Towards Trustworthy Explanation: On Causal Rationalization

Wenbo Zhang [1] **Tong Wu** [2] **Yunlong Wang** [2] **Yong Cai** [2] **Hengrui Cai** [1]

Abstract

With recent advances in natural language processing, rationalization becomes an essential self-explaining diagram to disentangle the black box by selecting a subset of input texts to account for the major variation in prediction. Yet, existing association-based approaches on rationalization cannot identify true rationales when two or more snippets are highly inter-correlated and thus provide a similar contribution to prediction accuracy, so-called spuriousness. To address this limitation, we novelly leverage two causal desiderata, non-spuriousness and efficiency, into rationalization from the causal inference perspective. We formally define a series of probabilities of causation based on a newly proposed structural causal model of rationalization, with its theoretical identification established as the main component of learning necessary and sufficient rationales. The superior performance of the proposed causal rationalization is demonstrated on real-world review and medical datasets with extensive experiments compared to state-of-the-art methods.

1. Introduction

Recent advancements in large language models have drawn increasing attention and have been widely used in extensive Natural Language Processing (NLP) tasks (see e.g., Vaswani et al., 2017; Kenton & Toutanova, 2019; Lewis et al., 2019; Brown et al., 2020). Although those deep learning-based models could provide incredibly outstanding prediction performance, it remains a daunting task in finding *trustworthy explanations* to interpret these models' behavior, which is particularly critical in high-stakes applications in various fields. In healthcare, the use of electronic health records (EHRs) with raw texts is increasingly common to forecast patients' disease progression and assist clinicians in making

decisions. These raw texts serve as abstracts or milestones of a patient's medical journey and characterize the patient's medical conditions (Estiri et al., 2021; Wu et al., 2021). Beyond simply predicting clinical outcomes, doctors are more interested in understanding the decision-making process of predictive models thereby building trust, as well as extracting clinically meaningful and relevant insights (Liu et al., 2015). Discovering faithful text information hence is particularly crucial for improving the early diagnostic of severe disease and for making efficient clinical decisions.

Disentangling the black box in deep NLP models, however, is a notoriously challenging task (Alvarez Melis & Jaakkola, 2018). There are a lot of research works focusing on providing trustworthy explanations for models, generally classified into post hoc techniques and self-explaining models (Danilevsky et al., 2020; Rajagopal et al., 2021). To provide better model interpretation, self-explaining models are of greater interest, and selective rationalization is one popular type of such a model by highlighting important tokens among input texts (see e.g., DeYoung et al., 2020; Paranjape et al., 2020; Jain et al., 2020; Antognini et al., 2021; Vafa et al., 2021; Chan et al., 2022). The general framework of selective rationalization as shown in Figure 1 consists of two components, a selector and a predictor. Those selected tokens by the trained selector are called *rationales* and they are required to provide similar prediction accuracy as the full input text based on the trained predictor. Besides, the selected rationales should reflect the model's true reasoning process (faithfulness) (DeYoung et al., 2020; Jain et al., 2020) and provide a convenient explanation to people (plausibility) (DeYoung et al., 2020; Chan et al., 2022). Most of the existing works (Lei et al., 2016; Bastings et al., 2019; Paranjape et al., 2020) found rationales by maximizing the prediction accuracy for the outcome of interest based on input texts, and thus are association-based models.

The major limitation of these association-based works (also see related works in Section 1.1) lies in falsely discovering *spurious* rationales that may be related to the outcome of interest but do not indeed cause the outcome. Specifically, when two or more snippets are highly intercorrelated and provide a similar high contribution to prediction accuracy, the association-based methods cannot identify the true rationales among them. Here, the true rationales, formally defined in Section 2 as the *causal rationales*, are the true

---

[1]Department of Statistics, University of California Irvine, California, USA [2]Advanced Analytics, IQVIA, Pennsylvania, USA. Correspondence to: Hengrui Cai <hengrc1@uci.edu>.

Figure 1: This figure shows a standard selective rationalization framework for a beer review, which can be seen as a select-predict pipeline where firstly rationales are selected and then fed into the predictor.
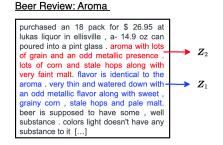

Figure 2: Motivating example from the Beer review data. Red highlights the text related to the aroma (denoted as Z_2) and blue highlights the text related to the palate (denoted as Z_1). The texts of aroma and palate are highly correlated with each other, which makes them indistinguishable in terms of predicting the sentiment of interest.

sufficient rationales to fully predict and explain the outcome without spurious information. As one example shown in Figure 2, one chunk of beer review comments covers two aspects, aroma, and palate. The reviewer has a negative sentiment toward the beer due to the unpleasant aroma (highlighted in red in Figure 2). While texts in terms of the palate as highlighted in blue can provide a similar prediction accuracy as it is highly correlated with the aroma but not cause the sentiment of interest. Thus, these two selected snippets can't be distinguished purely based on the association with the outcome when they all have high predictive power. Such troublesome spuriousness can be introduced in training as well owing to over-fitting to mislead the selector (Chang et al., 2020). The predictor relying on these spurious features fails to achieve high *generalization* performance when there is a large discrepancy between the training and testing data distributions (Schölkopf et al., 2021).

In this paper, we propose a novel approach called causal rationalization aiming to find trustworthy explanations for general NLP tasks. Beyond selecting rationales based on purely optimizing prediction performance, our goal is to identify rationales with causal meanings. To achieve this goal, we introduce a novel concept as causal rationales by considering two causal desiderata (Wang & Jordan, 2021): non-spuriousness and efficiency. Here, *non-spuriousness* means the selected rationales can capture features causally determining the outcome, and *efficiency* means only essential and no redundant features are chosen. Towards these causal desiderata, our main **contributions** are threefold.

● We first formally define a series of the probabilities of causation (POC) for rationales accounting for non-spuriousness and efficiency at different levels of language, based on a newly proposed structural causal model of rationalization.

● We systematically establish the theoretical results for identifications of the defined POC of rationalization at the individual token level - the conditional probability of necessity and sufficiency (CPNS) and derive the lower bound of CPNS under the relaxed identification assumptions for practical usage.

● To learn necessary and sufficient rationales, we propose a novel algorithm that utilizes the lower bound of CPNS as the criteria to select the causal rationales. More specifically, we add the lower bound as a causality constraint into the objective function and optimize the model in an end-to-end fashion, as shown in Figure 3.

With extensive experiments, the superior performance of our causality-based rationalization method is demonstrated in the NLP dataset under both out-of-distribution and in-distribution settings. The practical usefulness of our approach to providing trustworthy explanations for NLP tasks is demonstrated on real-world review and EHR datasets.

1.1. Related Work

Selective Rationalization. Selective rationalization was firstly introduced in Lei et al. (2016) and now becomes an important model for interpretability, especially in the NLP domain (Bao et al., 2018; Paranjape et al., 2020; Jain et al., 2020; Guerreiro & Martins, 2021; Vafa et al., 2021; Antognini & Faltings, 2021). To name a few recent developments, Yu et al. (2019) proposed an introspective model under a cooperative setting with a selector and a predictor. Chang et al. (2019) extended their model to extract class-wise explanations. During the training phase, the initial design of the framework was not end-to-end because the sampling from these selectors was not differentiable. To address this issue, some later works adopted differentiable sampling, like Gumbel-Softmax or other reparameterization tricks (see e.g., Bastings et al., 2019; Geng et al., 2020; Sha et al., 2021). To explicitly control the sparsity of the selected rationales, Paranjape et al. (2020) derived a sparsity-inducing objective by using the information bottleneck. Recently, Liu et al. (2022) developed a unified encoder to induce a better predictor by accessing valuable information blocked by the selector.
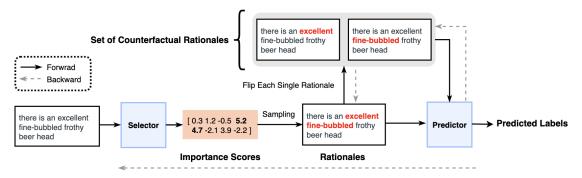
Figure 3: The framework of the proposed causal rationalization. Compared with the traditional selective rationalization in Figure 1, our method adds the causal component (highlighted in gray) to generate counterfactual rationales.

There are few recent studies that explored the issue of spuriousness in rationalization from the perspective of causal inference. Chang et al. (2020) proposed invariant causal rationales to avoid spurious correlation by considering data from multiple environments. Plyler et al. (2021) utilized counterfactuals produced in an unsupervised fashion using class-dependent generative models to address spuriousness. Yue et al. (2023) adopted a backdoor adjustment method to remove the spurious correlations in input and rationales. Our approach is notably different from the above methods. We only use a single environment, as opposed to multiple environments in Chang et al. (2020). This makes our method more applicable in real-world scenarios, where collecting data from different environments can be challenging. Additionally, our proposed CPNS regularizer offers a different perspective on handling spuriousness, leading to improved generalization in out-of-distribution scenarios. Our work differs from Plyler et al. (2021) and Yue et al. (2023) in that we focus on developing a regularizer (CPNS) to minimize spurious associations directly. This allows for a more straightforward and interpretable approach to the problem, which can be easily integrated into existing models.

Explainable Artificial Intelligence (XAI). Sufficiency and necessity can be regarded as the fundamentals of XAI since they are the building blocks of all successful explanations (Watson et al., 2021). Recently, many researchers have started to incorporate these properties into their models. Ribeiro et al. (2018) proposed to find features that are sufficient to preserve current classification results. Dhurandhar et al. (2018) developed an autoencoder framework to find pertinent positive (sufficient) and pertinent negative (non-necessary) features which can preserve the current results. Zhang et al. (2018) considered an approach to explain a neural network by generating minimal, stable, and symbolic corrections to change model outputs. Yet, sufficiency and necessity shown in the above methods are not defined from a causal perspective. Though there are a few works (Joshi et al., 2022; Balkir et al., 2022; Galhotra et al., 2021; Watson et al., 2021; Beckers, 2021) defining these two properties with causal interpretations, all these works focus on post hoc
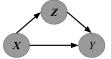

Figure 4: Causal diagram of rationalization. It describes the data-generating process for the text $\boldsymbol{X}$, the true rationales $\boldsymbol{Z}$, and the label Y. Solid arrows denote causal relationships.

analysis rather than a new model developed for rationalization. To the best of our knowledge, Wang & Jordan (2021) is the most relevant work that quantified sufficiency and necessity for high-dimensional representations by extending Pearl's POC (Pearl, 2000) and utilized those causal-inspired constraints to obtain a low-dimensional non-spurious and efficient representation. However, their method primarily assumed that the input variables are independent which allows a convenient estimation of their proposed POC. In our work, we generalize these causal concepts to rationalization and propose a more computationally efficient learning algorithm with relaxed assumptions.

2. Framework

Notations. Denote $\boldsymbol{X} = (X_1, \cdots, X_d)$ as the input text with d tokens, $\boldsymbol{Z} = (Z_1, \cdots, Z_d)$ as the corresponding selection where $Z_i \in \{0, 1\}$ indicates whether the i-th token is selected or not by the selector, and Y as the binary label of interest. Let $Y(\boldsymbol{Z} = \boldsymbol{z})$ denote the potential value of Y when setting $\boldsymbol{Z}$ as $\boldsymbol{z}$. Similarly, we define $Y(Z_i = z_i)$ as the potential outcome when setting the Z_i as z_i while keeping the rest of the selections unchanged.

Structural Causal Model for Rationalization. A structural causal model (SCM) (Schölkopf et al., 2021) is defined by a causal diagram (where nodes are variables and edges represent causal relationships between variables) and modeling of the variables in the graph. In this paper, we first propose an SCM for rationalization as follows with its causal diagram shown in Figure 4:

$$\boldsymbol{X} = f(N_X), \ \ \boldsymbol{Z} = g(\boldsymbol{X}, N_Z), \ \ Y = h(\boldsymbol{Z} \odot \boldsymbol{X}, N_Y), \quad (1)$$

where N_X, N_Y, N_Z are exogenous variables and f, g, h are unknown functions that represent the causal mechanisms
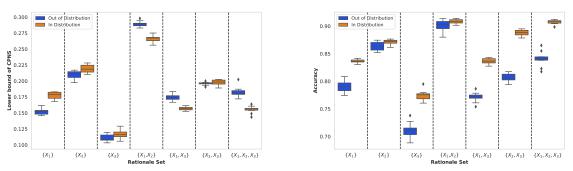
Figure 5: We evaluate the average lower bound of CPNS and the accuracy of simulated data under the in-distribution (ID) and the out-of-distribution (OOD) setting. (a) ID: True rationale set $\{X_1, X_2\}$ achieves the most accurate forecasting as the set $\{X_1, X_2, X_3\}$ with all variables but its highest CPNS score can help distinguish it from others. (b) OOD: $\{X_1, X_2\}$ doesn't only provide the most accurate predictions but also has the highest lower bound CPNS value among all the possible rationale sets, meaning that using CPNS can select the true rationale to achieve OOD generalization.

of $\boldsymbol{X}, \boldsymbol{Z}, Y$ respectively, with $\odot$ denoting the element-wise product. In this context, g and h can be regarded as the true selector and predictor respectively. Suppose we observe a data point with the text $\boldsymbol{X}$ and binary selections $\boldsymbol{Z}$, rationales can be represented by the event $\{\boldsymbol{X}_i \mathbb{I}(\boldsymbol{Z}_i = 1)\}_{1 \leq i \leq d}$, where $\mathbb{I}(\boldsymbol{Z}_i = 1)$ indicates if the i-th token is selected, $\boldsymbol{X}_i$ is the corresponding text, and d is the length of the text.

Remark 2.1. The data generation process in (1) matches many graphical models in previous work (see e.g., Chen et al., 2018; Paranjape et al., 2020). As a motivating example, consider the sentiment labeling process for the Beer review data. The labeler first locates all the important sub-sentences or words which encode sentiment information and marks them. After reading all the reviews, the labeler goes back to the previously marked text and makes the final judgment on sentiment. In this process, we can regard the first step of marking important locations of words as generating the selection of $\boldsymbol{Z}$ via reading texts $\boldsymbol{X}$. The second step is to combine the selected locations with raw text to generate rationales (equivalent to $\boldsymbol{Z} \odot \boldsymbol{X}$) and then the label Y is generated through a complex decision function h. Discussions of potential dependences in (1) are provided in Appendix B.

3. Probability of Causation for Rationales

In this section, we formally establish a series of the probabilities of causation (POC) for rationales accounting for non-spuriousness and efficiency at different levels of language, by extending Pearl (2000); Wang & Jordan (2021).

Definition 3.1. Probability of sufficiency (PS) for rationales:

$$\text{PS} \triangleq P\left(Y(\boldsymbol{Z} = \boldsymbol{z}) = y \mid \boldsymbol{Z} \neq \boldsymbol{z}, Y \neq y, \boldsymbol{X} = \boldsymbol{x}\right),$$

which indicates the sufficiency of rationales by evaluating the capacity of the rationales $\{\boldsymbol{X}_i \mathbb{I}(\boldsymbol{Z}_i = 1)\}_{1 \leq i \leq d}$ to "produce" the label if changing the selected rationales to the opposite. Spurious rationales shall have a low PS.

Definition 3.2. Probability of necessity (PN) for rationales:

$$\text{PN} \triangleq P\left(Y(\boldsymbol{Z} \neq \boldsymbol{z}) \neq y \mid \boldsymbol{Z} = \boldsymbol{z}, Y = y, \boldsymbol{X} = \boldsymbol{x}\right),$$

which is the probability of rationales $\{\boldsymbol{X}_i \mathbb{I}(\boldsymbol{Z}_i = 1)\}_{1 \leq i \leq d}$ being a necessary cause of the label $\mathbb{I}\{Y = y\}$. Non-spurious rationales shall have a high PN.

Desired true rationales should achieve non-spuriousness and efficiency simultaneously. This motivates us to define the probability of sufficiency and necessity as follows, with more explanations of these definitions in Appendix A.1.

Definition 3.3. Probability of necessity and sufficiency (PNS) for rationales:

$$\text{PNS} \triangleq P\left(Y(\boldsymbol{Z} \neq \boldsymbol{z}) \neq y, Y(\boldsymbol{Z} = \boldsymbol{z}) = y \mid \boldsymbol{X} = \boldsymbol{x}\right).$$

Here, PNS can be regarded as a good proxy of causality and we illustrate this in Appendix A.2. When both the selection $\boldsymbol{Z}$ and the label Y are univariate binary and $\boldsymbol{X}$ is removed, our defined PN, PS, and PNS boil-down into the classical definitions of POC in Definitions 9.2.1 to 9.2.3 of Pearl (2000), respectively. In addition, our definitions imply that the input texts are fixed and POC is mainly detected through interventions on selected rationales. This not only reflects the data generation process we proposed in Model (1) but also distinguishes our settings with existing works (Pearl, 2000; Wang & Jordan, 2021; Cai et al., 2023). To serve the role of guiding rationale selection, we extend the above definition to a conditional version for a single rationale.

Definition 3.4. Conditional probability of necessity and sufficiency (CPNS) for the jth selection:

$$\text{CPNS}_j \triangleq P(Y(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}) = y,$$
$$Y(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}) \neq y \mid \boldsymbol{X} = \boldsymbol{x}).$$

Definition 3.4 mainly focuses on a single rationale. The importance of Definitions 3.1-3.4 can be found in Appendix A.2. We further define CPNS over all the selected rationales as follows.

Definition 3.5. The CPNS over selected rationales:

$$\text{CPNS} \triangleq \prod_{j \in \mathbf{r}} (\text{CPNS}_j)^{\frac{1}{|\mathbf{r}|}},$$

where $\mathbf{r}$ is the index set for the selected rationales and $|\mathbf{r}|$ is the number of selected rationales.

The proposed CPNS can be regarded as a geometric mean and we discuss why utilizing this formulation in Appendix A.3, with more connections among Definitions 3.4-3.5 in Appendix A.4. To generalize CPNS to data, we can utilize the average CPNS over the dataset to represent an overall measurement. We denote the lower bound of CPNS and CPNS_j by $\underline{\text{CPNS}}$ and $\underline{\text{CPNS}}_j$, respectively, with their theoretical derivation and identification in Section 4.

3.1. Example of Using POC to Find Causal Rationales

We utilize a toy example to demonstrate why CPNS is useful for in-distribution (ID) feature selection and out-of-distribution (OOD) generalization. Suppose there is a dataset of sequences, where each sequence can be represented as $\boldsymbol{X} = (X_1, \ldots, X_l)$ with an equal length as $l = 3$ and a binary label Y. Here, we set $\{X_1, X_2\}$ are true rationales and $\{X_3\}$ is an irrelevant/spurious feature.

The process of generating such a dataset is described below. Firstly, we generate rationale features $\{X_1, X_2\}$ following a bivariate normal distribution with positive correlations. To create spurious correlation, we generate irrelevant features $\{X_3\}$ by using mapping functions which map $\{X_1, X_2, \epsilon\}$ to $\{X_3\}$ where $\epsilon \sim N(0, 0.5)$. Here we use a linear mapping function $X_3 = X_1 + \epsilon$. Thus irrelevant features are highly correlated with X_1.

There are 3 simulated datasets: the training dataset, the in-distribution test data, and the out-of-distribution test data. The training data $\{\boldsymbol{x}_i^{\text{train}}\}_{i=1}^{n_{\text{train}}}$ and the in-distribution test data $\{\boldsymbol{x}_i^{\text{test-in}}\}_{i=1}^{n_{\text{test-in}}}$ follows the same generation process, but for out-of-distribution test data $\{\boldsymbol{x}_i^{\text{test-out}}\}_{i=1}^{n_{\text{test-out}}}$, we modify the covariance matrix of X_1, X_2 to create a different distribution of the features. Then we make the label Y only depends on (X_1, X_2). This is equivalent to assuming that all the rationales are in the same position and the purpose is to simplify the explanations. For a single $\boldsymbol{x}_i$, we simulate $\text{P}(y_i = 1|\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i)$ below:

$$\pi(\boldsymbol{x}_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})}}.$$

Then we use threshold value 0.5 to categorize the data into one of two classes: $y_i = 1$ if $\pi \geq 0.5$ and $y_i = 0$ if $\pi \leq 0.5$. Since the dataset includes 3 features, there are 6 combinations of the rationale (we ignore the rationale containing no features). For ith rationale, we would fit a logistic regression model by using only selected features and

refit a new logistic regression model with subset features to calculate $\underline{\text{CPNS}}_i$. In our simulation, we set $n_{\text{train}} = n_{\text{test-in}} = n_{\text{test-out}} = 2000$, $\beta_0 = 1$, $\beta_1 = 0.5$ and $\beta_2 = 1$.

The results of the average of $\underline{\text{CPNS}}_1$ and $\underline{\text{CPNS}}_2$, and the accuracy measured in OOD and ID test datasets are shown in Figure 5 over 10 replications. It can be seen that true rationales ($\{X_1, X_2\}$) yield the highest scores of the average $\underline{\text{CPNS}}$ and accuracy in both OOD and ID settings. This motivates us to identify true rationales by maximizing the score of CPNS or the lower bound.

4. Identifiability and Lower Bound of CPNS

As has been shown, the proposed CPNS helps to discover the true rationales that achieve high OOD generalization. Yet, due to the unobserved counterfactual events in the observational study, we need to identify CPNS as statistically estimated quantities. To this end, we generalize three common assumptions in causal inference (Pearl, 2000; VanderWeele & Vansteelandt, 2009; Imai et al., 2010) to rationalization, including consistency, ignorability, and monotonicity.

Assumption 4.1.

- Consistency: $\boldsymbol{Z} = \boldsymbol{z} \to Y(\boldsymbol{Z} = \boldsymbol{z}) = Y.$

$$(2)$$

- Ignorability:

$$\{Y(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}),$$
$$Y(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j})\} \perp \boldsymbol{Z} \mid \boldsymbol{X}. \quad (3)$$

- Monotonicity (for $\boldsymbol{X} = \boldsymbol{x}$, Y is monotonic relative to $\boldsymbol{Z}_i$):

$$\{Y(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}) = y\}$$
$$\wedge \{Y(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}) \neq y\} = \text{False}, \quad (4)$$

where $\wedge$ is the logical operation AND. For two events A and B, $A \wedge B = \text{True}$ if $A = B = \text{True}$, $A \wedge B = \text{False}$ otherwise.

Here in the first assumption, the left-hand side means the observed selection of tokens to be $Z = z$, and the right-hand side means the actual label observed with the observed selection $Z = z$ is identical to the potential label we would have observed by setting the selection of tokens to be $Z = z$. The second assumption in causal inference usually means no unmeasured confounders, which is automatically satisfied under randomized trials. For observational studies, we rely on domain experts to include as many features as possible to guarantee this assumption. In rationalization, it means our text already contains all information. For monotonicity assumption, it indicates that a change on the wrong selection can not, under any circumstance, make Y change to the true label. In other words, true selection can increase the likelihood of the true label. The theorem below shows the identification and the partial identification results for CPNS.

Theorem 4.2. *Assume the causal diagram in Figure 4 holds.*
• *If assumptions* (2), (3), *and* (4) *hold, then* CPNS_j *can be identified by*

$$\begin{aligned} \text{CPNS}_j =\, & P(Y = y \mid Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}, \boldsymbol{X} = \boldsymbol{x}) \\ & - P(Y = y \mid Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}, \boldsymbol{X} = \boldsymbol{x}). \end{aligned}$$

• *If only assumptions* (2) *and* (3) *hold,* CPNS_j *is not identifiable but its lower bound can be calculated by*

$$\begin{aligned} \underline{\text{CPNS}}_j = \max \Big[& 0, P(Y = y \mid Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}, \boldsymbol{X} = \boldsymbol{x}) \\ & - P(Y = y \mid Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}, \boldsymbol{X} = \boldsymbol{x}) \Big]. \end{aligned}$$

The detailed proof is provided in Appendix H. Theorem 4.2 generalizes Theorem 9.2.14 and 9.2.10 of Pearl (2000) to multivariate binary variables. This is similar to the single binary variable case since for each single Z_j, the conditional event $\{\boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}, \boldsymbol{X} = \boldsymbol{x}\}$ doesn't change. The first part of the theorem provides identification results for the counterfactual quantity CPNS_j and we can estimate it using observational data and the flipping operation as shown in Figure 3. For example, given a piece of text $\boldsymbol{x}$, $P(Y = y \mid Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}, \boldsymbol{X} = \boldsymbol{x})$ can be estimated by feeding the original rationales $\boldsymbol{z}$ produced by the selector to the predictor, and $P(Y = y \mid Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}, \boldsymbol{X} = \boldsymbol{x})$ can be estimated by predicting the label of the counterfactual rationals which is obtained by flipping the z_j. We can notice that Theorem 4.2 can be generalized when z_i represents whether to mask a clause/sentence for finding clause/sentence-level rationales. Since the monotonicity assumption (4) is not always satisfied, especially during the model training stage. Based on Theorem 4.2, we can relax the monotonicity assumption and derive the lower bound of CPNS_j. Since a larger lower bound can imply higher CPNS_j but a larger upper bound doesn't, we focus on the lower bound here and utilize it as a substitution for CPNS_j. Combining each rationale, we can get the lower bound of CPNS as $\underline{\text{CPNS}} = \prod_{j \in \mathbf{r}} (\underline{\text{CPNS}}_j)^{\frac{1}{|\mathbf{r}|}}$.

5. Learning Necessary and Sufficient Rationale

In this section, we propose to learn necessary and sufficient rationales by incorporating $\underline{\text{CPNS}}$ as the causality constraint into the objective function.

5.1. Learning Architecture

Our model framework consists of a selector $g_\theta(\cdot)$ and a predictor $h_\phi(\cdot)$ as standard in the traditional rationalization approach, where θ and ϕ denote their parameters. We can get the selection $\boldsymbol{Z} = g_\theta(\boldsymbol{X})$ and fed it into predictor to get $Y = h_\phi(\boldsymbol{Z} \odot \boldsymbol{X})$ as shown in Figures 3 and 4. One main difference between causal rationale and original rationale is that we generate a series of counterfactual selections by flipping each dimension of the selection $\boldsymbol{Z}$

Algorithm 1 Causal Rationalization

Require: Training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, parameters k, α, μ and λ.

  Begin: Initialize the parameters of selector $g_\theta(\cdot)$ and predictor $h_\phi(\cdot)$, where θ and ϕ denote their parameters
  while not converge **do**
    Sample a batch $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $\mathcal{D}$
    Generate selections $\mathcal{S} = \{\boldsymbol{z}_i\}_{i=1}^n$ through Gumbel-Softmax sampling
    for $i = 1, \ldots, n$ **do**
      Get a random sample $\mathbf{r}_i^{(k)}$ from index set $\mathbf{r}_i$ where $\mathbf{r}_i$ represents the set of rationales that are selected as 1 in $\boldsymbol{z}_i$ and its size equals $k\% \times \text{length}(\mathbf{x}_i)$
      for $j = 1, \ldots, |\mathbf{r}_i^{(k)}|$ **do**
        Generate counterfactual selections $\boldsymbol{z}_{i(j)}$ by flipping the jth index of the index set $\mathbf{r}_i^{(k)}$
      end for
    end for
    Get a new batch of selections $\tilde{\mathcal{S}} = \{\boldsymbol{z}_{i(j)}\}_{j=1,\cdots,|\mathbf{r}_i^{(k)}|}^{i=1,\cdots,n}$ and set $\mathcal{S}_{\text{all}} = \mathcal{S} \bigcup \tilde{\mathcal{S}}$
    Compute $\mathcal{L}$ via Eq(5) by using $\mathcal{S}_{\text{all}}$ and $\mathcal{D}$
    Update parameters $\theta = \theta - \alpha \nabla_\theta \mathcal{L}$; $\phi = \phi - \alpha \nabla_\phi \mathcal{L}$
  end while
Output: selector $g_\theta(\cdot)$ and predictor $h_\phi(\cdot)$

---

we obtained from the selector. Then we feed raw rationales with new counterfactual rationales into our predictor to make predictions. Considering the considerable cost of obtaining reliable rationale annotations from humans, we only focus on unsupervised settings. Our goal is to make the selector select rationales with the property of necessity and sufficiency and our predictor can simultaneously provide accurate predictions given such rationales.

5.2. Role of POC and Its Estimation

For the j-th token to be selected as a rationale, according to the results of Theorem 4.2, we expect $P(Y = y \mid Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}, \boldsymbol{X} = \boldsymbol{x})$ to be large while $P(Y = y \mid Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}, \boldsymbol{X} = \boldsymbol{x})$ to be small. Here, these two probabilities can be estimated from the predictor using selected rationales and flipped rationales, respectively, through the deep model. The empirical estimation of $\underline{\text{CPNS}}_j$ is denoted as $\widehat{\underline{\text{CPNS}}}_j$. The estimated lower bound is used as the causal constraint to reflect the necessity and sufficiency of a token of determining the outcome. If $\widehat{\underline{\text{CPNS}}}_j$ is large, we expect the corresponding token to be selected into the final set of rationales.

5.3. Learning Necessary and Sufficient Rationales

Given a training dataset $\mathcal{D}$, we consider the following objective function utilizing lower bound of CPNS to train the causal rational model:

$$\min_{\theta,\phi} \mathcal{L} = \min_{\theta,\phi} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[L\left(y,\widehat{y}\right) + \lambda\delta\left(\boldsymbol{z}\right) - \underbrace{\mu \sum_{j\in\mathbf{r}^{(k)}} \frac{\log\widehat{\mathrm{CPNS}}_j^+}{|\mathbf{r}^{(k)}|}}_{\text{Causality Constraint}} \right],$$
(5)

where $\widehat{y} = h_\phi\left(\boldsymbol{z}\odot\boldsymbol{x}\right)$, $\boldsymbol{z} = g_\theta(\boldsymbol{x})$, $L(\cdot,\cdot)$ defined as the cross-entropy loss, $\delta(\cdot)$ is the sparsity penalty to control sparseness of rationales, $\mathbf{r}_i^{(k)}$ denotes the a random subset with size equal $k\%$ of the sequence length, λ and μ are the tuning parameters. The reason we sample a random subset $\mathbf{r}_i^{(k)}$ is due to the computational cost of flipping each selected rationale. To avoid the negative infinity value of $\log\widehat{\mathrm{CPNS}}_j^+$ when the estimated $\widehat{\mathrm{CPNS}}_j$ is 0, we add a small constant c to get $\widehat{\mathrm{CPNS}}_j^+ = \widehat{\mathrm{CPNS}}_j + c$. Since the value of c has no influence on the optimization of the objective function, we set $c = 1$. Our proposed algorithm to solve (5) is presented in Algorithm 1.

Remark 5.1. One of our motivations is from medical data. In those data, important rationales are not necessarily consecutive medical records, and can very likely scatter all over patient longitudinal medical journeys. However, continuity is a natural property in text data. Therefore, we first conduct the experiments without the continuity constraint in the next section and then conduct experiments with the continuity constraint in Appendix G.2.

6. Experiments and Results

6.1. Datasets

Beer Review Data. The Beer review dataset is a multi-aspect sentiment analysis dataset with sentence-level annotations (McAuley et al., 2012). Considering our approach focuses on the token-level selection and we don't use continuity constraint, we utilize the Beer dataset, with three aspects: appearance, aroma, and palate, collected from Bao et al. (2018) where token-level true rationales are given.

Hotel Review Data. The Hotel Review dataset is a form of multi-aspect sentiment analysis from Wang et al. (2010) and we mainly focus on the location aspect.

Geographic Atrophy (GA) Dataset. The proprietary GA dataset used in this study includes the medical claim records who are diagnosed with Geographic Atrophy or have risk factors. Each claim records the date, the ICD-10 codes of the medical service, where the codes represent different medical conditions and diseases, and the description of the service. We are tasked to utilize the medical claim data to find high-risk GA patients and reveal important clinical indications using the rationalization framework. This dataset doesn't provide human annotations because it requires a huge amount of time and money to hire domain experts to annotate such a large dataset.

6.2. Baselines, Implementations, and Metrics

Baselines. We consider five baselines: rationalizing neural prediction (RNP), variational information bottleneck (VIB), folded rationalization (FR), attention to rationalization (A2R), and invariant rationalization (INVART). RNP is the first select-predict rationalization approach proposed by Lei et al. (2016). VIB utilizes a discrete bottleneck objective to select the mask (Paranjape et al., 2020). FR doesn't follow a two-stage training for the generator and the predictor, instead, it utilizes a unified encoder to share information among the two components (Liu et al., 2022). A2R (Yu et al., 2021) combines both hard and soft selections to build a more robust generator. INVART (Chang et al., 2020) enables the predictor to be optimal across different environments to reduce spurious selections.

Implementations. We utilize the same sparse constraint in VIB, and thus the comparison between our method and VIB is also an ablation study to verify the usefulness of our causal module. For a fair comparison, all the methods don't include the continuous constraint. Following Paranjape et al. (2020), we utilize BERT-base as the backbone of the selector and predictor for all the methods for a fair comparison. We set the hyperparameter $\mu = 0.1$ and $k = 5\%$ for the causality constraint. See more details in Appendix D.

Metrics. For real data experiments of the Beer review dataset, we utilize prediction accuracy (Acc), Precision (P), Recall (R), and F1 score (F1), where P, R, and F1 are utilized to measure how selected rationales align with human annotations, and of the GA dataset, the area under curve (AUC) is used. For synthetic experiments of the Beer review dataset, we adopt Acc, P, R, F1, and False Discovery Rate (FDR) where FRD evaluates the percentage of injected noisy tokens captured by the model which have a spurious correlation with labels.

6.3. Real Data Experiments

We evaluate our method on three real datasets, Beer review, Hotel review, and GA data. For Beer and Hotel review data, based on summary in Appendix C Table 6, we select Top-10% tokens in the test stage and the way of choosing k during evaluation is in Appendix E. It is shown in Table 1 that our method achieves a consistently better performance than baselines in most metrics for Beer review data. Results for hotel review data are in Appendix G.3. Specifically, our method demonstrates a significant improvement over VIB which is an ablation study and indicates our causal component contribution to the superior performance.

We calculate empirical CPNS on the test dataset as shown in Figure 6. We find that our approach always obtains the highest values in three aspects that matches our expectation because one goal of our objective function is to maximize

Table 1: Results on the Beer review dataset. Top-10% tokens are selected for the test datasets.

Methods	Appearance				Aroma				Palate			
	Acc (↑)	P (↑)	R (↑)	F1 (↑)	Acc (↑)	P (↑)	R (↑)	F1 (↑)	Acc (↑)	P (↑)	R (↑)	F1 (↑)
VIB	92.2(1.1)	42.7(2.1)	21.6(1.9)	26.8(2.0)	83.5(1.5)	43.4(1.1)	25.6(2.2)	30.1(1.0)	74.3(2.4)	27.6(1.4)	24.1(3.5)	23.6(2.4)
RNP	91.0(1.1)	48.7(4.5)	11.7(0,9)	20.0(1.5)	82.4(0.8)	44.2(2.6)	20.7(3.2)	27.6(4.3)	74.0(0.9)	25.1(1.8)	21.9(2.0)	22.8(4.9)
FR	**94.5(0.6)**	37.7(1.9)	19.1(0.8)	23.7(1.1)	86.0(1.6)	40.3(2.5)	22.3(3.5)	28.0(1.9)	**77.0(1.9)**	25.8(1.0)	23.6(1.3)	22.4(1.6)
A2R	92.2(1.3)	**49.1(1.0)**	18.9(1.6)	25.9(0.9)	82.5(1.8)	51.2(2.7)	21.2(1.9)	29.8(2.4)	74.3(3.6)	31.8(2.5)	24.3(2.1)	25.4(2.0)
INVRAT	94.0(0.9)	45.8(1.4)	20.6(1.3)	26.7(1.6)	84.3(2.0)	43.0(4.5)	23.1(3.2)	28.3(4.0)	76.8(3.8)	28.7(2.2)	23.5(1.0)	23.8(2.0)
CR(ours)	93.1(1.1)	45.3(1.7)	**22.0(1.1)**	**28.0(1.2)**	**86.6(1.7)**	**60.3(3.0)**	**35.4(2.1)**	**39.0(4.1)**	75.7(3.2)	**32.5(2.8)**	**25.9(0.5)**	**26.5(1.2)**

Table 2: Results on the Spurious Beer review dataset. Top-10% tokens are highlighted for evaluation.

Methods	Aroma					Palate				
	Acc (↑)	P (↑)	R (↑)	F1 (↑)	FDR (↓)	Acc (↑)	P (↑)	R (↑)	F1 (↑)	FDR (↓)
VIB	80.2(1.8)	30.0(4.3)	17.2(3.1)	20.1(3.5)	97.6(5.3)	75.2(3.8)	26.8(4.5)	20.7(4.0)	21.6(3.8)	23.2(1.6)
RNP	79.0(1.0)	44.1(5.3)	12.3(1.0)	18.4(1.5)	97.4(3.0)	72.0(2.2)	22.8(2.6)	18.1(1.2)	18.6(1.6)	55.7(1.9)
FR	84.5(1.0)	38.7(3.7)	22.7(2.2)	27.0(2.7)	41.2(2.2)	76.2(3.5)	25.4(3.7)	22.0(4.0)	21.7(4.3)	11.6(3.1)
A2R	82.0(1.5)	44.6(3.4)	20.8(1.7)	26.8(3.2)	60.7(5.9)	73.5(3.8)	28.9(1.2)	21.8(3.2)	22.0(3.4)	39.7(3.3)
INVRAT	84.0(1.5)	42.0(2.6)	22.9(1.8)	27.9(2.4)	40.3(3.1)	**76.5(3.6)**	28.2(2.4)	22.5(2.7)	22.5(3.0)	7.2(1.7)
CR(ours)	**85.0(2.1)**	**55.3(3.1)**	**31.5(1.8)**	**37.8(2.1)**	37.9(0.7)	73.0(2.3)	**29.4(3.5)**	**22.6(3.8)**	**23.8(3.3)**	**3.0(1.1)**

the CPNS. This explains why our method has superior performance and indicates that CPNS is effective to find true rationales under the in-distribution setting. Examples of generated rationales are shown in Table 4 and more examples are provided in Appendix G.5.1. We also conduct sensitivity analyses for hyperparameters of causality constraints in Appendix G.5.2 and the results show the performance is insensitive to k and μ when μ is not too large.

For the GA dataset, as shown in Table 3, our method is slightly better than baselines in terms of prediction performance. We further examine the generated rationales on GA patients and observe that our causal rationalization could provide better clinically meaningful explanations, with visualized examples in Appendix G.5.2. This shows that CR can provide more trustworthy explanations for EHR data.

6.4. Synthetic Experiments

Beer-Spurious. We include spurious correlation into the Beer dataset by randomly appending spurious punctuation. We follow a similar setup in Chang et al. (2020) and Yu et al. (2021) to append punctuation "," and "." at the beginning of the first sentence with the following distributions:

P(append "," | Y = 1) = P(append "." | Y = 0) = α_1,

P(append "." | Y = 1) = P(append "," | Y = 0) = $1 - \alpha_1$.

Here we set $\alpha_1 = 0.8$. Intuitively, since the first sentence contains the appended punctuation with a strong spurious correlation, we expect the association-based rationalization approach to capture such a clue and our causality-based method can avoid selecting spurious tokens. Since for many review comments, the first sentence is usually about the appearance aspect, here we only utilize aroma and palate aspects as Yu et al. (2021). We inject tokens into the training and validation set, then we can evaluate OOD performance on the unchanged test set. See more details in Appendix F.

Table 3: Results on the GA data. Top-5% tokens are highlighted for evaluation.

	CR (ours)	RNP	VIB	FR
AUC (↑)	84.3(±0.2)	83.3(±0.8)	84.00(±0.5)	84.2(±1.0)

Figure 6: Estimated lower bound of CPNS of ours and baseline methods on the test set of the Beer review data.

From Table 2, our causal rationalization outperforms baseline methods in most aspects and metrics. Especially, our method selects only 37.9% and 3.0% injected tokens on the validation set for two aspects. RNP and VIB don't align well with human annotations and have low prediction accuracy because they always select spurious tokens. We notice it's harder to avoid selecting spurious tokens for the aroma aspect than the palate. Our method is more robust when handling spurious correlation and shows better generalization performance, which indicates CPNS can help identify true rationales under an out-of-distribution setting.

Table 4: Examples of generated rationales. Human annotations are <u>underlined</u> and rationales obtained from ours and baseline methods are highlighted by different colors. We find that the rationales of our method better align with human annotations.

CR	VIB	RNP	FR
Aspect: Appearance **Label**: Positive **Pred**: Positive poured from bottle into shaker pint glass . a : pale tan-yellow with very thin white head that quickly disappears . s : malt , caramel ... and skunk t : very bad . hint of malt . mostly just tastes bad though . m : acidic , thin , and watery d : not drinkable . unbelievably bad . stay away ... you have several thousand other beers that taste better to spend you money on . do n't be a fool like me .	**Aspect**: Appearance **Label**: Positive **Pred**: Positive poured from bottle into shaker pint glass . a : pale tan-yellow with very thin white head that quickly disappears . s : malt , caramel ... and skunk t : very bad . hint of malt . mostly just tastes bad though . m : acidic , thin , and watery d : not drinkable . unbelievably bad . stay away ... you have several thousand other beers that taste better to spend you money on . do n't be a fool like me .	**Aspect**: Appearance **Label**: Positive **Pred**: Positive poured from bottle into shaker pint glass . a : pale tan-yellow with very thin white head that quickly disappears . s : malt , caramel ... and skunk t : very bad . hint of malt . mostly just tastes bad though . m : acidic , thin , and watery d : not drinkable . unbelievably bad . stay away ... you have several thousand other beers that taste better to spend you money on . do n't be a fool like me .	**Aspect**: Appearance **Label**: Positive **Pred**: Positive poured from bottle into shaker pint glass . a : pale tan-yellow with very thin white head that quickly disappears . s : malt , caramel ... and skunk t : very bad . hint of malt . mostly just tastes bad though . m : acidic , thin , and watery d : not drinkable . unbelievably bad . stay away ... you have several thousand other beers that taste better to spend you money on . do n't be a fool like me .

Figure 7: The lower bound of CPNS with different highlighted lengths during test.

We evaluate out-of-distribution performance on the unchanged testing set, so the results in Table 2 do suggest that FR is better at avoiding the selection of spurious tokens than VIB and RNP under noise-injected scenarios and has better generalization. The results that VIB is better than FR in Table 1 don't conflict with the previous finding because Table 1 evaluates in-distribution learning ability and VIB can be regarded as mo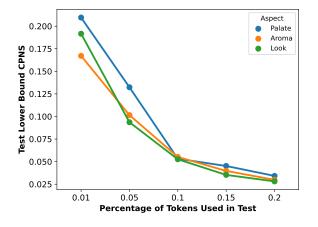re accurately extracting rationales without distribution shift. Figure 6 presents an estimated lower bound of the CPNS for our method and the baseline methods on the test set of the Beer review data. While it is true that FR has the lowest CPNS in Figure 6 this does not negate the importance of CPNS. It is crucial to note that the CPNS value estimates how well a model extracts necessary and sufficient rationales, but it is not the sole indicator of a model's overall performance. The primary goal of FR is to avoid selecting spurious tokens, which is evidenced by its lower FDR in Table 2, demonstrating its better generalization under noise-injected scenarios. We can argue that the discrepancy between CPNS and FDR for FR may arise due to the fact that FR is more conservative in selecting tokens

as rationales. This may lead to a lower CPNS, as FR may miss some necessary tokens, but at the same time, it avoids selecting spurious tokens, thus resulting in a lower FDR.

6.5. How Highlighted Length Influence CPNS

We want to know how highlighted length during evaluation can influence CPNS. We evaluate CPNS for three aspects with different percentages of tokens from $\{1\%, 5\%, 10\%, 15\%, 20\%\}$ as shown in Figure 7. It can be seen that as the highlighted length increase, the estimated value would decrease, which matches our expectations. For CPNS, it consists of two types of conditional probabilities $P(Y = y \mid Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}, \boldsymbol{X} = \boldsymbol{x})$ and $-P(Y = y \mid Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}, \boldsymbol{X} = \boldsymbol{x})$. If more events are given, namely the dimension of $\boldsymbol{Z}_{-j}$ increases, a flip of a single selection Z_j would bring less information, and hence the difference between two probabilities would decrease. The results indicate that we should always describe and compare CPNS of rationalization approaches using the same highlighted length.

7. Conclusion

This work proposes a novel rationalization approach to find causal interpretations for sentiment analysis tasks and clinical information extraction. We formally define the non-spuriousness and efficiency of rationales from a causal inference perspective and propose a practically useful algorithm. Moreover, we show the superior performance of our causality-based rationalization compared to state-of-the-art methods. The main limitation of our method is that CPNS is defined on the token-level and the computational cost is high when there are many tokens hence our method is not scalable well to long-text data. In future work, an interesting direction would be to define CPNS on the clause/sentence-level rationales. This would not only make the computation more feasible but also extracts higher-level units of meaning which improve the interpretability of the model's decisions.

References

Alvarez Melis, D. and Jaakkola, T. Towards robust interpretability with self-explaining neural networks. Advances in neural information processing systems, 31, 2018.

Antognini, D. and Faltings, B. Rationalization through concepts. arXiv preprint arXiv:2105.04837, 2021.

Antognini, D., Musat, C., and Faltings, B. Multi-dimensional explanation of target variables from documents. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pp. 12507–12515, 2021.

Balkir, E., Nejadgholi, I., Fraser, K., and Kiritchenko, S. Necessity and sufficiency for explaining text classifiers: A case study in hate speech detection. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2672–2686, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022. naacl-main.192. URL https://aclanthology.org/2022.naacl-main.192.

Bao, Y., Chang, S., Yu, M., and Barzilay, R. Deriving machine attention from human rationales. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1903–1913, 2018.

Bastings, J., Aziz, W., and Titov, I. Interpretable neural predictions with differentiable binary variables. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2963–2977, 2019.

Beckers, S. Causal explanations and xai. In First Conference on Causal Learning and Reasoning, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. Advances in neural information processing systems, 33: 1877–1901, 2020.

Cai, H., Wang, Y., Jordan, M., and Song, R. On learning necessary and sufficient causal graphs. arXiv preprint arXiv:2301.12389, 2023.

Chan, A., Sanjabi, M., Mathias, L., Tan, L., Nie, S., Peng, X., Ren, X., and Firooz, H. Unirex: A unified learning framework for language model rationale extraction. In International Conference on Machine Learning, pp. 2867–2889. PMLR, 2022.

Chang, S., Zhang, Y., Yu, M., and Jaakkola, T. A game theoretic approach to class-wise selective rationalization. Advances in neural information processing systems, 32, 2019.

Chang, S., Zhang, Y., Yu, M., and Jaakkola, T. Invariant rationalization. In International Conference on Machine Learning, pp. 1448–1458. PMLR, 2020.

Chen, H., He, J., Narasimhan, K., and Chen, D. Can rationalization improve robustness? In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3792–3805. Association for Computational Linguistics, 2022.

Chen, J., Song, L., Wainwright, M., and Jordan, M. Learning to explain: An information-theoretic perspective on model interpretation. In International Conference on Machine Learning, pp. 883–892. PMLR, 2018.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., and Sen, P. A survey of the state of explainable ai for natural language processing. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pp. 447–459, 2020.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., and Wallace, B. C. Eraser: A benchmark to evaluate rationalized nlp models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4443–4458, 2020.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. Advances in neural information processing systems, 31, 2018.

Estiri, H., Strasser, Z. H., and Murphy, S. N. High-throughput phenotyping with temporal sequences. Journal of the American Medical Informatics Association, 28(4):772–781, 2021.

Galhotra, S., Pradhan, R., and Salimi, B. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In Proceedings of the 2021 International Conference on Management of Data, pp. 577–590, 2021.

Geng, X., Wang, L., Wang, X., Qin, B., Liu, T., and Tu, Z. How does selective mechanism improve self-attention networks? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2986–2995, 2020.

Guerreiro, N. M. and Martins, A. F. Spectra: Sparse structured text rationalization. arXiv preprint arXiv:2109.04552, 2021.

Imai, K., Keele, L., and Tingley, D. A general approach to causal mediation analysis. Psychological methods, 15(4): 309, 2010.

Jain, S., Wiegreffe, S., Pinter, Y., and Wallace, B. C. Learning to faithfully rationalize by construction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4459–4473, 2020.

Joshi, N., Pan, X., and He, H. Are all spurious features in natural language alike? an analysis through a causal lens. arXiv preprint arXiv:2210.14011, 2022.

Kenton, J. D. M.-W. C. and Toutanova, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pp. 4171–4186, 2019.

Lei, T., Barzilay, R., and Jaakkola, T. Rationalizing neural predictions. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 107–117, 2016.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461, 2019.

Liu, C., Wang, F., Hu, J., and Xiong, H. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 705–714, 2015.

Liu, W., Wang, H., Wang, J., Li, R., Yue, C., and Zhang, Y. FR: Folded rationalization with a unified encoder. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), Advances in Neural Information Processing Systems, 2022. URL https://openreview.net/forum?id=ZPyKSBaKkiO.

McAuley, J., Leskovec, J., and Jurafsky, D. Learning attitudes and attributes from multi-aspect reviews. In 2012 IEEE 12th International Conference on Data Mining, pp. 1020–1025. IEEE, 2012.

Paranjape, B., Joshi, M., Thickstun, J., Hajishirzi, H., and Zettlemoyer, L. An information bottleneck approach for controlling conciseness in rationale extraction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1938–1952, 2020.

Pearl, J. Causality: models, reasoning, and inference. 2000.

Plyler, M., Green, M., and Chi, M. Making a (counterfactual) difference one rationale at a time. Advances in Neural Information Processing Systems, 34:28701–28713, 2021.

Rajagopal, D., Balachandran, V., Hovy, E. H., and Tsvetkov, Y. Selfexplain: A self-explaining architecture for neural text classifiers. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 836–850, 2021.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. Proceedings of the IEEE, 109 (5):612–634, 2021.

Sha, L., Camburu, O.-M., and Lukasiewicz, T. Learning from the best: Rationalizing predictions by adversarial information calibration. In AAAI, pp. 13771–13779, 2021.

Vafa, K., Deng, Y., Blei, D., and Rush, A. M. Rationales for sequential predictions. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 10314–10332, 2021.

VanderWeele, T. J. and Vansteelandt, S. Conceptual issues concerning mediation, interventions and composition. Statistics and its Interface, 2(4):457–468, 2009.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, H., Lu, Y., and Zhai, C. Latent aspect rating analysis on review text data: a rating regression approach. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 783–792, 2010.

Wang, Y. and Jordan, M. I. Desiderata for representation learning: A causal perspective. arXiv preprint arXiv:2109.03795, 2021.

Watson, D. S., Gultchin, L., Taly, A., and Floridi, L. Local explanations via necessity and sufficiency: Unifying theory and practice. In Uncertainty in Artificial Intelligence, pp. 1382–1392. PMLR, 2021.

Wu, T., Wang, Y., Wang, Y., Zhao, E., and Wang, G. Oamedsql: Order-aware medical sequence learning for clinical outcome prediction. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1585–1589. IEEE, 2021.

Yu, M., Chang, S., Zhang, Y., and Jaakkola, T. Rethinking cooperative rationalization: Introspective extraction and complement control. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4094–4103, 2019.

Yu, M., Zhang, Y., Chang, S., and Jaakkola, T. Understanding interlocking dynamics of cooperative rationalization. Advances in Neural Information Processing Systems, 34: 12822–12835, 2021.

Yue, L., Liu, Q., Wang, L., An, Y., Du, Y., and Huang, Z. Interventional rationalization, 2023. URL https://openreview.net/forum?id=KoEa6h1o6D1.

Zhang, X., Solar-Lezama, A., and Singh, R. Interpreting neural network judgments via minimal, stable, and symbolic corrections. Advances in neural information processing systems, 31, 2018.

A. More Details of Probability of Causation for Rationales

A.1. More Details of Definitions 3.1-3.3

The probability of sufficiency (PS) for a binary label Y and a binary feature Z is defined as $P\{Y(Z = True) = True | Y = False, Z = False)$, which means given the fact that we observed the false label Y with a false feature Z, what is the probability of the label turning to be true if we have had the chance to set the feature to be true. This probability thus describes the sufficiency of the feature Z to be true to obtain a true label. Now, moving to our Definition 3.1 proposed for the rationales, it means the probability of $Y = y$ when changing the selection to be $Z = z$ given text $X = x$ with the already observed selection $Z \neq z$ and the label $Y \neq y$. In other words, PS gives the probability of setting z would produce y in a situation where z and y are in fact absent given x. It describes the capacity of the rationales to "produce" the label. On the other hand, the probability of necessity (PN) for a binary label and a binary feature is $P\{Y(Z = False) = False | Y = True, Z = True)$, which means given the fact that we observed the true label Y with a true feature Z, what is the probability of the label turning to be false if we have had the chance to set the feature to be false. This probability thus describes the necessity of the feature Z to be true to obtain a true label. Following a similar logic, Definition 3.2 generalizes the PN score in the bivariate setting for rationales to describe the selection Z being a necessary cause of the label. Combining PN and PS yields the probability of necessity and sufficiency (PNS) for rationales in Definition 3.3 to comprehensively characterize the importance of rationales in causally determining the label. Those defined counterfactual quantities provide a formal way of measuring whether one event A is the necessary or sufficient cause of another event B.

A.2. Significance of Definitions 3.1-3.4 and Why PNS Means Causality

The significance of all the proposed definitions of the probability of causation (POC) is three-fold. First, we generalized the classical definitions in Pearl (2000) (where they only consider one binary outcome and one binary feature) for rationales to allow a selection of words as the feature input. Secondly, we align POC with the newly proposed structural causal model for rationalization, with additional conditioning on the texts. Our definitions imply that the input texts are fixed and thus POC is mainly detected through interventions on selected rationales. Lastly, to accommodate the task of rationalization, we further define $CPNS_j$ for the j-th selected token in Definition 3.4, which allows us to simultaneously quantify the sufficiency and necessity of an individual rationale.

We also illustrate why PNS Means Causality. First of all, we regard the underlying labeling process as a structure causal model shown in Figure 4, and thus the true selection Z can be seen as the cause of the label. Changing the underlying true Z selection to the opposite should also change the value of the label accordingly. This is aligned with the definition of PNS and thus it represents the causality of how the selected rationales determine the label. If a selection of tokens is necessary and sufficient causes the label, it should have a high PNS to reflect its high necessity and sufficiency in determining the label. We expect our rationalization approach can capture the underlying causal selection, so that's why we focus on optimizing PNS and CPNS.

A.3. Discussion of Geometric Average in Definition 3.5

We use the geometric mean because we want CPNS over selected rationales in Definition 3.5 to be a likelihood, which is a product of all individual rationales' CPNS scores in the selection. Yet, such a simple product is not ideal owing to the heterogeneous length of the texts. As the length increases, the number of selected rationales increases (because it's $k\%$ of raw text length), and the product would be lower by noting the probability ranging from 0 to 1. This is undesirable because we don't want to text length to be an influencing factor. Hence, we normalize the likelihood, leading to the geometric mean as presented in the current paper.

A.4. Clarification on Token-level CPNS and Reasoning

We propose Definition 3.4 to assess the causality of each token, which can be estimated by comparing the conditional probability of the label with and without this token (two counterfactual realities), as stated in Theorem 4.2. Therefore, to assess the causality of the rationale (i.e., a selection of tokens), the most natural way is to define the CPNS directly for this rationale by comparing the conditional probability of the label with the current selection versus that with counterfactual selections. However, for a selection consisting of r tokens, the counterfactual realities yield $2^r - 1$, which leads to a computational challenge when r is large. To overcome this issue, we alternatively propose Definition 3.5 to combine all

individual-level CPNS scores in the selection to reflect the causality of the rationale, which instead yields a polynomial computational complexity of $\mathcal{O}(r)$. Suppose the spillover effects (i.e., the words have dependence and the effect of the joint of words cannot simply be written as the sum of effects of single words) are uniform across words and sentences, with the Geometric Average in Definition 3.5, then we have the combined CPNS to approximate the CPNS of this rationale. Thus, this score helps to determine the causality of the rationale. When there is no spillover effect, Definition 3.4 can be viewed as the CPNS of a rationale directly. Notably, we **do not calculate** $CPNS_j$ for all tokens. Rather, we focus on a selection of tokens indexed by r, where the index set r can be determined by the selector g_θ. The demonstration of the toy example mainly explains why CPNS is useful in identifying true causes in the presence of spurious features.

B. Structural Causal Model for Rationalization under Potential Dependences

We firstly clarify that according to Figure 4 and the proposed structural causal model for rationalization, the entire rationale as a selection of true important tokens $\mathbf{Z}$ is determined by texts $\mathbf{X}$, however, we do allow exogenous variables N_Z to model possible dependence among the elements of $\mathbf{Z}$ (i.e., z_i). We understand that in practice, the rationale selection process may involve sequential labeling, where the selection of one token can influence the subsequent selections.

Impact of Potential Dependence on Our Method. To address the concern of whether the potential dependence or sequential labeling would affect our method, we argue that our method is still valid, but with conditions. Specifically, recall that we propose Definition 3.4 to assess the causality of each token, which can be estimated by comparing the conditional probability of the label with and without this token (two counterfactual realities), as stated in Theorem 4.2. Therefore, to assess the causality of the rationale (i.e., a selection of tokens), the most natural way is to define the CPNS directly for this rationale by comparing the conditional probability of the label with the current selection versus that with counterfactual selections. Such a definition would allow possible dependence among the elements of $\mathbf{Z}$ (i.e., z_i).

However, for a selection consisting of r tokens, the counterfactual realities yield $2^r - 1$, which leads to a computational challenge when r is large. To overcome this issue, we alternatively propose Definition 3.5 to combine all individual-level CPNS scores in the selection to reflect the causality of the rationale, which instead yields a polynomial computational complexity of $\mathcal{O}(r)$. The proposed alternative approach is valid under potential dependence including the following cases: 1. Suppose the spillover effects (i.e., the words have dependence and the effect of the joint of words cannot simply be written as the sum of effects of single words) are uniform across words and sentences, with the Geometric Average in Definition 3.5, then we have the combined CPNS to approximate the CPNS of this rationale. Thus, this score helps to determine the causality of the rationale. 2. When there is no spillover effect, Definition 3.5 can be viewed as the CPNS of a rationale directly. 3. In addition, suppose there exists conditional independence among words (considered in Joshi et al. (2022)), our proposed combined CPNS is equivalent to the CPNS of a rationale as well.

In conclusion, our method can still be applicable under the presence of potential dependence or sequential labeling, with certain conditions on the spillover effects of the dependency. Our approach to assessing the causality of rationales using the combined CPNS allows us to account for possible dependencies among the elements of $\mathbf{Z}$ while maintaining computational efficiency.

C. Data Prepossessing and Summary

Beer Review Data. We use the publicly available version of the Beer review dataset also adopted by Bao et al. (2018) and Chen et al. (2022). This dataset is cleaned by the previous authors and is a subset of the raw BeerAdvocate review dataset (McAuley et al., 2012). Following the same evaluation protocol of some previous works (see e.g., Bao et al., 2018; Yu et al., 2019; Chang et al., 2020; Chen et al., 2022), we convert their original scores which are in the scale of $[0, 1]$ into binary labels. Specifically, reviews with ratings ≤ 0.4 are labeled as negative and those with ≥ 0.6 are labeled as positive. We follow the same train/validation/test split as Chen et al. (2022) and it is summarized in Table 5. To make computation more feasible, except for the raw dataset, we create a short-text version of the dataset by filtering the texts over a length of 120. Table 6 summarizes the statistics of the Beer review dataset.

Hotel Review Data. The Hotel review data we used were first proposed by Wang et al. (2010) and we adopted processed one from Bao et al. (2018). Table 7 summarizes the statistics of the location aspect of Hotel review dataset.

GA Data. The proprietary GA dataset used in this study includes the medical claim records (diagnosis, prescriptions, and procedures) of 329,023 patients who are diagnosed as GA from 2018 to 2021 in the US, as well as those of additional

Table 5: The split of the dataset.

Short	Train	Val	Test
Beer (Appearance)	15932	3757	200
Beer (Aroma)	14085	2928	200
Beer (Palate)	9592	2294	200

Table 6: Dataset details, with rationale length ratios included for datasets where they are available.

Short	Len	Rationale(%)	N
Beer (Appearance)	88.94	19.2	19889
Beer (Aroma)	89.92	15.9	17033
Beer (Palate)	90.72	12.7	12086

Table 7: Dataset details for Hotel review data, with rationale length ratios included for datasets where they are available.

Aspect	Train	Val	Test	Len	Rationale(%)
Location	14472	1883	200	708.7	10.3

991,946 patients who have at least one of GA risk factors. Since patients have long medical sequences over years, here we only extract their most recent two years' visit. Then we further select patients with sequence lengths between 100 and 150. Finally, we sample 10000 positive patients and 30000 negative patients from the cohort to construct our dataset and Table 8 illustrates the division of the data into training, validation, and test sets.

Table 8: The split of the dataset.

Short	Train	Val	Test	Total	Len
GA	20000	10000	10000	40000	122.31

D. Implementation Details

For the Beer review data, we use two BERT-base-uncased as the selector and the predictor components for rationalization approaches. Those modules are initialized with pre-trained Bert.

A few challenges rise prior to directly including INVRAT for a fair comparison. Firstly, as discussed in the related work, INVRAT relies on multiple environments while our method and other baselines focus on a single environment. Second, their method highly depends on the construction of such multiple environments, yet, there is no principled guidance on how to select environments for INVRAT. Third, Chang et al. (2020) trained a simple linear regression model to predict the rating of the target aspect given the ratings of all the other aspects to generate the environments. This dataset is not publicly available. In contrast, for the public Beer dataset we used, a comment for one aspect doesn't have scores for other aspects, which means we can't simply utilize the experimental design of Chang et al. (2020) to select suitable environments. To address those issues, firstly we impute the missing scores based on aspect-specific prediction models which are trained on text data with their single provided aspect and then we follow the same way as Chang et al. (2020) to select environments.

For the GA data, we use the same architecture and the only difference is we replace the word embedding matrix with a randomly initialized health diagnosis code embedding and the embedding is trained jointly with other modules. For all experiments, we utilize a batch size of 256 and choose the learning rate $\alpha \in \{1e-5, 5e-4, 1e-4\}$. We train for 10 epochs all the datasets. For training the causal component, we tune the values of the Lagrangian multiplier $\mu \in \{0.01, 0.1, 1\}$ and set $k = 5$. We set the temperature of Gumbel-softmax to be 0.5. For our final evaluation, we highlight Top-10% tokens as the rationales for the Beer review data and Top-5% for the GA data. We conduct our experiments over 5 random seeds and calculate the mean and standard deviation of metrics. All of our experiments are conducted with PyTorch on 4 V100 GPU.

Our code is publicly available online.[1]

E. How to Select $k\%$ during Evaluation

Rationale for choosing $k = 10\%$ for the Beer review dataset: As shown in Appendix C Table 6, the true rationales in the Beer review dataset are between 10% and 20% of the total tokens. Therefore, we choose $k = 10\%$ as a reasonable threshold to represent the ground truth for this dataset, ensuring that we capture a significant portion of the true important tokens in our evaluation.

Rationale for choosing $k = 5\%$ for the GA dataset: The GA dataset consists of medical claim data, and a substantial portion of records (around 10%) are related to administrative and billing purposes, for example, codes for office visits or inpatient/outpatient admin records. These records offer limited insights into patients' disease progression and are less relevant as rationales. Given the smaller pool of meaningful rationales in the GA dataset compared to the Beer review dataset, we set a lower threshold ratio ($k = 5\%$) for this dataset.

In conclusion, the choice of different top token ratios for the Beer review dataset and GA dataset in Experiment 6.3 is based on the characteristics and ground truth of each dataset. Our aim is to ensure a fair evaluation of the models' performance in extracting meaningful rationales from the texts while taking into account the specific context and content of each dataset.

F. More Elaboration on Spurious Experiments

Punctuations as An OOD Scenario. While it is true that adding punctuation tokens does not change the meaning of the sentence, the distribution of spurious tokens is changed. The objective of the experiment in Section 6.4 is to evaluate the model's generalization under different conditions. By injecting punctuation tokens based on the label (as described in Section 6.4), we introduce spurious correlations that the model may exploit during training. These spurious tokens can be regarded as short-cuts that can potentially mislead rationalization methods. With and without punctuation are two scenarios representing whether short-cuts exit or not, we consider this as an OOD setting.

Clarification on Training/Validation/Test Data. In this experiment, we add punctuation tokens only to the training and validation data, keeping the test data unchanged. This setup allows us to examine the models' ability to generalize in an OOD setting, where the distribution of spurious tokens in the training and validation data is different from that in the test data. By keeping the test data free of injected punctuations, we can evaluate how well the models perform when faced with a scenario where the short-cuts present in the training and validation data are absent in the test data.

G. Experimental Results

G.1. Results Analyses

In conclusion, the experimental results demonstrate that FR is indeed better at avoiding the selection of spurious tokens and has better generalization under noise-injected scenarios. The differences between the findings in Tables 1, Table 2, and Figure 6 highlight the distinct evaluation scenarios (in-distribution vs. out-of-distribution) and emphasize the importance of considering multiple performance metrics (F1, FDR, and CPNS) to obtain a comprehensive understanding of the models' behavior.

G.2. Beer Review Results after Adding Continuous Constarint

From Table 9, we observe that our method, when applied with the continuity constraint, continues to perform well, suggesting that the continuity constraint does not negatively impact our method's effectiveness.

G.3. Hotel Review Results

Since Hotel review data have fewer continuous rationales, we compare all the baseline methods with CR without the continuity constraint. We don't include INVART because for Hotel review data, it doesn't has continuous scores and we can't follow the same way of selecting environments as Beer review data. From Table 10, upon conducting our analysis, we have observed that our approach outperforms other baseline methods in terms of capturing human annotations. Based on

---

[1]https://github.com/onepounchman/Causal-Retionalization.

Table 9: Results on the Beer review dataset after adding continuous constraint. Top-10% tokens are selected for the test datasets. Causal rationalization performs the best in all aspects in terms of capturing human annotations.

Methods	Appearance				Aroma				Palate			
	Acc (↑)	P (↑)	R (↑)	F1 (↑)	Acc (↑)	P (↑)	R (↑)	F1 (↑)	Acc (↑)	P (↑)	R (↑)	F1 (↑)
VIB	**93.8(2.4)**	52.6(2.0)	26.0(2.3)	32.9(2.1)	85.0(1.1)	54.2(2.9)	31.6(1.9)	37.7(2.8)	81.5(3.2)	41.2(2.1)	35.1(3.0)	35.2(2.8)
RNP	91.5(1.7)	40.0(1.4)	20.3(1.9)	25.2(1.7)	84.0(2.1)	49.1(3.2)	28.7(2.2)	32.0(2.5)	80.3(3.4)	38.6(1.8)	31.1(2.3)	29.7(2.0)
FR	93.5(1.0)	51.9(1.1)	25.1(2.0)	31.8(1.6)	88.0(1.8)	54.8(3.5)	33.7(2.6)	39.5(3.7)	82.0(2.1)	44.3(3.4)	32.5(2.7)	33.7(3.1)
A2R	91.5(2.2)	55.0(0.8)	25.8(1.6)	34.3(1.4)	85.5(1.9)	61.3(2.8)	34.8(3.1)	41.2(3.3)	80.5(2.4)	40.1(2.9)	34.2(3.2)	34.6(3.2)
INVRAT	91.0(3.1)	56.4(2.5)	27.3(1.2)	36.7(2.1)	**90.0(3.0)**	49.6(3.1)	27.5(1.9)	33.2(2.6)	80.0(1.8)	42.2(3.2)	32.2(1.6)	31.9(2.4)
CR(ours)	92.4(1.7)	**59.7(1.9)**	**31.6(1.6)**	**39.0(1.5)**	86.5(2.1)	**68.0(2.9)**	**42.0(3.0)**	**49.1(2.8)**	**82.5(2.3)**	**44.7(2.5)**	**37.3(2.0)**	**38.1(2.1)**

Table 10: Results on the Hotel review dataset. Top-10% tokens are highlighted for evaluation. Causal rationalization performs the best in all aspects in terms of capturing human annotations.

Methods	Location			
	Acc (↑)	P (↑)	R (↑)	F1 (↑)
VIB	93.3(1.8)	38.3(4.1)	41.6(6.4)	35.3(4.7)
RNP	94.9(1.7)	37.2(2.1)	39.8(3.3)	34.0(3.9)
FR	**97.3(1.8)**	35.5(1.7)	40.6(1.3)	33.5(1.2)
A2R	92.0(2.2)	37.8(2.9)	40.1(2.1)	34.4(3.2)
CR(ours)	94.0(2.1)	**39.4(1.0)**	**44.2(1.5)**	**36.9(1.0)**

these results, our method can be considered for long-text data.

G.4. Sensitivity Analyses

In the previous experiments, we set $\mu = 0.1$ and $k = 5\%$. To understand the sensitivity of the two parameters, we re-run the experiments on real Beer review data, with $\mu = \{0.01, 0.1, 1, 10\}$ and $k = \{1\%, 5\%, 10\%, 15\%\}$ while keeping the sparsity constraint to be 0.1. We select Top-10% tokens and use accuracy and F1 for the evaluation. Figure 8,9, and 10 summarize our results. It can be seen that our causal rationalization approach's performance is not sensitive to k and μ when μ is not too large e.g. $(0.01, 0.1, 1)$.

G.5. Visualization Examples

G.5.1. BEER REVIEW

We provide three examples for each aspect in terms of all the methods in Table 11.

G.5.2. GA

Since we are more concerned about positive patients who are diagnosed with GA, we present one example of them here. We have converted the medical codes to their corresponding descriptions. As the descriptions can be quite lengthy, we have only included selected codes. In instances where there are multiple consecutive codes, we have only displayed one.

Compared to the rationales found by baseline methods, the ones predicted by our proposed method hit more risk factors of GA. As shown in the first column of Table 9, the patient suffered eyesight defect (BILATERAL FIELD DETECT), irregular heart beat (ATRIAL FIBRILLATION), and diabetes (TYPE I DIABETES WITH DIABETIC POLYNEUROPATHY), all of which are clinically associated with GA as strong risk factors. In comparison, many of the rationales returned by baseline methods are clinically irrelevant to GA, hence less robust.

It is essential to consider the importance of model interpretability in high-stakes applications like healthcare. Although predictive accuracy is a vital aspect, the capacity of a model to provide self-explanatory and interpretable predictions is paramount in fostering trust among healthcare professionals. Furthermore, regulatory compliance necessitates interpretable models, as authorities may mandate explainability to ensure safety, efficacy, and fairness in healthcare algorithms. Given the domain-specific demands for algorithms in healthcare, model interpretability often takes precedence over predictive accuracy, as long as the accuracy is on par with other less interpretable algorithms. Consequently, we argue that the negligible discrepancy in clinical outcome predictions between our proposed method and baselines should be considered within the context of the critical role of interpretability in healthcare applications.
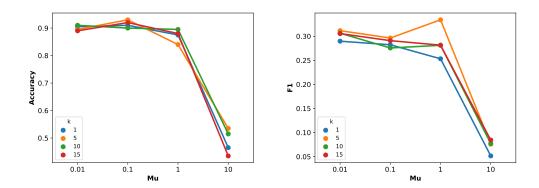
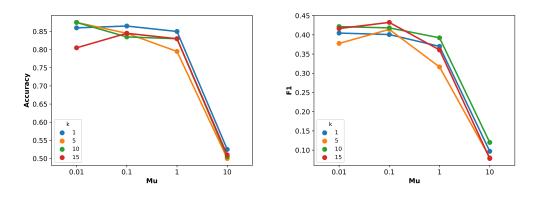Figure 8: Accuracy and F1 score for Appearance.
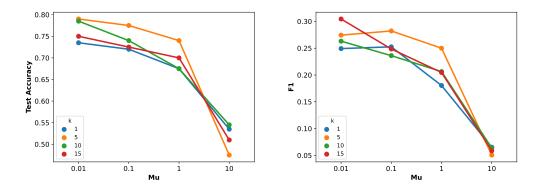

Figure 9: Accuracy and F1 score for Aroma.

Figure 10: Accuracy and F1 score for Palate.

Table 11: Visualization examples of Beer review data.

CR	VIB	RNP	FR
Aspect: Appearance **Label**: Positive **Pred**: Positive poured from bottle into shaker pint glass . a : pale tan-yellow with very thin white head that quickly disappears . s : malt , caramel ... and skunk t : very bad . hint of malt . mostly just tastes bad though . m : acidic , thin , and watery d : not drinkable . unbelievably bad . stay away ... you have several thousand other beers that taste better to spend you money on . do n't be a fool like me .	**Aspect**: Appearance **Label**: Positive **Pred**: Positive poured from bottle into shaker pint glass . a : pale tan-yellow with very thin white head that quickly disappears . s : malt , caramel ... and skunk t : very bad . hint of malt . mostly just tastes bad though . m : acidic , thin , and watery d : not drinkable . unbelievably bad . stay away ... you have several thousand other beers that taste better to spend you money on . do n't be a fool like me .	**Aspect**: Appearance **Label**: Positive **Pred**: Positive poured from bottle into shaker pint glass . a : pale tan-yellow with very thin white head that quickly disappears . s : malt , caramel ... and skunk t : very bad . hint of malt . mostly just tastes bad though . m : acidic , thin , and watery d : not drinkable . unbelievably bad . stay away ... you have several thousand other beers that taste better to spend you money on . do n't be a fool like me .	**Aspect**: Appearance **Label**: Positive **Pred**: Positive poured from bottle into shaker pint glass . a : pale tan-yellow with very thin white head that quickly disappears . s : malt , caramel ... and skunk t : very bad . hint of malt . mostly just tastes bad though . m : acidic , thin , and watery d : not drinkable . unbelievably bad . stay away ... you have several thousand other beers that taste better to spend you money on . do n't be a fool like me .
Aspect: Aroma **Label**: Negative **Pred**: Negative medium head that quickly disappears . lacing is spotty . the smell is rancid . it smells like swamp gas . i am going to assume it is from the can but with miller , who knows ? the best way to describe this brew is " sugar water with a slight watermelon taste " . this is a very sweet tasting beer hence it 's gloss on the can " the champagne of beers ! " . overall not bad for a macro but nothing exciting going on here . notes : shared with my old man at the kitchen table . he buys whatever is cheapest and i take the opportunity to review bad beer . i look at it this way . i wish the old man would buy better beer , but i get to review beer i normally would n't buy anyways .	**Aspect**: Aroma **Label**: Negative **Pred**: Negative medium head that quickly disappears . lacing is spotty . the smell is rancid . it smells like swamp gas . i am going to assume it is from the can but with miller , who knows ? the best way to describe this brew is " sugar water with a slight watermelon taste " . this is a very sweet tasting beer hence it 's gloss on the can " the champagne of beers ! " . overall not bad for a macro but nothing exciting going on here . notes : shared with my old man at the kitchen table . he buys whatever is cheapest and i take the opportunity to review bad beer . i look at it this way . i wish the old man would buy better beer , but i get to review beer i normally would n't buy anyways .	**Aspect**: Aroma **Label**: Negative **Pred**: Negative medium head that quickly disappears . lacing is spotty . the smell is rancid . it smells like swamp gas . i am going to assume it is from the can but with miller , who knows ? the best way to describe this brew is " sugar water with a slight watermelon taste " . this is a very sweet tasting beer hence it 's gloss on the can " the champagne of beers ! " . overall not bad for a macro but nothing exciting going on here . notes : shared with my old man at the kitchen table . he buys whatever is cheapest and i take the opportunity to review bad beer . i look at it this way . i wish the old man would buy better beer , but i get to review beer i normally would n't buy anyways .	**Aspect**: Aroma **Label**: Negative **Pred**: Negative medium head that quickly disappears . lacing is spotty . the smell is rancid . it smells like swamp gas . i am going to assume it is from the can but with miller , who knows ? the best way to describe this brew is " sugar water with a slight watermelon taste " . this is a very sweet tasting beer hence it 's gloss on the can " the champagne of beers ! " . overall not bad for a macro but nothing exciting going on here . notes : shared with my old man at the kitchen table . he buys whatever is cheapest and i take the opportunity to review bad beer . i look at it this way . i wish the old man would buy better beer , but i get to review beer i normally would n't buy anyways .
Aspect: Palate **Label**: Positive **Pred**: Positive sparkling yellow hue with large marshmallow head . light hops , corn , alcohol , grass in the nose . grass notes , bready , soapy hops in the finish . smooth but slick in mouthfeel . highly drinkable and enjoyable . sierra nevada is still the kings of hops even though this beer is a more average brew for them .	**Aspect**: Palate **Label**: Positive **Pred**: Positive sparkling yellow hue with large marshmallow head . light hops , corn , alcohol , grass in the nose . grass notes , bready , soapy hops in the finish . smooth but slick in mouthfeel . highly drinkable and enjoyable . sierra nevada is still the kings of hops even though this beer is a more average brew for them .	**Aspect**: Palate **Label**: Positive **Pred**: Positive sparkling yellow hue with large marshmallow head . light hops , corn , alcohol , grass in the nose . grass notes , bready , soapy hops in the finish . smooth but slick in mouthfeel . highly drinkable and enjoyable . sierra nevada is still the kings of hops even though this beer is a more average brew for them .	**Aspect**: Palate **Label**: Positive **Pred**: Positive sparkling yellow hue with large marshmallow head . light hops , corn , alcohol , grass in the nose . grass notes , bready , soapy hops in the finish . smooth but slick in mouthfeel . highly drinkable and enjoyable . sierra nevada is still the kings of hops even though this beer is a more average brew for them .

Table 12: Visualization examples of two GA patients.

CR	VIB	RNP	FR
Label: Positive **Pred**: Positive	**Label**: Positive **Pred**: Positive	**Label**: Positive **Pred**: Positive	**Label**: Positive **Pred**: Positive

CR	VIB	RNP	FR
✖ HEMORRHAGE, NOT ELSEWHERE CLASSIFIED	✖ ADVERSE EFFECT OF ANTICOAGULANTS, INITIAL ENCOUNTER	✖ ALCOHOL ABUSE WITH INTOXICATION, UNSPECIFIED	✖ HEMATURIA, UNSPECIFIED
✖ PAIN IN RIGHT HIP	✖ HEMORRHAGE, NOT ELSEWHERE CLASSIFIED	✖ PAIN IN RIGHT HIP	✖ ADVERSE EFFECT OF ANTICOAGULANTS, INITIAL ENCOUNTER
✔ HOMONYMOUS BILATERAL FIELD DEFECTS, LEFT SIDE	✖ DEHYDRATION	✖ SUBLUXATION COMPLEX (VERTEBRAL) OF LUMBAR REGION	✔ CHRONIC FATIGUE, UNSPECIFIED
✔ CHRONIC ATRIAL FIBRILLATION	✖ SUBLUXATION COMPLEX (VERTEBRAL) OF LUMBAR REGION	✖ PAIN IN RIGHT SHOULDER	✖ SUBLUXATION COMPLEX (VERTEBRAL) OF LUMBAR REGION
✔ HOMONYMOUS BILATERAL FIELD DEFECTS, UNSPECIFIED SIDE	✔ HOMONYMOUS BILATERAL FIELD DEFECTS, LEFT SIDE	✖ LOW BACK PAIN	✖ STRAIN OF MUSCLE, FASCIA AND TENDON OF LOWER BACK, INITIAL ENCOUNTER
✔ OTHER OPTIC ATROPHY, RIGHT EYE	✔ CHRONIC ATRIAL FIBRILLATION	✔ UNSPECIFIED ATRIAL FIBRILLATION	✔ HOMONYMOUS BILATERAL FIELD DEFECTS, LEFT SIDE
✔ TYPE 1 DIABETES MELLITUS WITH DIABETIC POLYNEUROPATHYSIDE	✔ HOMONYMOUS BILATERAL FIELD DEFECTS, UNSPECIFIED SIDE	✖ STRAIN OF MUSCLE(S) AND TENDON(S) OF THE ROTATOR CUFF OF RIGHT SHOULDER, INITIAL ENCOUNTER	✖ SUBLUXATION COMPLEX (VERTEBRAL) OF LUMBAR REGION
✖ ERECTILE DYSFUNCTION DUE TO ARTERIAL INSUFFICIENCYSIDE	✖ SUBLUXATION COMPLEX (VERTEBRAL) OF LUMBAR REGION	✖ LOW BACK PAIN	✖ BARRETT'S ESOPHAGUS WITH DYSPLASIA, UNSPECIFIED
✔ TYPE 1 DIABETES MELLITUS WITH DIABETIC POLYNEUROPATHY	✔ TYPE 1 DIABETES MELLITUS WITHOUT COMPLICATIONS	✔ TYPE 1 DIABETES MELLITUS WITH DIABETIC POLYNEUROPATHY	✖ STRAIN OF MUSCLE(S) AND TENDON(S) OF THE ROTATOR CUFF OF RIGHT SHOULDER, INITIAL ENCOUNTER
✖ ACQUIRED KERATOSIS [KERATODERMA] PALMARIS ET PLANTARIS	✖ STRAIN OF MUSCLE(S) AND TENDON(S) OF THE ROTATOR CUFF OF RIGHT SHOULDER, INITIAL ENCOUNTER	✖ STRAIN OF MUSCLE(S) AND TENDON(S) OF THE ROTATOR CUFF OF RIGHT SHOULDER, INITIAL ENCOUNTER	✔ TYPE 1 DIABETES MELLITUS WITH DIABETIC POLYNEUROPATHY
	✖ XEROSIS CUTIS	✖ PERIPHERAL VASCULAR DISEASE, UNSPECIFIED	✖ PAIN IN LEFT SHOULDER
	✖ PERIPHERAL VASCULAR DISEASE, UNSPECIFIED‖	✔ UNSPECIFIED ATRIAL FIBRILLATION	✖ ACQUIRED KERATOSIS [KERATODERMA] PALMARIS ET PLANTARIS
			✖ UNILATERAL PRIMARY OSTEOARTHRITIS OF FIRST CARPOMETACARPAL JOINT, LEFT HAND

H. Proof of Theorem 4.2

H.1. Identification of CPNS

Proof: Let's denote the logical operators *and*, *or* as $\wedge, \vee$, respectively. Given $\boldsymbol{X} = \boldsymbol{x}$, firstly we know

$$\{Y\left(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}\right) = y\} \vee \{Y\left(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i},\right) \neq y\} = \text{True}. \tag{6}$$

Then we have

$$
\begin{aligned}
&\{Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y\} \\
\overset{(6)}{=}\quad & \{Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y\} \\
& \wedge \left[\{Y\left(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}, \boldsymbol{X} = \boldsymbol{x}\right) = y\} \vee \{Y\left(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) \neq y\}\right] \\
=\quad & \left[\{Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y\} \wedge \{Y\left(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y\}\right] \\
& \vee \left[\{Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-i}\right) = y\} \wedge \{Y\left(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) \neq y\}\right] \\
\overset{(4)}{=}\quad & \left[\{Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y\} \wedge \{Y\left(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y\}\right],
\end{aligned}
\tag{7}
$$

where we use the monotonicity assumption in (4).

Also, we know

$$\{Y\left(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i},\right) = y\} \vee \{Y\left(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}\right) \neq y\} = \text{True}. \tag{8}$$

Then we can get

$$
\begin{aligned}
&\{Y\left(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y\} \\
\overset{(8)}{=}\quad & \{Y\left(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y\} \\
& \wedge \left[\{Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y\} \vee \{Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) \neq y\}\right] \\
=\quad & \left[\{Y\left(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y\} \wedge \{Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y\}\right] \\
& \vee \left[\{Y\left(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-i}\right) = y\} \wedge \{Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) \neq y\}\right] \\
\overset{(7)}{=}\quad & \{Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y\} \\
& \vee \left[\{Y\left(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-i}\right) = y\} \wedge \{Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) \neq y\}\right].
\end{aligned}
\tag{9}
$$

Based on the consistency assumption in (2), we either have $\{Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y\}$ or $\{Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) \neq y\}$ holds. Therefore, we know the two events in the last line of (9) are disjoint and further take the probability on both sides to get:

$$
\begin{aligned}
& P\left(Y\left(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y \mid \boldsymbol{X} = \boldsymbol{x}\right) \\
=\quad & P\left(Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y \mid \boldsymbol{X} = \boldsymbol{x}\right) \\
& + P\left(Y\left(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-i}\right) = y, Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) \neq y \mid \boldsymbol{X} = \boldsymbol{x}\right),
\end{aligned}
\tag{10}
$$

where the last term is exactly CPNS_j which we want to identify.

Finally with our ignorability assumption (3) we get:

$$
\begin{aligned}
& \text{CPNS}_j \\
=\quad & P\left(Y\left(Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y \mid \boldsymbol{X} = \boldsymbol{x}\right) - P\left(Y\left(Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}\right) = y \mid \boldsymbol{X} = \boldsymbol{x}\right) \\
\overset{(3)}{=}\quad & P(Y = y \mid Z_j = z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}, \boldsymbol{X} = \boldsymbol{x}) - P(Y = y \mid Z_j \neq z_j, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{-j}, \boldsymbol{X} = \boldsymbol{x}).
\end{aligned}
\tag{11}
$$

H.2. Lower Bound of CPNS

Proof: To find the lower bound of CPNS, for any three events A, B, and C, we know that

$$P(A, B \mid C) \geq \max[0, P(A \mid C) + P(B \mid C) - 1]. \tag{12}$$

We substitute A for $\{Y(Z_j = z_j, \mathbf{Z}_{-j} = \mathbf{z}_{-j}) = y\}$, B for $\{Y(Z_j \neq z_j, \mathbf{Z}_{-j} = \mathbf{z}_{-j}) \neq y\}$ and C for $\{\mathbf{X} = \mathbf{x}\}$.

Also similar to (11) with ignorability assumption (3), we can get

$$P(A \mid C) = P(Y(Z_j = z_j, \mathbf{Z}_{-j} = \mathbf{z}_{-j}) = y \mid \mathbf{X} = \mathbf{x}) = P(Y = y \mid Z_j = z_j, \mathbf{Z}_{-j} = \mathbf{z}_{-j}, \mathbf{X} = \mathbf{x}). \tag{13}$$
$$P(B \mid C) = P(Y(Z_j \neq z_j, \mathbf{Z}_{-j} = \mathbf{z}_{-j}) \neq y \mid \mathbf{X} = \mathbf{x}) = P(Y \neq y \mid Z_j \neq z_j, \mathbf{Z}_{-j} = \mathbf{z}_{-j}, \mathbf{X} = \mathbf{x}). \tag{14}$$

Then combining (13) and (14):

$$
\begin{aligned}
& P(A \mid C) + P(B \mid C) - 1 \\
=& P(Y = y \mid Z_j = z_j, \mathbf{Z}_{-j} = \mathbf{z}_{-j}, \mathbf{X} = \mathbf{x}) + P(Y \neq y \mid Z_j \neq z_j, \mathbf{Z}_{-j} = \mathbf{z}_{-j}, \mathbf{X} = \mathbf{x}) - 1 \\
=& P(Y = y \mid Z_j = z_j, \mathbf{Z}_{-j} = \mathbf{z}_{-j}, \mathbf{X} = \mathbf{x}) - P(Y = y \mid Z_j \neq z_j, \mathbf{Z}_{-j} = \mathbf{z}_{-j}, \mathbf{X} = \mathbf{x}).
\end{aligned}
\tag{15}
$$

Finally, the lower bound can be obtained by replacing $P(A \mid C) + P(B \mid C) - 1$ in (12) by (15).