World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

Generating high-quality and diverse synthetic datasets with large language models: A survey

Abinandaraj Rajendran *

Raleigh, USA.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 1145-1149

Publication history: Received on 26 March 2025; revised on 03 May 2025; accepted on 06 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.2.0652

Abstract

Large Language Models (LLMs) are increasingly leveraged to generate synthetic datasets that overcome challenges in real-world data collection, including privacy risks, imbalance, and scarcity. This paper surveys recent developments in LLM-based synthetic data generation, emphasizing techniques that improve diversity, task alignment, and reliability—crucial factors in high-stakes domains such as predictive maintenance. We categorize state-of-the-art approaches into four methodological pillars: prompt engineering, multi-step generation pipelines, quality control through data curation, and rigorous evaluation methods. Structured generation workflows and controlled prompting strategies significantly enhance output coherence and domain relevance, while self-correction mechanisms and diversity-aware metrics contribute to higher dataset fidelity. Despite progress, open challenges persist, including bias propagation, limited generalization across tasks and modalities, and the need for robust ethical safeguards. We outline promising future directions—such as integrating external knowledge, expanding to multilingual and multimodal settings, and fostering human-AI collaboration—for advancing synthetic data generation using LLMs.

Keywords: Synthetic Data Generation; Large Language Models; Predictive Maintenance; Anomaly Detection; Disk Failure Prediction; Cloud Storage Systems

1. Introduction

The generation of diverse synthetic datasets using Large Language Models (LLMs) has emerged as a critical area of research in artificial intelligence. This approach addresses several limitations of real-world data collection, including cost, scarcity, privacy concerns, and inherent biases.

However, synthetic data generated by LLMs also presents challenges such as bias amplification, hallucinations, limited task alignment, and insufficient diversity—factors that can undermine model reliability and generalization. These limitations are important to address, particularly in data-sensitive fields like healthcare, finance, and tasks such as predictive maintenance, where high-quality, contextually accurate data is essential. Applications within predictive maintenance, such as anomaly detection and disk failure prediction in cloud storage systems, rely on precise data signals and can greatly benefit from enhanced synthetic data generation techniques.

This survey reviews recent advancements in LLM-based synthetic data generation, with an emphasis on improving diversity, quality, and alignment of generated datasets. It organizes the landscape into four key methodological categories: prompt engineering for output control, multi-step generation for logical coherence, data curation for quality assurance, and evaluation strategies for utility and diversity assessment.

* Corresponding author: Abinandaraj Rajendran

By highlighting emerging techniques and quality control mechanisms, this paper aims to guide researchers and practitioners toward more effective, ethical, and domain-agnostic synthetic data generation practices.

2. Key Methodologies and Techniques

The generation of high-quality synthetic datasets using LLMs has evolved substantially, with research emphasizing four primary areas: prompt engineering, multi-step generation, data curation and quality control, and evaluation strategies.

2.1. Prompt engineering

Prompt engineering plays a foundational role in guiding LLM outputs. Techniques such as attribute-controlled prompts (e.g., AttrPrompt) enable the specification of properties like style and length, which enhance diversity and facilitate bias analysis and control in generated datasets [1]. Generator prompts introduce multi-step prompting mechanisms through option lists and random selections, further promoting output variation [2]. Additionally, meta-prompting provides a task-agnostic scaffolding approach where a meta-model coordinates expert models to manage complex generation tasks effectively [3].

2.2. Multi-step generation

Multi-step generation involves decomposing complex tasks into simpler subtasks, offering notable benefits. Chain-of-thought prompting structures reasoning steps to improve logical coherence within generated outputs [4]. More structured frameworks, such as TarGEN, implement a seedless, context-aware four-step generation pipeline with integrated self-correction mechanisms [5]. Similarly, the UniGen framework offers an end-to-end LLM system with built-in diversity and accuracy control, enabling high-fidelity and varied synthetic datasets [6].

2.3. Data curation and quality control

Data curation and quality control have become critical areas of focus. Selective annotation & uncertainty-based filtering methods use heuristics or re-weighting techniques to prioritize superior synthetic samples [4]. Selective annotation with human-in-the-loop feedback strategies involves correcting labels through human feedback or auxiliary models, while self-correction techniques leverage LLMs to autonomously identify and amend mislabeled data [4, 5]. Beyond generating new samples, some methods retrieve and transform existing datasets to tailor outputs for specific target tasks, a strategy that has proven effective in creating domain-aligned synthetic data [7].

2.4. Evaluation methods

Evaluation methods are crucial for assessing synthetic data utility. Direct evaluation relies on human annotators or auxiliary models to measure quality, while indirect evaluation involves benchmarking model performance on downstream tasks using the generated data [4]. Diversity analysis, employing measures such as lexical diversity and cosine similarity, provides insights into the variation within datasets [1, 5]. New tools like the LLM cluster-agent metric have been developed to specifically quantify the impact of data diversity on LLM behavior [8].

2.5. Summary

To aid quick reference and comparison, Table 1 below consolidates the key methodologies and techniques discussed in Sections 2.1 through 2.4. It categorizes each method by its functional domain—prompt engineering, multi-step generation, data curation, and evaluation—while providing a concise description and citation. This summary serves as a compact resource for researchers and practitioners to identify applicable strategies for LLM-driven synthetic data generation.

Table 1 Summary of Key Methodologies & Techniques

Category	Methodology	Description	References
Prompt Engineering	Attribute-controlled prompts (AttrPrompt)	Specifies multiple attributes (e.g., style, length, sentiment) to guide and diversify outputs	[1]
	Generator prompts	Generates options lists and applies random sampling to inject controlled randomness	[2]
	Meta-prompting	Uses a meta-controller LLM to orchestrate expert models specialized for subtasks	[3]
Multi-Step Generation	Chain-of-Thought prompting	Prompts models to generate intermediate reasoning steps before final output	[4]
	TarGEN	Four-phase generation: context initialization, instance seed generation, label-constrained instance generation, self-correction	[5]
	UniGen framework	Combines conditional sampling, diversity boosting, and validation checkpoints	[6]
Data Curation & Quality	Selective annotation & uncertainty-based filtering	Selects or reweights samples based on fluency, relevance, or heuristic quality scores	[4]
	Selective annotation with human-in-the-loop feedback	Corrects noisy labels via external human annotation or small auxiliary models	[4]
	Self-correction	Allows models to flag and amend errors during or post generation using self-evaluation modules	[5]
	Retrieval + transformation	Retrieves relevant examples from existing datasets and modifies them for new tasks	[7]
Evaluation Methods	Direct evaluation	Human annotators or pretrained models judge quality (fluency, consistency, task fit)	[4]
	Indirect evaluation	Assesses downstream task performance improvements using synthetic training data	[4]
	Diversity analysis	Uses token distribution metrics and semantic similarity (cosine distance)	[1, 5]
	LLM cluster-agent metric	Groups model behaviors and measures diversity's effect on generalization and performance	[8]

3. Key Findings and Insights

Recent research demonstrates that advanced prompting techniques like AttrPrompt and generator prompts substantially improve synthetic data quality, outperforming simple prompting baselines [1, 2]. These approaches allow for finer control over outputs, enhancing both diversity and alignment with target distributions.

Multi-step generation frameworks such as TarGEN and UniGen further strengthen the logical consistency and quality of generated data through structured workflows [5, 6].

Efficiency is another key advantage: methods like AttrPrompt show that high diversity can be achieved with lower computational overhead, enabling more scalable synthetic data pipelines [1].

Research also highlights the robustness of these methods across LLM architectures and tasks, indicating strong potential for broad application [1, 5].

Moreover, synthetic datasets generated by these techniques effectively augment low-resource scenarios, addressing data scarcity and enabling better model performance on long-tail distributions [9].

Finally, robust quality control mechanisms such as self-correction and direct evaluation protocols have proven essential to maintaining the integrity and usability of synthetic datasets [4, 5].

Task-specific alignment techniques like CodecLM and SyntheT2C show that contextualizing synthetic data to match application needs (e.g., predictive maintenance) significantly improves downstream task performance [10, 11].

4. Challenges and Future Directions

Despite significant progress, several challenges remain in generating synthetic data using LLMs. One persistent issue is mitigating bias and hallucinations. Because LLMs inherit statistical biases and inaccuracies from their pretraining data, generated synthetic datasets can reflect and even amplify these problems, posing risks to the fairness and reliability of downstream models [9].

Another critical limitation is the lack of generalization. Synthetic datasets often perform well within narrowly defined contexts but struggle in broader, unseen scenarios. This challenge becomes particularly acute during multi-turn generation tasks, where maintaining coherence over extended interactions remains difficult [12]. Improving the generalization capabilities of synthetic data is an essential area for future research.

The integration of external knowledge sources into the generation pipeline represents a promising development. Methods that can incorporate structured knowledge (such as knowledge graphs) or unstructured corpora into LLM generation workflows could enhance the factual grounding and domain specificity of synthetic datasets, leading to higher-quality outputs [4].

Ethical considerations are also paramount. As synthetic data becomes more capable and realistic, the risks of privacy violations, misinformation, and malicious use increase. Developing robust ethical frameworks and governance mechanisms for synthetic data generation and deployment is a necessary complement to technical innovation. For instance, [9] discuss privacy and high-level ethical guidelines, but do not address misinformation or detailed governance mechanisms.

Further, the need to expand synthetic data generation across modalities and languages remains a key open direction. Most current approaches focus heavily on English text generation, limiting applicability to global and multimodal contexts. Advances that extend synthetic generation to cross-modal (e.g., text-image, text-audio) and multilingual settings will be essential to broaden the impact of these technologies [5].

The field is also increasingly recognizing the importance of human-AI collaboration. Human-in-the-loop frameworks, where humans guide or validate the generation process, can significantly enhance the quality, relevance, and trustworthiness of synthetic datasets [4]. Such frameworks are particularly valuable in specialized domains requiring domain expertise.

Finally, enhancing instruction-following capabilities through synthetic data remains a promising frontier. By leveraging improved synthetic datasets that better mimic real-world instruction patterns, researchers can fine-tune LLMs to achieve superior understanding and execution of user instructions, which is vital for interactive and application-driven AI systems [12].

5. Conclusion

The field of LLM-driven synthetic data generation is evolving rapidly, with significant advancements in prompting techniques, generation workflows, and quality assurance. These innovations have made synthetic data a viable supplement—or even alternative—to real-world datasets in critical domains such as healthcare and finance, as well as in cross-domain applications like predictive maintenance. Nevertheless, addressing remaining challenges around bias, generalization, and ethics will be essential for realizing the full potential of synthetic datasets in real-world deployments.

References

[1] Yu, Y., Zhuang, Y., Zhang, J., Meng, Y., Ratner, A., Krishna, R., Shen, J., & Zhang, C. (2023). Large language model as attributed training data generator: A tale of diversity and bias. arXiv preprint arXiv:2306.15895.

[2] Chen, J., Qadri, R., Wen, Y., Jain, N., Kirchenbauer, J., Zhou, T., & Goldstein, T. (2024). GenQA: Generating millions of instructions from a handful of prompts. arXiv preprint arXiv:2406.10323.

[3] Suzgun, M., & Kalai, A. T. (2024). Meta-prompting: Enhancing language models with task-agnostic scaffolding. arXiv preprint arXiv:2401.12954.

[4] Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., & Wang, H. (2024). On LLMs-driven synthetic data generation, curation, and evaluation: A survey. arXiv preprint arXiv:2406.15126.

[5] Gupta, H., Scaria, K., Anantheswaran, U., Verma, S., Parmar, M., Sawant, S. A., Baral, C., & Mishra, S. (2023). TarGEN: Targeted data generation with large language models. arXiv preprint arXiv:2310.17876.

[6] Wu, S., Huang, Y., Gao, C., Chen, D., Zhang, Q., Wan, Y., Zhou, T., Zhang, X., Gao, J., Xiao, C., & Sun, L. (2024). UniGen: A unified framework for textual dataset generation using large language models. arXiv preprint arXiv:2406.18966.

[7] Gandhi, S., Kulkarni, V., & Shenoy, A. (2024). Better synthetic data by retrieving and transforming existing datasets. arXiv preprint arXiv:2404.14361.

[8] Chen, H., Waheed, A., Li, X., Wang, Y., Wang, J., Raj, B., & Abdin, M. I. (2024). On the diversity of synthetic data and its impact on training large language models. arXiv preprint arXiv:2410.15226.

[9] Guo, X., & Chen, Y. (2024). Generative AI for synthetic data generation: Methods, challenges and the future. arXiv preprint arXiv:2403.04190.

[10] Wang, Z., Li, C.-L., Perot, V., Le, L. T., Miao, J., Zhang, Z., Lee, C.-Y., & Pfister, T. (2024). CodecLM: Aligning language models with tailored synthetic data. arXiv preprint arXiv:2404.05875.

[11] Zhong, Z., Zhong, L., Sun, Z., Jin, Q., Qin, Z., & Zhang, X. (2024). SyntheT2C: Generating synthetic data for fine-tuning large language models on the Text2Cypher task. arXiv preprint arXiv:2406.10710.

[12] Zhao, H., Andriushchenko, M., Croce, F., & Flammarion, N. (2024). Is in-context learning sufficient for instruction following in LLMs? arXiv preprint arXiv:2405.19874.