **Unpacking the Synergy of Data and AI:**

The Data Deluge: In today's world, a vast amount of data is being produced by devices and connectivity. It comes with a volume that is difficult to manage; otherwise, AI can process and unearth insights at an unimaginable scale.

The AI Revolution: What was once a concept is more of a reality, as AI now has become the feature that lifts our world. It covers doings/moves such as machine learning and natural language processing, which refers to the capability of machines to learn with data execution tasks requiring human-level cognition. AI and data are related to each other by:

- **Data Feeds AI:** High-quality data trains AI models to recognize patterns and make informed decisions.
- **AI Enhances Data:** AI processes data faster than humanly possible, organizing and extracting key insights.

Table 2 illustrates how the use of natural data in AI development is influenced by trends towards open-source, the high computational and environmental costs of training models, the financial focus on generative AI technologies, and growing concerns about AI ethics and fairness. These elements highlight the ongoing evolution and challenges of using natural data for AI advancements.

Figure 1: The critical roles of data and AI

Table 2: The key data points on the critical role of natural data in AI development for 2023

Category	Value	Description	References
Foundation models released	149	Number of foundation models released in 2023	[40]
Generative AI investment (Billion)	22.4%	Investment in generative AI in 2023 (in billions)	[41]
Environmental impact	41 yrs	Years of power for an average American home provided by the energy used to train GPT-3 (General Pre-trained Transformer-3)	[40]
AI ethics submissions increase	10	Tenfold increase in submissions to AI ethics conferences since 2018	[42]

3.2 The Growing Demand for Data in AI

Now that AI has become associated with everything from personalized product recommendations to medical diagnostics, there is a demand for enormous stores of diverse data. Large models (like, for instance, GPT-4) [15,43] need the order of billions of words and images to operate at their best levels in terms of both accuracy and reliability. The burgeoning number of applications in different domains further expedites this demand. It is no secret that industries, be it finance, healthcare, or e-commerce, rely heavily on AI-powered analytics for better insights, to take control of trends, and to make strategic decisions [43].

On the one hand, they count on this level of demand. On the one hand, it hammers out rapid iteration as companies fight to make their AI shine brighter than others. On the other hand, it puts enormous pressure on natural data reserves, leading to a supply-demand imbalance that eventually slows down AI progress. Synthetic data, an alternative to genuine transaction data for testing and developing new analytics such as advanced Machine Learning (ML) [44] or AI, has, of late, become many organizations' answer where the

continued advancement, albeit very slowly brings with positive outcomes in part because it provides test cases where otherwise notary from real-world legitimate events were are even less use for plan these type algorithms. The absence of suitable sources for real data can reduce AI applications' reliability and ethical integrity, making them less attractive to use in critical sectors.

Table 3 summarizes the latest data on the growing demand for data in AI as of 2023, along with key trends and developments across various industries:

Table 3: The growing demand for data in AI as of 2023

Sector	Key data and trends	References
Overall AI	72% of all new foundation models are developed by industry.	[43]
Machine learning	Logged models grew 54%, and registered models grew 411% since February 2022.	[45]
Generative AI	Investment surged to $25.2 billion in 2023.	[46]
Healthcare	Big Data enables personalized medicine and predictive healthcare.	[47]
Retail	Big Data drives insights into consumer behavior and market trends.	[48]
Finance	Big Data is used for fraud detection and risk assessment.	[47]
Manufacturing	Big Data optimizes supply chain operations and production efficiency.	[49]

These insights demonstrate the critical role of data in driving advancements in AI across various sectors, highlighting the immense growth in data demand and the evolving capabilities of AI technologies.

3.3 Challenges of Data Scarcity for AI

AI has been training AI systems on ever-larger datasets, which is why we now have high-performing models such as ChatGPT or DALL-E 3. Simultaneously, research indicates that the growth of online data stocks is significantly slower than that of datasets utilized for AI training. In a paper published last year, researchers predicted we would run out of high-quality text data by 2026 if the current AI training trends continue. They also estimated that low-quality language data would be exhausted between 2030 and 2050 and low-quality image data between 2030 and 2060. According to the accounting and consulting firm PwC, artificial intelligence has the potential to contribute a staggering US$15.7 trillion (A$24.1 trillion) to the global economy by 2030. However, running short of usable data could slow down its development [50].

However, it is not just a by-product of its limitations as it has ethical and socio-economic implications. Without natural, high-quality data, AI models cannot remain perpetually privy to the dynamics of change and instead become captives in their synthetic ivory towers. For example, large language models require continuous access to diverse, up-to-date data to remain relevant and accurate. Otherwise, their outputs might become obsolete or inaccurate in time.

In addition, without ample natural data, AI systems may also become biased. Because these models are trained on incomplete or homogeneous datasets, they learn and replicate the biases built into that data, e.g., leading to discriminatory/unfair outcomes. One example might include an AI system used in hiring, which may lack diverse data, leading to certain demographics being preferred over others and ultimately favoring one group, thereby perpetuating societal imbalance. The consequences of these biases could be devastating as AI systems become central to different decision-making processes, from court judgments to healthcare recommendations [51].

Data scarcity is also very expensive from a monetary perspective. Many may struggle to innovate as companies who rely on AI for efficient operations, cost reduction, or enhanced customer experiences find it difficult when they are not sure about the origin of their data. This, in turn, could stall AI deployment in industry and, with it, the development of technologies that improve economic advancement. Table 4 lists key challenges of data scarcity in AI—i.e., data exhaustion, risk of bias, regulatory constraints, and cost—and their potential negative implications on model performance, fairness, and innovation.

Table 4: Challenges posed by data scarcity in AI

Challenge	Description	Potential impact
Data exhaustion	High-quality language and image data are projected to be depleted within the next two decades.	Slower AI model development, reduced innovation, and performance stagnation.
Increasing bias risk	Limited data diversity can reinforce existing biases in AI models.	Biased decision-making in sectors like hiring, law, and healthcare.
Regulatory constraints	Strict data regulations limit access to high-quality, real-world data.	Reduced availability of natural data, affecting model accuracy and fairness.
Resource imbalance	Smaller companies may struggle more with limited data access than larger, resource-rich organizations.	Possible monopolization of AI advancements by well-funded companies.
Cost of data collection	High costs associated with sourcing, curating, and annotating high-quality data.	Increased operational expenses, slowing down AI project deployment.

3.4 Risks of Synthetic Data as a Solution

As natural data becomes less available, synthetic data has become a possible solution. Synthetic data is fake; it is artificially created to mirror real-world data, and the possibilities of being generated are endless. However, replacing natural data comes with some risks and imperfections. One is synthetic data because it cannot replicate the nuanced complexities that humans experience in interactions. AI models do great in a controlled environment but will struggle to generalize on chaotic, real-life situations. Many synthetic data also risk exacerbating pre-existing biases by default, as they tend to be generated based on patterns in existing data that often exclude specific user groups.

Furthermore, using synthetic data touches on its ethical dilemmas. This is a problem in machine learning as the AI may generate data that it assumes is correct by using another self-trained algorithm. Concepts like "AI training AI" are not accountable or transparent, so these models' authenticity and accuracy cannot be easily proven. The biases contained within synthetic data will eventually dilute into the AI systems themselves, and in situations where accuracy is critical for a design, this could have adverse effects. Table 5 highlights how synthetic data can address data scarcity by increasing accessibility, affordability, control over bias, scalability, and ethics while also outlining associated risks like loss of authenticity, potential for bias, and ethical concerns.

3.5 Ethical and Privacy Concerns in the Data-Driven AI Landscape

The lack of natural data is also hindered by privacy and ethical challenges. Human-generated data must be acquired and employed, often at the cost of sensitive private-output validation with significant end-user

privacy implications. Instances like the Cambridge Analytica scandal have seeped into internet culture and public consciousness concerning data privacy, rising to quadruple scrutiny among nations. This has led to severe regulations and left little scope for vendors or site owners to snoop around without permission.

Table 5: Synthetic data as a solution to data scarcity

Aspect	Benefits	Risks
Accessibility	Synthetic data can be generated to fill data gaps.	Generated data may lack real-world authenticity, affecting model performance.
Cost-effectiveness	Reduces costs associated with data collection and annotation.	Quality concerns may require additional validation, adding costs.
Bias management	Synthetic data can be tailored to improve dataset diversity.	Potential for new biases introduced if synthetic data is derived from biased data sources.
Scalability	Easy to produce large volumes for training.	Excessive reliance on synthetic data risks a feedback loop in machine training, limiting diversity.
Ethical considerations	Avoids privacy concerns associated with real-world data.	Ethical ambiguity around training models without real-world grounding.

AI companies have to navigate data regulations that change dramatically between countries; they must also balance this with their deep wish for more user data and our demand for a private age. When the data is used to train AI systems that were not explicitly consented to or carried out with explicit permissions, ethical challenges start showing their face. Others are calling for more robust frameworks, especially concerning data collection, reiterating the importance of ethical approaches to AI regarding standards, as shown in Table 6.

Table 6: Ethical and privacy considerations in AI data usage

Ethical concern	Description	Importance of AI development
Data privacy	Ensuring data collection and usage comply with privacy laws and respect individual rights.	Builds public trust in AI systems and prevents legal repercussions.
Bias reduction	Avoiding biases that could lead to discrimination or unfair treatment.	Ensures AI applications serve all demographic groups equitably.
Transparency	Providing clarity on data sources and AI training methodologies.	Fosters trust and accountability in AI applications.
Accountability	Responsibility for ethical AI outcomes, especially in sensitive sectors like healthcare.	Minimizes risks of harm from biased or erroneous model outputs.
Public consent	Involving public opinion and securing consent for data usage.	Increases societal acceptance of AI and aligns AI development with societal values.

4 Technological Innovations to Address Data Scarcity

Advances in technology offer promising solutions to the challenges of data scarcity in AI. This section outlines several innovative approaches:

- **Advanced Machine Learning Algorithms:** Techniques such as few-shot learning, transfer learning, and self-supervised learning enable AI models to learn effectively from limited data. For instance, few-shot learning has been successfully employed in natural language processing (NLP) tasks, achieving up to a 20% improvement in accuracy with only a handful of training examples.
- **Data Compression and Augmentation Techniques:** Data compression reduces dataset size while maintaining data integrity, enabling efficient storage and processing. Meanwhile, data augmentation artificially enhances dataset variety without requiring new data collection. For example, image augmentation techniques such as rotation, scaling, and color adjustment generate diverse training samples from a single image.
- **Novel Data Acquisition Methods:** Leveraging unconventional data sources like IoT devices and user-generated content can substantially increase data availability. Examples include using smartphones and wearable devices to collect real-time health data for personalized medicine applications.

Table 7 illustrates various innovative technologies and their applications in AI, showcasing their impacts and providing key definitions, impact, and examples where applicable.

Table 7: Examples of technological innovations in AI

Technology	Definition	Impact	Example
Few-shot learning	Training AI models with only a few examples instead of thousands allows them to recognize patterns efficiently with minimal data [52].	A novel strategy that leverages (GANs, Generative Adversarial Network) and advanced optimization techniques	Bridges data scarcity with high-performing model adaptability and generalization
Data augmentation	Making small picture changes (flipping, rotating, changing brightness) to help AI learn better from limited data [53].	Enhanced training set diversity	Training autonomous driving systems with modified real-world images
IoT devices	Smartwatches or medical devices that track heart rate and send alerts if something is wrong [54].	Real-time health monitoring	Using wearable devices to monitor patient vitals in real-time
Synthetic data generation	Creating fake but realistic data so AI can learn without using real people's sensitive information [55].	Training without exposing personal data	Creating synthetic financial profiles for fraud detection testing
Self-supervised learning	AI teaches itself using raw data, like a person learning from experience instead of reading a manual [56].	Reduces the need for labeled datasets	Content moderation on social media platforms without predefined labels
Transfer learning	Taking what an AI learned in one area and using it elsewhere, like teaching a soccer player how to play basketball [57].	Adapting models to new areas without retraining	Applying financial market predictions to healthcare trends

5 Strategic Solutions to the Data Crisis

To address data scarcity, AI developers and companies can adopt strategic approaches that optimize data efficiency, expand data availability, and maintain ethical standards.

- **Optimizing Data Efficiency:** When data is limited, models should be optimized to extract maximum insight from the available information. Techniques like data augmentation, transfer learning, and reinforcement learning (See Appendix A, Table A1) help reduce the dependency on large datasets while maintaining model accuracy.
- **Collaborative Data Sharing:** A competitive way to democratize data access is via company partnerships. In pooling resources, groups will realize a more balanced and diverse training dataset in the models they develop, leading to less bias in general. A quintessential open-source initiative is Uber, which has made available its self-driving dataset to developers around the globe, thus facilitating a co-working environment of innovation and data dissemination subject to regulated ethical narratives.
- **Integrating Synthetic and Natural Data:** Although synthetic data is not sufficient by itself, a hybrid of natural and artificial may overcome the weaknesses inherent in both types. If natural data are used as the base, synthetic data can fill any gaps, especially in niche or underrepresented areas. When companies take this more balanced approach, they can respond to the demands for data without degrading selection accuracy or equity.
- **Exploring Alternative Data Sources:** Companies are already looking to newer organic data channels, such as customer feedback, offline footprints, and proprietary datasets. By digitizing and analyzing their resources, companies can create useful training data without stepping over the privacy boundary.
- Table 8 lists the primary approaches to AI data scarcity by enhancing training efficiency, data sharing, hybrid data usage, alternative sources, and regulation support for enabling improved availability, fairness, and ethics compliance.

Table 8: Strategic solutions to address data scarcity in AI

Solution	Description	Benefits
Data efficiency techniques	Focus on enhancing model training through data augmentation, transfer learning, and reinforcement learning.	Reduces reliance on extensive datasets, enabling effective learning with limited data resources [11].
Collaborative data sharing	Companies partner to share anonymized datasets, expanding diverse data pools.	Enhances data availability, mitigates bias risks, and fosters AI innovation.
Hybrid data use	Combines real-world and synthetic data to expand AI training capabilities.	Maintains data authenticity, improves model adaptability, and enhances fairness.
Exploring new data sources	Alternative sources like customer feedback, sensor data, and offline repositories are used.	Expands available data diversity, improving real-world model applications.
Policy and regulatory support	Establishing responsible data-sharing frameworks in partnership with governments and policymakers.	Ensures ethical AI deployment while maintaining compliance with legal standards.

5.1 The Road Ahead: Building Sustainable AI with Limited Data

Given these challenges, the AI industry should also work to maintain energy—and data-efficient systems. With so many seemingly successful algorithms, developing data-efficient algorithms that extract as much value from each data point will be vital to progress once natural resources are scarce. Policymakers will also have a role in making that happen, such as through regulations to facilitate responsible data use and ensure businesses maintain high ethical standards.

Much like oil, data is a resource that could be the key to unlocking even more potential human progress—but as we move forward into an uncertain future where unlimited data may no longer be possible or generally allowed, people can turn their attention from the acquisition of massive amounts of low-quality convenience samples and direct our efforts on authoring and cataloging diversity in "good" taste if it were. Ultimately, it comes down to combining lawful data quality with trustworthy AI systems and responsible development to ensure top-notch security features. If the industry keeps these values in mind, it should be able to innovate and grow responsibly, with AI as a positive force for good.

5.2 Strategic Approaches and Partnerships

Collaboration and strategic partnerships between organizations can be pivotal in overcoming data scarcity:

Data Sharing Initiatives: Companies like Google and IBM have pioneered sharing large datasets. For example, Google's release of the ImageNet database has revolutionized computer vision research.

Public-Private Partnerships: Collaborations between governments and tech companies can facilitate the development of AI technologies with shared datasets. An example is the partnership between the U.S. Department of Health and AI startups to analyze health data securely.

Table 9 highlights significant partnerships between organizations aimed at leveraging collaboration to address challenges, including data scarcity.

Table 9: Key strategic partnerships in AI

Partners	Initiative	Purpose	Contribution	References
Google and academic institutions	ImageNet database	Boost research in computer vision	Pioneered advancements in image recognition	[58,59]
U.S. department of health and startups	Health data analysis	Enhance predictive capabilities in healthcare	Improved diagnostics and treatment plans	[60]
IBM and weather channel	Weather data collaboration	Enhance meteorological predictions	Refined forecasting models in meteorology	[61]
Facebook and universities	Social data analysis	Study behavioral patterns	Provided insights into user interaction dynamics	[62]
Automotive companies and tech firms	Autonomous vehicle data sharing	Accelerate autonomous vehicle technology	Enhanced safety and navigation systems	[63]

5.3 Policy and Regulation Considerations

Effective policy and regulation are crucial to ensuring data is used responsibly:

Data Privacy Regulations: The implementation of GDPR in Europe [64] and the CCPA in California has set new benchmarks for data privacy, influencing AI data handling practices worldwide.

Incentives for Data Sharing: Governments can incentivize data sharing with tax breaks and grants, as seen in the EU's Data Governance Act, which aims to foster data sharing while ensuring privacy.

Table 10 explores how specific data privacy regulations affect AI development and application across various regions.

Table 10: Impact of data privacy regulations on AI

Regulation	Region	Impact	Details	References
GDPR (General Data Protection Regulation)	Europe	Tightened data protection	Requires stringent consent for data use in AI	[64]
CCPA (California Consumer Privacy Act)	California	Strengthened consumer data rights	Enables consumers to opt out of data selling	[65]
LGPD (General Data Protection Law)	Brazil	Enhanced privacy protections similar to GDPR	Mandates transparent data usage policies	[40]
PIPL (Personal Information Protection Law)	China	Strict data management and export controls	Imposes controls on cross-border data transfers	[66]
HIPAA (Health Insurance Portability and Accountability Act)	USA	Privacy Rule permits important uses of information	These regulations impose strict requirements on data handling and user consent, thereby influencing how AI systems are developed and implemented	[67]

5.4 Case Study: Addressing Data Scarcity in Fraud Detection

This is a common challenge where rare but critical events, such as fraudulent transactions, need to be identified, like fraudulent transactions. For a hands-on understanding of synthetic data generation, we used the Credit Card Fraud Detection Dataset [68], a benchmark dataset in which only 0.17% of transactions are marked as fraudulent. This case study is the application of synthetic oversampling techniques, such as SMOTE (Synthetic Minority Oversampling Technique), to mitigate data imbalance (See Appendix A, Table A1) and enhance AI model accuracy.

Dataset Overview: The dataset [68] consists of anonymized transaction data with significant class imbalance, as shown in Eqs. (1) and (2):

- **Normal Transactions (Support: 85,295):**

$$\text{Percentage} = \frac{85295}{85443} \times 100 = 99.83\% \tag{1}$$

- **Fraudulent Transactions (Support: 148):**

$$\text{Percentage} = \frac{148}{85443} \times 100 = 0.17\% \tag{2}$$

This imbalance is a clear representation of practical data scarcity, where the availability of examples for critical classifications (fraudulent cases) is minimal.

5.4.1 Experimental Results

We used SMOTE to create synthetic examples for the minority class. A Random Forest Classifier was trained on both the original and the augmented dataset, and its quality was assessed. Table 11 shows the main performance for the augmented dataset:

Table 11: Model performance metrics after applying SMOTE

	Precision %	Recall %	F1-score %	Support
0	100.0	100.0	100.0	85,295
1	89.2	78.4	83.5	148
Accuracy	–	–	99.9	–
Macro avg	94.6	89.2	91.7	85,443
Weighted avg	99.9	99.9	99.9	85,443

- 0 represents the Normal (Non-Fraudulent Transactions) class.
- 1 represents the Fraudulent Transactions class.
- Support refers to the number of true instances for each class in the dataset. It represents the count of actual samples in the test set for each class (0 for non-fraudulent transactions and 1 for fraudulent transactions).

5.4.2 Metric Definitions

A. Precision:

Precision measures the number of correctly identified instances of fraud divided by the total number of cases labeled as fraudulent. This particular measure focuses on how often the model is correct when it says it is positive, as shown in Eq. (3).

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

For example, in the Credit Card Fraud Detection Dataset, 0.17% of transactions are fraudulent. Second, precision is essential so that flagged transactions are indeed fraudulent and the false alarm rate (fraudulent transactions classified as non-fraudulent transactions) is minimized.

Application to Experimental Results:

The model achieved a precision of 89.00% for fraudulent transactions. That means 89% of transactions that the model flagged as fraudulent were fraudulent, showing it to be effective in reducing the false positive rate. This level of precision can be beneficial in fraud detection systems, where false positives can waste lots of time and resources.

B. Recall:

Recall the number of actual fraudulent transactions detected by the model. This metric is critical for grasping how well the model can capture low-probability events, as shown in Eq. (4).

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

In fraud detection, recall is essential because the cost of missing fraudulent transactions (false negatives) could lead to financial losses and undermine trust in the system.

It resulted in a 78.00% recall for fraudulent transactions. This means the model successfully detected 78% of all fraudulent cases. While this shows improvement, some fraudulent cases were still missed, highlighting the need for further optimization.

C. F1-Score:

The F1-Score, designed as a harmonic mean of precision and recall, offers a means to manage the trade-off between these measures. This is especially useful in imbalanced datasets where both of these metrics are important; the formula for the F1-score is typically given as Eq. (5), is:

$$F1\text{-score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{5}$$

A common metric used for diagnosis datasets with high-class imbalance (for example, fraud detection) is the F1-score because it provides a comprehensive evaluation of the model performance without weighing precision or recall too heavily.

The F1-Score of fraudulent transactions is 83.00%, indicating that the trade-off between precision and recall is balanced. This also shows that our model can detect fraud but does not raise many false positives.

These results confirm the efficiency of SMOTE, which improves the model's ability to recognize rare events while preserving high accuracy for all other cases.

D. Accuracy

Accuracy measures the proportion of all transactions (both normal and fraudulent) that the model correctly classified; the formula for accuracy is given in Eq. (6) below.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Because of the imbalance in the dataset, the accuracy can be misleading on its own, as it depends too much on the majority class.

The model reached 99.94% accuracy, showing that almost all transactions were classified correctly. However, such a high accuracy only represents the model's performance on the majority class (normal transactions) and is not indicative of its capability to detect fraudulent transactions.

5.5 Implications of the Case Study

Addressing Practical Data Scarcity

This case study shows that data scarcity can be alleviated substantially through synthetic data generation techniques like SMOTE. By generating more samples of the minority class, the model was able to:

1. **Reduce Bias (Recall—Fraud Transactions):** Overcome the inherent bias of the model towards the majority class.
2. **Enhance Generalization**: The model achieved high precision and improved recall for rare events, enhancing its ability to generalize across datasets.

Applicability across Domains

The success of SMOTE, in this case, has broader implications:

- In **healthcare,** synthetic data can help detect rare diseases with limited availability.
- In **manufacturing**, synthetic data can be used to enhance anomaly detection when a few failures occur in a production line.
- In **finance**, fraud detection systems are improved through oversampling imbalanced datasets.

5.6 Case Study: Evaluating AI Performance in Medical Diagnosis

Though theoretically, the development of AI appears promising, a number of real-world applications regarding health care have shown the effectiveness of these technologies in the real world. Applying a RandomForestClassifier to the Breast Cancer Wisconsin dataset represents another case study demonstrating how AI approaches the dual challenge of data sparsity and decision ethics in diagnostics. The model yielded high precision and recall with a balanced dataset of benign and malignant cases. This could be helpful in the early detection of a disease and its accurate diagnosis, which is one of the most important aspects of any patient care and treatment. Applications of this type constitute milestones in transforming healthcare by equipping medical professionals with reliable tools for decision-making. This case study describes an application of RandomForestClassifier using the Breast Cancer Wisconsin dataset [69] to estimate a diagnosis as benign or malignant. The dataset [69] is a multivariate one, consisting of 569 samples with 30 features each, which are further divided into a training set of 70% and a test set of 30%. The model RandomForest was trained and afterward tested on the test set in order to calculate the precision, recall, f1-score, and support of both diagnostic classes. Results, represented in Table 12, suggest very high accuracy regarding the diagnostic capability of the model: 98.33% precision and 93.65% recall for malignant cases and precision of 96.39% with a recall of 99.07% for benign cases. These metrics underline the robustness of the model, with F1-scores of 95.93% for malignant and 97.72% for benign diagnoses, therefore postulating that this model is accurate and reliable in distinguishing the two conditions quite well, as seen from Table 12. The support values are 63 for malignant and 108 for benign, showing the number of instances evaluated in each class and the balance in the dataset used.

Table 12: Classifier performance metrics

Diagnosis	Precision %	Recall %	F1-score %	Support
Malignant	98.33	93.65	95.93	63
Benign	96.39	99.07	97.72	108
Overall	Accuracy: 97.07	Macro avg: 97.36	Weighted avg: 97.11	

The above case study shows the vast potential of machine learning to improve diagnostic accuracy in the medical field. Therefore, it justifies the development of AI tools that can support clinical decision-making and hopefully improve patient outcomes.

Ethical Considerations

Although synthetic data resolves most scarcity-related issues, it introduces its own set of potential risks:

- **Bias Propagation:** Synthetic samples can also inherit biases from the original data.
- **Reduced Realism:** Generative data may not capture the intricacies of actual environmental interactions, leading to diminished fidelity of predictions when used for real-world applications.

This raises the need for synthetic data to be utilized along with a strong validation mechanism and an ethical framework, as discussed in Section 3.4.

5.7 Proposed Solutions and Their Applications

The pursuit of Artificial General Intelligence (AGI) and agentic AIs entering many industries as a new type of workforce are some of the hottest topics in AI-related spaces in the mid-2020s. Such well-known AI personas as Altman, Huang, Sutskever et al. claim that AGI has already been achieved and/or they know exactly how to do it [70–74]. While this is currently a top-secret, the possible solutions are those presented in Fig. 2 or a combination of them, which is even more likely. Top Large Language Models (LLMs) confirmed the possible solutions through the short survey to gather their input.

Figure 2: Top possible solutions to the data scarcity problem

As can be seen from Fig. 2, the top option with the highest weight is Synthetic Data Generation. The authors agree with this direction [74] and believe that creating high-quality synthetic data is impossible without AI-human collaboration or, in other words, a Human-in-the-loop.

5.7.1 Synthetic Data Creation with Human-in-the-Loop

While some models can be specially trained to generate synthetic data, others can verify them and then be used by top LLMs; without people involved, this process will make no sense. We will see something like software development without a client. Fig. 3 represents a Human-in-the-Loop Synthetic Data Workflow.

As apparent in Fig. 3, humanity will collaborate with AI as top content creators and algorithm developers to achieve the best possible outcome. This highly will likely include quantum computing and other ways of compressing models and learning more and better from fewer data; all approaches to distributed machine learning, such as federated learning (See Appendix A, Table A1), are obviously on a plate.

The integration of AI systems should be embedded, at the development stage itself, with privacy design principles for better practicality. This could include techniques such as differential privacy (See Appendix A, Table A1) in data processing, for which no trace of individual data points can be traced to owners. Similarly, clear policies on data governance, including consent management, data access rights, and transparency regarding data usage, would help foster trust and enforce ethics. Another benefit would be setting up an Ethics Review Committee for AI projects to ensure that the highest ethics are considered while overseeing complex issues.

5.7.2 Smaller Language Models

One of the most promising solutions for optimizing AI efficiency is Smaller Language Models (SLMs), which have gained significant attention due to their ability to operate effectively in resource-limited environments. Unlike Large Language Models (LLMs), which require extensive computational power, SLMs provide a more practical, cost-effective, and scalable approach to AI implementation. This makes them particularly

suitable for deployment in low-resource settings, such as mobile devices, edge computing, and cloud-based platforms. Recently released by Microsoft, Phi-4 stands out as one of the leading SLMs demonstrating the potential of efficient AI models. It is a 40-layer, transformer-based model with a hidden size of 5120, supporting over 100k token embeddings and containing 14.1 billion trainable parameters. Unlike traditional LLMs, Phi-4 achieves high performance while significantly reducing resource consumption, making it a viable alternative for real-world applications. Fig. 4 illustrates Phi-4 running on Google Colab, where the model was executed on the PRO+ tier of the notebook (on an A100 runtime). The entire process was completed within 751.6 s, demonstrating high efficiency.

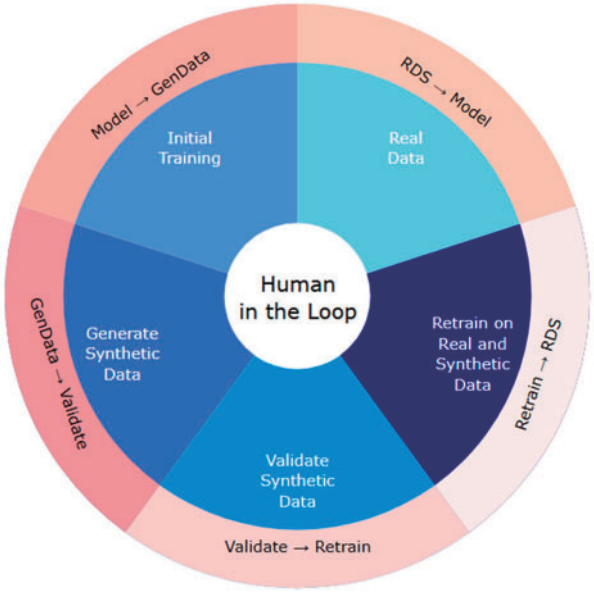

Figure 3: Human-in-the-loop synthetic data workflow

```python
import transformers

pipeline = transformers.pipeline(
    "text-generation",
    model="microsoft/phi-4",
    model_kwargs={"torch_dtype": "auto"},
    device_map="auto",
)

messages = [
    {"role": "system", "content": "You are an AI expert specializing in solutions for data scarcity in machine learning."},
    {"role": "user", "content": "What are the top solutions for addressing data scarcity in AI?"},
]

outputs = pipeline(messages, max_new_tokens=128)
print(outputs[0]["generated_text"][-1])
```

Figure 4: Phi-4 SLM running on Google Collab

The accessibility of SLMs like Phi-4 is a major advantage, allowing researchers and developers to run highly capable AI models on cloud platforms at a lower cost. For example, the Google Colab PRO+ subscription provides access to high-performance GPUs (Graphics Processing Units) for just USD 50 monthly, making SLMs an affordable research, development, and small-scale production solution. Phi-4 has demonstrated strong performance in practical applications across multiple NLP tasks, including text classification, summarization, question-answering (Q&A), and Retrieval-Augmented Generation (RAG). Notably, Phi-4 can suggest solutions to data scarcity challenges, utilizing techniques such as Synthetic Data Generation, Few-Shot and Transfer Learning, Data Augmentation, Self-Supervised Learning, Federated Learning, Zero-Shot Learning, and Privacy-Preserving AI Techniques (See Appendix A, Table A1). Additionally, SLMs have proven highly effective in RAG-based applications, significantly enhancing document summarization and knowledge retrieval. Researchers have successfully integrated Phi-4 into RAG, GraphRAG, and LazyGraphRAG applications [19,45,75,76], enabling AI models to interact with structured knowledge bases. These applications allow users to upload documents (e.g., PDFs, TXT files) and query them interactively, making AI more versatile and practical for real-world data processing.

Fig. 5 presents a visual representation of the GraphRAG approach [19], illustrating how it structures relationships within a document to enhance AI's ability to retrieve, comprehend, and summarize structured knowledge. The diagram showcases interconnected nodes, representing key concepts and their relationships, which help improve contextual understanding and efficient retrieval of information.

Figure 5: A visual representation of the GraphRAG approach, illustrating how a document's content is structured into a graph-based format. This diagram highlights key concepts and their relationships, improving AI-driven retrieval, comprehension, and summarization of structured knowledge

Table 13 summarizes all three RAG-based approaches, comparing their core concepts, storage utilization, retrieval mechanisms, and efficiency trade-offs.

The emergence of Smaller Language Models (SLMs) marks a significant shift in AI development, offering a balance between performance, efficiency, and accessibility. Models like Phi-4 exemplify how resource-friendly AI can power advanced applications, such as Retrieval-Augmented Generation (RAG),

document summarization, and interactive knowledge retrieval. While challenges remain, ongoing research in knowledge adaptation, quantization, and efficient training methodologies will further enhance the capabilities and widespread adoption of SLMs across various domains. As AI evolves, SLMs will play an increasingly critical role in democratizing machine learning access, making AI applications more scalable, efficient, and environmentally sustainable.

Table 13: RAG methods comparison

Feature	RAG	Full graph RAG	Lazy graph RAG
Concept	It uses a retriever-generator model to fetch and process text chunks.	Organizes information in a graph structure, improving relational understanding.	"Lazily" explores or expands the graph at query time, retrieving only the necessary subgraph.
Storage	Uses dense vector indexes for direct chunk retrieval.	Stores entities, documents, and relationships as graph nodes & edges.	Minimizes memory footprint by loading only necessary segments.
Retrieval	Searches for top-k text chunks and generates an answer.	Traverses graph relationships to extract relevant context.	Selects relevant nodes dynamically, reducing unnecessary retrieval overhead.
Efficiency	Fast, but lacks deep contextual relationships.	It is more resource-intensive, as graph traversal requires extra computations.	Optimized for efficiency, balancing context depth and computational cost.
Context quality	Depending on the chunk ranking, it may lose relational meaning.	Captures document relationships, improving contextual understanding.	Retains graph-based advantages while reducing computational load.

5.7.3 Quantization and Pruning

Quantization and pruning are key optimization techniques that significantly reduce machine learning models' size and computational cost, making AI more accessible and sustainable. These methods are beneficial for Small Language Models (SLMs) like Phi-4, which aim to achieve high efficiency without compromising performance. Quantization converts high-precision numerical values (32-bit floating points) into lower-precision formats (8-bit or 4-bit), reducing memory usage and power consumption. Meanwhile, pruning removes unnecessary parameters (weights, neurons, or channels) from a model, reducing complexity and computational load while maintaining accuracy.

Both methods align with the Green AI approach [18], ensuring lower energy consumption, faster inference speeds, and minimal hardware requirements while keeping AI models scalable for real-world applications. Table 14 presents the latest quantization and pruning approaches applied to Phi-4 (as previously tested on Phi-1.5 and Phi-2 models) and detailed explanations of their key features.

The significance of these techniques lies in their ability to reduce model size and computational overhead, leading to faster real-time responses. Additionally, they play a crucial role in minimizing energy consumption, contributing to the sustainability goals outlined in Green AI research. As AI models continue to evolve, smaller, more efficient architectures will lower the cost of AI deployment, making advanced machine learning more widely accessible; even though quantization allows AI models to store and process numerical data efficiently while pruning eliminates redundant connections, further research is required to evaluate their energy efficiency across different AI architectures. This is especially relevant for Large Language Models (LLMs) and SLMs, where balancing performance, computational efficiency, and accuracy remains an ongoing challenge.

Table 14: Quantization and pruning methods comparison

Approach	Library/Tool	Precision/Sparsity	Key features
4-bit Quant (NF4)	BitsAndBytes (bnb) + Hugging Face Transformers	4-bit Weights (NormalFloat4)	Maximizes memory savings while maintaining good accuracy retention. Used in LLMs & SLMs for extreme efficiency.
8-bit Quant (LLM.int8())	BitsAndBytes + Accelerate/HF Transformers	8-bit Matrix Multiplications	Reduces GPU memory usage significantly with a minor accuracy drop vs. FP16. Best for general AI applications.
Dynamic Quant (8-bit/16-bit)	Native PyTorch Quantization	8-bit or 16-bit (activations/weights)	Applies on-the-fly quantization, requiring minimal code changes. Accuracy may vary depending on the model's sensitivity. Suitable for low-power devices.
Quantization-Aware Training (QAT)	PyTorch or TF Model Optimization	8-bit or 16-bit (weights + activations)	Simulates quantization during forward/backward, yields higher accuracy, more complex setup.
Pruning	PyTorch Pruning Utilities	Any model/layer (set weights to 0)	Simulates quantization effects during training, improving accuracy in low-precision models. Used in production AI applications.

5.8 The Future of AI with Synthetic Data

This study underscores the significance of synthetic data generation as a pivotal technological innovation. Synthetic data addresses one of the most pressing challenges in AI development by enabling AI to function effectively in domains with limited access to natural data. Focusing on enhancing recall and overall model efficacy further emphasizes the benefits of complementing conventional model training with synthetic data to surmount challenges associated with natural data availability. As AI systems continue to extend into the areas where lack of data becomes a bottleneck, hybrid schemes utilizing synthetic data, transfer learning, and few-shot learning will gain even more traction. This study illustrates that you can bring AI to reality in low-resource environments and achieve responsible and sustainable AI development goals.

Developers can overcome data bottlenecks and improve model generalization by carefully generating diverse and representative synthetic samples.

Small language models [13–15] are designed with fewer parameters, reducing their reliance on massive datasets compared to larger models. Their compact architecture allows for effective training with limited data, making them suitable for niche applications or domains where data collection is challenging. SLMs also provide advantages in terms of reduced computational cost and faster inference, further enhancing their practicality.

Quantization and pruning help mitigate data scarcity by enabling the deployment of models with reduced memory footprints and lower computational demands. Quantization achieves this by representing model weights and activations with lower precision, while pruning removes less important connections in the network. Consequently, these techniques allow for efficient training and deployment of models even with limited data, as the reduced model complexity lessens the risk of overfitting and improves generalization on smaller datasets.

5.9 Future Research & Considerations

Future research could explore how AI integrates with emerging technologies such as quantum computing, which might enhance model training efficiency and solve complex optimization problems faster than classical methods. This integration could be particularly beneficial in data-intensive domains like drug discovery [29] and financial modeling [74]. However, practical implementations remain a challenge and require further investigation. Additionally, research should focus on dynamic AI governance frameworks (See Appendix A, Table A1), such as the EU's Data Governance Act, which can adapt to the rapid pace of technological change [16]. Future work could explore how regulatory sandbox environments (See Appendix A, Table A1) allow AI systems to be tested while ensuring ethical compliance and risk mitigation. Another promising area is advancing synthetic data generation techniques. While current methods like Generative Adversarial Networks (GANs) [77] and diffusion models are effective, they still struggle with maintaining diversity, realism, and fairness. Future research could focus on hybrid approaches that combine synthetic data with human feedback (Human-in-the-Loop AI) to improve data quality [78]. This collaborative approach leverages human expertise to guide AI systems, enhancing the realism and applicability of synthetic data [79]. Moreover, Green AI, as discussed in this paper, is gaining importance in developing computationally efficient models [76]. Investigating techniques like model pruning, quantization, and smaller AI architectures like Small Language Models (SLMs) like Phi-4 and Mistral could provide sustainable AI solutions for low-resource environments. These approaches aim to reduce the environmental impact of AI development while maintaining performance and adaptability. Addressing these areas will ensure that AI remains ethical, efficient, and adaptable in overcoming data scarcity challenges.

6 Conclusion

As artificial intelligence continues to evolve, the industry must proactively develop sustainable, data-efficient systems to address the challenges posed by data scarcity. Developing efficient algorithms that maximize the value of each data point will be essential as access to high-quality natural data becomes increasingly limited. This requires shifting the focus of AI development from simply accumulating large datasets to curating diverse, ethically sourced, high-quality data that ensures fairness, reliability, and adaptability.

To achieve this, AI systems must be designed with dynamic governance frameworks that can adapt to evolving ethical and regulatory landscapes, ensuring responsible development and deployment. Regulatory sandbox environments, as discussed, offer structured mechanisms to test AI models under controlled conditions, fostering both compliance and innovation.

Moreover, the AI industry must embrace computationally efficient techniques to minimize environmental impact while maintaining scalability. Strategies such as model pruning, quantization, and Small Language Models (SLMs) provide sustainable solutions, ensuring AI remains accessible even in low-resource environments. Simultaneously, quantum computing is expected to revolutionize AI, particularly in data-intensive fields. While these advancements hold promise, they also introduce new technical and ethical challenges that necessitate ongoing research and interdisciplinary collaboration.

Ultimately, the future of AI depends on a balanced integration of technological innovation, ethical supervision (See Appendix A, Table A1), and sustainable data strategies. By prioritizing responsible AI governance, optimized data utilization, and energy-efficient models, the AI industry can ensure equitable, ethical, and adaptive progress in overcoming data scarcity challenges and fostering AI's long-term sustainability.

Acknowledgement: The authors gratefully acknowledge the financial support from Wenzhou-Kean University.

Funding Statement: This work was supported by Internal Research Support Program (IRSPG202202). It is important to note that the USA team involved in this project was not funded by any China-based grants.

Author Contributions: Conceptualization: Hemn Barzan Abdalla, Yulia Kumar. Investigation: Hemn Barzan Abdalla, Ardalan Awlla, Maryam Cheraghy, Jose Marchena, Mehdi Gheisari. Methodology: Hemn Barzan Abdalla, Jose Marchena, Stephany Guzman. Formal analysis: Hemn Barzan Abdalla, Yulia Kumar, Jose Marchena, Stephany Guzman. Writing—original draft: Hemn Barzan Abdalla, Yulia Kumar, Jose Marchena, Stephany Guzman, Maryam Cheraghy. Writing—review & editing: Hemn Barzan Abdalla, Yulia Kumar, Ardalan Awlla. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All data based on references.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Appendix A

Table A1: Key vocabulary definitions

Term	Definition
AI governance frameworks	Policies, regulations, and ethical guidelines that ensure AI systems are developed and deployed responsibly.
Data imbalance	A common issue in AI datasets where one class of data (e.g., fraudulent transactions) is significantly underrepresented compared to another, leading to biased model predictions.
Data scarcity	The lack of sufficient labeled training data challenges the AI model performance and requires alternative strategies like transfer learning, synthetic data, and self-supervised learning.
Differential privacy	A data protection technique that ensures individual user data remains anonymous while still enabling AI model training.
Ethical supervision	Rules and guidelines ensure that AI systems are fair, unbiased, and used responsibly to avoid harm.
Federated learning	A method that allows AI to learn from data on different devices without collecting or storing it in one place, keeping user information private.

(Continued)

Table A1 (continued)

Term	Definition
Green AI	An approach to AI that prioritizes energy efficiency, sustainability, and minimizing environmental impact.
Privacy-preserving AI techniques	AI systems are designed to process and analyze data while maintaining user privacy, using techniques like federated learning and differential privacy to minimize risks.
Reinforcement learning	A machine learning paradigm is one in which an AI agent learns by interacting with its environment and receiving feedback in the form of rewards or penalties.
Regulatory sandbox	A controlled testing environment that allows AI developers to experiment with models while ensuring good performance with ethical standards.
SMOTE	SMOTE (Synthetic Minority Over-sampling Technique) is a method for balancing imbalanced datasets by generating synthetic samples for underrepresented classes.
Zero-shot learning	An AI capability that allows models to make predictions on data they have never seen before by transferring knowledge from related tasks.

References

1. Singh J. The rise of synthetic data: enhancing AI and machine learning model training to address data scarcity and mitigate privacy risks. J Artif Intell Res Appl. 2021;1(2):292–332.
2. Panagiotou E, Qian H, Marx S, Ntoutsi E. Generative AI based augmentation for offshore jacket design: an integrated approach for mixed tabular data generation under data scarcity and imbalance. [cited 2025 Jan 1]. Available from: https://doi.org/10.2139/ssrn.4703856.
3. Zhang CW, Pan R, Goh TN. Reliability assessment of high-quality new products with data scarcity. Int J Prod Res. 2021;59(14):4175–87. doi:10.1080/00207543.2020.1758355.
4. Danaher J, Nyholm S. Digital duplicates and the scarcity problem: might AI make us less scarce and therefore less valuable? Philos Technol. 2024;37(3):106. doi:10.1007/s13347-024-00795-z.
5. Shen JT. AI in education: effective machine learning methods to improve data scarcity and knowledge generalization. University Park, PA, USA: The Pennsylvania State University; 2023.
6. Bai J, Alzubaidi L, Wang Q, Kuhl E, Bennamoun M, Gu Y. Utilising physics-guided deep learning to overcome data scarcity. arXiv:2211.15664. 2022.
7. Kaai K. Addressing data scarcity in domain generalization for computer vision applications in image classification. Waterloo, ON, Canada: University of Waterloo; 2024.
8. Sufi F. Addressing data scarcity in the medical domain: a GPT-based approach for synthetic data generation and feature extraction. Information. 2024;15(5):264. doi:10.3390/info15050264.
9. Ramakrishnan R. How to build good AI solutions when data is scarce. MIT Sloan Manag Rev. 2022;64(1):1–9.
10. Aldoseri A, Al-Khalifa KN, Hamouda AM. Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges. Appl Sci. 2023;13(12):7082. doi:10.3390/app13127082.
11. Nandy A, Duan C, Kulik HJ. Audacity of huge: overcoming challenges of data scarcity and data quality for machine learning in computational materials discovery. Curr Opin Chem Eng. 2022;36:100778. doi:10.1016/j.coche.2021.100778.
12. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. arXiv:2303.08774. 2023.

13. Aghajanyan A, Yu L, Conneau A, Hsu WN, Hambardzumyan K, Zhang S, et al. Scaling laws for generative mixed-modal language models. In: International Conference on Machine Learning; 2023 Jul 23–29; Honolulu, HI, USA. p. 265–79.
14. Villalobos P, Ho A, Sevilla J, Besiroglu T, Heim L, Hobbhahn M. Position: will we run out of data? Limits of LLM scaling based on human-generated data. In: Forty-First International Conference on Machine Learning; 2024 Jul 21–27; Vienna, Austria.
15. Villalobos P, Ho A, Sevilla J, Besiroglu T, Heim L, Hobbhahn M. Will we run out of data? Limits of LLM scaling based on human-generated data. arXiv:2211.04325. 2022. doi:10.48550/arXiv.2211.04325.
16. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: scaling language modeling with pathways. J Mach Learn Res. 2023;24(240):1–13.
17. Hausen R, Azarbonyad H. Discovering data sets through machine learning: an ensemble approach to uncovering the prevalence of government-funded data sets. Harv Data Sci Rev. 2024;4:1–18. doi:10.1162/99608f92.
18. Babbar R, Schölkopf B. Data scarcity, robustness and extreme multi-label classification. Mach Learn. 2019;108(8):1329–51. doi:10.1007/s10994-019-05791-5.
19. Larson J, Steven Truitt S. GraphRAG: unlocking LLM discovery on narrative private data. [cited 2025 Jan 15]. Available from: https://www.microsoft.com/en-us/research/blog/graphrag-unlocking-llm-discovery-on-narrative-private-data/.
20. Pons G, Bilalli B, Abelló A, Blanco Sánchez S. On the use of trajectory data for tackling data scarcity. Inf Syst. 2025;130(5):102523. doi:10.1016/j.is.2025.102523.
21. Karst F, Li M, Reinhard P, Leimeister J. Not enough data to be fair? Evaluating fairness implications of data scarcity solutions. In: Proceedings of the 58th Hawaii International Conference on System Sciences; 2025 Jan 7–10; Big Island, HI, USA.
22. Chen Z, Gan W, Wu J, Hu K, Lin H. Data scarcity in recommendation systems: a survey. ACM Trans Recomm Syst. 2025;3(3):1–31. doi:10.1145/3700890.
23. Khalil M, Vadiee F, Shakya R, Liu Q. Creating artificial students that never existed: leveraging large language models and CTGANs for synthetic data generation. arXiv:2501.01793. 2025. doi:10.1145/3706468.
24. White J, Madaan P, Shenoy N, Agnihotri A, Sharma M, Doshi J. A case for rejection in low resource ML deployment. arXiv:2208.06359. 2022.
25. Singh P. Systematic review of data-centric approaches in artificial intelligence and machine learning. Data Sci Manag. 2023;6(3):144–57. doi:10.1016/j.dsm.2023.06.001.
26. Li D, Akashi K, Nozue H, Tayama K. A mirror environment to produce artificial intelligence training data. IEEE Access. 2022;10:24578–86. doi:10.1109/ACCESS.2022.3154825.
27. Lautrup AD, Hyrup T, Zimek A, Schneider-Kamp P. SynthEval: a framework for detailed utility and privacy evaluation of tabular synthetic data. Data Min Knowl Discov. 2025;39(1):1–25. doi:10.1007/s10618-024-01081-4.
28. Gangwal A, Ansari A, Ahmad I, Azad AK, Wan Sulaiman WMA. Current strategies to address data scarcity in artificial intelligence-based drug discovery: a comprehensive review. Comput Biol Med. 2024;179:108734. doi:10.1016/j.compbiomed.2024.108734.
29. Niel O. A novel algorithm can generate data to train machine learning models in conditions of extreme scarcity of real world data. arXiv:2305.00987. 2023.
30. Chen YT, Hsu CY, Yu CM, Barhamgi M, Perera C. On the private data synthesis through deep generative models for data scarcity of industrial Internet of Things. IEEE Trans Ind Inform. 2023;19(1):551–60. doi:10.1109/TII.2021.3133625.
31. Bansal MA, Sharma DR, Kathuria DM. A systematic review on data scarcity problem in deep learning: solution and applications. ACM Comput Surv. 2022;54(10s):1–29. doi:10.1145/3502287.
32. Nahid MMH, Bin Hasan S. SafeSynthDP: leveraging large language models for privacy-preserving synthetic data generation using differential privacy. arXiv:2412.20641. 2024.
33. Zou MX, Lo CC, Lin CH, Shieh CS, Horng MF. 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; 2022 Dec 16–18; Kitakyushu, Japan. Singapore: Springer Nature Singapore; 2022. p. 347–57. doi:10.1007/978-981-99-0105-0_31.
34. Hoang V. Mitigating data scarcity for large language models. arXiv:2302.01806. 2023.
35. Alzubaidi L, Bai J, Al-Sabaawi A, Santamaría J, Albahri AS, Al-dabbagh BSN, et al. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. J Big Data. 2023;10(1):46. doi:10.1186/s40537-023-00727-2.
36. Laurer M, van Atteveldt W, Casas A, Welbers K. Less annotating, more classifying: addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. Polit Anal. 2024;32(1):84–100. doi:10.1017/pan.2023.20.
37. Quito JA, Andrade LJ. Proposal for the generation of profiles using a synthetic database. In: AHFE, 2022 Conference on Applied Human Factors and Ergonomics International; 2022 Jul 24–28; New York City, NY, USA. doi:10.54941/ahfe1001462.
38. Oxford Analytica. Data scarcity challenges AI developers globally. Bingley, UK: Emerald Expert Briefings; 2024.
39. Huang T, Luo T, Yan M, Zhou JT, Goh R. RCT: resource constrained training for edge AI. IEEE Trans Neural Netw Learn Syst. 2022;35(2):2575–87. doi:10.1109/TNNLS.2022.3190451.
40. [cited 2025 Mar 12]. Available from: https://brusselsprivacyhub.com/wp-content/uploads/2024/02/Personal-Data-Protection-in-Brazil.pdf.
41. [cited 2025 Mar 12]. Available from: https://www.novaoneadvisor.com/report/artificial-intelligence-market.
42. [cited 2025 Mar 12]. Available from: https://ourworldindata.org/data-insights/investment-in-generative-ai-has-surged-recently.
43. [cited 2025 Mar 12]. Available from: https://hai.stanford.edu/news/ai-index-state-ai-13-charts.
44. Healy M, Baum A, Musumeci F. Addressing data scarcity in ML-based failure-cause identification in optical networks through generative models. Opt Fiber Technol. 2025;90(12):104137. doi:10.1016/j.yofte.2025.104137.
45. [cited 2025 Mar 12]. Available from: https://www.fbcinc.com/source/virtualhall_images/2024_Virtual_Events/USDA_Innovation/Databricks/State_of_Data___AI_Resource.pdf.
46. Belkada Y, Dettmers T, Pagnoni A, Gugger S, Mangrulkar S. Making LLMs even more accessible with bitsandbytes, 4-bit quantization and QLoRA. [cited 2025 Jan 15]. Available from: https://huggingface.co/blog/4bit-transformers-bitsandbytes.
47. Abdalla HB. A brief survey on big data: technologies, terminologies and data-intensive applications. J Big Data. 2022;9(1):107. doi:10.1186/s40537-022-00659-3.
48. Abdalla HB, Abuhaija B. Comprehensive analysis of various big data classification techniques: a challenging overview. J Info Know Mgmt. 2023;22(1):2250083. doi:10.1142/S0219649222500836.
49. Abdalla HB, Awlla AH, Kumar Y, Cheraghy M. Big data: past, present, and future insights. In: Proceedings of the 2024 Asia Pacific Conference on Computing Technologies, Communications and Networking; 2024 Jul 26–27; Chengdu, China. p. 60–70. doi:10.1145/3685767.3685777.
50. [cited 2025 Mar 12]. Available from: https://www.sciencealert.com/the-world-is-running-out-of-data-to-feed-ai-experts-warn.
51. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ Digit Med. 2023;6(1):186. doi:10.1038/s41746-023-00927-3.
52. Dang H, Mecke L, Lehmann F, Goller S, Buschek D. How to prompt? Opportunities and challenges of zero- and few-shot learning for human-AI interaction in creative applications of generative models. arXiv:2209.01390. 2022.
53. Yang S, Xiao W, Zhang M, Guo S, Zhao J, Shen F. Image data augmentation for deep learning: A survey. arXiv:2204.08610. 2022.
54. Abdulmalek S, Nasir A, Jabbar WA, Almuhaya MAM, Bairagi AK, Khan MA, et al. IoT-based healthcare-monitoring system towards improving quality of life: a review. Healthcare. 2022;10(10):1993. doi:10.3390/healthcare10101993.
55. Goyal M, Mahmoud QH. A systematic review of synthetic data generation techniques using generative AI. Electronics. 2024;13(17):3509. doi:10.3390/electronics13173509.

56. Rani V, Nabi ST, Kumar M, Mittal A, Kumar K. Self-supervised learning: a succinct review. Arch Comput Methods Eng. 2023;30(4):2761–75. doi:10.1007/s11831-023-09884-2.
57. Li W, Huang R, Li J, Liao Y, Chen Z, He G, et al. A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: theories, applications and challenges. Mech Syst Signal Process. 2022;167(12):108487. doi:10.1016/j.ymssp.2021.108487.
58. [cited 2025 Mar 12]. Available from: https://technologymagazine.com/articles/how-googles-ai-breakthroughs-earned-nobel-recognition.
59. [cited 2025 Mar 12]. Available from: https://research.google/blog/automl-for-large-scale-image-classification-and-object-detection.
60. [cited 2025 Mar 12]. Available from: https://innovationexchange.mayoclinic.org/an-introduction-to-federal-funding-for-innovation/.
61. [cited 2025 Mar 12]. Available from: https://www.techmonitor.ai/risks/extreme-weather-events/ibm-weather-and-climate-model?cf-view.
62. Lund B. Universities engaging social media users: an investigation of quantitative relationships between universities' Facebook followers/interactions and university attributes. J Mark High Educ. 2019;29(2):251–67. doi:10.1080/08841241.2019.1641875.
63. Miller T, Durlik I, Kostecka E, Borkowski P, Łobodzińska A. A critical AI view on autonomous vehicle navigation: the growing danger. Electronics. 2024;13(18):3660. doi:10.3390/electronics13183660.
64. [cited 2025 Mar 12]. Available from: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf.
65. [cited 2025 Mar 12]. Available from: https://oag.ca.gov/privacy/ccpa.
66. [cited 2025 Mar 12]. Available from: https://www.twobirds.com/en/insights/2024/china/ai-governance-in-china-strategies-initiatives-and-key.
67. [cited 2025 Mar 12]. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/laws.
68. [cited 2025 Mar 12]. Available from: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data.
69. [cited 2025 Mar 12]. Available from: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.
70. Kumar Y, Lin M, Paredes C, Li D, Yang G, Kruger D, et al. A comprehensive review of AI advancement using testFAILS and testFAILS-2 for the pursuit of AGI. Electronics. 2024;13(24):4991. doi:10.3390/electronics13244991.
71. Haruni R. Nvidia CEO maps out bold vision: AGI and robotics set to merge. [cited 2025 Jan 15]. Available from: https://wallstreetpit.com/120150-nvidia-ceo-maps-out-bold-vision-agi-and-robotics-set-to-merge/.
72. Barlow G. Altman predicts artificial superintelligence (AGI) will happen this year. [cited 2025 Jan 15]. Available from: https://www.techradar.com/computing/artificial-intelligence/sam-altman-predicts-artificial-superintelligence-agi-will-happen-this-year.
73. Robison K. OpenAI cofounder Ilya Sutskever says the way AI is built is about to change. [cited 2025 Jan 15]. Available from: https://www.theverge.com/2024/12/13/24320811/what-ilya-sutskever-sees-openai-model-data-training.
74. Kumar Y, Huang K, Perez A, Yang G, Li JJ, Morreale P, et al. Bias and cyberbullying detection and data generation using transformer artificial intelligence models and top large language models. Electronics. 2024;13(17):3431. doi:10.3390/electronics13173431.
75. Kumar Y, Marchena J, Awlla AH, Li JJ, Abdalla HB. The AI-powered evolution of big data. Appl Sci. 2024;14(22):10176. doi:10.3390/app142210176.
76. Edge D, Larson J. LazyGraphRAG: setting a new standard for quality and cost. [cited 2025 Jan 15]. Available from: https://www.microsoft.com/en-us/research/blog/lazygraphrag-setting-a-new-standard-for-quality-and-cost/.
77. Feng Y, Shen A, Hu J, Liang Y, Wang S, Du J. Enhancing few-shot learning with integrated data and GAN model approaches. arXiv:2411.16567. 2024.
78. Wiethof C, Bittner EA. Hybrid intelligence—combining the human in the loop with the computer in the loop: a systematic literature review. [cited 2025 Jan 1]. Available from: https://www.researchgate.net/publication/356209722.
79. Verdecchia R, Sallou J, Cruz L. A systematic review of green AI. Wiley Interdiscip Rev Data Min Knowl Discov. 2023;13(4):e1507. doi:10.1002/widm.1507.