Learnable Game-theoretic Policy Optimization for Data-centric Self-explanation Rationalization

Yunxiao Zhao ⓘ, Zhiqiang Wang ⓘ, Xingtong Yu ⓘ, Xiaoli Li ⓘ, Jiye Liang ⓘ, and Ru Li ⓘ,

Abstract—Rationalization, a data-centric framework, aims to build self-explanatory models to explain the prediction outcome by generating a subset of human-intelligible pieces of the input data. It involves a cooperative game model where a generator generates the most human-intelligible parts of the input (i.e., rationales), followed by a predictor that makes predictions based on these generated rationales. Conventional rationalization methods typically impose constraints via regularization terms to calibrate or penalize undesired generation. However, these methods are suffering from a problem called mode collapse, in which the predictor produces correct predictions yet the generator consistently outputs rationales with collapsed patterns. Moreover, existing studies are typically designed separately for specific collapsed patterns, lacking a unified consideration. In this paper, we systematically revisit cooperative rationalization from a novel game-theoretic perspective and identify the fundamental cause of this problem: the generator no longer tends to explore new strategies to uncover informative rationales, ultimately leading the system to converge to a suboptimal game equilibrium (correct predictions *v.s* collapsed rationales). To solve this problem, we then propose a novel approach, Game-theoretic **P**olicy **O**ptimization oriented **RAT**ionalization (PORAT), which progressively introduces policy interventions to address the game equilibrium in the cooperative game process, thereby guiding the model toward a more optimal solution state. We theoretically analyse the cause of such a suboptimal equilibrium and prove the feasibility of the proposed method. Furthermore, we validate our method on nine widely used real-world datasets and two synthetic settings, where PORAT achieves up to 8.1% performance improvements over existing state-of-the-art methods.

Index Terms—Data-centric Explainability, Self-explanation, Rationale Mining, Game-theoretic Policy Optimization.

---❖---

1 INTRODUCTION

With the success of deep learning in processing large-scale data, the demand for model interpretability has garnered significant attention in recent years [1]. Ideally, model explanations should be both plausible and faithful, which means they should be aligned with human understanding and can reflect the model's predictive behaviour simultaneously [2], [3]. Early studies of explainability [4]–[8], focusing on model-centric explanations, try to leverage post-hoc explanation by approximating important features through machine learning models to explain predictions. Despite appearing plausible, this may not faithfully represent an agent's decision [1], since the explanation generation process is trained separately from the model predictions. In contrast to model-centric post-hoc methods, data-centric self-explanation techniques typically offer increased transparency and faithfulness [9], as the prediction is made based on the explanation itself [10], [11].

In this study, our primary focus is on investigating a general data-centric self-explaining framework called Rationalizing Neural Predictions (RNP, also known as rationalization) [12], which with its variants has become mainstream approach for facilitating the interpretability of models [13]–
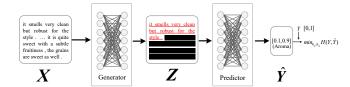
Fig. 1: A standard RNP framework on the binary sentiment analysis, where X, Z, $\hat{Y}$, Y represent the input data, rationale, prediction and the groundtruth label, respectively.

[21], and also has the potential to be applied to downstream tasks such as sentiment analysis [12], [22], image classification [23], graph neural networks [24], legal judgment [25], and the recommendation system [26]. As illustrated in Fig.1, there is a standard rationalization RNP framework, which aims to generate a small and human-intelligible subset (i.e., rationale) from the input data to support and explain the prediction results when yielding them. Here, they highlight key textual spans for input data and utilize a cooperative game with two players (a generator and a predictor) to maximize prediction accuracy through the computation of the maximum mutual information (MMI) loss [27]. As a result, this principle aims to faithfully provide explanations to explain the coupled connection between the input and the model-agnostic task label [28].

Despite such a rationalization model can ensure the faithfulness of the model [27] (i.e., certification of exclusion [29]), the cooperative game is difficult to train if the generator and the predictor are not well coordinated. In this paper, we identify two key challenges that constrain the learning and optimization of the rationale within this

self-explaining framework. **i) Mode collapse of rationale generation.** Mode collapse refers to the phenomenon where, during the process of generating self-explanations, the predictor produces correct predictions, yet the generator consistently outputs collapsed rationale patterns. It becomes fixated on a few dominant modes in the training data and fails to capture the rational rationale distribution. As illustrated in Fig.2, N different rationale modes prevent the generator from generating meaningful rationales with high plausibility. The generator may produce some meaningless fragments that are decorrelated information (e.g., Pattern 1) or not human-intelligible (e.g., Pattern N) to explain the predictor's prediction on the Aroma aspect. Though the predictor infers correct predictions, the generator yields uninformative rationales, converging to a sub-optimal state (correct predictions *v.s* collapsed rationales). The core idea lies in the fact that the generator initially produces a specific pattern (maybe bad patterns), when passed to the predictor, which still leads to a correct label. In such cases, the generator can receive positive feedback and is thus encouraged to overfit to that particular pattern. **ii) Unified modeling of rationale patterns.** Most existing research focuses on individual cases, lacking of unified modeling for data-intrinsic rationale patterns. For example, a series of studies adopts causal inference and a rectified criterion to exclude the mode collapse of spurious correlations (Pattern 1). Chang et al. [30] use an environment-agnostic predictor to recognize spurious correlations; Yue et al. [31] aim to remove spurious correlations based on backdoor adjustment; Liu et al. [32] propose the minimum conditional dependence criterion to uncover causal rationales rather than spurious features. In addition, some studies use additional information to regularize the predictor to address the partial degeneration problem (Pattern 2). Yu et al. [27] uses soft attention from the generator to input full text information into the predictor; Huang et al. [33] and Liu et al. [15] follow the importance of full input and align from different points of view; Liu et al. [21] uses the same encoder between the generator and the predictor to transmit information. Although the above studies have made progress on these two patterns, the patterns present in the data are not finite in variety, as evidenced by recently identified rationalization failure [16]. Moreover, some studies [28], [32] also indicate that various data features may compete with the true rationale for extraction opportunities, thereby hindering the interpretability of the data. This can easily lead to the generator developing diverse rationales selected without human understanding.

To address the above problems, we unify spurious correlations, degeneration and other collapsed rationale fragments, treating them collectively as suboptimal rationales; and systematically revisit the cooperative mechanism of rationalization from a novel game-theoretic perspective and present the existing cooperative problem in the rationalization framework. We theoretically analyze that the fundamental cause of collapsed rationales is that the generator no longer tends to explore new strategies and falls into a sub-optimal game equilibrium. Therefore, to solve this problem, we propose a novel rationalization method and prove its feasibility from a game-theoretic perspective, termed **P**olicy **O**ptimization oriented **RAT**ionalization (**PORAT**), which aims to guide the rationalization model to cope with such a
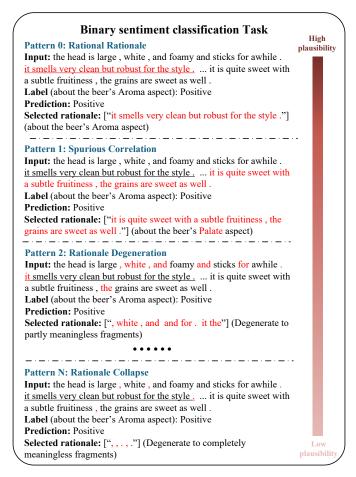

Fig. 2: Rationales of different patterns illustrate the rationales of rational, spurious correlation, partial degeneration, and complete degeneration, which are caused by the generator. Human-annotated rationales are underlined; "red text" indicates rationales generated by models.

suboptimal equilibrium and to promote the mode collapse problem of cooperative rationalization. The contributions of this paper can be summarized as follows:

- New perspective: We unify the concept of collapsed rationales, systematically revisit the cooperative game mechanism of rationalization from a novel perspective, and reveal the game-theoretic problem between two players, i.e., sub-optimal game equilibrium.
- Theoretical insights: We theoretically analyze the fundamental causes of the sub-optimal game equilibrium problem between two players in rationalization.
- New methodology: We propose a novel method called PORAT, which progressively introduces policy interventions to address the sub-optimal game equilibrium problem in the cooperative game process. Moreover, we prove the feasibility of the proposed PORAT.
- Empirical results: Extensive experiments on real-world benchmarks (nine widely used datasets) and synthetic settings (two synthetic settings) demonstrate the effectiveness of PORAT, which improves the F1 score by up to 8.1% as compared to the state-of-the-art method.

The remainder of this paper is organized as follows. Sect. 2 summarizes the related works. The problem definition of rationalization is given in Sect. 3. The revisiting of cooperative rationalization is specified in Sect. 4, including game-

theoretic mechanisms and problems. The proposed method and theoretical insights are presented in Sect. 5 Besides, the experimental results and analysis are in Sect. 6. Finally, we conclude this study in Sect. 7. To streamline the main text for better readability, we have moved some non-critical technical proofs and setups to the Appendices.

2 RELATED WORK

Explainability is a critical research area in the fields of data science and artificial intelligence. In this section, we categorize the explainability research into three main types: data-centric explanations, model-centric explanations and generative explanations with large language models. our primary focus will be on the methods of rationalization in the domain of data-centric explanation research.

2.1 Data-centric rationalization explanations

Data-centric rationalization has received increasing attention in recent years, aiming to answer the question: Which parts of the input data drive the prediction made by deep neural networks (DNNs)? [13]. Typically, this research consists of a generator and a predictor, which produces task-specific predictions using the predictor, while the generator identifies a short and coherent subset of the original input (namely rationale) that is sufficient to explain and support the prediction. There are two main lines of research: supervised rationalization and unsupervised rationalization.

Supervised rationalization. The supervised rationalization framework jointly utilizes rationales and class labels during training. Representative works mainly focus on benchmarks and proposed methods [3], [34]–[36]. For example, DeYoung et al. [34] propose an ERASER benchmark which contains several datasets with both task labels and gold rationales. Lehman et al. [35] propose a pipeline approach as a supervised baseline. Chan et al. [3] develop a unified framework to replace previous works' heuristic design choices with a generic learned rationale generator. Li et al. [36] propose to employ mixed adversarial training and boundary match constraint to improve supervised rationales. These studies usually rely on gold rationales annotated by humans during model training, and formulate rationalization as a multitask learning problem, optimizing the joint likelihood of both class labels and extractive rationales. However, for most tasks, obtaining large-scale annotated rationales is impractical, which limits the applicability in real-world scenarios.

Unsupervised rationalization. The other major line of research is initiated by Lei et al. [12], who propose a unsupervised framework for self-explainable rationale learning. This approach also employs a generator and a predictor component. Since the predictor makes its decision solely based on the explanation produced by the generator, the resulting rationale is faithful to the model's behavior [27], [29], [37]. However, optimizing unsupervised rationales remains a challenging problem. Some early studies [38]–[40] mainly focus on how to mitigate this blocking problem from a rationale sample perspective. Most recent studies have focused separately on individual collapsed problems.

For example, a series of studies uses additional information alignment to regularize the predictor, aiming to directly improve the degeneration. Yu et al. [9] add a complementary predictor that uses text not selected as the rationale, and use soft attention from the generator to input full text information into the predictor [27]. Huang et al. [33] and Liu et al. [15] follow the importance of full input and align from different points of view. Yue et al. [41] improve rational representations by reducing the mutual information between rational and non-rational parts of the input. Liu et al. [21] use the same encoder between the generator and the predictor; they also introduce lipschitz continuity to model asymmetric learning rates, with the aim of decoupling the optimization frequency of two players [20]. Zhao et al. [14] and Hu et al. [16] introduce an extra reinforced causal agent and a guidance module to guide the generator regulate the degeneration process of rationalization, respectively.

On the other hand, some work introduces causal theory and a rectified criterion to address the problem of spurious correlations. Chang et al. [30] use an environment-agnostic predictor to recognize spurious correlations. Yue et al. [31] aim to remove spurious correlations based on backdoor adjustment. Liu et al. [32] propose the minimum conditional dependence criterion to uncover causal rationales rather than spurious features, also introduce a new criterion that treats spurious features as equivalent to plain noise.

Although the above methods achieve improvements on individual collapsed mode problems, few studies have investigated the underlying nature of these collapsed rationales and conducted unified modeling. Moreover, the interaction between the two players in rationalization has not been fully explored. In this paper, we analyze the coordination mechanism of rationalization using game-theoretic methodology from a novel perspective, aiming to systematically reveal relationships and underlying problem between two players, and propose a solution to address this problem.

2.2 Model-centric post-hoc explanations

Different from data-centric methods, model-centric methods aim to approximate the important features used by machine learning models to generate predictions, which has also been widely explored [4], such as LIME [5], SHAP [6], Anchors [7] and so on. Recently, Menon et al. [8] propose MaNtLE, a natural language explainer, to analyze a set of classifier predictions and generate natural language explanations for structured classification tasks. From the perspective of explanation provenance, these methods are commonly known as post-hoc explanation approaches, as the explanations they provide are generated independently of the well-trained predictor. As a result, post-hoc explanation usually provides less transparency [1] and faithfulness [9]. Therefore, these explanation methods are a research line that is related but orthogonal to our research.

2.3 Generative explanation with large language models

With the great success of large language models, a new line of research in explainability has emerged: in-context learning (ICL)-based chain-of-thought (CoT) reasoning [42]. Instead of identifying rationales from the input data, these methods generate intermediate reasoning steps before producing an answer, treating the reasoning process itself as an explanation. This compelling CoT technique has inspired several variants such as IRCoT [43], Self-ASK [44], FLARE

[45] and DRAGIN [46], all of which have shown promising progress. However, due to unpredictable failure problems [47] and hallucinated reasoning [48], CoT-based generative explanations produced by large language models are often unreliable in high-stakes scenarios. Recent studies suggest that language models still struggle with unsupervised, self-explanatory tasks [18], [32], also CoT-based language model reasoning is not always faithful [49], [50].

3 PROBLEM DEFINITION

Notations. In this study, without losing generality, we consider the classification problem and denote the generator and predictor as $f_G(\cdot)$ and $f_P(\cdot)$, with θ_g and θ_p representing their parameters. Here, to ensure clarity and facilitate better comparison with mainstream methods, we consider input X as text data. Therefore, the input can be represented as $X = [x_1, x_2, ..., x_l](1 \leq i \leq l)$ consisting of text tokens x_i, where l is the number of tokens. The label of X is a one-hot vector $Y \in \{0, 1\}^c$, where c is the number of categories.

Self-explanation rationalization. The standard rationalization framework RNP consists of a generator $f_G(\cdot)$ and a predictor $f_P(\cdot)$, where the generator aims to select the most informative pieces from the input X. For each $(X, Y) \in \mathcal{D}$, the generator first gets a sequence of binary masks $M = [m_1, m_2, ..., m_k] \in \{0, 1\}^l$. Then, it forms the rationale $\hat{Z}$ by the element-wise product of X and the binary mask M:

$$\hat{Z} = M \odot X = [m_1 x_1, m_2 x_2, ..., m_k x_k]. \tag{1}$$

Subsequently, the informativeness of the rationale $\hat{Z}$ generated by $f_G(\cdot)$ is measured by the negative cross entropy $-H(Y, Y_{\hat{z}})$, where $Y_{\hat{z}}$ is the output of $f_P(\cdot)$ with the input being $\hat{Z}$. Consequently, the generator and the predictor are usually optimized cooperatively:

$$\min_{\theta_G, \theta_P} \mathcal{H}(Y, \hat{Y} \mid f_G(X)), s.t. \hat{Y} = f_P(f_G(X)). \tag{2}$$

Here, we denote the rationale $\hat{Z}$ generated by $f_G(X)$, i.e., $\hat{Z} = f_G(X)$. Ideally, the rationale $\hat{Z}$ by the model should consist of meaningful rationales with best plausibility, which we denote best rationale as Z (named golden rationale[1]).

Sparsity and continuity constraints. To make the rationales generated by $f_G(\cdot)$ human-intelligible, RNP methods usually constrain the rationales by compact and coherent regularization terms. Thus we also adopt the same constraints used in most research:

$$\Omega(M) = \lambda_1 \left| \frac{||M||_1}{l} - s \right| + \lambda_2 \sum_t |m_t - m_{t-1}| \tag{3}$$

where l denotes the number of tokens in the input. The first term encourages that the percentage of the tokens being selected as rationales is close to a pre-defined level s. The second term encourages the rationales to be coherent. Finally, the overall objective learned is defined as

$$\min \mathbb{L} = \min_{\theta_G, \theta_P} \mathcal{H}(Y, \hat{Y}) + \Omega(M), s.t. \hat{Y} = f_P(f_G(X)). \tag{4}$$

---

[1] Note that the golden rationale called refers only to the rationale under the assumed ideal conditions, the actual task is unsupervised rationalization during model training [17], [51].
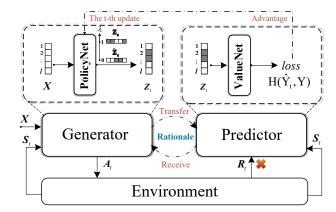

Fig. 3: The coordination mechanism in RNP framework.

4 REVISITING COOPERATIVE RATIONALIZATION

In this section, we first present the knowledge of preliminaries (Sect. 4.1) to illustrate the coordination mechanism of the RNP models (Sect. 4.2). Then, we analyze relationships and the underlying problem between two players from a game-theoretic perspective (Sect. 4.3).

4.1 Preliminaries

To intuitively illustrate the game-theoretic dynamics of RNP models, we first define the rationale optimization process, following prior work. We model the training process of rationale as a Markov decision process $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}\}$ *from the generator perspective* [14], where $\mathcal{S} = \{s^t\}$ represents set of states abstracting the process of optimizing rationale during training, and $\mathcal{A} = \{a^t\}$ indicate the set of actions that update a rationale to the one state. In particular, the transition dynamics $\mathcal{P}(s^{t+1}|s^t, a^{t+1})$ specify how the state s^{t+1} is updated from the prior state s^t by taking action a^{t+1}. Besides, $\mathcal{R}(s^t, a^{t+1})$ quantifies the reward obtained after taking action a^{t+1} based on the prior state s^t. Therefore, cooperative training for rationale can be depicted as the sequence process $(s^0, a^1, r^1, s^1, ..., a^K, r^K, s^K)$, where the state s^t at timestep t can be formulated by $s^t = \hat{Z}_t$ in the *t-th* update. However, previous work [14], [27] neglects the involvement of the predictor. To this end, we introduce the following definitions, which enable us to derive game-theoretic policies of both the generator and the predictor during the game.

Definition 1 (Two-Agent Markov Games for Rationalization). *Let $\mathcal{M}$ be a markov game process with two agents. It can be defined as a 7-tuple $< \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \boldsymbol{\rho}_0, \gamma, \mathcal{R} >$ of states $\mathcal{S}$, actions $\mathcal{A}$, transition probability function $\mathcal{P}(s^{t+1} \mid s^t, a)$, the initial state distribution ρ_0, a discount factor γ, and the joint reward function $\mathcal{R}$, where $\mathcal{N} = \{1, 2\}$.*

Definition 2 (Game-theoretic Policy for Rationalization). *A game-theoretic policy $\pi(a \mid s^t)$ is a probability distribution defined as $\pi : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ indicating the probability of choosing an action given the state s at timestep t. In this paper, we follow previous work [14], [27] and target on generate-predict strategies to learn a conditional policy $\pi(a \mid s^t)$, which indicates a policy for the generation or prediction of a candidate rationale.*

4.2 Game-theoretic Mechanism for Rationalization

As shown in Fig.3, a standard RNP game consists of a generator f_G, a predictor f_P and an environment $\mathbb{E}$. The f_G produces a rationale and feeds this action back to the environment. Meanwhile, f_P receives a rationale and performs a fitting, guiding f_G's policy update. Furthermore, $\mathbb{E}$ provides the corresponding observations and rewards based on the actions of f_G, respectively. However, it lacks a direct reward for f_P since model faithfulness is a key concern. f_P can only receive those rationales from f_G [20], [27], [29]. This leads to a key characteristic: *the quality of rationales generated by the f_G relies on the f_P's supervision; and whatever the f_G transmits, the f_P receives.* Therefore, we can derive a proposition for the nature of rationalization as follows,

Proposition 1 (Dependence and Non-discriminatory). *Given an RNP model, which consists of a generator f_G and a predictor f_P. Let X, Y and Z be the input data, label and rationale, where $\hat{Z}_i$ and $\hat{Z}_j$ are two candidate rationales. Then we have*

- *Dependence (for Generator):* $\min_{\theta_G} \mathcal{H}(Y, \hat{Y}), s.t. \hat{Y} = f_P(\hat{Z}); \hat{Z} = f_G(X)$, *which means learnable parameter θ_G depends on the f_P's supervised loss.*
- *Non-discriminatory (for Predictor):* $\hat{Y} = f_P(\hat{Z}_i)$ *and* $\hat{Z}_i = Z$, *which is satisfied for f_P; however,* $\hat{Y} = f_P(\hat{Z}_j)$ *and* $\hat{Z}_j \neq Z$, *which is also satisfied for f_P.*

This implies that to ensure interpretability (faithfulness), the rationalization model indirectly compromises the reward optimization process inherent in the game. We further investigate this mechanism by providing insights from both *probabilistic* and *suboptimality* perspectives.

Ideally, the rationale candidate generated by the generator at time t should be the most informative text segment, while simultaneously constrained by Equation 3, if and only if there is exactly one, i.e., the gold rationale (we denote that Z_c). In this context, the generator's policy network produces an action profile based on its distribution, thereby obtaining a candidate rationale Z_t as the current state s_t. Considering probabilistic events, the goal is to learn a high-quality of rationale Z_c from l masked tokens. The total number of possible events in the sample space is 2^l, while the probability of the model learning Z_c is only $\frac{1}{2^l}$. *Therefore, there is a low probability of identifying a high-quality rationale in an unsupervised setting.* On the other hand, Z_c at time t contains the higher informative piece, and the failed rationale $\tilde{Z}_c$ (Z_c's complementary set for X), which contains the lowest informative one. Generally, for the MMI loss, we have $L_{MMI}(Z_c) << L_{MMI}(\tilde{Z}_c)$ [28]. When the generator samples a rationale candidate Z_t, the loss can be expressed as

$$L_{MMI}(Z_t) = \lambda L_{MMI}(Z_c) + (1 - \lambda)L_{MMI}(\tilde{Z}_c), \quad (5)$$

where $\lambda = d(Z_t, Z_c)$ represents the distance between the generated candidate rationale Z_t and the gold rationale Z_c [20]. Based on the above, we derive the following lemma.

Lemma 1. *Let $L_{MMI}(Z_c)$ and $L_{MMI}(\tilde{Z}_c)$ represent the MMI loss corresponding to the gold rationale and the failed rationale, respectively. Then, existing at least one suboptimal state Z_t such that,*

$$L_{MMI}(Z_c) < L_{MMI}(Z_t) < L_{MMI}(\tilde{Z}_c). \quad (6)$$

When $\lambda = 1$ holds, $L_{MMI}(Z_t) = L_{MMI}(Z_c)$ that is relatively small; when $\lambda = 0$ holds, $L_{MMI}(Z_t) = L_{MMI}(\tilde{Z}_c)$ that is relatively large; and when $0 < \lambda < 1$ holds, the loss of Z_t falls between the two.

In summary, within the game-theoretic framework for rationalization, unsupervised optimization rarely leads directly to a high-quality rationale. *Once a low-quality rationale is selected, the model is likely to converge to a suboptimal state.*

4.3 Game-theoretic Problem for Rationalization

Then, what leads to such a sub-optimal state? To analyze the cooperative correlation between the generator f_G and the predictor f_P, we use the actor-critic-based process [52]–[54] to present rationale generation and optimization. Here, we denote the f_G as an actor and the f_P as a critic.

State-action value and state value learning. The actor f_G and the critic f_P collaboratively optimize the rationale generation policy, where f_G is responsible for generating rationale candidates (generation policy), and f_P is responsible for estimating the policy and guiding the optimization of the policy. Formally, the state-action value function Q and the state value function V at time step t can be expressed as:

$$\begin{aligned} Q^\pi(s^t, a) &= V^\pi(s^{t+1}), \\ V^\pi(s^t) &= R^t + \mathcal{H}(Y_{s^t}, \hat{Y}_{s^t}), \end{aligned} \quad (7)$$

where $Q(s, a)$ measures the expected return when taking action a in state s under policy π; $V(s)$ measures the expected return when following policy π from state s. In rationalization, the optimization of the actor $Q(s, a)$ and the critic $V(s)$ are mutually dependent. Specifically, $Q(s, a)$ is influenced by $V(s)$ in the next time step, while $V(s)$ is affected by immediate reward R^t and supervised label loss $\mathcal{H}$. Here, R^t is the sparsity and continuity constraints (Eq.3); $\mathcal{H}$ represent the calculated loss by the groundtruth label Y_{s^t} and prediction $\hat{Y}_{s^t}$ at timestep t.

Policy update and estimation. Given a policy π at timestep t, we can formalize the policy update for actor f_G by computing an advantage function [55] in rationalization.

$$\begin{aligned} \nabla_\theta J(\pi_\theta) &= \mathbb{E}_{s \sim d^\pi, a \sim \pi} \left[\nabla_\theta \log \pi_\theta(a|s) A^\pi(s, a) \right], \\ &= \mathbb{E}_{X \sim d^\pi, Z \sim \pi} \left[\nabla_\theta \log \pi_\theta(Z|X) A^\pi(X, Z) \right], \end{aligned} \quad (8)$$

where $d^\pi(s)$ represents the state distribution under policy π, and ∇_θ denotes the direction of gradient update; $A^\pi(s^t, a) = Q^\pi(s^t, a) - V^\pi(s^t)$ indicates the advantage of the action a compared to the current return of the state s at timestep t. In rationalization, the state-action function $Q^\pi(s^t, a)$ is estimated by the state-value function $V^\pi(s^{t+1})$ at the next time step, where

$$V_{s^{t+1} \sim X}^\pi(s^{t+1}) = \Omega(M) + \mathcal{H}(Y_{s^{t+1}}, \hat{Y}_{s^{t+1}}), \hat{Y}_{s^{t+1}} = f_P(s^{t+1}). \quad (9)$$

Coordinated Game-theoretic Problem: The actor who no longer tends to explore new strategies leads to a suboptimal equilibrium. To intuitively illustrate the coordinated problem between two players, we first reveal a phenomenon about empirical bias. As shown in Table 1, regardless of whether the given rationale mode r is collapsed, once the critic function $V^\pi(s)$ introduces an erroneous bias in its estimation, it will lead to the advantage function $A^\pi(s, a)$ guiding the actor $Q^\pi(s, a)$ towards a near-zero. Moreover,
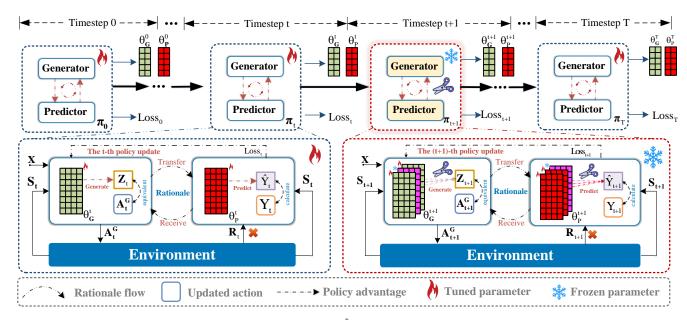
Fig. 4: The overview architecture of PORAT, where $X, Z, \hat{Y}, Y$ represent the input text, rationale, prediction and the groundtruth label, respectively. Here, we provide policy intervention at $(t+1)$-th timestep, which is a progressive policy optimization to help the whole model escape a suboptimal state.

TABLE 1: A toy example payoff (negative entropy) table of the optimization in accordance game, where Coll. rati. and Rati. rati. represent collapsed rationales and rational rationales, respectively; Real. loss indicates L_{MMI} in Eq.5.

Rati. mode r	Real. loss	$V(s^t)$	$\epsilon(s^t)$	$Q(s^t, a)$	$A(s^t, a)$
Rati. rati.	1.0	10000.0	-9999.0	9999.4	-0.6≈0
Coll. rati.(1)	8500.0	1.0	8499.0	1.3	0.3≈0
Coll. rati.(2)	9000.0	1.0	8999.0	1.1	0.1≈0
Coll. rati.(3)	10000.0	1.0	9999.0	0.9	-0.1≈0

$Q^\pi(s, a)$ learns this error $\epsilon(s)$ introduced by $V^\pi(s)$. With Proposition 1, we have that the actor f_G relies solely on the critic f_P's estimation. Errors in the critic's estimation will prevent the policy from converging to the optimal solution. Formally, we can express error $\epsilon(s)$ at timestep t by

$$V^\pi(s) = V^*(s) - \epsilon(s), \qquad (10)$$

where $V^*(s)$ is the optimal state value, and $\epsilon(s)$ represents the estimation error of the critic f_P. Therefore, an incorrect advantage function is generated under a given policy,

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s) = Q^*(s, a) - (V^*(s) - \epsilon(s)),$$
$$= Q^*(s, a) - V^*(s) + \epsilon(s). \qquad (11)$$

If $\epsilon(s)$ is too large or consistently negative at timestep t, then $A^\pi(s, a)$ may become excessively small, preventing the actor f_G from exploring new strategies. Therefore, we can establish the following theorem.

Theorem 1. *Given an RNP model with f_G and f_P. Let $\epsilon(s^t)$ be the estimation error of the f_P for a candidate rationale $\hat{Z}_t$ at timestep t. If exist $\epsilon(s^t) \neq 0$, then the f_G no longer tends to explore new strategies to uncover more informative rationale. That is to say, the policy of f_G π satisfies that*

$$\exists \pi \neq \pi^*, \quad \forall (s^t, a), A^\pi(s^t, a) = 0$$
$$\Rightarrow \nabla_\theta J(\pi) = \mathbb{E}_{s^t, a} \left[\nabla_\theta \log \pi(a|s^t) A^\pi(s^t, a) \right] = 0. \qquad (12)$$

Theorem 1 suggests that *regardless of the policy profile of f_G, if the estimation of f_P is biased, the RNP model will no longer tends to explore new strategies, which leads to a continual degeneration.*

5 METHODOLOGY AND THEORETICAL ANALYSIS

To address the above problem, we propose PORAT (Fig.4), a game-theoretic policy optimization for self-explanation rationalization, including the proposed method (Sect. 5.1) and theoretical insights (Sect. 5.2).

5.1 The Proposed Method

As shown in Equation 10, intuitively, if we introduce a regularization penalty term, the error of the critic can be alleviated. Recent work [15], [21], [27], [33], [41] has explored this through calibrating or penalizing the predictor. However, the penalty factor is difficult to control, which could lead to longer optimization paths or introduce extra local optima [28]. When the model converges to local optima, these approaches also encounter a bottleneck.

Diverging from previous research, we aim to develop a method to help the model cope with such continual degeneration so that, regardless of the strategy chosen by the model, gradient-based descent can guide it out of local optima. We assume that at timestep t, the RNP model is in a suboptimal state. According to Theorem 1, we can derive that the f_G's policy gradient is nearly zero,

$$\nabla_\theta J(\pi) = \mathbb{E}_{s^t, a} \left[\nabla_\theta \log \pi(a|s^t) A^\pi(s^t, a) \right] = 0, \qquad (13)$$

which means that the f_G no longer explores new actions, falling into a continual degeneration. Furthermore, we have

$$A^\pi(s^t, a) = Q^\pi(s^t, a) - V^*(s^t) + \epsilon(s^t) = 0. \qquad (14)$$

If we can ensure that $A^\pi(s^{t+1}, a) \neq 0$, f_G will be able to escape the suboptimal equilibrium at time $t + m$ $(m > 0)$. Formally, this can be expressed as:

$$A^\pi(s^{t+1}, a) = Q^\pi(s^{t+1}, a) - V^\pi(s^{t+1}) \neq 0. \quad (15)$$

We first need to confirm whether there is a more optimal policy selection. Here, we establish the following lemma.

Lemma 2. *Let $S = \{X_1, \ldots, X_{2^l}\}$ be the set of all candidate rationales for a given input X, and let $+C$ and $-C$ suggest a best rationale and a suboptimal one. Suppose f_G satisfies a suboptimal state s^t at timestep t, there exists at least one state induced by the corresponding policy profile $\pi(a|s^{t+1})$ that enables f_G to escape the s^t, that is,*

$$\forall s^t \sim d^\pi \in S, \nabla_\theta J(\pi) = 0 \Rightarrow \exists \pi(a|s^{t+1}), s.t. \nabla_\theta J(\pi) \neq 0, \quad (16)$$

and $\pi(a|s^{t+1}) = \{\pi_{+C}^G \times \pi_{-C}^P\}$ and $\{\pi_{-C}^G \times \pi_{+C}^P\}$ are two solutions for the policies of f_G and f_P.

Lemma 2 means that: there exists a strategy π_{t+1} that enables the model to escape the suboptimal state s^t, and the policy π_{t+1} is from $\pi_j^i (i \in \{f_G, f_P\}, j \in \{+C, -C\})$. However, according to the Proposition 1, we have the non-discriminability for predictor, which means if $f_G \Rightarrow R$, then $R \Rightarrow f_P$.

Parameter Freezing as Intervention. To this end, we disentangle the game between the generator and the predictor from the policy optimization perspective, as shown in Fig.4. Specifically, we first freeze the generator while keeping the predictor active, which allows the generator to block the predictor's suboptimal feedback and generate diverse candidate rationales as optional strategies. Formally, let $V^\pi(s^{t+1}) = 0$, we can rewrite the Equation 15 as

$$A^\pi(s^{t+1}, a) = Q^\pi(s^{t+1}, a) - V^\pi(s^{t+1}) = Q^\pi(s^{t+1}, a) \neq 0 \quad (17)$$

Since the model is in a suboptimal state at timestep t, Equation 17 is equivalent to the generator selecting suboptimal rationale at time $t + 1$, while the predictor does not further fit it. In addition, we have $A^\pi(s^{t+1}, a) \neq 0$, allowing the generator to continue exploring new actions. However, by continuously optimizing Equation 17, the error induced by the predictor's estimation will be learned by the new $Q^\pi(s^{t+1}, a)$. Therefore, we further freeze the predictor to mitigate the impact of errors arising from the suboptimal state. This allows the predictor to block the continuously degenerating parameter updates. According to Equation 7, we have $Q^\pi(s^t, a) = V^\pi(s^{t+1})$, so,

$$A^\pi(s^{t+1}, a) = Q^\pi(s^t, a) - V^\pi(s^t) = V^\pi(s^{t+1}) - V^\pi(s^t). \quad (18)$$

Intuitively, if $V^\pi(s^{t+1}) - V^\pi(s^t) = 0$, then f_P will overfit the state s^t. Therefore, to address the problem, we let $V^\pi(s^{t+1}) = 0$, and freeze the predictor f_P at timestep $t + 1$ in practice. Finally, following the general setup of RNP, we simultaneously activate both the f_G and the f_P, enabling them to collaborate once again.

Policy Optimization. Based on above, given the input X at timestep $t + 1$, the learning objective of the model $J(\pi)$ can

be represented as

$$
\begin{aligned}
J(\pi(a|s^{t+1})) &= \mathbb{E}_{s\sim d^\pi, a\sim\pi}^{(1)} \left\{ \log \pi_\theta(a|s^{t+1})[Q^\pi(s^{t+1}, a)] \right\} \\
&+ \mathbb{E}_{s\sim d^\pi, a\sim\pi}^{(2)} \left\{ \log \pi_\theta(a|s^{t+1})[-V^\pi(s^t)] \right\} \\
&+ \mathbb{E}_{s\sim d^\pi, a\sim\pi}^{(3)} \left\{ \log \pi_\theta(a|s^{t+1})[Q^\pi(s^{t+1}, a) - V^\pi(s^{t+1})] \right\}
\end{aligned}
\quad (19)
$$

further derivation, we have

$$
\begin{aligned}
J(\pi_\theta) &= \min_{\theta_p^{t+1}} \mathcal{H}(Y, f_P^t(f_G^t(X^{t+1})) + \Omega(M) \\
&+ \min_{\theta_g^{t+1}} \mathcal{H}(Y, f_P^{t+1}(f_G^t(X^{t+1}))) + \Omega(M) \\
&+ \min_{\theta_g^{t+1}, \theta_p^{t+1}} \mathcal{H}(Y, f_P^{t+1}(f_G^{t+1}(X^{t+1}))) + \Omega(M)
\end{aligned}
\quad (20)
$$

Iteration and Inference. Equation 20 ensures that the model can temporarily adjust a state. However, the local minima explored by multilayer networks are not unique [56], [57]. To address this issue, we introduce a progressive optimization process [58] and set the update timestep to N. After N time steps, we reintroduce the aforementioned policy optimization for model iteration. The final optimization process can be expressed as follows:

$$L = \sum_{i=0, i\notin\{k*N\}}^{T} L_1(Y, \hat{Y}) + \sum_{i\in\{k*N\}}^{T} L_2(Y, \hat{Y}), k = \{1, \ldots, n\} \quad (21)$$

where $L_1(Y, \hat{Y}) = \min_{\theta_g, \theta_p} \mathcal{H}(Y, \hat{Y}) + \Omega(M)$, $L_2(Y, \hat{Y}) = J(\pi_\theta^{k*N})$. During the inference phase, following the general self-explanation setup [12], [14], [20], the f_P only uses the rationale generated by the f_G for model prediction.

Algorithm. To help readers better understand the process, we detail the main steps for the training and inference phase in Algorithm 1, in which the input is a dataset $\mathcal{D}$, and the output is a model θ_f^* capable of providing both predictions $\hat{y}_j$ and rationales $\hat{z}_j$ for a data sample x_j.

5.2 Theoretical Analysis

Theorem 2. *Given an RNP model with two players (i.e., f_G and f_P) and a suboptimal state s^t at timestep t, where s^t indicates an collapsed rationale candidate $\hat{Z}_t \in X$ from f_G at time t. Suppose f_G satisfies the following condition,*

$$\nabla_\theta J(\pi) = \mathbb{E}_{s,a} \left[\nabla_\theta \log \pi(a|s^t) A^\pi(s^t, a) \right] = 0, \quad (23)$$

then, we have that after m time steps, the fusion and optimization of additional policy actions can be such that

$$A^\pi(s^{t+m}, a) = Q^\pi(s^{t+m}, a) - V^\pi(s^{t+m}) \neq 0. \quad (24)$$

Proof. We first denote a suboptimal state as s^t. According to Theorem 1, we can obtain

$$\nabla_\theta J(\pi) = \mathbb{E}_{s,a} \left[\nabla_\theta \log \pi(a|s^t) A^\pi(s^t, a) \right] = 0, \quad (25)$$

which means $A^\pi(s^t, a) = 0$, the generator f_G no longer tends to explore new strategies. According to Theorem 1, we define the gain change function of both players as $\varphi_{i,a_i}(\pi) = \max\{0, J_i(a_i, \pi_{-i}) - J_i(\pi)\}$ where i indicates the i-th player. When $\varphi_{i,a_i}(\pi) > 0$,

$$
\begin{aligned}
&\max\{0, J_i(a_i, \pi_{-i}) - J_i(\pi)\} > 0, \\
&J_i(a_i, \pi_{-i}) > J_i(\pi),
\end{aligned}
\quad (26)
$$

Algorithm 1 PORAT Algorithm

1: **Input:** A dataset $\mathcal{D}$, including $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$
2: **Output:** A self-explanation model θ_f^*
3: // Training Phase
4: **while** not converged **do**
5:    **for** $x_i \in \mathcal{D}_{\text{train}}$, **in epoch do**
6:       Compute the output of f_G and f_P:

$$\hat{y}_i = f_P(\theta_p^i, f_G(\theta_g^i, x_i)). \quad (22)$$

7:       Minimize $\mathcal{H}(y_i, \hat{y}_i)$.
8:       Update θ_p^i, θ_g^i through gradient decent.
9:       $\theta_g^* \leftarrow \theta_g^i; \theta_p^* \leftarrow \theta_p^i$
10:      // Game-theoretic Policy Intervention
11:       Execute Eq.(17): freezing θ_g^i, update θ_p^i.
12:       Execute Eq.(18): freezing θ_p^i, update θ_g^i.
13:       Update θ_g^i, update θ_p^i.
14:    **end for**
15: **end while**
16: // Inference Phase
17: **for** $x_j \in \mathcal{D}_{\text{test}}$ **do**
18:    Compute the predicting label $\hat{y}_j$ and generated explanation $\hat{z}_j$ using Eq.(22) by substituting the parameters θ_g^* and θ_p^*.
19: **end for**
20: **Output:** $\theta_f^* = \theta_g^* \bigcup \theta_p^*$; $(x_j, \hat{z}_j, \hat{y}_j), \forall x_j \in D_{\text{test}}$

TABLE 2: Statistics of datasets where Pos and Neg denote the number of positive and negative examples in each set.

Benchmarks	Datasets	Train		Dev		Annotation	
		Pos	Neg	Pos	Neg	Pos	Neg
BeerAdvocate [60]	Appearance	202385	12897	28488	1318	923	13
	Aroma	172299	30564	24494	3396	848	29
	Palate	176038	27639	24837	3203	785	20
BeerAdvocate* [12]	Appearance*	16891	16891	6628	2103	923	13
	Aroma*	15169	15169	6579	2218	848	29
	Palate*	13652	13652	6740	2000	785	20
HotelReview [61]	Location	7236	7236	906	906	104	96
	Service	50742	50742	6344	6344	101	99
	Cleanliness	75049	75049	9382	9382	99	101

used in rationalization. Note that each of them contains three distinct aspects, which are trained independently in our experiments. Consequently, these three benchmarks can be considered as nine distinct datasets to some extent. Following previous research [14], [20], [33], we obtain Beer-Advocate [60], BeerAdvocate* [12], and HotelReview [61] datasets, which are all publicly available. As shown in Table 2, the specific splitting details of the nine datasets are presented. In particular, BeerAdvocate is a correlated dataset on beer reviews that can be regarded as addressing the spurious correlation problem, while BeerAdvocate* is a dataset decorrelated by Lei et al. [12] that focuses on the degeneration problem. For HotelReview, it is another benchmark also widely used in rationalization. In synthetic settings, we use the same experiment setup as Yu et al. [27], Liu et al. [20] and Wu et al. [62] did.

Baselines. To validate the effectiveness of PORAT in a rationalization framework, we compare with seven latest methods for BeerAdvocate, including one standard rationalization method: RNP [12]; two calibration-based methods: DARE [41], FR [21]; two causal-based methods: INVRAT [30], MCD [32]; and two recent guidance-based methods: AGR [14] and G-RAT [16]. For BeerAdvocate* and HotelReview benchmarks, we compare with one standard method and five recent models, including RNP [12], DMR [33], A2R [27], FR [21], DR [20] and G-RAT [16].

Evaluation Metrics. Following previous methods [27], [29], [31], [33], we focus on the quality of rationales and adopt Precision (P), Recall (R), and F1-score (F1) as metrics. To fairly compare, we perform the best results on the validation set before testing on the test set. Here, Acc denotes the accuracy of the prediction task based on the selected rationales, while S represents the average ratio of selected tokens to the total length of the original text.

Implementations. We utilize one-layer 200-dimension bi-directional gated recurrent units (GRUs) [63] followed by one linear layer for each of the players, and the word embedding is 100-dimension Glove [64]. We use Adam [65] as the optimizer. The reparameterization trick for binarized sampling is Gumbel-softmax, which is also the same as existing research [20], [32], [38]. To verify the effectiveness of our PORAT by intervening policy, we perform ablation studies by intervening in the policies of both the prefix generator and the prefix predictor. To minimize the influence of other factors, we conduct the ablation experiments using the same hyperparameters as the baseline. In experiments, we use two different architectures (DR [20] and AGR [14]) as the backbone models to validate PORAT on different benchmarks, respectively. All our experiments are run on NVIDIA RTX 6000 Ada GPUs with 48GB.

we have $A^\pi(s^t, a) \neq 0$. However, according to Theorem 1, the joint action candidate of f_G and f_P under the RNP game cannot find a better joint policy that improves the payoff of the RNP model. Therefore, we need to introduce additional policies to enable the model to find the global optimum s^*, but finding the global optimum directly is difficult [59]. According to Lemma 2, we can identify at least two intermediate policy points that guide the model to escape the suboptimal state.

Therefore, if we can integrate these two policy profiles, we will be able to guide the advantage function such that $A^\pi(s^t, a) \neq 0$. Furthermore, with Equation 7, we have

$$A^\pi(s^{t+1}, a) = V^\pi(s^{t+1}) - V^\pi(s^t). \quad (27)$$

This indicates that we can learn the aforementioned intermediate policy points in the RNP game by controlling the value function $V^\pi(s)$, such that,

$$A^\pi(s^{t+m}, a) = Q^\pi(s^{t+m}, a) - V^\pi(s^{t+m}) \neq 0. \quad (28)$$

where m indicates the learning period. The proof of Theorem 2 is finished, which suggests that we can let the RNP models learn policy optimization to solve the suboptimal state for self-explanation rationalization.

6 EXPERIMENTS

In this section, we evaluate our method PORAT in various settings to demonstrate its effectiveness.

6.1 Experimental Setup

Datasets. We compare PORAT using BeerAdvocate [60], BeerAdvocate* [12] and HotelReview [61], which are three multi-aspect sentiment classification benchmarks widely

TABLE 3: Comparison with previous methods on BeerAdvocate [60] benchmark. S is a hyperparameter, which encourages that the percentage of the tokens being generated as rationales is close to a pre-defined level. The **bold** numbers are the best results for our proposed method. The same applies below.

Methods	BeerAdvocate-Appearance					BeerAdvocate-Aroma					BeerAdvocate-Plate				
	S	Acc	P	R	F1	S	Acc	P	R	F1	S	Acc	P	R	F1
RNP	10.0	-	32.4	18.6	23.6	10.0	-	44.8	32.4	37.6	10.0	-	24.6	23.5	24.0
INVRAT	10.0	-	42.6	31.5	36.2	10.0	-	41.2	39.1	40.1	10.0	-	34.9	45.6	39.5
DARE	10.0	-	63.9	42.8	51.3	10.0	-	50.5	44.8	47.5	10.0	-	33.1	45.8	38.4
FR	11.1	75.8	70.4	42.0	52.6	9.7	87.7	68.1	42.2	52.1	11.7	87.9	43.7	40.9	42.3
MCD	9.5	81.5	94.2	48.4	63.9	9.9	87.5	84.6	53.9	65.8	9.4	87.3	60.9	47.1	53.1
AGR	12.4	81.3	80.4	55.4	65.6	12.3	87.8	68.4	54.1	60.4	12.4	86.2	54.4	55.9	55.1
G-RAT	10.5	82.4	81.8	46.3	59.1	10.5	85.2	82.0	55.4	66.2	9.5	89.2	56.2	43.1	48.8
PORAT (Ours)	13.9	83.9	79.1	59.6	**68.0**	11.5	87.5	80.2	59.0	**68.0**	11.6	88.0	65.4	61.1	**63.2**
RNP	20.0	-	39.4	44.9	42.0	20.0	-	37.5	51.9	43.5	20.0	-	21.6	38.9	27.8
INVRAT	20.0	-	58.9	67.2	62.8	20.0	-	29.3	52.1	37.5	20.0	-	24.0	55.2	33.5
DARE	20.0	-	63.7	71.8	67.5	20.0	-	41.0	61.5	49.3	20.0	-	24.4	54.9	33.8
FR	20.9	84.6	74.9	84.9	79.6	19.5	89.3	58.7	73.3	65.2	20.2	88.2	36.6	59.4	45.3
MCD	20.0	85.5	79.3	85.5	82.3	19.3	88.4	65.8	81.4	72.8	19.6	87.7	41.3	65.0	50.5
AGR	19.7	85.2	83.3	88.4	85.8	19.6	89.2	65.7	82.7	73.2	18.0	87.0	45.2	65.6	53.5
G-RAT	19.7	85.0	80.2	85.2	82.6	20.2	88.1	60.5	78.2	68.2	20.3	86.1	38.4	62.7	47.6
PORAT (Ours)	19.3	85.4	84.6	88.1	**86.3**	19.2	89.4	69.4	85.3	**76.5**	19.3	86.7	50.6	78.6	**61.6**
RNP	30.0	-	24.2	41.2	30.5	30.0	-	27.1	55.7	36.4	30.0	-	15.4	42.2	22.6
INVRAT	30.0	-	41.5	74.8	53.4	30.0	-	22.8	65.1	33.8	30.0	-	20.9	71.6	32.3
DARE	30.0	-	45.5	80.6	58.1	30.0	-	32.7	68.2	44.2	30.0	-	19.7	70.5	30.8
FR	29.6	86.4	50.6	81.4	62.3	30.8	88.1	37.4	75.0	49.9	30.1	87.0	24.5	58.8	34.6
MCD	29.7	86.7	59.6	95.6	73.4	29.6	90.2	46.1	87.5	60.4	29.4	87.0	30.5	72.4	42.9
AGR	28.0	87.4	61.6	93.3	74.2	30.6	89.7	43.5	85.3	57.6	30.4	88.3	32.1	78.5	45.6
G-RAT	29.6	87.2	56.0	89.4	68.9	29.8	90.4	42.4	81.1	55.7	29.7	86.2	27.0	64.4	38.0
PORAT (Ours)	28.7	86.1	61.9	95.8	**75.2**	28.3	90.3	48.9	88.7	**63.0**	29.6	89.3	33.1	79.0	**46.7**

TABLE 4: Comparison with previous methods on BeerAdvocate* [12] benchmark.

Methods	BeerAdvocate*-Appearance*					BeerAdvocate*-Aroma*					BeerAdvocate*-Plate*				
	S	Acc	P	R	F1	S	Acc	P	R	F1	S	Acc	P	R	F1
RNP	18.2	83.3	73.8	72.7	73.2	16.0	85.2	64.1	65.9	64.9	13.0	85.2	60.1	63.1	61.5
DMR	18.2	-	71.1	70.2	70.7	15.4	-	59.8	58.9	59.3	11.9	-	53.2	50.9	52.0
A2R	18.4	83.9	72.7	72.3	72.5	15.4	86.3	63.6	62.9	63.2	12.4	81.2	57.4	57.3	57.4
FR	18.4	87.2	82.9	82.6	82.8	15.0	88.6	74.7	72.1	73.4	12.1	89.7	67.8	66.2	67.0
DR	18.3	81.1	82.4	81.6	82.0	15.4	86.2	77.7	76.8	77.2	12.5	85.0	65.9	66.0	66.0
G-RAT	18.5	-	84.8	83.2	84.0	15.5	-	79.1	74.3	76.6	12.3	-	63.4	67.2	65.2
PORAT (Ours)	18.1	83.1	85.2	83.1	**84.2**	15.6	88.0	77.9	78.2	**78.0**	12.5	84.0	69.0	69.0	**69.0**

TABLE 5: Comparison with previous methods on HotelReview [61] benchmark.

Methods	HotelReview-Location					HotelReview-Service					HotelReview-Cleanliness				
	S	Acc	P	R	F1	S	Acc	P	R	F1	S	Acc	P	R	F1
RNP	8.8	97.5	46.2	48.2	47.1	11.0	97.5	34.2	32.9	33.5	10.5	96.0	29.1	34.6	31.6
DMR	10.7	-	47.5	60.1	53.1	11.6	-	43.0	43.6	43.3	10.3	-	31.4	36.4	33.7
A2R	8.5	87.5	43.1	43.2	43.1	11.4	96.5	37.3	37.2	37.2	8.9	94.5	33.2	33.3	33.3
FR	9.0	93.5	55.5	58.9	57.1	11.5	94.5	44.8	44.7	44.8	11.0	96.0	34.9	43.4	38.7
DR	10.5	93.5	51.7	63.7	57.1	11.8	96.5	45.0	50.2	47.5	10.3	94.5	38.6	45.1	41.6
G-RAT	10.1	-	56.1	59.3	57.6	12.1	-	48.8	44.1	46.3	11.9	-	41.4	37.3	39.2
PORAT (Ours)	10.2	94.0	53.4	63.1	**57.8**	13.2	95.5	45.0	51.8	**48.2**	10.6	93.5	38.7	46.4	**42.2**

6.2 Evaluation on Standard Benchmarks

(1) Results on BeerAdvocate benchmark [60]. We first set the rationale sparsity S to approximately 10%, 20%, and 30% [29], [31], [32]. As shown in Table 3, we achieve significant improvements in F1 scores across various aspects, with an increase of up to 8.1% in the Palate aspect ($s \approx 10\%$). The significant improvement shows the superiority of our proposed game-theoretic policy optimization for PORAT in solving suboptimal rationalization, which can help RNP models to explore more optimal policies for games.

(2) Results on BeerAdvocate* benchmark [12]. As shown in Table 4, the results on BeerAdvocate* are illustrated, which focuses more on the research problem of decorrelation [20]. We can observe that our proposed method once again obtains the best performance across all three aspects of the decorrelated beer dataset consistently.

(3) Results on HotelReview benchmark [61]. Table 5 presents the experimental results on the HotelReview. In this benchmark, we set the rationale sparsity close to the human-annotated rationales. We can find that our proposed method also achieves varying degrees of improvement in the Location, Service and Cleanliness three datasets.

In conclusion, we demonstrate that our proposed method PORAT outperforms the best existing methods in

TABLE 6: Experimental results that induces degeneration on synthetic settings. "skew$k$" means that the predictor is pre-trained for k epochs.

Aspect	Setting	RNP				A2R				FR				DR				PORAT (Ours)			
		Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Aroma*	skew10	82.6	68.5	63.7	61.5	84.5	78.3	70.6	69.2	87.1	73.9	71.7	72.8	85.0	77.3	75.7	76.5	86.7	77.0	80.0	78.5
	skew15	80.4	54.5	51.6	49.3	81.8	58.1	53.3	51.7	86.7	71.3	68.0	69.6	85.4	76.1	77.2	76.6	86.6	77.1	79.2	78.1
	skew20	76.8	10.8	14.1	11.0	80.0	51.7	47.9	46.3	85.5	72.3	69.0	70.6	85.5	77.3	76.2	76.8	86.1	77.6	79.6	78.6
Palate*	skew10	77.3	5.6	7.4	5.5	82.8	50.3	48.0	45.5	75.8	54.6	61.2	57.7	85.8	67.7	68.6	68.2	84.7	68.3	68.2	68.3
	skew15	77.1	1.2	2.5	1.3	80.9	30.2	29.9	27.7	81.7	51.0	58.4	54.5	83.9	66.3	66.7	66.5	84.7	68.0	68.4	68.2
	skew20	75.6	0.4	1.4	0.6	76.7	0.4	1.6	0.6	83.1	48.0	58.9	52.9	85.0	59.4	62.6	61.0	85.7	66.6	67.6	67.1

TABLE 7: Experimental results that induces spurious correlation on synthetic settings. Here, following the same setting [62], we report the results of three random seeds across nine distinct setups where "bias=c" indicates the distinct biases of spurious correlations to use Spurious-Motif datasets [66].

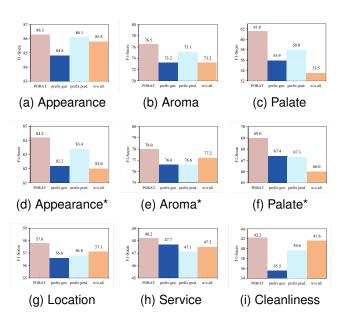| Methods	Suprious-Motif Datasets									Avg
	bias=0.1	bias=0.2	bias=0.3	bias=0.4	bias=0.5	bias=0.6	bias=0.7	bias=0.8	bias=0.9	
Attention	-	-	-	-	$18.3_{\pm 13.0}$	-	$18.2_{\pm 1.4}$	-	$13.4_{\pm 1.3}$	-
ASAP	-	-	-	-	$18.8_{\pm 2.3}$	-	$18.6_{\pm 2.7}$	-	$12.1_{\pm 2.1}$	-
Topk Pool	-	-	-	-	$20.7_{\pm 5.7}$	-	$21.2_{\pm 5.6}$	-	$14.8_{\pm 1.8}$	-
SAG Pool	-	-	-	-	$19.8_{\pm 6.2}$	-	$20.1_{\pm 6.4}$	-	$13.6_{\pm 1.4}$	-
DIR	$26.2_{\pm 1.4}$	$28.0_{\pm 2.8}$	$29.8_{\pm 3.6}$	$28.8_{\pm 3.0}$	$30.2_{\pm 3.3}$	$29.9_{\pm 2.4}$	$30.8_{\pm 1.6}$	$28.7_{\pm 4.3}$	$24.4_{\pm 1.3}$	$28.5_{\pm 2.6}$
DIR-DR	$26.0_{\pm 3.1}\downarrow$	$28.2_{\pm 3.9}$	$29.8_{\pm 4.3}\downarrow$	$29.2_{\pm 4.4}$	$29.1_{\pm 4.1}\downarrow$	$29.0_{\pm 1.7}\downarrow$	$28.6_{\pm 1.3}\downarrow$	$28.1_{\pm 0.7}\downarrow$	$26.3_{\pm 1.6}$	$28.2_{\pm 2.8}\downarrow$
DIR-PORAT	$27.2_{\pm 2.8}\uparrow$	$29.7_{\pm 1.5}\uparrow$	$31.3_{\pm 1.2}\uparrow$	$29.9_{\pm 3.1}\uparrow$	$30.9_{\pm 1.7}\uparrow$	$30.8_{\pm 1.8}\uparrow$	$31.3_{\pm 1.9}\uparrow$	$30.4_{\pm 2.5}\uparrow$	$25.5_{\pm 1.5}$	$29.7_{\pm 1.9}\uparrow$
w/o p.	$26.2_{\pm 0.5}$	$28.1_{\pm 2.2}$	$28.9_{\pm 2.7}$	$28.3_{\pm 3.8}$	$30.3_{\pm 3.0}$	$30.6_{\pm 2.7}$	$31.1_{\pm 2.0}$	$27.7_{\pm 2.6}$	$20.3_{\pm 1.7}$	$27.9_{\pm 2.4}$
w/o g.	$25.2_{\pm 1.2}$	$26.5_{\pm 1.5}$	$29.4_{\pm 2.4}$	$28.4_{\pm 1.9}$	$29.9_{\pm 3.2}$	$28.1_{\pm 0.6}$	$29.0_{\pm 2.1}$	$28.6_{\pm 3.6}$	$25.2_{\pm 3.0}$	$27.8_{\pm 2.2}$

Fig. 5: Ablation Studies on (a-c) BeerAdvocate [60], (d-f) BeerAdvocate* [12] and (g-i) HotelReview [61] Benchmarks.

terms of F1 score across nine datasets from three benchmark datasets (fifteen experiment settings), while maintaining competitive accuracy. This highlights that proposed PORAT method not only offers more accurate explanations than the existing methods but also exhibits strong generalizability.

6.3 Evaluation on Synthetic Settings

To better show the influence of sub-optimal state equilibrium, we further conduct two synthetic experiments.

(1) Degeneration. It is a typical phenomenon of suboptimal states. To show that even if the predictor overfits to trivial patterns and falls in a suboptimal equilibrium, PORAT can still escape, we conduct the same synthetic experiment as Yu et al. [27] and Liu et al. [20] did (The specific setting can be

found in Appendices). The results of inducing degeneration are shown in Table 6. We find that PORAT achieves effective improvements across different experimental settings in both aspect-based datasets. In particular, for the Palate task, under the condition where the skew predictor is trained for 20 epochs, the performance of the previous SOTA model DR, exhibits a degradation trend ($68.2 \rightarrow 66.5 \rightarrow 61.0$). In contrast, PORAT maintains stable performance ($68.3 \rightarrow 68.2 \rightarrow 67.1$). Compared to DR, PORAT achieves up to a 6.1% improvement in mitigating degeneration.

(2) Spurious correlations. We also conduct synthetic experiments to evaluate the effectiveness in addressing spurious correlations as Wu et al. [62] did (The specific setting can be found in Appendices). As shown in Table 7, we can observe that, compared to the previous SOTA method DR, PORAT achieves significant improvements in 8 out of 9 experimental settings. In particular, compared to the backbone model, PORAT does not exhibit any performance degradation, whereas DR shows varying degrees of performance decline in six settings (bias=0.1, 0.3, 0.5, 0.6, 0.7, 0.8). This further indicates that DR is sensitive to data distributions with prominent spurious patterns, while PORAT remains robust under such conditions. Moreover, we can see that PORAT also demonstrates more stable performance since it achieves a smaller mean variance.

6.4 Experiment Analysis

(1) Ablation analysis. To further verify the effectiveness of PORAT, as well as to investigate the impact of the policies adopted by the predictor and generator players, we remove the optimized policies on three benchmarks to conduct experiments. For a fair comparison, we do not specifically tune the hyperparameters and use the same hyperparameters for both PORAT and the modules to be ablated. As depicted in Fig.5, by removing all policies, we observe varying degrees of declination across multiply aspects in the nine datasets, which highlights the effectiveness of the proposed

TABLE 8: Results of methods with low sparsity on BeerAdvocate* benchmark.

Methods	Beer-Appearance*					Beer-Aroma*					Beer-Plate*				
	S	Acc	P	R	F1	S	Acc	P	R	F1	S	Acc	P	R	F1
RNP	11.9	-	72.0	46.1	56.2	10.7	-	70.5	48.3	57.3	10.0	-	53.1	42.8	47.5
CAR	11.9	-	76.2	49.3	59.9	10.3	-	50.3	33.3	40.1	10.2	-	56.6	46.2	50.9
DMR	11.7	-	83.6	52.8	64.7	11.7	-	63.1	47.6	54.3	10.7	-	55.8	48.1	51.7
FR	12.7	83.9	77.6	53.3	63.2	10.8	87.6	82.9	57.9	68.2	10.0	84.5	69.3	55.8	61.8
DR	12.8	83.8	81.8	56.6	66.9	11.5	83.5	65.0	60.6	62.7	11.2	82.3	73.2	58.6	65.1
PORAT (Ours)	11.7	84.2	**90.5**	**57.3**	**70.2**	11.6	84.0	**84.3**	**62.9**	**72.0**	10.8	86.7	71.9	**62.4**	**66.8**

TABLE 9: Experiments on large language model encoder.

Methods	Hotel-Location				Hotel-Service				Hotel-Cleanliness			
	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
In Context Learning (ICL)												
Llama-3.2-1B (ICL)	74.2	5.7	5.7	5.7	83.3	5.6	5.6	5.6	81.0	6.2	6.2	6.2
Llama-3.2-3B (ICL)	75.9	7.7	7.9	7.8	91.3	10.0	10.0	10.0	91.8	6.1	6.1	6.1
Llama-3.1-8B (ICL)	95.9	42.8	42.8	42.8	97.2	35.9	36.0	36.0	94.3	23.5	23.6	23.6
Supervised Fine-Tuning (SFT)												
Llama-3.2-1B (SFT)	56.8	11.0	10.8	10.9	57.6	11.3	11.5	11.4	57.9	8.8	8.8	8.8
Llama-3.2-3B (SFT)	95.2	37.4	37.4	37.4	88.5	30.2	30.3	30.2	95.5	17.5	17.5	17.5
Llama-3.1-8B (SFT)	84.8	34.0	34.1	34.0	90.5	35.3	35.4	35.4	92.0	24.9	25.0	25.0
PORAT (Ours)	94.0	**53.4**	**63.1**	**57.8**	95.5	**45.0**	**51.8**	**48.2**	93.5	**38.7**	**46.4**	**42.2**

TABLE 10: Experiments on small language model encoder.

Methods	Hotel-Location				
	S	Acc	P	R	F1
MCD-BERT-Tiny	9.4	85.0	14.7	16.4	15.5
MCD-BERT-Tiny-OOD	10.1	89.0 ↑	**8.5**	**10.4**	**9.4** ↓
MCD-BERT-Tiny-PORAT	9.8	86.5 ↑	**16.2**	**18.9**	**17.4** ↑

Methods	Hotel-Service				
	S	Acc	P	R	F1
MCD-BERT-Tiny	10.7	87.5	14.5	13.5	14.0
MCD-BERT-Tiny-OOD	10.6	94.0 ↑	**12.1**	**11.3**	**11.7** ↓
MCD-BERT-Tiny-PORAT	10.6	94.5 ↑	**18.0**	**18.2**	**18.1** ↑

Methods	Hotel-Cleanliness				
	S	Acc	P	R	F1
MCD-BERT-Tiny	10.4	89.5	17.8	20.6	19.1
MCD-BERT-Tiny-OOD	10.0	97.0 ↑	**7.8**	**8.7**	**8.3** ↓
MCD-BERT-Tiny-PORAT	9.6	91.0 ↑	**20.4**	**21.8**	**21.1** ↑

Fig. 6: Low sparsity ablation results.

framework. In addition, additional ablation studies on the spurious correlations synthetic experiments (Table 7) also highlight the importance of both the generator (g.) and the predictor (p.) policy interventions in PORAT.

(2) Low-sparsity analysis. The low-sparsity experiments demonstrate the RNP model's robustness [20]. Using the same settings [20], [33], [67], we also conduct an experiment where the sparsity of selected rationales is extremely low. The results are presented in Table 8. We can observe that PORAT also effectively improves the models' robustness compared to previous models. Besides, we also conduct the ablation results on low-sparsity experiments and observe the impact across multiple aspects (Fig.6). This consistently demonstrates the improvement of PORAT in terms of robustness, which indicates the diversity advantage.

(3) Analysis with pretrained language model encoder. Then, how do pretrained language models perform in terms of self-explanation? So, we further compare with language models encoders, including fine-tuning (FT)-based, prompt-based and supervised FT (SFT)-based methods, with ≤1B, 3B, and 8B parameters. As shown in Table 9, we ob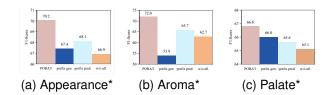serve that although various types of models can achieve high predictive accuracy, generating self-explainable rationales remains a challenge for language models, and our method outperforms language models with fewer than 8 billion parameters. Meanwhile, as shown in Table 10, we also employ language models as backbones to show the competitiveness of our PORAT. Consistent with most research [10], [11], [32], here we also use the BERT encoder as a backbone for a fair comparison. We observe that our proposed method PORAt not only improves predictive performance but also substantially enhances the self-explanation performance.

6.5 Optimization and Parameter Analysis

(1) Analysis of different prefix-player policy optimization. We also investigate the results of intervening with different prefix player policies. We discover that the impact of policy optimization for different players varies across different datasets. As shown in Fig.5, in the BeerAdvocate benchmark, intervening in the prefix predictor of the baseline model effectively helps the model escape the suboptimal state, leading to a significant improvement in model performance (from 85.8/73.2/53.5 to 86.1/75.1/58.0). However, this is not the case on the other two benchmarks, which may be related to the inherent difficulty of the datasets and their distribution. The experiments with the out-of-distribution (OOD) method in Table 10 further validate that while the distribution improves predictive performance, it also impacts the self-explanatory capability. Therefore, optimizing

TABLE 11: Examples of generated rationales from DR and PORAT. Human-annotated rationales are <u>underlined</u>. Rationales selected by the generator are highlighted in red, where "a", "s" and "t" in the input text indicate the rationales annotated by annotators from appearance, aroma and palate aspects.

DR	PORAT
Aspect: Beer-Aroma **Label:** Positive, **Prediction:** Positive **Input Text:** ... a : pours a clear golden color with a huge 4-finger white head that lasts forever s : spicy phenolic aroma with hints of hops . t : smooth heavy malt brew with sweet spices and alcohol in the background . herbal hops and tad fruity finish . m : medium body and high carbonation . o : very sweet beer - unique - but nothing i would grab again .	**Aspect:** Beer-Aroma **Label:** Positive, **Prediction:** Positive **Input Text:** ... a : pours a clear golden color with a huge 4-finger white head that lasts forever s : spicy phenolic aroma with hints of hops . t : smooth heavy malt brew with sweet spices and alcohol in the background . herbal hops and tad fruity finish . m : medium body and high carbonation . o : very sweet beer - unique - but nothing i would grab again .
Aspect: Beer-Aroma **Label:** Positive, **Prediction:** Positive **Input Text:** ... the pour a clear deep amber , the head is mediorce , the lace spare , the color off white . nose is malt , citrus tones , light hints of bubble gum . front is malt , sweet , the top is medium , the finish is acerbic , dry , the 10 % abv , is felt in the 'tummy ' and the long lasting alcohol bitter aftertaste . works for me ! , as i like my beers pungent and brawny . ranks # 504 on my current 1000 beer master list .	**Aspect:** Beer-Aroma **Label:** Positive, **Prediction:** Positive **Input Text:** ... the pour a clear deep amber , the head is mediorce , the lace spare , the color off white . nose is malt , citrus tones , light hints of bubble gum . front is malt , sweet , the top is medium , the finish is acerbic , dry , the 10 % abv , is felt in the 'tummy ' and the long lasting alcohol bitter aftertaste . works for me ! , as i like my beers pungent and brawny . ranks # 504 on my current 1000 beer master list .
Aspect: Beer-Aroma **Label:** Positive, **Prediction:** Positive **Input Text:** ... reviewed halloween evening , 2009 . poured a very nice deep copper color with fantastic head and lacing . great scent , very deep bitter aromas , a lot of citrus tones and a slight pine tinge . great taste , a nice deep maltiness with a fantastic bitter ending ; very nice american hops (citrus) with a nice earthy undertone to it . goes down very nice , with just the slightest hop roughness . great beer .	**Aspect:** Beer-Aroma **Label:** Positive, **Prediction:** Positive **Input Text:** ... reviewed halloween evening , 2009 . poured a very nice deep copper color with fantastic head and lacing . great scent , very deep bitter aromas , a lot of citrus tones and a slight pine tinge . great taste , a nice deep maltiness with a fantastic bitter ending ; very nice american hops (citrus) with a nice earthy undertone to it . goes down very nice , with just the slightest hop roughness . great beer .

the strategy for different pattern distributions helps improve performance in rationalization.

(2) Analysis of time interval for optimization. To gain an insight into the effects of selecting different interval of timestep N, we also conduct the analysis experiments varying N from 1 to 10. As Fig.7(a-b) show, we can observe that the impact of policy intervention time intervals is relatively minor. In contrast, applying policy interventions solely to either the generator or the predictor results in poorer model stability, whereas their joint presence leads to more stable performance. This further strengthens the effectiveness of our proposed method and validates the feasibility of the theoretical framework. In addition, we also provide analysis with the longer interval from 20 to 200 in Fig.7(c-d), which once again validates the previous conclusion.

6.6 Visualization Analysis

In Table 11, we also visualize the rationales generated using recent model DR and our PORAT intuitively. We find that although both models provide correct predictions, DR instead correlates with other aspects. This indicates that DR has already been able to focus on explanations in the Aroma aspect. However, it fails to simultaneously address spurious correlations or other forms of suboptimal rationales well, while PORAT demonstrates superior capability. One possible reason is that DR only focuses on degeneration to employ a smaller lipschitz constant to capture semantically closer rationale candidates [20]. But this still limits its scalability in addressing other pattern problems.

7 CONCLUSION AND FUTURE WORK

In this paper, we propose PORAT, a policy optimization-based data-centric self-explanation rationalization method.

(a) Aroma*-10   (b) Palate*-10

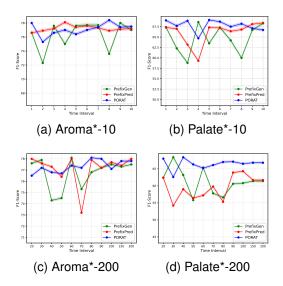(c) Aroma*-200   (d) Palate*-200

Fig. 7: Analysis of policy optimization timestep.

We first systematically revisit the cooperative game mechanism of rationalization in a novel game-theoretic perspective, and reveal the game-theoretic problem between two players in rationalization. Then we theoretically analyze the causes of the game-theoretic problem between two players in rationalization and also prove the feasibility of the proposed method. Extensive experiments on nine widely used real-world datasets and two synthetic settings show that our proposed method significantly improves performance and outperforms several recently published SOTA methods. Furthermore, experiments on ablation studies, low-sparsity analysis, and language models analysis demonstrate the effectiveness and diversity of PORAT. Moving forward, we plan to explore the feasibility of rationalizing predictions for

large generative language models such as self-explanation foundation model, and further study other way to address the collapsed self-rationales in the field of rationalization.

REFERENCES

[1] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[2] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 4198–4205, Association for Computational Linguistics, July 2020.

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

[4] A. Bhattacharya, *Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more.* Packt Publishing Ltd, 2022.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[8] R. Menon, K. Zaman, and S. Srivastava, "MaNtLE: Model-agnostic natural language explainer," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 13493–13511, Association for Computational Linguistics, Dec. 2023.

[9] M. Yu, S. Chang, Y. Zhang, and T. S. Jaakkola, "Rethinking cooperative rationalization: Introspective extraction and complement control," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[10] W. Liu, Z. Deng, Z. Niu, J. Wang, H. Wang, and R. Li, "Exploring practical gaps in using cross entropy to implement maximum mutual information criterion for rationalization," *Transactions of the Association for Computational Linguistics*, vol. 13, pp. 577–594, 2025.

[11] W. Liu, Z. Niu, L. Gao, Z. Deng, J. Wang, H. Wang, and R. Li, "Adversarial cooperative rationalization: The risk of spurious correlations in even clean datasets," *arXiv preprint arXiv:2505.02118*, 2025.

[12] T. Lei, R. Barzilay, and T. Jaakkola, "Rationalizing neural predictions," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (J. Su, K. Duh, and X. Carreras, eds.), (Austin, Texas), pp. 107–117, Association for Computational Linguistics, Nov. 2016.

[13] L. Yue, Q. Liu, Y. Du, L. Wang, W. Gao, and Y. An, "Towards faithful explanations: Boosting rationalization with shortcuts discovery," in *The International Conference on Learning Representations*, 2024.

[14] Y. Zhao, Z. Wang, X. Li, J. Liang, and R. Li, "AGR: Reinforced causal agent-guided self-explaining rationalization," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (L.-W. Ku, A. Martins, and V. Srikumar, eds.), pp. 510–518, Aug. 2024.

[15] W. Liu, H. Wang, J. Wang, Z. Deng, Y. Zhang, C. Wang, and R. Li, "Enhancing the rationale-input alignment for self-explaining rationalization," in *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 2218–2230, IEEE, 2024.

[16] S. Hu and K. Yu, "Learning robust rationales for model explainability: A guidance-based approach," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 18243–18251, 2024.

[17] H. Jiang, J. Duan, Z. Qu, and J. Wang, "MARE: Multi-aspect rationale extractor on unsupervised rationale extraction," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, eds.), (Miami, Florida, USA), pp. 11734–11745, Association for Computational Linguistics, Nov. 2024.

[18] W. Zhang, T. Wu, Y. Wang, Y. Cai, and H. Cai, "Towards trustworthy explanation: on causal rationalization," in *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, JMLR.org, 2023.

[19] A. Storek, M. Subbiah, and K. McKeown, "Unsupervised selective rationalization with noise injection," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 12647–12659, Association for Computational Linguistics, July 2023.

[20] W. Liu, J. Wang, H. Wang, R. Li, Y. Qiu, Y. Zhang, J. Han, and Y. Zou, "Decoupled rationalization with asymmetric learning rates: A flexible lipschitz restraint," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1535–1547, 2023.

[21] W. Liu, H. Wang, J. Wang, R. Li, C. Yue, and Y. Zhang, "Fr: Folded rationalization with a unified encoder," *Advances in Neural Information Processing Systems*, vol. 35, pp. 6954–6966, 2022.

[22] L. Sha, O.-M. Camburu, and T. Lukasiewicz, "Rationalizing predictions by adversarial information calibration," *AI Magazine*, vol. 315, p. 103828, 2023.

[23] H. Yuan, L. Cai, X. Hu, J. Wang, and S. Ji, "Interpreting image classifiers by generating discrete masks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2019–2030, 2020.

[24] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized explainer for graph neural network," *Advances in neural information processing systems*, vol. 33, pp. 19620–19631, 2020.

[25] L. Yue, Q. Liu, B. Jin, H. Wu, and Y. An, "A circumstance-aware neural framework for explainable legal judgment prediction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 5453–5467, 2024.

[26] Z. Deng, J. Li, Z. Guo, and G. Li, "Multi-aspect interest neighbor-augmented network for next-basket recommendation," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

[27] M. Yu, Y. Zhang, S. Chang, and T. Jaakkola, "Understanding interlocking dynamics of cooperative rationalization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12822–12835, 2021.

[28] W. Liu, Z. Deng, Z. Niu, J. Wang, H. Wang, R. Li, and Y. Zhang, "Is the mmi criterion necessary for explanation? degenerating non-causal features to plain noise," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024.

[29] W. Liu, H. Wang, J. Wang, R. Li, X. Li, Y. Zhang, and Y. Qiu, "MGR: Multi-generator based rationalization," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 12771–12787, Association for Computational Linguistics, July 2023.

[30] S. Chang, Y. Zhang, M. Yu, and T. Jaakkola, "Invariant rationalization," in *International Conference on Machine Learning*, pp. 1448–1458, PMLR, 2020.

[31] L. Yue, Q. Liu, L. Wang, Y. An, Y. Du, and Z. Huang, "Interventional rationalization," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 11404–11418, Association for Computational Linguistics, Dec. 2023.

[32] W. Liu, J. Wang, H. Wang, R. Li, Z. Deng, Y. Zhang, and Y. Qiu, "D-separation for causal self-explanation," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[33] Y. Huang, Y. Chen, Y. Du, and Z. Yang, "Distribution matching for rationalization," in *AAAI Conference on Artificial Intelligence*, 2021.

[34] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, "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.

[35] E. Lehman, J. DeYoung, R. Barzilay, and B. C. Wallace, "Inferring which medical treatments work from reports of clinical trials," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 3705–3717, Association for Computational Linguistics, June 2019.

[36] D. Li, B. Hu, Q. Chen, T. Xu, J. Tao, and Y. Zhang, "Unifying model explainability and robustness for joint text classification and rationale extraction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 10947–10955, 2022.

[37] Q. Lyu, M. Apidianaki, and C. Callison-Burch, "Towards faithful model explanation in NLP: A survey," *Computational Linguistics*, vol. 50, pp. 657–723, June 2024.

[38] Y. Bao, S. Chang, M. Yu, and R. Barzilay, "Deriving machine attention from human rationales," in *Proceedings of Empirical Methods in Natural Language Processing*, pp. 1903–1913, 2018.

[39] J. Bastings, W. Aziz, and I. Titov, "Interpretable neural predictions with differentiable binary variables," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A. Korhonen, D. Traum, and L. Màrquez, eds.), pp. 2963–2977, July 2019.

[40] B. Paranjape, M. Joshi, J. Thickstun, H. Hajishirzi, and L. Zettlemoyer, "An information bottleneck approach for controlling conciseness in rationale extraction," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (B. Webber, T. Cohn, Y. He, and Y. Liu, eds.), (Online), pp. 1938–1952, Association for Computational Linguistics, Nov. 2020.

[41] L. Yue, Q. Liu, Y. Du, Y. An, L. Wang, and E. Chen, "Dare: Disentanglement-augmented rationale extraction," *Advances in Neural Information Processing Systems*, vol. 35, pp. 26603–26617, 2022.

[42] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[43] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 10014–10037, Association for Computational Linguistics, July 2023.

[44] O. Press, M. Zhang, S. Min, L. Schmidt, N. Smith, and M. Lewis, "Measuring and narrowing the compositionality gap in language models," in *Findings of the Association for Computational Linguistics: EMNLP 2023* (H. Bouamor, J. Pino, and K. Bali, eds.), pp. 5687–5711, Dec. 2023.

[45] Z. Jiang, F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig, "Active retrieval augmented generation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (H. Bouamor, J. Pino, and K. Bali, eds.), pp. 7969–7992, Dec. 2023.

[46] W. Su, Y. Tang, Q. Ai, Z. Wu, and Y. Liu, "DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (L.-W. Ku, A. Martins, and V. Srikumar, eds.), pp. 12991–13013, Aug. 2024.

[47] E. Kiciman, R. Ness, A. Sharma, and C. Tan, "Causal reasoning and large language models: Opening a new frontier for causality," *Transactions on Machine Learning Research*, 2023.

[48] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM computing surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[49] I. Arcuschin, J. Janiak, R. Krzyzanowski, S. Rajamanoharan, N. Nanda, and A. Conmy, "Chain-of-thought reasoning in the wild is not always faithful," *arXiv preprint arXiv:2503.08679*, 2025.

[50] M. Turpin, J. Michael, E. Perez, and S. Bowman, "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting," *Advances in Neural Information Processing Systems*, vol. 36, pp. 74952–74965, 2023.

[51] L. Yuan, S. Hu, K. Yu, and L. Wu, "Boosting explainability through selective rationalization in pre-trained language models," *arXiv preprint arXiv:2501.03182*, 2025.

[52] R. S. Sutton, "Reinforcement learning: An introduction," *A Bradford Book*, 2018.

[53] T. Degris, P. M. Pilarski, and R. S. Sutton, "Model-free reinforcement learning with continuous action in practice," in *2012 American control conference (ACC)*, pp. 2177–2182, IEEE, 2012.

[54] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.), vol. 48 of *Proceedings of Machine Learning Research*, (New York, New York, USA), pp. 1928–1937, PMLR, 20–22 Jun 2016.

[55] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in neural information processing systems*, vol. 12, 1999.

[56] K. Kawaguchi, "Deep learning without poor local minima," *Advances in neural information processing systems*, vol. 29, 2016.

[57] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Artificial intelligence and statistics*, pp. 192–204, PMLR, 2015.

[58] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018.

[59] C. Yun, S. Sra, and A. Jadbabaie, "Global optimality conditions for deep neural networks," in *International Conference on Learning Representations*, 2018.

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

[61] H. Wang, Y. Lu, and C. Zhai, "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*, KDD '10, (New York, NY, USA), p. 783–792, Association for Computing Machinery, 2010.

[62] Y. Wu, X. Wang, A. Zhang, X. He, and T.-S. Chua, "Discovering invariant rationales for graph neural networks," in *International Conference on Learning Representations*, 2022.

[63] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (A. Moschitti, B. Pang, and W. Daelemans, eds.), pp. 1724–1734, Oct. 2014.

[64] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (A. Moschitti, B. Pang, and W. Daelemans, eds.), pp. 1532–1543, Oct. 2014.

[65] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, (San Diega, CA, USA), 2015.

[66] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," *Advances in neural information processing systems*, vol. 32, 2019.

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