
Proceedings Paper:
Zhao, Z. and Aletras, N. (2023) Incorporating attribution importance for improving faithfulness metrics. In: Rogers, A., Boyd-Graber, J. and Okazaki, N., (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 61st Annual Meeting of the Association for Computational Linguistics, 09-14 Jul 2023, Toronto, Canada. Association for Computational Linguistics , pp. 4732-4745. ISBN 9781959429722

https://doi.org/10.18653/v1/2023.acl-long.261

Incorporating Attribution Importance for Improving Faithfulness Metrics

Zhixue Zhao **Nikolaos Aletras**
Department of Computer Science, University of Sheffield
United Kingdom
{zhixue.zhao, n.aletras}@sheffield.ac.uk

Abstract

Feature attribution methods (FAs) are popular approaches for providing insights into the model reasoning process of making predictions. The more faithful a FA is, the more accurately it reflects which parts of the input are more important for the prediction. Widely used faithfulness metrics, such as sufficiency and comprehensiveness use a hard erasure criterion, i.e. entirely removing or retaining the top most important tokens ranked by a given FA and observing the changes in predictive likelihood. However, this hard criterion ignores the importance of each individual token, treating them all equally for computing sufficiency and comprehensiveness. In this paper, we propose a simple yet effective soft erasure criterion. Instead of entirely removing or retaining tokens from the input, we randomly mask parts of the token vector representations proportionately to their FA importance. Extensive experiments across various natural language processing tasks and different FAs show that our soft-sufficiency and soft-comprehensiveness metrics consistently prefer more faithful explanations compared to hard sufficiency and comprehensiveness.[1]

1 Introduction

Feature attribution methods (FAs) are popular post-hoc explanation methods that are applied after model training to assign an importance score to each token in the input (Kindermans et al., 2016; Sundararajan et al., 2017). These scores indicate how much each token contributes to the model prediction. Typically, the top-k ranked tokens are then selected to form an explanation, i.e. rationale (DeYoung et al., 2020). However, it is an important challenge to choose a FA for a natural language processing (NLP) task at hand (Chalkidis et al., 2021; Fomicheva et al., 2022) since there is no single FA that is consistently more faithful (Atanasova et al., 2020).

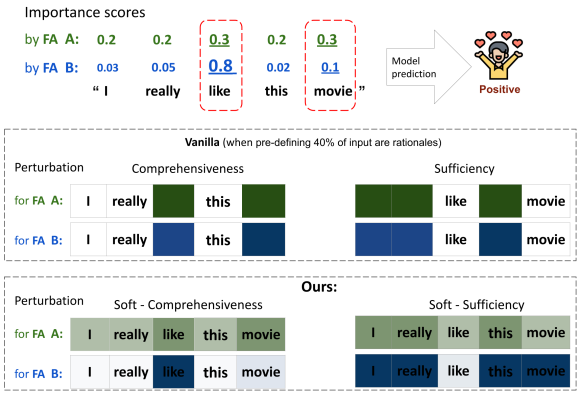

Figure 1: Hard and soft erasure criteria for comprehensiveness and sufficiency for two toy feature attribution (FA) methods A and B.

To assess whether a rationale extracted with a given FA is faithful, i.e. actually reflects the true model reasoning (Jacovi and Goldberg, 2020), various faithfulness metrics have been proposed (Arras et al., 2017; Serrano and Smith, 2019; Jain and Wallace, 2019; DeYoung et al., 2020). Sufficiency and comprehensiveness (DeYoung et al., 2020), also referred to as fidelity metrics (Carton et al., 2020), are two widely used metrics which have been found to be effective in capturing rationale faithfulness (Chrysostomou and Aletras, 2021a; Chan et al., 2022). Both metrics use a hard erasure criterion for perturbing the input by entirely removing (i.e. comprehensiveness) or retaining (i.e. sufficiency) the rationale to observe changes in predictive likelihood.

However, the hard erasure criterion ignores the different importance of each individual token, treating them all equally for computing sufficiency and comprehensiveness. Moreover, the hard-perturbed input is likely to fall out of the distribution the model is trained on, leading to inaccurate measurements of faithfulness (Bastings and Filippova,

[1]Our code: https://github.com/casszhao/SoftFaith

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4732–4745
July 9-14, 2023

2020; Yin et al., 2022; Chrysostomou and Aletras, 2022a; Zhao et al., 2022). Figure 1 shows an example of two toy FAs, A and B, identifying the same top two tokens ("like", "movie") as a rationale for the prediction. Still, each of them assigns different importance scores to the two tokens resulting into different rankings. According to the hard erasure criterion, comprehensiveness and sufficiency will assign the same faithfulness score to the two rationales extracted by the two FAs.

In this paper, we aim to improve sufficiency and comprehensiveness in capturing the faithfulness of a FA. We achieve this by replacing the hard token perturbation with a simple yet effective soft erasure criterion (see Figure 1 for an intuitive example). Instead of entirely removing or retaining tokens from the input, we randomly mask parts of token vector representations proportionately to their FA importance.

Our main contributions are as follows:

- We propose two new faithfulness metrics, soft-comprehensiveness and soft-sufficiency that rely on soft perturbations of the input. Our metrics are more robust to distribution shifts by avoiding entirely masking whole tokens;
- We demonstrate that our metrics are consistently more effective in terms of preferring more faithful rather than unfaithful (i.e. random) FAs (Chan et al., 2022), compared to their "hard" counterparts across various NLP tasks and different FAs.
- We advocate for evaluating the faithfulness of FAs by taking into account the entire input rather than manually pre-defining rationale lengths.

2 Related Work

2.1 Feature Attribution Methods

A popular approach to assign token importance is by computing the gradients of the predictions with respect to the input (Kindermans et al., 2016; Shrikumar et al., 2017; Sundararajan et al., 2017). A different approach is based on making perturbations in the input or individual neurons aiming to capture their impact on later neurons (Zeiler and Fergus, 2014). In NLP, attention mechanism scores have been extensively used for assigning token importance (Jain and Wallace, 2019; Serrano and Smith, 2019; Treviso and Martins, 2020; Chrysostomou and Aletras, 2021b). Finally, a widely used group of FA methods is based on training simpler linear meta-models to assign token importance (Ribeiro et al., 2016).

Given the large variety of approaches, it is often hard to choose an optimal FA for a given task. Previous work has demonstrated that different FAs generate inconsistent or conflicting explanations for the same model on the same input (Atanasova et al., 2020; Zhao et al., 2022).

2.2 Measuring Faithfulness

One standard approach to compare FAs and their rationales is faithfulness. A faithful model explanation is expected to accurately represent the true reasoning process of the model (Jacovi and Goldberg, 2020).

The majority of existing methods for quantitatively evaluating faithfulness is based on input perturbation (Nguyen, 2018; DeYoung et al., 2020; Ju et al., 2022). The main idea is to modify the input by entirely removing or retaining tokens according to their FA scores aiming to measure the difference in predictive likelihood .

Commonly-used perturbation methods include comprehensiveness, i.e. removing the rationale from the input), and sufficiency, i.e. retaining only the rationale (DeYoung et al., 2020). Another common approach is to remove a number of tokens and observe the number of times the predicted label changes, i.e. Decision Flip (Serrano and Smith, 2019). On the other hand, Monotonicity incrementally adds more important tokens while Correlation between Importance and Output Probability (CORR) continuously removes the most important tokens (Arya et al., 2021). (In)fidelity perturbs the input by dropping a number of tokens in a decreasing order of attribution scores until the prediction changes (Zafar et al., 2021). Additionally, Yin et al. (2022) proposed sensitivity and stability, which do not directly remove or keep tokens. Sensitivity adds noise to the entire rationale set aiming to find a minimum noise threshold for causing a prediction flip. Stability compares the predictions on semantically similar inputs.

One limitation of the metrics above is that they ignore the relative importance of each individual token within the selected rationale, treating all of them equally. Despite the fact that some of them might take the FA ranking into account, the relative importance is still not considered. Jacovi

and Goldberg (2020) have emphasized that faithfulness should be evaluated on a "grayscale" rather than "binary" (i.e. faithful or not) manner. However, current perturbation-based metrics, such as comprehensiveness and sufficiency, do not reflect a "grayscale" fashion as tokens are entirely removed or retained (e.g. comprehensiveness, sufficiency), or the rationale is entirely perturbed as a whole (e.g. sensitivity).

2.3 Evaluating Faithfulness Metrics

Quantitatively measuring the faithfulness of model explanations is an open research problem with several recent efforts focusing on highlighting the main issues of current metrics (Bastings and Filippova, 2020; Ju et al., 2022; Yin et al., 2022) and comparing their effectiveness (Chan et al., 2022).

A main challenge in comparing faithfulness metrics is that there is no access to ground truth, i.e. the true rationale for a model prediction (Jacovi and Goldberg, 2020; Ye et al., 2021; Lyu et al., 2022; Ju et al., 2022). Additionally, Ju et al. (2022) argue that it is risky to design faithfulness metrics based on the assumption that a faithful FA will generate consistent or similar explanations for similar inputs and inconsistent explanations for adversarial inputs (Alvarez-Melis and Jaakkola, 2018; Sinha et al., 2021; Yin et al., 2022).

Chan et al. (2022) introduced diagnosticity for comparing the effectiveness of faithfulness metrics. Diagnosticity measures the ability of a metric on separating random explanations (non-faithful) and non-random ones (faithful). They empirically showed that two perturbation metrics, sufficiency and comprehensiveness, are more 'diagnostic', i.e. effective in choosing faithful rationales compared to other metrics.

Despite the fact that sufficiency and comprehensiveness are in general more effective, they suffer from an out-of-distribution issue (Ancona et al., 2018; Bastings and Filippova, 2020; Yin et al., 2022). More specifically, the hard perturbation (i.e. entirely removing or retaining tokens) creates a discretely corrupted version of the original input which might fall out of the distribution the model was trained on. It is unlikely that the model predictions over the corrupted input sentences share the same reasoning process with the original full sentences which might be misleading for uncovering the model's true reasoning mechanisms.

3 Faithfulness Evaluation Metrics

3.1 Sufficiency and Comprehensiveness

We begin by formally defining sufficiency and comprehensiveness (DeYoung et al., 2020), and their corresponding normalized versions that allow for a fairer comparison across models and tasks proposed by Carton et al. (2020).

Normalized Sufficiency (NS): Sufficiency (S) aims to capture the difference in predictive likelihood between retaining only the rationale $p(\hat{y}|\mathcal{R})$ and the full text model $p(\hat{y}|\mathbf{X})$. We use the normalized version:

$$\begin{aligned} \mathrm{S}(\mathbf{X}, \hat{y}, \mathcal{R}) &= 1 - max(0, p(\hat{y}|\mathbf{X}) - p(\hat{y}|\mathcal{R})) \\ \mathrm{NS}(\mathbf{X}, \hat{y}, \mathcal{R}) &= \frac{\mathrm{S}(\mathbf{X}, \hat{y}, \mathcal{R}) - \mathrm{S}(\mathbf{X}, \hat{y}, 0)}{1 - \mathrm{S}(\mathbf{X}, \hat{y}, 0)} \end{aligned} \quad (1)$$

where $\mathrm{S}(\mathbf{x}, \hat{y}, 0)$ is the sufficiency of a baseline input (zeroed out sequence) and $\hat{y}$ is the model predicted class using the full text $\mathbf{x}$ as input.

Normalized Comprehensiveness (NC): Comprehensiveness (C) assesses how much information the rationale holds by measuring changes in predictive likelihoods when removing the rationale $p(\hat{y}|\mathbf{X}_{\setminus\mathcal{R}})$. The normalized version is defined as:

$$\begin{aligned} \mathrm{C}(\mathbf{X}, \hat{y}, \mathcal{R}) &= max(0, p(\hat{y}|\mathbf{X}) - p(\hat{y}|\mathbf{X}_{\setminus\mathcal{R}})) \\ \mathrm{NC}(\mathbf{X}, \hat{y}, \mathcal{R}) &= \frac{\mathrm{C}(\mathbf{X}, \hat{y}, \mathcal{R})}{1 - \mathrm{S}(\mathbf{X}, \hat{y}, 0)} \end{aligned} \quad (2)$$

3.2 Soft Nomralized Sufficiency and Comprehensiveness

Inspired by recent work that highlights the out-of-distribution issues of hard input perturbation (Bastings and Filippova, 2020; Yin et al., 2022; Zhao et al., 2022), our goal is to induce to sufficiency and comprehensiveness the relative importance of all tokens determined by a given FA. For this purpose, we propose Soft Normalized Sufficiency (Soft-NS) and Soft Normalized Comprehensiveness (Soft-NC) that apply a soft-erasure criterion to perturb the input.

Soft Input Perturbation: Given the vector representation of an input token, we aim to retain or remove vector elements proportionately to the token importance assigned by a FA by applying a Bernoulli distribution mask to the token embedding. Given a token vector $\mathbf{x}_i$ from the input $\mathbf{X}$ and its FA score a_i, we soft-perturb the input as follows:

$$\mathbf{x_i'} = \mathbf{x_i} \odot \mathbf{e_i}, \ \mathbf{e_i} \sim \mathrm{Ber}(q) \quad (3)$$

where Ber a Bernoulli distribution and $\mathbf{e}$ a binary mask vector of size n. Ber is parameterized with probability q:

$$q = \begin{cases} a, & \text{if retaining elements} \\ 1-a, & \text{if removing elements} \end{cases}$$

We repeat the soft-perturbation for all token embeddings in the input to obtain $\mathbf{x}'$. Our approach is a special case of dropout (Srivastava et al., 2014) on the embedding level.

Following Lakshmi Narayan et al. (2019), we have also tested two other approaches to soft perturbation in early-experimentation: (1) adding Gaussian noise to the embeddings; and (2) perturbing the attention scores, both in proportion to the FA scores. However, we found that dropout outperforms these two methods. Perhaps this is due to their sensitivity to hyperparameter tuning (e.g. standard deviation) which potentially contributes to their poor performance. Hence, we only conduct full experiments using dropout-based soft perturbation. Details on these alternative methods to perturb the input are included in Appendix C.

Soft Normalized Sufficiency (Soft-NS): The main assumption of Soft-NS is that the more important a token is, the larger number of embedding elements should be retained. On the other hand, if a token is not important most of its elements should be dropped. This way Soft-NS takes into account the complete ranking and importance scores of the FA while NS only keeps the top-k important tokens by ignoring their FA scores. We compute Soft-NS as follows:

$$\text{Soft-S}(\mathbf{X}, \hat{y}, \mathbf{X}') = 1 - max(0, p(\hat{y}|\mathbf{X}) - p(\hat{y}|\mathbf{X}'))$$

$$\text{Soft-NS}(\mathbf{X}, \hat{y}, \mathbf{X}') = \frac{\text{Soft-S}(\mathbf{X}, \hat{y}, \mathbf{X}') - \text{S}(\mathbf{X}, \hat{y}, 0)}{1 - \text{S}(\mathbf{X}, \hat{y}, 0)} \quad (4)$$

where $\mathbf{X}'$ is obtained by using $q = a_i$ in Eq. 3 for each token vector $\mathbf{x}'_\mathbf{i}$.

Soft Normalized Comprehensiveness (Soft-NC): For Soft-NC, we assume that the more important a token is to the model prediction, the heavier the perturbation to its embedding should be. Soft-NS is computed as:

$$\text{Soft-C}(\mathbf{X}, \hat{y}, \mathbf{X}') = max(0, p(\hat{y}|\mathbf{X}) - p(\hat{y}|\mathbf{X}'))$$

$$\text{Soft-NC}(\mathbf{X}, \hat{y}, \mathbf{X}') = \frac{\text{Soft-C}(\mathbf{X}, \hat{y}, \mathbf{X}')}{1 - \text{S}(\mathbf{X}, \hat{y}, 0)} \quad (5)$$

Dataset	Avg. Length	Classes	Size (Train/Dev/Test)	Avg. F1
SST	18	2	6,920 / 872 / 1,821	90.4 ± 0.5
AG	36	4	102,000 / 18,000 / 7,600	93.6 ± 0.2
Ev.Inf	363	3	5,789 / 684 / 720	82.3 ± 2.2
M.RC	305	2	24,029 / 3,214 / 4,848	74.0 ± 2.5

Table 1: Dataset statistics and mode prediction performance (average over five runs)

where $\mathbf{X}'$ is obtained by using $q = 1 - a_i$ in Eq. 3 for each token vector $\mathbf{x}'_\mathbf{i}$.

4 Experimental Setup

4.1 Tasks

Following related work on interpretability (Jain et al., 2020; Chrysostomou and Aletras, 2022b), we experiment with the following datasets:

- **SST**: Binary sentiment classification into positive and negative classes (Socher et al., 2013).
- **AG**: News articles categorized in Science, Sports, Business, and World topics (Del Corso et al., 2005).
- **Evidence Inference (Ev.Inf.)**: Abstract-only biomedical articles describing randomized controlled trials. The task is to infer the relationship between a given intervention and comparator with respect to an outcome (Lehman et al., 2019).
- **MultiRC (M.RC)**: A reading comprehension task with questions having multiple correct answers that should inferred from information from multiple sentences (Khashabi et al., 2018). Following DeYoung et al. (2020) and Jain et al. (2020), we convert this to a binary classification task where each rationale/question/answer triplet forms an instance and each candidate answer is labelled as True/False.

4.2 Models

Following Jain et al. (2020), we use BERT (Devlin et al., 2019) for SST and AG; SCIBERT (Beltagy et al., 2019) for EV.INF. and RoBERTa (Liu et al., 2019) for M.RC. See App. A for hyperparameters. Dataset statistics and model prediction performance are shown in Table 1.

4.3 Feature Attribution Methods

We experiment with several popular feature attribution methods to compare faithfulness metrics. We

do not focus on benchmarking various FAs but to improve faithfulness evaluation metrics.

- **Attention (α):** Token importance is computed using the corresponding normalized attention score (Jain et al., 2020).
- **Scaled attention ($\alpha\nabla\alpha$):** Attention scores scaled by their corresponding gradients (Serrano and Smith, 2019).
- **InputXGrad ($x\nabla x$):** It attributes importance by multiplying the input with its gradient computed with respect to the predicted class (Kindermans et al., 2016; Atanasova et al., 2020).
- **Integrated Gradients (IG):** This FA ranks input tokens by computing the integral of the gradients taken along a straight path from a baseline input (i.e. zero embedding vector) to the original input (Sundararajan et al., 2017).
- **DeepLift (DL):** It computes token importance according to the difference between the activation of each neuron and a reference activation, i.e. zero embedding vector (Shrikumar et al., 2017).

4.4 Computing Faithfulness with Normalized Sufficiency and Comprehensiveness

Following DeYoung et al. (2020), we compute the Area Over the Perturbation Curve (AOPC) for normalized sufficiency (NS) and comprehensiveness (NC) across different rationale lengths. AOPC provides a better overall estimate of faithfulness (DeYoung et al., 2020). We evaluate five different rationale ratios set to 1%, 5%, 10%, 20% and 50%, similar to DeYoung et al. (2020) and Chan et al. (2022).

4.5 Comparing the Diagnosticity of Faithfulness Metrics

Comparing faithfulness metrics is a challenging task because there is no a priori ground truth rationales that can be used.

Diagnosticity: Chan et al. (2022) proposed diagnosticity to measure the degree of a given faithfulness metric favors more faithful rationales over less faithful ones. The assumption behind this metric is that the importance scores assigned by a FA are highly likely to be more faithful than simply assigning random importance scores to tokens. Given an explanation pair (u, v), the diagnosticity is measured as the probability of u being a more faithful explanation than v given the same faithfulness metric F. u is an explanation determined by a FA, while v is a randomly generated explanation for the same input. For example the NC score of u should be higher than v when evaluating the diagnosticity of using NC as the faithfulness metric. More formally, diagnosticity $D_\varepsilon(F)$ is computed as follows:[2]

$$D_\varepsilon(F) \approx \frac{1}{|Z_\varepsilon|} \sum_{(u,v)\in Z_\varepsilon} \mathbb{1}(u \succ_F v) \quad (6)$$

where F is a faithfulness metric, Z_ε is a set of explanation pairs, also called ε-faithfulness golden set, $0 \leq \varepsilon \leq 1$. $\mathbb{1}\cdot$ is the indicator function which takes a value 1 when the input statement is true and a value 0 when it is false.

Chan et al. (2022) randomly sample a subset of explanation pairs (u, v) for each dataset and also randomly sample a FA for each pair. In our experiments, we do not sample but we consider all the possible combinations of data points and FAs across datasets.

5 Results

5.1 Diagnosticity of Faithfulness Metrics

We compare the diagnosticity of faithfulness metrics introduced in Section 3. Tables 2 and 3 show average diagnosticity scores across FAs and tasks, respectively. See App. B for individual results for each faithfulness metric, FA and dataset.

In general, we observe that Soft-NC and Soft-NS achieve significantly higher diagnosticity scores (Wilcoxon Rank Sum, $p < .01$) than NC and NS across FAs and datasets. The average diagnosticity of Soft-NC is 0.529 compared to 0.394 of NC while the diagnosticity of Soft-NS is 0.462 compared to NS (0.349). Our faithfulness metrics outperform NC and NS in 16 out of 18 cases, with the exception of Soft-NC on AG and Soft-NS on M.RC.

In Table 2, we note that both NC and Soft-NC consistently outperform Soft-NS and NS, which corroborates findings by Chan et al. (2022). We also see that using different FAs result into different diagnosticity scores. For example, diagnosticity ranges from 0.514 to .561 for Soft-NC while Soft-NS ranges from .441 to .480. We also observe similar behavior for NC and NS confirming results

[2]For a proof of Eq. 6, refer to Chan et al. (2022).

	α	$\alpha\nabla\alpha$	$x\nabla x$	IG	DL	Average
NC	.404	.405	.358	.428	.372	.394 (.025)
Soft-NC	**.525**	**.514**	**.526**	**.516**	**.561**	**.529*** (.017)
NS	.400	.383	.300	.368	.294	.349 (.044)
Soft-NS	**.479**	**.480**	**.444**	**.467**	**.441**	**.462*** (.017)

Table 2: Diagnosticity of soft normalized comprehensiveness (Soft-NC) and sufficiency (Soft-NS) compared to AOPC (hard) normalized comprehensiveness (NC) and sufficiency (NS) across FAs. * denotes a significant difference compared to its counterpart on the same FA, $p < .01$.

	SST	Ev.Inf	AG	M.RC	Average
NC	.409	.315	.416	**.434**	.394 (.046)
Soft-NC	**.431**	**.628***	**.649***	.406*	**.529*** (.111)
NS	.384	.344	**.385**	.282	.349 (.042)
Soft-NS	**.467**	**.560***	.294	**.527***	**.462*** (.102)

Table 3: Diagnosticity of faithfulness metrics across tasks. * denotes a significant difference compared to its counterpart on the same task, $p < .01$.

from Atanasova et al. (2020). Furthermore, we surprisingly see that various faithfulness metrics disagree on the rankings of FAs. For example DL is the most faithful FA measured by Soft-NC (.561) while NC ranks it as one of the least faithful (.372). However, Soft-NC and Soft-NS appear to be more robust by having less variance.

In Table 3, we observe that the diagnosticity of all four faithfulness metrics is more sensitive across tasks than FAs (i.e. wider range and higher variance). Also, we notice that in AG and M.RC, there is a trade-off between (Soft-)NS and (Soft-)NC. For example, on AG, Soft-NC is .649, the highest among all tasks but Soft-NS is the lowest. This result may be explained by the larger training sets of AG (102,000) and M.RC (24,029), compared to SST (6,920) and Ev.Inf (5,789) which might make the model more sensitive to the task-specific tokens.

5.2 Qualitative Analysis

We further conduct a qualitative analysis to shed light on the behavior of faithfulness metrics for different explanation pairs consisting of real and random attribution scores. Table 4 shows three examples from Ev.Inf, SST and AG respectively.

Repetitions in rationales affect faithfulness: Examining Example 1 (i.e. a biomedical abstract from Ev.Inf), we observe that the rationale (top 20% most important tokens) identified by DL contains repetitions of specific tokens, e.g. "aliskiren", "from", "in". On one hand, "aliskiren" (i.e. a drug for treating high blood pressure) is the main subject of the biomedical abstract and have been correctly identified by DL. On the other hand, we observe that many of these repeated tokens might not be very informative (e.g. many of them are stop words), however they have been selected as part of the rationale. This might happen due to their proximity to other informative tokens such as "aliskiren" due to the information mixing happening because of the contextualized transformer encoder (Tutek and Snajder, 2020).

We also notice that the random attribution baseline (Rand) selects a more diverse set of tokens that appear to have no connection between each other as expected. The random rationale also contains a smaller proportion of token repetitions. These may be the reasons why the random rationales may, in some cases, provide better information compared to the rationales selected by DL (or other FAs), leading to lower diagnosticity. Furthermore, NC between DL (.813) and Rand (.853) is very close (similar for NS) which indicates similar changes to predictive likelihood when retaining or removing rationales by DL and Rand. However, this may misleadingly suggest a similar model reasoning on the two rationales. We observe similar patterns using other FAs. Incorporating the FA importance scores in the input embeddings helps Soft-NC and Soft-S to mitigate the impact of issues above as they use all tokens during the evaluation.

Evenly distributed FA scores affect NC and NS: We also notice that for some inputs, the token importance assigned by FAs is very close to each other as demonstrated in Example 3, i.e. a news article from AG. The evenly distributed importance scores lead to similar low NC and NS between the FA (IG) and the random baseline attribution. Considering that the FA scores and ranking truly reflect the model reasoning process (i.e. the model made this prediction by equally weighing all tokens), then the faithfulness measurements provided by NS and NC might be biased.

We conjecture that this is likely to happen because these metrics entirely ignore the rest of the tokens even though these could represent a non-

	Text	FA	Metric	Faith.
1	TITLE: Long-term effects of aliskiren on blood pressure and the renin angiotensin - aldosterone system in hypertensive hemodialysis patients. ABSTRACT .OBJECTIVE: The long-term effects of aliskiren in hypertensive hemodialysis patients remain to be elucidated. ABSTRACT.DESIGN: In this post hoc analysis , we followed up 25 hypertensive hemodialysis patients who completed 8-week aliskiren treatment in a previous study for 20 months to investigate the blood pressure - lowering effect	DL	NC	.813
			Soft-NC	.984
			NS	.159
			Soff-NS	.904
		Rand	NC	.853
			Soft-NC	.351
			NS	.116
			Soff-NS	.055
2	by the end i was looking for something hard with which to bludgeon myself unconscious	$x\nabla x$	NC	.131
			Soft-NC	.339
			NS	.743
			Soff-NS	.975
		Rand	NC	.097
			Soft-NC	.101
			NS	.787
			Soff-NS	.557
3	ATHENS , Greece - Right now, the Americans are n't just a Dream Team - they 're more like the Perfect Team. Lisa Fernandez pitched a three - hitter Sunday and Crystl Bustos drove in two runs as the Americans rolled to their eighth shutout in eight days , 5-0 over Australia , putting them into the gold medal game ...	IG	NC	.186
			Soft-NC	.997
			NS	.016
			Soff-NS	.962
		Rand	NC	.194
			Soft-NC	.003
			NS	.020
			Suff-NS	.315

Table 4: Examples of inputs with their rationales (when taking the top 20% important tokens) and their different faithfulness metrics scores. Highlighted tokens are the rationales by a given FA and the random baseline. The tints indicate their importance scores, the lighter the less important. The three examples are from Ev.Inf, SST and AG, respectively.

negligible percentage of the FA scores distribution. However, Soft-NC and Soft-NS take into account the whole FA distribution without removing or retaining any specific tokens, hence they do not suffer from this limitation.

Different part of speech preferences for tasks We find that FAs tend to favor different parts of speech for different tasks. In Example 1 where the task is to reason about the relationship between a given intervention and a comparator in the biomedical domain, FAs tend to select proper nouns (e.g. "aliskiren") and prepositions (e.g. "on", "in" and "to"). On the other hand, in Example 2 which shows a text from SST, FAs favor adjectives (e.g. "unconscious" and "hard") for the sentiment analysis task. In Example 3, we see that articles such as "the" and proper nouns such as "Greece" and "Bustos" are selected.

6 Impact of Rationale Length on Faithfulness and Diagnosticity

Up to this point, we have only considered computing cumulative AOPC NC and NS by evaluating faithfulness scores at multiple rationale lengths together (see Section 3). Here, we explore how faithfulness and diagnosticity of NC and NS at individual rationale lengths compare to Soft-NC and Soft-NS. We note that both 'soft' metrics do not take the rationale length into account.

6.1 Faithfulness

Figure 2 shows the faithfulness scores of NC and NS at different rationale lengths for all FAs including random baseline attribution in each dataset.[3] We observe that the faithfulness scores of NC and NS follow an upward trend as the rationale length increases. This is somewhat expected because using information from an increasing number of tokens makes the rationale more similar to the original input.

In AG and SST, NC and NS lines appear close by or overlap. One possible reason is that the input text in SST and AG is relatively short (average length of 18 and 36 respectively), possibly leading to higher contextualization across all tokens. There-

[3]For brevity, we do not highlight the different FAs as they follow similar patterns.

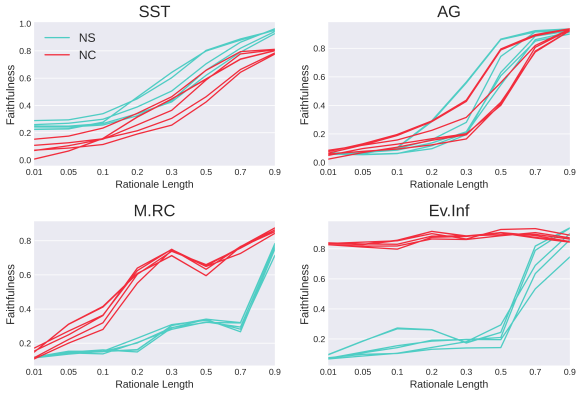

Figure 2: The impact of rationale length on normalized comprehensiveness (NC) and sufficiency (NS). Each line represents a FA.

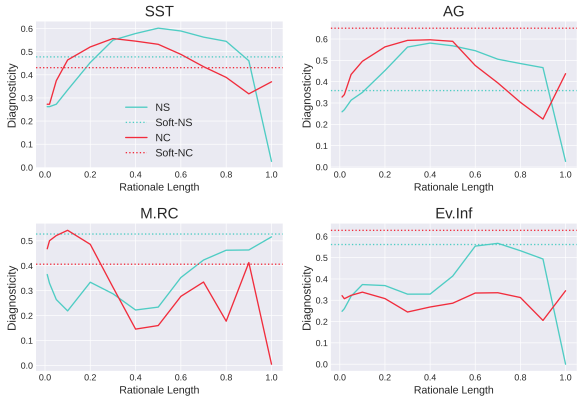

Figure 3: The impact of rationale length (shown in ratio) on Diagnosticity scores.

fore, removing or retaining more tokens results in a similar magnitude of changes in predictive likelihood.

In M.RC and Ev.Inf, two comprehension tasks that consist of longer inputs (average length is 305 and 365 respectively), we observe a different relationship between NC and NS. For instance, NC in Ev.Inf tends to be less impacted by the rationale length. This maybe due to the token repetitions in rationales discussed in Section 5.2. For example, when taking 2% of the top-k tokens out, e.g. 6 out of 300 tokens, all the task-related tokens may have been removed already.

6.2 Diagnosticity

Figure 3 shows the diagnosticity scores of NS and NC on different rationale lengths (average across FAs) together with the diagnosticity of Soft-NC and Soft-NS. Overall in all datasets, we see that the diagnosticity of NC and NS does not monotonically increase as we expected. In SST and AG, the diagnosticity of NS and NC both initially increase and then decrease. This happens because after increasing to a certain rationale length, the random selected rationales (used in the diagnosticity metric) contain sufficient information making it hard for FAs to beat. In M.RC and Ev.Inf, Soft-NC and Soft-NS have higher diagnosticity than NC and NS. One possible reason is that the corrupted version of input could fall out-of-distribution, confusing the model. Our 'soft' metrics mitigate this issue by taking all tokens into account.

Based on the observations on Figures 2 and 3, we conclude that it is hard to define an optimal rationale length for NC and NS which also has been demonstrated in previous work (Chrysostomou and Aletras, 2022b). In general, we see that diagnosticity decreases along with longer rationale length for NC and NS. On the other hand, faithfulness measured by NC and NS increases for longer rationales (Figure 2). Therefore, this might be problematic for selecting optimal rationale length for NC and NS. For example, if we want to select an optimal rationale length for M.RC by looking at its relation to faithfulness, we might choose a length of 30% over 20% because it shows higher NC and NS. However, the diagnosticity of NC and NS is lower at 30%, which means the higher NC and NS results to less trustful rationales. Our metrics bypass these issues because they focus on evaluating the FA scores and ranking as a whole considering all the input tokens. Soft-NC and Soft-NS do not require a pre-defined rationale length or evaluating faithfulness across different lengths.

We suggest that *it is more important to identify the most faithful FA given a model and task by taking into account all tokens rather than pre-defining a rationale of a specific length that ignores a fraction of the input tokens when evaluating faithfulness. The choice of how the FA importance scores will be presented (e.g. a top-k subset of the input tokens or all of them using a saliency map) should only serve practical purposes (e.g. better visualization, summarization of model rationales).*

7 Conclusion

In this work, we have proposed a new soft-perturbation approach for evaluating the faithfulness of input token importance assigned by FAs. Instead of perturbing the input by entirely removing or retaining tokens for measuring faithfulness, we incorporate the attribution importance by ran-

domly masking parts of the token embeddings. Our soft-sufficiency and soft-comprehensiveness metrics are consistently more effective in capturing more faithful FAs across various NLP tasks. In the future, we plan to experiment with sequence labeling tasks. Exploring differences in faithfulness metrics across different languages is also an interesting avenue for future work.

Limitations

This work focuses on binary and multi-class classification settings using data in English. Benchmarking faithfulness metrics in sequence labeling tasks as well as in multi-lingual settings should be explored in future work.

Acknowledgements

ZZ and NA are supported by EPSRC grant EP/V055712/1, part of the European Commission CHIST-ERA programme, call 2019 XAI: Explainable Machine Learning-based Artificial Intelligence. This project made use of time on Tier 2 HPC facility JADE2, funded by EPSRC (EP/T022205/1). We thank Huiyin Xue for her invaluable feedback.

References

David Alvarez-Melis and Tommi S Jaakkola. 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations (ICLR)*, 1711.06104.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. What is relevant in a text document?: An interpretable machine learning approach. *PloS one*, 12(8):e0181142.

Vijay Arya, Rachel K Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilovic, et al. 2021. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. In *INFORMS Annual Meeting*.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

Chun Sik Chan, Huanqi Kong, and Liang Guanqing. 2022. A comparative study of faithfulness metrics for model interpretability methods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5029–5038, Dublin, Ireland. Association for Computational Linguistics.

George Chrysostomou and Nikolaos Aletras. 2021a. Enjoy the salience: Towards better transformer-based faithful explanations with word salience. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8189–8200, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

George Chrysostomou and Nikolaos Aletras. 2021b. Improving the faithfulness of attention-based explanations with task-specific information for text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 477–488, Online. Association for Computational Linguistics.

George Chrysostomou and Nikolaos Aletras. 2022a. An empirical study on explanations in out-of-domain settings. In *Proceedings of the 60th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6920–6938, Dublin, Ireland. Association for Computational Linguistics.

George Chrysostomou and Nikolaos Aletras. 2022b. Flexible instance-specific rationalization of nlp models. In *The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*.

Gianna M Del Corso, Antonio Gulli, and Francesco Romani. 2005. Ranking a stream of news. In *Proceedings of the 14th international conference on World Wide Web*, pages 97–106.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 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.

Marina Fomicheva, Lucia Specia, and Nikolaos Aletras. 2022. Translation error detection as rationale extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4148–4159, Dublin, Ireland. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 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, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473.

Yiming Ju, Yuanzhe Zhang, Zhao Yang, Zhongtao Jiang, Kang Liu, and Jun Zhao. 2022. Logic traps in evaluating attribution scores. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5911–5922, Dublin, Ireland. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. *arXiv preprint arXiv:1611.07270*.

Pooja Lakshmi Narayan, Ajay Nagesh, and Mihai Surdeanu. 2019. Exploration of noise strategies in semi-supervised named entity classification. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 186–191, Minneapolis, Minnesota. Association for Computational Linguistics.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. 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)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Seventh International Conference on Learning Representations ICLR 2019*.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2022. Towards faithful model explanation in nlp: A survey. *arXiv preprint arXiv:2209.11326*.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.

Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji, and Yanjun Qi. 2021. Perturbing inputs for fragile interpretations in deep natural language processing. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 420–434, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Marcos Treviso and André F. T. Martins. 2020. The explanation game: Towards prediction explainability through sparse communication. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 107–118, Online. Association for Computational Linguistics.

Martin Tutek and Jan Snajder. 2020. Staying true to your word: (how) can attention become explanation? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 131–142, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xi Ye, Rohan Nair, and Greg Durrett. 2021. Connecting attributions and QA model behavior on realistic counterfactuals. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5496–5512, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. 2022. On the sensitivity and stability of model interpretations in NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2631–2647, Dublin, Ireland. Association for Computational Linguistics.

Muhammad Bilal Zafar, Michele Donini, Dylan Slack, Cedric Archambeau, Sanjiv Das, and Krishnaram Kenthapadi. 2021. On the lack of robust interpretability of neural text classifiers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3730–3740, Online. Association for Computational Linguistics.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Dongxu Zhang and Zhichao Yang. 2018. Word embedding perturbation for sentence classification. *arXiv preprint arXiv:1804.08166*.

Zhixue Zhao, George Chrysostomou, Kalina Bontcheva, and Nikolaos Aletras. 2022. On the impact of temporal concept drift on model explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4039–4054, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Model Hyperparameters

Dataset	Model	Batch Size	Learning Rate	Learning Rate (linear)
SST	bert-base-uncased	8	1e-5	1e-4
AG	bert-base-uncased	8	1e-5	1e-4
Ev.Inf	scibert_scivocab_uncased	4	1e-5	1e-4
M.RC	roberta-base	4	1e-5	1e-4

Table 5: Mode implementation details.

We use pre-trained models from the Huggingface library (Wolf et al., 2020). We use the AdamW optimizer (Loshchilov and Hutter, 2019) with an initial learning rate of $1e^{-5}$ for fine-tuning BERT. We fine-tune all models for 3 epochs using a linear scheduler, with 10% of the data in the first epoch as warming up. We also use a grad-norm of 1.0. The model with the lowest loss on the development set is selected. All models are trained across 5 random seeds, and we report the average. Experiments are run on a single Nvidia Tesla V100 GPU. Table 5 shows an overview of models and hyperparameters.

B Detailed Diagnosticity Results

Dataset	Feature	NS	Soft-NS	NC	Soft-NC
SST	Attention	0.406	0.496	0.349	0.407
SST	Scaled attention	0.387	0.509	0.352	0.396
SST	Gradients	0.324	0.495	0.394	0.394
SST	IG	0.437	0.489	0.535	0.395
SST	Deeplift	0.367	0.347	0.413	0.562
Ev.Inf	Attention	0.437	0.583	0.334	0.632
Ev.Inf	Scaled attention	0.448	0.576	0.329	0.624
Ev.Inf	Gradients	0.280	0.494	0.282	0.638
Ev.Inf	IG	0.294	0.564	0.298	0.615
Ev.Inf	Deeplift	0.263	0.582	0.331	0.633
AG	Attention	0.465	0.294	0.505	0.654
AG	Scaled attention	0.432	0.302	0.512	0.640
AG	Gradients	0.320	0.294	0.314	0.658
AG	IG	0.452	0.283	0.435	0.647
AG	Deeplift	0.256	0.296	0.315	0.648
M.RC	Attention	0.292	0.541	0.427	0.408
M.RC	Scaled attention	0.266	0.533	0.428	0.397
M.RC	Gradients	0.276	0.493	0.443	0.415
M.RC	IG	0.288	0.529	0.445	0.411
M.RC	Deeplift	0.290	0.538	0.428	0.400

Table 6: The diagnosticity of faithfulness metrics.

C Alternative implementations for soft perturbation

Adding Gaussian noise We perturb the pre-trained word embeddings with standard Gaussian noise. This Gaussian noise-based embedding perturbation is similar to the "statistical noise" used by Zhang and Yang (2018) and Lakshmi Narayan et al. (2019) for data augmentation. Specifically, we:

1. Multiply the token embedding with the token importance score, adding Gaussian noise. The resulting embedding is $\gamma\lambda \odot \mathbf{x_i}$ in Equation 7, where $\mathbf{x_i}$ is the original input embedding and λ is the FA scores (importance degree), γ is the hyperparameters based on the FA scores. $\odot$ is element-wise multiplication. As demonstrated by Lakshmi Narayan et al. (2019), adding Gaussian noise to the embedding requires tuning the standard deviation. Similarly, we tune the standard deviation $\sigma^2 \in \{0.005, 0.01, 0.05, 0.1, 0.5, 1, 2\}$ for soft-comprehensiveness and soft-sufficiency separately.

2. Add the embedding $\gamma\lambda \odot \mathbf{x_i}$, to the token embedding ($\mathbf{x_i}$) to obtain a perturbed embedding ($\mathbf{x'_i}$).

$$\mathbf{x'_i} = \mathbf{x_i} + \gamma\lambda \odot \mathbf{x_i}, \gamma \sim \mathcal{N}(\mu, \sigma^2) \quad (7)$$

An alternative way to add noise is to:

1. Generate a noise embedding by multiplying the token embedding with Gaussian noise with standard deviation, σ^2, associated with the importance score of the token. The embedding $\gamma \odot \mathbf{x_i}$ in Equation 8, where $\mathbf{x_i}$ is the original input embedding and λ is the importance score.

2. Add $\gamma \odot \mathbf{x_i}$, to the token embedding ($\mathbf{x_i}$) to get the perturbed embedding ($\mathbf{x'_i}$).

$$\mathbf{x'_i} = \mathbf{x_i} + \gamma \odot \mathbf{x_i}, \gamma \sim \mathcal{N}(\mu, \sigma^2) \quad (8)$$

Continuous attention mask We simply replace the binary-valued attention mask with a continuous-valued mask, where the continuous value is associated with the FA score for each token. The remaining part of the embeddings and the model remain the same.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4

☑ B1. Did you cite the creators of artifacts you used?
Yes, we use multiple datasets which are cited and presented in section 4

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Open source

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Use of public datasets

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 and Appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
also will provide a comprehensive list for the environment, packages and dependencies in the to-be-released git repo

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.