Evaluating the Usefulness of Large Language Models for Synthetic Samples Generation via Few-shot Learning

Maynara Donato de Souza
Universidade Federal de Pernambuco
Recife, Brazil
mds3@cin.ufpe.br

Flavio Arthur Oliveira Santos
Universidade Federal de Pernambuco
Recife, Brazil
faos@cin.ufpe.br

Cleber Zanchettin
Universidade Federal de Pernambuco
Recife, Brazil
cz@cin.ufpe.br

Abstract

This paper investigates the potential of Large Language Models (LLMs), such as GPT-4, Cohere, and Gemini, to generate synthetic samples of time-series sensor data using only a few-shot learning (3 samples) without fine-tuning. We aim to highlight their viability in augmenting datasets with minimal data, addressing data scarcity and class imbalance challenges. We evaluate Human Activity Recognition (HAR) tasks from wearable device sensors as a use case in this investigation. Our findings demonstrate that LLMs can produce high-quality synthetic samples in less imbalanced datasets, achieving competitive results compared to traditional generative models. However, the LLM performance decreases with more imbalanced datasets, where the generated synthetic data lacks diversity. We also observed that classification models trained with LLM-generated samples showed more stability in terms of confidence intervals, with the Gemini model consistently producing more stable data. We also present a framework for evaluating synthetic data generation methods, showing the trade-off between synthetic and real-world data and suggesting practical directions for future work addressing data scarcity and balance limitations.

1 Introduction

Developing accurate Machine Learning and Deep Learning models often requires large, well-labeled datasets [1], which are time-consuming and expensive to collect. In response to these challenges, data augmentation techniques have gained attention as a way to artificially expand datasets, thus improving model performance [2]. Some of them have relied on methods such as time-series transformations, signal processing techniques [3], and noise injection. However, recent advancements in natural language processing (NLP) and the advent of Large Language Models (LLMs) have opened new avenues for data generation beyond text, including synthesizing realistic sensor data [4, 5].

LLMs, particularly models like GPT-4 [6], have shown promise in generating high-quality data across a range of domains (see [7, 8]), leveraging contextual understanding and pattern recognition [9]. This work investigates whether LLMs can generate useful synthetic samples for time-series sensor data with only few (three) samples provided as input. This setting poses a special challenge, as it requires the generative models to capture the underlying distributions of complex sensor data

Latinx in AI @ NeurIPS-24 (LXAI 2024).

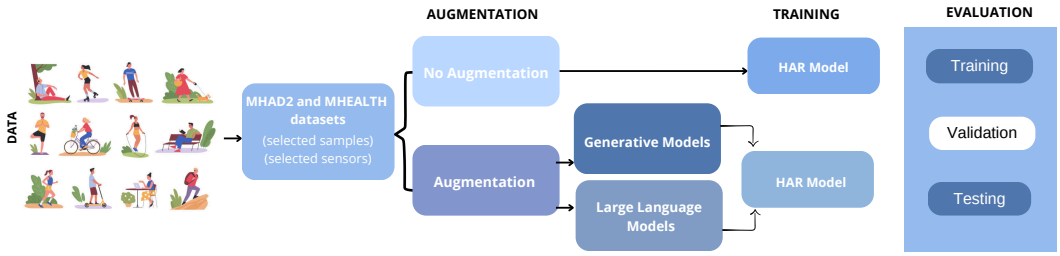

Figure 1: **A summary of our proposal:** Starting from an original labeled dataset, we compare the performance of a classification model with and without the use of data augmentation. In this work, we examine the performance of classifiers of human activities. The classification is based on sensor data from wearable devices. To augment data, we used synthetic samples generated by two distinct types of generative models - traditional generative models and large language models. This evaluation is conducted through cross-validation, with the analyses incorporating both quantitative and qualitative metrics.

with minimal examples. We also investigate the potential of LLMs to enhance Human Activity Recognition (HAR) datasets, what remains largely unexplored, especially in scenarios with limited data— such as few-shot learning— we selected HAR as our case study. We selected the HAR scenario as it is a critical field for advancing healthcare and well-being [10].

In this study, we evaluate the effectiveness of LLM-generated synthetic data in augmenting datasets using the metrics: Mixed, TSTR (Train on Synthetic, Test on Real), and TRTS (Train on Real, Test on Synthetic) [11] and Predictive Capacity (PC) [12]. These metrics provide a comprehensive evaluation of the predictive power, generalizability, and robustness of the synthetic data when integrated with real-world datasets. By comparing the performance of models trained on synthetic data generated by different LLMs, including GPT-4 [6], Command R+ (Cohere) [13], and Gemini [14], without fine-tuning, we aim to assess whether LLMs can effectively support data augmentation in a samples and computational resources scarcity scenario. Figure 1 summarizes our approach. Specifically, we address the key question: *Are LLMs useful for time-series data augmentation?*. The key contributions of this paper are as follows:

- We demonstrated that LLMs, such as GPT-4 and Cohere, can generate synthetic samples for time-series sensors from only 3-shot samples without fine-tuning, relying solely on the LLM's inference capabilities.
- By leveraging different metrics like F1-score, accuracy, and analyzing the confidence intervals, our experiments showed where each LLM excelled (e.g., Gemini providing more stable data) and identified their limitations, particularly in more unbalanced datasets.
- We evaluated the diversity of the synthetic data and its ability to capture the characteristics of time-serie sensors. Our findings underscore the representativeness and practical utility of LLM-generated data in improving HAR model performance.
- We compared the performance of various LLMs and traditional generative models, assessing the practical value of these approaches in HAR tasks, particularly in augmenting limited training data.
- We examine under which conditions LLMs outperform or complement generative models like Generative Adversarial Networks (GANs) or diffusion models.
- We analyzed the trade-offs between the costs of acquiring and annotating real-world data versus the use of LLM-generated synthetic data. This comparison provides practical insights into LLMs' scalability and resource efficiency for data augmentation.

2 Related Works

The TimeGAN model [15] is a foundational approach for generating time-series data, especially in HAR. An advancement on this is the DroppelGANger [16], which enhances TimeGAN's capabilities. A recent method using transformers is the TTS-GAN [17], which effectively generates high-quality temporal data. Diffusion Models [18], such as the SSSD^{S4} (SSSD) [19], have gained attention for their ability to manage complex, high-dimensional data distributions and incorporate temporal

information. In the context of large language models (LLMs), pretrained models like Gemini [14] and GPT-4 [6] have emerged as influential tools. These models have significantly impacted various applications, including healthcare [20] and time-series tasks [4]. Additionally, the Command R+ model from Cohere [13] is a new entrant LLM with an ongoing exploration of its potential (see more in supplementary materials).

3 Methodology

3.1 Evaluation

We used the evaluation framework proposed by Souza et al. (2023) [11], a protocol that combines both quantitative (employing recall, accuracy, and the F1 score) and qualitative metrics [21]. Focusing on the analysis of the samples generated from the LLMs is useful to address HAR scarcity we analyzed the metric Predictive Capacity (PC) [12]. The protocols can be summarized as follows:

- **Train on Synthetic, Test on Real data (TSTR):** Verifying the efficiency of the synthetic samples to model the original data distribution, allowing training an ML classifier using only synthetic samples [21].
- **Train on Real, Test on Synthetic (TRTS):** Assures the realism of the synthetic samples. It evaluates if the generated sample is able to mirror the original training data [21].
- **MIXED:** Evaluating the capacity of synthetic data to augment the original dataset. The evaluation ensures the validity of this artificial data for expanding the training dataset. This protocol evaluates if the synthetic data does not fully capture the data's diversity due to inherent biases. This challenge is frequently referred to as the Domain Shift dilemma [22][23].
- **Data quality:** Based on three criteria proposed by Fekri et al. (2019) [24], we evaluate the quality of the synthetic data samples. The criteria are Fidelity (accuracy), Diversity, and Generalization. Souza et al. (2023) demonstrate that adequate data samples met all of them.
- **Predictive Capacity (PC):** The metric [12] assesses the capacity and quality of the synthetic data to train the model. PC ≤ 0.9 as having synthetic data with inferior quality and PC ≥ 0.9 as having synthetic data with good quality. PC is measured as:

$$PC(Score) = \frac{\textit{Score of model trained on synthetic data}}{\textit{Score of model trained on real data}}$$

In TSTR, we trained the HAR using 30 synthetic samples in each fold (for each class), and the baseline for comparison was also trained using only 30 samples (chosen randomly). Due to the cost, this was the maximum amount of data we could collect from the LLMs. In TRTS, we used 30 samples for each category in train set, for a fairer comparison with the LLMs, for which we requested only this amount of samples. For the MIXED protocol, we systematically combined synthetic samples with the original training sample and randomly selected the samples for inclusion. The Tables that summarize the results present the "#Baseline" model, which means we adopt the approach of Training the model on Real data and Testing on Real data (TRTR). Due to space constraints in this paper, we only present the average from the protocols when comparing LLMs to traditional generative models. Nevertheless, complete results will be made available in the supplementary materials. Usually, the PC is measured on accuracy, but we assessed for accuracy and F1 scores due to the class imbalance.

Table 1: Results per model and datasets under the Mixed protocol.

Dataset	Model	Added	Accuracy	Recall	F1	Dataset	Model	Added	Accuracy	Recall	F1
MHEALTH	Baseline	–	94.74±0.87	95.20±0.79	95.14±0.80	MHAD2	Baseline	–	69.55±2.42	68.40±2.48	68.53±2.41
	Cohere	100	84.11 ± 2.28	81.27 ± 3.10	81.31 ± 2.99		Cohere	100	68.75 ± 2.36	62.52 ± 3.66	62.38 ± 2.83
		150	83.30 ± 3.11	80.5 ± 3.11	80.51 ± 2.99			150	69.48 ± 3.30	63.21 ± 4.91	62.89 ± 4.42
		300	81.91 ± 2.15	79.52 ± 3.12	79.74 ± 3.02			300	69.17 ± 3.49	64.36 ± 4.62	63.50 ± 4.47
	Gemini	100	83.31 ±0.94	80.37 ±2.66	80.32±2.58		Gemini	100	70.42 ± 2.38	63.99 ± 4.76	63.7 ± 4.14
		150	84.27 ± 2.23	80.53 ± 3.06	80.73 ± 2.95			150	69.69 ± 2.55	66.26 ± 4.74	65.83 ± 4.20
		300	81.68 ±1.88	79.01 ±2.72	78.65 ±1.88			300	71.35 ± 4.07	70.1 ± 5.8	67.16 ± 5.36
	GPT-4	100	95.26 ±1.05	95.68 ±0.93	95.60 ±0.98		GPT-4	100	70.54±2.59	69.57±3.06	69.79±3.14
		150	**95.46** ±0.87	**95.86** ±0.84	**95.78** ±0.82			**150**	**70.80**±2.98	**69.81**±3.19	**69.84**±3.68
		300	96.13 ±1.29	96.54 ±1.13	96.44 ±1.15			300	68.57±2.23	67.88±1.39	67.53±2.88

3.2 LLM Prompt

We used a prompt to request LLMs to generate time-series data from a wearable device. The generated data samples were designed to follow the same distribution as the original data provided within the prompt. To achieve this, we randomly selected three data instances from the desired class and the same dataset fold, resulting in nd-arrays with a shape of (3, temporal window, 3). We used these arrays as the context for the prompt given to LLMs (see supplementary materials for more details).

4 Experiment Setup

4.1 Dataset

The study employed the UTD-MHAD dataset [25], featuring 27 different actions performed by eight individuals. This dataset comprised data from accelerometer and gyroscope sensors and divided into two subsets, UTD-MHAD1 (21 activities) and UTD-MHAD2 (6 activities). The experiments specifically focused on the MHAD2 subset and accelerometer data. We also incorporated the Mobile HEALTH (MHEALTH) dataset, which includes 12 distinct activities performed by ten participants each. Data was collected from various sensors, including an accelerometer and ECG. Both data processing followed the procedures detailed in Dclassifier [26], specifically focusing on accelerometer data.

4.2 Evaluated Models

We utilized three large language models (LLMs) in the data generation experiments: GPT-4 [6], Command R+ (Cohere) [13], and Gemini 1.5 flash [14]. All models were employed with their default temperature settings to maintain consistency across the tests. To evaluate the effectiveness of these models, we also compared their performance with traditional generative approaches, selecting four state-of-the-art Generative Adversarial Networks (GANs): TimeGAN [15], DGAN [16], Time-LogCosh-GAN (TLCGAN) [11], and TTS-GAN [27], as well as a diffusion model, SS [19]. These GANs were trained for 200 epochs, while the diffusion model was trained for 1,000 iterations. Additionally, we assessed performance using a DeepConv LSTM classifier (Dclassifier) [26] the state-of-art HAR for various datasets, including MHAD2 and MHEALTH, employing a 10-fold cross-validation method over 16 epochs, with all models set to their default parameters. For further technical details, please refer to the supplementary material.

5 Results and Discussion

5.1 Do LLMs contribute to the classifier's performance ?

LLMs, such as GPT-4, have been shown to positively contribute to classification tasks. However, their impact varies depending on the dataset and the performance metric being considered. As shown in Table 1, GPT-4 is particularly useful for improving metrics like accuracy, F1-score, recall, and stability, as indicated by narrower confidence intervals. Nevertheless, there is a limit to how much synthetic data should be added before performance begins to degrade due to saturation. As traditional models (see [11]), beyond a certain point, adding more data can lead to redundant or low-diversity samples, which negatively impact the model's effectiveness. In contrast, the other models tested, such as Gemini and Cohere, were less effective at generating high-quality synthetic data.

In some cases, they even reduced overall performance. This may be because, unlike GPT-4, these models are less capable of following prompt instructions [9] accurately and fail to generate data that meaningfully balances the dataset or improves the representation of minority classes. In particular, Cohere produced synthetic data that harmed performance by introducing more redundancy and less diversity (see Figure 7 in the Section A.5), which hindered the model's ability to learn effectively. When applied to more complex and unbalanced datasets like MHAD2 (more details in Section A.4), GPT-4 still contributed positively by improving the F1-score. However, its overall impact diminished in terms of accuracy and robustness. This suggests that while GPT-4 excels at capturing general patterns, it struggles with properly representing minority classes in skewed distributions. As a result, the model's ability to improve performance in such datasets is limited [28]. In summary, while GPT-4 improves HAR performance by generating useful synthetic data, its benefits are constrained by the

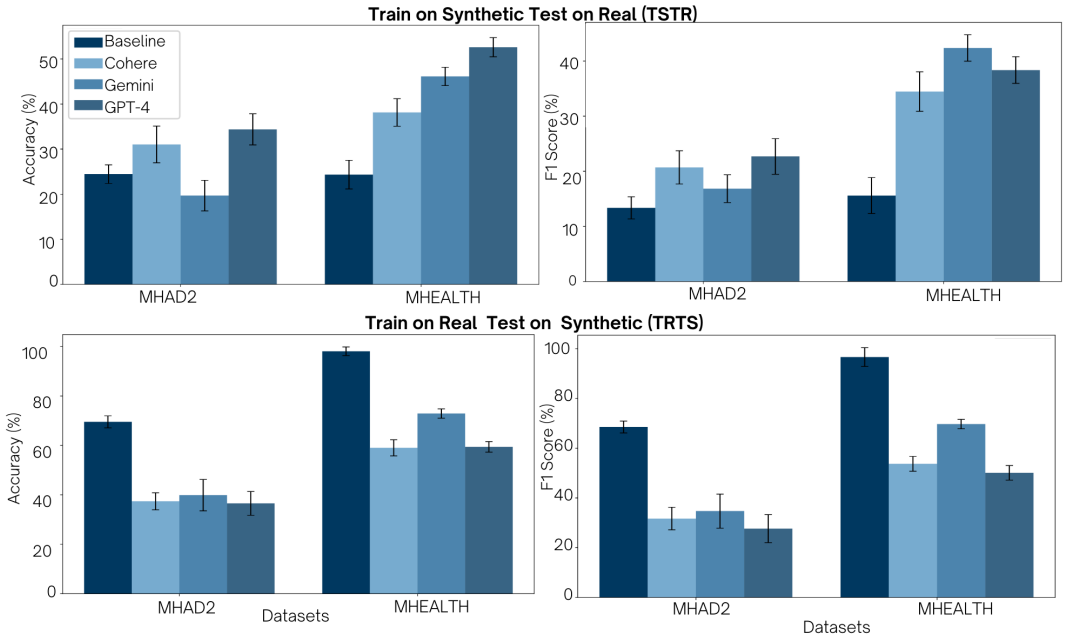

Figure 2: Comparison of model performance across different datasets using TSTR and TRTS protocols. The thin lines centered on the bars represent the confidence intervals, while the bars depict the assessed metrics. The y-axis shows the metric values, and the x-axis indicates the evaluated datasets. A notable disparity between accuracy and F1 scores highlights how class imbalance affects model performance. The Gemini model stands out, having the smallest gap between the two metrics.

complexity and balance of the dataset. Care must be taken to avoid data saturation, and challenges with minority class representation persist, particularly in more difficult scenarios like MHAD2.

5.2 How do synthetic data generated by LLMs impact the classifier's stability?

The analysis (see Table 1) suggests that the synthetic data generated impacts the stability of metrics like recall and F1-score. In the experiments, adding synthetic data led to increased variability in model performance. However, GPT-4 demonstrated greater consistency, showing less variation in the metrics, with narrower confidence intervals compared to the other models. This suggests that the data generated by GPT-4 is of higher quality, contributing more reliably to the models' performance. In contrast, synthetic data generated by Gemini and Cohere introduced higher variability, especially in the MHAD2 dataset. This indicates that their synthetic data may not contribute as effectively and could negatively impact the robustness of the model's predictions [29]. In certain scenarios, synthetic data provided more stability, as seen in the MHEALTH dataset under the TSTR (see Figure 2). In this case, most results showed narrower confidence intervals than the baseline, indicating that the synthetic data contributed positively to model stability [30]. However, in datasets with more complex or unbalanced distributions, such as MHAD2, synthetic data tended to introduce greater variability in the model's predictions. This was particularly evident in the TRTS metric (see Figure 2), where wider confidence intervals suggested that models trained on synthetic data exhibited greater uncertainty when tested on real data. This suggests that while LLMs like GPT-4 can capture general patterns effectively, their ability to represent minority or less frequent classes may be limited, especially in datasets with skewed class distributions, leading to decreased stability [31].

5.3 Does synthetic data help improve minority class classification?

Yes, synthetic data generated by LLMs has proven useful for balancing datasets and improving the classification of minority classes. In both MHEALTH and MHAD2 datasets, models trained with synthetic data showed improved recall and F1-scores, indicating dealing better with minority class [32], particularly when using the GPT-4 model. However, despite these improvements in class

distinction, a noticeable discrepancy between accuracy and F1 scores was observed. This suggests that while synthetic data helps in distinguishing between classes (improving F1), it does not always lead to better overall predictive accuracy. This discrepancy highlights the complexity of the data and the impact of the imbalance. In the more imbalanced MHAD2 dataset, GPT-4 achieved the highest F1 score but also exhibited a larger discrepancy between accuracy and F1 (around 10%). This indicates that, while the model performs well on minority class identification, it struggles to maintain overall accuracy, likely due to the dataset's class imbalance. On the other hand, although the Gemini model did not achieve the highest F1 score, showed a much smaller discrepancy between F1 and accuracy (around 3%), suggesting it handles imbalanced data more robustly. The impact of synthetic data was more consistent in the less imbalanced MHEALTH dataset, where models like GPT-4 significantly improved accuracy and F1-score compared to the baseline. This reinforces the idea that synthetic data is more effective when the dataset is more balanced [33]. In summary, while synthetic data helps improve minority class classification, the degree of dataset imbalance and the specific model used significantly affect the overall performance.

5.4 What are the trade-offs between the synthetic data amount and the classifier's performance?

LLMs, such as GPT-4, can effectively generate useful augmentation data for HAR tasks. However, this contribution is sensitive to the volume of data added. GPT-4 demonstrates better generalization and stability compared to other models, but the quality of synthetic data must be closely monitored to avoid saturation and performance degradation. The experiments reveal a critical balance point regarding synthetic data generation (see Table 1). Moderate amounts of data—specifically between 100 and 150 samples per class—significantly enhance performance metrics. However, increasing the amount beyond 300 samples can lead to diminished returns and, in some cases, a decrease in performance. This decline is often due to reduced diversity or the creation of less relevant data, highlighting the importance of regulating the amount of generated data to prevent overfitting or loss of quality. For GPT-4, adding between 150 to 300 samples can improve performance on certain metrics. Conversely, other models like Gemini and Cohere experienced a drop in performance when faced with excessive synthetic data, likely due to the lack of diversity among samples. This suggests a saturation point in synthetic data generation, where additional data begins to negatively impact overall performance.

Table 2: Predictive Capacity (PC) for Accuracy and F1

Dataset	Model	PC (Accuracy)	PC (F1)
MHAD2	GPT-4	≈ 1.41	≈ 1.69
	Gemini	≈ 0.80	≈ 1.26
	Cohere	≈ 1.27	≈ 1.55
MHEALTH	GPT-4	≈ 2.16	≈ 2.46
	Gemini	≈ 1.90	≈ 2.72
	Cohere	≈ 1.57	≈ 2.21

5.5 Is it possible to generate diverse synthetic data using only three-shot samples ?

The experiments demonstrated that while LLMs, such as GPT-4, can generate synthetic data from a limited number of initial samples, the effectiveness of this data is highly dependent on the task - Figure 2 and Table 1 express that - and the complexity of the original dataset. In certain scenarios, the generated data captures relevant aspects of the original distribution (see Figures 7 and 6 in the Section A.5); however, in others, it exhibits limited diversity and representativeness, negatively impacting model learning. For data augmentation to be effective, the generated synthetic data must provide diverse and novel information to the model [11, 34]. GPT-4 largely succeeds in this regard, as evidenced by performance improvements observed when synthetic data is added, particularly within specific thresholds—300 samples for the MHEALTH dataset and 150 samples for the MHAD2 dataset. Nonetheless, the results indicate that exceeding these amounts can degrade performance, suggesting a limit to the diversity and usefulness of the generated data.

In contrast, the Cohere model appears to generate data with reduced diversity. This is reflected in performance declines as more data is added across both datasets, implying that the generated data may be excessively redundant or poorly aligned with the classification model. Additionally, the observed decrease in F1-score and recall relative to accuracy indicates that Cohere's synthetic data

might distort class balance or fail to be effectively utilized by the model. Gemini exhibited similar trends, showing a slight decline in quality with increased data, reinforcing concerns regarding the lack of diversity in the synthetic samples produced by various models. While it is feasible to generate synthetic data from only three initial samples, the resultant diversity and effectiveness depend on the model used and the management of the data volume.

5.6 Can LLMs generate useful synthetic data ?

Generating synthetic data from only three samples is challenging due to the limited context available for the model to learn from. This often results in a lack of diversity in the generated data, as observed in several scenarios. However, LLMs like GPT-4 have demonstrated an impressive ability to generalize, even with minimal data. Their large-scale pretraining allows them to infer key patterns and characteristics from a few samples, generating synthetic data that retains essential aspects of the original dataset. In experiments, LLM-generated data showed good fidelity (see Section A.5), especially when evaluated using the TSTR metric, which revealed that the real data's characteristics were well reproduced. However, despite this fidelity, the synthetic data often lacked diversity, limiting its usefulness in data augmentation and model robustness. For instance, the TRTS metric indicated that while the synthetic data captured certain patterns, it did not surpass the baseline in complex or unbalanced datasets like MHAD2, where diversity is crucial for better generalization. LLM-generated data has been more effective for simpler and more balanced datasets, such as MHEALTH. GPT-4, in particular, was able to generate useful data with just three initial samples, improving model performance in specific metrics. This suggests that, under the right conditions, synthetic data generated by LLMs can still be valuable for training. However, models like Gemini struggled to provide diverse synthetic data, leading to a drop in performance when more samples were added, suggesting redundancy or lack of variation. In contrast, GPT-4 showed greater diversity in its synthetic data, contributing to improved model robustness, especially in scenarios requiring mixed data sources. Overall, metrics like TSTR and PC(Table 2) indicated that LLMs are capable of generating high-quality synthetic data, with several models achieving a PC above 0.9. This suggests that, while synthetic data may not always surpass real data in performance, it is still useful for training [12], particularly when generated by GPT-4.

5.7 Does class balancing affect LLMs ?

The previous sections show that the synthetic data tested achieved the best performance on the MHEALTH dataset. Both datasets are imbalanced, but MHAD2 is more imbalanced than MHEALTH (see Section A.4), which impacts the variability of the results. When randomly selecting three samples from each class in the MHEALTH dataset, the sampling process favors overrepresented classes. This increases the likelihood of producing a subset that accurately represents the overall data distribution [35]. Such a representative subset allows LLMs to generate synthetic data that better captures key patterns and features of the original data, leading to superior augmentation and enhanced classifier performance. Conversely, in a more imbalanced dataset like MHAD2, randomly selecting three samples per class is less likely to capture a representative cross-section of the data [35]. In these datasets, certain classes have far fewer examples, which increases the risk that the chosen samples may not adequately reflect the diversity and distribution of the dataset. Consequently, the synthetic data generated by LLMs may fail to capture the full spectrum of the data's complexity, resulting in poorer performance. Thus, the degree of dataset imbalance directly impacts the representativeness of the samples available to the LLMs. In more balanced datasets (like MHEALTH), the synthetic data generated from a small, random sample set is more likely to reflect the overall dataset, leading to better model performance. In contrast, in more imbalanced datasets (like MHAD2), the same number of randomly selected samples is less likely to be representative, resulting in less effective synthetic data and lower performance.

5.8 Traditional models vs. GPTs

Traditional models still tend to outperform LLMs in some contexts. However, considering that a small number of samples is available, LLMs provide a significant advantage by generating additional data to improve classifier performance. GPT-4 outperforms the baseline, introducing synthetic samples to the MHAD2 training set, in two of the three demonstrated scenarios (see Figure 3). In these scenarios, considering accuracy, it surpasses both TimeGAN and SSSD. However, when considering

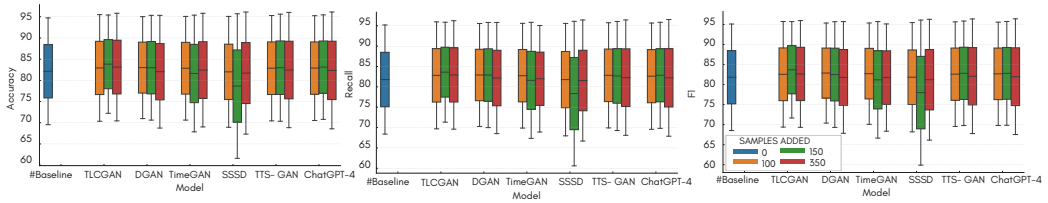

Figure 3: Average performance of the models per dataset. Each plot corresponds to one of the metrics evaluated; each color represents the incremental volume of synthetic data added to both original datasets, while each row corresponds to the method applied for blending synthetic and real data. The metrics correspond to the average of the metric in each dataset.

the arithmetic mean, TTS-GAN also falls within this category. TLCGAN demonstrated the highest performance, with nearly 3% more accuracy and a higher average. With respect to the MHEALTH dataset, GPT-4 outperforms the baseline in all scenarios. Within these, it surpasses TLCGAN in two out of three scenarios, thus becoming the top-performing model. This holds true both in terms of average and accuracy, implying that GPT's performance is notably high since the average was maintained and the datasets are unbalanced. GPT-4 is not the best-performing model on TSTR at the MHAD2 and MHEALTH datasets. Still, it enhances the baseline by approximately 10% on the MHAD2 TSTR evaluation and almost 28% on the MHEALTH TSTR evaluation. It outperforms SSSD (with an improvement of around 14%) and TTS-GAN (with a gain of roughly 8%) on MHAD2 and TimeGAN (with a gain of roughly 32%), DGAN (more than 40%), and TTS-GAN (more than 30%) on MHEALTH. The TLCGAN showed the most exceptional performance, exceeding the baseline in both scenarios in MHAD2. In MHEALTH, it was the second best on TSTR but the best model on TRTS, even without outperforming the baseline. Even though it is not the top-performing model in all scenarios, GPT-4 is competitive compared to other models. Like the others, augmenting data improves the model's performance to a certain extent, suggesting a consistency in the results produced when employing this model. Section A.6 provides more discussion about it.

5.9 Data Acquisition Costs and LLM Usage

As highlighted earlier, we employed a Large Language Model (LLM) to generate synthetic data without fine-tuning, using just three samples per class as input. Despite the limited initial data, the LLM's ability to generate useful synthetic datasets proves to be a cost-effective solution in data-scarce environments. Our results suggest that LLMs can sometimes rival or even outperform traditional models, which typically require extensive datasets for comparable generalization performance [36]. This highlights a significant trade-off: while traditional models rely heavily on large-scale data collection, incurring higher computational and financial costs, LLMs achieve similar results with minimal data and computational resources. LLMs also present an advantage in terms of monetary investment. Synthetic data can be generated using trial API keys from various LLM providers. For instance, using the Cohere API, it is possible to generate two and a half folds of synthetic data for a dataset with twelve classes, collecting 30 samples per class, entirely with a free key. Similarly, Gemini allows the creation of five folds under the same conditions, while GPT-4 can generate only half a fold before reaching usage limits. Therefore, in scenarios where data is limited, the cost-effectiveness and reduced resource requirements of LLMs make them a viable and attractive option for synthetic data generation, offering a compelling balance between data acquisition costs and computational efficiency.

6 Conclusion

Our study investigated the effectiveness of LLMs like GPT-4, Cohere, and Gemini in generating synthetic samples for data augmentation through a few-shot learning (three samples). The results show that LLMs, while not consistently outperforming real-world data, offer certain advantages and can be considered useful for HAR augmentation in specific contexts. For datasets like MHEALTH, LLMs— especially Gemini—generated data that closely resembled real distributions, leading to more stable performance with narrower confidence intervals. However, in more unbalanced datasets such as MHAD2, the effectiveness of LLM-generated data diminished, indicating limitations in handling

class imbalance. GPT-4, in particular, improved the baseline performance when considering metrics like F1-score, suggesting its potential for enhancing classifier performance. While LLM-generated data may not always surpass the utility of real-world data in terms of diversity, they offer a viable solution in data and computational resource scarcity scenarios. Additionally, their ability to generate useful data without fine-tuning presents a cost-effective method for augmenting HAR datasets.

References

[1] Aleksandr Ometov, Viktoriia Shubina, Lucie Klus, Justyna Skibińska, Salwa Saafi, Pavel Pascacio, Laura Flueratoru, Darwin Quezada Gaibor, Nadezhda Chukhno, Olga Chukhno, Asad Ali, Asma Channa, Ekaterina Svertoka, Waleed Bin Qaim, Raúl Casanova-Marqués, Sylvia Holcer, Joaquín Torres-Sospedra, Sven Casteleyn, Giuseppe Ruggeri, Giuseppe Araniti, Radim Burget, Jiri Hosek, and Elena Simona Lohan. A survey on wearable technology: History, state-of-the-art and current challenges. *Computer Networks*, 193:108074, 2021.

[2] Bohan Li, Yutai Hou, and Wanxiang Che. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90, 2022.

[3] Mengchen Liu, Jiaxin Shi, Kelei Cao, Jun Zhu, and Shixia Liu. Analyzing the training processes of deep generative models. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):77–87, 2018.

[4] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Health-llm: Large language models for health prediction via wearable sensor data, 2024.

[5] Emilio Ferrara. Large language models for wearable sensor-based human activity recognition, health monitoring, and behavioral modeling: A survey of early trends, datasets, and challenges. *Sensors*, 24(15), 2024.

[6] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[7] Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024.

[8] Zhengyang Xiao, Wenyu Li, Hannah Moon, Garrett W Roell, Yixin Chen, and Yinjie J Tang. Generative artificial intelligence gpt-4 accelerates knowledge mining and machine learning for synthetic biology. *ACS synthetic biology*, 12(10):2973–2982, 2023.

[9] Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Lida Chen, Xintao Wang, Yuncheng Huang, Haoning Ye, Zihan Li, Shisong Chen, Yikai Zhang, Zhouhong Gu, Jiaqing Liang, and Yanghua Xiao. Can large language models understand real-world complex instructions?, 2024.

[10] Md Zia Uddin and Ahmet Soylu. Human activity recognition using wearable sensors, discriminant analysis, and long short-term memory-based neural structured learning. *Scientific Reports*, 11(1):1–15, 2021.

[11] Maynara Donato de Souza, Clesson Roberto Silva Junior, Jonysberg Quintino, André Luis Santos, Fabio Q B da Silva, and Cleber Zanchettin. Exploring the impact of synthetic data on human activity recognition tasks. *Procedia Computer Science*, 222:656–665, 2023. International Neural Network Society Workshop on Deep Learning Innovations and Applications (INNS DLIA 2023).

[12] Rabindra Khadka. How to evaluate the robustness of synthetic data, June 9 2021. ClearBox AI.

[13] Cohere Team. Introducing coral, the knowledge assistant for enterprises. `https://cohere.com/blog/introducing-coral`, 2023. Accessed: 2024-09-16.

[14] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung,

Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha

Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-

Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2024.

[15] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.

[16] Zinan Lin, Alankar Jain, Chen Wang, Giulia Fanti, and Vyas Sekar. Using gans for sharing networked time series data: Challenges, initial promise, and open questions. In *Proceedings of the ACM Internet Measurement Conference*, pages 464–483, 2020.

[17] Xiaomin Li, Vangelis Metsis, Huangyingrui Wang, and Anne Hee Hiong Ngu. Tts-gan: A transformer-based time-series generative adversarial network. *arXiv preprint arXiv:2202.02691*, 2022.

[18] Jonathan Ho, Xi Chen, and Arnaud Doucet. Time series generation using diffusion models. *arXiv preprint arXiv:2006.11257*, 2020.

[19] Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models. *arXiv preprint arXiv:2208.09399*, 2022.

[20] Som S Biswas. Role of chat gpt in public health. *Annals of biomedical engineering*, 51(5):868–869, 2023.

[21] Mohammad Fekri, Ananda Mohon Ghosh, and Katarina Grolinger. Generating energy data for machine learning with recurrent generative adversarial networks. *Energies*, 13, 12 2019.

[22] Viktor Seib, Benjamin Lange, and Stefan Wirtz. Mixing real and synthetic data to enhance neural network training–a review of current approaches. *arXiv preprint arXiv:2007.08781*, 2020.

[23] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018.

[24] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022.

[25] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 168–172, 2015.

[26] Satya P Singh, Madan Kumar Sharma, Aimé Lay-Ekuakille, Deepak Gangwar, and Sukrit Gupta. Deep convlstm with self-attention for human activity decoding using wearable sensors. *IEEE Sensors Journal*, 21(6):8575–8582, 2020.

[27] Xiaomin Li, Vangelis Metsis, Huangyingrui Wang, and Anne Hee Hiong Ngu. Tts-gan: A transformer-based time-series generative adversarial network, 2022.

[28] Fadi Thabtah, Suhel Hammoud, Firuz Kamalov, and Amanda Gonsalves. Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513:429–441, 2020.

[29] R Confalonieri, S Bregaglio, and M Acutis. A proposal of an indicator for quantifying model robustness based on the relationship between variability of errors and of explored conditions. *Ecological Modelling*, 221(6):960–964, 2010.

[30] Chris Chatfield. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3):419–466, 1995.

[31] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

[32] Forhad An Naim, Ummae Hamida Hannan, and Md Humayun Kabir. Effective rate of minority class over-sampling for maximizing the imbalanced dataset model performance. In *Proceedings of Data Analytics and Management: ICDAM 2021, Volume 2*, pages 9–20. Springer, 2022.

[33] Ayesha Siddiqua Dina, AB Siddique, and D Manivannan. Effect of balancing data using synthetic data on the performance of machine learning classifiers for intrusion detection in computer networks. *IEEE Access*, 10:96731–96747, 2022.

[34] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best practices and lessons learned on synthetic data, 2024.

[35] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[36] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *International Journal of Computer Vision*, 128(10):2586–2606, 2020.

[37] Allyson Ettinger. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 01 2020.

[38] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

[39] Luciano Rossoni and GPT Chat. A inteligência artificial e eu: escrevendo o editorial juntamente com o chatgpt. *Revista eletrônica de ciência administrativa*, 21(3):399–405, 2022.

[40] Nigar M Shafiq Surameery and Mohammed Y Shakor. Use chat gpt to solve programming bugs. *International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290*, 3(01):17–22, 2023.

[41] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

A Supplementary Material

A.1 Related Works

The TimeGAN model [15] represents a foundational approach to generating time-series data and is widely employed in the field of Human Activity Recognition (HAR). It has significantly influenced subsequent research as one of the pioneering works in applying Generative Adversarial Networks (GANs) to time-series data. An advancement in this area is the DroppelGANger model [16], which enhances the TimeGAN framework by introducing techniques that improve data generation quality and efficiency.

A more recent trend in this domain involves the use of transformers, exemplified by the TTS-GAN [17]. This approach leverages the strengths of transformer architectures to generate high-quality temporal data, reflecting the evolving landscape of generative models.

Another noteworthy model gaining traction in the context of temporal data is the Diffusion Model [18]. Its ability to modulate complex and high-dimensional data distributions makes it particularly effective for handling the inherent uncertainty present in time-series data. The SSSD^{S4} (SSSD) model [19] serves as a prime example of how diffusion models can incorporate temporal information.

In recent years, the practice of pretraining Natural Language Processing (NLP) models using a language modeling objective has gained substantial attention [37]. Notable instances include Gemini [14] and the GPT series [6], with GPT-4 emerging as one of the most popular variants. These models have been widely utilized across various tasks, impacting numerous fields beyond academia, as evidenced by studies on their applications in education [38], editorial processes [39], software debugging [40], and healthcare [20].

Additionally, the Command R+ model from Cohere [13] represents a new and influential development in large language models (LLMs). Its recent introduction has sparked community interest as researchers explore its potential applications and capabilities.

A.2 Models details

This section presents some details about the models employed in the main paper.

The GPT-4 [6] is an LLM developed by OpenAI that excels at natural language processing (NLP) tasks, such as text generation, summarization, translation, and more. It improves upon its predecessors with enhanced reasoning capabilities, a larger context window, and better performance in nuanced tasks like dialogue comprehension and text-based reasoning.

Command R+ by Cohere [13] is an advanced version of Cohere's language models focused on retrieval-augmented generation (RAG). It specializes in tasks that involve recalling external knowledge, offering better precision in generating text-based responses grounded in factual retrieval. This model is particularly optimized for business applications like customer support, content generation, and summarization.

Gemini 1.5 Flash [14] is part of Google's Gemini family of AI models, combining strengths in both generative and reasoning tasks. It focuses on multimodal applications, integrating language, vision, and reasoning in a powerful and efficient framework. The "Flash" version likely indicates improvements in speed and adaptability, making it suitable for real-time tasks and interactive AI applications.

Time-LogCosh-GAN (TLCGAN) is a traditional GAN model that uses two separate random noise inputs, z_1 and z_2, generated in a time-series format[11]. This model was trained using 10-fold stratified data over 200 epochs, with a learning rate of $lr = 0.0001$ and a batch size of 5.

The Time-series GAN (TimeGAN) is a generative model trained adversarially and jointly. It uses a learned embedding space with both supervised and unsupervised losses. It comprises four network components: an embedding function, a recovery function, a sequence generator, and a sequence discriminator. In this study, we set the maximum sequence length and the hidden dimension size to 50. The model underwent training over 200 epochs with all other parameters set to their default values.

Transformer-based Time-Series Generative Adversarial Network (TTS-GAN) architecture comprises two primary components: a generator and a discriminator. Both components are built using the transformer encoder architecture [6]. The encoder comprises two compound blocks. The first block utilizes a multi-head self-attention module, while the second block is a feed-forward MLP with a GELU activation function. The model was trained for 200 iterations, with the sequence length matching the size of the temporal windows from the utilized dataset. All other parameters were kept at their default values.

The Structured State Space Diffusion SSSD^{S4} (SSSD) model [19] is a diffusion model designed for time-series data, drawing inspiration from a generative model for audio - DiffWave [41]. The model was trained using the default configuration for 1,000 iterations.

DroppelGANger (DGAN) is a synthetic data generation framework based on generative adversarial networks (GANs) [16], consisting of a metadata generator and a min/max generator. We adjusted it to set the max sequence length and sample length equal to the temporal window length from the used dataset. The batch size was 5, and the learning rate was $lr = 10^{-4}$ for both the discriminator and the generator. The model was trained for 200 epochs.

A.3 Prompting details

As previously mentioned, we selected three data instances from the desired class and from the same dataset fold, which resulted in nd-arrays with a shape of (3, temporal window, 3). We used these arrays as the context for the prompt given to the LLMs (Figure 4 illustrates it). All the LLMs received the same samples as context to ensure fairness.

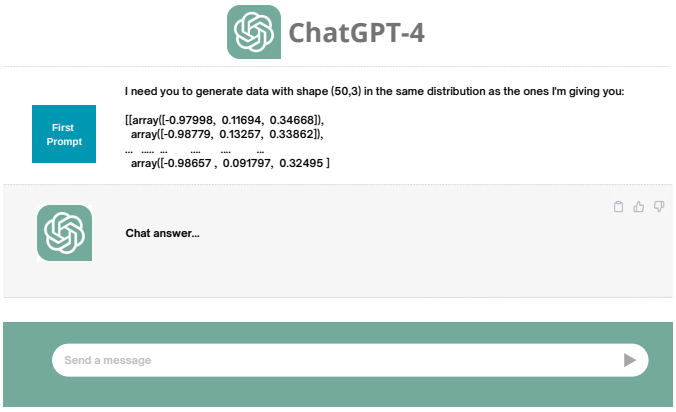

Figure 4: Illustration of the prompt given to ChatGPT-4: We provide an nd-array data with a shape of (3,50,3). Here, '3' represents the number of examples, '50' denotes the time-series window size, and the final '3' corresponds to the number of accelerometer attributes. We then ask it to generate a set of points with dimensions (50,3) that follow the same distribution as the input data. We intentionally maintain the prompt straightforward to evaluate the zero-shot performance of the language model in understanding the data distribution. The strategy to generate one sample at a time is employed to reduce the chance of errors or misunderstandings by the LLM when interpreting the prompt.

A.4 On the datasets imbalacement

In this section, we discuss the improvement of the adopted datasets. Figure5 displays two histograms representing the distribution of two variables, MHAD2 and MHEALTH. The x-axis likely represents different categories or classes, while the y-axis represents the count of observations within each class.

In both datasets, some classes are significantly overrepresented (e.g., class 11 in MHAD1 and class 4 in MHAD2). This imbalance can lead to biased model performance. For instance, a model might perform well in the majority classes but fail to learn adequately in the minority classes.

For the MHAD2 dataset, the imbalance is more pronounced, with class 4 having the highest count (around 220), followed by class 5 with a slightly lower count. On the other hand, classes 2 and 3 have

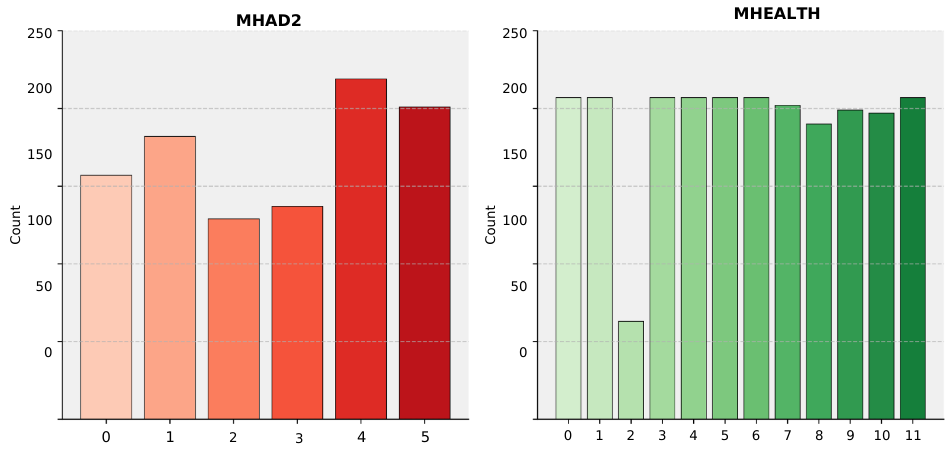

Figure 5: Mean number of samples per class across datasets. The datasets are divided into stratified 10-folds, with the bars representing the mean number of samples per class across all folds.

much lower counts (around 100), while classes 0 and 1 show an intermediate count. The imbalance could likely bias models towards predicting classes 4 and 5 more frequently, as they dominate the dataset. Minority classes like 2 and 3 could be underrepresented in the model's learning process, leading to poor classification for those specific classes.

The imbalance in the MHEALTH model is less severe compared to MHAD2. Most classes (except class 0 and class 3) have around 250 examples. Class 0 is significantly underrepresented, with less than 50 examples, and class 3 has around 100 examples. Although there is still some imbalance, especially with class 0, it is not as extreme as in MHAD2. The model is likely to perform more evenly across most classes, though it might struggle to correctly classify examples of class 0 due to the low representation.

A.5 LLMs data's distribution

In this section, we provide plots about the LLMs data's distributions. It is possible to compare the original distribution of the data from fold 0 and category 0 with the synthetic for the same fold and category. For a better visualization, we divided the figure in two, allowing the reader to see the details and avoiding shrinking the figures.

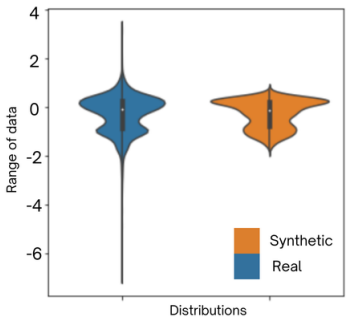

Figure 6: Comparative analysis of synthetic (orange) and real (blue) distributions using the MHAD2 Dataset. The GPT-4 model provides the synthetic data. There are noticeable similarities in trends between synthetic and real distributions. However, these synthetic datasets diverge from each other in terms of data range and statistics metrics.

It is evident that the synthetic data generally follows the same distribution pattern as the real data, although some discrepancies can be observed. For instance, in the case of Gemini (see Figure 7), the synthetic data has a lower minimum value and a higher maximum compared to the original.

Conversely, for Cohere and GPT, the minimum values are higher, and the maximum values are lower than the original data. Notably, the data generated by Cohere and Cohere are quite similar, while GPT's data stands out with more divergence (see Figure 6). Despite this, both Cohere and Gemini closely resemble the real data, indicating a strong fidelity to the original distribution, with some variability—albeit confined to specific regions where the shape of the violin plot shows noticeable differences. On the other hand, GPT's data appears less similar and exhibits greater diversity, as indicated by the shape of its violin plot, a finding that quantitative metrics can corroborate. The graphs also reinforce our previous observations: LLM-generated data tends to lack diversity, which is understandable given the model's limited access to only three samples as examples of the target distribution.

The real and synthetic data medians are quite similar for all three groups. This suggests that the synthetic data generators have captured the real data's central tendency. The IQR, which represents the spread of the middle 50% of the data, also seems to be comparable between real and synthetic data. This indicates that the synthetic data generators have maintained the overall variability of the real data.

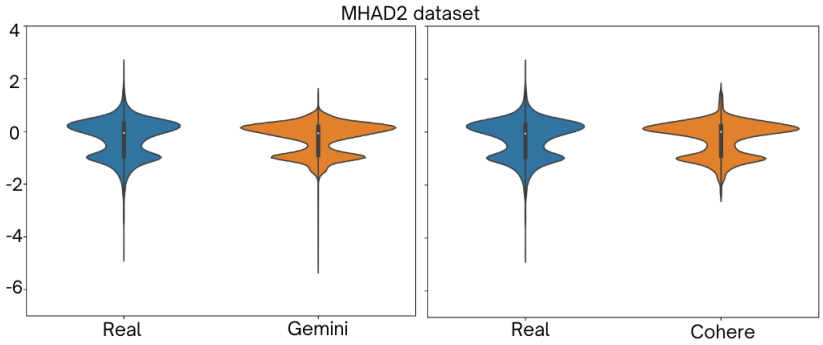

Figure 7: Comparative analysis of synthetic (orange) and real (blue) distributions using the MHAD2 Dataset.

However, the violin plots reveal some differences in the distribution shapes between real and synthetic data. For instance, the real data for Gemini appears to have a slightly wider distribution than the synthetic data, while Cohere's synthetic data seems to have a slightly narrower distribution than the real data.

Overall, the violin plots suggest that the synthetic data generated by Gemini and Cohere is reasonably similar to the real data in terms of central tendency and variability. However, there are some differences in the distribution shapes and the presence of outliers.

A.6 GPT-4 vs. Traditional models

In this section, we provide the full results of the comparison between the best-performing LLM and traditional generative models.

A.6.1 TRTS and TSTR protocols

In TSTR, we trained the classifier in each fold using 30 synthetic samples (for each class), and the baseline for comparison was also trained using only 30 samples (chosen randomly). This was the maximum amount of data we could collect from GPT-4. In TRTS, we used the number of synthetic samples as the number of real test samples. Table 3 presents the performance of all methods since the number of samples in the testing set is fewer than 30.

GPT-4 is not the best-performing model on TSTR at the MHAD2 and MHEALTH datasets. Still, it enhances the baseline by approximately 10% on the MHAD2 TSTR evaluation and almost 28% on the MHEALTH TSTR evaluation. It significantly outperforms SSSD (with an improvement of around 14%) and TTS-GAN (with a gain of roughly 8%) on MHAD2 and TimeGAN (with a gain of roughly 32%), DGAN (more than 40%), and TTS-GAN (more than 30%) on MHEALTH. Regarding TRTS, it does not exceed the baseline but still proves superior to most models in the TSTR protocol.

Table 3: Results of each model under the evaluated protocols.

Eval.	Dataset	Model	Add	Accuracy	Eval.	Dataset	Model	Add	Accuracy
MIXED	MHAD2	#Baseline	-	69.55 ± 2.42	TSTR	MHAD2	#Baseline	-	24.46 ±2.07
		GPT-4	150	70.80 ± 2.98			TLCGAN	-	**47.08 ±4.50**
		DGAN	100	70.98 ± 1.88			DGAN	-	45.63±6.16
		SSSD	100	68.93 ± 2.65			TimeGAN	-	45.52±2.83
		TLCGAN	150	**72.23 ± 3.14**			SSSD	-	20.73±2.73
		TTS-GAN	100	70.45 ± 2.91			TTS-GAN	-	26.15±5.30
		TimeGAN	100	70.63 ± 3.11			GPT-4	-	34.38±3.46
	MHEALTH	#Baseline	-	94.74 ± 0.87		MHEALTH	#Baseline	-	24.33±3.17
		GPT-4	300	**96.13 ± 1.29**			TLCGAN	-	55.25±3.37
		DGAN	150	95.34 ± 0.75			DGAN	-	11.61±2.73
		SSSD	300	96.12 ± 0.57			TimeGAN	-	20.80±2.56
		TLCGAN	300	95.83 ± 0.33			SSSD	-	**63.47 ±1.66**
		TTS-GAN	300	96.04 ± 0.91			TTS-GAN	-	12.94 ±3.30
		TimeGAN	300	95.83 ± 0.94			GPT-4	-	52.61 ±2.13

Eval.	Dataset	Model	Add	Accuracy	Recall	F1
TRTS	MHAD2	#Baseline	-	69.55±2.42	68.40±2.48	68.53±2.41
		TLCGAN	-	**69.17 ±4.59**	**68.20±4.81**	**68.17±4.94**
		DGAN	-	*55.83*±3.54	*54.81*±3.74	*53.96*±4.17
		TimeGAN	-	*68.33*±6.62	*65.96*±7.27	*60.96*±8.05
		SSSD	-	*17.92*±0.94	*19.38*±1.16	*10.58*±2.25
		TTS-GAN	-	20.73±3.55	23.29±3.98	12.21±3.78
		GPT-4	-	36.56±4.82	36.06±4.47	27.66±5.63
	MHEALTH	#Baseline	-	98.12±1.75	96.82±3.71	96.66±3.76
		TLCGAN	-	**78.26 ±3.77**	**78.28 ±3.78**	**74.79 ±4.53**
		DGAN	-	70.03±2.73	70.09±2.73	68.44±3.00
		TimeGAN	-	25.99±4.21	25.83±4.23	18.66±3.53
		SSSD	-	70.03 ±2.73	70.10±2.73	68.44±3.00
		TTS-GAN	-	8.92 ±0.49	8.80 ±0.48	1.88±0.47
		GPT-4	-	59.43 ±2.12	59.44 2.17	50.10 ±2.93

In the "#Baseline" model, we adopt the approach of Training the model on Real data and Testing on Real data (TRTR). For both the MHAD2 and MHEALTH datasets, we utilize 30 samples from each class for training. However, in the case of the MHEALTH dataset, we restrict our testing to only 30 samples. The term 'add' refers to the number of synthetic examples incorporated into the existing training set. Due to space constraints in this paper, we only present the best results from the MIXED protocol. Nevertheless, complete results will be made available in the supplementary materials accompanying this paper for a comprehensive view.

TLCGAN showed the most exceptional performance, exceeding the baseline in both scenarios in MHAD2. In MHEALTH, it was the second best on TSTR but the best model on TRTS, even without outperforming the baseline.

A.6.2 Mixed

We systematically combined synthetic samples with the original training samples and evaluated the performance of the Dclassifier [26] when trained on this mixed data. Given that the generated dataset is larger than the data to be incorporated, we randomly selected the samples for inclusion. Figures 3 and 3 summarize this procedure. Due to being the best LLM in the tasks evaluated in the paper, our analysis will concentrate on the data generated by GPT-4 (our primary focus) and TLCGAN (the baseline).

Upon introducing synthetic samples to the MHAD2 training set (refer to figure 3), GPT-4 outperforms the baseline in two of the three demonstrated scenarios. In these scenarios, considering accuracy, it surpasses both TimeGAN and SSSD. However, when considering the arithmetic mean, TTS-GAN also falls within this category. TLCGAN demonstrated the highest performance, with nearly 3% more accuracy and a higher average.

With respect to the MHEALTH dataset, GPT-4 outperforms the baseline in all scenarios. Within these, it surpasses TLCGAN in two out of three scenarios, thus becoming the top-performing model. This holds true both in terms of average and accuracy, implying that GPT's performance is notably high since the average was maintained and the datasets are unbalanced.

The performance of both models varies according to the amount of data added, and depending on the dataset, it may negatively affect model performance as more synthetic data is added. Too many synthetic samples might introduce noise or overfitting, diminishing the benefits of augmentation (see table 3 for the best results in this evaluation). Even though it is not the top-performing model

in all scenarios, GPT-4 is competitive compared to other models. Like the others, augmenting data improves the model's performance to a certain extent, suggesting a consistency in the results produced when employing this model.

A.7 Limitations and Future works

Our current work primarily focuses on three LLMs. While GPT-4 demonstrated strong performance in synthetic data generation, the study would benefit from evaluating additional LLM models, including earlier versions and other state-of-the-art generative models. Also, fine-tuning LLM models on domain-specific data could potentially enhance their performance and generate more accurate and representative synthetic data. Future research should explore the impact of fine-tuning on synthetic data quality and in the classifier's performance.