ISSN 0315-3681 Volume 122 (2), 2025

Leveraging Large Language Models to Generate Training Data in Low-Resource Scenarios

Guneet Singh Kohli
Independent Researcher, USA

Abstract

The persistent scarcity of training data continues to be a major roadblock for machine learning applications in specialized domains and underrepresented languages. This article explores how Large Language Models (LLMs) are emerging as a promising solution to this problem by serving as synthetic data generators. When human-annotated data is unavailable or too expensive to obtain, LLMs can produce labeled examples that bootstrap model development. This paper examines several methodological approaches to synthetic data generation—instruction tuning, few-shot prompting, and chain-of-thought techniques—and their applications in zero-shot and few-shot learning contexts. The article also investigates how LLMs can facilitate data augmentation, error case generation, and counterfactual testing to improve model robustness and fairness. Despite their potential, significant challenges remain, including bias amplification, distribution mismatch between synthetic and real data, quality assurance concerns, and computational demands. Looking ahead, the article identifies promising research directions in controlled generation, hybrid data approaches, adaptive generation systems, standardized quality metrics, and domain adaptation techniques that may fundamentally change how businesses develop machine learning systems for specialized applications and underrepresented domains.

Keywords: Synthetic data generation, Large Language Models, Low-resource domains, Few-shot learning, Bias mitigation

1. Introduction

Machine learning practitioners face a persistent challenge: getting enough high-quality annotated data for specialized or low-resource domains. Research shows that LLMs can be effectively taught to use various tools through appropriate prompting strategies, suggesting they're flexible enough to generate structured outputs that conform to specific task requirements [1].

The synthetic data approach is particularly valuable for classification tasks, information extraction, and question-answering systems where human-labeled data remains prohibitively expensive or simply unavailable. The data scarcity problem appears across numerous specialized domains, including legal document analysis, medical entity recognition, and financial sentiment classification. In multilingual contexts, this challenge is magnified exponentially. While thousands of languages exist worldwide, robust training data is available for only a small fraction [10], creating significant disparities in AI accessibility and effectiveness across linguistic boundaries. This disparity extends beyond purely academic concerns—organizations implementing machine learning solutions frequently report that data acquisition and annotation represent the most substantial portion of project budgets and timeline extensions.

Recent studies have shown that synthetic data generated by LLMs can meaningfully address these challenges [2]. The ability of LLMs to generate diverse, contextually appropriate examples makes them valuable tools for expanding limited datasets or creating entirely new ones where none

previously existed. When properly instructed, these models can produce examples that capture the nuances and distribution characteristics needed for effective training of downstream models. Researchers have examined this capability in depth, comparing human annotations with those generated by LLMs across various natural language understanding tasks, finding that synthetic data can serve as a viable alternative or supplement to human-annotated data in many scenarios [2].

The economic benefits extend well beyond cost reduction. Faster development, shorter development cycles, and more effective resource distribution are also possible due to lowered dependence on costly human processes of annotations. This strategy will be noteworthy as it can democratize access to advanced machine learning capabilities. Research teams and developers that deal with underrepresented languages or interesting fields, thus far ignored by large technology enterprises, can now put in place advanced machine learning applications based on significantly fewer assets. Moreover, development process is fundamentally transformed. Teams can quickly prototype, test, and iterate on models with synthetic data instead of waiting for annotated data. Some research groups report significant reductions in development time when incorporating synthetic data into the workflow, highlighting its potential to accelerate prototyping and iteration.

This paper examines in detail the methodological approaches for utilizing Large Language Models as synthetic data generators – the methods that work, real-world applications, the thorny challenges still being wrestled with, and where this field is headed. Synthetic data has the potential to fundamentally transform how ML systems are developed for specialized domains and historically underserved languages. It is also among the most practical and impactful applications of large language models beyond conversational AI. The technology is evolving rapidly, so getting a handle on what these models can do, where they fall short, and how to get the most out of them isn't just academically interesting – it's becoming essential knowledge for anyone trying to break through the data bottlenecks that have held back progress for too long.

2. Methodology of Synthetic Data Generation

LLMs can be leveraged to generate training data through several established techniques. Instruction tuning involves providing the model with clear directives about the desired output format and content characteristics. Few-shot prompting demonstrates examples of the target task, allowing the model to infer patterns and produce similar instances. Chain-of-thought prompting guides the model through a reasoning process to generate more complex examples with appropriate rationales or explanations. Studies have shown that chain-of-thought prompting significantly enhances the quality of synthetic data by encouraging LLMs to articulate intermediate reasoning steps, creating a form of self-taught reasoning that results in more accurate and nuanced outputs across complex tasks, including arithmetic, commonsense reasoning, and symbolic manipulation [3].

The quality of synthetic data depends significantly on prompt engineering. Succinct prompts should give information about the domain knowledge, output form, the required variety of results, and the restrictions on the resulting content. As an example, a prompt may ask to create 10 instances of customer support inquiry regarding the network connectivity problem with sentiment labels (positive, negative, neutral). This methodology allows precise control over the structure and content of the generated dataset. Recent advances in prompt engineering have identified several critical factors that influence synthetic data quality, including prompt specificity, exemplar

selection, and instruction clarity. Studies examining various prompting strategies across both open and closed-source LLMs have revealed significant performance variations based on how instructions are formulated, with properly constructed prompts enhancing the alignment between generated data and the desired task specifications [4].

To ensure data quality, synthetic examples should undergo validation through several rigorous techniques. Cross-validation against available real data samples provides a benchmark for assessing the distribution alignment between synthetic and authentic examples. In this process, one usually calculates the similarity scores between features, e.g., KL-divergence or Earth Mover 2 states between feature distributions on each dataset. Consistency checks across generations can help identify and remove unstable or contradictory examples that might introduce noise during training. This approach leverages statistical analysis to detect outliers or examples that demonstrate high variance across repeated generations under identical prompting conditions [4]. For particularly sensitive applications, evaluation by domain experts remains essential despite its resource intensity. Expert validation has been shown to identify subtly incorrect examples that automated methods often miss, particularly in specialized domains such as healthcare, where factual accuracy carries significant consequences. Additionally, automated filtering based on predefined quality metrics offers a scalable approach to maintaining data integrity. Research examining instruction tuning across various model architectures suggests that implementing systematic quality control mechanisms is essential for maintaining consistency in synthetically generated datasets, particularly when scaling to large volumes of examples [4].

Synthetic data generation methodologies continue to evolve, with new techniques offering greater control over the generation process. Further advances can be expected in QoS of synthetic data quality using techniques such as controlled decoding, i.e., dynamic control of the parameters of the sampling scheme to retain characteristics of the desired distributions. The latter can be just as well illustrated in similar hierarchical generation tasks where complex examples are decomposed into manageable entities and subsequently re-composed with the use of such tasks including dialogue generation or multi-step logical reasoning. All these methodological breakthroughs answer a central challenge of coming up with synthetic examples that are close enough to the real thing to use, but diverse enough to allow good model training. Figure 1 summarizes these methodological stages of synthetic data generation via different approaches.

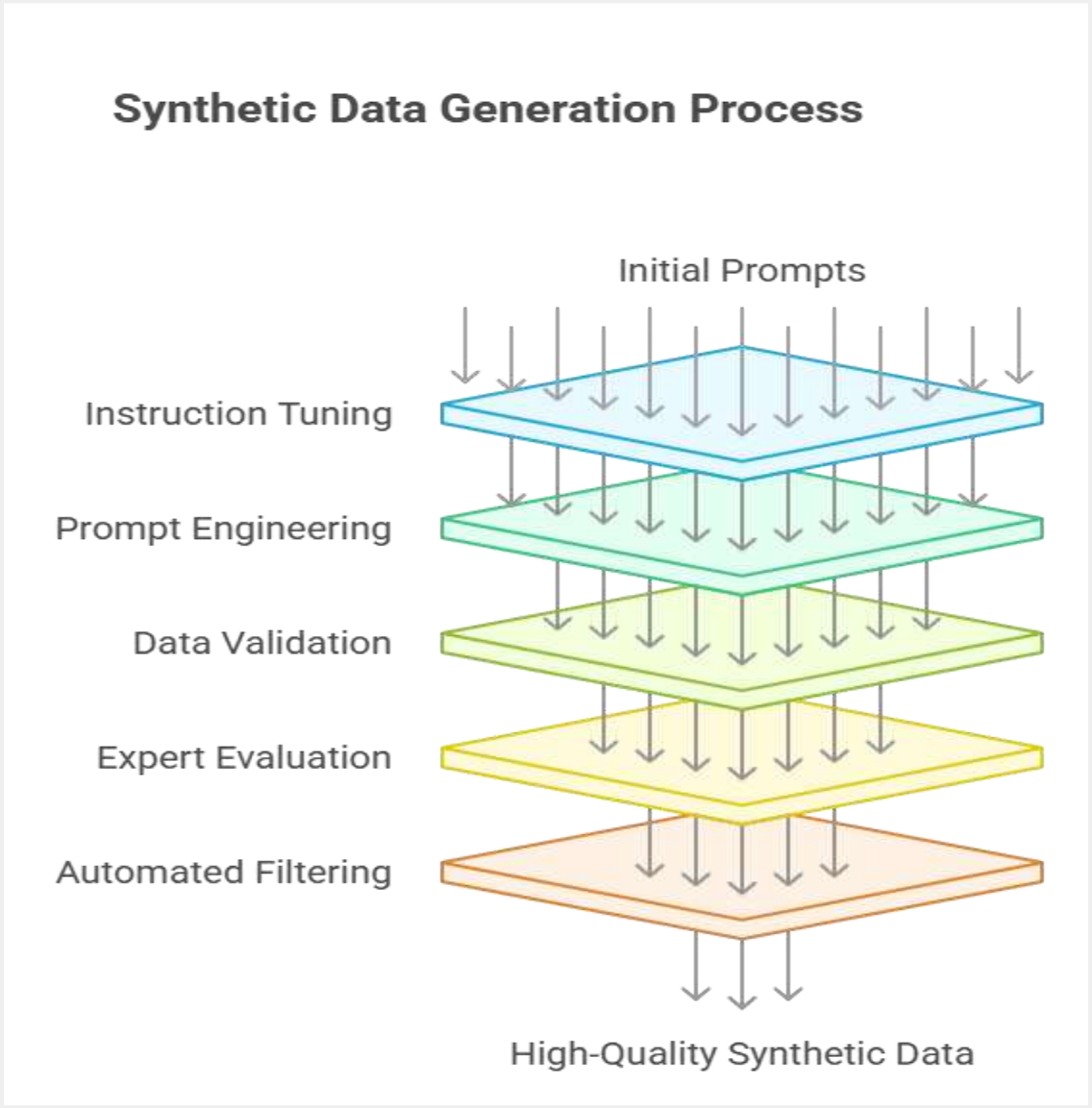

Figure 1: Synthetic Data Generation Process [3, 4]

3. Applications in Zero and Few-Shot Learning

In zero-shot learning scenarios, where no labeled examples exist for a target task, LLM-generated synthetic data serves as an initial bootstrap mechanism. This synthetic dataset provides a foundation for model training that would otherwise be impossible. For example, in rare disease diagnosis from clinical notes, an LLM can generate synthetic patient descriptions with appropriate diagnostic labels based on medical literature, enabling the development of specialized classification models despite the scarcity of real patient data. Research has demonstrated the efficacy of this approach across multiple domains, showing that large language models could generate synthetic examples for tasks they had never explicitly been trained on, producing data that captured essential patterns and relationships necessary for downstream model training [5]. Comprehensive evaluation across dozens of NLP tasks demonstrated that language models with sufficient scale can

perform zero-shot learning without explicit examples, suggesting these same capabilities can be leveraged to generate synthetic training data for novel tasks.

Zero-shot synthetic data generation offers particular advantages for low-resource languages and specialized technical fields. Through the cross-lingual and cross-domain knowledge transfer mechanisms of multilingual LLMs, the practitioners will be able to create training data sets on the languages or domains where very little resources are available. This approach can help democratize access to machine learning for underrepresented languages and specialized domains.It usually concerns the gradual formulation of questions that determine the required input-output correlation, domain context, and the nature of the examples. Recent improvements in instruction tuning and prompt engineering have enabled LLMs to produce high-quality synthetic data even without the example of the target task, which makes zero-shot generation of synthetic data more and more realistic to apply in practice.

Few-shot learning scenarios benefit from synthetic data augmentation. When only a limited number of real examples are available, LLMs can generate additional examples that follow similar patterns but introduce controlled variations. This expansion of the training set helps prevent overfitting to the small real dataset and improves model generalization. Research indicates that combining a small set of high-quality real examples with a larger corpus of synthetic data often yields superior performance compared to using either data source alone [6]. Studies have explored this hybrid approach extensively, examining how proper prompt formatting and example selection significantly influence few-shot learning performance [6]. Gao et al. [6] introduced Pattern-Exploiting Training (PET) that reformulates tasks into cloze-style templates, demonstrating how few-shot examples could be effectively leveraged both for direct inference and for generating additional training examples that maintain consistent patterns.

The mechanisms underlying these improvements relate to how synthetic data complements real examples. While real examples provide ground truth that reflects true data distributions, synthetic examples contribute controllable diversity that improves the representativeness of training data. In few-shot settings, this complementarity is particularly valuable, as it helps prevent models from overfitting to the limited real examples and encourages learning more generalizable features. Also, artificial or synthetically generated data can at times be created based on the strategy to remedy weaknesses or gaps in existing real data. As an example, whenever real examples share a distribution centred on common cases, LLMs may be encouraged to produce examples of edge cases or of very rare phenomena, making a more representative training distribution. This method of targeted augmentation can enable effective utilisation of the abilities of generating synthetic data, with the added benefit of devoting the computation power to the most useful spheres. Figure 2 illustrates how synthetic data generation supports both zero-shot and few-shot learning by supplementing limited real datasets with diverse, task-aligned examples that improve generalization.

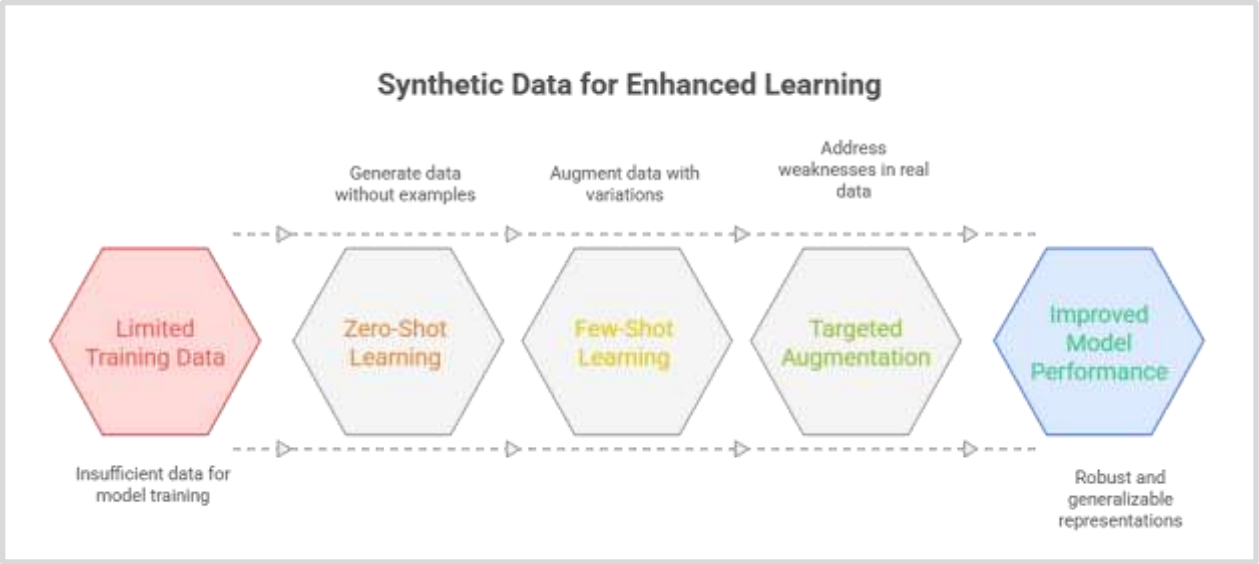

Figure 2: Synthetic Data for Enhanced Learning [5, 6]

4. Data Augmentation and Error Case Generation

LLMs are highly effective at generating variations of existing examples, making them valuable for systematic data augmentation. By instructing an LLM to preserve meaning while varying sentence structure, vocabulary, or style, practitioners can generate multiple versions of each original example. The technique is successful in scaling the size of the available training data without any more human annotation. Studies prove that data augmentation methods like Easy Data Augmentation (EDA) can greatly enhance model performance even with low resources by adding graduated variation so as to maintain semantic context but diversify existing data [7]. This paper laid principles that have been generalised to work on LLM-based methods and demonstrate that even small forms of augmentation done systematically can provide significant performance gains.

Efficient LLM-based data expansion relies on several supporting mechanisms, including lexical variation, distributional balancing, and contextual coherence. First, LLMs have a vast background of linguistic variability and generate semantically allied instances that add lexical and syntactic varieties to retain significant meaning. Such heterogeneity assists downstream models with building more secure feature representations that are not as sensitive to superficial variations. Second, LLMs have the ability to sample according to distribution properties, so that practitioners can counteract training data imbalances by generating examples of underrepresented classes or phenomena. Third, the contextual comprehension of LLMs facilitates generation of examples that preserve coherence and plausibility with meaningful variations, and thus will not run the risk of artifacts that may occur due to naive methods of rule-based augmentation.

Particularly valuable is the ability to generate challenging edge cases and potential error scenarios. By prompting LLMs to create examples that might confuse a model or represent rare but important situations, developers can build more robust systems. For instance, an LLM might be asked to generate examples of ambiguous queries that could be interpreted in multiple ways, helping to identify and address potential weaknesses in a question-answering system. Research has explored this application extensively in natural language processing, with frameworks such as TextAttack providing systematic methods for generating adversarial examples to test model robustness [8].

These approaches demonstrate how carefully crafted examples can expose weaknesses in models and have since been extended to LLMs for natural language tasks.

Another potential strong use is counterfactual testing, in which LLMs are used to come up with examples that are variations of training data in a controlled manner. The examples can assist in the measurement of model fairness and determination of biases by testing the influence of changes in sensitive attributes on the predictions. This is one of the methods of creating pairs of examples in which the variation is limited to the relevant criterion, like gender, ethnicity, age referencing, etc., whereas other contextual features are kept similar. Model predictions can be used to analyze the models, and based on how those changes affect model behavior and cause deviations or mischaracterizations, and developers can take corrective actions. These augmentation and counterfactual testing workflows are summarized in Figure 3, which depicts how LLMs enhance model robustness through systematic data variation and error case generation.

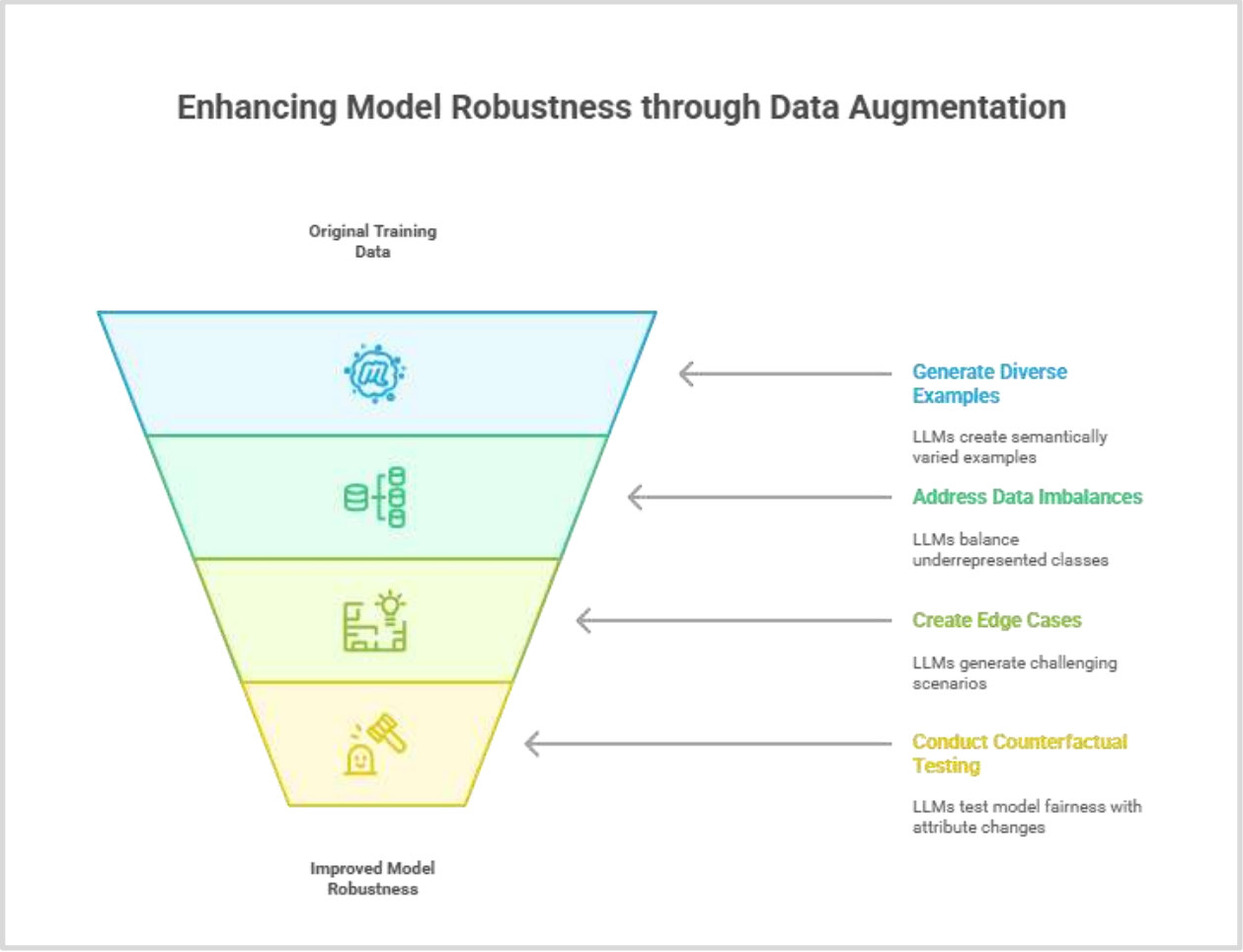

Figure 3: Enhancing Model Robustness through Data Augmentation [7, 8]

5. Challenges and Limitations

Several significant challenges arise when using LLMs to generate synthetic data. The most drastic issue is that of bias amplification. These models accidentally use the biases that exist in their training data and might exponentiate it through the generation process. Studies have also recorded

cases of seemingly neutral prompts that resulted in outputs that reinforce harmful stereotypes [9]. This aspect raises a basic challenge to the aspect of fairness and equity in the machine learning application that utilizes the synthetic data.

A number of methodological procedures have been offered in dealing with such biases. Counterfactual data augmentation methods create examples that neutralize known biases, and in effect, redistribute the encounters that various demographic categories get in the training data. Filtering-based systems involve an automatic way of detecting examples to be filtered based on the ability of such examples to strengthen stereotypical associations. Further, there are fairness constraints they could apply when generating solutions, as protection achieving balance in the representation. Nonetheless, bias management is also a consistent phenomenon that needs new research and creation of more advanced approaches.

Another significant limitation is distributional mismatch between synthetic and real-world data. It is common that synthetic cannot perfectly reproduce the complex distributions and edge cases found in real-world data, especially in specialty areas. Such a mismatch may lead to models that will do fairly well on such artificial validation sets with the risk of radical performance degradation in production settings with realistic data. The problem is especially prominent when discussing rare occurrences or corner cases in which even sophisticated LLMs have not seen enough to produce representative examples.

There are more complexities associated with quality assurance particularly in technical areas. Synthetic examples often contain factual errors and logical inconsistencies that degrade downstream model performance [10]. Strict validation processes, and human curation are also part and parcel of a powerful synthetic data pipeline. This necessity is especially important in such specific disciplines as medicine or law, where insignificant inaccuracies can slip unnoticed without the necessary knowledge of the field. Although researchers have come up with automated metric of evaluation such as perplexity score, consistency checks across multiple generation attempts and fact-verification against knowledge bases, these strategies sometimes fail to detect domain-related problems that can seriously affect application performances. It is now widely accepted that automated screening, supplemented with strategic human review, offers the most effective quality assurance.

Computational resource needs are a big hindering force to mass application. Training multiple synthetic training datasets with state-of-the-art LLMs will demand a very high amount of computational apparatus, the expense of which might go up to thousands of dollars in the instances where big-scale works occur [10]. This economic barrier limits accessibility and risks concentrating the benefits of synthetic data generation within well-resourced institutions. Several strategies have been proposed to mitigate this challenge, including more efficient sampling methods, domain-specific model paradigms that enable smaller models to generate data quickly, and collaborative resource pools that broaden access to computational infrastructure. Despite these challenges, ongoing innovation continues to make synthetic data generation methods more transparent, equitable, and predictable. Figure 4 summarizes the key challenges and limitations of LLM-based synthetic data generation, including bias amplification, distribution mismatch, quality control, and computational constraints.

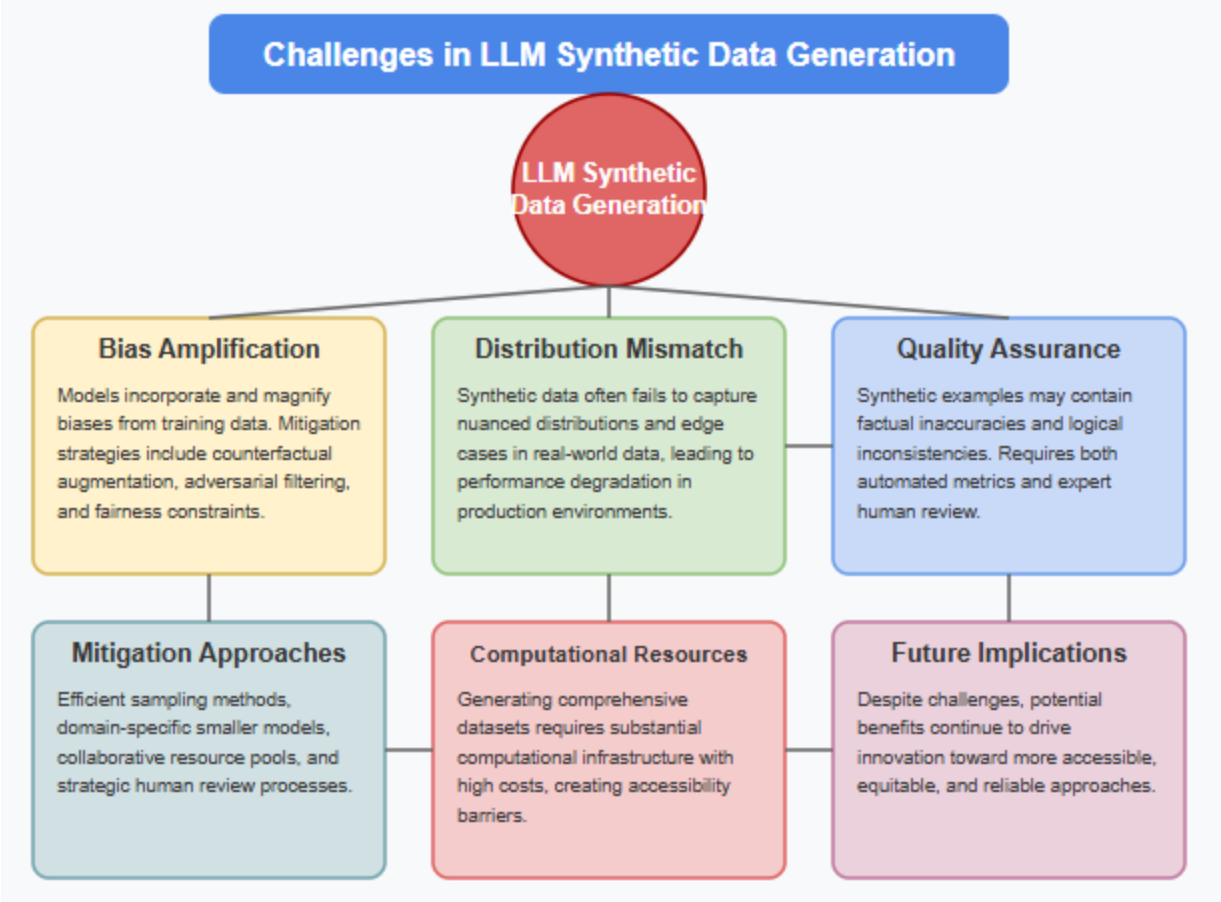

Figure 4: Challenges and Limitations in LLM Synthetic Data Generation [9, 10]

6. Future Directions and Research Opportunities

The field of synthetic data generation using LLMs presents numerous promising research directions. Controlled Generation represents a critical frontier, focusing on developing more precise methods for controlling the distribution and characteristics of generated data, ensuring better alignment with target domains and use cases. Research has established foundational approaches for controlled text generation by incorporating control codes that guide language models to produce outputs with specific attributes [11]. The Conditional Transformer Language Model (CTRL), developed by Keskar et al. [11], demonstrates how large-scale language models can be conditioned on control codes representing domain, style, topics, and other attributes to generate text with desired characteristics. This approach provides a framework that researchers are now extending to synthetic training data generation, where fine-grained control over generated examples is essential for creating balanced, diverse, and representative datasets.

The evolution of controlled generation techniques addresses several key challenges in synthetic data applications. First, distribution matching requires methods that can precisely align synthetic data distributions with target domain characteristics, particularly for specialized fields with unique linguistic patterns. Second, diversity control mechanisms aim to ensure that generated datasets contain appropriate variation while avoiding redundancy or mode collapse. Third, attribute-conditional generation enables practitioners to systematically vary specific dataset characteristics

while maintaining coherence and realism. Recent advances include techniques such as controlled decoding strategies that dynamically adjust sampling parameters, prompt engineering frameworks specifically designed for data generation tasks, and hybrid approaches that combine explicit attribute control with implicit distribution matching. These approaches collectively move toward synthetic data generation systems that offer the precision and reliability needed for mission-critical applications.

Studies explore this complementarity in depth through work on relation extraction in underexplored biomedical domains, demonstrating how synthetic data can effectively supplement limited human-labeled examples to improve performance in specialized domains [12]. Research on biomedical relation extraction provides concrete evidence that strategically combining authentic and synthetic examples can overcome the constraints of limited domain-specific annotations while maintaining high performance.

Adaptive Generation focuses on creating systems that dynamically identify gaps in training data and generate synthetic examples specifically designed to address those weaknesses. This approach moves beyond static data generation toward feedback loops where model performance analysis guides subsequent synthetic data creation. Current research explores techniques including uncertainty sampling to identify areas where models demonstrate low confidence, error analysis to target synthetic data toward frequently misclassified cases, and distribution analysis to identify underrepresented regions in feature space. The potential of this approach lies in its efficiency, focusing computational resources on generating the most valuable examples rather than producing homogeneous data.

Quality Metrics research aims to establish standardized metrics and evaluation frameworks for assessing synthetic data quality across different domains and applications. This direction addresses the critical need for reliable methods to determine whether synthetic data is suitable for specific downstream applications. The predominant methods involve distribution-based measures which involve statistical comparison of synthetic and real data, task-based analysis which optimally evaluates down stream performance of models trained on synthetic data, and human evaluation guidelines which capture qualities of synthetic data on a qualitative basis. Robust, transferable criteria on quality could be crucial in helping practitioners gain confidence that they can use synthetic data in their machine learning pipelines.

Domain Adaptation investigates techniques to better adapt general-purpose LLMs to generate highly specialized data for niche domains without requiring extensive domain-specific training. This direction recognizes that many valuable applications exist in specialized domains where domain-specific data is scarce, presenting both challenges and opportunities for synthetic data generation. Current methods include few-shot adaptation using small domain-specific datasets, knowledge distillation to incorporate domain expertise into generation, and hybrid models that combine general language capabilities with specialized domain modules. Continued advances in these areas are likely to strengthen the contribution of LLMs to overcoming data scarcity, and in time, could fundamentally reshape how machine learning systems are designed to meet the needs of specialized and underrepresented domains.

Conclusion

Innovative approaches, such as leveraging Large Language Models for synthetic data generation, provide promising solutions to the persistent problem of data scarcity in machine learning. This

approach is especially valuable for less widely studied languages and domains that have historically received little attention. Despite ongoing challenges in bias mitigation, distribution alignment, quality control, and computational demand, research in this area continues to advance methodologically. It is reasonable to expect that synthetic data generation will gradually become a standard part of machine learning development processes, as techniques are further refined (e.g. in their ability to control generation, hybrid data systems, and domain adaptation). These receive potential not only in the enhancement of technical performance but also democratization of access to artificial intelligence technologies in various communities and areas of application. As methods for synthetic data generation continue to mature, they are poised to become a standard component of machine learning pipelines, driving greater efficiency, fairness, and accessibility in AI development.

References

[1] Schick, Timo, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. *Toolformer: Language Models Can Teach Themselves to Use Tools.*arXiv:2302.04761. https://doi.org/10.48550/arXiv.2302.04761

[2] Ding, Bosheng, Canwen Xu, Yujia Qin, Zhiyang Teng, Yusheng Su, Xipeng Qiu, and Ting Liu. 2023. *Is GPT-3 a Good Data Annotator?* In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Volume 1: Long Papers*, 11173–11195. https://doi.org/10.18653/v1/2023.acl-long.626

[3] Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.* arXiv:2201.11903. https://doi.org/10.48550/arXiv.2201.11903

[4] Wang, Yizhong, Yifan Jiang, Tianyi Zhang, Ruiqi Zhong, and Noah A. Smith. 2023. *How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources.* arXiv:2306.04751. https://doi.org/10.48550/arXiv.2306.04751

[5] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. *Language Models Are Few-Shot Learners. Advances in Neural Information Processing Systems (NeurIPS 2020)*, arXiv:2005.14165. https://doi.org/10.48550/arXiv.2005.14165

[6] Gao, Tianyu, Adam Fisch, and Danqi Chen. 2021. *Making Pre-trained Language Models Better Few-Shot Learners.*In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021)*, 3816–3830. https://doi.org/10.18653/v1/2021.acl-long.295 (Preprint: https://arxiv.org/abs/2012.15723)

[7] Wei, Jason, and Kai Zou. 2019. *EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks.* In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6382–6388. https://doi.org/10.18653/v1/D19-1670

[8] Morris, John X., Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. *TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP.* In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2020)*, 119–126. https://doi.org/10.18653/v1/2020.emnlp-demos.16

[9] Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623. https://doi.org/10.1145/3442188.3445922

[10] Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 2022. *On the Opportunities and Risks of Foundation Models.* arXiv:2108.07258. https://doi.org/10.48550/arXiv.2108.07258

[11] Keskar, Nitish Shirish, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. *CTRL: A Conditional Transformer Language Model for Controllable Generation.* arXiv:1909.05858. https://doi.org/10.48550/arXiv.1909.05858

[12] Delmas, Morgane, Martyna Wysocka, and André Freitas. 2024. *Relation Extraction in Underexplored Biomedical Domains: A Diversity-Optimized Sampling and Synthetic Data Generation Approach. Computational Linguistics* 50 (3): 953–1000. https://doi.org/10.1162/coli_a_00520