Interpreting Dataset Shift in Clinical Notes

Shariar Vaez-Ghaemi* SHARIAR.VAEZGHAEMI@DUKE.EDU
Duke University

Furong Jia* FLORA.JIA@DUKE.EDU
Duke University

Monica Agrawal MONICA.AGRAWAL@DUKE.EDU
Duke University

Abstract

Distribution shift can lead to degradation in the performance of machine learning models. This concern is particularly salient in medicine, in which several forces can lead to shifts in Electronic Health Record (EHR) data. Distribution shift in the text domain is vastly understudied, but increasingly important, given the widespread integration of large language models into clinical workflows. Identifying the existence of a shift is necessary but insufficient; actionability often requires understanding the nature of the shift. To address this challenge, we establish an extensible benchmark suite that induces synthetic distribution shifts using real clinical notes and develop two methods to assess generated shift explanations. We further introduce **SIReNs**, a general-domain end-to-end approach that explains distributional differences between two datasets by selecting representative notes from each. The SIReNs method was evaluated on both binary and continuous feature shifts, and the results show that it recovers salient binary shifts well, but struggles with more subtle shifts. A substantial gap remains to a ground-truth oracle for continuous shifts, suggesting room for improvement in future methods.

Keywords: clinical notes, distribution shift, shift explanation, benchmark, interpretability, large language models

Data and Code Availability This paper leverages the MIMIC-IV dataset (Johnson et al., 2020, 2023). The code implemented for this paper can be found here: Github Repo

Institutional Review Board (IRB) This research does not require IRB approval.

*

1. Introduction

Machine learning models frequently encounter distribution shift, in which test-time data significantly differs from training data, often leading to eroded model performance (Koh et al., 2021). Distribution shift poses a particularly acute threat in medicine, as medical knowledge, treatment protocols, documentation styles, and patient demographics differ across institutions and evolve over time (Sohn et al., 2018; Subbaswamy and Saria, 2020; Finlayson et al., 2021; Guo et al., 2023). Prior work on distribution shift detection and mitigation has largely focused on images and structured data, with comparatively less attention on the text domain (Malinin et al., 2021; Yao et al., 2022; Gardner et al., 2023; Taori et al., 2020).

However, with the rapid adoption of NLP in healthcare, textual shift could affect a wide range of deployed NLP tasks, from predictive (e.g., readmission prediction) to extractive (e.g., phenotyping) and generative tasks (e.g., summarization, patient question answering) (Tai-Seale et al., 2024; Van Veen et al., 2023; Jiang et al., 2023). While the performance of predictive models can often be monitored based on future ground truth, the evaluation of extractive and generative algorithms in medicine often requires manual labor-intensive effort, making proactive shift identification essential. Given the high stakes of medicine, failures to generalize could have serious downstream consequences. Further motivations include understanding (i) how synthetic clinical notes diverge from their real-world counterparts and (ii) how ambient documentation tools may change the content of clinical notes (Kweon et al., 2023; Li et al., 2021; Tierney et al., 2024).

The ability to detect and interpret shifts in clinical text inputs is essential to anticipate model failures and guide remediation before or during deploy-

© 2025 S. Vaez-Ghaemi, F. Jia & M. Agrawal.

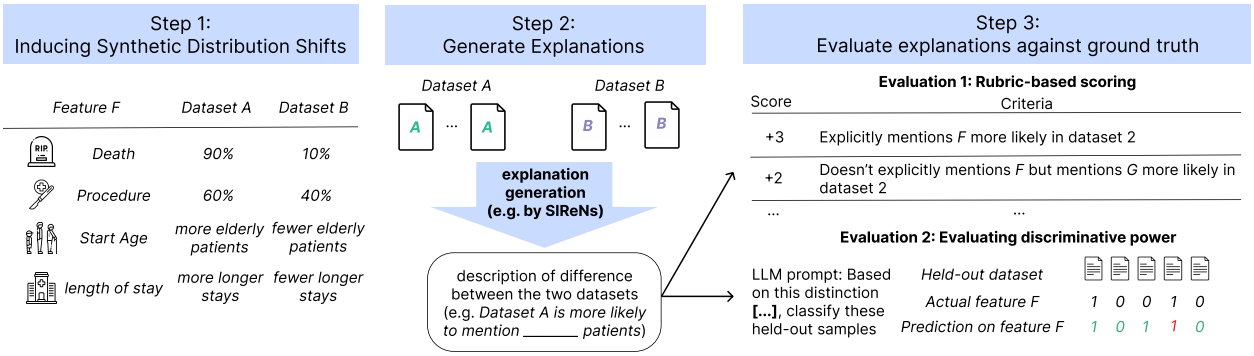

Figure 1: An overview of our benchmark pipeline: 1) Constructing text datasets that contain synthetic distribution shifts based on a feature, 2) Generating explanations for the shift using various methods, including SIReNs, 3) Evaluating explanations against ground truth, using both rubric-based scores and the discriminative power of the explained feature against the ground truth feature we shifted on

ment. Because a distribution shift does not necessarily imply degradation in task performance, explaining the nature of the shift is critical for identifying those shifts that necessitate intervention, e.g., new in-context examples, further fine-tuning, or data augmentation. **Therefore, this work aims to introduce and address dataset shift interpretation in clinical notes.** Since there are no existing ground truth datasets in this space, we establish the existence of these shifts, design benchmark software to synthetically induce interpretable shifts in the MIMIC-IV dataset (Johnson et al., 2023), and introduce a principled approach for tackling the problem. Specifically, we make the following contributions:

1. **Demonstration of shift in MIMIC notes** (Section 3): We demonstrate real-world shifts in MIMIC-IV's Brief Hospital Course summaries.
2. **Benchmark software suite for synthetically inducing shifts in MIMIC** (Section 4): We introduce an extensible, end-to-end benchmark suite (Figure 1) that induces controlled shifts of user-specified magnitude in user-selected structured variables, supporting both binary (prevalence) and continuous (distributional) features. It further evaluates explanations via a rubric-based scoring and an accuracy score on the discriminative efficacy in classifying the data source.
3. **SIReNs: a lightweight paradigm for explaining distribution shift in text** (Section 5): We introduce **SIReNs**, a general-domain framework that explains distributional differences between two datasets by selecting a small set of representative notes, using interchangeable selection modules (multiple instantiations provided), and prompting an LLM for a natural language explanation. While SIReNs is domain-agnostic, in this paper, we evaluate its effiacy on clinical notes displaying both binary and continuous shifts. SIReNs recovers salient binary shifts but struggles with continuous attributes and broad concepts, highlighting the challenge of aggregating weak textual signals.

2. Related Work

2.1. Detection of Distribution Shift in Text

Research on temporal semantic change focuses on how meaning and usage of individual words evolve over time, from time-sliced distributional methods with PPMI/SVD and word2vec (Hamilton et al., 2016), to contextualized, usage-sensitive analyses with BERT-style representations (Giulianelli et al., 2020), and community benchmarks (Schlechtweg et al., 2020). In parallel, early methods for detecting shifts between text distributions have been largely motivated by the task of distinguishing human-generated text from machine-generated text. Earlier methods employed property-specific statistics to measure individual text characteristics like repetition (Holtzman et al., 2019; Welleck et al., 2019) or sentence-level verifiability against Wikipedia. (Massarelli et al., 2019). MAUVE compares the distribution of model-generated text to human text using divergence frontiers between language model probability distributions (Pillutla et al., 2021). Other methods similarly measure the divergence between the distributions of cluster assignments of text embeddings (Gupta et al., 2023).

While these methods offer increasingly adaptable ways to detect and measure shifts between text distributions, they do not explain the nature of such shifts in a human-interpretable way that points to underlying causes.

2.2. Distribution Shift in Clinical Text

Research on distribution shift in clinical notes remains limited, and prior work has focused on surface-level or predefined structural metrics rather than open-ended semantic characterization. For example, Sohn et al. (2018) examined frequencies of predetermined asthma keywords across two hospitals and the corresponding degradation in an NLP asthma model transferred between the two sites. Others have found different rates of pre-defined sentiment words in clinical notes of patients across races (Bilotta et al., 2024). Longitudinally, Rahimian et al. found oncology notes became longer and more readable after the implementation of OpenNotes. Similarly, Peterson and Liu found that the embeddings learned from old vs. new clinical notes show semantic shift (e.g. 'portal' used to be associated with portal vein and shifted to largely refer to patient portals). While existing work has typically been confined to predefined categories, real-wirld semantic shifts can be localized, compositional, and multi-scale.

2.3. Explanation Generation

Rationale extraction work focuses on instance-level prediction explanation by selecting minimal, sufficient spans that preserve a model's label. Early joint selectors learn contiguous rationales without gold label supervision for the rationales (Lei et al., 2016). Extensions include incorporating more diverse objectives towards robustness, faithfulness, and plausibility (Li et al., 2022; Madani and Minervini, 2023), or unified training frameworks (Chan et al., 2022). Despite differing optimization goals, these approaches all assume labeled examples and relatively short inputs, rather than unlabeled, corpus-level comparison.

Closer to our goal are methods that verbalize differences or characterize datasets globally. VisDiff generates natural language descriptions for differences between two image sets under an assumption that all items in Group B share a characteristic not found in Group A (Dunlap et al., 2024). In particular, VisDiff samples captions generates caption for each image, has an LLM propose differences, and uses a reranker to choose the best explanation. Similarly, Zhong et al. (2024) has an LLM suggest human-legible clusters and cluster assignments, and then sees how predictive cluster assignment is of the label. However, both works deal with inputs that were less multifaceted (e.g., distinguishing articles of politicians from articles of athletes).

3. Data

3.1. Dataset Overview

We use the Brief Hospital Course (BHC) summaries from MIMIC-IV (Johnson et al., 2024), a large database of ICU admissions. While exact admission dates are de-identified, a three-year window for the start year of each note allows for approximate analysis. The dataset also spans the pivotal October 2015 transition on International Classification of Diseases (ICD) codes from ICD-9 to ICD-10, a known source of distributional shift.

3.2. Examples of Real-World Dataset Shift

To motivate our work, we first demonstrate that real-world, impactful shifts are present within this corpus. We identified two types of temporal shifts in BHC notes from 2008 to 2020.

Structural Shift BHC notes underwent marked structural change (2008–2020): mean length increased by over 34%. This was partially driven by the widespread adoption of standardized templates. A key example is the structured "Issues" section, found in $\leq 1\%$ of notes in 2008 and over 50% in 2020. A detailed analysis and visualization are provided in Appendix B.

Semantic Shift in the Depression Cohort More subtly, we found a significant semantic shift related to the ICD coding transition on the single diagnosis of "Major Depressive (Affective) Disorder, Single Episode, Unspecified," the most common billing code for depression in the dataset. Figure 2 reveals that note embeddings with this ICD-9 depression code have much more clustered representations, compared to notes with the ICD-10 depression code, which resemble a random set of notes. Quantitatively, we compute pairwise Maximum Mean Discrepancies (MMDs) between Gatortron embeddings of notes in each of the four groups (Yang et al., 2022). The MMD between ICD-9 depression and each of the other three groups is larger than any other pairwise MMD among these groups, with more results shown in Table 9

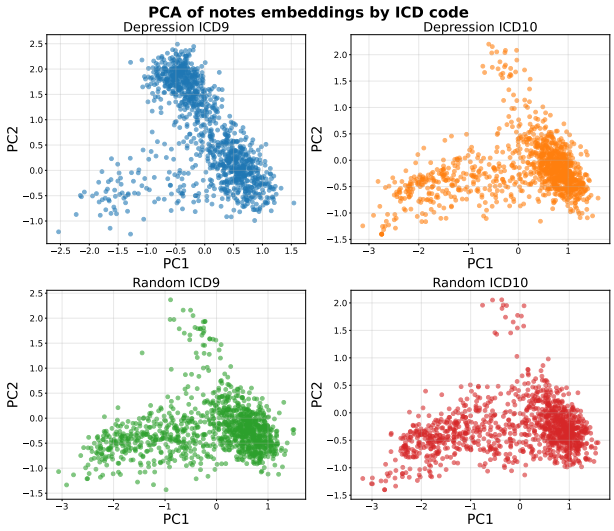

Figure 2: ICD-9 depression notes form a tight cluster; ICD-10 depression notes shows a broader semantic dispersion, which mimics the wider distribution.

in Appendix B. This marks a shift, as the ICD-9 code was primarily used when the chief complaint was downstream of depression, whereas the ICD-10 code was used for patients of all chief complaints, even when the depression was just incidental. We later show in Appendix C that our method can recover this shift.

4. Benchmarking Explanation of Shift

Evaluating the quality of a shift explanation requires a pair of datasets with a known, underlying distinction. To circumvent the lack of ground truth data, we create a benchmark where we induce controllable, synthetic shifts in MIMIC-IV, providing a quantitative ground truth against which we can evaluate generated explanations. Our benchmark is designed to assess if a natural language explanation is both *accurate* (correctly identifies the induced shift) and *powerful* (is discriminative for the underlying feature). The rich structured data in MIMIC-IV, linked to each clinical note, allows us to induce precise shifts based on features F and treat the nature and magnitude of this induced change as the ground truth. We describe our shift induction methods and two complementary evaluation frameworks below.

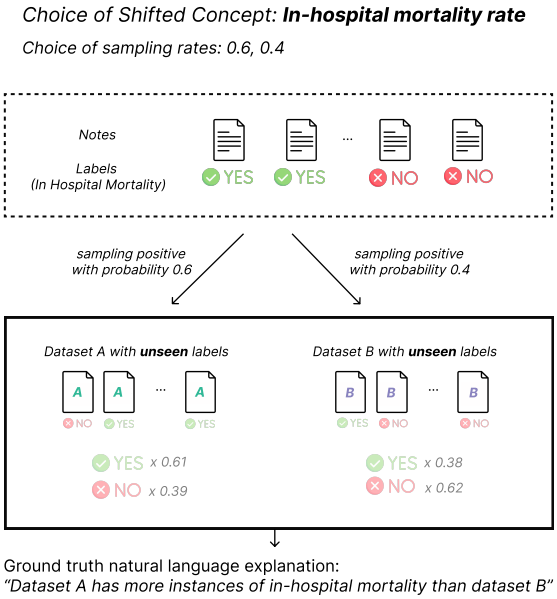

Figure 3: Construction of datasets with a prevalence shift on binary feature: Dataset A has more instances of in-hospital mortality than Dataset B

4.1. Inducing Synthetic Distribution Shifts

We begin with a large set of hospital admission notes U. To create datasets with shift, we sample two datasets, $D_A \subset U$ and $D_B \subset U$, such that their underlying distributions differ with respect to a chosen feature F. Our objective would be to generate a natural language explanation $E(D_A, D_B)$ that accurately describes the shift on feature F.

Shift on Binary Variable Let $F : U \to \{0, 1\}$ be a binary feature (e.g., in-hospital-mortality). We construct two datasets, dataset A (D_A) and dataset B (D_B), with target prevalences p_A and p_B where $p_A > p_B$. We sample without replacement from U using stratified sampling over the strata $F = 1$ and $F = 0$ to satisfy $P(F(x) = 1|x \in D_A) = p_A$ and $P(F(x) = 1|x \in D_B) = p_B$. Figure 3 illustrates an example with $p_A = 0.60$ and $p_B = 0.40$ for in-hospital mortality.

Shift on Continuous Variable Let $F : U \to \mathbb{R}$ be a continuous feature (e.g., patient age at admission), and let N be the number of notes in U. We first sort U in ascending order of F so that x_1 has the lowest value and x_N the highest. To induce a controlled shift toward higher F values in dataset A, we assign a sampling weight to each x_i via a temperature-controlled softmax: $\pi(x_i) = \frac{\exp(i/T)}{\sum_{j=1}^{N} \exp(j/T)}$ for each note x_i in

the full corpus U, where $T > 0$ modulates concentration (lower T places more mass on high-ranked items). We then sample n_A items without replacement from U according to π to form D_A. Dataset D_B is obtained by sampling n_B items uniformly without replacement from the remaining pool $U \setminus D_A$. This procedure yields D_A with a distribution of F shifted toward higher values relative to D_B. In the extremes, $T \to \infty$ approaches uniform sampling for D_A, while $T \to 0$ concentrates D_A almost entirely on the notes with the highest values of F. A concise discussion of how T affects the realized separation between D_A and D_B measured through Maximum Mean Discrepancy (MMD) appears in Appendix A.4 and Table 8.

4.2. Evaluating Explanations against ground-truth shift

Evaluation 1: Rubric-based Scoring To assess an explanation's semantic alignment with the ground-truth shift, we employ a rubric-based scoring system that can be consistently applied at scale by a large language model (LLM), similar to the evaluation of HealthBench (Arora et al., 2025). The rubric awards points based on how accurately the explanation identifies the shifted feature and the direction of the shift. For example, Table 1 shows the rubric for a shift in in-hospital mortality. A high score (+3) is given for explicitly naming the feature and its direction, while partial credit (+2) is given for identifying highly correlated concepts (e.g., "palliative care," "end-of-life discussions"). This method is intuitive but may depend on the LLM's domain knowledge to recognize correlated features.

Evaluation 2: Evaluating Discriminative Power An effective explanation should not just name the shifted feature F but also capture the underlying distinction in data with different F. A powerful explanation $E(D_A, D_B)$ should be convertible into a strong classifier for the ground-truth feature F on held-out data. However, other features G may be highly correlated with F (e.g., palliative care is correlated with mortality). In such cases, an explanation based on G is also valid.

To provide an objective measure of an explanation's discriminative power that is robust to such correlations, we introduce the second evaluation for the explanation's discriminative power on the ground-truth feature F. This measures whether the explanation is an actionable summary that has distilled the key discriminative pattern from the notes. We construct a held-out test set that is class-balanced for binary F (equal number of $F = 0$ and $F = 1$) and, for continuous F, comprises notes from the lower and upper quartiles (bottom and top 25%). The procedure is detailed in Algorithm 1. Examples and prompts for both evaluation methods are detailed in Appendix E.

Score	Criteria
+3	Explicitly mentions *in-hospital death*, and says that it is more likely in dataset A
+2	Doesn't explicitly mention *in-hospital death*, but mentions characteristics that are correlated with it, and says those characteristics are more likely in dataset A
+1	Suggests there is a difference between the datasets, but doesn't make an effort to explain it
0	Mentions a difference completely unrelated to *in-hospital death*
-1	Fails to present a difference between datasets, or says they are the same
-2	Mentions a characteristic related to *in-hospital death*, but says it is more common in dataset B
-3	Mentions *in-hospital death*, but says it is more common in dataset B

Table 1: Evaluation 1: Rubric-based scoring criteria for in-hospital mortality shift. The template applies to all features by substituting the *feature name*.

Algorithm 1 Evaluation 2: Discriminative Power (Feature Classification)

Input: Explanation E, ground-truth feature F, full data corpus U.
Output: Accuracy

Construct a held-out test set $D_h \subset U$ of notes with balanced values of feature F.
`// Binary: equal counts for` F`=0 and` F`=1; Continuous: samples from bottom and top 25% quantiles.`
Use the explanation $E(D_A, D_B)$ to guide an LLM in generating feature predictions Y_{pred} for all notes in D_h.
Let Y_{true} be the set of corresponding ground-truth feature labels for notes in D_h.
Calculate classification metrics by comparing Y_{pred} and Y_{true}.
return *Accuracy*

5. Explanation Generation with SIReNs

We introduce **SIReNs** (**S**hift **I**nterpretation via **R**epresentative **N**otes), a pipeline that produces natural language explanations for distributional shift between sets of clinical notes. As illustrated in Figure 4, SIReNs first reduce the two large datasets to small, highly illustrative subsets of notes and then use these subsets to prompt an LLM for a summary of the underlying shift.

5.1. Extracting Representative Notes

Clinical notes are quite long compared to typical inputs in NLP, and they tend to cover a large number of concepts. In real-world cases of subtle dataset shift, only some notes will provide evidence of the shifting feature, and the signal may be concentrated in only a small portion of that note. Our goal therefore is to identify a small subset of k notes from each dataset (D_A and D_B) that are maximally *indicative* of the shift.

To achieve this, each clinical note $x \in D_A \cup D_B$ is encoded into an embedding of dimension m: $h(x) \in \mathbb{R}^m$ using GatorTron (Yang et al., 2022). We obtain $h(x)$ by mean-pooling the last hidden states. We seek a scoring function $f : \mathbb{R}^m \to \mathbb{R}$ that, when applied to a note's embedding, reflects its likelihood of belonging to D_A. For all the notes in $D_A \cup D_B$, the top-k notes under f form the representative subset for D_A, and the bottom-k notes form the subset for D_B. We explore three methods for deriving the scoring function f below:

Difference of means First, we define a linear scoring function derived from the difference between the mean embeddings of the two datasets (Appendix G). Motivated by the view that embedding dimensions approximate near-orthogonal latent concepts (Elhage et al., 2022), a shift in concept prevalence should manifest as a shift in the mean vector. Let $v_{\text{diff}} = \text{mean}(\{h(x) : x \in D_A\}) - \text{mean}(\{h(x) : x \in D_B\})$. The score for any note is therefore the projection of its embedding onto this difference vector: $f(x) = h(x) \cdot v_{\text{diff}}$. The top-$k$ and bottom-k notes are respectively selected as representatives of D_A and D_B.

Linear Probing We also train an L2-regularized logistic regression classifier on frozen GatorTron embeddings to discriminate between D_A and D_B, labeling $y = 1$ for D_A. The scoring function for a note x is the predicted probability $f(x) = P(y = 1|h(x))$. We take the top-k and bottom-k among all notes as representative subsets.

Rationale-Based Finetuning Because clinical notes are long and often contain substantial irrelevant text, we adopt a rationale-based selector–predictor framework. Starting from a GatorTron base model, we fine-tune only the final l_{tuning} transformer layers. A selector module assigns soft selection weights to tokens, while a predictor module takes in an embedding of the selected span(s) to classify dataset membership. Training minimizes the loss function $L = L_{\text{pred}} + \alpha \cdot L_{\text{sparsity}} + \beta \cdot L_{\text{continuity}}$, where L_{pred} is cross-entropy, L_{sparsity} encourages short rationales, and $L_{\text{continuity}}$ discourages fragmentation. During training time, we use Gumbel-Softmax for differentiable token selection (Lei et al., 2016). At inference, for each note we keep the top $p\%$ tokens (hyperparameter) by selection probability, form the rationale embedding, and obtain $f(e) = P(y = 1|h(\text{rationale}))$. Representative sets are chosen exactly as before (top-k form subset for D_A and bottom-k form subset for D_B), but only the extracted rationale snippets are retained for explanation.

5.2. Using representative notes to generate a shift explanation

With small sets of representative notes or text snippets, a human could feasibly identify the underlying shift. To automate this and enable evaluation at scale, we provide these sets to a powerful LLM. The prompt template can be found in Appendix E.

6. Experiments

6.1. Experiment Settings

Our benchmark can induce shifts along **arbitrary** structured attributes in MIMIC-IV. This allows for the creation of a wide range of evaluation scenarios by selecting different attributes (e.g., demographics, diagnoses, lab values) as the feature to induce a shift on. While the benchmark supports a wide selection of features, here we focus on six representative features to evaluate SIReN's performance:

- **Binary Features: Death** (patient death in-hospital), **Procedure** (whether a patient had any procedure), and **Palliative care** (whether the patient received palliative care).

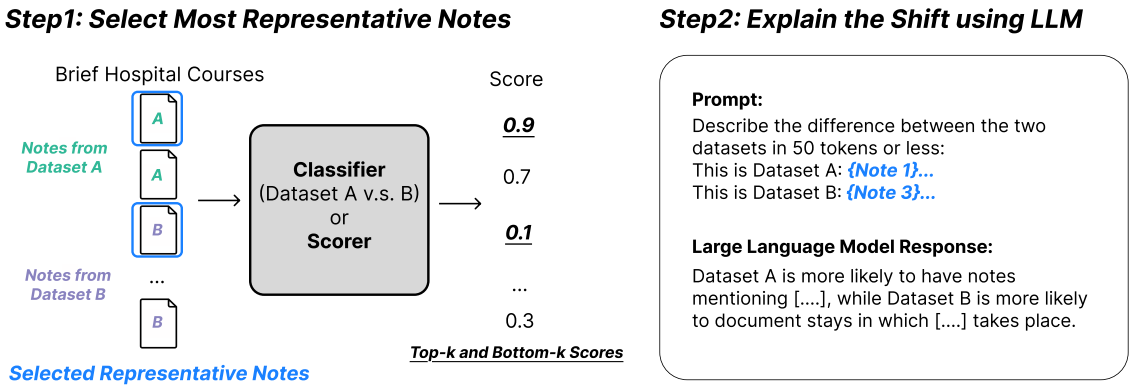

Figure 4: The SIReNs pipeline: From two large datasets of clinical notes (D_A, D_B), we first identify k representative notes that best capture the distributional shift for each dataset. These notes are then provided to a large language model (LLM) to generate a concise, natural language explanation of the difference.

	Mean-Diff	Linear Probing	Rationale (Snippet)	Rationale (Context)	Random
Binary features					
Death	$\mathbf{2.62 \pm 0.06}$	2.33 ± 0.07	2.12 ± 0.07	2.37 ± 0.08	1.83 ± 0.15
Procedure	$\mathbf{2.77 \pm 0.06}$	2.72 ± 0.11	2.73 ± 0.12	2.73 ± 0.06	2.05 ± 0.17
Palliative	2.80 ± 0.05	$\mathbf{2.87 \pm 0.04}$	2.53 ± 0.11	2.60 ± 0.10	2.37 ± 0.16
Continuous features					
Length of stay	$\mathbf{2.09 \pm 0.04}$	1.78 ± 0.09	1.04 ± 0.11	1.65 ± 0.11	1.85 ± 0.07
Diagnosis count	$\mathbf{1.98 \pm 0.03}$	1.88 ± 0.05	1.37 ± 0.11	1.85 ± 0.06	1.95 ± 0.04
Start age	$\mathbf{1.71 \pm 0.08}$	1.43 ± 0.11	0.86 ± 0.11	1.39 ± 0.10	1.47 ± 0.10

Table 2: (Evaluation Method 1 in Section 4) Mean rubric score ± standard error for each method, averaged over all settings and all trials. **Best score** is in bold and second-best is underlined; higher is better.

	Mean-Diff	Linear Probing	Rationale (Snippet)	Rationale (Context)	Random	Ground Truth
Binary features						
Death	$\mathbf{0.95 \pm 0.01}$	0.93 ± 0.01	0.91 ± 0.02	0.93 ± 0.01	0.56 ± 0.05	0.97 ± 0.00
Procedure	$\mathbf{0.66 \pm 0.01}$	0.63 ± 0.02	0.63 ± 0.01	0.62 ± 0.01	0.50 ± 0.02	0.68 ± 0.01
Palliative	$\mathbf{0.88 \pm 0.01}$	0.86 ± 0.02	0.85 ± 0.02	0.87 ± 0.02	0.46 ± 0.04	0.92 ± 0.00
Continuous features						
Length of stay	$\mathbf{0.60 \pm 0.02}$	0.58 ± 0.02	0.54 ± 0.02	0.56 ± 0.02	0.51 ± 0.02	0.83 ± 0.00
Diagnosis count	$\mathbf{0.64 \pm 0.02}$	0.63 ± 0.02	0.61 ± 0.02	0.64 ± 0.02	0.55 ± 0.03	0.76 ± 0.01
Start age	0.54 ± 0.01	0.52 ± 0.01	0.56 ± 0.01	$\mathbf{0.58 \pm 0.01}$	0.53 ± 0.02	0.73 ± 0.01

Table 3: (Evaluation Method 2 in Section 4) Mean accuracy ± standard error for each method, averaged over all settings and all trials. **Best score** is in bold and second-best is underlined (ground truth excluded); higher is better. The Ground Truth column shows discriminative performance when the LLM is given an oracle explanation that directly states the actual shifted feature (approximate upper bound).

- **Continuous Features: Diagnosis Count** (the total count of diagnoses), **Length of Stay** (the length of hospital stay in days), and **Age** (the patient's age at the start of the admission).

We test a range of shift magnitudes to simulate different real-world scenarios:

- **Binary Shifts:** For each binary feature, we induce three distinct shifts in prevalence $(p_A \to p_B)$:
  - A large, eminent change $(0.9 \to 0.1)$: verifies recovery of clear prevalence differences.
  - A more difficult, realistic, and subtle change $(0.6 \to 0.4)$: probes sensitivity under low signal.
  - The disappearance of a feature $(0.3 \to 0.0)$: tests robustness when explicit evidence is absent in one cohort.
- **Continuous Shifts:** Shifts are induced using our rank-based sampling procedure with four different temperature settings: $T \in \{0.8, 1, 3, 5\}$. We report empirical divergence measured in Mean Maximum

Discrepancy scores at each temperature and discuss observed separability in Appendix A.4.

For each combination of feature, shift magnitude, and method, we conduct 20 trials. In each trial, we sample datasets D_A and D_B of 500 notes each. We then use each method to select $k = 5$ representative notes (or rationales) from each dataset to generate an explanation, which is then evaluated by our two evaluation methods. For the second evaluation on the explanation's discriminative power, we sample 20 held-out notes in total for each trial.

Implementation Details. For the rationale-based method, the loss weights are set to $\alpha = 0.001$ and $\beta = 0.001$. At inference time, we extract rationales by selecting the top $p = 20\%$ of tokens with the highest selector probabilities. We freeze the pretrained Gatrotron model and finetune only the last four layers as well as the selector and predictor module. We evaluate two variants of this method: Rationale (Snippets), which uses only the selected tokens, and Rationale (Context), which includes a 20-token window before and after each snippet. In all experiments, explanations are generated by a Gemini model (Gemini-2.0-Flash-001) accessed via Vertex AI (usage path compliant with PhysioNet data policies); prompt templates are detailed in Appendix E.

6.2. Methods and Baselines

We evaluate four SIReNs methods (Difference of Means, Linear Probing, Rationale (Snippets), and Rationale (Context)) along with two additional baselines:

- **Ground Truth:** An upper-bound where the true shifted attribute is provided to LLM as explanation (e.g., "Dataset 1 has more instances of in-hospital mortality than dataset 2").
- **Random Selection:** A lower-bound baseline where the $k = 5$ notes are randomly selected from each dataset.

6.3. Results and Analysis

Tables 2 and 3 report per-feature performance averaged over shift magnitudes. Figure 5 compares the same methods on binary features under different shift magnitudes. More detailed results appear in Appendix A.1. All SIReNs methods substantially outperform Random on both evaluation methods, indicating the importance of a strong selection mechanism for surfacing informative evidence. Difference-of-means was generally the most consistent, except for Age, a hyper-localized feature, for which rationale methods perform the best. Generally, the snippets derived from rationale selector methods are insufficient alone, but can be much stronger when additional context is provided. We do note that Evaluation Methods 1 and 2 are highly correlated, but do not necessarily always provide the same ranking across methods. An additional ablation on the number of representative notes ($k \in \{3, 5, 10\}$) is provided in Appendix A.3, showing minimal sensitivity to k across methods.

Binary Shifts Across the three binary attributes, SIReNs achieves consistently high rubric scores (Table 2), indicating that the generated explanation captures the intended semantics. For Death and Palliative care, strong lexical cues (e.g., "death", "cancer", "end-of-life care") yield near-oracle accuracies for the discriminative power. However, Procedure behaves differently. Although rubric scores generally exceed 2.7, accuracies stagnate around 0.65, even for the Ground Truth oracle. To disentangle selection from explanation, we report *selector correctness* (top-/bottom-k purity): of the k notes flagged as most indicative of D_A, how many truly come from D_A (and analogously for D_B). Purity is high ($\sim$ 4.3–5.0 of 5, as shown in Appx. A.2), indicating that while the selection of note is mostly correct, *Procedure*'s lower accuracy may stem from weaker/indirect cues in BHC notes.

Figure 5 illustrates the impact of shift magnitude: all methods exhibit their sharpest degradation at the subtle prevalence change ($0.6 \rightarrow 0.4$). The Difference-of-Means variant is the most stable across magnitudes, whereas others show larger variance.

Continuous Shifts Continuous shifts are harder for SIReNs than binary ones, with uniformly lower rubric scores and accuracies (Tables 2-3). All methods remain tightly clustered and well below the ground truth oracle, suggesting a shared ceiling rather than isolated model weakness. This pattern implies that lexical or conceptual traces of gradual intensity changes are diffuse and easily drowned out, limiting any selector's ability to surface high-signal cues. On the rubric benchmark, mean-diff continues to excel, while on the accuracy benchmark, rationale-based methods and linear probing perform comparably. Overall, the continuous setting exposes a clear ceiling for current SIReNs methods and motivates the need for additional method development to aggregate

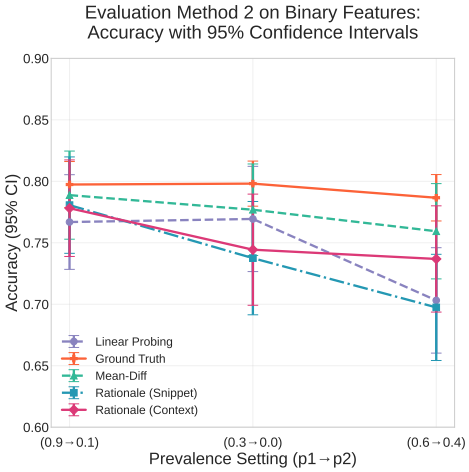

Figure 5: Discriminative accuracy across different shift magnitudes for binary features.

many weak, spatially and semantically dispersed indicators of shift.

7. Conclusion

In this paper, we formalize and benchmark the task of explaining dataset shift in clinical notes by releasing a benchmarking software suite that induces controlled, feature-based shifts in MIMIC-IV alongside ground truth explanations. We introduce SIReNs, which retrieve representative notes from the two datasets and prompt an LLM to explain the underlying shift. Empirical results using two different facets of evaluation show that SIReNs variants can successfully provide cogent explanations; the best variant for shift identification can depend on the localization of the underlying shift, but suggests it may be useful to try multiple methods, as the true shift will be unknown. Methods do struggle more on subtler continuous shifts, underscoring the need for techniques that pool distributed weak signals.

Future Work There are clear next steps to extend our foray into interpreting text shift, in addition to improving continuous shift identification. First, our benchmark suite was built to be extensible, enabling the simple creation of arbitrary additional shifts for further exploration on a wider set of features. Second, exploration into multi-factorial shifts is key. We hypothesize rationale-based methods may help isolate the effect of multiple via iterative masking. In addition, all analyses use single-site MIMIC-IV BHC notes. Therefore, extending to cross-institutional datasets and broader sections is an important next step. Finally, given our promising initial results, further work should be conducted on using our software to find real-world shifts in clinical notes, as they may have implications for model performance, including a user study evaluating whether the explanations measurably improve monitoring and mitigation decisions.

References

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Isabel Bilotta, Scott Tonidandel, Winston R Liaw, Eden King, Diana N Carvajal, Ayana Taylor, Julie Thamby, Yang Xiang, Cui Tao, Michael Hansen, et al. Examining linguistic differences in electronic health records for diverse patients with diabetes: natural language processing analysis. *JMIR Medical Informatics*, 12(1):e50428, 2024.

Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, and Hamed Firooz. Unirex: A unified learning framework for language model rationale extraction. In *International Conference on Machine Learning*, pages 2867–2889. PMLR, 2022.

Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24199–24208, 2024.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385 (3):283–286, 2021.

Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with tableshift. *Advances in Neural Information Processing Systems*, 36:53385–53432, 2023.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. Analysing lexical semantic change with contextualised word representations. *arXiv preprint arXiv:2004.14118*, 2020.

Lin Lawrence Guo, Ethan Steinberg, Scott Lanyon Fleming, Jose Posada, Joshua Lemmon, Stephen R Pfohl, Nigam Shah, Jason Fries, and Lillian Sung. Ehr foundation models improve robustness in the presence of temporal distribution shift. *Scientific Reports*, 13(1):3767, 2023.

Gyandev Gupta, Bashir Rastegarpanah, Amalendu Iyer, Joshua Rubin, and Krishnaram Kenthapadi. Measuring distributional shifts in text: the advantage of language model-based embeddings. *arXiv preprint arXiv:2312.02337*, 2023.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*, 2016.

Deborah S Hasin, Aaron L Sarvet, Jacquelyn L Meyers, Tulshi D Saha, W June Ruan, Malka Stohl, and Bridget F Grant. Epidemiology of adult dsm-5 major depressive disorder and its specifiers in the united states. *JAMA psychiatry*, 75(4):336–346, 2018.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, et al. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969):357–362, 2023.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, pages 49–55, 2020.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv, Oct 2024. URL https://physionet.org/content/mimiciv/3.1/.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1, 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, et al. Publicly shareable clinical large language model built on synthetic clinical notes. *arXiv preprint arXiv:2309.00237*, 2023.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*, 2016.

Dongfang Li, Baotian Hu, Qingcai Chen, Tujie Xu, Jingcong Tao, and Yunan Zhang. Unifying model explainability and robustness for joint text classification and rationale extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10947–10955, 2022.

Jianfu Li, Yujia Zhou, Xiaoqian Jiang, Karthik Natarajan, Serguei Vs Pakhomov, Hongfang Liu, and Hua Xu. Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. *Journal of the American Medical Informatics Association*, 28(10): 2193–2201, 2021.

Mohammad Reza Ghasemi Madani and Pasquale Minervini. Refer: an end-to-end rationale extraction framework for explanation regularization. *arXiv preprint arXiv:2310.14418*, 2023.

Andrey Malinin, Neil Band, German Chesnokov, Yarin Gal, Mark JF Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*, 2021.

Luca Massarelli, Fabio Petroni, Aleksandra Piktus, Myle Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel. How decoding strategies affect the verifiability of generated text. *arXiv preprint arXiv:1911.03587*, 2019.

Ramin Mojtabai, Mark Olfson, and Beth Han. National trends in the prevalence and treatment of depression in adolescents and young adults. *Pediatrics*, 138(6):e20161878, 2016.

Kevin J Peterson and Hongfang Liu. An examination of the statistical laws of semantic change in clinical notes. *AMIA Summits on Translational Science Proceedings*, 2021:515, 2021.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers, 2021. URL https://arxiv.org/abs/2102.01454.

Maryam Rahimian, Jeremy L Warner, Liz Salmi, S Trent Rosenbloom, Roger B Davis, and Robin M Joyce. Open notes sounds great, but will a provider's documentation change? an exploratory study of the effect of open notes on oncology documentation. *JAMIA open*, 4(3):ooab051, 2021.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. Semeval-2020 task 1: Unsupervised lexical semantic change detection. *arXiv preprint arXiv:2007.11464*, 2020.

Sunghwan Sohn, Yanshan Wang, Chung-Il Wi, Elizabeth A Krusemark, Euijung Ryu, Mir H Ali, Young J Juhn, and Hongfang Liu. Clinical documentation variations and nlp system portability: a case study in asthma birth cohorts across institutions. *Journal of the American Medical Informatics Association*, 25(3):353–359, 2018.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: Dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020. doi: 10.1093/biostatistics/kxz041. URL https://doi.org/10.1093/biostatistics/kxz041.

Ming Tai-Seale, Sally L Baxter, Florin Vaida, Amanda Walker, Amy M Sitapati, Chad Osborne, Joseph Diaz, Nimit Desai, Sophie Webb, Gregory

Polston, et al. Ai-generated draft replies integrated into health records and physicians' electronic communication. *JAMA Network Open*, 7(4):e246565–e246565, 2024.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Aaron A Tierney, Gregg Gayre, Brian Hoberman, Britt Mattern, Manuel Ballesca, Patricia Kipnis, Vincent Liu, and Kristine Lee. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery*, 5(3):CAT–23, 2024.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. Clinical text summarization: adapting large language models can outperform human experts. *Research square*, pages rs–3, 2023.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records, 2022. URL https://arxiv.org/abs/2203.03540.

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei W Koh, and Chelsea Finn. Wild-time: A benchmark of in-the-wild distribution shift over time. *Advances in Neural Information Processing Systems*, 35:10309–10324, 2022.

Ruiqi Zhong, Heng Wang, Dan Klein, and Jacob Steinhardt. Explaining datasets in words: Statistical models with natural language parameters. *Advances in Neural Information Processing Systems*, 37:79350–79380, 2024.

	Mean-Diff	Linear Probing	Rationale(Snippet)	Rationale(Context)	Random	Ground Truth
Binary variable shift (prevalence change $p_1 \to p_2$)						
Death ($0.9 \to 0.1$)	$\mathbf{2.65 \pm 0.109}$	2.35 ± 0.109	2.10 ± 0.069	2.35 ± 0.109	1.70 ± 0.300	3.00 ± 0.00
Death ($0.6 \to 0.4$)	$\mathbf{2.55 \pm 0.114}$	2.15 ± 0.150	2.10 ± 0.191	2.45 ± 0.170	1.95 ± 0.276	3.00 ± 0.00
Death ($0.3 \to 0.0$)	$\mathbf{2.65 \pm 0.109}$	2.50 ± 0.115	2.15 ± 0.082	2.30 ± 0.105	1.85 ± 0.196	3.00 ± 0.00
Procedure ($0.9 \to 0.1$)	2.85 ± 0.082	$\mathbf{3.00 \pm 0.000}$	2.95 ± 0.050	2.65 ± 0.109	1.84 ± 0.308	3.00 ± 0.00
Procedure ($0.6 \to 0.4$)	2.80 ± 0.092	2.20 ± 0.304	2.35 ± 0.350	$\mathbf{2.81 \pm 0.101}$	2.25 ± 0.323	3.00 ± 0.00
Procedure ($0.3 \to 0.0$)	2.65 ± 0.109	$\mathbf{2.95 \pm 0.050}$	2.90 ± 0.069	2.75 ± 0.099	2.05 ± 0.223	3.00 ± 0.00
Palliative ($0.9 \to 0.1$)	$\mathbf{2.85 \pm 0.082}$	$\mathbf{2.85 \pm 0.082}$	$\mathbf{2.85 \pm 0.082}$	2.80 ± 0.092	2.50 ± 0.323	3.00 ± 0.00
Palliative ($0.6 \to 0.4$)	2.70 ± 0.105	$\mathbf{2.90 \pm 0.069}$	2.00 ± 0.281	2.45 ± 0.211	2.55 ± 0.312	3.00 ± 0.00
Palliative ($0.3 \to 0.0$)	$\mathbf{2.85 \pm 0.082}$	$\mathbf{2.85 \pm 0.082}$	2.75 ± 0.099	2.55 ± 0.170	2.30 ± 0.164	3.00 ± 0.00
Continuous variable shift (temperature T)						
Length of Stay (0.8)	$\mathbf{2.10 \pm 0.143}$	1.70 ± 0.164	1.05 ± 0.235	1.65 ± 0.196	2.00 ± 0.000	$\mathbf{3.00 \pm 0.00}$
Length of Stay (1)	$\mathbf{2.00 \pm 0.000}$	1.85 ± 0.150	0.60 ± 0.210	1.40 ± 0.210	1.90 ± 0.100	$\mathbf{3.00 \pm 0.00}$
Length of Stay (3)	$\mathbf{2.10 \pm 0.069}$	1.85 ± 0.150	0.90 ± 0.228	1.70 ± 0.300	1.80 ± 0.138	$\mathbf{3.00 \pm 0.00}$
Length of Stay (5)	$\mathbf{2.15 \pm 0.082}$	1.70 ± 0.272	1.60 ± 0.184	1.85 ± 0.150	1.70 ± 0.219	$\mathbf{3.00 \pm 0.00}$
Diagnosis Count (0.8)	1.90 ± 0.100	1.80 ± 0.138	1.50 ± 0.199	$\mathbf{2.00 \pm 0.000}$	$\mathbf{2.00 \pm 0.000}$	$\mathbf{3.00 \pm 0.00}$
Diagnosis Count (1)	$\mathbf{2.00 \pm 0.000}$	1.90 ± 0.100	1.00 ± 0.229	1.70 ± 0.164	$\mathbf{2.00 \pm 0.000}$	$\mathbf{3.00 \pm 0.00}$
Diagnosis Count (3)	$\mathbf{2.00 \pm 0.000}$	1.90 ± 0.100	1.30 ± 0.219	1.90 ± 0.100	1.90 ± 0.100	$\mathbf{3.00 \pm 0.00}$
Diagnosis Count (5)	$\mathbf{2.00 \pm 0.000}$	1.90 ± 0.100	1.68 ± 0.172	1.80 ± 0.138	1.90 ± 0.100	$\mathbf{3.00 \pm 0.00}$
Start Age (0.8)	$\mathbf{1.65 \pm 0.196}$	1.50 ± 0.199	0.20 ± 0.138	0.84 ± 0.233	1.30 ± 0.219	$\mathbf{3.00 \pm 0.00}$
Start Age (1)	$\mathbf{1.80 \pm 0.138}$	1.20 ± 0.225	0.80 ± 0.225	1.40 ± 0.210	1.70 ± 0.164	$\mathbf{3.00 \pm 0.00}$
Start Age (3)	$\mathbf{1.90 \pm 0.100}$	1.35 ± 0.233	1.00 ± 0.229	1.80 ± 0.138	1.60 ± 0.184	$\mathbf{3.00 \pm 0.00}$
Start Age (5)	1.50 ± 0.199	$\mathbf{1.65 \pm 0.209}$	1.47 ± 0.208	1.50 ± 0.199	1.26 ± 0.227	$\mathbf{3.00 \pm 0.00}$

Table 4: (Evaluation Method 1 in Section 4) Mean rubric score ± standard error for each method. The upper block shows binary-variable shifts (prevalence difference $p_1 \to p_2$) and the lower block shows continuous-variable shifts (temperature T). **Best score** is in bold and second-best is underlined (ground-truth score ignored); higher is better.

	Mean-Diff	Linear Probing	Rationale(Snippet)	Rationale(Context)	Random	Ground Truth
Binary variable shift (prevalence change $p_1 \to p_2$)						
Death ($0.9 \to 0.1$)	$\mathbf{0.970 \pm 0.010}$	0.940 ± 0.017	0.945 ± 0.016	0.948 ± 0.013	0.557 ± 0.090	0.968 ± 0.004
Death ($0.6 \to 0.4$)	$\mathbf{0.945 \pm 0.014}$	0.880 ± 0.035	0.870 ± 0.036	0.902 ± 0.036	0.580 ± 0.088	0.968 ± 0.004
Death ($0.3 \to 0.0$)	0.943 ± 0.017	$\mathbf{0.957 \pm 0.011}$	0.912 ± 0.029	0.945 ± 0.011	0.545 ± 0.074	0.973 ± 0.004
Procedure ($0.9 \to 0.1$)	$\mathbf{0.685 \pm 0.020}$	0.655 ± 0.025	0.682 ± 0.022	0.645 ± 0.028	0.524 ± 0.034	0.691 ± 0.015
Procedure ($0.6 \to 0.4$)	0.610 ± 0.020	0.590 ± 0.028	0.585 ± 0.025	$\mathbf{0.616 \pm 0.020}$	0.487 ± 0.031	0.665 ± 0.013
Procedure ($0.3 \to 0.0$)	$\mathbf{0.682 \pm 0.026}$	0.640 ± 0.024	0.620 ± 0.022	0.602 ± 0.024	0.492 ± 0.032	0.683 ± 0.013
Palliative ($0.9 \to 0.1$)	0.870 ± 0.022	0.885 ± 0.014	$\mathbf{0.915 \pm 0.016}$	0.905 ± 0.014	0.470 ± 0.075	0.925 ± 0.005
Palliative ($0.6 \to 0.4$)	$\mathbf{0.883 \pm 0.018}$	0.797 ± 0.035	0.755 ± 0.043	0.833 ± 0.034	0.478 ± 0.084	0.918 ± 0.007
Palliative ($0.3 \to 0.0$)	0.878 ± 0.020	$\mathbf{0.905 \pm 0.016}$	0.887 ± 0.030	0.885 ± 0.029	0.417 ± 0.058	0.924 ± 0.007
Continuous variable shift (temperature T)						
Length of Stay (0.8)	0.555 ± 0.031	$\mathbf{0.565 \pm 0.035}$	0.539 ± 0.022	0.538 ± 0.022	0.512 ± 0.041	0.821 ± 0.009
Length of Stay (1)	$\mathbf{0.565 \pm 0.027}$	0.535 ± 0.040	0.523 ± 0.031	0.523 ± 0.039	0.478 ± 0.029	0.824 ± 0.009
Length of Stay (3)	0.615 ± 0.041	$\mathbf{0.625 \pm 0.030}$	0.490 ± 0.026	0.530 ± 0.031	0.510 ± 0.023	0.831 ± 0.007
Length of Stay (5)	$\mathbf{0.667 \pm 0.027}$	0.588 ± 0.036	0.625 ± 0.030	0.665 ± 0.022	0.555 ± 0.041	0.828 ± 0.009
Diagnosis Count (0.8)	0.555 ± 0.043	0.543 ± 0.029	0.573 ± 0.026	$\mathbf{0.590 \pm 0.048}$	0.507 ± 0.055	0.741 ± 0.011
Diagnosis Count (1)	0.550 ± 0.046	$\mathbf{0.585 \pm 0.041}$	0.533 ± 0.031	0.575 ± 0.031	0.555 ± 0.051	0.747 ± 0.010
Diagnosis Count (3)	$\mathbf{0.720 \pm 0.038}$	0.633 ± 0.045	0.622 ± 0.033	0.647 ± 0.035	0.537 ± 0.046	0.767 ± 0.011
Diagnosis Count (5)	$\mathbf{0.753 \pm 0.022}$	0.747 ± 0.025	0.718 ± 0.030	0.743 ± 0.027	0.580 ± 0.053	0.763 ± 0.011
Start Age (0.8)	0.518 ± 0.032	0.510 ± 0.025	0.522 ± 0.026	$\mathbf{0.563 \pm 0.022}$	0.508 ± 0.036	0.718 ± 0.010
Start Age (1)	0.502 ± 0.033	0.470 ± 0.029	0.537 ± 0.025	$\mathbf{0.542 \pm 0.022}$	0.518 ± 0.032	0.722 ± 0.012
Start Age (3)	0.562 ± 0.024	0.508 ± 0.028	0.578 ± 0.031	$\mathbf{0.580 \pm 0.033}$	0.543 ± 0.037	0.735 ± 0.011
Start Age (5)	0.573 ± 0.025	0.602 ± 0.026	0.618 ± 0.024	$\mathbf{0.635 \pm 0.023}$	0.571 ± 0.028	0.741 ± 0.012

Table 5: (Evaluation Method 2 in Section 4) Mean accuracy score ± standard error for each method. The upper block shows binary-variable shifts (prevalence difference $p_1 \to p_2$) and the lower block shows continuous-variable shifts (temperature T). **Best score** is in bold and second-best is underlined (ground-truth score ignored); higher is better.

Appendix A. Additional Experiment Results

A.1. Full Experimental Results

Here we provide the comprehensive results of our benchmarking experiments, serving as a detailed reference that supports the summarized findings presented in the main paper (Section 6). The tables and figures below contain the full breakdown of performance for each method across all tested features and shift magnitudes.

Table 4 presents the mean rubric scores and standard errors for each method on both binary and continuous variable shifts. Table 5 shows the corresponding mean classification accuracies and standard errors, which evaluate the discriminative power of the generated explanations.

Additionally, we include figures that visually represent these results. While Figure 5 in Section 6 already illustrates the performance of each method on the discriminative accuracy, Figure 6 further demonstrates the performance of each method on rubric scores across different shift magnitudes for binary features. Similar to the accuracy scores, under the setting with the most subtle shift, the methods perform worse. These detailed results provide a complete picture of our experimental outcomes and allow for a thorough analysis of each method's strengths and weaknesses under various conditions.

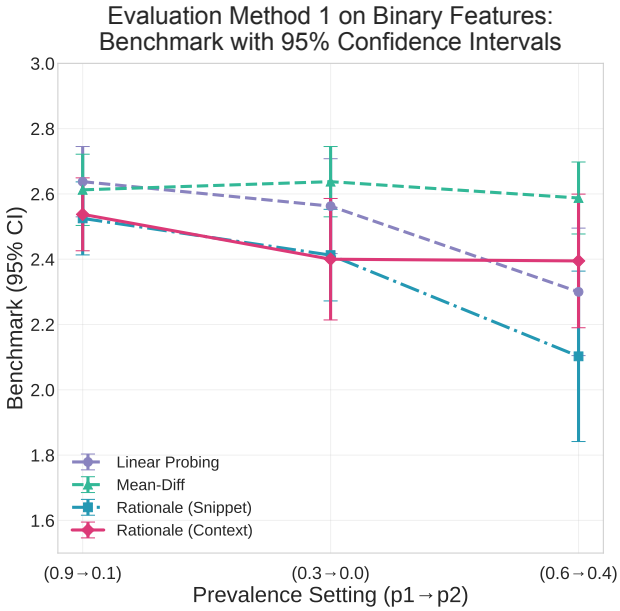

Figure 6: Method Performance on rubric scores across different shift magnitudes for binary features.

A.2. Selector Correctness

For binary features with k=5 representatives per side, we report *selector correctness* (also called *purity*) for: (i) *top-k purity* = the number of the k notes selected as most indicative of D_A that truly come from D_A; (ii) *bottom-k purity* = the number of the k notes selected as most indicative of D_B that truly come from D_B. The ideal value is 5/5. We report mean±SE across seeds.

Purity is consistently high ($\sim$ 4.3–5.0/5) for *death* and *palliative*, and remains high for *procedure*. Together with Table 3, this indicates that lower oracle accuracy for *procedure* reflects weaker or indirect mentions in BHC, not selection failure. We do note that an example may be correct, without necessarily being the clearest exemplar.

For *procedure*, the gap to oracle despite high purity likely reflects weaker surface evidence in BHC: procedure information is often indirect (peri-procedural course, templated phrases) and heterogeneous, and

Feature	Linear Probing	Mean-Diff	Random	Rationale (Context)	Rationale (Snippet)
death	4.533 ± 0.115 / 4.717 ± 0.079	5.000 ± 0.000 / 5.000 ± 0.000	1.817 ± 0.167 / 3.067 ± 0.181	4.817 ± 0.102 / 4.933 ± 0.052	4.817 ± 0.102 / 4.933 ± 0.052
procedure	4.383 ± 0.098 / 4.467 ± 0.110	4.367 ± 0.109 / 4.450 ± 0.110	2.153 ± 0.182 / 3.254 ± 0.138	4.518 ± 0.108 / 4.536 ± 0.132	4.517 ± 0.102 / 4.517 ± 0.125
palliative	4.300 ± 0.107 / 4.767 ± 0.069	4.517 ± 0.118 / 4.917 ± 0.043	1.833 ± 0.158 / 2.917 ± 0.198	4.400 ± 0.147 / 4.933 ± 0.040	4.400 ± 0.147 / 4.933 ± 0.040

Table 6: Selector correctness (**top-k / bottom-k**; k=5; mean±SE) on binary features.

admissions may involve multiple procedures, diluting explicit cues. Hence, the lower discriminative power stems from textual signaling rather than selection failure.

A.3. Number of representative notes k

We varied $k \in \{3, 5, 10\}$ on *Palliative care* and *Procedure* and observed minimal sensitivity across methods in the discriminative-power evaluation (Evaluation 2). Performance fluctuates only slightly across k, suggesting robustness to the choice of k. In practice, k is bounded by the LLM context window; we use k=5 as a token/compute trade-off.

Feature	k	Ground Truth	Linear Probing	Mean-Diff	Random	Rationale (Context)	Rationale (Snippets)
Palliative	3	0.918±0.004	0.868±0.017	0.869±0.012	0.501±0.040	0.866±0.015	0.849±0.020
Palliative	5	0.922±0.004	0.863±0.015	0.877±0.011	0.455±0.042	0.874±0.016	0.852±0.020
Palliative	10	0.917±0.004	0.893±0.014	0.863±0.012	0.459±0.043	0.865±0.017	0.860±0.018
Procedure	3	0.598±0.008	0.554±0.015	0.606±0.013	0.458±0.016	0.551±0.015	0.583±0.015
Procedure	5	0.601±0.008	0.569±0.014	0.612±0.011	0.496±0.015	0.577±0.014	0.559±0.015
Procedure	10	0.599±0.008	0.575±0.012	0.613±0.013	0.486±0.016	0.571±0.014	0.557±0.015

Table 7: Ablation on the number of representative notes k (Evaluation 2 accuracy).

A.4. Temperature parameter and induced divergence

For continuous features, T controls how aggressively we bias sampling of D_A toward high-rank (large-F) notes. Lower T increases this bias. Because we sample without replacement and then draw D_B from the remaining pool, the head is depleted for D_B, so overlap between D_A and D_B does not vary monotonically with T. Consequently, performance is not strictly inverse in T; it follows the realized separation. To report the actual shift achieved at each T, we include Maximum Mean Discrepancy (MMD) values in Table 8. Within a feature, larger MMD mostly indicates more separation (e.g., for Length of Stay and Diagnosis Count, T=5 yields the largest MMD).

Feature	$T = 0.1$	$T = 0.8$	$T = 1.0$	$T = 3.0$	$T = 5.0$	$T = 10.0$
Length of Stay	0.001571± 0.000245	0.001668 ± 0.000294	0.001516 ± 0.000176	0.001648 ± 0.000230	0.002143 ± 0.000420	0.002504 ± 0.000572
Diagnosis Count	0.001581 ± 0.000277	0.001587 ± 0.000258	0.001600 ± 0.000212	0.001865 ± 0.000410	0.001962 ± 0.000312	0.003021 ± 0.000814
Start Age	0.001500 ± 0.000197	0.001636 ± 0.000350	0.001734 ± 0.000396	0.001692 ± 0.000289	0.001869 ± 0.000384	0.002171± 0.000582

Table 8: Realized separation between D_A and D_B for continuous features, measured by MMD (mean ± std over 20 seeds). Larger is more separation.

Appendix B. Proof of Real-World Dataset Shift

We analyzed the temporal evolution of Brief Hospital Course (BHC) notes in MIMIC-IV from 2008 to 2020 to demonstrate the non-stationarity of clinical documentation. Figure 7 reveals two significant trends:

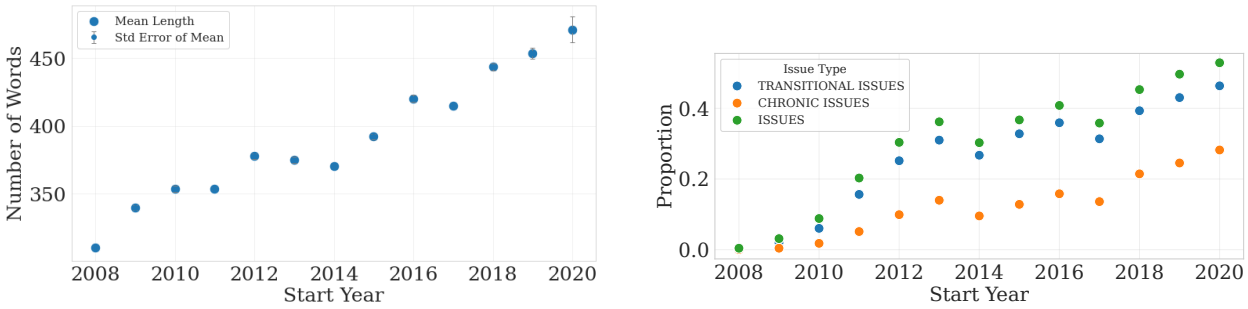

Figure 7: Temporal shift in Brief Hospital Course (BHC) notes from 2008 to 2020. The average note length rose sharply, and the prevalence of a structured "Issues" section grew from under 1% to over 50%.

Quantitative Changes Between 2008 and 2020, mean BHC note length increased by 34% (310 to 417 words), corresponding to a 0.72 standard deviation shift. This expansion co-occurred with pronounced structural change in template shown in the left panel in Figure 7: the structured "Issues" section was present in less than 1% of notes in 2008 but in more than 50% by 2020. These patterns indicate that the observed lengthening reflects not solely elaborated narrative content but also progressive standardization of documentation structure.

Compositional Changes This growth in length is not arbitrary but is largely attributable to a structural change in documentation practices, namely the adoption of new, standardized note templates. The right panel of Figure 7 details the emergence and proliferation of the "Issues" section, a structured component for documenting patient problems. These sections, which were virtually absent (prevalence ¡ 1%) in 2008, became a standard feature, appearing in over 50% of BHC notes by 2020. The most common of these are "Transitional Issues" and "Chronic Issues," which provide clearly delineated summaries of patient conditions for easier review by clinicians.

This evolution from free-form to structured documentation exemplifies the distribution shifts that clinical NLP models might encounter and where explanations for the shift would be helpful for adaptation to downstream tasks.

Semantic Shift in Depression Cohort To quantify the visual pattern from Fig. 2, we compute Maximum Mean Discrepancy (MMD) on mean-pooled GatorTron embeddings across four cohorts: *ICD-9 depression*, *ICD-10 depression*, *random ICD-9*, and *random ICD-10*. As shown in Table 9, ICD-9 depression exhibits the largest separation from every other group. By contrast, ICD-10 depression is very close to random ICD-10, indicating that the ICD-10 cohort resembles contemporaneous notes overall. The non-zero random ICD-9 vs. random ICD-10 baseline reflects generic temporal drift, but the larger ICD-9-depression gaps show a *diagnosis-specific* shift beyond time alone. Together with Fig. 2, this supports the interpretation that ICD-9 depression notes concentrate on depression-centric admissions, while ICD-10 depression more often appears incidentally across diverse complaints.

Group 1	Group 2	Raw MMD
Depression ICD9	Depression ICD10	0.0327
Depression ICD9	Random ICD9	0.0302
Depression ICD9	Random ICD10	0.0367
Depression ICD10	Random ICD9	0.0115
Depression ICD10	Random ICD10	0.0030
Random ICD9	Random ICD10	0.0109

Table 9: MMD between depression and random cohorts across ICD coding.

Appendix C. Interpreting the Semantic Shift in the Depression Cohort with SIReNs

In Section 3.2, we demonstrated a semantic shift between clinical notes from the ICD-9 and ICD-10 depression cohorts. While a PCA visualization confirmed a difference in the embedding space, it did not *explain the nature of the shift*. Here, we apply our SIReNs method to interpret this real-world change. For this experiment, we sampled two datasets of 500 notes each from the ICD-9 and ICD-10 cohorts, selecting 10 representative notes from each to generate an explanation.

- **Difference of Means**: "Dataset 1 focuses on psychiatric and substance use issues, including medication adjustments and psychosocial interventions. Dataset 2 emphasizes surgical procedures and post-operative recovery, mentioning pain management and discharge planning."
- **Linear Probing**: "Dataset 1 predominantly features psychiatric conditions and substance use, while Dataset 2 primarily focuses on chronic medical illnesses and end-of-life care."
- **Random**: "Dataset 1 shows a higher prevalence of mental health and neurological issues, while Dataset 2 has more cases related to surgical aftercare and chronic medical conditions."

While each method has its own subtle phrasing, all three SIReNs approaches consistently identify the same core shift: the ICD-9 depression notes (Dataset 1) are more centered on psychiatric care and mental health issues, while the ICD-10 notes (Dataset 2) are more focused on surgical procedures and general medical conditions. This corroborating evidence from different selection methods increases confidence in the explanation and mitigates the risk of artifacts.

This observed shift aligns with broader trends in mental health documentation and clinical practices. Mojtabai et al. reported an increase in major depressive episode prevalence since 2005, particularly among young adults and adolescents. Hasin et al. partly attributes this to differential diagnostic standards introduced by the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), released in 2013, approximately 1.5 years before the ICD shift. While our findings do not confirm whether changes in diagnostic conditions or a broader mandate to diagnose mental health conditions are solely responsible for the increased prevalence of major depressive disorder, they do show that within our evaluated dataset, post-ICD-10 hospital admissions with depression diagnoses were less likely to describe severe psychiatric conditions than pre-ICD-10 admissions with such diagnoses.

Appendix D. Implementation Details and Computational Resource Use

In order to ensure reproducibility, we disclose the exact number of trials that we ran for our benchmarking pipeline, as well as the cost in GPU-hours and Vertex API calls that would be required to replicate it.

In our benchmarking procedure, we ran a total of 2,100 trials, each independently using one NVIDIA A6000 GPU for an average of 16.55 seconds, for a total of 13.24 GPU hours. The bulk of our GPU usage was for training and inference on Gatortron. The model has 345 million parameters; for the rationale-based methods, we fine-tuned the last four layers of the Gatortron model.

Each individual trial required three different calls to the Vertex API platform: one for obtaining an explanation, and one for performing each of the two benchmarks. This resulted in a total of 6,300 API calls.

Appendix E. Example Prompts and Completions

Prompts for generating shift explanation using representative notes Here is the prompt for LLM to generate a shift explanation between the two datasets mentioned in Section 5:

> You are a hospital administrator seeking to understand the distributional difference between two datasets of hospital stays, based on discharge summaries of each. Describe the difference between the two datasets in [max tokens] words or less: Here is dataset 1: [...] Here is dataset 2: [...].

Prompts for Obtaining Rubric Scores Here is an example of how an external LLM grades an explanation to obtain the rubric score as introduced in Section 4.2.

USER: You are asked to grade the following response with the given criteria. In this example, the underlying shift is in in-hospital mortality (Death).

- +3: Explicitly mentions **in-hospital mortality**, and says that it is more likely in dataset 1.
- +2: Doesn't explicitly mention **in-hospital mortality**, but mentions related characteristics, and says they are more likely in dataset 1.
- +1: Suggests there is a difference between the datasets, but doesn't make an effort to explain it.
- +0: Mentions a difference completely unrelated to **in-hospital mortality**.
- -1: Fails to present a difference between datasets, or says they are the same.
- -2: Mentions a characteristic related to **in-hospital mortality**, but say it is more common in dataset 2.
- -3: Mentions **in-hospital mortality**, but says it is more common in dataset 2.

Provide grade in a single number, with no explanation.
Response: **Dataset 1 describes patient deaths with multi-system failure and end-of-life care. Dataset 2 focuses on resolving acute conditions and discharge planning.**

LLM OUTPUT: +2.

Here are some other examples of explanations generated for this feature (under different shift magnitudes but they share the same rubric standards), and their accompanying rubric scores graded:

- **Random**: "Dataset 1 describes hospital stays with a focus on active management of chronic conditions and acute illnesses. The cases involve complex patients with multiple comorbidities, often requiring ongoing treatment and follow-up. There's an emphasis on diagnostic workups, medication adjustments, and planning for transitional care, including specialist referrals and further investigations. Dataset 2, in contrast, presents cases with more acute, often terminal, events. There's a higher incidence of neurological emergencies, cardiac arrests, and advanced metastatic cancer leading to comfort care and death in the hospital. While chronic conditions are present, the focus shifts to managing acute decompensation and end-of-life care, with less emphasis on long-term management and outpatient follow-up." (-3).
- **Rationale(Snippet)**: "Dataset 1 primarily describes patients presenting with stroke, evidenced by terms like 'ischemic stroke,' 'tPA,' 'MCA infarct,' 'aphasia,' and 'hemiparesis.' The notes detail stroke characteristics, treatment (TPA, thrombectomy), affected brain regions, and risk factors. Dataset 2 focuses on patients with cardiac issues, including 'unstable angina,' 'chest pain,' 'coronary artery disease,' 'EKG changes,' 'troponin,' and 'NSTMI.' The notes discuss diagnostic findings (EKG, cardiac enzymes), potential causes (anxiety, GERD), and interventions (stents, CABG).summary, Dataset 1 centers on cerebrovascular events (strokes), while Dataset 2 revolves around cardiovascular problems, particularly ischemic heart disease." (0)
- **Linear Probing**: "Dataset 1 describes hospital stays primarily ending in patient death, often following critical events like cardiac arrest, respiratory failure, or complications from severe underlying conditions such as metastatic cancer or pancreatitis. The summaries detail aggressive resuscitative efforts, multiple

organ dysfunction, and eventual transition to comfort measures.Dataset 2 describes hospital stays with more positive outcomes, focusing on post-operative care, resolution of acute issues like hyponatremia or melena, and management of chronic conditions. Discharges are often to rehab facilities or with follow-up appointments, indicating a focus on recovery and continued care outside the hospital." (+3)

- **Rationale(Snippet)**: "Dataset 1 overwhelmingly describes patients receiving comfort measures only (CMO) and dying in the hospital. Key phrases include 'DNR/DNI,' 'comfort measures only,' 'expired peacefully,' 'withdrawal of pressors,' and 'poor prognosis.' The notes focus on end-of-life care decisions and the patient's final moments.2, in contrast, describes patients undergoing active treatment and diagnostic procedures. Notes mention specific interventions like 'pacer wire,' 'sphincterotomy,' 'cholecystectomy,' 'burr hole,' 'subdural drain,' and 'endoscopic harvesting.' The focus is on identifying the cause of the patient's condition and implementing treatments to improve their health. There is no mention of end-of-life care." (+3)

- **Rationale(Context)**: "Dataset 1 describes hospital stays primarily ending in patient death, often following decisions to make the patient comfort measures only (CMO) or do-not-resuscitate/intubate (DNR/DNI). The notes detail severe conditions like septic shock, pontine hemorrhage, and unrecoverable metabolic acidosis, with interventions often deemed futile. Dataset 2, in contrast, describes hospital stays related to gynecological procedures and conditions. The notes focus on post-operative recovery, pain management, antibiotic treatment for infections like pyosalpinx and endometritis, and successful discharge home. The overall tone is positive, emphasizing uncomplicated procedures and patient improvement." (+3)

Prompts for Obtaining Accuracy Scores Here is an example of how an external LLM grades an explanation to obtain the accuracy score used to evaluate the explanation's discriminative power, as introduced in Section 4.2.

USER: Based on the following distinction: {*actual generated shift explanation*}, respond with the notes most likely to be in dataset 1. Format as a list with no explanation: e.g. [1,3,5,8,9,10,..]:
Note 1: {*Actual note 1*}
Note 2: {*Actual note 2*}
... (omitted here due to length consideration)
Note 20: {*Actual note 20*}

Appendix F. Failure Examples

This appendix provides concrete examples of suboptimal explanations generated by the SIReNs methods to illustrate their respective failure modes. Here, we use the Procedure feature under the $0.6 \rightarrow 0.4$ shift magnitude, as it has the poorest performance in the of discriminative accuracy.

- **Difference of Means**: "Dataset 1 primarily describes post-operative care and recovery from acute injuries. The notes focus on surgical procedures, pain management, mobilization, and discharge planning, often involving specific surgical interventions and monitoring of physiological parameters. The patients generally progress towards functional recovery and are discharged with specific instructions and follow-up.2, in contrast, focuses on the management of complex medical conditions and psychiatric issues. The notes detail diagnostic workups, medication adjustments, and management of chronic conditions like cirrhosis, diabetes, and mental health disorders. There's a greater emphasis on identifying the underlying causes of symptoms, coordinating care with multiple specialists, and addressing psychosocial factors. The patients often have multiple comorbidities and require ongoing management and follow-up for complex medical and psychiatric needs."

- **Linear Probing**: "Dataset 1 appears to represent a broader range of medical specialties and patient conditions, including cardiology (TAVR, CAD), neurosurgery (SDH evacuation), trauma (renal laceration), and general medicine (hyperlipidemia, hypertension). The notes detail specific procedures, medication adjustments, and follow-up plans across these specialties.2 seems more focused on psychiatric and gynecological/urological cases. It includes detailed psychiatric assessments (borderline personality disorder, paranoia, substance abuse), surgical gynecology (hysterectomy, lymphadenectomy), and urology (ureteral stone extraction). The notes emphasize mental health treatment plans, surgical procedures, and post-operative recovery.summary, Dataset 1 is more diverse in medical specialties, while Dataset 2 is concentrated on psychiatric, gynecological, and urological conditions."

- **Rationale(Snippet)**: "Dataset 1 primarily describes patients undergoing surgical procedures. The notes detail pre-operative conditions (e.g., skull fracture, pancreatic mass), the procedures themselves (e.g., aortic valve replacement, neck dissection), and post-operative care (e.g., extubation, pain management, drain placement). The focus is on physical ailments and surgical interventions.2, in contrast, focuses on patients with psychiatric conditions. The notes describe symptoms of psychosis, mood disorders, substance use, suicidal ideation, and homicidal ideation. Treatment plans involve antipsychotic medications (e.g., olanzapine, risperidone) and therapeutic interventions. The emphasis is on mental health diagnoses and management."

- **Rationale(Context)**: "Dataset 1 primarily contains discharge summaries related to surgical and trauma cases. Common themes include post-operative recovery from procedures like CABG, valve replacements, and cancer resections, as well as management of traumatic injuries like SDH and skull fractures. The summaries detail surgical interventions, medication management, and progress in physical therapy.2 focuses on psychiatric admissions and management of chronic conditions. Summaries describe patients with diagnoses like schizophrenia, bipolar disorder, and substance use disorders. The notes detail mental status exams, medication adjustments (antipsychotics, mood stabilizers), and management of behavioral issues. There is also a note about atrial fibrillation management and another about diabetes and chronic diarrhea."

In these examples, the representative notes generated all referred to surgical procedures, but the LLM was overly specific in describing the differences, making them less useful as a discriminative explanation for the sources.

Appendix G. Theoretical Justification for Difference of Means

Suppose the underlying concept is $\boldsymbol{v} \in \{0,1\}^K$, where $K >> D$, and each v_i is independently distributed from concept probability $p_i \in [0,1]$. Further suppose the transformer embedding of a document $\boldsymbol{e} \in \mathbb{R}^D$ with underlying concept v comes from the product $\boldsymbol{e} = M\boldsymbol{v}$, with $M \in \mathbb{D} \times \mathbb{K}$. Then, if the transformer embeddings in dataset D_1 and D_2 come from underlying concept probability vectors $\boldsymbol{p}$, $\boldsymbol{q}$, where p and q differ only in dimension i, then the average difference between the datasets will be along the ith column of M, or m_i.

The matrix M is unseen, and therefore so is the column vector m_i. However, if we multiply m_i with a transformer embedding $\boldsymbol{e}' = M\boldsymbol{v}'$, the result is the product

$$\boldsymbol{m}_i^T \boldsymbol{e}' = \boldsymbol{m}_i^T M \boldsymbol{v}' = \sum_{j=1}^{D} v_j \boldsymbol{m}_i^T m_j$$

In the setting where the column vectors m_i are mutually near-orthogonal, $v_i m_i^T m_i$ comes to dominate this term. The expression denotes how much the concept i is expressed times a norm of the expression vector, which may not necessarily be 1.