TREU : A Trainable Evaluation Metric for Natural Language Rationales

Anonymous ACL submission

Abstract

Explanable AI (XAI) and Natural Language Processing (NLP) researchers often rely on humans to annotate both labels and natural language rationales (explanations) with the goal that models can utilize these rationales to improve model performance, or can generate human-understandable explanations. However, human-annotated rationales are very subjective and could be low-quality, as some recent works discovered. The vital question arises: **how can we evaluate the quality of the human-annotated natural language rationales?** In this paper, we propose TREU , a **t**rainable **e**valuation metric that can evaluate the helpfulness of natural language **r**ationales towards models' prediction performances for a wide range of NLP tasks and models with the help of a **u**nified data structure. Our evaluation experiment on five popular datasets with two different model architectures demonstrates that TREU can coherently and faithfully evaluate the quality of rationales among datasets while the `Simulatability` metric fails. TREU score can also reveal rationale's quality towards specific classes in a multi-class classification task.

1 Introduction

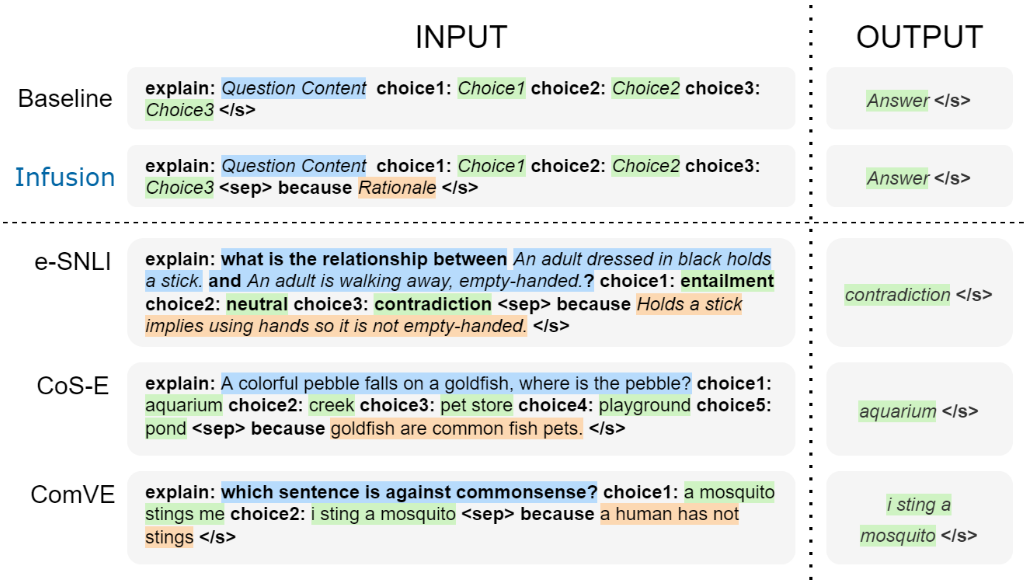
Despite today's large-scale language models (LLM) (Devlin et al., 2019; Radford et al., 2019; Lewis et al., 2019; Raffel et al., 2020) can exhibit close-to-human performance on many natural language processing (NLP) tasks (e.g., Question Answering (Rajpurkar et al., 2016; Kočiskỳ et al., 2018; Mou et al., 2021), Natural Language Inference (Bowman et al., 2015; Williams et al., 2017; Wang et al., 2018), and Text Generation (Duan et al., 2017; Yao et al., 2022)), human are eager to know how these State-of-the-Art (SOTA) models arrive at a prediction. Researchers working around natural language rationales[1] turned to human annotators for help by recruiting crowd-workers or domain experts to annotate both the labels and corresponding natural language rationales as explanations to their label annotation (Camburu et al., 2018; Rajani et al., 2019; Aggarwal et al., 2021), with which they can then leverage these human-annotated rationales to boost up models' performance or train models to generate explanations that people can understand.

However, the quality issue around such human-annotated rationales has been under-explored. Researchers intuitively leverage popular NLG metrics such as `BLEU` (Papineni et al., 2002), `ROUGE` (Lin, 2004) and `Simulatability` to evaluate the coherence and similarity between model-generated and human-annotated rationales, with a strong assumption that human-annotated rationales are the gold standard. We argue that the core value of rationales is how much helpfulness they can provide for the model's prediction performance instead of semantic similarity between each other.

Unlike labeling for the classification or multiple choice tasks, different people may come up with distinct and subjective natural language rationales for the same observation, and such rationales are task-dependent. As a result, human-annotated natural language rationales should not be simply treated as the gold standard, and the community is eager for a coherent metric that can automatically and truthfully evaluate the helpfulness of rationales towards models' prediction performance.

To fill this gap, we propose TREU score, a trainable evaluation metric for rationales to eval-

[1] In this paper, we use "rationales" and "natural language rationales" to refer to the collective concepts of "free-form rationale", "free-text explanation", and "natural language explanation", which differs from "rule-based" or "extractive" explanations.



	INPUT	OUTPUT
Baseline	**explain:** *Question Content* **choice1:** *Choice1* **choice2:** *Choice2* **choice3:** *Choice3* **</s>**	*Answer* **</s>**
Infusion	**explain:** *Question Content* **choice1:** *Choice1* **choice2:** *Choice2* **choice3:** *Choice3* **<sep> because** *Rationale* **</s>**	*Answer* **</s>**
e-SNLI	**explain: what is the relationship between** *An adult dressed in black holds a stick.* **and** *An adult is walking away, empty-handed.* **? choice1: entailment choice2: neutral choice3: contradiction <sep> because** *Holds a stick implies using hands so it is not empty-handed.* **</s>**	*contradiction* **</s>**
CoS-E	**explain:** A colorful pebble falls on a goldfish, where is the pebble? **choice1:** aquarium **choice2:** creek **choice3:** pet store **choice4:** playground **choice5:** pond **<sep> because** goldfish are common fish pets. **</s>**	*aquarium* **</s>**
ComVE	**explain: which sentence is against commonsense? choice1:** a mosquito stings me **choice2:** i sting a mosquito **<sep> because** a human has not stings **</s>**	*i sting a mosquito* **</s>**

Figure 1: Unified structure of Baseline and Infusion settings. Bold text are fixed prompts for each dataset. We show corresponding Infusion data format of classification task like e-SNLI and multiple choice task like CoS-E and ComVE into our unified structure. ECQA will share the same structure as CoS-E. The color schema follows: blue denotes the question content; green denotes the choice content; orange denotes the rationales.

uate the helpfulness of rationales towards models' performance faithfully. Furthermore, inspired by SOTA sequence-to-sequence language models (e.g., T5 (Raffel et al., 2020) and BART (Lewis et al., 2019)), we also propose a unified data format with template-based prompts to be used together with Treu metric, which can convert any classification or multiple choice tasks into a unified multiple choice generation task format. The benefit of the unified data format is that we can minimize the influence of structural variations across different tasks towards models' prediction performance so that Treu score can evaluate the helpfulness of rationales faithfully. We provide two settings for the unified data structure where researchers can decide to include rationales (Infusion hereinafter) or not include the rationales (Baseline hereinafter) into the input. Details are shown in Figure 1.

We conduct an experiment to compare the proposed Treu score against the current practice of the Simulatability score (Doshi-Velez and Kim, 2017) when evaluating the human-annotated rationale's quality on five popular datasets. The result shows that the Treu score can provide consistent evaluation ranks of the rationale's quality across all five datasets on two benchmark model architectures, while the Simulatability score fails. Our Treu takes into account the helpfulness of rationales during prediction for both models fine-tuned with Baseline and Infusion , while the Simulatability score only reflects the helpfulness of rationales on baseline models. As a result, in the case of two datasets with low Simulatability scores, our Treu metric suggests that the rationales in both datasets can provide helpfulness to prediction performance when the model is fine-tuned with these rationales under the Infusion setting. Furthermore, our Treu score can truthfully reflect quality issues with rationales for the specific class(es) in a classification task dataset with class-level Treu scores. We speculate that SOTA models have limited capabilities for interpreting the negation connotations that appear in large numbers in rationales of those datasets with low Treu scores. We conclude our paper with limitations and future research directions.

2 Related Work

2.1 Datasets with Natural Language Rationales

Despite the development of new model architectures and potentially more significant parameters, they still lack the ability to explain their prediction, which leads to the whole community being eager for human-annotated rationales to teach models either leverage rationales during training or be able to self-rationalize during prediction. For example, Wiegreffe and Marasovic (2021) recently reviewed 65 datasets and provided a 3-class taxonomy of explanations: highlights, free-text, and structured.

Dataset	Task	Task Format	Data Instances			Average Rationale Length (token)
			Train	Valid	Test	
CoS-E v1.0	Commonsense QA	3-choice Multiple-Choice	7610	950	-	16.148
CoS-E v1.11	Commonsense QA	5-choice Multiple-Choice	9741	1221	-	8.996
ECQA	Commonsense QA	5-choice Multiple-Choice	7598	1098	2194	63.572
e-SNLI	Natural Language Inference	3-label Classification	549367	9842	9824	15.977
ComVE	Commonsense Validation	2-choice Multiple-Choice	10000	1000	1000	10.288

Table 1: Task description and core statistics for popular large scale datasets with human-annotated natural language rationales that are included in the evaluation using our proposed TREU metric.

We focus on five large publicly available datasets that have human-annotated rationales at the instance level (Table 1). We double-checked these datasets' licenses, and there is no personally identifiable information.

The most prominent dataset is CoS-E and its two variants **CoS-E v1.0** and **CoS-E v1.11**(Rajani et al., 2019)). It extended the Commonsense Question-Answering (CQA v1.0 and v1.11 versions) dataset (Talmor et al., 2018) by adding human-annotated rationales to the single correct answer choice. However, a few recent works suggest that the CoS-E's rationale quality is not good, as Narang et al. (2020) independently hand-labeled some new rationales for CoS-E and found a very low BLEU score between its original rationales and the new ones. To improve the rationale's quality, **ECQA** (Aggarwal et al., 2021) recruited human annotators to add a single-sentence explanation for every answer option, then summarized them into a natural language rationale for every data instance in the CQA v1.11 dataset. Sun et al. (2022) proved that CoS-E rationales are not as good as ECQA rationales as human evaluators do not believe CoS-E rationales can provide additional information to support their decision makings. The fourth dataset is **e-SNLI** (Camburu et al., 2018), which consists with rationales for the Stanford Natural Language (SNLI) dataset (Bowman et al., 2015). The fifth dataset is **ComVE** (Wang et al., 2020) that asks which one of two sentences is against commonsense. Later we evaluate the proposed TREU metric against the baseline metric for the quality of human-annotated natural language rationales using all these five datasets.

Worth mentioning we do not include datasets such as **SBIC** (Sap et al., 2019) or **E-δ-NLI** (Brahman et al., 2021) because the former does not provide rationales for all the data, while the latter approaches to generate rationales through various sources to augment the δ-NLI (Rudinger et al., 2020) dataset instead of human annotations.

2.2 Evaluation Metric for Rationales

Many commonly used evaluation metrics for text-based content like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) treat human-annotated answers as the absolute gold standard, which cannot evaluate the quality of them. One established evaluation metric called `Simulatability` score derives from Human Simulatability (Doshi-Velez and Kim, 2017) and can examine gold rationales. It simply measures the change in a baseline model's prediction performance, whether the rationale is provided as the input or not. Previous works (Chandrasekaran et al., 2018; Yeung et al., 2020; Hase et al., 2020; Wiegreffe et al., 2020; Poursabzi-Sangdeh et al., 2021; Rajagopal et al., 2021) have demonstrated the usefulness of `Simulatability` score for evaluating rationale quality. However, this metric has a couple of inherent disadvantages. First, it only considers the helpfulness of rationales as input during prediction on a baseline model, where we prove that rationales provide different helpfulness during fine-tuning and prediction in our preliminary experiment. Besides, a model's performance could also differ when we form the original task with the same data into other tasks, such as a classification task into a multiple-choice task or a generation task with different input data prompts and structures. In contrast, our proposed TREU evaluation metric complements both drawbacks of the `Simulatability` score by considering the helpfulness of rationales both at fine-tuning and predicting with the help of a unified structure to minimize the impact of task differences.

2.3 Usage of Rationales for SOTA models

Existing works have been exploring circumstances in which rationales can or cannot improve model performance; for example, Hase and Bansal (2021) argues that rationales are most suitable for use as model input for predicting. Some recent works have been trying to generate better rationales with a self-rationalization setting (Wiegreffe et al., 2020;

Marasović et al., 2021), where a model is asked to generate the prediction label and rationale at the same time. We conduct a preliminary experiment to find the best setting for models to leverage rationales for better prediction performance in Section 4.1. There also exist many recent works (Paranjape et al., 2021; Liu et al., 2021; Chen et al., 2022) that explore the usage of prompts to complete rationales, generate knowledge as additional information for the original task, or examine whether generated rationales can provide robustness to adversarial attacks. Another related line of research focuses on extracting or generating rationales with a unified framework (Chan et al., 2022) or with a teachable reasoning system that generates chains of reasoning (Dalvi et al., 2022).

3 Unified Structure

While popular metrics like BLEU and ROUGE can evaluate text coherence and similarity, what is vital to rationales is how much helpfulness they can provide for the model's prediction. The desiderata are to develop a metric that can faithfully evaluate rationales' utility towards model performance. We expect an excellent metric can systematically demonstrate how good or bad the rationales are, for example, what does 'noisy' mean in a human study from previous works on CoS-E rationales.

With the advantage of sequence-to-sequence models like T5 that can map different types of language tasks into generation tasks, we can control and minimize the influence of varying task formats on model performance while evaluating the helpfulness of rationales by leveraging a unified data format. We realize that existing datasets with human-annotated rationales are mostly either multiple choices tasks or classification tasks, and the classification task could be viewed as a multiple-choice task where the labels are indeed choices. Inspired by several previous works that manipulated prompts for sequence-to-sequence models (Marasović et al., 2021; Liu et al., 2021), we incorporate a few well-defined words as a template-based prompt for the unified data structure to indicate the task content and corresponding rationale.

Examples are shown in Figure 1 to explain how to map various tasks into a unified multiple choice generation task. We propose two settings: no rationales (`Baseline`) and rationales as additional input (`Infusion`). Several template-based prompt words across all datasets infer the data's basic structure, such as 1). '*explain*' is the leading word followed by the question content, 2). everything after '*choicen*' will be a candidate answer to be generated, and 3). special token '<*sep*>' separates the rationales from the task content, and the rationales in `Infusion` will be led by 'because' so that the model can infer the rationales are going to explain the task content. For datasets like CoS-E and ECQA, we leverage the original task as the question content. On the other hand, we define a fixed question prompt for e-SNLI: "*what is the relation between [Premise] and [Hypothesis]?*" and for ComVE: "*which sentence is against commonsense?*" to specify corresponding tasks to models.

Rationales as Input vs Output	Fine-tune Setting		
	`Baseline`	Self-rationalization	`Infusion`
CoS-E v1.0	0.695	0.646	**0.878**
ECQA	0.572	0.513	**0.989**

Table 2: Preliminary experiment results of using rationales as part of Input(`Infusion`) vs Output(Self-rationalization) vs without rationales (`Baseline`) on CoS-E and ECQA datasets.

4 Preliminary Experiment

4.1 Utilizing Rationales as Part of Input V.S. Part of Output

Recent works have been exploring various circumstances that human-annotated rationales could help in different aspects; for example, Hase and Bansal (2021) argued that explanation as additional input would best suit performance improvement. Marasović et al. (2021) proposed self-rationalize models, which generates rationales along with prediction label, and can generate more reasonable rationales. However, they do not provide prediction accuracy compared with the baseline. We hypothesize that leveraging rationales as additional input information with the original task input allows models to use rationales for better prediction, while the self-rationalization setting complicates the prediction task for the models and may lead to a prediction performance decrease. We conduct a preliminary experiment on CoS-E v1.0 and ECQA datasets to justify our hypothesis.

We fine-tune a T5-base model on each dataset with three different settings: `Baseline` , `Infusion` , and rationales as additional output (Self-rationalization hereinafter) For each model, we maintain the same setting during fine-tuning and inference. For example, the model fine-tuned with

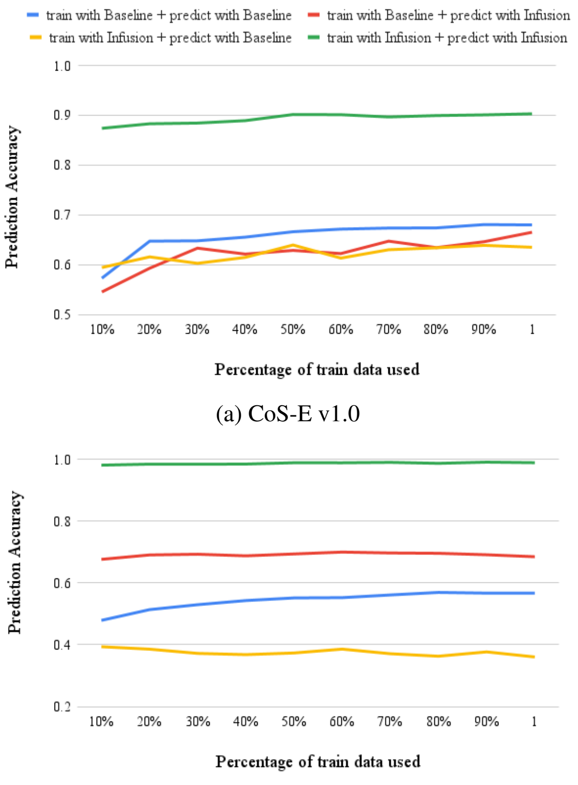
Infusion will also take data under Infusion during inference. We leverage the unified structure for Baseline and Infusion shown in Figure 1 and make minor adjustments for the self-rationalization setting accordingly (shown in Appendix A).

The experiment results are shown in Table 2. We notice that the self-rationalization setting performs worse than the Baseline , which is aligned with our assumption. On the other hand, the Infusion setting surprisingly achieves significant improvement on CoS-E, which was considered 'noisy' by previous works, demonstrating that the CoS-E rationales still provide helpfulness to the model's performance. The Infusion setting also approaches nearly complete correctness on the ECQA dataset.

4.2 Rationales as Partial Input During Fine-Tuning

In order to examine what is the utility of rationales to the models during fine-tuning, we perform an in-depth experiment with the Baseline and Infusion setting. First, we fine-tune a series of models with gradually increased training data and analyze the models' prediction performance. More specifically, we randomly shuffle and select 9 sub-datasets of varying amounts of data ranging from 10% to 90% of the training data in each dataset we used in the first preliminary experiment. Then, for each sub-dataset, we fine-tune three different models with randomly shuffled random seeds for sampling and fine-tuning, then acquire the average prediction performance over three models. As a result, for each CoS-E v1.0 and ECQA dataset, we get 60 models fine-tuned with varying amounts of data for both the Baseline and Infusion setting, including the models fine-tuned on full training data, then perform prediction with the Baseline and Infusion settings. We maintain the same hyper-parameters across the models fine-tuned for this experiment and report them in Appendix B.1.

From the two prediction performance diagrams in Figure 2 (detailed results in Appendix 4), we notice that the addition of training data does not consistently improve the performance of models fine-tuned with the Infusion setting (yellow and green line), proving that the fine-tuning process is not teaching the model with new knowledge that was supposed to be conveyed in the rationales. Besides, the models fine-tuned with Infusion setting perform worse than baselines when no rationales are provided during inference (yellow and blue line correspondingly), demonstrating that the fine-tuning of Infusion setting teaches the models to rely on the rationale part of the input to predict. Additionally, we observe that the baseline models for CoS-E perform worse while predicting with Infusion setting than with Baseline . In contrast, the baseline models for ECQA consistently exceed baseline performance by a significant margin while predicting with the Infusion setting (the red lines). This observation is aligned with previous works that many of the rationales in the CoS-E dataset are low-quality, while the rationales in ECQA have much better quality. The preliminary experiment demonstrates that the rationales provide different helpfulness during fine-tuning and inference. Thus, both situations should be considered while evaluating the quality of rationales.



(a) CoS-E v1.0

(b) ECQA

Figure 2: Rationales as partial input during fine-tuning on CoS-E v1.0 (top) and ECQA (bottom) with different amount of training data. We perform fine-tuning and predicting for both Baseline and Infusion settings.

5 TREU Evaluation Metric

5.1 Definition

From the preliminary experiment, we have observed that 1) rationales provide the most helpfulness as additional input, and 2) rationales pro-

$$\textsc{Treu} = \underbrace{\left(\mathrm{Accu}(\mathrm{M}_{\texttt{Infusion}\,|\,\texttt{Infusion}}) - \mathrm{Accu}(\mathrm{M}_{\mathrm{Baseline}\,|\,\mathrm{Baseline}}) \right)}_{\textit{Helpfulness @ fine-tuning \& inference}}$$

$$+ \underbrace{\left(\mathrm{Accu}(\mathrm{M}_{\texttt{Infusion}\,|\,\mathrm{Baseline}}) - \mathrm{Accu}(\mathrm{M}_{\mathrm{Baseline}\,|\,\mathrm{Baseline}}) \right)}_{\textit{Helpfulness @ inference only}}$$

Figure 3: The formula of our TREU metric. *M* denotes a model and the subscript denotes (inference setting | fine-tune setting). Our score incorporates the helpfulness of rationales at both fine-tuning and inference.

vide different helpfulness to the model's prediction performance during fine-tuning and inference. As a result, we propose TREU score along with the unified structure we proposed in Section 3. Figure 3 shows the formula of TREU score. TREU score evaluates the quality of rationales with the sum of two parts: helpfulness at inference only, where the model is fine-tuned with `Baseline` setting, and at fine-tuning as well as inference, where the model is fine-tuned with `Infusion` setting. In each part, the helpfulness is calculated by the prediction performance difference between the `Infusion` and `Baseline` settings on the same model.

A positive score demonstrates that the rationales provide overall helpfulness for better prediction, while a negative score does not necessarily mean the rationales are not helpful. Instead, a negative score indicates the rationales lead to the model's prediction performance drop in at least one part of the evaluation. By further analyzing the intermediate score for each part, researchers can locate when rationales do not help improve the model's performance. As a result, the TREU score ranges theoretically from -2 to 2. In comparison, the `Simulatability` score only considers the second part within our TREU score formula and without the control of influence from various task structures.

5.2 Experiment

We evaluate human-annotated natural language rationales across five popular datasets using our evaluation metric and the `Simulatability` score. To justify that our TREU score is independent of the specific model architecture and to examine the influence of different pre-trained models over the prediction performance, we perform experiments with five datasets on T5 and BART with the base models as the backbone for fine-tuning. During our evaluation, we also leverage the unified data structure for the `Simulatability` score to make it a much stronger baseline.

We maintain the same fine-tuning hyper-parameter for all the models in the experiment (details in Appendix B.2). The only exception is the e-SNLI dataset, which has about 10x the size (549,367 data instances) of training data compared to the other datasets. Therefore, we only fine-tune models on e-SNLI dataset with two epochs. Furthermore, for the experiments with BART, we leverage the special token '<*s*>' used during the pre-training process instead of '<*sep*>' and ask the BART tokenizer to add special tokens during tokenization automatically.

Table 3 presents the evaluation results. The ordering of datasets in each table is based on our TREU score, while the inconsistent ranking of the `Simulatability` score is marked red. We further provide TREU score by class for the dataset of classification task, which is e-SNLI, to examine the difference of helpfulness per class, where *class1* / *class2* / *class3* refers to *entailment* / *neutral* / *contradiction* correspondingly.

5.3 Observation

By first comparing the models' prediction results based on two architectures, we notice performance differences among all datasets. More specifically, all models fine-tuned on T5-base outperform the ones fine-tuned on BART-base with the same setting, mostly with a significant margin. Despite apparent performance differences between model architectures, by looking at the orderings of datasets in both tables, which are based on our TREU score, We can easily observe that TREU score provides the same ranking result for the quality of rationales in 5 datasets over two model architectures. Based on TREU scores in Table 3, relatively speaking, rationale quality varies in the different datasets roughly in the following order:

ECQA > CoS-E v1.11 > CoS-E v1.0 > e-SNLI > ComVE

According to TREU score, rationales in ECQA have the best quality among five datasets. Especially, rationales in ECQA are much better than the ones in both CoS-E datasets, which is consistent with previous works' consensus that rationales in CoS-E are much worse than ECQA. It is worth noticing that both CoS-E datasets achieve positive TREU scores, though significantly lower than the ones for ECQA, demonstrating that rationales in CoS-E datasets still have positive overall help-

T5-base	Fine-tune with Baseline		Fine-tune with Infusion	Simulatability Score	TREU Score	TREU score by class		
	predict with Baseline	predict with Infusion	predict with Infusion			class1	class2	class3
ECQA	0.572	0.746	0.989	0.174	**0.591**			
CoS-E v1.11	0.608	0.610	0.803	0.002	**0.197**			
CoS-E v1.0	0.695	0.645	0.878	-0.05	**0.133**			
e-SNLI	0.907	0.676	0.981	-0.231	**-0.157**	**0.13**	**-0.483**	**0.094**
ComVE	0.88	0.527	0.949	-0.353	**-0.284**			

BART-base	Fine-tune with Baseline		Fine-tune with Infusion	Simulatability Score	TREU Score	TREU score by class		
	predict with Baseline	predict with Infusion	predict with Infusion			class1	class2	class3
ECQA	0.428	0.438	0.901	0.010	**0.483**			
CoS-E v1.11	0.443	0.449	0.700	0.006	**0.263**			
CoS-E v1.0	0.512	0.486	0.790	-0.026	**0.252**			
e-SNLI	0.888	0.658	0.978	-0.23	**-0.14**	**0.115**	**-0.277**	**-0.271**
ComVE	0.812	0.596	0.864	-0.216	**-0.164**			

Table 3: Evaluation results of human-annotated rationales in 5 datasets with our TREU score and Simulatability score. The tables above and below correspond to models fine-tuned on T5-base and BART-base, respectively. The ordering of datasets is based on our TREU score, and the inconsistent ranking of Simulatability score is marked red. For e-SNLI which is the only classification task in the experiment, *class1* / *class2* / *class3* refers to *entailment* / *neutral* / *contradiction* respectively.

fulness for models' prediction performance even though they are considered 'low quality and noisy' by human experiments in previous works.

On the other hand, the Simulatability score cannot provide a consistent ranking of rationale quality among five datasets on two model architectures. Based on the two models, Simulatability score provides two distinct rankings:

T5-base:

ECQA > CoS-E v1.11 >

CoS-E v1.0 > e-SNLI > ComVE

BART-base:

ECQA > CoS-E v1.11 >

CoS-E v1.0 > ComVE> e-SNLI

From Table 3, the Simulatability score ranks e-SNLI and ComVE reversely on fine-tuned BART models compared with fine-tuned T5 models, indicating Simulatability score could be more affected by different model architectures even with the unified data structure.

We notice that ComVE ranks worst among five datasets in both tables, indicating the rationales in ComVE are the least helpful for the models to either fine-tuned or predict. Since the ComVE task asks models to predict which sentence is more likely ***against*** commonsense, the question itself implies a negation connotation. Likewise, many ComVE rationales contain negation, such as the one in Figure 1. The concept of negation has always been a relatively complex concept for machines. Although both T5 and BART models fine-tuned with the Baseline setting are able to perform relatively well on ComVE, the addition of rationales that contain negation during inference is likely to create more difficulties for the models to interpret the information, which eventually leads to a significant performance drop.

One advantage of using our TREU score to evaluate the quality of rationales is that we can further decompose and analyze the score by class or intermediate results from different fine-tuning and predicting settings. For instance, we observe that the TREU scores for the e-SNLI dataset with T5 and BART models are both negative, indicating that the quality of rationales within e-SNLI could be poor. By looking into the intermediate results, though the baseline models receive significant performance drops while predicting with Infusion compared with the Baseline setting, the models that are fine-tuned with Infusion still outperform the baseline models while predicting with Infusion , justifying the rationales indeed provide improvements under this setting. Looking further into the decomposed TREU score of e-SNLI on class level, we notice that rationales for the specific class(s) in the e-SNLI dataset provide a much lower TREU score by class than the others, which causes the overall score to decline. More specifically, the models fine-tuned on T5 and BART have more than 40% prediction accuracy drop on data with the ground-truth label 'neutral' when they are fine-tuned with Baseline and predicted with Infusion . Moreover, TREU score by class indicates that the fine-tuned BART models have about 40% prediction accuracy drop on data with ground-truth 'contradiction' labels.

We suspect human annotators behave differently while providing rationales for data with various categories in e-SNLI. For instance, human annotators may explain why two sentences are 'entailment' by describing the shared information conveyed by both sentences. However, by inspection, we notice humans tend to provide counter-examples to explain why two sentences are 'neutral' or 'contradiction' classes. Besides, humans also like to negate the universal correctness of the contents described in two sentences for these classes. We provide representative examples for each class in Appendix 5. Such behavior's tendency to use many negation connotations for rationales may cause difficulty for the baseline models to interpret the information and falsely make predictions.

Nevertheless, these models can correctly understand rationales for all categories after being fine-tuned with rationales under the Infusion setting. Worth pointing out that ECQA rationales are summarized from positive and negative properties for each candidate choice which also contains negation words, but those negation words mostly appear in negative properties for wrong choices instead of the positive property for correct choices. As a result, we notice the pre-trained baseline models are able to leverage ECQA rationales with Infusion during the predicting process and achieve performance improvement. Since we are the first to discover such a class-level drop on e-SNLI by using TREU score, we only propose our hypothetical assumption and leave a definitive study for future work.

6 Limitations and Risks

Our paper shows the proposed TREU score can be used to measure the quality of rationales towards the models' prediction performance on multiple experiment datasets. However, our evaluation are only on the human-created natural language rationales, and it is a natural next step that the TREU could be used for evaluating the helpfulness of model-generated rationales. We would like to caution readers of this paper when they apply the TREU to the model-generated rationales: This metric and our evaluation experiment require the model to generate rationales for the training data split in the datasets, and then use the train split with generated rationales to fine-tune the model with the Infusion setting. Last but not least, we acknowledge the proposed TREU is only one way of automatically evaluating the natural language rationale's quality, and there may be many other ways; besides, the high TREU score may not necessarily reflect the human-perceived quality if we ask human to rate it, as our calculation only measure its helpfulness from the modeling perspective.

7 Conclusion and Future Work

In this paper, we propose the TREU score as a faithful evaluation metric for human-annotated natural language rationales regarding the helpfulness to models' performance for a variety of nlp tasks. We design the TREU score to consider rationales' helpfulness at both fine-tuning with inference and inference-only settings which is based on the discoveries of two preliminary experiments: 1). rationales provide the most helpfulness while being used as additional input and 2). the helpfulness of rationales differs significantly between models fine-tuned and not fine-tuned with rationales. We also propose a unified data structure for TREU that minimizes the influence of tasks' differences by mapping various tasks to a unified multiple choice generation task. Finally, we perform the evaluation of human-annotated rationales in 5 popular large-scale datasets with two different sequence-to-sequence model architectures.

Evaluation results demonstrate that TREU score can consistently reflect the relative rank of rationale qualities among five datasets while an established metric fails, and the reflected quality of rationales by TREU score is aligned with previous works. we also hypothesize SOTA models have limited ability to interpret the negation connotations or counter-examples that appear in large numbers in rationales with low TREU scores. To the best of our knowledge, we are the first to propose a faithful evaluation metric for human-annotated rationales, which leads to envisioning many avenues for future work. We would expand the evaluation on other datasets with human-labeled rationales and suggest researchers leverage our TREU metric as an essential quality check while collecting rationales in the future. We would also continue to evaluate and analyze human natural language rationales, which may lead to other inconspicuous properties that could be beneficial for developing better reasoning systems.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for commonsenseqa: New dataset and models. In *Workshop on Commonsense Reasoning and Knowledge Bases*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. 2021. Learning to rationalize for nonmonotonic reasoning with distant supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12592–12601.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

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

Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make vqa models more predictable to a human? *arXiv preprint arXiv:1810.12366*.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can rationalization improve robustness? *arXiv preprint arXiv:2204.11790*.

Bhavana Dalvi, Oyvind Tafjord, and Peter Clark. 2022. Towards teachable reasoning systems. *arXiv preprint arXiv:2204.13074*.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 866–874.

Peter Hase and Mohit Bansal. 2021. When can models learn from explanations? a formal framework for understanding the roles of explanation data. *arXiv preprint arXiv:2102.02201*.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? *arXiv preprint arXiv:2010.04119*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2021. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*.

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E Peters. 2021. Few-shot self-rationalization with natural language prompts. *arXiv preprint arXiv:2111.08284*.

Xiangyang Mou, Chenghao Yang, Mo Yu, Bingsheng Yao, Xiaoxiao Guo, Saloni Potdar, and Hui Su. 2021. Narrative question answering with cutting-edge open-domain QA techniques: A comprehensive study. *Transactions of the Association for Computational Linguistics*, 9:1032–1046.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Prompting contrastive explanations for commonsense reasoning tasks. *arXiv preprint arXiv:2106.06823*.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability.

In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard Hovy, and Yulia Tsvetkov. 2021. Selfexplain: A self-explaining architecture for neural text classifiers. *arXiv preprint arXiv:2103.12279*.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Rachel Rudinger, Vered Shwartz, Jena D Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

Jiao Sun, Swabha Swayamdipta, Jonathan May, and Xuezhe Ma. 2022. Investigating the benefits of free-form rationales. *arXiv preprint arXiv:2206.11083*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. Semeval-2020 task 4: Commonsense validation and explanation. *arXiv preprint arXiv:2007.00236*.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Sarah Wiegreffe, Ana Marasović, and Noah A Smith. 2020. Measuring association between labels and free-text rationales. *arXiv preprint arXiv:2010.12762*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Li, Mo Yu, and Ying Xu. 2022. It is AI's turn to ask humans a question: Question-answer pair generation for children's story books. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–744, Dublin, Ireland. Association for Computational Linguistics.

Arnold Yeung, Shalmali Joshi, Joseph Jay Williams, and Frank Rudzicz. 2020. Sequential explanations with mental model-based policies. *arXiv preprint arXiv:2007.09028*.

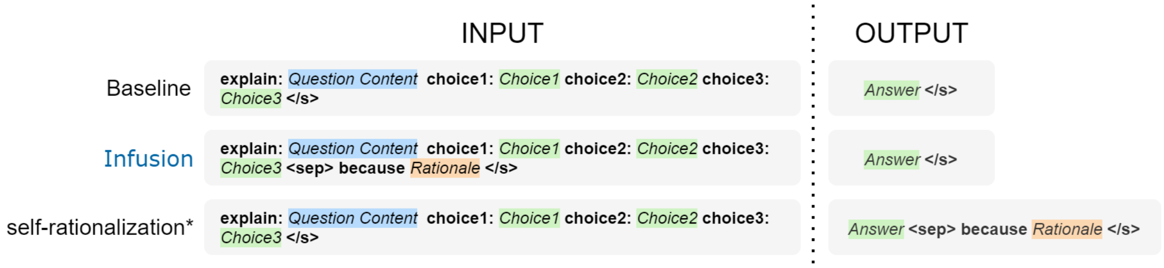

INPUT
OUTPUT
Baseline
explain: Question Content choice1: Choice1 choice2: Choice2 choice3: Choice3 </s>
Answer </s>
Infusion
explain: Question Content choice1: Choice1 choice2: Choice2 choice3: Choice3 <sep> because Rationale </s>
Answer </s>
self-rationalization*
explain: Question Content choice1: Choice1 choice2: Choice2 choice3: Choice3 </s>
Answer <sep> because Rationale </s>

Figure 4: Unified structure of `Baseline` , `Infusion` , and self-rationalization settings. Bold text are fixed prompts for each dataset.

Appendix

A Implementation of self-rationalization format

We show the implementation of self-rationalization setting proposed by Marasović et al. (2021) and put it together in Figure 4 with our proposed unified structure of the `Baseline` and `Infusion` setting.

B Experiment Hyper-Parameters

We perform all the computational experiments on a Google Colab instance with a single Nvidia V100 GPU and 50 Gigabytes of RAM.

B.1 Hyper-parameter for Preliminary Experiment

For the preliminary experiment of utilizing rationales as part of input V.S. part of output, we leverage the following hyper-parameters for all models with different data structures: *max_len* : 512, *target_max_len* : 64, *train_batch_size* : 1, *learning_rate* : $5e^{-5}$, *num_train_epochs* : 12.

For the preliminary experiment of rationales as partial input during fine-tuning, we maintain the following hyper-parameters for all models fine-tuned with partial/full train data of CoS-E and ECQA datasets: *max_len* : 512, *target_max_len* : 16, *train_batch_size* : 1, *learning_rate* : $1e^{-4}$, *num_train_epochs* : 6.

B.2 Hyper-parameter for Rationale Evaluation with five Datasets

For the evaluation of human-annotated rationales on 5 different datasets, we maintain the following hyper-parameters for all the models: *max_len* : 512, *target_max_len* : 64, *train_batch_size* : 1, *learning_rate* : $5e^{-5}$, *num_train_epochs* : 12. The only exception is the e-SNLI dataset, which has about 10x size (549,367 data instances) of training data compared to the other datasets. Therefore, we only fine-tune models on e-SNLI dataset with 2 epochs.

C Results for Preliminary Experiment - Rationales as Partial Input During Fine-tuning

For the preliminary experiment of rationales as partial input during fine-tuning, we randomly shuffle 3 seeds to select the subset of data and fine-tune the model. The detailed results of each experiment and average accuracy is reported in Table 4.

D Examples of different rationales for each category in e-SNLI dataset

From our evaluation results, we suspect human annotators behave differently while providing rationales for data with various categories in e-SNLI. For instance, human annotators may explain why two sentences are ‘entailment’ by describing the shared information or similarities conveyed by both sentences. However, humans tend to provide counter-examples or negations to explain why two sentences are not related (neutral) or contradiction. Here in Table 5, we show representative examples of data with corresponding rationales for each class.

	Fine-tune with Baseline on CoS-E v1.0									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	1
Predict Baseline	0.583	0.656	0.638	0.658	0.661	0.670	0.674	0.678	0.697	0.676
	0.550	0.644	0.664	0.650	0.666	0.667	0.667	0.682	0.668	0.682
	0.584	0.64	0.64	0.655	0.670	0.675	0.677	0.66	0.674	0.68
Average	0.572	0.647	0.647	0.655	0.665	0.671	0.673	0.673	0.680	0.679
Predict Infusion	0.586	0.586	0.625	0.633	0.596	0.621	0.663	0.655	0.649	0.676
	0.561	0.591	0.642	0.609	0.656	0.630	0.618	0.650	0.641	0.652
	0.525	0.6	0.631	0.62	0.631	0.614	0.658	0.595	0.647	0.665
Average	0.545	0.592	0.632	0.621	0.628	0.622	0.647	0.634	0.645	0.664

	Fine-tune with Infusion on CoS-E v1.0									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	1
Predict Baseline	0.588	0.622	0.617	0.613	0.635	0.616	0.615	0.625	0.652	0.629
	0.592	0.614	0.573	0.610	0.650	0.592	0.632	0.64	0.610	0.64
	0.601	0.609	0.615	0.618	0.631	0.629	0.641	0.635	0.652	0.634
Average	0.594	0.615	0.602	0.614	0.639	0.612	0.629	0.633	0.638	0.634
Predict Infusion	0.867	0.874	0.884	0.889	0.902	0.894	0.890	0.886	0.910	0.904
	0.875	0.888	0.881	0.890	0.898	0.901	0.9	0.901	0.896	0.895
	0.877	0.885	0.887	0.887	0.903	0.907	0.898	0.910	0.894	0.908
Average	0.873	0.882	0.884	0.889	0.901	0.901	0.896	0.899	0.900	0.902

	Fine-tune with Baseline on ECQA									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	1
Predict Baseline	0.495	0.522	0.528	0.553	0.550	0.550	0.554	0.569	0.561	0.562
	0.471	0.505	0.525	0.533	0.549	0.561	0.558	0.572	0.572	0.572
	0.469	0.511	0.533	0.541	0.553	0.545	0.569	0.564	0.566	0.565
Average	0.478	0.513	0.529	0.542	0.551	0.552	0.560	0.568	0.566	0.566
Predict Infusion	0.664	0.672	0.710	0.716	0.692	0.702	0.708	0.722	0.684	0.701
	0.685	0.682	0.673	0.697	0.681	0.682	0.694	0.677	0.699	0.641
	0.678	0.715	0.693	0.648	0.706	0.713	0.686	0.685	0.688	0.711
Average	0.675	0.690	0.692	0.687	0.693	0.699	0.696	0.695	0.690	0.684

	Fine-tune with Infusion on ECQA									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	1
Predict Baseline	0.417	0.406	0.402	0.395	0.381	0.379	0.365	0.379	0.375	0.374
	0.381	0.363	0.367	0.366	0.368	0.400	0.385	0.349	0.368	0.371
	0.381	0.386	0.345	0.341	0.369	0.376	0.361	0.359	0.386	0.334
Average	0.393	0.385	0.371	0.367	0.373	0.385	0.370	0.362	0.376	0.360
Predict Infusion	0.974	0.983	0.983	0.989	0.985	0.988	0.989	0.984	0.990	0.992
	0.984	0.985	0.983	0.981	0.990	0.989	0.991	0.985	0.990	0.983
	0.984	0.982	0.984	0.981	0.989	0.987	0.988	0.989	0.989	0.989
Average	0.980	0.983	0.983	0.984	0.988	0.988	0.989	0.986	0.990	0.988

Table 4: Detailed results for the preliminary experiment of rationales as partial input during fine-tuning.

Category	Premise	Hypothesis	Rationale
entailment	A young family enjoys feeling ocean waves lap at their feet.	A family is at the beach.	Ocean waves implies the beach.
	An old man with a package poses in front of an advertisement.	A man poses in front of an ad.	The word " ad " is short for the word " advertisement ".
	A man reads the paper in a bar with green lighting.	The man is inside.	In a bar means the man could be inside.
neutral	An old man with a package poses in front of an advertisement.	A man poses in front of an ad for beer.	Not all advertisements are ad for beer.
	A woman with a green headscarf, blue shirt and a very big grin.	The woman is young.	the woman could've been old rather than young
	A man reads the paper in a bar with green lighting.	The man is reading the sportspage.	The man could be reading something other than the sportspage.
contradiction	A woman with a green headscarf, blue shirt and a very big grin.	The woman has been shot.	There can be either a woman with a very big grin or a woman who has been shot.
	A man playing an electric guitar on stage.	A man playing banjo on the floor.	The man can't play on stage if he is on the floor.
	A couple walk hand in hand down a street.	A couple is sitting on a bench.	The couple cannot be walking and sitting a the same time.

Table 5: Representative examples of data with corresponding rationales for each class in e-SNLI.