A Shapley-value Guided Rationale Editor for Rationale Learning

Zixin Kuang
zkua236@aucklanduni.ac.nz

Meng-Fen Chiang
meng.chiang@nycu.edu.tw

Wang-Chien Lee
wlee@cse.psu.edu

Abstract

Rationale learning aims to automatically uncover the underlying explanations for NLP predictions. Previous studies in rationale learning mainly focus on the relevance of independent tokens with the predictions without considering their marginal contribution and the collective readability of extracted rationales. Through an empirical analysis, we argue that the sufficiency, informativeness, and readability of rationales are essential for explaining diverse end-task predictions. Accordingly, we propose **Sha**pley-value Guided **R**ationale **E**ditor (SHARE), an unsupervised approach that refines editable rationales while predicting task outcomes. SHARE extracts a sequence of tokens as a rationale, providing a collective explanation that is sufficient, informative, and readable. SHARE is highly adaptable for tasks like sentiment analysis, claim verification, and question answering, and can integrate seamlessly with various language models to provide explainability. Extensive experiments demonstrate its effectiveness in balancing sufficiency, informativeness, and readability across diverse applications. Our code and datasets are available at https://github.com/zixinK/SHARE.

1 INTRODUCTION

Modern language models are increasingly outpacing traditional machine learning techniques in natural language processing (NLP) (Zini and Awad, 2022; Lauriola et al., 2022). Despite their effectiveness, these models are often perceived as black boxes due to their limited interpretability. To tackle the challenge of explaining NLP model predictions, researchers have explored automatic learning methods (Carton et al., 2022; Lei et al., 2016; Chen et al., 2018; Liu et al., 2023a). The goal is to extract a subset of input tokens, termed 'rationale', which can provide insights into the decisions made by language models (Carton et al., 2022). Previous research has concentrated on assessing the significance of individual tokens for inclusion in the rationale (Lei et al., 2016; DeYoung et al., 2020; Chan et al., 2022b; Paranjape et al., 2020). Some existing methods treat each token as an independent unit, basing decisions solely on individual significance. This approach often results in disjointed sentences and incoherent logic within the extracted rationales. Additionally, other studies prioritize sufficiency while neglecting critical aspects of extracted rationales, such as the potential for information leakage from non-rationales and the overall readability of the rationales, as highlighted in (Chen et al., 2019).

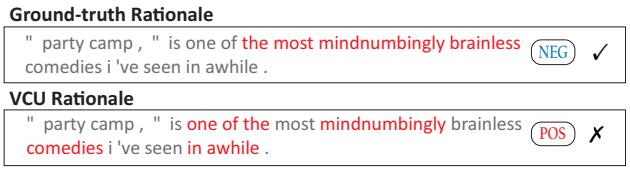

Figure 1: Comparison of Rationale Quality on Movie Review sentiment analysis (DeYoung et al., 2020): Ground-truth vs. VCU-extracted rationales (Schuster et al., 2021). POS/NEG are the predicted end-task labels using the extracted rationales. ✓and ✗indicate the correctness of the predictions.

To demonstrate the advantages of linked phrases and adjacent tokens for comprehension compared to isolated tokens, we analyze human-annotated 'ground-truth' rationales alongside those extracted by the notable VCU model (Schuster et al., 2021) on the Movie Review dataset (DeYoung et al., 2020).[1] Our quantitative analysis, which meticulously evaluates multiple aspects, is presented in Table 1. In contrast to the ground-truth rationale depicted in Figure 1, which provides meaningful insights and logical coherence for understanding negative sentiment predictions, our analysis reveals the following observations regarding the rationales extracted by VCU. (i) **Information Leak from Non-rationale Content:** We observe that the average prediction accuracy from the ground-truth non-rationale (0.527) is sig-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

[1]The unsupervised rationale extraction model presented in (Schuster et al., 2021) is referred to as VCU.

Table 1: Rationale Quality Comparison on Movie Review (DeYoung et al., 2020): Ground-truth rationale (GT) vs. extracted rationales by VCU (Schuster et al., 2021) in terms of Rationale ACC, Non-rationale ACC, and Rationale TD, which refer to rationale prediction accuracy, non-rationale prediction accuracy, and average token distance (an indicator of continuity in rationale), respectively.

	Rationale ACC	Non-rationale ACC	Rationale TD
GT	0.976	0.527	0.253
VCU	0.769	0.875	0.668

nificantly lower than that from the VCU-determined non-rationale (0.875). This indicates that the VCU-determined non-rationale still contains critical information relevant to the target prediction, leading to potential information leakage. In contrast, the ground-truth non-rationale provides limited explanatory value for the target predictions. (ii) **Information Loss from Extracted Rationale Content:** The average prediction accuracy of the VCU-extracted rationale (0.769) is lower than that of the ground-truth rationale (0.976), suggesting that VCU-extracted rationales inadequately support prediction choices (DeYoung et al., 2020). This comparison confirms the presence of information loss in the VCU-extracted rationale. (iii) **Fragmented Rationale Token Sequence:** Continuity in rationales is another critical aspect often overlooked. Our analysis evaluates the continuity of both ground-truth and VCU-extracted rationales using the token distance metric (TD). The comparison reveals that VCU-extracted rationales exhibit a fragmented token sequence, with a score of 0.668, compared to the ground-truth rationales' score of 0.253.

Research Objective. Motivated by the aforementioned considerations, we propose a novel unsupervised framework called **Sha**pley-value Guided **R**ationale **E**ditor (**SHARE**). SHARE extracts readable rationales that both sufficiently support and informatively explain predictions. The rationale content is iteratively refined to maximize relevant information while preventing information leakage from non-rationale content. Specifically, SHARE comprises three key components: (i) The **Rationale Extractor**, inspired by (Schuster et al., 2021; Chen et al., 2019), extracts rationales from articles to support corresponding end-task labels. The rationale extractor captures token interactions by combining individual token relevance with marginal contributions, resulting in collective tokens that serve as sufficient and informative rationales. (ii) The **Rationale Editor** iteratively refines the readability, in terms of continuity, of initial rationales, using four editing operations: insertion, deletion, replacement, and no operation. (iii) The **End-task Pseudo Prediction** component considers sufficiency, informativeness, and readability in a joint learning objective to optimize the rationales during learning. Note that SHARE is model-agnostic, compatible with state-of-the-art language models, and task-agnostic, adaptable to various tasks such as sentiment analysis, claim verification, and question answering. Our contributions can be summarized as follows.

- We propose SHARE, an unsupervised self-explaining framework that extracts readable rationales to sufficiently and informatively explain task predictions. SHARE seamlessly integrates into diverse applications requiring explainability and accommodates any chosen language model.
- SHARE enhances the sufficiency, informativeness, and readability of rationales by considering both independent and marginal relevance in explaining predictions.
- The proposed rationale editor, designed as a plug-and-play component, extracts editable rationales by balancing continuity and integrated token scores through four editing operations.
- Extensive experiments, case studies, and thorough analyses validate SHARE's effectiveness, explainability, and rationality across various end-tasks.

2 PRELIMINARIES

2.1 Problem Statement

Given articles $\mathcal{X}$ with ground-truth labels $\mathcal{Y}$, we aim to learn the function $\mathcal{F}_R : X_i \rightarrow X_i^R$ where $X_i \in \mathcal{X}$. $\mathcal{F}_R$ generates a continuous rationale X_i^R from X_i that provides a logically coherent explanation for the prediction made by the end-task classifier $\mathcal{F}_C : X_i^R \rightarrow Y_i$, i.e., the end-task classifier uses the learned rationale X_i^R to predict the labels $Y_i \in \mathcal{Y}$.

2.2 Independent Token Relevance

To understand the reasoning behind tasks like claim verification, token selection plays a crucial role in forming a comprehensible interpretation. Recent studies have introduced various "independent token relevance" (Lei et al., 2016; Voskarides et al., 2015; Chen et al., 2018) while leveraging Transformer attentions (Voita et al., 2019) to finalize token selections.

Given a pre-trained classifier $\mathcal{F}_C$, an article $X_i \in \mathcal{X}$ and its end-task ground-truth label $Y_i \in \mathcal{Y}$, let X_{ij} denote the j-th token within the article X_i. The independent relevance of a token X_{ij} in relation to the end-task label Y_i can be measured using the conditional probability $P\left(Y_i \mid X_{ij}\right)$ for Y_i, given the token X_{ij} and the classifier model $\mathcal{F}_C$. The independent relevance of token X_{ij}, as described in (Chen et al., 2018), is formulated as follows.

$$v_{X_i,Y_i}(j) := \hat{E}\left[-\log \frac{1}{P\left(Y_i \mid X_{ij}\right)} \mid X_i\right] \qquad (1)$$

where $\hat{E}\left[\cdot|X_i\right]$ denotes the expectation over $P\left(\cdot \mid X_i\right)$. The token relevance, as discussed, solely concentrates on the

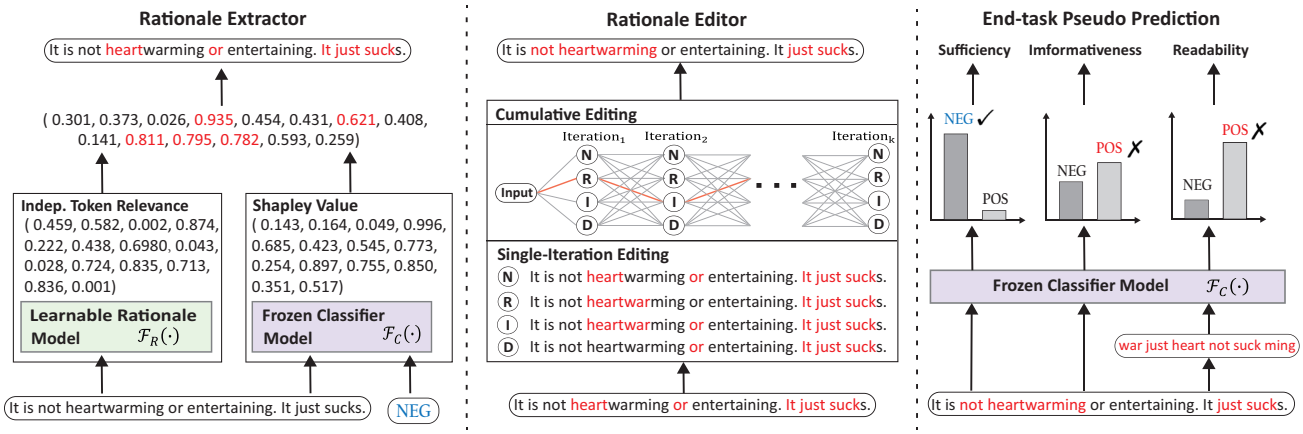

Figure 2: The overview of the SHARE framework. The Rationale Extractor, based on the independent token relevance and the Shapley value, generates an initial rationale. The Rationale Editor enhances continuity and token scores within the rationale with four operations: *D* (deletion), *I* (insertion), *R* (replacement), and *N* (no operation). The End-task Pseudo Prediction uses the extracted rationales for end-task prediction, calibrating subsequent training iterations.

isolated relevance of each individual token concerning label prediction, neglecting other potential factors. Thus, in Section 2.3, we introduce the concept of marginal contribution to account for interactions among relevance.

2.3 Marginal Token Relevance

Using independent relevance to assess the importance of a given token X_{ij} for an article X_i is a standard practice. However, independent relevance fails to account for interactions between tokens, which can significantly influence the final outcome. For instance, consider the calculation of independent relevance for the token "not" in the sentences:

It is not a bad thing.
It is not a good thing.

The word "not" considerably affects the likelihood of the prediction label being positive or negative, indicating its importance for label prediction. However, when computing the independent relevance score, "not" is treated as a standalone entity. Under this approach, it is perceived as a neutral token since it does not convey emotional information. Consequently, the independent token relevance score of "not" would be zero, contrary to its actual significance in the given context. The *Shapley value*, introduced in (Shapley, 1953), quantifies the collective contribution of each player in a coalition to the overall worth. The Shapley value, applied in various fields, offers a principled approach to distributing marginal rewards in a cooperative setting (Rozemberczki et al., 2022). Treating each token as a player, we define the token's marginal contribution of relevance to the overall value of the token coalition (i.e., rationale) using the notion of marginal contribution in rationale learning. Our objective is to evaluate the Shapley value for each token X_{ij}, in response to the target variable $Y_i \in \mathcal{Y}$. Notably, this objective incurs a high computational cost of $2^{|X_i|-1}$ due to the complexity involved.

Marginal Contribution. The current form of the independent token relevance overlooks token interactions, which significantly affects the final outcome. For instance, assume a token subset X_{iS}, where $S \subseteq \{1, \cdots, |X_i|\}$ represents an arbitrary subset of token indices in the article X_i. When analyzing token subset X_{iS} containing token X_{ij} in article X_i, the marginal contribution of X_{ij} to X_{iS} represents the difference in the independent token relevance with and without X_{ij}. To account for these differences, the marginal contribution is defined as follows.

$$m_{X_i,Y_i}(S, j) := v_{X_i,Y_i}(S) - v_{X_i,Y_i}(S \backslash \{j\}). \quad (2)$$

Marginal Contribution of Neighbors (Shapley Value). Considering all possible token subsets in X_i, the brute-force Shapley value of token X_{ij} is calculated by averaging its marginal contributions across all token subsets in X_i. However, this computation is expensive. To address this, we approximate the Shapley value using the *connected Shapley value* (Chen et al., 2019). Let $C_k(p)$ denote the set of token subsets that include up to k-grams in the neighborhood of token X_{ij}. The connected Shapley value is defined as follows.

$$\tilde{\phi}^k_{X_i,Y_i}(j) := \sum_{S \in \mathcal{C}_k(p)} \frac{2 \cdot m_{X_i,Y_i}(S, j)}{(|S|+2)(|S|+1)|S|}. \quad (3)$$

3 METHODOLOGY

We argue that a sound rationale should possess the following qualities: (i) Sufficiency: It should effectively and sufficiently reveal the prediction for the end-task. (ii) Informativeness: It should contain meaningful information, facilitating comprehension of the prediction for the end-task at

hand. (iii) Readability: It should exhibit logical coherence, while enhancing readability and facilitating understanding.

To address these aspects, we propose the SHARE framework (see Figure 2), which consists of three phases: *Rationale Extractor*, *Rationale Editor*, and *End-task Pseudo Prediction*. The *Rationale Extractor* is a tunable model F_R responsible for rationale extraction, while the *End-task Pseudo Prediction* employs a fixed classifier F_C for the final classification task. In the initial phase, each token in an article receives a multi-faceted score, and the rationale is formed by selecting tokens with scores above a threshold, treating the rest as noise. In the second phase, the rationale editor improves the rationale's continuity while maintaining satisfactory token scores through token deletion, insertion, and replacement, guided by sufficiency, informativeness, and readability. Finally, the end-task pseudo prediction utilizes the refined rationales for task prediction, calibrating subsequent training iterations. The final learning objective is formulated, from an information-theoretic perspective, as follows.

$$\min \lambda_S \mathcal{L}_S + \lambda_I \mathcal{L}_I + \lambda_R \mathcal{L}_R \tag{4}$$

where $\mathcal{L}_S$ denotes the sufficiency loss, which quantifies the rationale's capacity to convey the decision-making process of the end-task model $\mathcal{F}_C(\cdot)$. $\mathcal{L}_I$ denotes the information loss, reflecting the rationale's informativeness. Lastly, $\mathcal{L}_R$ captures the readability loss, evaluating the coherence of token ordering within the rationale. The coefficients λ_S, λ_I, and λ_R, regulate the influence of the rationale quality factors on the overall objective and are empirically determined through extensive experiments, taking into account the rationale length in relation to the specific tasks and data characteristics.

3.1 Rationale Extractor

To account for multiple facets, it is crucial to distinguish between rationale and non-rationale portions of an article. The Rationale Extractor phase aims to identify a sequence of tokens that collectively represents a rationale of desired quality. To accomplish this, we derive a multi-faceted token score for each token X_{ij} in an article $X_i \in \mathcal{X}$. The token score $s_{X_i}(j)$ is determined by averaging the independent token relevance $\alpha_{X_i,Y}(j)$ and the marginal token relevance $\beta_{X_i,Y}(j)$, as detailed in subsequent subsections, where Y is $\overline{Y_i}$ or Y_i. This combined approach maximizes information extraction by integrating feature perturbation and attention-based importance values. Tokens with scores exceeding a threshold (γ) are classified as rationale (X^R), while those below the threshold are considered noise (X^N).

3.1.1 Independent Token Relevance Score

Let $S = \{X_{ij}\}$, Eq.(1) is used to measure the relevance of token X_{ij} in article X_i to the prediction $\hat{Y}_i$. It quantifies the importance of the j-th token in X_i with respect to the ground-truth label Y_i as $v_{X_i,Y_i}(j)$, and its importance with respect to the false label $\overline{Y_i}$ as $v_{X_i,\overline{Y_i}}(j)$. To regulate these independent token relevance scores, the Gumbel-Softmax function $GS(\cdot)$ is used. $GS(\cdot)$ introduces random perturbation and ensures the regularized differentiability of the rationale learning process (Jang et al., 2017; Chen et al., 2018).

$$\alpha_{X_i,Y}(j) = GS(v_{X_i,Y}(j)), \tag{5}$$

where Y is $\overline{Y_i}$ or Y_i. The resulting independent token relevance scores $\alpha_{X_i,\overline{Y_i}}(j)$ and $\alpha_{X_i,Y_i}(j)$ indicate the likelihood of token X_{ij} being unselected or selected as part of the rationale, respectively. These scores range between 0 and 1. A token with high independent relevance may contain redundant information already present in the selected tokens. Therefore, the independent token relevance of marginal impact lies in determining the priorities of token candidates based on their potential contribution to the overall collection of information from the selected candidates.

3.1.2 Shapley Value

To capture token interactions within the article X_i, we consider the marginal contribution of each token X_{ij} in X_i using Eq.(2). The Shapley value quantifies the overall marginal contribution of each token from a collective perspective. To study collective decision-making in token selection, we freeze the classifier $\mathcal{F}_C(\cdot)$. This keeps the Shapley values of the tokens constant during the training process, allowing the rationale extractor $\mathcal{F}_R(\cdot)$ to focus on optimizing the token collection.

Given an end-task classifier $\mathcal{F}_C(\cdot)$, we calculate the connected Shapley values $\tilde{\phi}^k_{X_i,Y_i}(j)$ and $\tilde{\phi}^k_{X_i,\overline{Y_i}}(j)$ for each token. These values correspond to the target and false labels, respectively, computed based on Eq.(3) with the neighborhood scope k. Further hyperparameter details are available in the Appendix. To ensure smooth differentiability in the learning process, we utilize the Gumbel-Softmax function $GS(\cdot)$ as follows (Gumbel, 1954; Jang et al., 2017).

$$\beta_{X_i,Y}(j) = GS(\tilde{\phi}^k_{X_i,Y}(j)), \tag{6}$$

where Y is $\overline{Y_i}$ or Y_i. The regularized Shapley values $\beta_{X_i,\overline{Y_i}}(j)$ and $\beta_{X_i,Y_i}(j)$ indicate the likelihood of token X_{ij} being marginally unselected or selected as part of the rationale, respectively. These values range between 0 and 1.

3.2 Rationale Editor

Our analysis of human-annotated rationales reveals the collective effect of linked phrases and adjacent tokens, emphasizing the importance of continuity. In light of this insight, we design a sub-token level rationale editor aimed at minimizing the overall distance between rationale tokens in an article X_i, while ensuring satisfactory token scores for each token X_{ij} for all $j \in [1, |X_i^R|]$. In this editing phase, the

Table 2: An example, in a single-step editing of the Rationale Editor, on a movie review snippet with a negative sentiment. ↑ (↓) indicates the increase (decrease) of the estimator and its components in Eq.(9), compared to no operation (N).

	Article	Score	Distance	Benefit
(D)	this film is **extra**ordin**arily** hor**re**ndous and i 'm not going to waste any more words on it .	2.19 (↓)	7 (↓)	0.31 (↑)
(I)	this film is **extra**or**dinarily** hor**re**ndous and i 'm not going to **waste** any more words on it .	3.29 (↑)	17 (↓)	0.19 (↑)
(R)	this film is **extra**or**dinarily** hor**re**ndous and i 'm not going to waste any more words on it .	2.66 (↓)	6 (↓)	**0.44** (↑)
(N)	this film is **extra**ordin**arily** hor**re**ndous and i 'm not going to **waste** any more words on it .	2.83	18	0.16

extent of modification at the sub-token level indicates higher relevance in the explanation, with words undergoing more editing operations more likely to be part of the rationale.

3.2.1 Token Distance

To measure the overall distance between rationale tokens within an article, we introduce the concept of *token distance*, which indicates the continuity of the extracted rationale. For each token $X_{ij} \in X_i$, the distance of token X_{ij} to other tokens in the rationale X_i^R is defined as follows.

$$D_{X_i}(j) = \min(\{d(X_{ij}, X_{ip}) \mid \forall X_{ip} \in X_i^R\}, X_{ij} \neq X_{ip}), \tag{7}$$

where token distance (d) is defined as the index difference between the j-th and p-th tokens in X_i^R. The overall token distance for the extracted rationale X_i^R is calculated by the sum of token distance $D_{X_i}(j)$ for each token $X_{ij} \in X_i^R$ as follows.

$$\delta_{X_i} = \sum_{X_{ij} \in X_i^R} D_{X_i}(j). \tag{8}$$

A smaller δ_{X_i} indicates higher continuity in the extracted rationale X_i^R, and vice versa.

3.2.2 Rationale Editor Operations

To minimize the average token distance within a rationale X_i^R while maintaining a satisfactory overall token score, our rationale editor iteratively employs four editing operations on the extracted rationale X_i^R from the first phase: token deletion (**D**), token insertion (**I**), token replacement (**R**), and no operation (**N**). Token deletion removes the token with the highest distance , using the lowest token score as a tie-breaker based on its anticipated impact on the rationale. Token insertion adds the highest-scoring token that can reduce the overall token distance. Token replacement balances continuity and overall token scores by performing token deletion and token insertion simultaneously. No operation indicates that the current rationale requires no further improvement.

To select the optimal operation at each editing iteration, we evaluate the *benefit* of each operation (o) based on token score and overall token distance, which represent token interactions and continuity, respectively. This allows the Rationale Editor to choose the operation that maximizes the benefit at each iteration. Given an operation o, the benefit estimator $B_{X_i}(o)$ for the resulting rationale $X_i^R(o)$ is formally defined as follows.

$$B_{X_i}(o) = \frac{\sum_{X_{ij} \in X_i^R(o)} s_{X_i}(j)}{\sum_{X_{ij} \in X_i^R(o)} D_{X_i}(j)}, \tag{9}$$

where the denominator represents the total token distance, and the numerator denotes the overall token scores from the resulting rationale $X_i^R(o)$. The Rationale Editor performs multiple iterations based on available computational resources.

Example: Table 2 presents an editing iteration for a movie review snippet with negative sentiment, highlighting the selected rationales in bold. The operation with the highest estimated benefit, $B_i(o)$, is the replacement (R) operation, which is chosen for execution.

3.3 End-task Pseudo Prediction

To generate high-quality rationales, our joint learning objective incorporates sufficiency ($\mathcal{L}_S$), informativeness ($\mathcal{L}_I$), and readability ($\mathcal{L}_R$) as defined in Eq.(4). At each training step, all non-rationale tokens in the input are replaced with the ¡MASK¿ token. The classifier is then evaluated exclusively based on the extracted rationale, ensuring that its predictions are conditioned solely on the highlighted evidence. This enables an end-to-end training paradigm in which the model iteratively refines both its rationale extraction and classification decisions, allowing the predictive outputs to dynamically adapt as the rationale evolves. These facets ensure that the rationale contributes to accurate predictions, conveys information, and preserves logical coherence, respectively. Formally, each facet is defined as follows.

3.3.1 Sufficiency

The concept of sufficiency is employed to assess whether the extracted rationales $\mathcal{X}^R = \{X_i^R | 1 \leq i \leq |\mathcal{X}|\}$ effectively predicts the end-task label through $\mathcal{F}_C(\cdot)$. Our goal is to align the probability distribution obtained from the learned rationale with the probability distribution from the original articles. The similarity between the two distributions is measured using cross-entropy. Formally, the sufficiency loss $\mathcal{L}_S$ is defined as follows.

$$\mathcal{L}_S = H(P(\hat{\mathcal{Y}} \mid \mathcal{X}^R), \mathcal{Y}) \tag{10}$$

where $H(\cdot)$ denotes the cross entropy function, $P(\hat{\mathcal{Y}} \mid \mathcal{X}^R)$ denotes the label prediction distribution made by the extracted rationales $\mathcal{X}^R$, and $\mathcal{Y}$ denotes the ground-truth label distribution.

3.3.2 Informativeness

Informativeness of a rationale refers to the disparity in information between the predominantly informative tokens in the extracted rationale and the remaining non-rationale tokens for the target label. This requires a clear distinction between the predictions made by the initial phase: informative tokens (rationale) and non-informative tokens (noise).

Let the tokens that have little indication of the correct label be considered as *noise*, denoted by $\mathcal{X}^N = \{X_i^N | 1 \leq i \leq |\mathcal{X}|\}$. We argue that the end-task predictions made solely based on the noise should exhibit a substantial deviation from the ground-truth label. Hence, the deviation can be formally defined as follows.

$$\mathcal{L}_N = -H(P(\hat{\mathcal{Y}} \mid \mathcal{X}^N), \mathcal{Y}) \tag{11}$$

where $P(\hat{\mathcal{Y}} \mid \mathcal{X}^N)$ represents the label prediction distribution made by the disentangled noise $\mathcal{X}^N$. To distinguish the label prediction distributions between the rationale and noise, we define the max-margin loss as follows.

$$\mathcal{L}_I = \max(\mathcal{L}_N - \mathcal{L}_S - h, 0) \tag{12}$$

where $\mathcal{L}_I$ enforces the disparity between $\mathcal{L}_N$ and $\mathcal{L}_S$ exceeds a specified margin h.

3.3.3 Readability

To ensure logical fluency for readability, we evaluate the impact of permuted rationales $\mathcal{X}^S = \{X_i^S | 1 \leq i \leq |\mathcal{X}|\}$, which are shuffled versions of X^R. Shuffling the rationales involves randomly permuting the tokens within the extracted rationale while keeping all non-rationale tokens masked (i.e., replaced with the $\langle MASK \rangle$ token). This perturbation is designed to assess the model's sensitivity to token ordering within the rationale. Since a shuffled rationale disrupts the original semantic and syntactic structure, it may impair comprehension and lead to degraded model performance, thereby providing insights into the extent to which the model relies on coherence and logical flow in its reasoning process. The readability loss $\mathcal{L}_R$ is formulated to measure the discrepancy between the predicted distribution of the disordered rationale $\mathcal{X}^S$ and the distribution of the ground-truth labels.

$$\mathcal{L}_R = -H(P(\hat{\mathcal{Y}} \mid \mathcal{X}^S), \mathcal{Y}) \tag{13}$$

where $P(\hat{\mathcal{Y}} \mid \mathcal{X}^S)$ represents the label prediction distribution made by the permuted rationale $\mathcal{X}^S$. A higher sensitivity to the ordering of rationale tokens indicates that the model's predictive performance is strongly influenced by the coherence and structured reasoning within the rationale. This suggests that the model effectively leverages logical consistency and sequential dependencies to derive accurate predictions.

4 EXPERIMENTS

We conduct experiments to address four research questions and validate the efficacy of SHARE.

4.1 Experimental Settings

Datasets. We evaluate our approach on three real-world datasets with ground-truth rationales: *VitaminC* (Schuster et al., 2021) [2] for claim verification, *Movie Review* for sentiment analysis, and *MultiRC* (DeYoung et al., 2020) [3] for question answering.

Baselines. To evaluate SHARE, we compare it against several baselines. AllenNLP (Shah et al., 2020) conceals important tokens to extract sufficient arguments for accurate predictions. VCU (Schuster et al., 2021), based on ALBERT (Schuster et al., 2021), focuses on the sufficiency of extracted rationales in an unsupervised manner. C-Shapley (Chen et al., 2019) and L-Shapley (Chen et al., 2019) incorporate the Shapley principle into rationale learning, with C-Shapley accounting for the continuity of neighboring tokens and L-Shapley not considering continuity. For comparison in generative rationale learning models, we utilize LLaMA2-7B[4] (Touvron et al., 2023) as the baseline model.

Metrics. We extensively conduct performance evaluation on several aspects, including sufficiency, informativeness and readability. (i) *sufficiency* is assessed by utilizing the classifier to assess whether the extracted rationales effectively predict the end-task label by measuring prediction accuracy (**ACC**) based on the extracted rationale (Herman, 2017; Jacovi and Goldberg, 2020; Rudin, 2019). (ii) To gauge *informativeness*, we compare the extracted rationale to the ground-truth rationale using token-level F1 (**TF1**) (Bodria et al., 2020; Fürnkranz et al., 2020). If the extracted rationale accurately explains the prediction outcome, it should closely resemble the ground-truth rationale. (iii) To measure *continuity*, we calculate the average minimal token distance (**TD**) between rationales. Additionally, we assess *logical fluency* using a logical coherence score (**LF**) to evaluate readability (Chen et al., 2020a).

$$LF = \frac{\sum_{i=1}^{|\mathcal{X}|} \sum_{q=1}^{Q} \left(P(\hat{\mathcal{Y}}_i \mid \mathcal{X}_i^R) - P(\hat{\mathcal{Y}}_i \mid \mathcal{X}_i^S) \right)}{|\mathcal{X}| \cdot Q} \tag{14}$$

where Q represents the number of permutations.

[2] https://github.com/TalSchuster/VitaminC

[3] https://www.eraserbenchmark.com/

[4] https://huggingface.co/meta-llama/Llama-2-7b-hf

Table 3: End-task performance comparison. Best and second-best results are highlighted in bold and underlined. The number in brackets represents the performance change compared to the ground-truth rationale. S-A, S-G, and S-T represent SHARE with ALBERT, GPT-2, and T5, respectively. A dash ("-") indicates that LLaMA2 fails to generate a rationale, resulting in an empty output.

Model	VitaminC	Movie Review	MultiRC
C-Shapley	0.566	0.236	0.595
L-Shapley	0.559	0.236	0.595
AllenNLP	**0.587**	0.429	0.540
VCU	0.569	<u>0.769</u>	0.597
LLaMA2	-	0.523	0.478
S-A	0.571 (↑ 1.4%)	**0.894** (↓ 8.4%)	<u>0.626</u> (↓ 2.8%)
S-G	0.576 (↓ 0.2%)	0.524 (↑ 13.9%)	0.597 (↑ 1.7%)
S-T	<u>0.585</u> (↓ 1.3%)	0.631 (↓ 3.1%)	**0.791** (↑ 9.8%)

Table 4: Comparison of rationale quality. A dash ("-") indicates that LLaMA2 fails to generate a rationale, resulting in an empty output.

Model	VitaminC		Movie Review		MultiRC	
	TF1 (↑)	TD (↓)	TF1 (↑)	TD (↓)	TF1 (↑)	TD (↓)
C-Shapley	0.720	0.455	0.539	0.517	0.657	0.354
L-Shapley	0.721	0.395	0.539	0.518	0.657	0.346
AllenNLP	0.305	0.478	0.387	0.608	0.275	0.740
VCU	0.739	0.412	0.467	0.668	<u>0.800</u>	<u>0.249</u>
LLaMA2	-	-	<u>0.558</u>	**0.052**	**0.805**	**0.019**
S-A	0.677	0.343	0.530	0.558	0.708	0.426
S-G	**0.758**	**0.131**	0.516	0.627	0.630	0.496
S-T	<u>0.750</u>	<u>0.222</u>	**0.668**	<u>0.451</u>	0.703	0.410

Implementation Details. We fine-tune and pre-train the classifiers (F_C) using widely adopted language models (ALBERT-base, GPT-2, and T5-small) across three datasets. The rationale extractor (F_R) utilizes a pre-trained model. We empirically determine rationale threshold (γ) based on the desired length for specific user applications. Further hyperparameters details are provided in the Appendix.

4.2 End-task Performance (RQ1)

Setup. We assess the sufficiency of the extracted rationale in the end-task using classification accuracy (ACC). Our model $\mathcal{F}_R$ extracts a rationale for each $X_i \in \mathcal{X}$, which is then used as input for the dedicated end-task classifier $\mathcal{F}_C$ to predict $\hat{Y}_i \in \mathcal{Y}$.

Results. In Table 3, SHARE-generated rationales demonstrate superior sufficiency for end tasks involving lengthy documents, such as Movie Review and MultiRC, while achieving results comparable to the baselines for short documents like VitaminC. For example, in the sentiment analysis task (Movie Review), SHARE-generated rationales coupled with ALBERT significantly outperform various baselines in terms of accuracy. Moreover, the accuracy of predictions based on SHARE-generated rationales closely aligns with, and in some cases surpasses, that of human-annotated ground-truth rationales (GT). In the question-answering task (MultiRC), SHARE combined with T5 achieves a 9.8% improvement, while our GPT-2-based model improves by 1.7% over the ground truth. For the short document task, our ALBERT-based model achieves a 1.4% improvement over the ground truth. Compared to the rationales generated by LLaMA2, those selected by SHARE demonstrate superior performance. Specifically, on the Movie Review dataset, SHARE combined with ALBERT achieves a 70.9% relative improvement over LLaMA2, while on the MultiRC dataset, SHARE with T5 yields a 65.5% relative improvement over LLaMA2. These results highlight the effectiveness of SHARE in selecting higher-quality rationales that enhance model performance across diverse datasets.

Table 5: Comparison of rationale readability in classification accuracy (ACC), and logical coherence score (LF). A dash ("-") indicates that LLaMA2 fails to generate a rationale, resulting in an empty output.

Model	VitaminC		Movie Review		MultiRC	
	ACC (↑)	LF (↑)	ACC (↑)	LF (↑)	ACC (↑)	LF (↑)
LLaMA2	-	-	0.498	-0.022	0.405	-0.149
VCU	0.569	-0.012	0.769	0.118	0.597	-0.006
S-A	0.571	-9.02e-5	0.894	0.075	0.626	0.009
GT	0.563	-1.28e-4	0.976	0.038	0.644	0.014

4.3 Rationale Quality (RQ2)

Setup. We evaluate the informativeness and readability of the generated rationale using three metrics: token-level F1 (TF1), token distance (TD), and logical coherence score (LF). A higher TF1 indicates closer alignment with the ground-truth rationale, while a lower TD reflects improved continuity. Additionally, a higher LF signifies a more meaningful token ordering within the rationale.

Results. Table 4 shows that SHARE enhances the informativeness and continuity of rationales, aligning more closely with ground truth than baseline models. Specifically, S-G and S-T, which integrate SHARE with GPT-2 and T5, respectively, achieve higher informativeness (TF1) in the VitaminC and Movie Review datasets and demonstrate significantly improved continuity, as indicated by a substantial reduction in token distance (TD). While VCU generates more aligned and continuous rationales in the MultiRC dataset, its end-task prediction accuracy (ACC) of 0.597 is significantly lower than that of SHARE-compatible models like S-T, which achieves an ACC of 0.791. Overall, SHARE consistently presents compatibility with various language models, resulting in higher informativeness (↑ TF1) and improved continuity (↓ TD). Compared to other models, the rationales generated by LLaMA2 exhibit greater continuity and higher alignment with the ground-truth rationales. However, further analysis suggests that this alignment is primarily an artifact of the lower token-level similarity be-

Table 6: Rationale Case Study for Claim Verification Task. The ground-truth rationale (GT) and true label are provided in bold, with the extracted rationale highlighted in yellow. Incorrect predictions (PRE) are marked in red.

"*Claim: More than 1,101 767s had been delivered to Boeing 's customers by July 2017.*"

Model	Highlighted Rationale	PRE	Truth
GT	As of **June** 2017 , Boeing has received 1,204 orders for the 767 from 74 customers ; **1,101** have been delivered .	**REF**	**REF**
L-Shapley	**As of June** 2017 , Boeing has received 1,204 orders for the 767 from 74 customers ; 1,101 have been **delivered** .	SUP	REF
AllenNLP	**As** of **June 2017** , Boeing has **received 1,204 orders** for the 767 from 74 customers ; 1,101 have been **delivered** .	SUP	REF
SHARE	As of **June 2017** , Boeing has received 1,204 orders for the 767 from 74 customers ; **1,101** have been delivered .	REF	REF

"*Claim: Baadshah was only dubbed into French.*"

Model	Highlighted Rationale	PRE	Truth
GT	Later , this **movie** was dubbed into **Malayalam** , **Japanese** , and **Korean** under the same name .	**REF**	**REF**
L-Shapley	Later , this movie was dubbed into **Malayalam , Japanese** , and Korean **under** the same name .	REF	REF
AllenNLP	**Later ,** this movie **was dubbed** into **Malayalam , Japanese** , and **Korean** under the same name .	REF	REF
SHARE	Later , this movie was dubbed into Malayalam , Japanese , and Korean under **the same** name .	REF	REF

Table 7: Ablation Study.

Model	Movie Review			
	ACC (↑)	TF1 (↑)	TD (↓)	LF (↑)
SHARE w/o Shapley	0.909	0.249	0.994	0.081
SHARE w/o Indep. Relevance	0.742	0.633	0.499	0.021
SHARE w/o Rationale Editor	0.878	0.522	0.649	0.070
SHARE w/o Sufficiency	0.834	0.524	0.610	0.063
SHARE w/o Informativeness	0.749	0.605	0.530	0.025
SHARE w/o Readability	0.874	0.523	0.613	0.063
SHARE (Full)	0.894	0.577	0.605	0.075

tween LLaMA2-generated rationales and the ground-truth. For instance, in the Movie Review dataset, the average token overlap (computed as the ratio of matched tokens to the length of the ground-truth rationale) is only 0.038, while in the MultiRC dataset, it is 0.020. In contrast, SHARE with ALBERT achieves superior performance, demonstrating a stronger ability to extract high-quality rationales while maintaining meaningful alignment with the ground-truth. We further assess the logical coherence (LF) alongside ACC, as relying solely on LF may not accurately reflect overall quality. For example, a perfectly continuous and readable token sequence may still fail to predict outcomes accurately. In Table 5, SHARE with ALBERT achieves a higher LF than VCU in the VitaminC and MultiRC datasets and surpasses the ground truth (GT) in VitaminC and Movie Review datasets, indicating more meaningful ordering of the extracted rationales. However, in the Movie Review dataset, SHARE has a lower LF than VCU, as VCU selects over 50% of tokens, which may lead to confusing interpretations after permutations.

4.4 Ablation Study (RQ3)

Due to space constraints, we analyze SHARE's key components on the Movie Review dataset using the ALBERT model in Table 7.(i) Rationale Extractor: We first assess the impact of Shapley and independent token relevance scores on the rationale extractor's performance metrics. The variant without the Shapley score excels in sufficiency and logical fluency but suffers from reduced informativeness and continuity, resulting in the undesirable selection of nearly all tokens in an article as the rationale. Conversely, the variant without the independent token relevance score shows improvements in F1 and reduced token distance compared to the full model. However, the full model, combining independent token relevance and Shapley scores, achieves superior results in sufficiency (ACC), informativeness (TF1), and readability (TD, LF). (ii) Rationale Editor: The rationale editor enhances continuity compared to the model without it, as evidenced by a shorter token distance for rationales extracted by SHARE than for those extracted without the editor. (iii) Quality Aspects: When sufficiency is preserved, both accuracy and F1 decrease. If informativeness is preserved, accuracy declines significantly. Similarly, maintaining readability results in deterioration across all metrics.

4.5 Case Study and Analysis (RQ4)

We conduct case studies to compare the extracted rationales against ground-truth rationales (GT). An example from the claim verification task is presented in Table 6. The results indicate that SHARE with ALBERT produces rationales closely aligned with the ground truth, outperforming the baselines. In contrast, AllenNLP extracts misleading information, while L-Shapley overlooks key details. Moreover, SHARE facilitates accurate end-task predictions, whereas both L-Shapley and AllenNLP yield incorrect predictions. Additionally, SHARE's rationales consistently exhibit greater continuity across all scenarios.

5 LITERATURE REVIEW

Extractive Rationale Learning. Some rationale learning models are self-explaining, providing explanations alongside predictions(Yue et al., 2022; Liu et al., 2023b), while others are post-hoc, generating explanations after predictions (Danilevsky et al., 2020; Linardatos et al., 2020). Post-hoc models, which operate independently of the pre-

diction model, may produce less reliable interpretations (Ribeiro et al., 2016; Zafar and Khan, 2021; Pryzant et al., 2018). Models such as fact-guided AllenNLP (Shah et al., 2020) and VitaminC (Schuster et al., 2021) function as self-explaining models; however, their generalizability is limited by the lack of human-provided explanations in unsupervised settings. Therefore, it is essential to develop an unsupervised, self-explaining model for extracting rationales when ground-truth rationales are unavailable. Moreover, existing models often prioritize rationale sufficiency while neglecting aspects like information discrepancy and readability (Yu et al., 2019; Chen et al., 2020b; Carton et al., 2022; Liu et al., 2022). This highlights the need to address multiple facets of rationale quality in an unsupervised manner.

Generative Rationale Learning. Generative (Chan et al., 2022a; Wang et al., 2022; Chen et al., 2022, 2023) and extractive rationale learning differ fundamentally and require distinct techniques to address their respective challenges. First, free-text rationales heavily depend on external pre-trained knowledge, rendering comparisons between the two approaches unfair due to differing assumptions about prior knowledge. Second, generative rationale learning requires a unique evaluation framework to address its specific challenges. While free-text rationales are typically coherent and readable, they often lack faithfulness (Wiegreffe et al., 2022) as external knowledge may lead reasoning astray from the context, even when the answers appear correct. Thus, our framework emphasizes extractive rationales, selecting tokens from the provided context to explain predictions. This approach optimizes informativeness and readability, enhancing rationale quality and mitigating the hallucination issues prevalent in generative rationale learning.

Attention Mechanism. Attention mechanisms are widely used for rationale extraction (Cho et al., 2014; Barbieri et al., 2018; Vaswani et al., 2017). While some argue that random attention weights may not significantly impact predictions (Jain and Wallace, 2019), others show that attention weights can provide robust explanations in certain contexts (Wiegreffe and Pinter, 2019). However, attention alone may not offer comprehensive explanations (Wiegreffe and Pinter, 2019). Our method extends beyond existing mechanism by considering the marginal contributions of tokens for a more holistic explanation.

6 CONCLUSION

We present SHARE, a novel editable framework for self-explaining rationale extraction that effectively explains predictions by considering the relevance of individual tokens and the marginal contribution of neighboring tokens. The rationale editor within SHARE is a flexible, plug-and-play component adaptable to various tasks. Extensive experiments demonstrate its effectiveness in balancing sufficiency, informativeness, and readability across diverse applications.

References

Barbieri, F., Espinosa-Anke, L., Camacho-Collados, J., Schockaert, S., and Saggion, H. (2018). Interpretable emoji prediction via label-wise attention LSTMs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4766–4771.

Bodria, F., Panisson, A., Perotti, A., and Piaggesi, S. (2020). Explainability Methods for Natural Language Processing: Applications to Sentiment Analysis. In *Sistemi Evoluti per Basi di Dati*.

Carton, S., Kanoria, S., and Tan, C. (2022). What to Learn, and How: Toward Effective Learning from Rationales. In *Findings of the Association for Computational Linguistics*, pages 1075–1088.

Chan, A., Nie, S., Tan, L., Peng, X., Firooz, H., Sanjabi, M., and Ren, X. (2022a). Frame: Evaluating Rationale-Label Consistency Metrics for Free-Text Rationales. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Chan, A., Sanjabi, M., Mathias, L., Tan, L., Nie, S., Peng, X., Ren, X., and Firooz, H. (2022b). UNIREX: A Unified Learning Framework for Language Model Rationale Extraction. In *Proceedings of the 39th International Conference on Machine Learning*, pages 2867–2889.

Chen, H., Brahman, F., Ren, X., Ji, Y., Choi, Y., and Swayamdipta, S. (2022). Information-Theoretic Evaluation of Free-Text Rationales with Conditional V-Information. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

Chen, H., Brahman, F., Ren, X., Ji, Y., Choi, Y., and Swayamdipta, S. (2023). REV: Information-theoretic evaluation of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Chen, H., Zheng, G., and Ji, Y. (2020a). Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection. In *Annual Meeting of the Association for Computational Linguistics*.

Chen, H., Zheng, G., and Ji, Y. (2020b). Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5578–5593.

Chen, J., Song, L., Wainwright, M., and Jordan, M. (2018). Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 883–892.

Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. (2019). L-shapley and C-shapley: Efficient Model Interpretation for Structured Data. In *International Conference on Learning Representations*.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., and Sen, P. (2020). 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*, pages 447–459.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., and Wallace, B. C. (2020). ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458.

Fürnkranz, J., Kliegr, T., and Paulheim, H. (2020). On Cognitive Preferences and the Plausibility of Rule-Based Models. *Machine Learning*, pages 853–898.

Gumbel, E. J. (1954). Statistical Theory of Extreme Values and Some Practical Applications: A Series of Lectures. *The Journal of the Royal Aeronautical Society*.

Herman, B. (2017). The Promise and Peril of Human Evaluation for Model Interpretability. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Jacovi, A. and Goldberg, Y. (2020). 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*, pages 4198–4205.

Jain, S. and Wallace, B. C. (2019). Attention is not explanation. In *North American Chapter of the Association for Computational Linguistics*.

Jang, E., Gu, S., and Poole, B. (2017). Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*.

Lauriola, I., Lavelli, A., and Aiolli, F. (2022). An Introduction to Deep Learning in Natural Language Processing: Models, Techniques, and Tools. *Neurocomputing*, pages 443–456.

Lei, T., Barzilay, R., and Jaakkola, T. (2016). Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*.

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

Liu, W., Wang, H., Wang, J., Li, R., Yue, C., and Zhang, Y. (2022). FR: Folded rationalization with a unified encoder. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Liu, W., Wang, J., Wang, H., Li, R., Deng, Z., Zhang, Y., and Qiu, Y. (2023b). D-separation for causal self-explanation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Paranjape, B., Joshi, M., Thickstun, J., Hajishirzi, H., and Zettlemoyer, L. (2020). An Information Bottleneck Approach for Controlling Conciseness in Rationale Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1938–1952.

Pryzant, R., Basu, S., and Sone, K. (2018). Interpretable Neural Architectures for Attributing an Ad's Performance to its Writing Style. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 125–135.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "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*, pages 1135–1144.

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., and Sarkar, R. (2022). The Shapley Value in Machine Learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 5572–5579.

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, pages 206–215.

Schuster, T., Fisch, A., and Barzilay, R. (2021). Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643.

Shah, D. J., Schuster, T., and Barzilay, R. (2020). Automatic Fact-guided Sentence Modification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8791–8798.

Shapley, L. S. (1953). A Value for N-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–317.

Touvron, H., Martin, L., Stone, K. R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D. M., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A. S., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I. M., Korenev, A. V., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M. H. M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.

Voskarides, N., Meij, E., Tsagkias, M., de Rijke, M., and Weerkamp, W. (2015). Learning to Explain Entity Relationships in Knowledge Graphs. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 564–574.

Wang, P., Chan, A., Ilievski, F., Chen, M., and Ren, X. (2022). PINTO: Faithful Language Reasoning Using Prompt-Generated Rationales. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

Wiegreffe, S., Hessel, J., Swayamdipta, S., Riedl, M., and Choi, Y. (2022). Reframing human-ai collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. 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)*, pages 11–20.

Yu, M., Chang, S., Zhang, Y., and Jaakkola, T. S. (2019). Rethinking Cooperative Rationalization: Introspective Extraction and Complement Control. In *Empirical Methods in Natural Language Processing*.

Yue, L., Liu, Q., Du, Y., An, Y., Wang, L., and Chen, E. (2022). DARE: Disentanglement-augmented rationale extraction. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Zafar, M. R. and Khan, N. (2021). Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability. *Machine Learning and Knowledge Extraction*, pages 525–541.

Zini, J. E. and Awad, M. (2022). On the Explainability of Natural Language Processing Deep Models. *ACM Computing Surveys*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Not Applicable]
   (b) Complete proofs of all theoretical results. [Not Applicable]
   (c) Clear explanations of any assumptions. [Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No, all source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper.]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [No, all source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper.]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

[twoside]article

aistats2025

times soul url [hidelinks]hyperref [utf8]inputenc [small]caption graphicx amsmath amsthm booktabs algorithm algorithmic xcolor subfigure multirow longtable microtype amsmath bbding

A Shapley-value Guided Rationale Editor for Rationale Learning: Supplementary Materials

1 ADDITIONAL EXPERIMENTAL DETAILS

In addition to the state-of-the-art extractive baselines, we also present further experimental results comparing the quality of SHARE with LLM-generated rationales (LLaMA2) across four key research questions.

1.1 Implementation Details

We fine-tune and pre-trained the classifiers (F_C) using widely adopted language models such as ALBERT-base-v2, GPT-2, and T5-small architectures in HuggingFace across the three datasets. F_R utilizes a pre-trained model from HuggingFace. We intentionally keep F_R and F_C independent for two reasons. Firstly, joint learning may introduce interference between them. To maintain the integrity of predictions made by F_C and focus solely on extracting rationales for explanation, F_R is designed to sequentially select the most informative tokens contributing to F_C's prediction outcome. Secondly, the independent nature of F_R allows it to concentrate solely on extracting rationales, making it adaptable to various tasks and language models for broader applications. The rationale model, F_R, for LLaMA2 corresponds to LLaMA2-7B, while the classifier model, F_c, for LLaMA2 is the pre-trained classifier used in SHARE. Our model is implemented in PyTorch on a 32GB NVIDIA Tesla V100 GPU. LLaMA2 is implemented in PyTorch on a 24GB RTX 3090 GPU. The key hyperparameters are listed in Table 8. We observe that the rationale threshold (γ) directly influences the length of the rationale. Consequently, we empirically determine the threshold (γ) based on the desired rationale length within the users' application of interest. We provide an example for rigorous coefficient search on Movie Review dataset using T5 model in Table 9.

Table 8: Hyperparameters Settings.

Parameters	VitaminC	Movie Review	MultiRC
#Permutations (Q)	50	50	50
Shapley Neighbors (k)	2	2	2
Rationale Editor Times	2	3	3
Comprehensiveness (h)	0.03	0.05	0.03
Rationale Threshold γ	0.9	0.85	0.9
Sufficiency (λ_S)	1.0	1.0	1.0
Informativeness (λ_I)	adaptive	adaptive	adaptive
Readability (λ_R)	0.2	0.2	0.2

Table 9: Coefficient Search for T5 on Movie Review Dataset.

λ_S	λ_I	λ_R	ACC (↑)	TF1 (↑)	TD (↓)	LF (↑)
0.5	0.2	0.2	0.815	0.672	0.543	0.045
0.5	0.2	0.5	0.816	0.659	0.551	0.046
0.5	0.5	0.2	0.816	0.657	0.551	0.044
0.5	0.5	0.5	0.667	0.931	0.001	0.001
1	0.2	0.2	0.824	0.658	0.551	0.048
1	0.2	0.5	0.814	0.658	0.551	0.049
1	0.5	0.2	0.809	0.658	0.552	0.046
1	0.5	0.5	0.674	0.931	0.001	0.002

1.2 Additional Quantitative Results on FEVER (RQ2)

We evaluate the informativeness and readability of the generated rationale using three metrics: token-level F1 (TF1), token distance (TD), and logical coherence score (LF) on FEVER (DeYoung et al., 2020) [5].

Table 10 shows that SHARE improves the informativeness and continuity of rationales, aligning more closely with the ground truth compared to baseline models. Specifically, S-A, which integrates SHARE with ALBERT, achieves higher informativeness (TF1) in the FEVER dataset. However, it exhibits slightly lower continuity, as reflected by an increase in token distance (TD).

Table 10: Quantitative Results on Fever Dataset.

Model	ACC	TF1	TD
C-Shapley	0.565	0.756	0.308
L-Shapley	0.63	0.75	0.33
VCU	0.01	0.896	0.01
S-A	0.431	0.826	0.418

Table 11: Ablation Study. ACC, TF1, TD, and LF represent the end-task prediction accuracy, token-level F1, token distance, and logical coherence score, respectively.

Model	MultiRC			
	ACC (↑)	TF1 (↑)	TD (↓)	LF (↑)
SHARE w/o Shapely	0.601	0.530	0.590	0.002
SHARE w/o Significance	0.638	0.578	0.558	0.009
SHARE w/o Rationale Editor	0.621	0.712	0.549	0.008
SHARE w/o Sufficiency	0.613	0.585	0.546	0.009
SHARE w/o Informativeness	0.615	0.674	0.463	0.000
SHARE w/o Readability	0.623	0.672	0.463	0.002
SHARE (Full)	0.626	0.708	0.426	0.009

1.3 Additional Ablation Study on MultiRC (RQ3)

Table 11 presents the additional ablation study on the MultiRC dataset.

Rationale Generator: SHARE without Shapley value and without significance score both lead to a decrease of informativeness (TF1) and continuity (TD). Additionally, when considering prediction accuracy and token-level F1 together, the full model outperforms the model without significance score, with a slight decrease in accuracy but a significant increase in token-level F1.

Rationale Editor: Including a rationale editor positively impacts continuity compared to the model without a rationale editor. Specifically, in MultiRC, the token distance of the extracted rationale by SHARE is much lower than that of SHARE without a rationale editor.

Quality Aspects: Without one of the loss components, i.e., sufficiency, informativeness, or readability, all metrics deteriorate.

1.4 Additional Qualitative Results (RQ4)

In this section, we present additional qualitative examples in Tables 12 and 13, focusing on the Movie Review and MultiRC datasets, respectively, to evaluate the quality of SHARE against SOTA baselines and LLM-generated rationales. As shown in the first case on Movie Review dataset in Table 12, the rationale generated by LLaMA2 bears little resemblance to the original review, providing only a generic explanation for a positive movie review. Similarly, we also provide qualitative cases in Table 13. the LLaMA2-generated rationale diverges from the document. For example, while the document states "all five friends showed up," LLaMA2 incorrectly generates "the fifth friend didn't show up." From the examples in both

[5] https://www.eraserbenchmark.com/

tables, it is evident that although LLaMA2-generated rationales appear more continuous, they often introduce information that is not present in the source, leading to hallucinations.

Table 12: Case Study on Movie Review dataset. We present additional examples comparing the quality of SHARE against SOTA extractive baselines and LLM-generated rationales.

Human Rationale	Baseline VitaminC	Baselie C-Shapley	SHARE	LLaMA2
[CLS] Trees lounge is the directoral debut from one of my favorite actors , steve buscemi . He gave memorable performences in in the soup , fargo , and reservoir dogs . Now he tries his hand at writing , directing and acting all in the same flick . The movie starts out awfully slow with tommy (buscemi) hanging around a local bar the " trees lounge " and him pestering his brother . It 's obvious he a loser . But as he says " it 's better i 'm a loser and know i am , then being a loser and not thinking i am . " well put . The story starts to take off when his uncle dies , and tommy , not having a job , decides to drive an ice cream truck . Well , the movie starts to pick up with him finding a love interest in a 17 year old girl named debbie (chloe sevigny) and . . . I liked this movie alot even though it did not reach my expectation . After you 've seen him in fargo and reservoir dogs , you know he is capable of a better performence . I think his brother , michael , did an excellent job for his debut performence . Mr . buscemi is off to a good career as a director ! [SEP]	[CLS] Trees lounge is the directoral debut from one of my favorite actors , steve buscemi . He gave memorable perform ences in in the soup , fargo , and reservoir dogs . Now he tries his hand at writing , directing and acting all in the same flick . The movie starts out awfully slow with tommy (buscemi) hanging around a local bar the " trees lounge " and him pestering his brother . It' s obvious he a loser . But as he says " it 's better i 'm a loser and know i am , then being a loser and not thinking i am . " well put . The story starts to take off when his uncle dies , and tommy , not having a job , decides to drive an ice cream truck . Well , the movie starts to pick up with him finding a love interest in a 17 year old girl named debbie (chloe sevigny) and . . . I liked this movie alot even though it did not reach my expectation . After you 've seen him in fargo and reservoir dogs , you know he is capable of a better performence . I think his brother , michael , did an excellent job for his debut performence . Mr . buscemi is off to a good career as a director ! [SEP]	[CLS] Trees lounge is the directoral debut from one of my favorite actors , steve buscemi . He gave memorable performences in in the soup , fargo , and reservoir dogs . Now he tries his hand at writing , directing and acting all in the same flick . The movie starts out awfully slow with tommy (buscemi) hanging around a local bar the " trees lounge " and him pestering his brother . It 's obvious he a loser . But as he says " it 's better i 'm a loser and know i am , then being a loser and not thinking i am . " well put . The story starts to take off when his uncle dies , and tommy , not having a job , decides to drive an ice cream truck . Well , the movie starts to pick up with him finding a love interest in a 17 year old girl named debbie (chloe sevigny) and . . . I liked this movie alot even though it did not reach my expectation . After you 've seen him in fargo and reservoir dogs , you know he is capable of a better performence . I think his brother , Michael , did an excellent job for his debut performence . Mr . buscemi is off to a good career as a director ! [SEP]	[CLS] Trees lounge is the directoral debut from one of my favorite actors , steve buscemi . He gave memorable performances in in the soup , fargo , and reservoir dogs . Now he tries his hand at writing , directing and acting all in the same flick . The movie starts out awfully slow with tommy (buscemi) hanging around a local bar the " trees lounge " and him pestering his brother . It 's obvious he a loser . But as he says " it 's better i 'm a loser and know i am , then being a loser and not thinking i am . " well put . The story starts to take off when his uncle dies , and tommy , not having a job , decides to drive an ice cream truck . Well , the movie starts to pick up with him finding a love interest in a 17 year old girl named debbie (chloe sevigny) and . . . I liked this movie alot even though it did not reach my expectation . After you 've seen him in fargo and reservoir dogs , you know he is capable of a better performence . I think his brother , michael , did an excellent job for his debut performence . Mr . buscemi is off to a good career as a director ! [SEP]	The sentiments is positive because the movie was very funny and entertaining. The characters were likable and the plot was interesting. The movie had a lot of heart and soul which made it enjoyable to watch. The movie also had some great performances by the cast members which added to the overall enjoyment of the film.Overall, I would recommend this movie to anyone who wants to have a good time at the movies.
Lable: POSITIVE	**Label: NEGATIVE**	**Label: NEGATIVE**	**Label: POSITIVE**	**Label: POSITIVE**

Table 13: Case Study on MultiRC dataset. We present additional examples comparing the quality of SHARE against SOTA extractive baselines and LLM-generated rationales.

Human Rationale	Baseline VitaminC	Baselie C-Shapley	SHARE	LLaMA2
[CLS] Susan wanted to have a birthday party . She called all of her friends . She has five friends . Her mom said that Susan can invite them all to the party . Her first friend could not go to the party because she was sick . Her second friend was going out of town . Her third friend was not so sure if her parents would let her . The fourth friend said maybe . The fifth friend could go to the party for sure . Susan was a little sad . On the day of the party , all five friends showed up . Each friend had a present for Susan . Susan was happy and sent each friend a thank you card the next week . [SEP] How many people did Susan call ? || All of her firends [SEP]	[CLS] Susan wanted to have a birthday party . She called all of her friends . She has five friends . Her mom said that Susan can invite them all to the party . Her first friend could not go to the party because she was sick . Her second friend was going out of town . Her third friend was not so sure if her parents would let her . The fourth friend said maybe . The fifth friend could go to the party for sure . Susan was a little sad . On the day of the party , all five friends showed up . Each friend had a present for Susan . Susan was happy and sent each friend a thank you card the next week . [SEP] How many people did Susan call ? || All of her firends [SEP]	[CLS] Susan wanted to have a birthday party . She called all of her friends . She has five friends . Her mom said that Susan can invite them all to the party . Her first friend could not go to the party because she was sick . Her second friend was going out of town . Her third friend was not so sure if her parents would let her . The fourth friend said maybe . The fifth friend could go to the party for sure . Susan was a little sad . On the day of the party , all five friends showed up . Each friend had a present for Susan . Susan was happy and sent each friend a thank you card the next week . [SEP] How many people did Susan call ? || All of her firends [SEP]	[CLS] Susan wanted to have a birthday party . She called all of her friends . She has five friends . Her mom said that Susan can invite them all to the party . Her first friend could not go to the party because she was sick . Her second friend was going out of town . Her third friend was not so sure if her parents would let her . The fourth friend said maybe . The fifth friend could go to the party for sure . Susan was a little sad . On the day of the party , all five friends showed up . Each friend had a present for Susan . Susan was happy and sent each friend a thank you card the next week . [SEP] How many people did Susan call ? || All of her firends [SEP]	The given answer is false because Susan only invited four friends to the party. The fifth friend could go to the party for sure but she didn't show up at the party.
Label: FALSE	**Label: FALSE**	**Label: FALSE**	**Label: FALSE**	**Label: TRUE**