SenseCF: LLM-Prompted Counterfactuals for Intervention and Sensor Data Augmentation

Shovito Barua Soumma[1,2], Asiful Arefeen[1,2], Stephanie M. Carpenter[1], Melanie Hingle[3] and Hassan Ghasemzadeh[1]

***Abstract*—Counterfactual explanations (CFs) offer human-centric insights into machine learning predictions by highlighting minimal changes required to alter an outcome. Therefore, CFs can be used as (i) interventions for abnormality prevention and (ii) augmented data for training robust models. In this work, we explore large language models (LLMs), specifically GPT-4o-mini, for generating CFs in a zero-shot and three-shot setting. We evaluate our approach on two datasets: the AI-Readi flagship dataset for stress prediction and a public dataset for heart disease detection. Compared to traditional methods such as DiCE, CFNOW, and NICE, our few-shot LLM-based approach achieves high plausibility (up to 99%), strong validity (up to 0.99), and competitive sparsity. Moreover, using LLM-generated CFs as augmented samples improves downstream classifier performance (an average accuracy gain of 5%), especially in low-data regimes. This demonstrates the potential of prompt-based generative techniques to enhance explainability and robustness in clinical and physiological prediction tasks. Code base: github.com/shovito66/SenseCF.**

***Index Terms*—Counterfactual explanations, Diabetes, Digital health, Explainable AI, Metabolic health, Wearable sensors, LLM**

I. Introduction and Related Work

Accurate and interpretable predictions from machine learning (ML) models are increasingly vital in healthcare applications such as disease risk forecasting and sleep efficiency estimation using physiological and sensor data. While these models excel at outcome prediction, they often fall short in guiding actionable interventions to reverse adverse outcomes—especially in black-box settings.

Counterfactual explanations (CFs) offer a powerful solution by revealing the minimal changes needed to flip a model's prediction. Traditional CFE methods like DiCE, CFNOW, and NICE rely on optimization procedures that often require model internals or gradient access, limiting their real-world applicability and struggling with categorical plausibility. In contrast, large language models (LLMs) provide a promising alternative: leveraging zero- and few-shot prompting, they can generate realistic, coherent counterfactuals using only input-output context [1]. This paradigm not only removes the dependence on gradients or model access but also opens the door for scalable, interpretable explanations across diverse datasets.

Recent work highlights LLMs' innate counterfactual reasoning capabilities without fine-tuning [2], [3], yet their use in structured, multimodal health data remains underexplored. Importantly, CFs not only enhance interpretability but can also serve as data augmenters, introducing label-flipping samples to strengthen models—particularly in imbalanced medical datasets. For data augmentation, CFs enhance robustness by introducing label-flipping variations while preserving data distributions. Optimization methods show promise in medical or low-data contexts but struggle with categorical coherence—a gap addressed by LLMs' semantic understanding [4].

However, several critical gaps persist in the current literature: first, the effectiveness of LLM-based CFs has not been comprehensively evaluated on multimodal clinical datasets; second, standardized evaluation metrics comparing optimization-based and generative approaches remain limited; third, CFs' potential as data augmenters in healthcare scenarios remains underexplored.

To address this gap, we introduce a systematic evaluation of zero- and few-shot LLM-generated counterfactuals across two real-world clinical datasets. Our contributions extend beyond existing LLM-focused studies that primarily evaluate natural language processing (NLP) tasks, providing a rigorous and quantitative comparison in multimodal clinical settings [2], [3]. We benchmark their plausibility, diversity, and impact on model performance against state-of-the-art baselines. To the best of our knowledge, this is the first study to explore LLMs as counterfactual generators for both explanation and augmentation in sensor-driven health contexts, moving toward AI systems that can inform not just prediction, but intervention.

II. Methods

In this section, we detail our approach for generating CFs using GPT-4o, structured as illustrated in Fig 1 and Fig 2. Our methodology aims at: (1) producing actionable counterfactual (CF) interventions by reversing the predictions of trained ML models, and (2) leveraging these CFs as augmented training data to enhance model performance, specifically addressing potential data imbalance. We represent our input data as a set

[1]College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA. Email: {shovito, aarefeen, stephanie.m.carpenter, hghasemz}@asu.edu.

[2]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ 85281, USA.

[3]School of Nutritional Sciences and Wellness, University of Arizona, Tucson, AZ, USA. Email: hinglem@arizona.edu

This work was supported in part by the National Science Foundation (NSF) under grant IIS-2402650 and A. Arefeen is supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (NIH) under Award T32DK137525. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF and NIH.

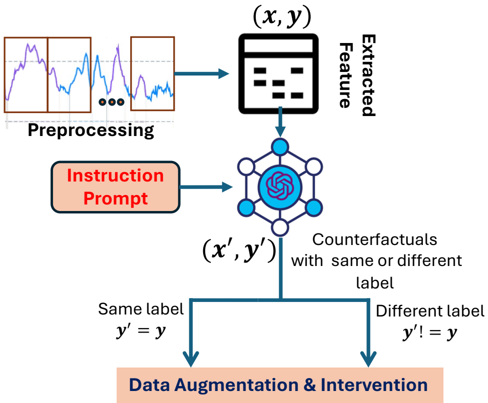


Fig. 1: Counterfactual generation using LLMs from sensor-derived features.

of tuples (x_i, y_i) , where $x_i \in X$ is a feature vector representing either clinical or physiological data and $y_i \in \{0, 1\}$ denotes the ground truth label. The trained predictive model $f(\cdot)$ outputs predictions $\hat{y_i} = f(x_i)$. Our preprocessing stage transforms raw data into structured, tabular feature vectors suitable for prompting the LLM.

A. Counterfactual Generation

We used GPT-4o as an off-the-shelf counterfactual generator using specifically crafted instruction prompts in both zero-shot and few-shot settings. Formally, given a feature vector x_i and the prediction $\hat{y_i}$, GPT-4o generates a modified vector $x_i^{'}$, where the model's prediction changes from $\hat{y_i}$ to a desired opposite outcome $y_i \neq \hat{y_i}$. We also explicitly constrain the LLM from altering immutable or clinically fixed features (e.g., age, sex, or medication type), ensuring that generated counterfactuals remain actionable and plausible within domain constraints. The generation of CFs can be described as:

$$x'_i = \text{LLM}(x_i, \text{prompt}), \quad \text{subject to } f(x'_i) \neq f(x_i)$$

The instructional prompt explicitly constrains GPT-4o to minimally alter feature values to achieve a realistic and actionable counterfactual, ensuring the plausibility and feasibility of generated CFs.

B. Intervention and Data Augmentation

Generated CFs serve dual purposes: (1) acting as plausible intervention points, and (2) augmenting training data. Specifically, CFs that successfully reverse predictions are included in the training dataset to enhance the robustness of the predictive model $f(\cdot)$. Formally, the augmented training set X_{aug} is defined as:

$$X_{\text{aug}} = X \cup \{(x'_i, y'_i) \mid f(x'_i) \neq f(x_i)\}$$

Augmentation with CFs is particularly effective in handling imbalanced datasets, introducing valuable variability and balanced class representation, thus improving predictive model performance.

III. Experiment

A. Data

1. The **AI-READI** data, part of an NIH-funded initiative, combines rich data sources—like vitals, ECG signals, continuous glucose monitoring (CGM), physical activity, and eye imaging—collected from 1,067 participants, including both diabetic and non-diabetic. In this analysis, we used twelve features to design counterfactual interventions for transitioning individuals from high stress ($y = 1$) to moderate stress levels ($y = 0$). Of the twelve features, we considered age, sex, and medication as immutable features; the rest are either raw or derived features from the sleep and daily glucose information collected using sensors (Garmin Vivosmart 5 and Dexcom G6).

2. Modified **Heart Disease** dataset from [5] has 918 instances. We used four categorical and five continuous features to map whether an individual will have a heart disease ($y = 1$) or not ($y = 0$) and design interventions based on five features. A subset of the features, including fasting blood sugar level, resting blood pressure, resting ECG, and exercise-induced S-T depression, are derived using wearable sensors.

B. Baselines

We have identified the following techniques to compare against SenseCF.

DiCE [6] identifies a set of CFs by optimizing for proximity, diversity and sparsity. ***CFNOW*** [7] searches an optimal point close to the factual point where the classification differs from the original. CFNOW performs greedy optimization for metrics like speed, coverage, distance, and sparsity. ***NICE*** [4] CFs are not necessarily adversarial data points but nearby instances in the data that reflect the desired outcome.

C. Evaluation Metrics

We assess the CFs using some standard metrics found in the literature:

Validity assesses whether the produced CFs genuinely belong to the desired class. High validity indicates the technique's effectiveness in generating valid CF examples.

$$\mathit{validity} = \frac{\#|f(X_T^*) \neq f(X_T)|}{\|CF\|}$$

Distance between the CF and the factual sample is calculated from the L_2 normalized distance of the continuous features and the hamming distance of the categorical features.

Sparsity is the average number of feature changes per CF. A low sparsity ensures better user understanding of the CFs.

$$\mathit{sparsity} = \frac{\sum_{X_T^* \in CF} \sum_{i=1}^{d} \mathbb{1}(x_T^{*i} \neq x_T^i)}{\|CF\|} \tag{1}$$

Plausibility quantifies the fraction of explanations that fall within the feature ranges derived from the data-

$$\mathit{plausibility} = \frac{\sum_{X_T^* \in CF} \mathbb{1}(\text{dist}(X_T^*) \subseteq \text{dist}(X))}{\|CF\|}$$

where, $\text{dist}(X_T^*)$ and $\text{dist}(X)$ represent the distribution of feature values in the CF instances X_T^* and in the training data, respectively. $\|CF\|$ is the total number of CF instances.

IV. Results

Our evaluation highlights the dual role of LLM-generated CFs—as highly plausible interventions and as impactful data augmenters for robust model training in digital health contexts.

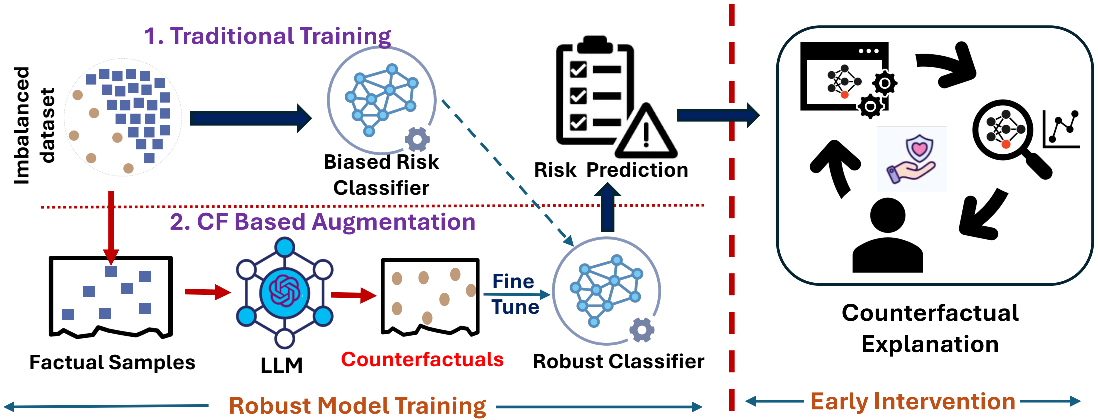
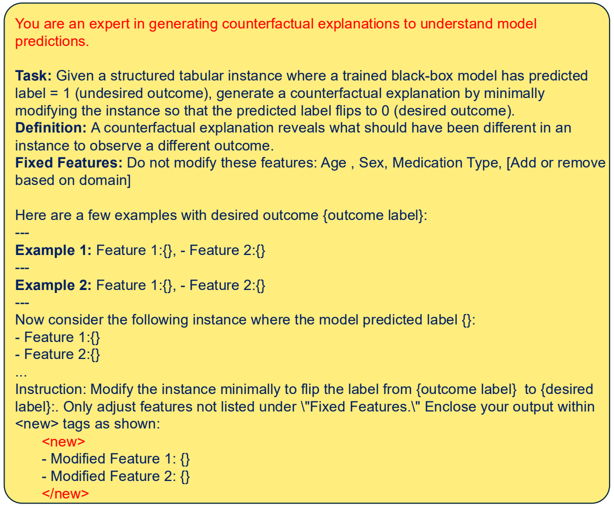


Fig. 2: SenseCF pipeline: LLM-generated counterfactuals are used both for augmenting imbalanced training data (left) and for model interpretability (right).



You are an expert in generating counterfactual explanations to understand model predictions.

Task: Given a structured tabular instance where a trained black-box model has predicted label = 1 (undesired outcome), generate a counterfactual explanation by minimally modifying the instance so that the predicted label flips to 0 (desired outcome).
Definition: A counterfactual explanation reveals what should have been different in an instance to observe a different outcome.
Fixed Features: Do not modify these features: Age , Sex, Medication Type, [Add or remove based on domain]

Here are a few examples with desired outcome {outcome label}:

Example 1: Feature 1:{}, - Feature 2:{}

Example 2: Feature 1:{}, - Feature 2:{}

Now consider the following instance where the model predicted label {}:
- Feature 1:{}
- Feature 2:{}
...
Instruction: Modify the instance minimally to flip the label from {outcome label} to {desired label}:. Only adjust features not listed under \"Fixed Features.\" Enclose your output within <new> tags as shown:
<new>
- Modified Feature 1: {}
- Modified Feature 2: {}
</new>

Fig. 3: Prompt template for counterfactual generation.

Condition	Intervention
A 81-year-old patient labeled as "stressed" showed low deep sleep (30.1%), moderate REM (15.4%), high blood glucose (210.8 mg/dL), and limited activity (5.95 steps). Stress level was high (85.25), with poor glucose control (TIR: 12.5%, 1 hyper event).	The LLM suggests ***increasing deep sleep*** to ↑35% and ***REM sleep*** to ↑20%, which could help reduce physiological and emotional stress. It also recommends ***lowering blood glucose*** from 210.8 to ↓180 mg/dL, aligning with better metabolic control. These changes reflect clinically actionable strategies such as sleep hygiene improvement and tighter glucose management.

TABLE I: Example of LLM-suggested counterfactual intervention for a high-stress patient

A. Intervention

A representative counterfactual intervention generated by the SenseCF is illustrated in Table I. As shown in Tables II and III, CFs generated by GPT-4o using zero-shot and few-shot prompting achieve consistently high validity scores (up to 0.99), while maintaining competitive sparsity and distance. Notably, few-shot prompting improves realism and interpretability (e.g., 99% plausibility on AI-READI, 98.1% on HR), underscoring the LLM's semantic alignment with the target domain. Unlike optimization-based methods (e.g., DICE, CFNOW), which rely on access to model internals, our LLM-based approach generates CFs in a model-agnostic fashion while remaining interpretable and actionable.

TABLE II: Evaluating the CFs on AI-READI Dataset.

Method	validity ↑	distance ↓	sparsity ↓	plausibility ↑
Zero-shot	0.91	1.1	3.6	85
3-shot	0.99	1.2	4.4	99
DiCE	0.67	0.2	2.27	100
NICE	0.85	0.1	2.9	100
CFNOW	0.44	0.02	1.12	33

TABLE III: Evaluating the CFs on Heart Disease Dataset.

Method	validity ↑	distance ↓	sparsity ↓	plausibility ↑
Zero-shot	0.88	4.2	7.9	97.1
3-shot	0.97	2.6	5.2	98.1
DiCE	0.60	0.2	2.3	100
NICE	0.85	0.07	2.5	100
CFNOW	0.44	0.02	1.13	33

Figure 4 shows the diversity of the CFs from different methods across all the mutable features. SenseCF-generated CFs exhibit the least diversity for all features, except that the 3-shot variant shows higher diversity in Average Step counts.

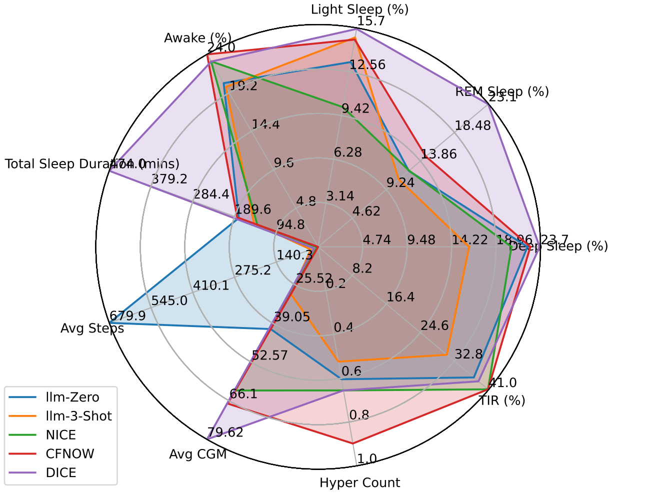


Fig. 4: Feature diversity in the generated CFs for AI-Readi data. Avg: Average, Hyper: No of hyperglycemia

B. Impact as Data Augmenters

Beyond explanation, Tables IV and V demonstrate that LLM-generated CFs significantly enhance model performance when used for data augmentation. Across classifiers—including Random Forest (RF), SVC, XGB, and NN—we observe performance improvements in terms of accuracy, precision, recall, and AUC. While the original models already perform well on balanced datasets (e.g., HR), the gains are especially relevant in realistic healthcare scenarios with data imbalance, as reflected in AI-READI outcomes. For instance, RF performance on AI-READI improves from 0.74 to 0.80 in accuracy with counterfactual augmentation.

The few-shot LLM prompts, in particular, lead to consistently high classification scores that match or exceed those obtained from traditional SOTA augmentation techniques like DICE, NICE, and CFNOW. This means that LLMs capture not only the structure of the data but also the semantic boundaries between classes—resulting in augmentations that are both label-flipping and label-preserving depending on the use case.

TABLE IV: Performance Impact of LLM-Generated and SOTA Counterfactuals on AI-Readi Classifiers

Model	Method	ACC	PRE	REC	F1	AUC
RF	Zero	0.79	0.79	0.77	0.78	0.86
	Few	0.793	0.79	0.78	0.79	0.87
	Both	0.8	0.8	0.79	0.79	0.86
	DICE	0.788	0.79	0.76	0.78	0.87
	NICE	0.787	0.78	0.77	0.78	0.86
	CFNOW	0.79	0.78	0.78	0.78	0.86
	×	0.74	0.72	0.77	0.74	0.82
SVC	Zero	0.63	0.63	0.57	0.6	0.67
	Few	0.65	0.64	0.61	0.62	0.7
	Both	0.65	0.64	0.6	0.62	0.7
	DICE	0.65	0.64	0.6	0.62	0.7
	NICE	0.65	0.64	0.6	0.62	0.7
	CFNOW	0.65	0.64	0.61	0.62	0.7
	×	0.63	0.63	0.58	0.6	0.67
XGB	Zero	0.77	0.77	0.75	0.76	0.85
	Few	0.78	0.78	0.75	0.76	0.86
	Both	0.78	0.78	0.77	0.77	0.86
	DICE	0.78	0.79	0.75	0.77	0.86
	NICE	0.78	0.79	0.75	0.77	0.85
	CFNOW	0.76	0.77	0.74	0.76	0.85
	×	0.74	0.72	0.77	0.74	0.81
NN	Zero	0.64	0.63	0.6	0.62	0.67
	Few	0.642	0.64	0.64	0.63	0.69
	Both	0.633	0.62	0.62	0.62	0.67
	DICE	0.635	0.64	0.56	0.6	0.68
	CFNOW	0.64	0.64	0.57	0.6	0.69
	NICE	0.638	0.63	0.63	0.63	0.69
	×	0.63	0.6	0.57	0.58	0.67

V. Conclusion & Future Work

In this work, we introduce a novel framework for generating CFs using large language models (LLMs), with a focus on structured sensor-derived datasets in health and physiological monitoring. Our method leverages zero-shot and few-shot prompting with GPT-4o to generate semantically valid, plausible CFs that flip model predictions while respecting domain-specific constraints such as immutable features.

Through extensive evaluation on two real-world datasets, we show that LLM-generated CFs are not only effective for model interpretability and intervention design but also significantly enhance model performance when used for data augmentation, particularly in contexts affected by data imbalance. The method is model-agnostic, requires no gradient access, and generalizes across clinical and wearable domains. One limitation is the higher feature distance compared to baselines; future work will focus on refining prompts to improve counterfactual compactness and quality.

TABLE V: Performance Impact of LLM and SOTA Counterfactual Augmentation on Heart Disease Classification Models

Model	Method	ACC	PRE	REC	F1	AUC
RF	Zero	0.985	1	0.97	0.99	1
	Few	0.985	1	0.97	0.99	1
	Both	0.985	1	0.97	0.99	1
	DICE	0.985	1	0.97	0.99	1
	NICE	0.985	1	0.97	0.99	1
	CFNOW	0.985	1	0.97	0.99	1
	×	0.9854	1	0.97	0.99	1
SVC	Zero	0.79	0.77	0.83	0.8	0.88
	Few	0.82	0.78	0.89	0.83	0.87
	Both	0.8	0.78	0.83	0.8	0.88
	DICE	0.81	0.76	0.9	0.83	0.87
	NICE	0.82	0.76	0.93	0.84	0.87
	CFNOW	0.80	0.76	0.87	0.81	0.87
	×	0.82	0.76	0.92	0.83	0.87
XGB	Zero	0.985	1	0.97	0.99	0.99
	Few	0.985	1	0.97	0.99	0.99
	Both	0.985	1	0.97	0.99	0.99
	DICE	0.91	0.98	0.83	0.9	0.96
	NICE	0.91	0.92	0.87	0.9	0.95
	CFNOW	0.91	0.99	0.83	0.9	0.97
	×	0.98	0.98	0.97	0.98	0.98
NN	Zero	0.81	0.78	0.84	0.81	0.88
	Few	0.78	0.75	0.84	0.79	0.88
	Both	0.8	0.79	0.82	0.81	0.88
	DICE	0.77	0.75	0.83	0.78	0.87
	CFNOW	0.76	0.74	0.82	0.77	0.87
	NICE	0.77	0.74	0.84	0.79	0.87
	×	0.78	0.74	0.85	0.79	0.87

To the best of our knowledge, this is the first systematic exploration of LLM-based CFs in sensor-driven data under both zero- and few-shot settings. We believe this opens a promising direction for integrating generative AI into trustworthy, intervention-oriented healthcare ML pipelines.

References

[1] B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, vol. 1, p. 3, 2020.

[2] A. Bhattacharjee, R. Moraffah, J. Garland, and H. Liu, " Zero-shot LLM-guided Counterfactual Generation: A Case Study on NLP Model Evaluation ," in *2024 IEEE International Conference on Big Data (BigData)*. Los Alamitos, CA, USA: IEEE Computer Society, Dec. 2024, pp. 1243–1248. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/BigData62323.2024.10825537

[3] Y. Li, M. Xu, X. Miao, S. Zhou, and T. Qian, "Prompting large language models for counterfactual generation: An empirical study," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 13 201–13 221. [Online]. Available: https://aclanthology.org/2024.lrec-main.1156/

[4] D. Brughmans and D. Martens, "Nice: an algorithm for nearest instance counterfactual explanations," *Data Mining and Knowledge Discovery*, pp. 1–39, 2021.

[5] Fedesoriano, "Heart Failure Prediction Dataset," https://www.kaggle.com/fedesoriano/heart-failure-prediction, September 2021, retrieved April 2024.

[6] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 607–617.

[7] R. M. B. de Oliveira, K. Sörensen, and D. Martens, "A model-agnostic and data-independent tabu search algorithm to generate counterfactuals for tabular, image, and text data," *European Journal of Operational Research*, 2023.