Evaluating the Fairness, Quality, and Performance of Synthetic Data Generation Using Large Language Models

Software Engineering For AI Report

Benedetto Scala
Id: 0522501794
b.scala1@studenti.unisa.it

Leopoldo Todisco
Id: 0522501795
l.todisco4@studenti.unisa.it

Carlo Venditto
Id: 0522501796
c.venditto@studenti.unisa.it

ACM Reference Format:
Benedetto Scala, Leopoldo Todisco, and Carlo Venditto. 2024. Evaluating the Fairness, Quality, and Performance of Synthetic Data Generation Using Large Language Models: Software Engineering For AI Report. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ABSTRACT

In this study, we evaluate the capabilities of Large Language Models (LLMs) in generating high-quality synthetic datasets, focusing on three main aspects: data quality, model performance, and fairness. Given the challenge of data scarcity in various domains, synthetic data generation offers a promising solution, with LLMs showing potential for creating realistic and diverse datasets. Using the German Credit dataset as a basis, we applied different prompting techniques (0-shot, 1-shot, and 2-shot) to generate synthetic data and assessed their structural quality through metrics such as completeness, readability, uniqueness, and consistency. Our findings indicate that while LLMs can produce high-quality data structurally, they often generate a significant number of duplicate rows, necessitating human oversight.

We also evaluated the performance of machine learning models trained on these synthetic datasets, achieving competitive results in terms of F1 score and accuracy. The 1-shot prompt technique yielded the highest performance metrics, though its generalizability to real-world data was limited. Additionally, we examined the fairness of these datasets across protected attributes such as age and sex. Our results showed mixed outcomes, with the 0-shot and 2-shot techniques maintaining better fairness metrics compared to the 1-shot technique.

Overall, our research demonstrates the potential of LLMs in synthetic data generation but also highlights the need for further refinement to address issues like data duplication and bias. Future work should explore other LLMs and datasets to enhance the generalizability and ethical implications of synthetic data generation.

1 INTRODUCTION

Machine learning (ML) has transformed numerous industries, but its wider adoption is hindered by a pervasive roadblock: insufficient data. Specifically, the use of ML algorithms presumes the availability and access to large datasets for training, be it labeled or unlabeled. Unfortunately, many real-world domains are often data scarce, such as healthcare and finance [11].

Synthetic data generation has emerged as a promising solution to these challenges, offering a cost-effective way to enrich datasets. Traditional methods, such as SMOTE [5], focus on creating new data points, but they usually fail in understanding the correlation between features [1].

More recently, Large language models (LLMs) have also shown great potential in the realm of synthetic data generation, generating high quality realistic data with various feature types [2].

In the current work we provide an extensive analysis of how good are LLMs at generating synthetic data, analyzing the quality of datasets in terms of reliability, fairness and privacy.

The link to the Github repository is the following: https://github.com/leotodisco/QUALITY

2 PROJECT GOAL

The goal of the study is to address wether LLMs are good tabular dataset generators, in terms of data quality, fairness and performances.

To formalize the main goal of our study, we applied the Goal-Question-Metric approach proposed by Caldiera et al. [4].

The *goal* of our research was to *evaluate* the quality, performance during training, and fairness of a synthetically generated dataset with large language models (LLMs). The *perspective* was that of both researchers and practitioners.

2.1 Research Question

Starting from the goal of this study, which is to analyze the extent to which synthetically generated datasets can be employed in real-world scenarios, we defined four research questions.

The generated datasets have been collected starting from a real-world dataset —GERMAN CREDIT [6]— through the use of CHAT GPT 4O. For this reason, before proceding further with the analysis of metrics computed on models trained with these datasets, we needed to be sure that the datasets could be applied in real-world scenarios. Hence, we had to assess the quality of these datasets. To perform such task, we analyzed *Data Quality Metrics* defined by Elouataoui et al [7]. These aspects were evaluated in the context of our first and preliminary research question:

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

> 🔍 **RQ$_0$.** *To what extent are datasets generated by language models of high quality in terms of absence of duplicate values, missing values, and similar issues?*

After assessing the quality of the datasets, we proceeded with the analysis of performance metrics by calculating key indicators such as Precision, Accuracy, Recall, and F1 score.

> 🔍 **RQ$_1$.** *Can the datasets generated using Large Language Models (LLM) achieve good performance in terms of F1 score and Accuracy?*

Finally, in light of the increasing emphasis on fairness in the creation and training of machine learning models, we are committed to rigorously evaluating and ensuring fairness when generating synthetic data with large language models. This commitment includes implementing robust fairness metrics, thereby ensuring ethical and equitable outcomes in our data generation efforts.

> 🔍 **RQ$_2$.** *To what extent are datasets generated by language models fair across dimensions of equity, such as representation, accuracy, and impartiality towards demographic groups?*

3 METHODOLOGY

First, we selected three different prompts for the LLM, i.e., the input or instruction given to the model to generate a response. In this regard, we decided to apply the *few-shot* prompting technique [3], i.e., providing demonstrations in the prompt to steer the model to better performances. We selected the prompts by following the guidelines by Xi et al. [8], which demonstrated a 30% increase in performance when using a 1-shot prompt compared to a 0-shot prompt. Moreover, they demonstrated that a 2-shot prompt provides benefits but results get worse beyond that point. Hence, we decided to execute our study using a 0, 1 and 2-shot prompt. For each prompt technique, we will generate three datasets, acknowledging the inherent non-deterministic nature of large language models (LLMs) [9]. The decision to limit this to three datasets is driven by token and time constraints.

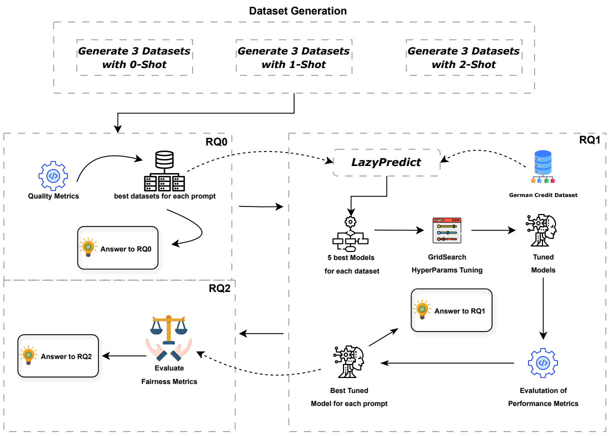
Dataset Generation
Generate 3 Datasets with 0-Shot
Generate 3 Datasets with 1-Shot
Generate 3 Datasets with 2-Shot
RQ0
Quality Metrics
best datasets for each prompt
Answer to RQ0
RQ1
LazyPredict
German Credit Dataset
5 best Models for each dataset
GridSearch HyperParams Tuning
Tuned Models
Answer to RQ1
Best Tuned Model for each prompt
Evalutation of Performance Metrics
RQ2
Answer to RQ2
Evaluate Fairness Metrics

Figure 1: An overview of the proposed approach

Metodological steps to assess RQ0 To evaluate the qualitative aspects of the generated datasets, we will employ the following metrics:

- **Completeness**, that is a measure of how many values are incomplete and is defined as:
$$\frac{number\_of\_non\_empty\_values}{total\_values} \times 100$$
- **Uniqueness**, that is the measure of how many redundant values are there and is defined as:
$$\frac{number\_of\_unique\_rows}{total\_values} \times 100$$
- **Consistency**, that refers to information that adheres to uniform structures, types, and aligns with established data schemas and standards [7] and is defined as:
$$\frac{number\_of\_value\_with\_consistent\_types}{total\_values} \times 100$$
- **Readability**, that refers to the number of mispelled values, or nonsense words and is defined as:
$$\frac{number\_of\_non\_mispelled\_values}{total\_values} \times 100$$

After measuring these metrics, we will select the best dataset out of the three generated for each prompt. The selection will be based on the mean value of the four metrics ensuring that we utilize the highest quality dataset generated by the language models in the subsequent research questions.

Additionally, we will employ data visualization techniques to check the distribution of the data and the relationships between the generated dataset and the original German credit dataset. By plotting the data, we aim to visually assess how closely the generated data matches the original dataset in terms of distribution and key relationships.

Furthermore, we anticipated that ChatGPT, the model used to generate the dataset, might forget the context while generating, potentially resulting in duplicated rows. To address this, we conducted the same analysis of uniqueness, completeness, consistency, and readability on the firts 500 rows of the generated datasets to check if the metrics changed significantly due to the potentially presence of duplicate rows.

Metodological steps to assess RQ1 To address RQ1, starting with the dataset identified as the best for each prompting technique in RQ0, we aim to compare the metrics such as accuracy, and F1 score that the models achieve on the synthetic dataset with those they achieve on the original dataset.

First, we will use LazyPredict[1], a library that allows us to train and evaluate a variety of machine learning models, to identify the top three performers. For each of the three prompting techniques, we will select the best five models using k-fold cross-validation. We will then perform hyperparameter tuning on these top models using grid search to identify the optimal parameters, which helps reduce the risk of overfitting and improves model performance. Finally, we will compare the performance of these tuned models on synthetic datasets against their performance (accuracy, f1-score)

[1] https://pypi.org/project/lazypredict/

on the original datasets (evaluating on the entire original dataset) and the synthetic datasets (using an 80-20 split).

It is important to note that we will not employ any preprocessing techniques other than one-hot encoding. This decision is based on our objective to assess the performance of the raw dataset. By avoiding extensive preprocessing, we aim to evaluate the models' ability to handle unprocessed, real-world data, which provides insights into their robustness and generalizability.

Metodological steps to assess RQ2 To answer RQ2, we will utilize the top models trained and optimized through the grid search conducted in RQ1. We will then use the predictions obtained from each of the four best models (1 shot, 2 shot, 3 shot, and the original German Credit dataset) to calculate various fairness metrics.

To do this we first identified the protected attributes (e.g., age, sex), privileged (e.g., older individuals, male) and unprivileged groups (e.g., younger individuals, female).

Specifically, we will calculate the following metrics, using the `IBM AIF-360` library:

- **Average Odds Difference (AOD)**: Average of difference in False Positive Rate and True Positive Rate for unprivileged and privileged groups.
- **Statistical Partity Difference (SPD)**: This metric quantifies the disparity in predicted positive outcomes between a privileged group and an unprivileged group within a predictive model.
- **Equal Opportunity Difference (EOD)**: Measures the deviation from the equality of opportunity, which means that the same proportion of each population receives the favorable outcome.

4 RESULTS

RQ0 - Assessing the quality of LLM generated datasets

At this stage of our study, we had three datasets for each prompt engineering technique. We used the scripts coming from Recupito et al. [10] to compute the metrics described in section 3, taking the mean value of each metric for each dataset. The values for the metrics of the best datasets are shown in table 1.

Please note that the Readability metric appears low due to the nature of the original dataset, where all categorical values are formatted as *A34, A35, ...*

	Completeness	Readability	Uniqueness	Consistency
Original Set	100	38.09	5.12	100
0-Shot	100	38.09	3.62	100
1-Shot	100	38.09	1.09	100
2-Shot	100	38.09	3.3	100

Table 1: Quality metrics of the best datasets

At this point, we can conclude that ChatGPT is an effective data generator when considering only structural quality metrics. However, we also examined the number of duplicate rows in each generated dataset and found that, on average, one-third of the dataset consists of duplicate rows.

To investigate further, we decided to assess whether the uniqueness of the data increased when considering only the first 500 rows. Our hypothesis was that the number of duplicates would rise after the initial 500 rows. The results are shown in table 2.

	Completeness	Readability	Uniqueness	Consistency
0-Shot	100	38.09	2.47	100
1-Shot	100	38.09	1.57	100
2-Shot	100	38.09	5.48	100

Table 2: Quality metrics when considering only 500 rows

However, our results showed no significant improvement in uniqueness when considering only the first 500 rows, indicating that limiting the dataset to the first 500 rows did not yield better results.

From the perspective of duplicate rows, focusing on only the first 500 rows significantly reduced the number of duplicates compared to the full 1000 rows. In the 0-shot dataset, there were only 123 duplicates; the 1-shot dataset had no duplicates, and the 2-shot dataset contained just 22 duplicates.

Finding 1. Using LLMs to generate data is a valuable technique for data augmentation. However, at least at the time we write this work, due to the high number of duplicate rows, it is less effective for generating entire datasets.

From the visual data analysis, which is available in the notebooks in our GitHub repository[2], it comes the following finding:

Finding 2. 0-Shot prompt engineering technique appears to be the best technique because the LLM tries to generate all possible values for every categorical feature, and all the values appear the same number of times.

Please note that while generating the initial 2-shot dataset, ChatGPT exhibited the so called hallucinations and produced less realistic data, such as people being 300 years old.

Answer to RQ$_0$. From a structural perspective, an LLM can generate good quality data. However, this process requires human supervision due to the potential for hallucinations and a high likelihood of duplicate rows.

RQ1 - Evaluating the model performances on LLM generated datasets.

The results obtained by the best models with the best hyperparameter are the following:

[2] https://github.com/leotodisco/QUALITY

	RandomForest (0-Shot Prompt)	RandomForest (1-Shot Prompt)	AdaBoost (2-Shot Prompt)
Performance on synthetic set without duplicates			
F1 Score	0.91	0.968	0.955
Accuracy	0.903	0.956	0.95
Performance on original set without duplicates			
F1 Score	0.764	0.508	0.696
Accuracy	0.647	0.660	0.59
Performance on synthetic set with duplicates			
F1 Score	0.948	0.956	0.95
Accuracy	0.945	0.940	0.955
Performance on original set with duplicates			
F1 Score	0.67	0.623	0.654
Accuracy	0.58	0.478	0.557

Table 3: Performance metrics of models trained on synthetic data, evaluated on different test sets and prompts, with and without duplicates.

All performance metrics reported in this table are derived from models trained on synthetic datasets. When evaluating performance on synthetic sets, the models were tested using 20% of the synthetic data following an 80-20 split paradigm. In contrast, when referring to the original set, the models were tested on the entire German Credit dataset.

Such results show that machine learning models trained on synthetic datasets generated using LLMs perform significantly well, even better than models trained on the original dataset, of which the best achieved 0.869 in F1 score and 0.8 in accuracy.

- The **1-Shot Prompt Technique** yielded the highest performance metrics on the test set without duplicates, with an F1 score of 0.968 and an accuracy of 0.956. However, its performance on the original set was lower, indicating that while the synthetic data is of high quality, it may not generalize as well when combined with the original dataset.
- The **0-Shot Prompt Technique** also showed strong performance, particularly on the test set without duplicates, suggesting it is effective in generating high-quality data for training models. On the original set, it had a higher F1 score compared to the 1-Shot technique.
- The **2-Shot Prompt Technique** performed lower than the 1-Shot prompt in the synthetic test set but still provided competitive results, indicating that LLMs can generate useful training data across different prompting techniques.

> 🔍 **Finding 3.** The choice of prompting technique plays a crucial role, for instance, the 1-Shot technique yielded the highest metrics on the synthetic test set out of all three prompt technique. However, it performed lower on the original dataset, indicating that while the synthetic data is of high quality, it may not generalize as well to the original data.

> ☞ **Answer to RQ$_1$.** An LLM can generate data on which models can achieve excellent results in terms of F1 score and Accuracy on the synthetic test set.

RQ2 - Evaluating ethics and Fairness of LLM generated Datasets Building on the machine learning models discussed in RQ1, we turn our focus towards evaluating the ethics and fairness of the datasets generated by large language models (LLMs). In this context, it is crucial to ensure that the datasets are not only effective for their intended tasks but also fair and unbiased.

To achieve this, we used the two protected attributes already identified by IBM.: age and sex, then we computed various fairness metrics as outlined in Section 3 of this paper. The detailed results of our fairness evaluations are as follows:

	AOD	SPD	EOD
Original Set	0.0	-0.019	0.0
0-Shot	0.0	-0.063	0.0
1-Shot	0.0	-0.165	0.0
2-Shot	0.0	-0.002	0.0

Table 4: Fairness Metrics for Sex

	AOD	SPD	EOD
Original Set	0.111	0.078	0.157
0-Shot	0.012	-0.006	0.026
1-Shot	0.13	0.2	0.061
2-Shot	-0.006	0.008	-0.014

Table 5: Fairness Metrics for Age

> ☞ **Answer to RQ$_2$.** For sex, the Original Set shows minimal bias. The 0-Shot and 2-Shot settings maintain similar fairness levels, while the 1-Shot setting increases bias (SPD = -0.165). For age, the Original Set has moderate bias, which is significantly reduced in the 0-Shot and 2-Shot datasets. However, the 1-Shot setting introduces a higher AOD (0.13). Overall, 2-Shot settings provided the most balanced fairness across both attributes.

5 DISCUSSION AND LIMITATIONS

Our research demonstrates that LLMs can be powerful tools for synthetic data generation, offering a promising solution to data scarcity in various domains. However, the current limitations, particularly the issue of duplicate rows and potential for generating unrealistic data (hallucinations), highlight the necessity for human supervision and additional refinement in the data generation process. As LLM technology continues to evolve, addressing these challenges will be vital to fully harness their potential for generating high-quality, fair, and reliable synthetic datasets.

There are also limitations in our study. The models were chosen based on the best performance in F1 and Accuracy; however, some studies in the literature [12] have shown a correlation between fairness and performance. Therefore, selecting different models or hyperparameters could result in different fairness metrics.

Another limitation of our study arises from the analysis being confined to a single dataset. Consequently, the results lack broad generalizability due to the dataset generation with each prompt being replicated only three times. This restricted replication may not sufficiently capture variability across multiple generations.

Furthermore, the scope of our analysis was limited to the behavior of ChatGPT-4o, excluding other large language models (LLMs) that could potentially exhibit different dynamics and provide additional insights.

Additionally, while ChatGPT-4o may already be familiar with the original dataset, our visual analyses indicate a substantial variation in data distribution between the real dataset and those synthetically generated. This suggests that despite potential prior knowledge, the model's response behavior can still significantly differ when interacting with new or altered data constructs.

6 CONCLUSION

In this study, we aimed to evaluate the capability of Large Language Models (LLMs) in generating high-quality synthetic datasets, focusing on data quality, performance, and fairness. By employing various prompting techniques (0-shot, 1-shot, and 2-shot), we generated multiple datasets from the German Credit dataset and assessed their quality using structural metrics. Additionally, we analyzed the performance of machine learning models trained on these datasets and evaluated fairness metrics to ensure ethical data generation practices.

Data Quality: Our analysis revealed that LLMs, particularly ChatGPT, can generate structurally sound datasets, exhibiting high completeness, consistency, and readability. However, a significant issue identified was the presence of duplicate rows, which constituted a considerable portion of the generated data. While focusing on the first 500 rows of the datasets reduced the number of duplicates, it did not significantly improve uniqueness. This highlights the need for human oversight and potential post-processing steps to handle duplicates effectively.

Model Performance: The machine learning models trained on synthetic datasets generated by LLMs showed promising results. The 1-shot prompt technique, in particular, demonstrated the best overall performance. Performance metrics such as F1 score and accuracy were competitive when compared to models trained on the original dataset, indicating that LLM-generated datasets can be viable substitutes for real data in training scenarios. However, the presence of duplicates inflated performance metrics, underscoring the importance of careful data handling and cleaning.

Fairness and Ethics: Evaluating fairness metrics across protected attributes such as age and sex revealed mixed results. While datasets generated by the 0-shot and 2-shot prompt techniques showed relatively better fairness metrics, the 1-shot technique displayed notable disparities, especially in terms of statistical parity difference (SPD). These findings suggest that while LLMs have the potential to generate fair data, the choice of prompting technique and careful monitoring of generated data are crucial to avoid introducing or exacerbating biases.

Future works: In future, we could investigate the behavior of other language models, such as LLAMA, to further understand their fairness characteristics. Additionally, analyzing more well-known datasets in the fairness research community could provide deeper insights. By examining a broader range of datasets, the results would become more generalizable, enhancing our ability to create equitable machine learning datasets.

REFERENCES

[1] Rok Blagus and Lara Lusa. 2013. SMOTE for high-dimensional class-imbalanced data. *BMC bioinformatics* 14 (2013), 1–16.

[2] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2022. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280* (2022).

[3] Tom B. Brown and et al. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[4] Victor R Basili1 Gianluigi Caldiera and H Dieter Rombach. 1994. The goal question metric approach. *Encyclopedia of software engineering* (1994), 528–532.

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[6] Dheeru Dua, Casey Graff, et al. 2017. UCI machine learning repository. (2017).

[7] Widad Elouataoui, Imane El Alaoui, Saida El Mendili, and Youssef Gahi. 2022. An advanced big data quality framework based on weighted metrics. *Big Data and Cognitive Computing* 6, 4 (2022), 153.

[8] Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. 2024. Large Language Models (LLMs) on Tabular Data: Predic-tion, Generation, and Understanding-A Survey. *arXiv preprint arXiv:2402.17944* (2024).

[9] Shuyin Ouyang, Jie M. Zhang, Mark Harman, and Meng Wang. 2023. LLM is Like a Box of Chocolates: the Non-determinism of ChatGPT in Code Generation. arXiv:2308.02828 [cs.SE]

[10] Gilberto Recupito, Raimondo Rapacciuolo, Dario Di Nucci, and Fabio Palomba. 2023. Unmasking Data Secrets: An Empirical Investigation into Data Smells and Their Impact on Data Quality. (2023).

[11] Nabeel Seedat, Nicolas Huynh, Boris van Breugel, and Mihaela van der Schaar. 2023. Curated LLM: Synergy of LLMs and Data Curation for tabular augmentation in ultra low-data regimes. *arXiv preprint arXiv:2312.12112* (2023).

[12] Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H Chi. 2021. Understanding and improving fairness-accuracy trade-offs in multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1748–1757.